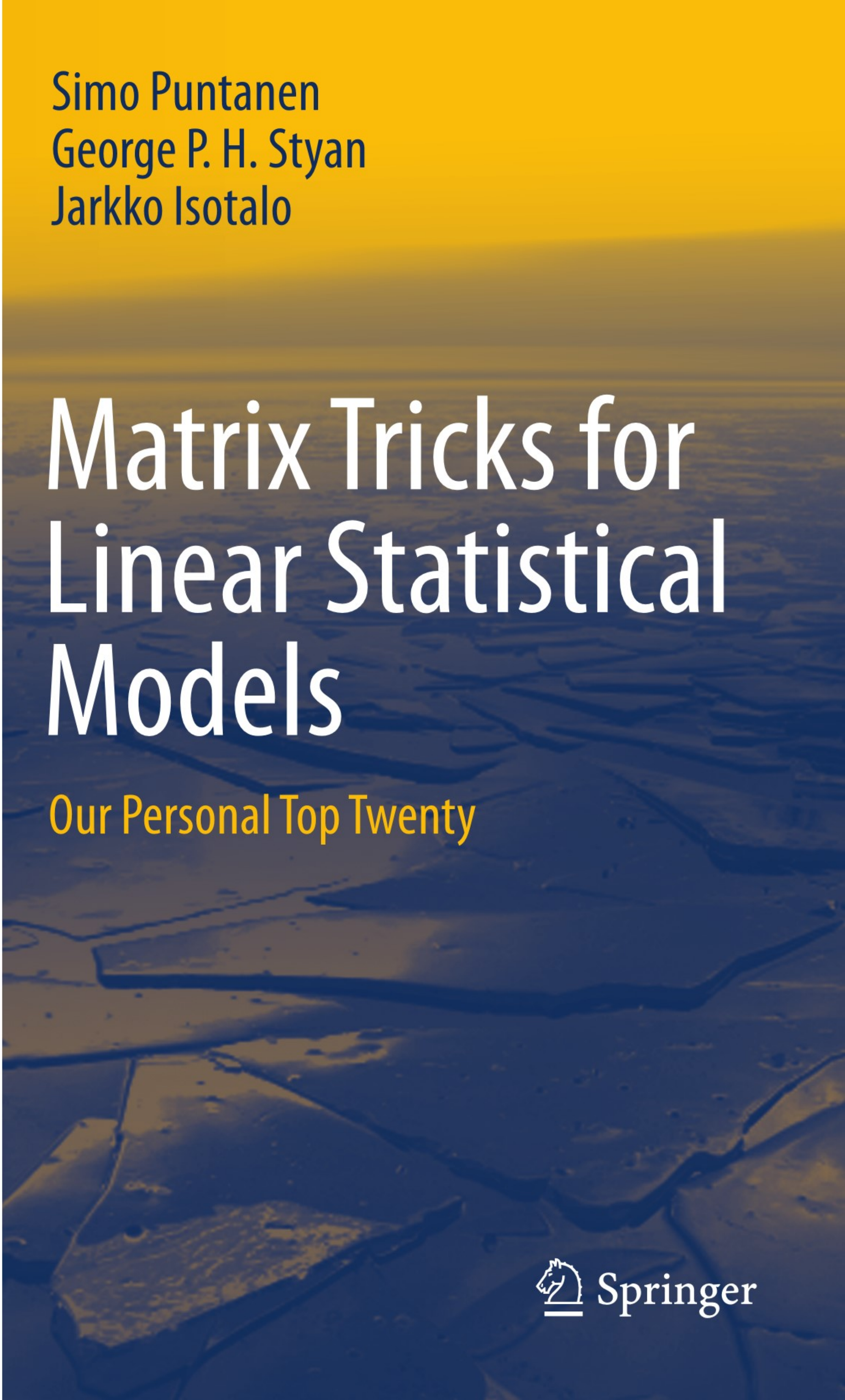


Simo Puntanen
George P. H. Styan
Jarkko Isotalo

Matrix Tricks for Linear Statistical Models

Our Personal Top Twenty

Springer

Matrix Tricks for Linear Statistical Models

Simo Puntanen
George P. H. Styan
Jarkko Isotalo

Matrix Tricks for Linear Statistical Models

Our Personal Top Twenty

Springer

Simo Puntanen
School of Information Sciences
FI-33014 University of Tampere
Finland
simo.puntanen@uta.fi

George P. H. Styan
Department of Mathematics & Statistics
McGill University
805 ouest, Sherbrooke Street West
Montréal (Québec) H3A 2K6
Canada
styan@math.mcgill.ca

Jarkko Isotalo
School of Information Sciences
FI-33014 University of Tampere
Finland
jarkko.isotalo@uta.fi

ISBN 978-3-642-10472-5 e-ISBN 978-3-642-10473-2
DOI 10.1007/978-3-642-10473-2
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011935544

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

In teaching linear statistical models to first-year graduate students or to final-year undergraduate students there is no way to proceed smoothly without matrices and related concepts of linear algebra; their use is really essential. Our experience is that making some particular matrix tricks very familiar to students can substantially increase their insight into linear statistical models (and also multivariate statistical analysis). In matrix algebra, there are handy, sometimes even very simple "tricks" which simplify and clarify the treatment of a problem—both for the student and for the professor. Of course, the concept of a *trick* is not uniquely defined—by a trick we simply mean here a useful important handy result. Notice the three adjectives we use here, useful, important, and handy, to describe the nature of the tricks to be considered.

In this book we collect together our Top Twenty favourite matrix tricks for linear statistical models. Interestingly, nobody can complain that our title is wrong; someone else could certainly write a different book with the same title.

Structure of the Book

Before presenting our Top Twenty, we offer a quick tour of the notation, linear algebraic preliminaries, data matrices, random vectors, and linear models. Browsing through these pages will familiarize the reader with our style. There may not be much in our Introduction that is new, but we feel it is extremely important to become familiar with the notation to be used. To take a concrete example, we feel that it is absolutely necessary for the reader of this book to remember that throughout we use $\mathbf{H}$ and $\mathbf{M}$, respectively, for the orthogonal projectors onto the column space of the model matrix $\mathbf{X}$ and onto its orthocomplement. A comprehensive list of symbols with explanations is given on pages 427–434.

In our book we have inserted photographs of some matricians[1] and statisticians. We have also included images of some philatelic items—for more about mathematical philatelic items we recommend the book by Wilson (2001) and the website by Miller (2010); see also Serre (2007). For "stochastic stamps"—postage stamps that are related in some way to chance, i.e., that have a connection with probability and/or statistics, see Puntanen & Styan (2008b).

We do not claim that there are many new things in this book even after the Introduction—most of the results can be found in the literature. However, we feel that there are some results that might have appeared in the literature in a somewhat concealed form and here our aim is to upgrade their appreciation—to put them into "business class".

Twenty chapters correspond to our Top Twenty tricks, one chapter for each trick. Almost every chapter includes sections which are really examples illustrating the use of the trick. Most but not all sections are statistical. Most chapters end with a set of exercises (without solutions). Each section itself, however, serves as an exercise—with solution! So if you want to develop your skills with matrices, please look at the main contents of a particular section that interests you and try to solve it on your own. This is highly recommended!

Our Tricks are not all of the same size[2] as is readily seen from the number of pages per chapter. There are some further matrix tools that might well have deserved a place in this book. For example, Lagrange multipliers and matrix derivatives, see, e.g., Ito & Kunisch (2008), Magnus & Neudecker (1999), as well as Hadamard products and Kronecker products, see, e.g., Graham (1981), Horn (1990), Styan (1973a), could have been added as four more Tricks.

Our twenty tricks are not necessarily presented in a strictly logical order in the sense that the reader must start from Chapter 1 and then proceed to Chapter 2, and so on. Indeed our book may not build an elegant mathematical apparatus by providing a construction piece by piece, with carefully presented definitions, lemmas, theorems, and corollaries. Our first three chapters have the word "easy" in their title: this is to encourage the reader to take a look at these chapters first. However, it may well be appropriate to jump directly into a particular chapter if the trick presented therein is of special interest to the reader.

---

[1] Farebrother (2000) observed that: "As the word matrician has not yet entered the English language, we may either use it to denote a member of the ruling class of a matriarchy or to denote a person who studies matrices in their medical, geological, or mathematical senses."

[2] Unlike Greenacre (2007, p. xii), who in his Preface says he "wanted each chapter to represent a fixed amount to read or teach, and there was no better way to do that than to limit the length of each chapter—each chapter is exactly eight pages in length."

Matrix Books for Statisticians

An interesting signal of the increasing importance of matrix methods for statistics is the recent publication of several books in this area—this is not to say that matrices were not used earlier—we merely wish to identify some of the many recently-published matrix-oriented books in statistics. The recent book by Seber (2008) should be mentioned in particular: it is a delightful handbook for a matrix-enthusiast-statistician. Other recent books include: Bapat (2000), Hadi (1996), Harville (1997, 2001), Healy (2000), Magnus & Neudecker (1999), Meyer (2000), A. R. Rao & Bhimasankaram (2000), C. R. Rao & M. B. Rao (1998), and Schott (2005) (first ed. 1996). Among somewhat older books we would like to mention Graybill (2002) (first ed. 1969), Seber (1980) (first ed. 1966), and Searle (1982).

There are also some other recent books using or dealing with matrix algebra which is helpful for statistics: for example, Abadir & Magnus (2005), Bernstein (2009) (first ed. 2005), Christensen (2001, 2002), Fujikoshi, Ulyanov & Shimizu (2010), Gentle (2007), Groß (2003), Härdle & Hlávka (2007), Kollo & von Rosen (2005), S. K. Mitra, Bhimasankaram & Malik (2010), Pan & Fang (2002), C. R. Rao, Toutenburg, Shalabh et al. (2008) (first ed. 1995), Seber & A. J. Lee (2003) (first ed. 1977), Sengupta & Jammalamadaka (2003), Srivastava (2002), S.-G. Wang & Chow (1994), F. Zhang (1999; 2009), and the *Handbook of Linear Algebra* by Hogben (2007).

New arrivals of books on linear statistical models are appearing at a regular pace: Khuri (2009), Monahan (2008), Rencher & Schaalje (2008) (first ed. 1999), Ryan (2009) (first ed. 1997), Stapleton (2009) (first ed. 1995), Casella (2008), and Toutenburg & Shalabh (2009), to mention a few; these last two books deal extensively with experimental design which we consider only minimally in this book. As Draper & Pukelsheim (1996, p. 1) point out, "the topic of statistical design of experiments could well have an entire encyclopedia devoted to it." In this connection we recommend the recent *Handbook of Combinatorial Designs* by Colbourn & Dinitz (2007).

There are also some books in statistics whose usefulness regarding matrix algebra has long been recognized: for example, the two classics, *An Introduction to Multivariate Statistical Analysis* by T. W. Anderson (2003) (first ed. 1958) and *Linear Statistical Inference and its Applications* by C. R. Rao (1973a) (first ed. 1965), should definitely be mentioned in this context. Both books also include excellent examples and exercises related to matrices in statistics. For generalized inverses, we would like to mention the books by C. R. Rao & S. K. Mitra (1971b) and by Ben-Israel & Greville (2003) (first ed. 1974), and the recent book by Piziak & Odell (2007). For Schur complements, see the recent book by Zhang (2005b) and the articles in this book by Puntanen & Styan (2005a,b).

The Role of Matrices in Statistics

For interesting remarks related to matrices in statistics, see the papers by Farebrother (1996, 1997), Olkin (1990, 1998), Puntanen, Seber & Styan (2007), Puntanen & Styan (2007), Searle (1999, 2000). Of special interest also are those papers in which statistical ideas are used to prove some matrix theorems, especially matrix inequalities; see, e.g., S. K. Mitra (1973a), S. K. Mitra & Puntanen (1991), Dey, Hande & Tiku (1994), and C. R. Rao (2000; 2006). As for the development of the use of matrices in statistics, we would like to refer to Searle (2005) and to the conversation by Wells (2009, p. 251), with Shayle R. Searle:

> WELLS: "You were an early advocate of using matrices in statistics, looking back this prospective seems obvious. Do you have a conjecture why early progress on the application of matrices was so slow?"
>
> SEARLE: "The first of my *Annals* papers of (1956), 1958 and 1961 was 'Matrix methods in variance and covariance components analysis'. Its title begs the question: Why has it taken so long for matrices to get widely adopted where they are so extremely useful? After all, matrices are two hundred and some years old and their use in statistics is only slowly becoming commonplace. But this was not so, even as recently as the 1950s. Even at Cambridge, in lectures on regression in 1952 there was no use of matrices."

Many thanks to Kimmo Vehkalahti for alerting us to the following interesting remarks by Bock (2007, p. 41):

> "The year 1934 and part of 1935 was a period of intense work for Thurstone. In 1935, the University of Chicago Press published *The Vectors of Mind*, his extended treatise on multiple factor analysis. [...] It also includes, for the first time in the psychological or statistical literature, an introductory section containing definitions and results of matrix algebra and their geometrical interpretations. [...] As a matter of interest, I reviewed both the *Journal of the American Statistical Association* and *The Annals of Mathematical Statistics* looking for applications of matrix algebra before 1935. Even as late as 1940, *JASA* contained not only no matrix algebra, but hardly any algebra at all; it was still largely a journal of statistics in the old sense—the presentation and analysis of tables of economic and social indicators. The earliest instance of matrix algebra I could find was in the *AMS*, Volume 6, 1935, in an article by Y. K. Wong entitled, 'Application of orthogonalization processes to the theory of least-squares'."

Our Aim

In summary: our aim is not to go through all the steps needed to develop the use of matrices in statistics. There are already several books that do that in a nice way. Our main aim is to present *our personal favourite tools* for the interested student or professor who wishes to develop matrix skills for linear models. We assume that the reader is somewhat familiar with linear algebra, matrix calculus, linear statistical models, and multivariate statistical analysis, although a thorough knowledge is not needed, one year of undergraduate study of linear algebra and statistics is expected. A short course in regression would also be welcome before going deeply into our book. Here are some examples of smooth introductions to regression: Chatterjee & Hadi (2006) (first ed. 1977), Draper & Smith (1998) (first ed. 1966), and Weisberg (2005) (first ed. 1980).

We have not included any real data or any discussion of computer software—these are beyond the scope of this book. In real life, when facing a class of statistics students, there is no way to ignore the need of modern high-calibre statistical software, allowing, in particular, access to flexible matrix manipulation. Nowadays, pen and paper are simply not enough—clever software can uncover or increase our understanding of lengthy formulas.

As regards the references, our attempt is to be rather generous—but not necessarily thorough. Speed (2009, p. 13) asks “Do you ever wonder how we found references before the www?” and follows with interesting remarks on today’s methods of finding references. We recommend *Current Index to Statistics*, *MathSciNet* and *ZMATH*, as well as *Google*!

The material of this book has been used in teaching statistics students at the University of Tampere for more than 10 years. Warm thanks to all those students for inspiring cooperation! The idea has been to encourage students to develop their skills by emphasizing the tricks that we have learnt to be used again and again. Our belief is that it is like practicing a sport: practice makes perfect.

One possible way to use this book as a textbook for a course might be just to go through the exercises in this book. Students should be encouraged to solve and discuss them on the blackboard without any notes. The idea would be to push the student into the waves of individual thinking.

Kiitos!

We are most grateful to Theodore W. Anderson for introducing matrix methods for statistics to the second author in the early 1960s, and for supervising his Ph.D. thesis (Styan, 1969) at Columbia University; this then led to the Ph.D. theses at the University of Tampere by the first author (Puntanen, 1987) and by the third author (Isotalo, 2007).

Much of the work reported in this book follows the extensive collaboration by the first authors with Jerzy K. Baksalary (1944–2005) during the past 30 years or so; we feel he had an enormous effect on us, and if he were still alive, he would certainly be very interested (and critical as always) in this book. Some of our joint adventures in the column and row spaces are described in Puntanen & Styan (2008a) and in Isotalo, Puntanen & Styan (2008b). For an appreciation of Jerzy by many scholars and for a complete bibliography of his publications see Oskar Maria Baksalary & Styan (2007).

Sincere thanks go also to Oskar Maria Baksalary, Ka Lok Chu, Maria Ruíz Delmàs, S. W. Drury, Katarzyna Filipiak, Stephen Haslett, Augustyn Markiewicz, Ville Puntanen, Daniel J.-H. Rosenthal, Evelyn Matheson Styan, Götz Trenkler, and Kimmo Vehkalahti, and to the anonymous reviewers for their help. Needless to say that ...[3]

We give special thanks to Jarmo Niemelä for his outstanding help in setting up the many versions of this book in LaTeX. The figures for scatter plots were prepared using Survo software, online at http://www.survo.fi, (again thanks go to Kimmo Vehkalahti) and the other figures using PSTricks (again thanks go to Jarmo Niemelä).

We are most grateful to John Kimmel for suggesting that we write this monograph, and to Niels Peter Thomas and Lilith Braun of Springer for advice and encouragement.

Almost all photographs were taken by Simo Puntanen and are based on his collection (Puntanen, 2010a); exceptions are the photographs of Gene H. Golub and Shayle R. Searle, which were taken by George P. H. Styan and Harold V. Henderson, respectively. The photograph of the three authors with their mentor (p. xi) was taken by Soile Puntanen. The photographs of the two Markovs are taken from Wikipedia and the photograph of a Tappara defenceman is based on a hockey card. The images of philatelic items are based on items in the collection of George P. H. Styan (2010); we owe our gratitude also to Somesh Das Gupta (1935–2006) for providing us with the first-day cover for Mahalanobis (p. 90). *Scott* catalogue numbers are as given in the *Scott Standard Postage Stamp Catalogue* (Kloetzel, 2010).

The website http://mtl.uta.fi/matrixtricks supports the book by additional material.

The research for this monograph was supported in part by the Natural Sciences and Engineering Research Council of Canada.

Dedication

To Soile and Evelyn
and
to Theodore W. Anderson—the academic grandfather, father and great grandfather, respectively, of the three authors.

SP, GPHS & JI
11 April 2011

MSC 2000: 15-01, 15-02, 15A09, 15A42, 15A99, 62H12, 62J05.

---

[3] cf. Flury (1997, p. ix).

Key words and phrases: Best linear unbiased estimation, Cauchy–Schwarz inequality, column space, eigenvalue decomposition, estimability, Gauss–Markov model, generalized inverse, idempotent matrix, linear model, linear regression, Löwner ordering, matrix inequalities, oblique projector, ordinary least squares, orthogonal projector, partitioned linear model, partitioned matrix, rank cancellation rule, reduced linear model, Schur complement, singular value decomposition.

Photograph 0.1 (From left to right) Jarkko Isotalo, Simo Puntanen, George P. H. Styan and Theodore W. Anderson, during the 15th International Workshop on Matrices and Statistics in Uppsala, Sweden: 15 June 2006.

Contents

Introduction

> Algernon (Languidly): *I don't know that I am much interested in your family life, Lane.*
> Lane: *No, sir; it is not a very interesting subject. I never think of it myself.*
>
> Oscar Wilde: *The Importance of Being Earnest*

Opening Figures

Let us start by considering the three scatter plots of Figures 0.1 and 0.2. In Figure 0.1a we have three data points $\binom{x_1}{y_1}$, $\binom{x_2}{y_2}$, $\binom{x_3}{y_3}$, while in Figure 0.2 there are 1000 data points. Suppose that we have two tasks:

- Task 1: Draw a line $\hat{y} = \alpha + \beta x$ into each scatter plot so that the sum of the squared vertical distances $(y_i - \hat{y}_i)^2$ would be as small as possible.
- Task 2: Define the line $\tilde{y} = \gamma + \delta x$ so that the sum of squares of the orthogonal distances of each point (x_i, y_i) from this line would be minimal.

These tasks are typical things that every statistics student should be able to do in no time. However, it is extremely clumsy to solve these problems without knowing some appropriate matrix algebra. If you the reader are not convinced about the usefulness of matrices, try to solve these problems by other means—that should finally convince you that the life becomes more comfortable after familiarizing yourself with handy matrix tricks.

The resulting lines, the regression line and the first major axis of a particular a confidence ellipse, are represented in Figures 0.8–0.9 (p. 56). By the confidence ellipse we mean here a constant-density contour assuming that this sample were taken from a normally distributed population.

The first two scatterplots are based on the following data matrices:

$$\mathbf{A} = (\mathbf{x} : \mathbf{y}) = \begin{pmatrix} 1 & 1 \\ 4 & 1 \\ 4 & 4 \end{pmatrix}, \tag{0.1}$$

$$\mathbf{B}' = (\mathbf{x} : \mathbf{y})' = \begin{pmatrix} \mathbf{x}' \\ \mathbf{y}' \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 2 & 2 & 2 & 3 & 3 & 3 & 4 & 4 & 4 \\ 2 & 3 & 4 & 1 & 3 & 4 & 1 & 2 & 4 & 1 & 2 & 3 \end{pmatrix}. \tag{0.2}$$

Above $\mathbf{A} = (\mathbf{x} : \mathbf{y})$ stands for a partitioned matrix; the columns of $\mathbf{A}$ are vectors $\mathbf{x}$ and $\mathbf{y}$. The matrix $\mathbf{B}$ is represented in the transposed form: the ith

S. Puntanen et al., *Matrix Tricks for Linear Statistical Models: Our Personal Top Twenty*,
DOI 10.1007/978-3-642-10473-2_1,

column of $\mathbf{B}'$ is the transpose of the ith row of $\mathbf{B}$. The third data matrix is a 1000×2 matrix including 1000 observations from a two-dimensional normal distribution.

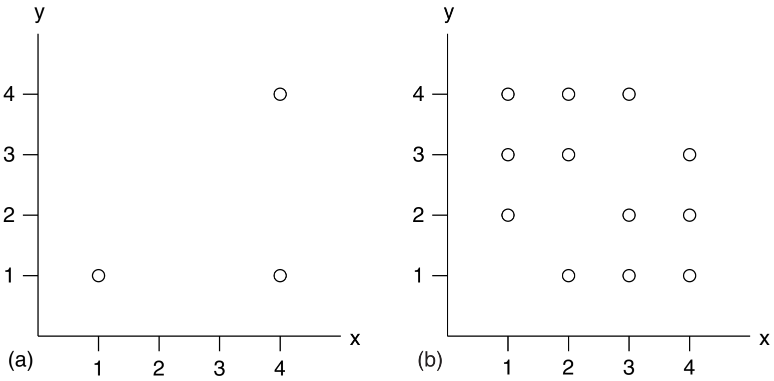


Figure 0.1 (a) Three observations, (b) Twelve observations.

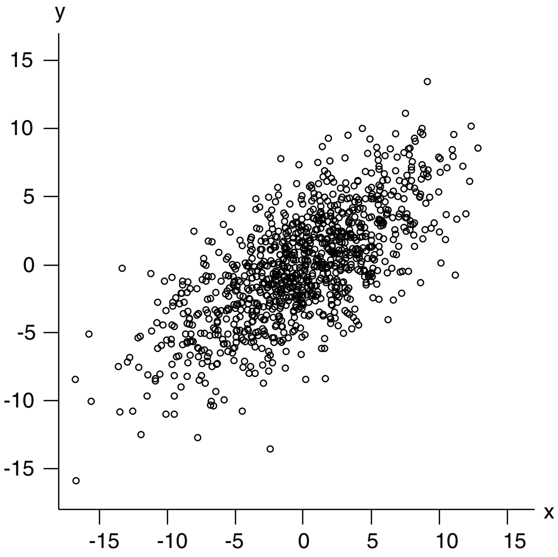


Figure 0.2 1000 observations from $\mathrm{N}_2(\mathbf{0}, \mathbf{\Sigma})$; $\sigma_x = 5$, $\sigma_y = 4$, $\varrho_{xy} = 0.7$.

Preliminaries to Linear Algebra and Statistics

First, let us shortly introduce some notation. A more comprehensive list of symbols with explanations is given on pages 427–434.

The symbols $\mathbf{A}'$, $\mathbf{A}^-$, $\mathbf{A}^+$, $\mathscr{C}(\mathbf{A})$, $\mathscr{C}(\mathbf{A})^\perp$, $\mathscr{N}(\mathbf{A})$, and $\operatorname{rank}(\mathbf{A})$ [or shortly $\operatorname{rk}(\mathbf{A})$] will stand for the transpose, a generalized inverse, the Moore–Penrose inverse, the column space, the orthogonal complement of the column space, the null space, and the rank, respectively, of $\mathbf{A}$. The set of $n \times m$ matrices is denoted as $\mathbb{R}^{n\times m}$. Occasionally we may denote $\mathbf{A}_{n\times m}$ indicating that $\mathbf{A}$ is an $n \times m$ matrix. We only consider matrices with real elements.

It may be worth mentioning that we frequently use the symbol n to indicate the number of rows in a matrix. This is a trivial fact but it may helpful to keep this in mind.

Vectors belonging to $\mathbb{R}^n$ are denoted by lower-case bold-face Roman letters like $\mathbf{a}$, $\mathbf{b}$ and the matrices belonging to $\mathbb{R}^{n\times m}$ in upper-case bold-face Roman letters like $\mathbf{A}$, $\mathbf{B}$. Random vectors are denoted in the same way as the vectors in $\mathbb{R}^n$; it is too complicated to devote a particular notation for them. When dealing with random variables and vectors, our attempt is to use symbols $\mathbf{x}, \mathbf{y}, \dots$ for the random vectors and $\mathbf{a}, \mathbf{b}, \dots$ for vectors with real elements. This rule will have several exceptions but from the context it should be clear whether we are dealing with a random or a nonrandom vector.

Column Space

The *column space* of an $n \times m$ matrix $\mathbf{A}$ is a subspace of $\mathbb{R}^n$ spanned by the columns of $\mathbf{A}$:

$$\mathscr{C}(\mathbf{A}_{n\times m}) = \{\, \mathbf{y} \in \mathbb{R}^n : \text{there exists } \mathbf{x} \in \mathbb{R}^m \text{ such that } \mathbf{y} = \mathbf{A}\mathbf{x} \,\}, \tag{0.3}$$

and, correspondingly, the *null space* of $\mathbf{A}$ is

$$\mathscr{N}(\mathbf{A}_{n\times m}) = \{\, \mathbf{x} \in \mathbb{R}^m : \mathbf{A}\mathbf{x} = \mathbf{0} \,\} \subset \mathbb{R}^m. \tag{0.4}$$

For example, if $\mathbf{1}_2 = \binom{1}{1}$, then $\mathscr{C}(\mathbf{1}_2)$ is a line in $\mathbb{R}^2$ going through the origin and the "point" $(1, 1)$, i.e., it is a line spanned by the vector $\mathbf{1}_2$. Notice that we will drop off the subscript from $\mathbf{1}_n \in \mathbb{R}^n$ if the dimension is obvious from the context; the same concerns also the identity matrix $\mathbf{I}_n$.

The columns of $\mathbf{A}_{n\times m} = (\mathbf{a}_1 : \dots : \mathbf{a}_m)$ are said to be linearly independent if (and only if) $\mathbf{A}\mathbf{x} = \mathbf{0}$ is satisfied only for the null vector $\mathbf{x} = \mathbf{0}$. In this situation we say that $\mathbf{A}$ has full column rank, i.e.,

$$\begin{aligned} \mathscr{N}(\mathbf{A}) = \{\mathbf{0}\} &\iff \mathbf{A} \text{ has full column rank} \\ &\iff \text{the columns of } \mathbf{A} \text{ are linearly independent.} \end{aligned} \tag{0.5}$$

The rank of the matrix $\mathbf{A}$ is the maximal number of linearly independent columns of $\mathbf{A}$, and the dimension of the column space $\mathscr{C}(\mathbf{A})$ is $\operatorname{rank}(\mathbf{A})$.

The Chapter 1 of this book, entitled Easy Column Space Tricks (pp. 57–70), is devoted to some essential properties of the column space. Without (most of) them, it is almost impossible to play with the matrices related to linear models—or at least it becomes clumsy. For example, in regression analysis we frequently need to know the rank of the matrix $\mathbf{X}'\mathbf{X}$. Then it really should be obvious[4] to us that this rank is precisely the same as the rank of $\mathbf{X}$; otherwise things proceed far too slowly, so to say.

Generalized Inverse

Hopefully the reader has met the generalized inverses before reading these lines, but if not, it is not the end of the world. Here we take a very quick look at them; more thoroughly we go through them in Chapter 4 (pp. 105–120).

The matrix $\mathbf{G}_{m\times n}$ is a *generalized inverse* of $\mathbf{A}_{n\times m}$ if it satisfies the equation

$$\mathbf{AGA} = \mathbf{A}\,, \tag{mp1}$$

and we denote $\mathbf{A}^- = \mathbf{G}$. An equivalent characterization (see Theorem 4, p. 106) is that $\mathbf{G}$ is a generalized inverse of $\mathbf{A}$ if $\mathbf{b} = \mathbf{Gy}$ is a solution of $\mathbf{Ab} = \mathbf{y}$ for any $\mathbf{y}$ which makes the equation consistent (solvable). The matrix $\mathbf{A}^+$, the *Moore–Penrose inverse* of $\mathbf{A}$, is defined as the unique solution to the four equations

$$\begin{aligned} &\text{(mp1)}\ \ \mathbf{AA}^+\mathbf{A} = \mathbf{A}\,, && \text{(mp2)}\ \ \mathbf{A}^+\mathbf{AA}^+ = \mathbf{A}^+, \\ &\text{(mp3)}\ \ \mathbf{AA}^+ = (\mathbf{AA}^+)', && \text{(mp4)}\ \ \mathbf{A}^+\mathbf{A} = (\mathbf{A}^+\mathbf{A})'. \end{aligned} \tag{0.6}$$

The set of all generalized inverses of $\mathbf{A}_{n\times m}$ is denoted as

$$\{\mathbf{A}^-\} = \{\, \mathbf{G} \in \mathbb{R}^{m\times n} : \mathbf{AGA} = \mathbf{A}\,\}\,. \tag{0.7}$$

For example, if $\mathbf{A}_{n\times m}$ is such that $(\mathbf{A}'\mathbf{A})^{-1}$ exists, then it is easy to confirm that

$$(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}' = \mathbf{A}^+. \tag{0.8}$$

It is also useful to know that always

$$(\mathbf{A}'\mathbf{A})^+\mathbf{A}' = \mathbf{A}'(\mathbf{AA}')^+ = \mathbf{A}^+, \quad \mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{A}' = \mathbf{AA}^+. \tag{0.9}$$

---

[4] According to an old (and possibly apocryphal) story, during one of his lectures David Hilbert once wrote a line on the blackboard and said, “It is obvious that . . . ,” but then Hilbert paused and thought for a moment. He then became noticeably perplexed, and he even left the room, returning only after an awkward passage of time. When Hilbert resumed his lecture, he began by saying “It is obvious that . . . ”. [Cited from Steele (2004, pp. 55–56).]

Photograph 0.2 Adi Ben-Israel (Windsor, Canada, 2007).

If the equation $\mathbf{Ab} = \mathbf{y}$ is solvable for $\mathbf{b}$ and $\mathbf{G}$ is a generalized inverse of $\mathbf{A}$, then $\mathbf{b} = \mathbf{Gy}$ is one solution. All solutions (the "general solution") to $\mathbf{Ab} = \mathbf{y}$ can be generated through

$$\mathbf{b}_0 = \mathbf{Gy} + (\mathbf{I}_m - \mathbf{GA})\mathbf{z}\,, \tag{0.10}$$

where $\mathbf{z} \in \mathbb{R}^m$ is free to vary. The other way to express $\mathbf{b}_0$ is to write it as

$$\begin{aligned}\mathbf{b}_0 &= \{\text{one solution to } \mathbf{Ab} = \mathbf{y}\} \\ &+ \{\text{the general solution to } \mathbf{Ab} = \mathbf{0}\}\,. \end{aligned}\tag{0.11}$$

The equation $\mathbf{A}_{n\times m}\mathbf{B}_{m\times q} = \mathbf{Y}_{n\times q}$ has a solution for $\mathbf{B}$ if $\mathscr{C}(\mathbf{Y}) \subset \mathscr{C}(\mathbf{A})$ in which case all solutions can be obtained from

$$\mathbf{B}_0 = \mathbf{GY} + (\mathbf{I}_m - \mathbf{GA})\mathbf{Z}\,, \tag{0.12}$$

where $\mathbf{Z} \in \mathbb{R}^{m\times q}$ is free to vary.

As for the terminology, Ben-Israel & Greville (2003) call a matrix $\mathbf{G}$ satisfying (mp1) a $\{1\}$-inverse, and $\mathbf{G}$ satisfying (mp1) and (mp2) a $\{12\}$-inverse. A $\{12\}$-inverse is often called a reflexive generalized inverse.

Projectors

If the column spaces $\mathscr{C}(\mathbf{A})$ and $\mathscr{C}(\mathbf{B})$ are disjoint (sometimes said essentially disjoint) in the sense that

$$\mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B}) = \{\mathbf{0}\}\,, \tag{0.13}$$

then we have the direct sum decomposition

$$\mathscr{C}(\mathbf{A} : \mathbf{B}) = \mathscr{C}(\mathbf{A}) \oplus \mathscr{C}(\mathbf{B})\,. \tag{0.14}$$

In particular, if $\mathscr{C}(\mathbf{A})$ and $\mathscr{C}(\mathbf{B})$ in (0.14) are orthogonal [cf. (0.35) p. 10], we denote $\mathscr{C}(\mathbf{A} : \mathbf{B}) = \mathscr{C}(\mathbf{A}) \boxplus \mathscr{C}(\mathbf{B})$. In (0.14), $\mathscr{C}(\mathbf{B})$ is a complement of $\mathscr{C}(\mathbf{A})$ in $\mathscr{C}(\mathbf{A} : \mathbf{B})$, and then every $\mathbf{y} \in \mathscr{C}(\mathbf{A} : \mathbf{B})$ has a unique representation as a sum

$$\mathbf{y} = \mathbf{x} + \mathbf{z}\,, \quad \text{where } \mathbf{x} \in \mathscr{C}(\mathbf{A}),\ \mathbf{z} \in \mathscr{C}(\mathbf{B})\,. \tag{0.15}$$

If $\mathscr{C}(\mathbf{B})$ is a complement of $\mathscr{C}(\mathbf{A})$ in the *whole* $\mathbb{R}^n$, not only in $\mathscr{C}(\mathbf{A} : \mathbf{B})$, i.e.,

$$\mathbb{R}^n = \mathscr{C}(\mathbf{A}) \oplus \mathscr{C}(\mathbf{B})\,, \tag{0.16}$$

then every $\mathbf{y} \in \mathbb{R}^n$ has a unique representation as a sum

$$\mathbf{y} = \mathbf{y}_A + \mathbf{y}_B \,, \quad \text{where } \mathbf{y}_A \in \mathscr{C}(\mathbf{A}),\ \mathbf{y}_B \in \mathscr{C}(\mathbf{B}) \,. \tag{0.17}$$

Photograph 0.3 C. Radhakrishna Rao (Istanbul, 1997).

The vector $\mathbf{y}_A$ in (0.17) is said to be the *projection* of $\mathbf{y}$ onto $\mathscr{C}(\mathbf{A})$ along $\mathscr{C}(\mathbf{B})$. A matrix $\mathbf{P}$ which transforms every $\mathbf{y} \in \mathbb{R}^n$ into its projection is called a *projector* onto $\mathscr{C}(\mathbf{A})$ along $\mathscr{C}(\mathbf{B})$; that is, the multiplication $\mathbf{Py}$ gives the projection of $\mathbf{y}$:

$$\mathbf{Py} = \mathbf{y}_A \,. \tag{0.18}$$

It appears that the projector $\mathbf{P} := \mathbf{P}_{\mathbf{A}|\mathbf{B}}$ onto $\mathscr{C}(\mathbf{A})$ along $\mathscr{C}(\mathbf{B})$ may be defined by the equation

$$\mathbf{P}_{\mathbf{A}|\mathbf{B}}(\mathbf{A} : \mathbf{B}) = (\mathbf{A} : \mathbf{0}) \,. \tag{0.19}$$

Once the decomposition $\mathbb{R}^n = \mathscr{C}(\mathbf{A}) \oplus \mathscr{C}(\mathbf{B})$ is fixed, then the corresponding projector is unique.

Rao (1974) considered also an extended version of projector in the sense that $\mathscr{C}(\mathbf{A}) \oplus \mathscr{C}(\mathbf{B})$ does not coincide with the entire $\mathbb{R}^n$; even then (0.19) defines the projector $\mathbf{P}_{\mathbf{A}|\mathbf{B}}$ onto $\mathscr{C}(\mathbf{A})$ along $\mathscr{C}(\mathbf{B})$. However, in this situation the projector $\mathbf{P}_{\mathbf{A}|\mathbf{B}}$ need not be unique and idempotent as is the case when (0.16) holds. We denote the set of matrices $\mathbf{G}$ satisfying $\mathbf{G}(\mathbf{A} : \mathbf{B}) = (\mathbf{A} : \mathbf{0})$ as $\{\mathbf{P}_{\mathbf{A}|\mathbf{B}}\}$:

$$\mathbf{G} \in \{\mathbf{P}_{\mathbf{A}|\mathbf{B}}\} \iff \mathbf{G}(\mathbf{A} : \mathbf{B}) = (\mathbf{A} : \mathbf{0}) \,. \tag{0.20}$$

Projectors belong to the set of very important tools in linear models. Their basic properties are gone through in Chapter 2 (pp. 71–89).

Inner Product and Norm

The *inner product* and the *norm* are central concepts in linear algebra—as well as their statistical "buddies" covariance and variance. The standard inner product between two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ is defined as

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}'\mathbf{y} = \sum_{i=1}^{n} x_i y_i \,, \tag{0.21}$$

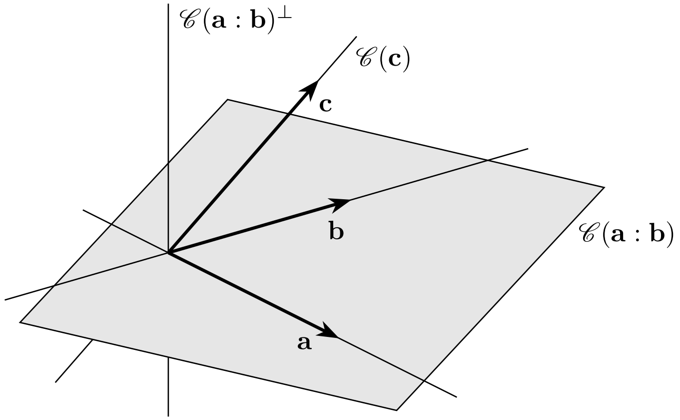


Figure 0.3 In this figure $\mathbb{R}^3 = \mathscr{C}(\mathbf{a}:\mathbf{b}) \oplus \mathscr{C}(\mathbf{c})$, $\mathbb{R}^3 = \mathscr{C}(\mathbf{a}:\mathbf{b}) \boxplus \mathscr{C}(\mathbf{a}:\mathbf{b})^\perp$.

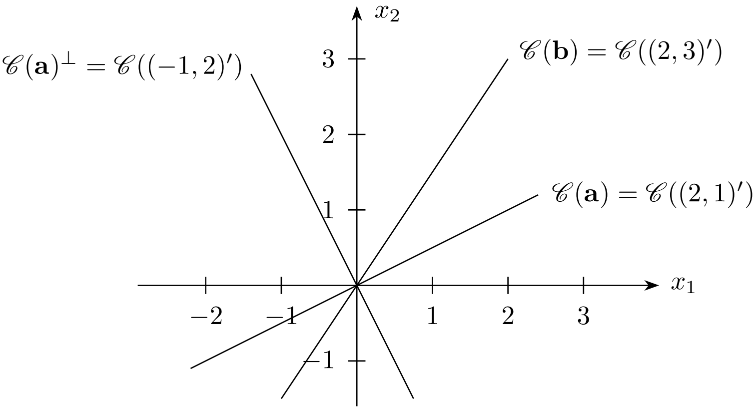


Figure 0.4 Here $\mathbb{R}^2 = \mathscr{C}(\mathbf{a}) \oplus \mathscr{C}(\mathbf{b})$, $\mathbb{R}^2 = \mathscr{C}(\mathbf{a}) \boxplus \mathscr{C}(\mathbf{a})^\perp$.

which is also called the Euclidean inner product in $\mathbb{R}^n$. The corresponding standard norm (length, 2-norm) in $\mathbb{R}^n$ is the positive square root of $\langle \mathbf{x}, \mathbf{x} \rangle$:

$$\|\mathbf{x}\| = \|\mathbf{x}\|_2 = +\sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} = +\sqrt{x_1^2 + \cdots + x_n^2}\,. \tag{0.22}$$

In general, a real-valued function $\langle \mathbf{x}, \mathbf{y} \rangle$ is an inner product in $\mathbb{R}^n$ if it satisfies the following three properties:

$$\langle \mathbf{x}, \mathbf{x} \rangle \geq 0 \text{ and the equality holds } \iff \mathbf{x} = \mathbf{0} \quad \text{(positivity)}, \tag{0.23a}$$

$$\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle \quad \text{(symmetry)}, \tag{0.23b}$$

$$\langle \alpha\mathbf{x} + \mathbf{y}, \mathbf{z} \rangle = \alpha\langle \mathbf{x}, \mathbf{z} \rangle + \langle \mathbf{y}, \mathbf{z} \rangle \quad \text{(linearity)}, \tag{0.23c}$$

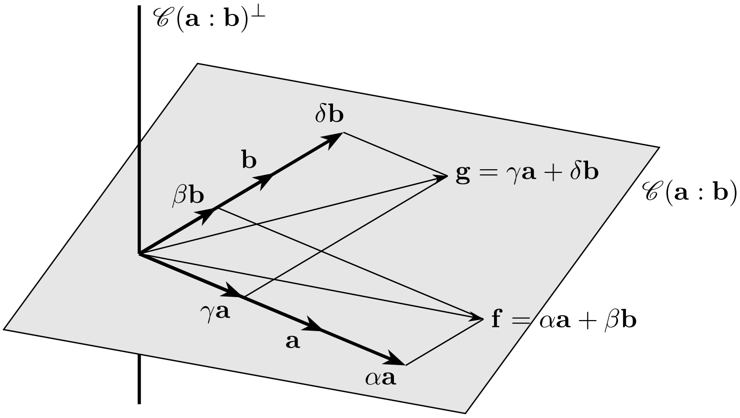


Figure 0.5 Geometric illustration of $\mathbf{f}=\alpha\mathbf{a}+\beta\mathbf{b}$ and $\mathbf{g}=\gamma\mathbf{a}+\delta\mathbf{b}$, $\alpha>1$, $\beta<1$, $\gamma<1$, $\delta>1$. Also the orthocomplement $\mathscr{C}(\mathbf{a}:\mathbf{b})^\perp$ is marked.

for all $\mathbf{x},\mathbf{y},\mathbf{z}\in\mathbb{R}^n$, and $\alpha\in\mathbb{R}$. If (0.23a) is replaced by the weaker condition $\langle\mathbf{x},\mathbf{x}\rangle\geq 0$, we then have a *semi-inner product*; now it is possible that $\langle\mathbf{x},\mathbf{x}\rangle=0$ but $\mathbf{x}\neq\mathbf{0}$.

Correspondingly, a general vector norm in $\mathbb{R}^n$ is defined as a function $\mathbb{R}^n\to\mathbb{R}$, denoted by $\|\mathbf{x}\|$, that satisfies the following three conditions:

$$\|\mathbf{x}\|\geq 0 \text{ and the equality holds } \iff \mathbf{x}=\mathbf{0} \quad \text{(positivity)}, \tag{0.24a}$$

$$\|\alpha\mathbf{x}\|=|\alpha|\|\mathbf{x}\| \quad \text{(positive homogeneity)}, \tag{0.24b}$$

$$\|\mathbf{x}+\mathbf{y}\|\leq\|\mathbf{x}\|+\|\mathbf{y}\| \quad \text{(triangle inequality)}, \tag{0.24c}$$

for all $\mathbf{x},\mathbf{y}\in\mathbb{R}^n$, and $\alpha\in\mathbb{R}$. If (0.24a) is replaced by $\|\mathbf{x}\|\geq 0$ for all $\mathbf{x}\in\mathbb{R}^n$, the norm may be called a *seminorm*.

Given an inner product $\langle\mathbf{x},\mathbf{y}\rangle$ and the corresponding norm $\|\mathbf{x}\|=\sqrt{\langle\mathbf{x},\mathbf{x}\rangle}$, the angle between two nonzero vectors $\mathbf{x},\mathbf{y}\in\mathbb{R}^n$ is the real number θ, $0\leq\theta\leq\pi$, such that

$$\cos\theta=\frac{\langle\mathbf{x},\mathbf{y}\rangle}{\|\mathbf{x}\|\cdot\|\mathbf{y}\|}\,. \tag{0.25a}$$

Instead of $\cos\angle(\mathbf{x},\mathbf{y})$, we will frequently use the shorter notation $\cos(\mathbf{x},\mathbf{y})$ to indicate the cosine of the angle between the nonzero vectors $\mathbf{x}$, $\mathbf{y}$:

$$\cos(\mathbf{x},\mathbf{y})=\cos\angle(\mathbf{x},\mathbf{y})=\cos\theta=\frac{\langle\mathbf{x},\mathbf{y}\rangle}{\|\mathbf{x}\|\cdot\|\mathbf{y}\|}\,. \tag{0.25b}$$

It is important to notice that the angle is dependent on the given inner product.

Instead of $\mathbb{R}^n$, consider now the vector space $\mathbb{R}^{n\times m}$, that is, the set of $n\times m$ real matrices. Let $\mathbf{A},\mathbf{B}\in\mathbb{R}^{n\times m}$ and denote

$$\operatorname{vec}(\mathbf{A}) = \operatorname{vec}(\mathbf{a}_1 : \ldots : \mathbf{a}_m) = \begin{pmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_m \end{pmatrix} . \tag{0.26}$$

Then the standard inner product of $\operatorname{vec}(\mathbf{A}) \in \mathbb{R}^{nm}$ and $\operatorname{vec}(\mathbf{B}) \in \mathbb{R}^{nm}$ is

$$\langle \operatorname{vec}(\mathbf{A}), \operatorname{vec}(\mathbf{B}) \rangle = \mathbf{a}_1'\mathbf{b}_1 + \cdots + \mathbf{a}_m'\mathbf{b}_m = \operatorname{tr}(\mathbf{A}'\mathbf{B}) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij} b_{ij} \, , \tag{0.27}$$

where $\operatorname{tr}(\mathbf{U})$ refers to the trace of the matrix $\mathbf{U}$:

$$\operatorname{tr}(\mathbf{U}_{n\times n}) = \sum_{i=1}^{n} u_{ii} \, . \tag{0.28}$$

Hence a natural way to define a matrix inner product is

$$\langle \mathbf{A}, \mathbf{B} \rangle = \operatorname{tr}(\mathbf{A}'\mathbf{B}) \, , \tag{0.29}$$

which is called the standard inner product for matrices. The corresponding standard (Euclidean) matrix norm is the positive square root of $\langle \mathbf{A}, \mathbf{A} \rangle$:

$$\|\mathbf{A}\|_F = +\sqrt{\operatorname{tr}(\mathbf{A}'\mathbf{A})} = \Big[\sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij}^2 \Big]^{1/2} , \tag{0.30}$$

where the subscript F refers to Frobenius; the standard matrix norm (0.30) is often called the *Frobenius norm*. Because the trace of $\mathbf{A}'\mathbf{A}$ equals the sum of the eigenvalues of $\mathbf{A}'\mathbf{A}$, see (18.5) (p. 358), we have

$$\|\mathbf{A}\|_F = \sqrt{\operatorname{ch}_1(\mathbf{A}'\mathbf{A}) + \cdots + \operatorname{ch}_m(\mathbf{A}'\mathbf{A})} \, , \tag{0.31}$$

where $\operatorname{ch}_i(\mathbf{A}'\mathbf{A})$ refers to the ith largest eigenvalue of $\mathbf{A}'\mathbf{A}$.

In passing we may refresh the reader's memory about the useful commutativity property of the trace: $\operatorname{tr}(\mathbf{K}_{n\times p}\mathbf{L}_{p\times n}) = \operatorname{tr}(\mathbf{L}_{p\times n}\mathbf{K}_{n\times p})$.

The *matrix 2-norm* (or the spectral norm) is defined as follows:

$$\begin{aligned} \|\mathbf{A}\|_2 = \max_{\|\mathbf{x}\|_2 = 1} \|\mathbf{A}\mathbf{x}\|_2 &= \max_{\mathbf{x} \neq \mathbf{0}} \bigg(\frac{\mathbf{x}'\mathbf{A}'\mathbf{A}\mathbf{x}}{\mathbf{x}'\mathbf{x}} \bigg)^{1/2} \\ &= \sqrt{\operatorname{ch}_1(\mathbf{A}'\mathbf{A})} = \operatorname{sg}_1(\mathbf{A}) \, , \end{aligned} \tag{0.32}$$

where $\|\mathbf{x}\|_2$ refers to the standard Euclidean vector norm, and

$$\operatorname{sg}_i(\mathbf{A}) = \sqrt{\operatorname{ch}_i(\mathbf{A}'\mathbf{A})} = \delta_i = \text{the } i\text{th largest singular value of } \mathbf{A} \, . \tag{0.33}$$

Note that the Frobenius norm and the spectral norm are (usually) indeed different:

$$\|\mathbf{A}\|_F = \sqrt{\delta_1^2 + \cdots + \delta_r^2}\,, \quad \|\mathbf{A}\|_2 = \delta_1\,, \tag{0.34}$$

where $r = \operatorname{rank}(\mathbf{A})$. For a definition of the general matrix norm, see, e.g., Ben-Israel & Greville (2003, p. 13) and Meyer (2000, §5.2).

Vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ are *orthogonal* if

$$\langle \mathbf{x}, \mathbf{y} \rangle = 0 \quad \text{which is denoted as } \mathbf{x} \perp \mathbf{y}\,. \tag{0.35}$$

It is important to pay attention to the fact that orthogonality depends on the given inner product. If $\mathcal{U}$ is a subspace of $\mathbb{R}^n$, the *orthocomplement* $\mathcal{U}^\perp$ of $\mathcal{U}$ is the set of all vectors of $\mathbb{R}^n$ which are orthogonal to every vector of $\mathcal{U}$:

$$\mathcal{U}^\perp = \{\, \mathbf{z} \in \mathbb{R}^n : \langle \mathbf{z}, \mathbf{x} \rangle = 0 \text{ for all } \mathbf{x} \in \mathcal{U} \,\}\,. \tag{0.36}$$

Again the subspace $\mathcal{U}^\perp$ depends on the given inner product. It is easy to conclude, see Theorem 2 (p. 71), that under the standard inner product in $\mathbb{R}^n$, the orthocomplement of $\mathscr{C}(\mathbf{A}_{n\times m})$ is

$$\begin{aligned} \mathscr{C}(\mathbf{A})^\perp &= \{\, \mathbf{y} \in \mathbb{R}^n : \mathbf{z}'\mathbf{A}'\mathbf{y} = 0 \text{ for all } \mathbf{z} \in \mathbb{R}^m \,\} \\ &= \{\, \mathbf{y} \in \mathbb{R}^n : \mathbf{A}'\mathbf{y} = \mathbf{0} \,\} = \mathscr{N}(\mathbf{A}')\,. \end{aligned} \tag{0.37}$$

The columns of matrix $\mathbf{A} = (\mathbf{x} : \mathbf{y}) \in \mathbb{R}^{n\times 2}$ are said to be orthonormal if $\mathbf{x}'\mathbf{y} = 0$ and they have unit length, i.e., $\mathbf{A}'\mathbf{A} = \mathbf{I}_2$. We say that a matrix $\mathbf{A}$ is orthogonal if $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{A}'\mathbf{A} = \mathbf{I}_n$ (in which case also $\mathbf{A}\mathbf{A}' = \mathbf{I}_n$).

Orthogonal Projector

Let $\mathbf{P}_{n\times n}$ be an idempotent matrix, i.e., $\mathbf{P}^2 = \mathbf{P}$. Then

$$\mathscr{C}(\mathbf{P}) \cap \mathscr{C}(\mathbf{I}_n - \mathbf{P}) = \{\mathbf{0}\}\,, \tag{0.38}$$

and $\mathscr{C}(\mathbf{I}_n - \mathbf{P}) = \mathscr{N}(\mathbf{P})$ (please confirm), and hence

$$\mathbb{R}^n = \mathscr{C}(\mathbf{P}) \oplus \mathscr{C}(\mathbf{I}_n - \mathbf{P}) = \mathscr{C}(\mathbf{P}) \oplus \mathscr{N}(\mathbf{P})\,. \tag{0.39}$$

The column spaces $\mathscr{C}(\mathbf{P})$ and $\mathscr{C}(\mathbf{I}_n - \mathbf{P})$ may not be orthogonal with respect to a given inner product, but if they are, we may denote

$$\mathscr{C}(\mathbf{P}) \perp \mathscr{C}(\mathbf{I}_n - \mathbf{P})\,, \tag{0.40}$$

and then $\mathscr{C}(\mathbf{I}_n - \mathbf{P})$ is the orthocomplement of $\mathscr{C}(\mathbf{P})$:

$$\mathbb{R}^n = \mathscr{C}(\mathbf{P}) \boxplus \mathscr{C}(\mathbf{I}_n - \mathbf{P}) = \mathscr{C}(\mathbf{P}) \boxplus \mathscr{C}(\mathbf{P})^\perp = \mathscr{C}(\mathbf{P}) \boxplus \mathscr{N}(\mathbf{P})\,. \tag{0.41}$$

Now the matrix $\mathbf{P}$ is an *orthogonal projector* onto $\mathscr{C}(\mathbf{P})$ (w.r.t. a given inner product). Sometimes the term oblique projector is used, to emphasize that (0.39) is valid but not necessarily (0.41).

Suppose $\mathbf{P}$ is an idempotent matrix. Then it satisfies (0.40) with respect to the standard inner product $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}'\mathbf{y} = \mathbf{x}'\mathbf{I}_n\mathbf{y}$, shortly said "w.r.t. $\mathbf{I}_n$", if and only if

$$\mathbf{x}'\mathbf{P}'(\mathbf{I}_n - \mathbf{P})\mathbf{y} = 0 \quad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n, \tag{0.42}$$

i.e., $\mathbf{P}' = \mathbf{P}'\mathbf{P}$, and hence $\mathbf{P}$ must be symmetric. We can conclude that the matrix $\mathbf{P}$ is an orthogonal projector if it is idempotent and symmetric:

$$\mathbf{P} \text{ is orthogonal projector (w.r.t. } \mathbf{I}) \iff \mathbf{P}^2 = \mathbf{P} \text{ and } \mathbf{P}' = \mathbf{P}\,. \tag{0.43}$$

If an idempotent symmetric $\mathbf{P}$ has the property $\mathscr{C}(\mathbf{P}) = \mathscr{C}(\mathbf{A})$, then $\mathbf{P}$ is the orthogonal projector onto $\mathscr{C}(\mathbf{A})$, denoted as $\mathbf{P}_{\mathbf{A}}$.

Let $\mathbf{A}\mathbf{b}_*$ denote the orthogonal projection of $\mathbf{y} \in \mathbb{R}^n$ onto $\mathscr{C}(\mathbf{A}_{n\times m})$. Then it is (at least geometrically) clear that the vector $\mathbf{y} - \mathbf{A}\mathbf{b}_*$ must be orthogonal to every column of $\mathbf{A}$: $\mathbf{A}'(\mathbf{y} - \mathbf{A}\mathbf{b}_*) = \mathbf{0}$, i.e., $\mathbf{A}'\mathbf{A}\mathbf{b}_* = \mathbf{A}'\mathbf{y}$, whose solution can be written as $\mathbf{b}_* = (\mathbf{A}'\mathbf{A})^-\mathbf{A}'\mathbf{y}$. Hence $\mathbf{A}\mathbf{b}_*$ can be expressed as $\mathbf{A}\mathbf{b}_* = \mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{A}'\mathbf{y}$. We will use intensively the fact that $\mathbf{P}_{\mathbf{A}}$ has the explicit representation as

$$\mathbf{P}_{\mathbf{A}} = \mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{A}' = \mathbf{A}\mathbf{A}^+, \tag{0.44}$$

which is invariant for any choice of $(\mathbf{A}'\mathbf{A})^-$. For detailed introduction of (0.44), see Theorem 8 (p. 155).

By $\mathbf{A}^\perp$ we denote any matrix whose columns span the orthocomplement of $\mathscr{C}(\mathbf{A})$:

$$\mathscr{C}(\mathbf{A}^\perp) = \mathscr{N}(\mathbf{A}') = \mathscr{C}(\mathbf{A})^\perp. \tag{0.45}$$

One choice for $\mathbf{A}^\perp$ is $\mathbf{I}_n - \mathbf{P}_{\mathbf{A}}$. Hence we have

$$\mathbb{R}^n = \mathscr{C}(\mathbf{A}) \boxplus \mathscr{C}(\mathbf{A})^\perp = \mathscr{C}(\mathbf{A}) \boxplus \mathscr{C}(\mathbf{I} - \mathbf{P}_{\mathbf{A}})\,. \tag{0.46}$$

It is important to remember that the matrix $\mathbf{A}^\perp$ is not necessarily unique (when is it unique?) but the column space $\mathscr{C}(\mathbf{A}^\perp)$ is unique.

Löwner Ordering

If a symmetric matrix $\mathbf{A}_{n\times n}$ is nonnegative definite, we denote $\mathbf{A} \geq_{\mathsf{L}} \mathbf{0}$, which means that

$$\text{there exists a matrix } \mathbf{L} \text{ such that } \mathbf{A} = \mathbf{L}\mathbf{L}'. \tag{0.47}$$

Instead of "nonnegative definite", the term "positive semidefinite" is often used. Notation $\mathbf{A} >_{\mathsf{L}} \mathbf{0}$ indicates that $\mathbf{A}$ is positive definite, i.e.,

$$\mathbf{A} = \mathbf{L}\mathbf{L}' \quad \text{for some nonsingular } \mathbf{L}_{n\times n}\,. \tag{0.48}$$

The sets of all symmetric nonnegative definite and positive definite $n \times n$ matrices are denoted as NND_n and PD_n, respectively. Moreover,

$$\mathbf{A} \geq_{\mathsf{L}} \mathbf{B} \text{ means that } \mathbf{A} - \mathbf{B} = \mathbf{K}\mathbf{K}' \text{ for some } \mathbf{K}\,, \tag{0.49}$$

and then we say that the matrix $\mathbf{B}$ is below $\mathbf{A}$ with respect to the *Löwner partial ordering*; this ordering is due to Löwner (1934).

In terms of quadratic forms, for a symmetric $\mathbf{A}_{n\times n}$:

$$\mathbf{A} \geq_{\mathsf{L}} \mathbf{0} \iff \mathbf{x}'\mathbf{A}\mathbf{x} \geq 0 \quad \text{for all } \mathbf{x} \in \mathbb{R}^n, \tag{0.50a}$$

$$\mathbf{A} >_{\mathsf{L}} \mathbf{0} \iff \mathbf{x}'\mathbf{A}\mathbf{x} > 0 \quad \text{for all nonzero } \mathbf{x} \in \mathbb{R}^n. \tag{0.50b}$$

It is also useful to know that using the principal minors (cf. page 429) and the ordered eigenvalues $\mathrm{ch}_1(\mathbf{A}) \geq \mathrm{ch}_2(\mathbf{A}) \geq \cdots \geq \mathrm{ch}_n(\mathbf{A})$ of a symmetric $\mathbf{A}$, we have

$$\mathbf{A} \geq_{\mathsf{L}} \mathbf{0} \iff \mathrm{ch}_n(\mathbf{A}) \geq 0 \iff \text{all principal minors of } \mathbf{A} \text{ are} \geq 0\,, \tag{0.51}$$

$$\begin{aligned} \mathbf{A} >_{\mathsf{L}} \mathbf{0} &\iff \mathrm{ch}_n(\mathbf{A}) > 0 \\ &\iff \text{all leading principal minors of } \mathbf{A} \text{ are} > 0\,. \end{aligned} \tag{0.52}$$

Notice that (0.47) means that

$$\mathbf{A} \text{ is nonnegative definite} \implies \mathbf{A} \text{ is necessarily symmetric.} \tag{0.53}$$

Whenever we talk about a nonnegative definite matrix $\mathbf{A}$ we assume that $\mathbf{A}$ is symmetric. Occasionally we may still clarify the situation by talking about "a symmetric nonnegative definite matrix $\mathbf{A}$" instead of "a nonnegative definite matrix $\mathbf{A}$" even though the latter implicitly indicates that $\mathbf{A}$ is symmetric.

The Löwner partial ordering is a surprisingly strong and useful property. For example, we observe immediately the following implications of the Löwner partial ordering:

$$\begin{aligned} \mathbf{A} \geq_{\mathsf{L}} \mathbf{B} \implies\ & a_{ii} \geq b_{ii}\,, \quad \mathrm{tr}(\mathbf{A}) \geq \mathrm{tr}(\mathbf{B})\,, \quad \mathrm{ch}_i(\mathbf{A}) \geq \mathrm{ch}_i(\mathbf{B})\,, \\ & \det(\mathbf{A}) \geq \det(\mathbf{B})\,, \quad \|\mathbf{A}\|_F \geq \|\mathbf{B}\|_F\,, \quad \|\mathbf{A}\|_2 \geq \|\mathbf{B}\|_2\,, \end{aligned} \tag{0.54}$$

where $i = 1, \dots, n$, and $\mathrm{ch}_i(\cdot)$ refers to the ith largest eigenvalue, and $\|\mathbf{A}\|_F = \sqrt{\mathrm{tr}(\mathbf{A}'\mathbf{A})}$ and $\|\mathbf{A}\|_2 = \sqrt{\mathrm{ch}_1(\mathbf{A}'\mathbf{A})}$ refer to the Frobenius norm and the spectral norm, respectively. Notice that if $\lambda_i = \mathrm{ch}_i(\mathbf{A})$ for a nonnegative definite (symmetric) $\mathbf{A}$, then

$$\|\mathbf{A}\|_F^2 = \mathrm{tr}(\mathbf{A}^2) = \lambda_1^2 + \cdots + \lambda_n^2\,, \quad \|\mathbf{A}\|_2^2 = \lambda_1^2\,. \tag{0.55}$$

We may also note that not all nonnegative definite matrices can be put into Löwner ordering. For example, $\mathbf{A} = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$ and $\mathbf{B} = \left(\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ are both nonnegative definite but neither $\mathbf{A} \geq_{\mathsf{L}} \mathbf{B}$ nor $\mathbf{B} \geq_{\mathsf{L}} \mathbf{A}$ holds.

Photograph 0.4 Ingram Olkin (Montréal, 1995).

It appears, see (0.82) (p. 18) that every correlation matrix is nonnegative definite and hence one may be inclined to believe that for example the matrix

$$\mathbf{A} = \begin{pmatrix} 1 & -2/3 & -2/3 \\ -2/3 & 1 & -2/3 \\ -2/3 & -2/3 & 1 \end{pmatrix} \tag{0.56}$$

is nonnegative definite because it "looks like a correlation matrix". However, $\mathbf{A}$ is not (why?) nonnegative definite and thereby it is not a correlation matrix; see, e.g., Yanai (2003), Puntanen & Styan (2005b, §6.2.1), and Section 14.1 (p. 307).

As a thorough reference to matrix orderings we may mention the book by Marshall, Olkin & Arnold (2011).

More on Orthogonal Projectors

Consider now the inner product in $\mathbb{R}^n$ defined as

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{V}} = \mathbf{x}'\mathbf{V}\mathbf{y}\,, \tag{0.57}$$

where $\mathbf{V}$ is a positive definite symmetric matrix. Ben-Israel & Greville (2003, p. 8) call (0.57) an ellipsoidal norm. Now (see Proposition 2.6, p. 81) the orthocomplement of $\mathscr{C}(\mathbf{A}_{n\times m})$ with respect to this inner product appears to be

$$\mathscr{C}(\mathbf{A})^{\perp}_{\mathbf{V}} = \{\, \mathbf{y} \in \mathbb{R}^n : \mathbf{z}'\mathbf{A}'\mathbf{V}\mathbf{y} = 0 \text{ for all } \mathbf{z} \in \mathbb{R}^m \,\} = \mathscr{C}(\mathbf{V}^{-1}\mathbf{A}^{\perp})\,. \tag{0.58}$$

By $\mathbf{A}^{\perp}_{\mathbf{V}}$ we will denote any matrix whose column space is $\mathscr{C}(\mathbf{A})^{\perp}_{\mathbf{V}}$. Recall that $\mathbf{A}^{\perp}_{\mathbf{I}}$ is shortly denoted as $\mathbf{A}^{\perp}$. Now we have the following decomposition:

$$\mathbb{R}^n = \mathscr{C}(\mathbf{A}) \boxplus \mathscr{C}(\mathbf{A})^{\perp}_{\mathbf{V}} = \mathscr{C}(\mathbf{A}) \boxplus \mathscr{C}(\mathbf{V}^{-1}\mathbf{A}^{\perp})\,. \tag{0.59}$$

Every $\mathbf{y} \in \mathbb{R}^n$ has now a unique representation as a sum

$$\mathbf{y} = \mathbf{A}\mathbf{b} + \mathbf{V}^{-1}\mathbf{A}^{\perp}\mathbf{c} = \mathbf{y}_* + \dot{\mathbf{y}}\,, \tag{0.60}$$

for some $\mathbf{b}$ and $\mathbf{c}$ (which are not necessarily unique).

The vector $\mathbf{y}_* = \mathbf{Ab}$ is the orthogonal projection of $\mathbf{y}$ onto $\mathscr{C}(\mathbf{A})$ along $\mathscr{C}(\mathbf{A})^{\perp}_{\mathbf{V}}$. The orthogonal projector $\mathbf{P}_{\mathbf{A};\mathbf{V}}$ is such a matrix which transforms $\mathbf{y}$ into its projection $\mathbf{y}_*$, i.e., $\mathbf{P}_{\mathbf{A};\mathbf{V}}\mathbf{y} = \mathbf{y}_* = \mathbf{Ab}$. Its explicit unique representation is

$$\mathbf{P}_{\mathbf{A};\mathbf{V}} = \mathbf{A}(\mathbf{A}'\mathbf{VA})^{-}\mathbf{A}'\mathbf{V}\,. \tag{0.61}$$

Finally we may recall the well-known fundamental minimizing property of the orthogonal projector:

$$\|\mathbf{y} - \mathbf{P}_{\mathbf{A};\mathbf{V}}\mathbf{y}\|_{\mathbf{V}} \leq \|\mathbf{y} - \mathbf{Ab}\|_{\mathbf{V}} \quad \text{for all } \mathbf{b},\, \mathbf{y}\,, \tag{0.62}$$

where the norm is defined as $\|\mathbf{z}\|^2_{\mathbf{V}} = \mathbf{z}'\mathbf{Vz}$. It is just the property (0.62) that makes the orthogonal projector such a powerful tool in linear models and multivariate analysis.

Eigenvalues and Singular Values

We recall that the scalar λ (real or complex) is an *eigenvalue* of $\mathbf{A}_{n\times n}$ if

$$\mathbf{At} = \lambda\mathbf{t} \quad \text{for some nonzero vector } \mathbf{t} \in \mathbb{R}^n, \tag{0.63}$$

in which case $\mathbf{t}$ is an *eigenvector* of $\mathbf{A}$ corresponding to λ. If $\mathbf{A}$ is symmetric, then all eigenvalues are real. The eigenvalues are the roots of the characteristic equation

$$\det(\mathbf{A} - \lambda\mathbf{I}_n) = 0\,. \tag{0.64}$$

Instead of $\det(\mathbf{A})$ we may occasionally use notation $|\mathbf{A}|$ (to save space).

As regards the matrix factorizations needed, the most important in this book is the *eigenvalue decomposition*[5] (EVD) (spectral decomposition): every symmetric $n \times n$ matrix $\mathbf{A}$ can be expressed as

$$\mathbf{A} = \mathbf{T\Lambda T}' = \lambda_1\mathbf{t}_1\mathbf{t}_1' + \cdots + \lambda_n\mathbf{t}_n\mathbf{t}_n'\,, \tag{EVD}$$

where $\mathbf{T}$ is orthogonal, $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$, and $\lambda_1 \geq \cdots \geq \lambda_n$ are the ordered eigenvalues of $\mathbf{A}$. The columns of $\mathbf{T}$ are the orthonormal eigenvectors of $\mathbf{A}$, and we denote $\operatorname{ch}_i(\mathbf{A}) = \lambda_i$.

In particular, for a symmetric nonnegative definite $n \times n$ matrix $\mathbf{A}$ with rank $r > 0$ we have

$$\begin{aligned}\mathbf{A} = \mathbf{T\Lambda T}' &= (\mathbf{T}_1 : \mathbf{T}_0)\begin{pmatrix}\mathbf{\Lambda}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{T}_1'\\ \mathbf{T}_0'\end{pmatrix}\\ &= \mathbf{T}_1\mathbf{\Lambda}_1\mathbf{T}_1' = \lambda_1\mathbf{t}_1\mathbf{t}_1' + \cdots + \lambda_r\mathbf{t}_r\mathbf{t}_r'\,,\end{aligned} \tag{0.65}$$

---

[5] According to Benzécri (1973, p. 289): “All in all, doing a data analysis, in good mathematics, is simply searching eigenvectors; all the science (or the art) of it is just to find the right matrix to diagonalize.”

where $\lambda_1 \geq \dots \geq \lambda_r > 0$, $\mathbf{\Lambda}_1 = \operatorname{diag}(\lambda_1, \dots, \lambda_r)$, and $\mathbf{T}_1 = (\mathbf{t}_1 : \ldots : \mathbf{t}_r)$, $\mathbf{T}_0 = (\mathbf{t}_{r+1} : \ldots : \mathbf{t}_n)$. Note that for a symmetric nonnegative definite $\mathbf{A}$ we have

$$\mathbf{A}^+ = \mathbf{T}_1 \mathbf{\Lambda}_1^{-1} \mathbf{T}_1' \,, \tag{0.66}$$

and the symmetric nonnegative definite *square root* of $\mathbf{A}$ is

$$\mathbf{A}^{1/2} = \mathbf{T}_1 \mathbf{\Lambda}_1^{1/2} \mathbf{T}_1' \,. \tag{0.67}$$

Moreover,

$$(\mathbf{A}^+)^{1/2} := \mathbf{A}^{+1/2} = \mathbf{T}_1 \mathbf{\Lambda}_1^{-1/2} \mathbf{T}_1' \,. \tag{0.68}$$

There is a longish Chapter 18 (pages 357–390) devoted to the eigenvalues.

The other extremely useful matrix factorization needed in statistics is of course the *singular value decomposition* (SVD):

$$\begin{aligned} \mathbf{A}_{n\times m} &= \mathbf{U}_{n\times n} \mathbf{\Delta}_{n\times m} \mathbf{V}'_{m\times m} = (\mathbf{U}_1 : \mathbf{U}_0) \begin{pmatrix} \mathbf{\Delta}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{V}_1' \\ \mathbf{V}_0' \end{pmatrix} \\ &= \mathbf{U}_1 \mathbf{\Delta}_1 \mathbf{V}_1' = \delta_1 \mathbf{u}_1 \mathbf{v}_1' + \dots + \delta_r \mathbf{u}_r \mathbf{v}_r' \,, \end{aligned} \tag{SVD}$$

where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal, $\operatorname{rank}(\mathbf{A}) = r$, and

$$\mathbf{\Delta}_{n\times m} = \begin{pmatrix} \mathbf{\Delta}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \quad \mathbf{\Delta}_1 = \operatorname{diag}(\delta_1, \dots, \delta_r) \,, \quad \delta_1 \geq \dots \geq \delta_r > 0 \,. \tag{0.69}$$

Above δ_i is the ith largest singular value, denoted also as $\operatorname{sg}_i(\mathbf{A})$. The eigenvalues and singular values have crucial properties when finding the extrema of quadratic forms (p. 360); Courant–Fischer minimax theorem (p. 389), Eckart–Young theorem (p. 400), Poincaré separation theorem (p. 398), to mention a few important results.

For example, for a statistician it is more than useful to know that for a partitioned nonnegative definite covariance matrix

$$\operatorname{cov}(\mathbf{z}) = \operatorname{cov}\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{\Sigma}_{\mathbf{xx}} & \mathbf{\Sigma}_{\mathbf{xy}} \\ \mathbf{\Sigma}_{\mathbf{yx}} & \mathbf{\Sigma}_{\mathbf{yy}} \end{pmatrix} = \mathbf{\Sigma} \tag{0.70a}$$

we have

$$\operatorname{ch}_1(\mathbf{\Sigma}) = \max_{\mathbf{a}\neq\mathbf{0}} \frac{\mathbf{a}'\mathbf{\Sigma}\mathbf{a}}{\mathbf{a}'\mathbf{a}} = \max_{\mathbf{a}'\mathbf{a}=1} \mathbf{a}'\mathbf{\Sigma}\mathbf{a} = \max_{\mathbf{a}'\mathbf{a}=1} \operatorname{var}(\mathbf{a}'\mathbf{z}) \,, \tag{0.70b}$$

and

$$\begin{aligned} \operatorname{sg}_1^2(\mathbf{\Sigma}_{\mathbf{xy}}) &= \max_{\mathbf{a}\neq\mathbf{0},\, \mathbf{b}\neq\mathbf{0}} \frac{(\mathbf{a}'\mathbf{\Sigma}_{\mathbf{xy}}\mathbf{b})^2}{\mathbf{a}'\mathbf{a} \cdot \mathbf{b}'\mathbf{b}} \\ &= \max_{\mathbf{c}'\mathbf{c}=\mathbf{d}'\mathbf{d}=1} (\mathbf{c}'\mathbf{\Sigma}_{\mathbf{xy}}\mathbf{d})^2 = \operatorname{ch}_1(\mathbf{\Sigma}_{\mathbf{xy}}\mathbf{\Sigma}_{\mathbf{yx}}) \,, \end{aligned} \tag{0.70c}$$

$$\begin{aligned}\operatorname{sg}_1^2(\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1/2}\boldsymbol{\Sigma}_{\mathbf{xy}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1/2}) &= \max_{\mathbf{a}\neq\mathbf{0},\,\mathbf{b}\neq\mathbf{0}} \frac{(\mathbf{a}'\boldsymbol{\Sigma}_{\mathbf{xy}}\mathbf{b})^2}{\mathbf{a}'\boldsymbol{\Sigma}_{\mathbf{xx}}\mathbf{a}\cdot\mathbf{b}'\boldsymbol{\Sigma}_{\mathbf{yy}}\mathbf{b}} \\ &= \max_{\mathbf{c}'\mathbf{c}=\mathbf{d}'\mathbf{d}=1}(\mathbf{c}'\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1/2}\boldsymbol{\Sigma}_{\mathbf{xy}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1/2}\mathbf{d})^2 \\ &= \operatorname{ch}_1(\boldsymbol{\Sigma}_{\mathbf{xy}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\boldsymbol{\Sigma}_{\mathbf{yx}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}) \\ &= \max_{\mathbf{a}\neq\mathbf{0},\,\mathbf{b}\neq\mathbf{0}} \operatorname{cor}^2(\mathbf{a}'\mathbf{x},\mathbf{b}'\mathbf{y})\,. \qquad (0.70\text{d})\end{aligned}$$

Chapter 19 (pages 391–413) is devoted to the singular values. A summary of some formulas useful for linear models and related matrix theory is presented in Isotalo, Puntanen & Styan (2008c).

Preliminaries to Random Vectors and Data Matrices

Here we recall some elementary properties of random vectors and data matrices.

An efficient quick tour through random vectors, data matrices, multivariate methods and linear models can be done via Puntanen & Styan (2007) and Puntanen, Seber & Styan (2007).

Random Vectors

An $n \times m$ random matrix $\mathbf{Z}$ is a matrix $\mathbf{Z} = \{z_{ij}\}$ of random variables $z_{11}, \dots, z_{nm}$. The expectation (mean) of a random matrix $\mathbf{Z}$ is $\operatorname{E}(\mathbf{Z}) = \{\operatorname{E}(z_{ij})\}$.

Photograph 0.5 George A. F. Seber (Auckland, 1999).

In the textbooks of mathematical statistics, frequently the scalar random variables are denoted as upper-case light-face italic letters such as X, Y and Z. We, however, do not follow this rule here. It gets too complicated to keep up the consistency of notation, in particular, when dealing with the random vectors. As was mentioned earlier in this chapter, random vectors are denoted in the same way as the vectors in $\mathbb{R}^n$. Our attempt is to use symbols $\mathbf{x}$, $\mathbf{y}, \dots$ for the random vectors and $\mathbf{a}$, $\mathbf{b}, \dots$ for vectors with real elements. From the context it should be clear whether we are dealing with a random or nonrandom vector.

The variance $\operatorname{var}(x)$ of the random variable x is

$$\operatorname{var}(x) = \sigma_x^2 = \mathrm{E}(x - \mu)^2, \quad \text{where } \mu = \mathrm{E}(x)\,, \tag{0.71}$$

and the covariance $\operatorname{cov}(x, y)$ between the two random variables x and y is

$$\operatorname{cov}(x, y) = \sigma_{xy} = \mathrm{E}(x - \mu)(y - \nu)\,, \quad \text{where } \mu = \mathrm{E}(x),\ \nu = \mathrm{E}(y)\,. \tag{0.72}$$

The covariance matrix (or the variance-covariance matrix or dispersion matrix) of the $p \times 1$ random vector $\mathbf{x} = (x_1, \dots, x_p)'$ is the $p \times p$ matrix $\operatorname{cov}(\mathbf{x}) = \mathbf{\Sigma}$ of variances and covariances of all the entries of $\mathbf{x}$:

$$\begin{aligned} \operatorname{cov}(\mathbf{x}) = \mathbf{\Sigma} = \{\sigma_{ij}\} &= \{\operatorname{cov}(x_i, x_j)\} = \{\mathrm{E}(x_i - \mu_i)(x_j - \mu_j)\} \\ &= \mathrm{E}(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})' = \mathrm{E}(\mathbf{x}\mathbf{x}') - \boldsymbol{\mu}\boldsymbol{\mu}', \end{aligned} \tag{0.73}$$

where $\boldsymbol{\mu} = \mathrm{E}(\mathbf{x})$. We will use the short notation

$$\mathrm{E}(\mathbf{x}) = \boldsymbol{\mu},\ \operatorname{cov}(\mathbf{x}) = \mathbf{\Sigma}: \quad \mathbf{x} \sim (\boldsymbol{\mu}, \mathbf{\Sigma})\,. \tag{0.74}$$

The cross-covariance matrix $\operatorname{cov}(\mathbf{x}, \mathbf{y})$ between the $p \times 1$ random vector $\mathbf{x} = (x_1, \dots, x_p)'$ and the $q \times 1$ random vector $\mathbf{y} = (y_1, \dots, y_q)'$ is the $p \times q$ matrix of all the covariances $\operatorname{cov}(x_i, y_j)$:

$$\begin{aligned} \operatorname{cov}(\mathbf{x}, \mathbf{y}) &= \{\operatorname{cov}(x_i, y_j)\} = \{\mathrm{E}(x_i - \mu_i)(y_j - \nu_j)\} \\ &= \mathrm{E}(\mathbf{x} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\nu})' = \mathrm{E}(\mathbf{x}\mathbf{y}') - \boldsymbol{\mu}\boldsymbol{\nu}', \end{aligned} \tag{0.75}$$

where $\boldsymbol{\mu} = \mathrm{E}(\mathbf{x})$ and $\boldsymbol{\nu} = \mathrm{E}(\mathbf{y})$. Notice that in our notation

$$\operatorname{cov}(\mathbf{x}) = \operatorname{cov}(\mathbf{x}, \mathbf{x})\,. \tag{0.76}$$

The correlation matrix $\operatorname{cor}(\mathbf{x}) = \boldsymbol{\rho}$, say, of the $p \times 1$ random vector $\mathbf{x} = (x_1, \dots, x_p)'$ is the $p \times p$ matrix of correlations of all the entries in $\mathbf{x}$:

$$\operatorname{cor}(\mathbf{x}) = \boldsymbol{\rho} = \{\varrho_{ij}\} = \{\operatorname{cor}(x_i, x_j)\} = \left\{ \frac{\sigma_{ij}}{\sigma_i \sigma_j} \right\}, \tag{0.77}$$

where $\sigma_i = \sqrt{\sigma_{ii}}$ = standard deviation of x_i; $\sigma_i, \sigma_j > 0$. Denoting

$$\mathbf{\Sigma}_\delta = \operatorname{diag}(\mathbf{\Sigma})\,, \tag{0.78}$$

we get

$$\operatorname{cor}(\mathbf{x}) = \boldsymbol{\rho} = \mathbf{\Sigma}_\delta^{-1/2} \mathbf{\Sigma} \mathbf{\Sigma}_\delta^{-1/2}, \quad \operatorname{cov}(\mathbf{x}) = \mathbf{\Sigma} = \mathbf{\Sigma}_\delta^{1/2} \boldsymbol{\rho} \mathbf{\Sigma}_\delta^{1/2}. \tag{0.79}$$

Let $\mathbf{x}$ be a $p \times 1$ random vector with the expectation $\boldsymbol{\mu} = \mathrm{E}(\mathbf{x})$, covariance matrix $\mathbf{\Sigma} = \operatorname{cov}(\mathbf{x})$, and let $\mathbf{A}$ be a $q \times p$ (nonrandom) matrix and vector $\mathbf{b} \in \mathbb{R}^q$. Then

$$\mathrm{E}(\mathbf{A}\mathbf{x} + \mathbf{b}) = \mathbf{A}\boldsymbol{\mu} + \mathbf{b}\,, \quad \operatorname{cov}(\mathbf{A}\mathbf{x} + \mathbf{b}) = \mathbf{A}\mathbf{\Sigma}\mathbf{A}'. \tag{0.80}$$

We note that

$$\operatorname{var}(\mathbf{a}'\mathbf{x}) = \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a} \quad \text{for all nonrandom } \mathbf{a} \in \mathbb{R}^p\,. \tag{0.81}$$

Because the variance is always nonnegative, (0.81) shows that

♠ every covariance matrix is necessarily nonnegative definite. (0.82)

The above conclusion concerns both theoretical covariance matrix and empirical sample covariance matrix.

When a p-dimensional random variable $\mathbf{z}$ (taking values in $\mathbb{R}^p$) follows the *multinormal distribution* with $\boldsymbol{\mu} = \mathrm{E}(\mathbf{z})$ and $\boldsymbol{\Sigma} = \operatorname{cov}(\mathbf{z})$, we denote

$$\mathbf{z} \sim \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})\,. \tag{0.83}$$

A p-dimensional random vector $\mathbf{z}$ is said to have a p-variate normal distribution N_p if every linear function $\mathbf{a}'\mathbf{z}$ has a univariate normal distribution. If $\mathbf{a}'\mathbf{z} = b$, where b is a constant, we define $\mathbf{a}'\mathbf{z} \sim \mathrm{N}(b, 0)$.

Alternatively, a p-dimensional random vector $\mathbf{z}$, with $\boldsymbol{\mu} = \mathrm{E}(\mathbf{z})$ and $\boldsymbol{\Sigma} = \operatorname{cov}(\mathbf{z})$, is said to have a p-variate normal distribution N_p if it can be expressed as $\mathbf{z} = \boldsymbol{\mu} + \mathbf{F}\mathbf{u}$, where $\mathbf{F}$ is an $p \times r$ matrix of rank r and $\mathbf{u}$ is a random vector of r independent univariate random variables. Notice that this definition (as well as the previous one) allows $\operatorname{cov}(\mathbf{z})$ to be singular.

Let $\mathbf{z} \sim \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is positive definite. Then $\mathbf{z}$ has the density

$$n(\mathbf{z}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \mathrm{e}^{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu})}. \tag{0.84}$$

For a reference to multinormal distribution, see, e.g., Rao (1973a, pp. 525–528), and Seber (2008, §20.5). Speed (2010) gives interesting remarks on how to define the multivariate normal distribution.

Let us complete this section with a short list of distributional facts; for further details, see Seber (2008, §20.5) and references therein:

(a) Central χ^2-distribution: $\mathbf{z} \sim \mathrm{N}_p(\mathbf{0}, \mathbf{I}_p)$: $\mathbf{z}'\mathbf{z} = \chi^2_p \sim \chi^2(p)\,.$

(b) Noncentral χ^2-distribution: $\mathbf{z} \sim \mathrm{N}_p(\boldsymbol{\mu}, \mathbf{I}_p)$: $\mathbf{z}'\mathbf{z} = \chi^2_{p,\delta} \sim \chi^2(p, \delta)$, $\delta = \boldsymbol{\mu}'\boldsymbol{\mu}\,.$

(c) Noncentral F-distribution: $F = \frac{\chi^2_{p,\delta}/p}{\chi^2_q/q} \sim \mathrm{F}(p, q, \delta)$, where $\chi^2_{p,\delta}$ and χ^2_q are independent.

(d) t-distribution: $\mathrm{t}^2(p) = \mathrm{F}(1, p)\,.$

(e) Let $\mathbf{z} \sim \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is positive definite and let $\mathbf{A}$ and $\mathbf{B}$ be symmetric and $\mathbf{b} \in \mathbb{R}^p$. Then (i)–(v) hold:

  (i) $\mathbf{z}'\mathbf{A}\mathbf{z} \sim \chi^2(r, \delta) \iff \mathbf{A}\boldsymbol{\Sigma}\mathbf{A} = \mathbf{A}$, in which case $r = \operatorname{tr}(\mathbf{A}\boldsymbol{\Sigma}) = \operatorname{rk}(\mathbf{A}\boldsymbol{\Sigma})$, $\delta = \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}\,,$

  (ii) $\mathbf{z}'\boldsymbol{\Sigma}^{-1}\mathbf{z} = \mathbf{z}'[\operatorname{cov}(\mathbf{z})]^{-1}\mathbf{z} \sim \chi^2(r, \delta)$, where $r = p$, $\delta = \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\,,$

(iii) $(\mathbf{z}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu}) \sim \chi^2(p)$,

(iv) $\mathbf{z}'\mathbf{A}\mathbf{z}$ and $\mathbf{z}'\mathbf{B}\mathbf{z}$ are independent $\iff \mathbf{A}\boldsymbol{\Sigma}\mathbf{B} = \mathbf{0}$,

(v) $\mathbf{z}'\mathbf{A}\mathbf{z}$ and $\mathbf{b}'\mathbf{z}$ independent $\iff \mathbf{A}\boldsymbol{\Sigma}\mathbf{b} = \mathbf{0}$.

If $\boldsymbol{\Sigma}$ is possibly singular and $\mathbf{z} \sim \mathrm{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$, then

(vi) $\mathbf{z}'\mathbf{A}\mathbf{z} \sim \chi^2(r) \iff \boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\Sigma} = \boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\Sigma}$, in which case $r = \mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}) = \mathrm{rk}(\mathbf{A}\boldsymbol{\Sigma})$,

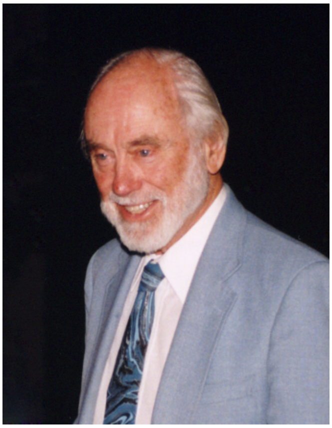
(vii) $\mathbf{z}'\mathbf{A}\mathbf{z}$ and $\mathbf{x}'\mathbf{B}\mathbf{x}$ are independent $\iff \boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\Sigma}\mathbf{B}\boldsymbol{\Sigma} = \mathbf{0}$,

(viii) $\mathbf{z}'\boldsymbol{\Sigma}^{-}\mathbf{z} = \mathbf{z}'[\mathrm{cov}(\mathbf{z})]^{-}\mathbf{z} \sim \chi^2(r)$ for any choice of $\boldsymbol{\Sigma}^{-}$ and $\mathrm{rank}(\boldsymbol{\Sigma}) = r$.

Above the independence between random variables means of course statistical (stochastic) independence: the random vectors $\mathbf{x}$ and $\mathbf{y}$ are statistically independent if and only if the joint distribution function of $\binom{\mathbf{x}}{\mathbf{y}}$ is the product of the distribution functions of $\mathbf{x}$ and $\mathbf{y}$. Intuitively this means that knowing something about the value of $\mathbf{x}$ does not provide any information about the value of $\mathbf{y}$. For example, if x and y are discrete random variables with values $x_1, \dots, x_r$ and $y_1, \dots, y_c$, then x and y are statistically independent if and only if

$$\mathrm{P}(x = x_i,\, y = y_j) = \mathrm{P}(x = x_i)\,\mathrm{P}(y = y_j)\,,\ i = 1, \dots, r,\ j = 1, \dots, c\,; \tag{0.85}$$

here P refers to the probability. For a deeper look at the statistical independence, see, e.g., Anderson (2003, pp. 10–11). Interesting remarks on the independence of the quadratic forms and so-called Craig–Sakamoto–Matusita theorem appear, e.g., in Driscoll & Gundberg (1986), Dumais (2000), Ogawa & Olkin (2008), and Randles (2007, pp. 140–141).



Photograph 0.6 Theodore W. Anderson (Fort Lauderdale, 1998).

Data Matrices

The matrix $\mathbf{U} \in \mathbb{R}^{n\times d}$ is called a *data matrix* if it contains the observed values of d variables $u_1, \dots, u_d$, each measured on n individuals (or corresponding units). We write

$$\mathbf{U}_{n\times d} = (\mathbf{u}_1 : \dots : \mathbf{u}_d) = \begin{pmatrix} \mathbf{u}'_{(1)} \\ \vdots \\ \mathbf{u}'_{(n)} \end{pmatrix}, \quad \mathbf{u}_j \in \mathbb{R}^n,\ \mathbf{u}_{(i)} \in \mathbb{R}^d. \tag{0.86}$$

Some authors use the transpose of $\mathbf{U}$ as the data matrix, in which case the data matrix $\mathbf{W}$, say, could be written as

$$\mathbf{W}_{d\times n} = (\mathbf{w}_1 : \ldots : \mathbf{w}_n), \quad \mathbf{w}_i \in \mathbb{R}^d. \tag{0.87}$$

The notation of (0.87) is handy in particular when the "primary units" are the observations as in multivariate methods, our notation being more convenient for linear statistical models. While reading a statistical article or book, the reader should indeed pay attention to such a trivially-feeling feature—whether the data matrix is defined as (0.86) or as its transpose.

In this section, and also elsewhere, we often denote the data matrix with a letter $\mathbf{U}$. This allows us then to label the columns of $\mathbf{U}_{n\times 2}$ for example as $\mathbf{x}$ and $\mathbf{y}$, which sometimes makes the notation easier to follow. We could (and we will) of course denote the data matrices as

$$\mathbf{X}_{n\times d} = (\mathbf{x}_1 : \ldots : \mathbf{x}_d) = (\mathbf{x}_{(1)} : \ldots : \mathbf{x}_{(n)})', \tag{0.88}$$

or

$$\mathbf{Y}_{n\times d} = (\mathbf{y}_1 : \ldots : \mathbf{x}_d) = (\mathbf{y}_{(1)} : \ldots : \mathbf{y}_{(n)})'. \tag{0.89}$$

We take minor efforts to save the symbol $\mathbf{X}$ for the model matrix in the linear model—however, these efforts are really mild.

Note that we use symbol d to indicate the number of variables. Here (there will be many exceptions) we follow the style initiated by George A. F. Seber, see, e.g., Seber (1984, p. 3), where he states that he departs from a more common but less convenient notation p for dimension. Indeed d is conveniently associated with the dimension of the (multidimensional) variable. Symbol n is quite uniformly used to indicate the number of observations.

In the data matrix $\mathbf{U}_{n\times d}$, the vector $\mathbf{u}_{(i)} \in \mathbb{R}^d$ represents the ith case or the ith observation in the observation space, and $\mathbf{u}_j \in \mathbb{R}^n$ represents the jth variable in the variable space. Sometimes the $\mathbf{u}_{(i)}$'s are called *observation vectors* and the $\mathbf{u}_j$'s are called *variable vectors*. A good treatment of the observation space and variable space appears, e.g., in Belsley (1991, §1.5).

The data matrix $\mathbf{U}$ can be thought to be a realization of the random matrix $\mathbf{U}$ which represents a random sample from some population. In other words, each $\mathbf{u}_{(i)}$ is a realization of the d-dimensional random vector $\mathbf{u}$. Mathematically this means each $\mathbf{u}_{(i)}$ has an identical distribution, and, as a consequence, they have identical means and identical covariance matrices:

$$\mathrm{E}(\mathbf{u}_{(i)}) = \boldsymbol{\mu} \in \mathbb{R}^d\,, \quad \operatorname{cov}(\mathbf{u}_{(i)}) = \mathbf{\Sigma}_{d\times d}\,, \quad i = 1, \ldots, n\,. \tag{0.90}$$

The second traditional assumption is that the random vectors (observations) $\mathbf{u}_{(i)}$ are statistically independent of each other, which implies

$$\operatorname{cov}(\mathbf{u}_{(r)}, \mathbf{u}_{(s)}) = \mathbf{0}\,, \quad \text{for } r \neq s\,. \tag{0.91}$$

The stochastic independence of the rows of $\mathbf{U}$ is not always holding, and then that fact should be carefully taken into account. The phrase

$$\text{“}\mathbf{U} = (\mathbf{u}_{(1)} : \ldots : \mathbf{u}_{(n)})' \text{ is a random sample”} \tag{0.92}$$

indicates that $\mathbf{u}_{(i)}$ are independent unless further information is provided.

Even though the observation vectors $\mathbf{u}_{(i)}$ are (usually) uncorrelated, the variable vectors $\mathbf{u}_j$ and $\mathbf{u}_k$ $(j \neq k)$ can be correlated:

$$\operatorname{cov}(\mathbf{u}_j, \mathbf{u}_k) = \sigma_{jk}\mathbf{I}_n\,, \ \text{while}\ \operatorname{E}(\mathbf{u}_j) = \mu_j \mathbf{1}_n\,, \quad j, k = 1, \ldots, d\,. \tag{0.93}$$

In terms of the vec-notation and the Kronecker product (for the definition, see page 429), we have

$$\mathbf{u}_* = \operatorname{vec}(\mathbf{U}) = \begin{pmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_d \end{pmatrix}, \quad \operatorname{E}(\mathbf{u}_*) = \begin{pmatrix} \mu_1 \mathbf{1}_n \\ \vdots \\ \mu_d \mathbf{1}_n \end{pmatrix} = \boldsymbol{\mu} \otimes \mathbf{1}_n \in \mathbb{R}^{dn}, \tag{0.94}$$

and

$$\operatorname{cov}(\mathbf{u}_*) = \begin{pmatrix} \sigma_1^2 \mathbf{I}_n & \sigma_{12}\mathbf{I}_n & \ldots & \sigma_{1d}\mathbf{I}_n \\ \vdots & \vdots & & \vdots \\ \sigma_{d1}\mathbf{I}_n & \sigma_{d2}\mathbf{I}_n & \ldots & \sigma_d^2 \mathbf{I}_n \end{pmatrix} = \mathbf{\Sigma} \otimes \mathbf{I}_n \in \mathrm{NND}_{dn}\,. \tag{0.95}$$

What happens if the rows of $\mathbf{U}_{n \times d}$, i.e., the random vectors $\mathbf{u}_{(i)}$, are not statistically independent? We could express this dependence by replacing $\mathbf{I}_n$ in (0.93) with an $n \times n$ nonnegative definite matrix $\mathbf{V}$:

$$\operatorname{cov}(\mathbf{u}_j, \mathbf{u}_k) = \sigma_{jk}\mathbf{V} \in \mathrm{NND}_n\,, \quad j, k = 1, \ldots, d\,. \tag{0.96}$$

Then (please confirm) for the rows of $\mathbf{U}$ we have

$$\operatorname{cov}(\mathbf{u}_{(r)}, \mathbf{u}_{(s)}) = v_{rs}\mathbf{\Sigma} \in \mathrm{NND}_d\,, \quad r, s = 1, \ldots, n\,, \tag{0.97}$$

and corresponding to (0.95), we get

$$\operatorname{cov}[\operatorname{vec}(\mathbf{U})] = \begin{pmatrix} \sigma_1^2 \mathbf{V} & \sigma_{12}\mathbf{V} & \ldots & \sigma_{1d}\mathbf{V} \\ \vdots & \vdots & & \vdots \\ \sigma_{d1}\mathbf{V} & \sigma_{d2}\mathbf{V} & \ldots & \sigma_d^2 \mathbf{V} \end{pmatrix} = \mathbf{\Sigma} \otimes \mathbf{V} \in \mathrm{NND}_{dn}\,. \tag{0.98}$$

Geometrically $\mathbf{U}$ can be interpreted as a collection of n points in $\mathbb{R}^d$, or as a collection of d points in $\mathbb{R}^n$.

The *sample variance* of the variable x whose observed values are in the vector $\mathbf{x}$ is defined as

$$\mathrm{var_s}(x) = \mathrm{var_d}(\mathbf{x}) = \tfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \tfrac{1}{n-1}t_{xx} = s_x^2\,, \tag{0.99}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and

$$t_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = \mathrm{SS}_x\,. \tag{0.100}$$

The symbol $\mathrm{var_s}(x)$ refers to the sample variance when the argument is the variable x while in $\mathrm{var_d}(\mathbf{x})$ the argument is the vector (data) comprising the observed values of the variable x. All in all, let's write down one more time the theoretical (population) variance and the sample variance:

$$\mathrm{var}(x) = \mathrm{E}(x-\mu)^2 = p_1(x_1-\mu)^2 + \dots + p_k(x_k-\mu)^2 = \sigma_x^2\,, \tag{0.101a}$$

$$\mathrm{var_s}(x) = \tfrac{1}{n-1}\big[(x_1-\bar{x})^2 + \dots + (x_n-\bar{x})^2\big] = s_x^2\,, \tag{0.101b}$$

where x in (0.101a) is a discrete random variable having values $x_1, \dots, x_k$ whose probabilities are $p_1, \dots, p_k$; $p_1 + \dots + p_k = 1$.

The *sample covariance* between x and y is

$$\begin{aligned}\mathrm{cov_s}(x,y) = \mathrm{cov_d}(\mathbf{x},\mathbf{y}) &= \tfrac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \\ &= \tfrac{1}{n-1}t_{xy} = \tfrac{1}{n-1}\,\mathrm{SP}_{xy} = s_{xy}\,. \end{aligned} \tag{0.102}$$

The *sample correlation* is

$$\mathrm{cor_s}(x,y) = \mathrm{cor_d}(\mathbf{x},\mathbf{y}) = \frac{\mathrm{SP}_{xy}}{\sqrt{\mathrm{SS}_x\,\mathrm{SS}_y}} = \frac{s_{xy}}{s_x s_y} = r_{xy}\,, \tag{0.103}$$

where we have to assume that $s_x > 0$, $s_y > 0$.

Let $\mathbf{1} \in \mathbb{R}^n$ denote the vector of ones and

$$\mathbf{J} = \mathbf{P_1} = \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}' = \tfrac{1}{n}\mathbf{1}\mathbf{1}', \quad \mathbf{C} = \mathbf{I}_n - \mathbf{J}\,, \tag{0.104}$$

where $\mathbf{C}$ is the centering matrix, and $\mathbf{J} = \mathbf{P_1}$ refers to the orthogonal projector (with respect to the standard inner product) onto $\mathscr{C}(\mathbf{1})$. Now

$$\mathbf{Jx} = \mathbf{1}\cdot\tfrac{1}{n}\mathbf{1}'\mathbf{x} = \bar{x}\mathbf{1} = \bar{\bar{\mathbf{x}}} \in \mathbb{R}^n, \tag{0.105}$$

and hence the variable vector $\mathbf{x}$ can be centered by premultiplying it by $\mathbf{C}$:

$$\mathbf{Cx} = \mathbf{x} - \mathbf{Jx} = \mathbf{x} - \bar{x}\mathbf{1} = \tilde{\mathbf{x}} \in \mathbb{R}^n. \tag{0.106}$$

In matrix terms,

$$\mathrm{SS}_x = \mathbf{x}'\mathbf{C}\mathbf{x}\,, \quad \mathrm{SP}_{xy} = \mathbf{x}'\mathbf{C}\mathbf{y}\,, \quad r_{xy} = \frac{\mathbf{x}'\mathbf{C}\mathbf{y}}{\sqrt{\mathbf{x}'\mathbf{C}\mathbf{x} \cdot \mathbf{y}'\mathbf{C}\mathbf{y}}}\,. \tag{0.107}$$

The data matrix $\mathbf{U}$ can be centered by premultiplying it by $\mathbf{C}$:

$$\begin{aligned}\widetilde{\mathbf{U}} = \mathbf{C}\mathbf{U} &= \mathbf{U} - \mathbf{J}\mathbf{U} = \mathbf{U} - \mathbf{1}\tfrac{1}{n}\mathbf{1}'\mathbf{U} = \mathbf{U} - \mathbf{1}\bar{\mathbf{u}}' \\ &= (\mathbf{C}\mathbf{u}_1 : \ldots : \mathbf{C}\mathbf{u}_d) \\ &= (\mathbf{u}_1 - \bar{u}_1\mathbf{1} : \ldots : \mathbf{u}_d - \bar{u}_d\mathbf{1}) = (\tilde{\mathbf{u}}_1 : \ldots : \tilde{\mathbf{u}}_d) \\ &= \begin{pmatrix} \mathbf{u}'_{(1)} - \bar{\mathbf{u}}' \\ \vdots \\ \mathbf{u}'_{(n)} - \bar{\mathbf{u}}' \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{u}}'_{(1)} \\ \vdots \\ \tilde{\mathbf{u}}'_{(n)} \end{pmatrix}, \end{aligned} \tag{0.108}$$

where

$$\bar{\mathbf{u}} = \tfrac{1}{n}\mathbf{U}'\mathbf{1} = \begin{pmatrix} \bar{u}_1 \\ \vdots \\ \bar{u}_d \end{pmatrix} \in \mathbb{R}^d, \tag{0.109}$$

and

$$\begin{aligned}\tilde{\mathbf{u}}_j &= j\text{th centered variable vector}\,, \quad j = 1, \ldots, d\,, \\ \tilde{\mathbf{u}}_{(i)} &= i\text{th centered observation vector}\,, \quad i = 1, \ldots, n\,.\end{aligned}$$

Hence the matrix of the sums of squares and products of the deviations about the mean is

$$\begin{aligned}\mathbf{T} = \widetilde{\mathbf{U}}'\widetilde{\mathbf{U}} = \mathbf{U}'\mathbf{C}\mathbf{U} &= \sum_{i=1}^{n}(\mathbf{u}_{(i)} - \bar{\mathbf{u}})(\mathbf{u}_{(i)} - \bar{\mathbf{u}})' \\ &= \sum_{i=1}^{n}\mathbf{u}_{(i)}\mathbf{u}'_{(i)} - n\bar{\mathbf{u}}\bar{\mathbf{u}}',\end{aligned} \tag{0.110}$$

and the sample covariance matrix from the data matrix $\mathbf{U}$ is

$$\mathbf{S} = \tfrac{1}{n-1}\mathbf{T} = \tfrac{1}{n-1}\mathbf{U}'\mathbf{C}\mathbf{U} = \tfrac{1}{n-1}\sum_{i=1}^{n}(\mathbf{u}_{(i)} - \bar{\mathbf{u}})(\mathbf{u}_{(i)} - \bar{\mathbf{u}})'. \tag{0.111}$$

Corresponding to the notation $\mathrm{var_d}(\mathbf{x}) = \frac{1}{n-1}\mathbf{x}'\mathbf{C}\mathbf{x} = s_x^2$, where the argument is the data, we may use the notation

$$\mathrm{cov_d}(\mathbf{U}) = \tfrac{1}{n-1}\mathbf{U}'\mathbf{C}\mathbf{U} = \mathbf{S}\,, \tag{0.112}$$

where the argument is the data matrix. Moreover, corresponding to $\mathrm{var_s}(x) = s_x^2$, where the argument is the variable x measured, we may denote

$$\mathrm{cov_s}(\mathbf{u}) = \mathbf{S}\,, \tag{0.113}$$

where the argument is the d-dimensional variable $\mathbf{u}$ which is being measured (n times, yielding the data matrix $\mathbf{U}$).

The sample correlation matrix is

$$\mathbf{R} = [\operatorname{diag}(\mathbf{S})]^{-1/2} \cdot \mathbf{S} \cdot [\operatorname{diag}(\mathbf{S})]^{-1/2}, \tag{0.114}$$

and

$$r_{ij} = \operatorname{cor}_{\mathrm{d}}(\mathbf{u}_i, \mathbf{u}_j) = \frac{t_{ij}}{\sqrt{t_{ii}t_{jj}}} = \frac{s_{ij}}{s_i s_j} = \frac{\mathbf{u}_i'\mathbf{C}\mathbf{u}_j}{\sqrt{\mathbf{u}_i'\mathbf{C}\mathbf{u}_i \cdot \mathbf{u}_j'\mathbf{C}\mathbf{u}_j}}\,. \tag{0.115}$$

If $\mathbf{A}$ is a $d \times q$ matrix then the covariance matrix calculated from the transformed data $\underline{\mathbf{U}}_{n\times q} = \mathbf{UA}$ is

$$\operatorname{cov}_{\mathrm{d}}(\mathbf{UA}) = \tfrac{1}{n-1}\mathbf{A}'\mathbf{U}'\mathbf{CUA} = \mathbf{A}' \operatorname{cov}_{\mathrm{d}}(\mathbf{U})\mathbf{A} = \mathbf{A}'\mathbf{SA}\,. \tag{0.116}$$

Notice that

$$\underline{\mathbf{U}} = \mathbf{UA} \implies \underline{\mathbf{u}}_{(i)}' = \mathbf{u}_{(i)}'\mathbf{A} \implies \underline{\mathbf{u}}_{(i)} = \mathbf{A}'\mathbf{u}_{(i)}\,. \tag{0.117}$$

If $\mathbf{u}$ refers to the d-dimensional variable yielding the data matrix $\mathbf{U}$ and $\underline{\mathbf{u}}$ yields $\mathbf{UA}$, then

$$\underline{\mathbf{u}} = \mathbf{A}'\mathbf{u}\,, \quad \operatorname{cov}_{\mathrm{s}}(\underline{\mathbf{u}}) = \operatorname{cov}_{\mathrm{s}}(\mathbf{A}'\mathbf{u}) = \mathbf{A}' \operatorname{cov}_{\mathrm{s}}(\mathbf{u})\mathbf{A} = \mathbf{A}'\mathbf{SA}\,. \tag{0.118}$$

Mahalanobis Distance

The Euclidean distance (squared) of the ith observation $\mathbf{u}_{(i)}$ from the mean $\bar{\mathbf{u}}$ is of course

$$\|\mathbf{u}_{(i)} - \bar{\mathbf{u}}\|^2 = (\mathbf{u}_{(i)} - \bar{\mathbf{u}})'(\mathbf{u}_{(i)} - \bar{\mathbf{u}}) = \sum_{j=1}^{d}(u_{ij} - \bar{u}_i)^2. \tag{0.119}$$

While describing how far $\mathbf{u}_{(i)}$ is from the mean $\bar{\mathbf{u}}$, the Euclidean distance is naturally doing a mathematically good job. But given the data matrix $\mathbf{U}$, one may wonder if there is a more informative way, particularly in a statistical sense, to measure the distance between $\mathbf{u}_{(i)}$ and $\bar{\mathbf{u}}$. In the univariate case, regarding the variable x, it is natural to calculate

$$D_i = \frac{|x_i - \bar{x}|}{\sqrt{\operatorname{var}_{\mathrm{s}}(x)}}\,, \tag{0.120}$$

because then the distance is measured in units of standard deviation and then we might well know that D_i is large if it exceeds 2 or 3. Moreover, D_i does not depend on the unit of measurement like the Euclidean distance.

One natural step towards a statistical distance in the multivariate case when, for simplicity, $\mathbf{U} = (\mathbf{x} : \mathbf{y})$, would be to calculate

$$\begin{aligned} D_i^2 &= \frac{(x_i - \bar{x})^2}{s_x^2} + \frac{(y_i - \bar{y})^2}{s_y^2} \\ &= (x_i - \bar{x}, y_i - \bar{y}) \begin{pmatrix} s_x^2 & 0 \\ 0 & s_y^2 \end{pmatrix}^{-1} \begin{pmatrix} x_i - \bar{x} \\ y_i - \bar{y} \end{pmatrix} \\ &= (\mathbf{u}_{(i)} - \bar{\mathbf{u}})' \begin{pmatrix} s_x^2 & 0 \\ 0 & s_y^2 \end{pmatrix}^{-1} (\mathbf{u}_{(i)} - \bar{\mathbf{u}})\,. \end{aligned} \tag{0.121}$$

However, it is easy to construct situations where the above D_i^2 is relatively small but at the same time the ith observation is "somehow different" from the rest of the data points; in particular this happens when x and y are highly correlated.

An improvement to (0.121), taking the correlations into account, is the squared (sample) *Mahalanobis distance*:

$$\mathrm{MHLN}^2(\mathbf{u}_{(i)}, \bar{\mathbf{u}}, \mathbf{S}) = (\mathbf{u}_{(i)} - \bar{\mathbf{u}})'\mathbf{S}^{-1}(\mathbf{u}_{(i)} - \bar{\mathbf{u}}) = \|\mathbf{S}^{-1/2}(\mathbf{u}_{(i)} - \bar{\mathbf{u}})\|^2, \tag{0.122}$$

where $\mathrm{cov_d}(\mathbf{U}) = \mathbf{S}$. For the history of the Mahalanobis distance, see Das Gupta (1993), and the first article on this concept by Mahalanobis (1936). (See also Philatelic Item 2.1, p. 90.)

Consider now the following set $\mathcal{A}$:

$$\mathcal{A} = \{\, \mathbf{u} \in \mathbb{R}^d : (\mathbf{u} - \bar{\mathbf{u}})'\mathbf{S}^{-1}(\mathbf{u} - \bar{\mathbf{u}}) = c^2 \,\}\,, \tag{0.123}$$

where c is a given nonzero constant. The set $\mathcal{A}$ is an ellipsoid centered at the mean $\bar{\mathbf{u}}$. Hence we have one fundamental feature of the Mahalanobis distance:

> ♠ those data points which have equal Mahalanobis distance from the mean, are located in the same ellipsoid centered at the mean. (0.124)

In Section 18.6 (p. 372), we discuss more thoroughly some interesting properties of the Mahalanobis distance.

The population Mahalanobis distance (squared) is defined as follows:

$$\mathrm{MHLN}^2(\mathbf{u}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\mathbf{u} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{u} - \boldsymbol{\mu})\,, \tag{0.125}$$

where $\mathrm{E}(\mathbf{u}) = \boldsymbol{\mu}$ and $\mathrm{cov}(\mathbf{u}) = \boldsymbol{\Sigma}$. Note that $\mathbf{u}$ in (0.125) is a random vector while $\mathbf{u}_{(i)}$ in (0.122) is an observed data point. [We can of course treat (0.122) as a random quantity, based on random sample $\mathbf{U}$.] Denoting

$$\mathbf{z} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{u} - \boldsymbol{\mu}) \tag{0.126}$$

we get $\mathrm{E}(\mathbf{z}) = \mathbf{0}$, $\mathrm{cov}(\mathbf{z}) = \mathbf{I}_d$, and if $\mathbf{u} \sim \mathrm{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then

$$\mathbf{z}'\mathbf{z} = z_1^2 + \cdots + z_d^2 = (\mathbf{u} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{u} - \boldsymbol{\mu}) \sim \chi^2(d)\,, \tag{0.127}$$

where z_i are independent and each $z_i \sim \mathrm{N}(0, 1)$, and $\chi^2(d)$ refers to the Chi-squared distribution with d degrees of freedom.

It might be useful to list a few distributional properties related to the random sample from the multinormal distribution—even though we do not use these properties very much in what follows. For further details, see, e.g., Anderson (2003, Ch. 5), Muirhead (1982, §3.2), and Seber (2008, §21.2.2), and references therein.

(a) Wishart distribution. Let $\mathbf{U}' = (\mathbf{u}_{(1)} : \ldots : \mathbf{u}_{(n)})$ be a random sample from $\mathrm{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$, i.e., $\mathbf{u}_{(i)}$'s are independent and each $\mathbf{u}_{(i)} \sim \mathrm{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$. Then $\mathbf{W} = \mathbf{U}'\mathbf{U} = \sum_{i=1}^n \mathbf{u}_{(i)}\mathbf{u}'_{(i)}$ is said to a have a Wishart distribution with n degrees of freedom and scale matrix $\boldsymbol{\Sigma}$, and we write $\mathbf{W} \sim \mathrm{Wishart}_p(n, \boldsymbol{\Sigma})$.

(b) Let $\mathbf{U}'$ be a random sample from $\mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then $\bar{\mathbf{u}} = \frac{1}{n}\mathbf{U}'\mathbf{1}_n$ and $\mathbf{T} = \mathbf{U}'(\mathbf{I} - \mathbf{J})\mathbf{U}$ are independent and $\mathbf{T} \sim \mathrm{Wishart}_p(n - 1, \boldsymbol{\Sigma})$.

(c) Hotelling's T^2 distribution. Suppose $\mathbf{v} \sim \mathrm{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$, $\mathbf{W} \sim \mathrm{Wishart}_p(m, \boldsymbol{\Sigma})$, $\mathbf{v}$ and $\mathbf{W}$ are independent, and that $\mathbf{W}$ is positive definite. Hotelling's T^2 distribution is the distribution of

$$T^2 = m \cdot \mathbf{v}'\mathbf{W}^{-1}\mathbf{v} = \mathbf{v}'\big(\tfrac{1}{m}\mathbf{W}\big)^{-1}\mathbf{v} \tag{0.128}$$

and is denoted as $T^2 \sim \mathrm{T}^2(p, m)$. See also Exercise 0.25 (p. 53).

Example 0.1 (Simple data set of three observations). As an example, let us consider the following (very) simple data of three observations and two variables; see the first Opening Example (p. 2). Figures 0.6 and 0.7 illustrate the situation.

$$\mathbf{U} = (\mathbf{x} : \mathbf{y}) = \begin{pmatrix} 1 & 1 \\ 4 & 1 \\ 4 & 4 \end{pmatrix} \begin{matrix} \text{Kalle} \\ \text{Ville} \\ \text{Maija} \end{matrix}\,, \tag{0.129a}$$

$$\bar{\mathbf{u}} = \begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix} = \frac{1}{3}\left[\begin{pmatrix} 1 \\ 1 \end{pmatrix} + \begin{pmatrix} 4 \\ 1 \end{pmatrix} + \begin{pmatrix} 4 \\ 4 \end{pmatrix}\right] = \begin{pmatrix} 3 \\ 2 \end{pmatrix}, \tag{0.129b}$$

$$\bar{\bar{\mathbf{x}}} = \bar{x}\mathbf{1} = \begin{pmatrix} 3 \\ 3 \\ 3 \end{pmatrix}, \quad \bar{\bar{\mathbf{y}}} = \bar{y}\mathbf{1} = \begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix}, \tag{0.129c}$$

$$\widetilde{\mathbf{U}} = (\tilde{\mathbf{x}} : \tilde{\mathbf{y}}) = \begin{pmatrix} 1 & 1 \\ 4 & 1 \\ 4 & 4 \end{pmatrix} - \begin{pmatrix} 3 & 2 \\ 3 & 2 \\ 3 & 2 \end{pmatrix} = \begin{pmatrix} -2 & -1 \\ 1 & -1 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{u}}'_{(1)} \\ \tilde{\mathbf{u}}'_{(2)} \\ \tilde{\mathbf{u}}'_{(3)} \end{pmatrix}, \tag{0.129d}$$

$$\mathbf{S} = \begin{pmatrix} 3 & 3/2 \\ 3/2 & 3 \end{pmatrix}, \quad \mathbf{S}^{-1} = \frac{4}{9}\begin{pmatrix} 1 & -1/2 \\ -1/2 & 1 \end{pmatrix}, \tag{0.129e}$$

where, e.g., the first centered observation vector is

$$\tilde{\mathbf{u}}_{(1)} = \mathbf{u}_{(1)} - \bar{\mathbf{u}} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} - \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} -2 \\ -1 \end{pmatrix}. \tag{0.129f}$$

Moreover,

$$\|\mathbf{u}_{(1)} - \bar{\mathbf{u}}\|^2 = \|\mathbf{u}_{(3)} - \bar{\mathbf{u}}\|^2 = 5\,, \quad \|\mathbf{u}_{(2)} - \bar{\mathbf{u}}\|^2 = 2\,, \tag{0.129g}$$

$$\mathrm{MHLN}^2(\mathbf{u}_{(i)}, \bar{\mathbf{u}}, \mathbf{S}) = 4/3\,, \quad i = 1, 2, 3\,. \tag{0.129h}$$

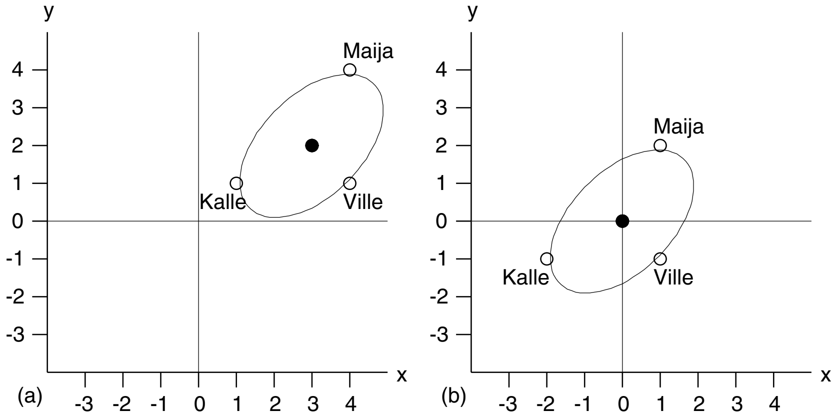


Figure 0.6 (Example 0.1) (a) Original data points, (b) centered data points. The mean point is marked as a filled circle. A confidence ellipse is also drawn; see Figure 0.8a (p. 56). Each observation has an equal Mahalanobis distance from the mean.

Keeping Observed Data as a Theoretical Distribution

Following Seber (2008, p. 433) we may make the following remark. Let x_1, …, x_n be the observed values of some empirical variable x, that is, x_1, …, x_n is a realized sample of size n. Let us imagine that x_* is a discrete uniform random variable whose values are x_1, …, x_n, i.e.,

$$\mathrm{P}(x_* = x_i) = \tfrac{1}{n}\,, \quad i = 1, \dots, n\,. \tag{0.130}$$

Then

$$\mathrm{E}(x_*) = \bar{x}, \quad \mathrm{var}(x_*) = \tfrac{n-1}{n} s_x^2\,. \tag{0.131}$$

More generally, consider a data matrix $\mathbf{U} = (\mathbf{u}_{(1)} : \ldots : \mathbf{u}_{(n)})'$ and define a discrete random vector $\mathbf{u}_*$ with probability function

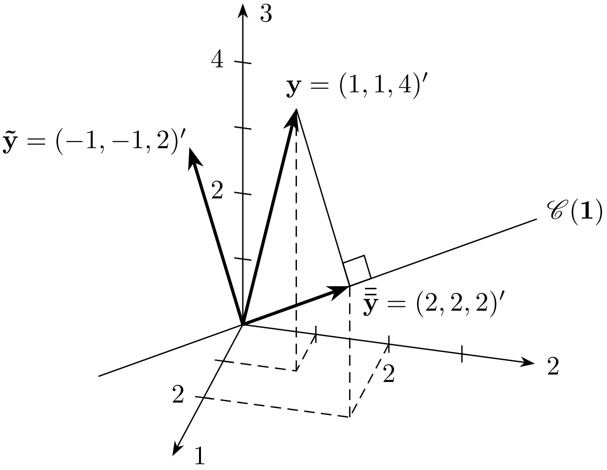


Figure 0.7 (Example 0.1) The variable vector $\mathbf{y}$ in the variable space $\mathbb{R}^3$; $\tilde{\mathbf{y}}$ is the centered $\mathbf{y}$ and $\bar{\bar{\mathbf{y}}} = \mathbf{J}\mathbf{y}$.

$$\mathrm{P}(\mathbf{u}_* = \mathbf{u}_{(i)}) = \tfrac{1}{n}, \quad i = 1, \dots, n\,, \tag{0.132}$$

i.e., every data point has the same probability to be the value of the random vector $\mathbf{u}_*$. Then

$$\mathrm{E}(\mathbf{u}_*) = \bar{\mathbf{u}}\,, \quad \operatorname{cov}(\mathbf{u}_*) = \tfrac{1}{n}\mathbf{U}'\mathbf{C}\mathbf{U} = \tfrac{n-1}{n}\mathbf{S}\,. \tag{0.133}$$

Moreover, the sample correlation matrix of data matrix $\mathbf{U}$ is the same as the (theoretical, population) correlation matrix of the random vector $\mathbf{u}_*$. Therefore any property shown for population correlation coefficients, holds for sample correlation coefficients and vice versa. According to Seber (2008, p. 433), we can "translate" sample properties into population properties using an appropriate discrete population.

Preliminaries to Linear Models

In this book, most of our examples are related to the general linear model, often called the Gauss–Markov model—named after Johann Carl Friedrich Gauss (1777–1855) and Andrey (Andrei) Andreyevich Markov (1856 N.S.–1922):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\,, \tag{0.134}$$

or in another notation,

$$\mathscr{M} = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{V}\}\,, \tag{0.135}$$

Photograph 0.7 Andrey (Andrei) Andreyevich Markov (1856 N.S.–1922) was a Russian mathematician, who is best known for his work on stochastic processes and what is now known as Markov chains.

Photograph 0.8 Andrei Viktorovich Markov (b. 1978) is a Russian professional ice hockey defenceman and an alternate captain of the Montréal Canadiens of the National Hockey League (NHL). (See also Tappara defenceman in Photograph 10.6 p. 252.)

where

$$\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} = \boldsymbol{\mu}\,, \quad \mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}\,, \quad \operatorname{cov}(\mathbf{y}) = \operatorname{cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{V}\,. \tag{0.136}$$

(For Gauss, see also Philatelic Item 9.1, p. 214.) For example, if we put $\mathbf{X} = (\mathbf{1} : \mathbf{x}_1 : \ldots : \mathbf{x}_k)$ and denote $p = k + 1$, we can write the model as

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \ldots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}. \tag{0.137}$$

Here

- vector $\mathbf{y}$ is an $n \times 1$ observable random vector,
- $\boldsymbol{\varepsilon}$ is an $n \times 1$ unobservable random error vector,
- $\mathbf{X}$ is a known $n \times p$ model (design) matrix,
- $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters,
- $\mathbf{V}$ is a known $n \times n$ nonnegative definite matrix, and

- σ^2 is an unknown nonzero constant.

The variable y with values in $\mathbf{y}$, is often called the response variable. The variables $x_1, \dots, x_k$, whose values are in the vectors $\mathbf{x}_1, \dots, \mathbf{x}_k$, are called predictors, regressors, explanatory variables, or independent variables. Of course, the constant variable "1" with values in $\mathbf{1}$, is also to be considered as a predictor. It is not necessary to have the vector $\mathbf{1}$ as a column of $\mathbf{X}$ but if we do have it there, we use the notation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = (\mathbf{1} : \mathbf{X}_0) \begin{pmatrix} \beta_0 \\ \boldsymbol{\beta}_{\mathbf{x}} \end{pmatrix} + \boldsymbol{\varepsilon} = \beta_0 \mathbf{1} + \mathbf{X}_0 \boldsymbol{\beta}_{\mathbf{x}} + \boldsymbol{\varepsilon} \,. \tag{0.138}$$

Notice that in this setup, the model matrix $\mathbf{X}$ is a matrix of fixed quantities, that is, the values of the predictors can basically be chosen, they are controllable. Hence, for example, the sample variance of x_i, $\mathrm{var}_{\mathrm{d}}(\mathbf{x}_i) = \mathbf{x}_i' \mathbf{C} \mathbf{x}_i / (n-1)$, is *not* necessarily an estimate of the corresponding population variance—because there may not exist any corresponding population. However, the situation changes if $x_1, \dots, x_k$ are random variables and $\mathbf{x}_1, \dots, \mathbf{x}_k$ represent their observed values resulting from a random sample. Then we might assume that the conditional expectation of the random vector $\mathbf{y}$, given that the random matrix $\mathbf{X}$ has the observed value $\underline{\mathbf{X}}$, say, is of the form $\mathrm{E}(\mathbf{y} \mid \mathbf{X} = \underline{\mathbf{X}}) = \underline{\mathbf{X}}\boldsymbol{\beta}$; similarly, $\mathrm{cov}(\mathbf{y} \mid \mathbf{X} = \underline{\mathbf{X}}) = \sigma^2 \mathbf{V}$.

Example 0.2 (Three treatments). Assume that we have three treatments A, B, and C, whose effects on the response variable y we wish to study using seven individuals. Let us suppose that one way to interpret the treatments is that A means running 1 hour per day, B means running 2 hours per day, and C means running 3 hours per day. Then we can consider the following models:

$$\mathscr{M}: \quad \mathbf{y} = \begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{31} \\ y_{32} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 2 \\ 1 & 2 \\ 1 & 2 \\ 1 & 3 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{pmatrix} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \,, \tag{0.139a}$$

$$\mathscr{M}_*: \quad \mathbf{y} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} + \boldsymbol{\varepsilon} = \mathbf{X}_* \boldsymbol{\beta}_* + \boldsymbol{\varepsilon} \,, \tag{0.139b}$$

$$\mathscr{M}_{\#}\colon \quad \mathbf{y} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{pmatrix} + \boldsymbol{\varepsilon} = \mathbf{X}_{\#}\boldsymbol{\beta}_{\#} + \boldsymbol{\varepsilon}\,. \tag{0.139c}$$

Notice that the model $\mathscr{M}$ differs essentially from the other two models $\mathscr{M}_*$ and $\mathscr{M}_{\#}$. These latter models simply divide the data into three groups—according to the treatments—and while doing so, the "numerical values" of treatments disappear. The regressors in $\mathscr{M}_*$ are simply dichotomous variables defined as $x_1 = 1$, if treatment is A, $x_1 = 0$ otherwise; x_2 and x_3 being defined in the corresponding way. Model $\mathscr{M}_{\#}$ is "overparametrized"; there are more parameters than necessary to indicate the treatment. Models $\mathscr{M}_*$ and $\mathscr{M}_{\#}$ are one-way analysis of variance, ANOVA, models. ☐

Example 0.3 (Block model). Let us take a quick look at the *block model*

$$\mathscr{M}\colon \quad \mathbf{y} = \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} = (\mathbf{1}_n : \mathbf{T} : \mathbf{B}) \begin{pmatrix} \mu \\ \boldsymbol{\tau} \\ \boldsymbol{\beta} \end{pmatrix} + \boldsymbol{\varepsilon}\,, \tag{0.140}$$

where

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_b \end{pmatrix}, \quad \mathbf{y}_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{ic_i} \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} \mathbf{1}_{c_1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{c_2} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{1}_{c_b} \end{pmatrix}, \tag{0.141}$$

and $n = c_1 + \dots + c_b$ = number of experimental units, $\boldsymbol{\tau} = (\tau_1, \dots, \tau_t)'$ is the vector of treatment effects (t treatments), $\boldsymbol{\beta} = (\beta_1, \dots, \beta_b)'$ is the vector of block effects (b blocks), $\mathbf{B}_{n\times b}$ is the design matrix of the block effects, $\mathbf{T}_{n\times t}$ is the design matrix of the treatment effects:

$$\mathbf{T} = (\mathbf{t}_1 : \mathbf{t}_2 : \dots : \mathbf{t}_t) = \begin{pmatrix} \mathbf{T}_1 \\ \mathbf{T}_2 \\ \vdots \\ \mathbf{T}_b \end{pmatrix}, \quad \mathbf{T}_i \in \mathbb{R}^{c_i \times t}. \tag{0.142}$$

The variable t_i whose values are in $\mathbf{t}_i$, has value 1 if the corresponding unit receives the treatment i, otherwise it is 0; each unit receives one treatment. Moreover, let r_i denote the number of units receiving the ith treatment:

$$\mathbf{r} = (r_1, r_2, \dots, r_t)' = \text{vector of replications}\,, \tag{0.143a}$$

$$\mathbf{c} = (c_1, c_2, \dots, c_b)' = \text{vector of block sizes}\,, \tag{0.143b}$$

so that $\mathbf{r}'\mathbf{1}_t = \mathbf{c}'\mathbf{1}_b = n$.

Let us denote the design, i.e., the arrangements of treatments on experimental units as a symbol d. For example, if we have $t = 3$ treatments A, B, C, $b = 2$ blocks, and $n = 6$ experimental units, we could have the following design d, say, (in the matrix form), where each row represents one block:

$$d = \begin{pmatrix} A & B & C \\ A & B & B \end{pmatrix} . \tag{0.144}$$

In this situation we have

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \mu + \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \tau_1 \\ \tau_2 \\ \tau_3 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \end{pmatrix} . \tag{0.145}$$

The main interest in this kind of model is in comparing the treatments, not in comparing the block effects.

We will consider properties of experimental design only very briefly in this book. Experimental design offers a huge area of linear models with a lot of literature. □

Example 0.4 (Multivariate linear model). In (0.137) (p. 29) we have only one response variable y whose observed values are in the vector $\mathbf{y}$. Suppose that we have two response variables y_1 and y_2 whose observed values are in the matrix $\mathbf{Y}$:

$$\mathbf{Y} = (\mathbf{y}_1 : \mathbf{y}_2) = \begin{pmatrix} \mathbf{y}'_{(1)} \\ \vdots \\ \mathbf{y}'_{(n)} \end{pmatrix} . \tag{0.146}$$

If we now believe that

$$\mathbf{y}_i = \mathbf{X}\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i \,,\ \mathrm{E}(\boldsymbol{\varepsilon}_i) = \mathbf{0} \,,\ \mathrm{cov}(\boldsymbol{\varepsilon}_i) = \sigma_{ii}\mathbf{I}_n \,,\ \mathrm{cov}(\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2) = \sigma_{12}\mathbf{I}_n \,, \tag{0.147}$$

where $i = 1, 2$, then we have a *multivariate* linear model which can be written as

$$\mathbf{Y} = \mathbf{X}\mathbb{B} + \boldsymbol{\mathcal{E}} \,, \tag{0.148}$$

where

$$\mathbb{B} = (\boldsymbol{\beta}_1 : \boldsymbol{\beta}_2) \,, \quad \boldsymbol{\mathcal{E}} = (\boldsymbol{\varepsilon}_1 : \boldsymbol{\varepsilon}_2) \,. \tag{0.149}$$

In this setup, the rows (individuals) are uncorrelated, i.e., $\mathrm{cov}(\mathbf{y}_{(r)}, \mathbf{y}_{(s)}) = \mathbf{0}$, $r \neq s$. Using the Kronecker product and the vec-notation, we can obviously express (0.148) as

$$\begin{aligned}\operatorname{vec}(\mathbf{Y}) = \begin{pmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{pmatrix} &= \begin{pmatrix}\mathbf{X} & \mathbf{0}\\ \mathbf{0} & \mathbf{X}\end{pmatrix}\begin{pmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{pmatrix} + \begin{pmatrix}\boldsymbol{\varepsilon}_1\\ \boldsymbol{\varepsilon}_2\end{pmatrix}\\ &= (\mathbf{I}_2 \otimes \mathbf{X})\operatorname{vec}(\mathbb{B}) + \operatorname{vec}(\boldsymbol{\mathcal{E}})\,. \end{aligned} \tag{0.150}$$

In view of (0.95) (p. 21), we have

$$\operatorname{cov}[\operatorname{vec}(\mathbf{Y})] = \begin{pmatrix}\sigma_{11}\mathbf{I}_n & \sigma_{12}\mathbf{I}_n\\ \sigma_{21}\mathbf{I}_n & \sigma_{22}\mathbf{I}_n\end{pmatrix} = \boldsymbol{\Sigma} \otimes \mathbf{I}_n\,, \tag{0.151}$$

and hence the multivariate model (0.148) can be expressed as a univariate model

$$\{\operatorname{vec}(\mathbf{Y}),\ (\mathbf{I}_2 \otimes \mathbf{X})\operatorname{vec}(\mathbb{B}),\ \boldsymbol{\Sigma} \otimes \mathbf{I}_n\}\,. \tag{0.152}$$

Section 10.5 (p. 229) deals (shortly) with the estimation under the multivariate linear model. □

Ordinary Least Squares

The idea in the least squares estimation is to find the vector $\boldsymbol{\beta}$ so that the vector $\mathbf{X}\boldsymbol{\beta}$ would be as close to the observed $\mathbf{y}$ as possible, i.e., we are minimizing the quantity

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{0.153}$$

with respect to $\boldsymbol{\beta}$, i.e., we

$$\text{minimize } \|\mathbf{y} - \boldsymbol{\mu}\|^2 \text{ under the condition } \boldsymbol{\mu} \in \mathscr{C}(\mathbf{X})\,. \tag{0.154}$$

Of course, here we use $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$ as mathematical variables—not as the parameters of the model $\mathscr{M}$. This sin is rather common in literature, and it should not be too confusing. The minimum of (0.153) is naturally obtained by projecting $\mathbf{y}$ onto the column space $\mathscr{C}(\mathbf{X})$. We will use the notation

$$\mathbf{H} = \mathbf{P}_{\mathbf{X}}\,, \quad \mathbf{M} = \mathbf{I}_n - \mathbf{H}\,, \tag{0.155}$$

thereby obtaining the *ordinary least squares* (OLS) *estimator* of $\mathbf{X}\boldsymbol{\beta}$ as

$$\operatorname{OLSE}(\mathbf{X}\boldsymbol{\beta}) = \mathbf{H}\mathbf{y} = \mathbf{P}_{\mathbf{X}}\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} = \widehat{\mathbf{X}\boldsymbol{\beta}} = \hat{\boldsymbol{\mu}} = \hat{\mathbf{y}}\,, \tag{0.156}$$

the corresponding vector of the residuals being

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{H}\mathbf{y} = \mathbf{M}\mathbf{y}\,. \tag{0.157}$$

The matrix $\mathbf{H}$ is often called the “hat matrix”. For a clarity, we may call the vector $\hat{\boldsymbol{\varepsilon}}$ as the OLSE’s residual. (Recall that there are several kinds of residuals, like e.g. the BLUE's and BLP's residual.) Keeping $\mathbf{y}$ as a random vector so that $\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and $\mathrm{cov}(\mathbf{y}) = \sigma^2\mathbf{V}$, we have

$$\mathrm{E}(\mathbf{Hy}) = \mathbf{HX}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta}\,, \quad \mathrm{cov}(\mathbf{Hy}) = \sigma^2\mathbf{HVH}\,, \tag{0.158}$$

and

$$\mathrm{E}(\mathbf{My}) = \mathbf{MX}\boldsymbol{\beta} = \mathbf{0}\,, \quad \mathrm{cov}(\mathbf{My}) = \sigma^2\mathbf{MVM}\,. \tag{0.159}$$

For a geometric illustration of the OLS method, see Figure 8.3 (p. 183).

Remark 0.1 (Remember H and M). The reader is strongly encouraged to put the short notations $\mathbf{H}$ and $\mathbf{M}$ into his/her active section of the memory. This investment will be profitable. Of course there is a bunch of notations worth remembering, but the importance of $\mathbf{H}$ and $\mathbf{M}$ makes their role worth emphasizing. □

In (0.156) the vector $\hat{\boldsymbol{\beta}}$ is an arbitrary solution to the *normal equation*

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}\,. \tag{0.160}$$

The general solution (the set of all solutions) to the normal equation (which is always consistent) is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{y} + [\mathbf{I}_p - (\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{X}]\mathbf{z}\,, \tag{0.161}$$

where $\mathbf{z} \in \mathbb{R}^p$ is free to vary and $(\mathbf{X}'\mathbf{X})^-$ is an arbitrary (but fixed) generalized inverse of $\mathbf{X}'\mathbf{X}$. On the other hand, every solution to the normal equation can be expressed as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{y}\,, \tag{0.162}$$

for some choice of $(\mathbf{X}'\mathbf{X})^-$.

When $\mathbf{X}$ does not have full column rank, then the vector $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{y}$ is not unique and it is not a *proper* estimator: it is merely a *solution* to the normal equations—"…this point cannot be overemphasized", as stated by Searle (1971, p. 169). The situation when $\hat{\boldsymbol{\beta}}$ is needed is often such that we are interested in the OLSE of an *estimable* parametric vector $\mathbf{K}'\boldsymbol{\beta}$, which is defined as $\mathbf{K}'\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is any solution to (0.160). In such a situation $\mathbf{K}'\hat{\boldsymbol{\beta}}$ is indeed unique. Of course $\mathbf{X}\hat{\boldsymbol{\beta}}$ is always unique. In this context we wish to cite also the following from Searle (2000, p. 26):

Photograph 0.9 Shayle R. Searle (Auckland, 2005).

> "One of the greatest contributions to understanding the apparent quirkiness of normal equations of non-full rank (as is customary with linear models), which

> have an infinity of solutions, is due to Rao (1962). Using the work of Moore (1920) and Penrose (1955), he showed how a generalized inverse matrix yields a solution to the normal equations and how that solution can be used to establish estimable functions and their estimators—and these results are invariant to whatever generalized inverse is being used. Although the arithmetic of generalized inverses is scarcely any less than that of regular inverses, the use of generalized inverses is of enormous help in understanding estimability and its consequences."

Example 0.5 (Three treatments, continued). Notice that in Example 0.2 (p. 30), the column spaces $\mathscr{C}(\mathbf{X}_*)$ and $\mathscr{C}(\mathbf{X}_\#)$ are obviously identical and thereby

$$\mathbf{P}_{\mathbf{X}_*} = \mathbf{P}_{\mathbf{X}_\#} = \mathbf{H} = \begin{pmatrix} \mathbf{J}_2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_3 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{J}_2 \end{pmatrix} = \begin{pmatrix} \frac12 & \frac12 & 0 & 0 & 0 & 0 & 0 \\ \frac12 & \frac12 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac13 & \frac13 & \frac13 & 0 & 0 \\ 0 & 0 & \frac13 & \frac13 & \frac13 & 0 & 0 \\ 0 & 0 & \frac13 & \frac13 & \frac13 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac12 & \frac12 \\ 0 & 0 & 0 & 0 & 0 & \frac12 & \frac12 \end{pmatrix} . \tag{0.163}$$

Hence

$$\operatorname{OLSE}(\mathbf{X}\boldsymbol{\beta}_* \mid \mathscr{M}_*) = \operatorname{OLSE}(\mathbf{X}\boldsymbol{\beta}_\# \mid \mathscr{M}_\#) = \mathbf{H}\mathbf{y} = \begin{pmatrix} \bar{y}_1 \mathbf{1}_2 \\ \bar{y}_2 \mathbf{1}_3 \\ \bar{y}_3 \mathbf{1}_2 \end{pmatrix}, \tag{0.164a}$$

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{H}\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 - \bar{y}_1 \mathbf{1}_2 \\ \mathbf{y}_2 - \bar{y}_2 \mathbf{1}_3 \\ \mathbf{y}_3 - \bar{y}_3 \mathbf{1}_2 \end{pmatrix} = \begin{pmatrix} y_{11} - \bar{y}_1 \\ y_{12} - \bar{y}_1 \\ y_{21} - \bar{y}_2 \\ y_{22} - \bar{y}_2 \\ y_{23} - \bar{y}_2 \\ y_{31} - \bar{y}_3 \\ y_{32} - \bar{y}_3 \end{pmatrix}, \tag{0.164b}$$

and

$$\begin{aligned} \mathrm{SSE} &= \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = \|\mathbf{y}_1 - \bar{y}_1 \mathbf{1}_2\|^2 + \|\mathbf{y}_2 - \bar{y}_2 \mathbf{1}_3\|^2 + \|\mathbf{y}_3 - \bar{y}_3 \mathbf{1}_2\|^2 \\ &= \sum_{j=1}^{2} (y_{1j} - \bar{y}_1)^2 + \sum_{j=1}^{3} (y_{2j} - \bar{y}_2)^2 + \sum_{j=1}^{2} (y_{3j} - \bar{y}_3)^2 \\ &:= \mathrm{SS}_1 + \mathrm{SS}_2 + \mathrm{SS}_3 := \mathrm{SS}_{\mathrm{Between}}, \end{aligned} \tag{0.165}$$

where $\bar{y}_i$ is the mean of the y-values in the ith group. The OLSE of $\boldsymbol{\beta}_*$ under $\mathscr{M}_*$ is unique but the OLSE of $\boldsymbol{\beta}_\#$ under $\mathscr{M}_\#$ is not unique. □

Remark 0.2 (Regression line and conditional means). Let the groups of the previous Example 0.5 be indicated by numbers 1, 2, 3, and let us make a scatter plot of seven data points, where x (treatment) is on the x-axis and y on the y-axis; see Exercise 8.11 (p. 186). Fitting a regression line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ means that we have to minimize

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = \sum_{j=1}^{2}(y_{1j} - \alpha_1)^2 + \sum_{j=1}^{3}(y_{2j} - \alpha_2)^2 + \sum_{j=1}^{2}(y_{3j} - \alpha_3)^2, \tag{0.166}$$

where $\alpha_1 = \beta_0 + \beta_1\,, \alpha_2 = \beta_0 + 2\beta_1\,, \alpha_3 = \beta_0 + 3\beta_1$. Because (please confirm if you are hesitant)

$$\sum_{j=1}^{n_i}(y_{ij} - \alpha_i)^2 \geq \sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 = \mathrm{SS}_i \quad \text{for all } \alpha_i,\ i = 1, 2, 3\,, \tag{0.167}$$

we confirm a very important fact:

> ♠ whenever possible, the regression line goes through the conditional means. (0.168)

In other words, if the conditional means lie on the same line, then that line is precisely the regression line. This information can be very helpful for quick drawing of regression lines. Notice that this is just the case in Figures 0.8 (p. 56) and 0.9. □

Three Important Projectors

When $\mathbf{X} = \mathbf{1} \in \mathbb{R}^n$, we will denote

$$\mathbf{J} = \mathbf{P}_{\mathbf{1}} = \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}' = \tfrac{1}{n}\mathbf{1}\mathbf{1}', \quad \mathbf{C} = \mathbf{I}_n - \mathbf{J}\,, \tag{0.169}$$

and hence under the simple basic model $\mathscr{M}_0 = \{\mathbf{y},\ \mathbf{1}\beta,\ \sigma^2\mathbf{V}\}$ we have

$$\mathrm{OLSE}(\beta) = \hat{\beta} = (\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'\mathbf{y} = \mathbf{1}^+\mathbf{y} = \bar{y}\,, \tag{0.170a}$$

$$\mathrm{OLSE}(\mathbf{1}\beta) = \mathbf{1}\hat{\beta} = \mathbf{J}\mathbf{y} = \bar{y}\mathbf{1}\,, \tag{0.170b}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. The residual vector under $\mathscr{M}_0$ is the centered $\mathbf{y}$:

$$\mathbf{y} - \mathbf{J}\mathbf{y} = \mathbf{C}\mathbf{y} = \tilde{\mathbf{y}}\,. \tag{0.171}$$

The three orthogonal projectors $\mathbf{H}$, $\mathbf{J}$, and $\mathbf{C}$ (the centering matrix), play crucial roles in many considerations related to linear regression. An important fact is also that the sample correlation coefficient between the variables x

and y, whose values are the elements of vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, is the cosine between the corresponding centered vectors:

$$\operatorname{cor}_{\mathrm{s}}(x, y) = \operatorname{cor}_{\mathrm{d}}(\mathbf{x}, \mathbf{y}) = \cos(\mathbf{Cx}, \mathbf{Cy}) = r_{xy} = \frac{\mathbf{x}'\mathbf{Cy}}{\sqrt{\mathbf{x}'\mathbf{Cx} \cdot \mathbf{y}'\mathbf{Cy}}} \,. \tag{0.172}$$

It is important to realize that centering can be interpreted as a projection:

$$\tilde{\mathbf{y}} = \mathbf{y} - \mathbf{Jy} = \mathbf{Cy} = (\mathbf{I} - \mathbf{P}_{\mathbf{1}})\mathbf{y} = \mathbf{P}_{\mathbf{1}^\perp}\mathbf{y} = \mathbf{P}_{\mathbf{C}}\mathbf{y} \,. \tag{0.173}$$

For the geometry of correlation, see Figure 3.1 (p. 93).

Estimability

Let $\mathbf{K}'\boldsymbol{\beta}$ be a given vector of parametric functions specified by $\mathbf{K}' \in \mathbb{R}^{q \times p}$ and let $\{\mathrm{LE}(\mathbf{K}'\boldsymbol{\beta}; \mathbf{y})\}$ denote the set of all *linear estimators* of $\mathbf{K}'\boldsymbol{\beta}$, that is,

$$\{\mathrm{LE}(\mathbf{K}'\boldsymbol{\beta}; \mathbf{y})\} = \{\, \mathbf{Ay} : \mathbf{A} \in \mathbb{R}^{q \times n} \,\} \,. \tag{0.174}$$

In (0.174) we actually consider only homogeneous linear estimators; inhomogeneous linear estimators are of the form

$$\mathbf{Ay} + \mathbf{a} : \quad \mathbf{A} \in \mathbb{R}^{q \times n}, \quad \mathbf{a} \in \mathbb{R}^q. \tag{0.175}$$

A parametric function $\mathbf{K}'\boldsymbol{\beta}$ is said to be *estimable* if it has a linear unbiased estimator, i.e., there exists a matrix $\mathbf{A}$ such that

$$\mathrm{E}(\mathbf{Ay}) = \mathbf{AX}\boldsymbol{\beta} = \mathbf{K}'\boldsymbol{\beta} \quad \text{for all } \boldsymbol{\beta} \in \mathbb{R}^p, \tag{0.176}$$

and hence

$$\mathbf{K}'\boldsymbol{\beta} \text{ is estimable} \iff \exists\, \mathbf{A} \colon \mathbf{K} = \mathbf{X}'\mathbf{A}' \iff \mathscr{C}(\mathbf{K}) \subset \mathscr{C}(\mathbf{X}') \,. \tag{0.177}$$

The OLSE of $\mathbf{K}'\boldsymbol{\beta}$ is defined as

$$\mathrm{OLSE}(\mathbf{K}'\boldsymbol{\beta}) = \widehat{\mathbf{K}'\boldsymbol{\beta}} = \mathbf{K}'\hat{\boldsymbol{\beta}} \,, \tag{0.178}$$

where $\hat{\boldsymbol{\beta}}$ is any solution to $\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}$. Now the condition $\mathscr{C}(\mathbf{K}) \subset \mathscr{C}(\mathbf{X}')$ guarantees that $\mathbf{K}'\hat{\boldsymbol{\beta}}$ is unique, even though $\hat{\boldsymbol{\beta}}$ may not be unique; see Section 12.2 (p. 284).

We can also conclude that

$$\{\, \mathbf{K}'\boldsymbol{\beta} : \mathbf{K}'\boldsymbol{\beta} \text{ is estimable} \,\} = \{\, \mathbf{AX}\boldsymbol{\beta} : \mathbf{A} \in \mathbb{R}^{q \times n} \,\} \,. \tag{0.179}$$

The fact that all estimable parametric functions $\mathbf{K}'\boldsymbol{\beta}$ are of the form $\mathbf{AX}\boldsymbol{\beta}$ (for some $\mathbf{A}$) means that it is the vector $\mathbf{X}\boldsymbol{\beta}$ whose estimation has a central role in linear estimation.

In particular, a scalar valued parametric function $\mathbf{k}'\boldsymbol{\beta}$ is estimable if

$$\mathbf{k} = \mathbf{X}'\mathbf{a} \quad \text{for some } \mathbf{a} \in \mathbb{R}^n. \tag{0.180}$$

Hence we can write

$$\{\, \mathbf{k}'\boldsymbol{\beta} : \mathbf{k}'\boldsymbol{\beta} \text{ estimable} \,\} = \{\, \mathbf{a}'\mathbf{X}\boldsymbol{\beta} : \mathbf{a} \in \mathbb{R}^n \,\}. \tag{0.181}$$

Example 0.6 (Two treatments). As an example, consider the model $\{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{V}\}$, where

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} = (\mathbf{1} : \mathbf{x}_1 : \mathbf{x}_2)\,, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \end{pmatrix}, \quad \mathbf{k} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}. \tag{0.182}$$

Then $\mathbf{k}'\boldsymbol{\beta} = \tau_2$ is not estimable because

$$\mathbf{k} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \notin \mathscr{C}(\mathbf{X}') = \mathscr{C}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \mathscr{C}\begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} = \left\{ \begin{pmatrix} a+b \\ a \\ b \end{pmatrix} \right\}, \tag{0.183}$$

where a and b can be any real numbers. It is easy to confirm, using the above technique, that each individual element of $\boldsymbol{\beta}$ is estimable under $\{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{V}\}$ if and only if $\mathbf{X}$ has full column rank. The other way to confirm this is done in (0.187) below. Notice that the normal equation $\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}$ in this situation becomes

$$\begin{pmatrix} 4 & 3 & 1 \\ 3 & 3 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \end{pmatrix} = \begin{pmatrix} y_1 + y_2 + y_3 + y_4 \\ y_1 + y_2 + y_3 \\ y_4 \end{pmatrix}. \tag{0.184}$$

Moreover,

$$(\mathbf{X}'\mathbf{X})^- = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1/3 & 0 \\ 0 & 0 & 1 \end{pmatrix} \implies \begin{pmatrix} \hat{\mu} \\ \hat{\tau_1} \\ \hat{\tau_2} \end{pmatrix} = \begin{pmatrix} 0 \\ (y_1 + y_2 + y_3)/3 \\ y_4 \end{pmatrix}. \tag{0.185}$$

□

The vector $\boldsymbol{\beta}$ itself is estimable if there exists a matrix $\mathbf{A}$ such that

$$\mathrm{E}(\mathbf{Ay}) = \mathbf{AX}\boldsymbol{\beta} = \boldsymbol{\beta} \quad \text{for all } \boldsymbol{\beta} \in \mathbb{R}^p, \tag{0.186}$$

i.e.,

$$\begin{aligned}\boldsymbol{\beta} \text{ is estimable } &\iff \exists\, \mathbf{A}\colon \mathbf{I}_p = \mathbf{AX} = \mathbf{X}'\mathbf{A}' \iff \mathbb{R}^p = \mathscr{C}(\mathbf{X}') \\ &\iff \mathbf{X} \text{ has full column rank}\,. \end{aligned} \tag{0.187}$$

The vector $\mathbf{X}\boldsymbol{\beta}$ is (trivially) estimable because $\mathrm{E}(\mathbf{I}\cdot\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$.

The estimator $\mathbf{Ay}$ is unbiased for $\mathbf{X}\boldsymbol{\beta}$ if $\mathbf{AX}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta}$ for all $\boldsymbol{\beta} \in \mathbb{R}^p$, i.e.,

$$\mathbf{AX} = \mathbf{X}\,. \tag{0.188}$$

It is of interest to note that if we consider such a matrix $\mathbf{A}$ which satisfies the condition

$$\mathscr{C}(\mathbf{A}) \subset \mathscr{C}(\mathbf{X})\,, \quad \text{i.e.,} \quad \mathbf{A} = \mathbf{XL} \quad \text{for some } \mathbf{L}\,, \tag{0.189}$$

then we can characterize the unbiasedness of $\mathbf{Ay} = \mathbf{XLy}$ as follows:

$$\mathbf{XLy} \text{ is unbiased estimator of } \mathbf{X}\boldsymbol{\beta} \iff \mathbf{L} \in \{\mathbf{X}^-\}\,. \tag{0.190}$$

Photograph 0.10 Jerzy K. Baksalary (Dortmund, 2003).

We might emphasize at this point that we overlook so called natural restrictions that occur when $\mathbf{V}$ is singular; they concern the freedom of $\boldsymbol{\beta}$ to vary through the whole $\mathbb{R}^p$. We refer to discussion in Puntanen & Styan (1990), Christensen (1990), Farebrother (1990), Harville (1990b), Baksalary, Rao & Markiewicz (1992), and Tian, Beisiegel, Dagenais & Haines (2008).

Best Linear Unbiased Estimator

A linear estimator $\mathbf{Gy}$ is an unbiased estimator of $\mathbf{X}\boldsymbol{\beta}$ if

$$\mathrm{E}(\mathbf{Gy}) = \mathbf{X}\boldsymbol{\beta} \quad \text{for all } \boldsymbol{\beta} \in \mathbb{R}^p, \tag{0.191}$$

and it is the *best linear unbiased estimator* (BLUE) of $\mathbf{X}\boldsymbol{\beta}$ if it has the smallest covariance matrix (in the Löwner sense) among all unbiased linear estimators:

$$\operatorname{cov}(\mathbf{Gy}) \leq_{\mathsf{L}} \operatorname{cov}(\mathbf{By}) \quad \text{for all } \mathbf{By} \text{ such that } \mathrm{E}(\mathbf{By}) = \mathbf{X}\boldsymbol{\beta}\,. \tag{0.192}$$

Since the unbiasedness of $\mathbf{Gy}$ means that $\mathbf{GX} = \mathbf{X}$, and under the model $\{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{V}\}$ we have $\operatorname{cov}(\mathbf{Gy}) = \sigma^2\mathbf{GVG}'$, we can rewrite (0.192) as

$$\mathbf{GVG}' \leq_{\mathsf{L}} \mathbf{BVB}' \quad \text{for all } \mathbf{B} \text{ such that } \mathbf{BX} = \mathbf{X}\,. \tag{0.193}$$

We denote the BLUE of $\mathbf{X}\boldsymbol{\beta}$ as

$$\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta}) = \widetilde{\mathbf{X}\boldsymbol{\beta}}\,, \tag{0.194a}$$

or

$$\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta}) = \mathbf{X}\tilde{\boldsymbol{\beta}}\,. \tag{0.194b}$$

Sometimes it is convenient to denote $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ and $\mathrm{BLUE}(\boldsymbol{\mu}) = \tilde{\boldsymbol{\mu}}$.

Remark 0.3. A gentle warning regarding notation (0.194b) may be worth giving. Namely in (0.194b) $\tilde{\boldsymbol{\beta}}$ refers now to any vector such that $\mathbf{X}\tilde{\boldsymbol{\beta}}$ is the $\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta})$. The vector $\tilde{\boldsymbol{\beta}}$ in (0.194b) need *not* be the $\mathrm{BLUE}(\boldsymbol{\beta})$—the parameter vector $\boldsymbol{\beta}$ may not even be estimable in which case it cannot have the BLUE! □

Let $\mathbf{X}$ have full column rank. Then $\boldsymbol{\beta}$ is estimable and $\mathbf{Ay}$ is the BLUE of $\boldsymbol{\beta}$ if

$$\mathbf{AVA}' \leq_{\mathsf{L}} \mathbf{BVB}' \quad \text{for all } \mathbf{B} \text{ such that } \mathbf{BX} = \mathbf{I}_p\,. \tag{0.195}$$

We denote the BLUE of $\boldsymbol{\beta}$ as $\mathrm{BLUE}(\boldsymbol{\beta}) = \tilde{\boldsymbol{\beta}}$. Because the Löwner ordering is so strong ordering, see (0.54) (p. 12), we have

$$\mathrm{var}(\tilde{\beta}_i) \leq \mathrm{var}(\beta_i^*), \quad i = 1,\dots,p\,, \tag{0.196a}$$

$$\mathrm{tr}\,\mathrm{cov}(\tilde{\boldsymbol{\beta}}) \leq \mathrm{tr}\,\mathrm{cov}(\boldsymbol{\beta}^*)\,, \tag{0.196b}$$

$$\det\mathrm{cov}(\tilde{\boldsymbol{\beta}}) \leq \det\mathrm{cov}(\boldsymbol{\beta}^*)\,, \tag{0.196c}$$

$$\mathrm{ch}_i[\mathrm{cov}(\tilde{\boldsymbol{\beta}})] \leq \mathrm{ch}_i[\mathrm{cov}(\boldsymbol{\beta}^*)]\,, \quad i = 1,\dots,p\,, \tag{0.196d}$$

$$\|\mathrm{cov}(\tilde{\boldsymbol{\beta}})\|_F \leq \|\mathrm{cov}(\boldsymbol{\beta}^*)\|_F\,, \tag{0.196e}$$

$$\|\mathrm{cov}(\tilde{\boldsymbol{\beta}})\|_2 \leq \|\mathrm{cov}(\boldsymbol{\beta}^*)\|_2\,, \tag{0.196f}$$

for any $\boldsymbol{\beta}^*$ which is a linear unbiased estimator of $\boldsymbol{\beta}$. Above $\|\cdot\|_F$ and $\|\cdot\|_2$ refer to the Frobenius norm and the spectral norm, respectively.

Consider now the model $\{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}\}$, where, for simplicity, we assume that $\sigma^2 = 1$. Then

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{X}^{+}\mathbf{y}\,, \quad \mathrm{cov}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}. \tag{0.197}$$

Let $\mathbf{By}$ be another linear unbiased estimator of $\boldsymbol{\beta}$, i.e., $\mathbf{B}$ satisfies $\mathbf{BX} = \mathbf{I}_p$. Then

$$\begin{aligned}\mathrm{cov}(\mathbf{By} - \hat{\boldsymbol{\beta}}) &= \mathrm{cov}(\mathbf{By}) + \mathrm{cov}(\hat{\boldsymbol{\beta}}) - \mathrm{cov}(\mathbf{By}, \hat{\boldsymbol{\beta}}) - \mathrm{cov}(\hat{\boldsymbol{\beta}}, \mathbf{By}) \\ &= \mathbf{BB}' + (\mathbf{X}'\mathbf{X})^{-1} - \mathbf{B}(\mathbf{X}^{+})' - \mathbf{X}^{+}\mathbf{B}' \\ &= \mathbf{BB}' + (\mathbf{X}'\mathbf{X})^{-1} - \mathbf{BX}(\mathbf{X}'\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{B}' \\ &= \mathbf{BB}' - (\mathbf{X}'\mathbf{X})^{-1},\end{aligned} \tag{0.198}$$

which implies

$$\mathbf{B}\mathbf{B}' - (\mathbf{X}'\mathbf{X})^{-1} = \operatorname{cov}(\mathbf{B}\mathbf{y}) - \operatorname{cov}(\hat{\boldsymbol{\beta}}) = \operatorname{cov}(\mathbf{B}\mathbf{y} - \hat{\boldsymbol{\beta}}) \geq_{\mathsf{L}} \mathbf{0}\,, \tag{0.199}$$

where the last (Löwner) inequality follows from the fact that every covariance matrix is nonnegative definite. Now (0.199) means that we have the Löwner ordering

$$\operatorname{cov}(\hat{\boldsymbol{\beta}}) \leq_{\mathsf{L}} \operatorname{cov}(\mathbf{B}\mathbf{y}) \quad \text{for all } \mathbf{B}\text{: } \mathbf{B}\mathbf{X} = \mathbf{I}_p\,. \tag{0.200}$$

Thus we have proved a simple version of the Gauss–Markov theorem, which says that

$$\operatorname{OLSE}(\boldsymbol{\beta}) = \operatorname{BLUE}(\boldsymbol{\beta}) \quad \text{under } \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}\}\,. \tag{0.201}$$

A generalized version of the Gauss–Markov theorem appears in (0.202) below in the next subsection and more thoroughly in Theorem 10 (p. 216). For a positive definite $\mathbf{V}$, see also Exercise 0.16 (p. 50).

The Fundamental BLUE Equation

What is the necessary and sufficient condition that $\mathbf{G}$ has to satisfy for $\mathbf{G}\mathbf{y}$ to be the BLUE of $\mathbf{X}\boldsymbol{\beta}$ under the model $\{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{V}\}$? The answer is given in Theorem 10 (p. 216):

$$\text{the fundamental BLUE equation:} \quad \mathbf{G}(\mathbf{X} : \mathbf{V}\mathbf{X}^{\perp}) = (\mathbf{X} : \mathbf{0})\,. \tag{0.202}$$

The corresponding condition for $\mathbf{F}\mathbf{y}$ to be the BLUE of an estimable parametric function $\mathbf{K}'\boldsymbol{\beta}$ is

$$\mathbf{F}(\mathbf{X} : \mathbf{V}\mathbf{X}^{\perp}) = (\mathbf{K}' : \mathbf{0})\,. \tag{0.203}$$

It is obvious that in (0.202) the matrix $\mathbf{X}^{\perp}$ can be replaced with $\mathbf{M} = \mathbf{I}_n - \mathbf{H} = \mathbf{I}_n - \mathbf{P}_{\mathbf{X}}$, and thus (0.202) becomes

$$\mathbf{G}(\mathbf{X} : \mathbf{V}\mathbf{M}) = (\mathbf{X} : \mathbf{0})\,. \tag{0.204}$$

For references to (0.202), see, e.g., Drygas (1970, p. 55) and Rao (1973b, p. 282).

Clearly if $\mathbf{G}_1$ and $\mathbf{G}_2$ satisfy (0.204), then

$$\mathbf{G}_1(\mathbf{X} : \mathbf{V}\mathbf{M}) = \mathbf{G}_2(\mathbf{X} : \mathbf{V}\mathbf{M})\,. \tag{0.205}$$

In particular, if $\mathbf{X}$ is partitioned as $\mathbf{X} = (\mathbf{X}_1 : \mathbf{X}_2)$, where $\mathbf{X}_1$ has p_1 and $\mathbf{X}_2$ has p_2 columns and

$$\mathbf{K}' = (\mathbf{X}_1 : \mathbf{0})\,, \quad \mathbf{K}'\boldsymbol{\beta} = \mathbf{X}_1\boldsymbol{\beta}_1\,, \tag{0.206}$$

then (0.203) becomes, assuming that $\mathbf{X}_1\boldsymbol{\beta}_1$ is estimable,

$$\mathbf{F}(\mathbf{X}_1 : \mathbf{X}_2 : \mathbf{VM}) = (\mathbf{X}_1 : \mathbf{0} : \mathbf{0})\,. \tag{0.207}$$

Hence we can also write

$$\mathbf{Gy} = \mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta}) \iff \mathbf{G} \in \{\, \mathbf{P}_{\mathbf{X}|\mathbf{VM}} \,\}\,, \tag{0.208a}$$

$$\mathbf{Fy} = \mathrm{BLUE}(\mathbf{X}_1\boldsymbol{\beta}_1) \iff \mathbf{F} \in \{\, \mathbf{P}_{\mathbf{X}_1|(\mathbf{X}_2:\mathbf{VM})} \,\}\,. \tag{0.208b}$$

The parametric function $\mathbf{X}_1\boldsymbol{\beta}_1$ is estimable if and only if the column spaces $\mathscr{C}(\mathbf{X}_1)$ and $\mathscr{C}(\mathbf{X}_2)$ are disjoint; see Proposition 16.1 (p. 345).

It is interesting to observe that the OLSE of $\mathbf{X}\boldsymbol{\beta}$ is defined as $\mathbf{Ay}$ if $\mathbf{A}$ satisfies

$$\mathbf{A}(\mathbf{X} : \mathbf{X}^{\perp}) = (\mathbf{X} : \mathbf{0})\,. \tag{0.209}$$

Similarly $\mathbf{By}$ is the OLSE of the estimable parametric function $\mathbf{K}'\boldsymbol{\beta}$ if

$$\mathbf{B}(\mathbf{X} : \mathbf{X}^{\perp}) = (\mathbf{K}' : \mathbf{0})\,. \tag{0.210}$$

Conditions (0.209) and (0.210) are simply the versions of (0.202) and (0.203) when $\mathbf{V}$ is replaced with $\mathbf{I}_n$.

For example, consider the estimation of $\mathbf{k}'\boldsymbol{\beta}$ via $\mathbf{a}'\mathbf{y}$ under the model $\{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \mathbf{V}\}$. Then $\mathbf{a}_*'\mathbf{y}$ is the BLUE of $\mathbf{k}'\boldsymbol{\beta}$ if and only if

$$\mathbf{a}_*'\mathbf{V}\mathbf{a}_* = \min_{\mathbf{a}} \mathbf{a}'\mathbf{V}\mathbf{a} = \min_{\mathbf{a}} \|\mathbf{a}\|_{\mathbf{V}}^2 \quad \text{subject to } \mathbf{X}'\mathbf{a} = \mathbf{k}\,, \tag{0.211}$$

i.e., we have to find such a solution to $\mathbf{X}'\mathbf{a} = \mathbf{k}$ which has the minimum norm when the inner product matrix is $\mathbf{V}$. Now according to (0.203), the minimizing $\mathbf{a}_*$ must satisfy

$$\mathbf{X}'\mathbf{a}_* = \mathbf{k} \quad \text{and} \quad \mathbf{a}_*'\mathbf{VM} = \mathbf{0}_n'\,, \tag{0.212}$$

which further can be written as

$$\mathbf{X}'\mathbf{a}_* = \mathbf{k} \quad \text{and} \quad \mathbf{V}\mathbf{a}_* \in \mathscr{C}(\mathbf{X})\,, \tag{0.213}$$

i.e.,

$$\mathrm{E}(\mathbf{a}_*'\mathbf{y}) = \mathbf{k}'\boldsymbol{\beta} \quad \text{and} \quad \mathrm{cov}(\mathbf{a}_*'\mathbf{y}, \mathbf{My}) = \mathbf{0}_n'\,. \tag{0.214}$$

All solutions to a consistent equation $\mathbf{X}'\mathbf{a} = \mathbf{k}$ can be generated through $(\mathbf{X}')^-\mathbf{k}$ by varying the generalized inverse $(\mathbf{X}')^-$. In order to get the BLUE, we should be able to pick up an appropriate $(\mathbf{X}')^-$. The correct choice is a so-called minimum norm g-inverse, denoted as $\mathbf{G} = (\mathbf{X}')^-_{m(\mathbf{V})}$. Then $\mathbf{k}'\mathbf{G}'\mathbf{y}$ is the BLUE for $\mathbf{k}'\boldsymbol{\beta}$.

One may wonder whether (0.202) actually has a solution for $\mathbf{G}$. No worries, the disjointness of $\mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{VM}) = \{\mathbf{0}\}$ guarantees the solution, see Section 5.1 (p. 123).

It is useful to observe that

$$\mathbf{Gy} = \mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta}) \implies \mathbf{HGy} = \mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta})\,, \tag{0.215}$$

which further means that

$$\mathbf{Gy} = \mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta}) \implies \text{there exists } \mathbf{L}\colon\, \mathbf{XLy} = \mathbf{Gy}\,. \tag{0.216}$$

Note that in equation (0.202) we can now interpret $\mathbf{G}$ as a projector—in the spirit of (0.19) (p. 6)—onto $\mathscr{C}(\mathbf{X})$ along $\mathscr{C}(\mathbf{VM})$. In particular, when $\mathbf{V}$ is positive definite, then $\mathbf{G}$ is the orthogonal projector onto $\mathscr{C}(\mathbf{X})$ when the inner product matrix is $\mathbf{V}^{-1}$. In this situation the projection direction is, see (0.58) (p. 13),

$$\mathscr{C}(\mathbf{VM}) = \mathscr{C}(\mathbf{X})^{\perp}_{\mathbf{V}^{-1}} = \mathscr{C}(\mathbf{I} - \mathbf{P}_{\mathbf{X};\mathbf{V}^{-1}}) = \mathscr{N}(\mathbf{X}'\mathbf{V}^{-1})\,, \tag{0.217}$$

where $\mathscr{C}(\mathbf{X})^{\perp}_{\mathbf{V}^{-1}}$ refers to the orthocomplement of $\mathscr{C}(\mathbf{X})$ with respect to the inner product defined via $\mathbf{V}^{-1}$.

When $\mathbf{V}$ is nonsingular then the BLUE of $\mathbf{X}\boldsymbol{\beta}$ is

$$\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta}) = \mathbf{X}\tilde{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{V}^{-1}\mathbf{y} = \mathbf{P}_{\mathbf{X};\mathbf{V}^{-1}}\mathbf{y}\,, \tag{0.218}$$

cf. Aitken (1935, p. 45), so that

$$\min_{\boldsymbol{\beta}}\,(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})\,. \tag{0.219}$$

We may recall here (see, e.g., Albert 1973, and Rao 1973b) three representations for the BLUE($\mathbf{X}\boldsymbol{\beta}$):

$$\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta}) = \mathbf{Hy} - \mathbf{HVM}(\mathbf{MVM})^{-}\mathbf{My} := \mathbf{G}_1\mathbf{y}\,, \tag{0.220a}$$

$$\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta}) = \mathbf{y} - \mathbf{VM}(\mathbf{MVM})^{-}\mathbf{My} := \mathbf{G}_2\mathbf{y}\,, \tag{0.220b}$$

$$\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta}) = \mathbf{X}(\mathbf{X}'\mathbf{W}^{-}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}^{-}\mathbf{y} := \mathbf{G}_3\mathbf{y}\,, \tag{0.220c}$$

where

$$\mathbf{W} = \mathbf{V} + \mathbf{XUX}', \tag{0.221a}$$

and $\mathbf{U}$ is an arbitrary matrix such that

$$\mathscr{C}(\mathbf{W}) = \mathscr{C}(\mathbf{X} : \mathbf{V})\,; \tag{0.221b}$$

see also Section 10.4 (p. 228). When $\mathbf{V}$ is nonsingular the matrix $\mathbf{G}$ such that $\mathbf{Gy}$ is the BLUE of $\mathbf{X}\boldsymbol{\beta}$ is unique, but when $\mathbf{V}$ is singular this may not be so. However, the numerical value of BLUE($\mathbf{X}\boldsymbol{\beta}$) is unique with probability 1. The matrices $\mathbf{X}$ and $\mathbf{V}$ can be of arbitrary rank but the model must be *consistent* in a sense that the realized value of random vector $\mathbf{y}$ satisfies

$$\text{the consistency condition of the model } \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \mathbf{V}\}\colon\quad \mathbf{y} \in \mathscr{C}(\mathbf{X} : \mathbf{V})\,. \tag{0.222}$$

The consistency condition means, for example, that whenever we have some statements where the random vector $\mathbf{y}$ is involved, these statements need to hold only for those values of $\mathbf{y}$ which belong to $\mathscr{C}(\mathbf{X} : \mathbf{V})$. Thus, if $\mathbf{Gy}$ and $\mathbf{Fy}$ are two presentations of the BLUE($\mathbf{X}\boldsymbol{\beta}$), then

$$\mathbf{Gy} = \mathbf{Fy} \quad \text{for all } \mathbf{y} \in \mathscr{C}(\mathbf{X} : \mathbf{V})\,. \tag{0.223}$$

In view of (0.216), there always exists a matrix $\mathbf{L}$ such that BLUE($\mathbf{X}\boldsymbol{\beta}$) can be expressed as $\mathbf{XLy}$. The unbiasedness condition (0.190) means that $\mathbf{L} \in \{\mathbf{X}^-\}$ [in which case $\operatorname{rank}(\mathbf{XL}) = \operatorname{rank}(\mathbf{X})$], but other than that, the vector $\mathbf{Ly}$ is not necessarily unique. We might denote $\mathbf{Ly} = \tilde{\boldsymbol{\beta}}$, and thereby BLUE($\mathbf{X}\boldsymbol{\beta}$) $= \mathbf{X}\tilde{\boldsymbol{\beta}}$, but it is important to realize that in this notation the vector $\tilde{\boldsymbol{\beta}}$ is unique (with probability 1) only if $\mathbf{X}$ has full column rank. In this case the vector $\boldsymbol{\beta}$ is estimable, i.e., it has an unbiased linear estimator, and one general expression for $\tilde{\boldsymbol{\beta}}$ is

$$\begin{aligned}\tilde{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{VM}(\mathbf{MVM})^-\mathbf{My} \\ &= \mathbf{X}^+\mathbf{y} - \mathbf{X}^+\mathbf{VM}(\mathbf{MVM})^-\mathbf{My}\,. \end{aligned} \tag{0.224}$$

We can also define $\tilde{\boldsymbol{\beta}}$ as a solution to the *generalized normal equations*

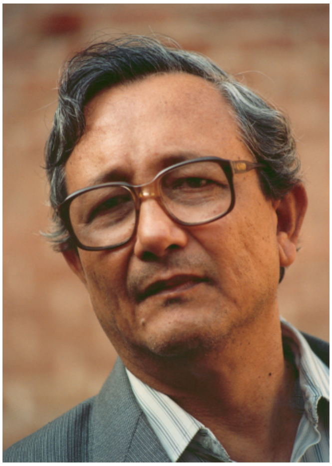
$$\mathbf{X}'\mathbf{W}^-\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{W}^-\mathbf{y}\,, \tag{0.225}$$

where $\mathbf{W}$ is defined as in (0.221); see, e.g., Baksalary & Puntanen (1989), Mitra (1973b), and Rao (1971c). The equation (0.225) is consistent for $\mathbf{y} \in \mathscr{C}(\mathbf{X} : \mathbf{V})$; see Exercise 12.1 (p. 288). If $\mathbf{W}$ is a symmetric nonnegative definite matrix and $\tilde{\boldsymbol{\beta}}$ is a solution to (0.225), then

$$\begin{aligned}&\min_{\boldsymbol{\beta}}\, (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{W}^-(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &\qquad = (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'\mathbf{W}^-(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})\,. \end{aligned} \tag{0.226}$$

Notice that once we decide to denote the BLUE of $\mathbf{X}\boldsymbol{\beta}$ as $\mathbf{X}\tilde{\boldsymbol{\beta}}$, then, in view of (0.220b) (p. 43), we also have

$$\begin{aligned}\mathbf{X}\tilde{\boldsymbol{\beta}} &= \mathbf{y} - \mathbf{VM}(\mathbf{MVM})^-\mathbf{My} \\ &= \mathbf{G}_2\mathbf{y}\,, \end{aligned} \tag{0.227}$$



Photograph 0.11 Sujit Kumar Mitra (New Delhi, 1992).

which may look somewhat confusing for a novice. However, as emphasized in Remark 0.3 (p. 40), (0.227) is simply a short notation for the fact that the matrix $\mathbf{G}_2$ satisfies the BLUE-equation (0.202) (p. 41). Perhaps notation $\widetilde{\mathbf{X}\boldsymbol{\beta}}$ would occasionally be preferable to $\mathbf{X}\tilde{\boldsymbol{\beta}}$.

In view of (0.227), we can define the BLUE's residual as

$$\tilde{\boldsymbol{\varepsilon}} = \mathbf{y} - \widetilde{\mathbf{X}\boldsymbol{\beta}} = \mathbf{VM}(\mathbf{MVM})^{-}\mathbf{My} = \mathbf{V}\dot{\mathbf{M}}\mathbf{y}\,, \tag{0.228}$$

where the matrix $\dot{\mathbf{M}}$ is defined as

$$\dot{\mathbf{M}} = \mathbf{M}(\mathbf{MVM})^{-}\mathbf{M}\,. \tag{0.229}$$

The matrix $\dot{\mathbf{M}}$ appears to be very handy in many considerations and hence we have devoted one chapter for it, see Chapter 15 (p. 317).

The weighted sum of squares of errors, which is needed when estimating σ^2, see, e.g., Rao (1971a,b), can now be written as

$$\begin{aligned}\mathrm{SSE}(\mathbf{V}) &= (\mathbf{y} - \widetilde{\mathbf{X}\boldsymbol{\beta}})'\mathbf{V}^{-}(\mathbf{y} - \widetilde{\mathbf{X}\boldsymbol{\beta}}) \\ &= (\mathbf{y} - \widetilde{\mathbf{X}\boldsymbol{\beta}})'\mathbf{W}^{-}(\mathbf{y} - \widetilde{\mathbf{X}\boldsymbol{\beta}}) \\ &= \tilde{\boldsymbol{\varepsilon}}'\mathbf{V}^{-}\tilde{\boldsymbol{\varepsilon}} = \mathbf{y}'\dot{\mathbf{M}}\mathbf{y}\,, \end{aligned} \tag{0.230}$$

where $\mathbf{W}$ is defined as in (0.221a). See also Section 15.9 (p. 338).

Finally we remark that on the basis of (0.220), it is easy to introduce the following general representations of the BLUE's covariance matrix:

$$\operatorname{cov}(\widetilde{\mathbf{X}\boldsymbol{\beta}}) = \mathbf{HVH} - \mathbf{HVM}(\mathbf{MVM})^{-}\mathbf{MVH} \tag{0.231a}$$

$$= \mathbf{V} - \mathbf{VM}(\mathbf{MVM})^{-}\mathbf{MV} \tag{0.231b}$$

$$= \mathbf{X}(\mathbf{X}'\mathbf{W}^{-}\mathbf{X})^{-}\mathbf{X}' - \mathbf{XUX}'. \tag{0.231c}$$

The covariance matrix of the BLUE's residual appears to be (confirm please!)

$$\begin{aligned}\operatorname{cov}(\tilde{\boldsymbol{\varepsilon}}) &= \operatorname{cov}(\mathbf{V}\dot{\mathbf{M}}\mathbf{y}) = \mathbf{V}\dot{\mathbf{M}}\mathbf{V}\dot{\mathbf{M}}'\mathbf{V} \\ &= \mathbf{VM}(\mathbf{MVM})^{-}\mathbf{MVM}[(\mathbf{MVM})^{-}]'\mathbf{MV} \\ &= \mathbf{VM}(\mathbf{MVM})^{-}\mathbf{MV} = \mathbf{V}\dot{\mathbf{M}}\mathbf{V}\,. \end{aligned} \tag{0.232}$$

If $\mathbf{X}$ has full column rank, then the expression for the covariance matrix of the BLUE of $\boldsymbol{\beta}$ can be written as

$$\begin{aligned}\operatorname{cov}(\tilde{\boldsymbol{\beta}}) &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\big[\mathbf{X}'\mathbf{VX} - \mathbf{X}'\mathbf{VM}(\mathbf{MVM})^{-}\mathbf{MVX}\big](\mathbf{X}'\mathbf{X})^{-1} \\ &= \operatorname{cov}(\hat{\boldsymbol{\beta}}) - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{VM}(\mathbf{MVM})^{-}\mathbf{MVX}(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned} \tag{0.233}$$

Example 0.7 (Getting a ticket for over-speeding). We end this chapter with a simple example in which the details are left to the reader.

Mr. Yrjö Kärkinen was stopped on a highway where the speed limit was 90 km/h. In the court Yrjö claimed that he was driving only 85 km/h. Judge Juhani Järvinen, dealing with this case, had called a witness, Ms. Maija Mäkinen, who told that she had been driving in the speed of 80 km/h when Yrjö heartlessly had passed her. She further explained that Yrjö had been driving 10 km/h faster than she. The traffic police Matti Virtanen who had caught Yrjö told that Yrjö had been driving 100 km/h and Maija 60 km/h.

(a) Identify the appropriate linear model assuming that all five observations are independent and have the same variance. Compute the ordinary least squares estimates for the speeds of Yrjö (β) and Maija (α). [$\hat{\beta} = 89.375$, $\hat{\alpha} = 73.125$]

(b) What about if Maija had claimed—instead of saying that Yrjö had been driving 10 km/h faster than she—that Yrjö had been driving 90 km/h?

(c) Suppose that the Maija's two observations are correlated (ϱ_1) and similarly Matti's two observations are correlated (ϱ_2). Write down the covariance matrix $\mathbf{V}$ of the five observations.

(d) Assuming that $\varrho_1 = \varrho_2 = 0.5$, compute the BLUEs for the speeds. Note that poor Yrjö now gets a ticket. [$\tilde{\beta} = 92.143$, $\tilde{\alpha} = 74.821$]

□

Exercises

These exercises are supposed to be warm-up exercises—to get rid of the possible rust in the reader's matrix engine.

0.1 (Contingency table, 2×2). Consider three frequency tables (contingency tables) below. In each table the row variable is x.

(a) Write up the original data matrices and calculate the correlation coefficients r_{xy}, r_{xz} and r_{xu}.

(b) What happens if the location (cell) of the zero frequency changes while other frequencies remain mutually equal?

(c) Explain why $r_{xy} = r_{xu}$ even if the u-values 2 and 5 are replaced with arbitrary a and b such that $a < b$.

		y	
		0	1
x	0	1	1
	1	0	1

	z	
	0	1
0	2	2
1	0	2

	u	
	2	5
0	1	1
1	0	1

0.2. Prove the following results concerning two dichotomous variables whose observed frequency table is given below.

$$\operatorname{var}_{\mathrm{s}}(y) = \frac{1}{n-1}\frac{\gamma\delta}{n} = \frac{n}{n-1} \cdot \frac{\delta}{n}\left(1 - \frac{\delta}{n}\right),$$

$$\operatorname{cov}_{\mathrm{s}}(x, y) = \frac{1}{n-1}\frac{ad - bc}{n}, \quad \operatorname{cor}_{\mathrm{s}}(x, y) = \frac{ad - bc}{\sqrt{\alpha\beta\gamma\delta}} = r\,,$$

$$\chi^2 = \frac{n(ad - bc)^2}{\alpha\beta\gamma\delta} = nr^2.$$

		y		
		0	1	total
x	0	a	b	α
	1	c	d	β
	total	γ	δ	n

0.3. Let $\mathbf{z} = \binom{x}{y}$ be a discrete 2-dimensional random vector which is obtained from the frequency table in Exercise 0.2 so that each observation has the same probability $1/n$. Prove that then

$$\mathrm{E}(y) = \frac{\delta}{n},\ \mathrm{var}(y) = \frac{\delta}{n}\left(1-\frac{\delta}{n}\right),\ \mathrm{cov}(x,y) = \frac{ad-bc}{n^2},\ \mathrm{cor}(x,y) = \frac{ad-bc}{\sqrt{\alpha\beta\gamma\delta}}.$$

0.4 (Continued ...). Show that in terms of the probabilities:

$$\begin{aligned} \mathrm{var}(y) &= p_{\cdot 1}p_{\cdot 2}\,, \\ \mathrm{cov}(x,y) &= p_{11}p_{22} - p_{12}p_{21}\,, \\ \mathrm{cor}(x,y) &= \frac{p_{11}p_{22} - p_{12}p_{21}}{\sqrt{p_{\cdot 1}p_{\cdot 2}p_{1\cdot}p_{2\cdot}}} = \varrho_{xy}\,. \end{aligned}$$

		y		
		0	1	total
x	0	p_{11}	p_{12}	$p_{1\cdot}$
	1	p_{21}	p_{22}	$p_{2\cdot}$
total		$p_{\cdot 1}$	$p_{\cdot 2}$	1

0.5 (Continued ...). Confirm:

$$\varrho_{xy} = 0 \Leftrightarrow \det\begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} = \det\begin{pmatrix} a & b \\ c & d \end{pmatrix} = 0 \Leftrightarrow \frac{p_{11}}{p_{21}} = \frac{p_{12}}{p_{22}} \Leftrightarrow \frac{a}{c} = \frac{b}{d}\,.$$

0.6 (Continued ...). Show, using (0.85) (p. 19), that the dichotomous random variables x and y are statistically independent if and only if $\varrho_{xy} = 0$. By the way, for interesting comments on 2×2 tables, see Speed (2008b).

0.7 (Continued, in a way ...). Consider dichotomous variables x and y whose values are A_1, A_2 and B_1, B_2, respectively, and suppose we have n observations from these variables. Let us define new variables in the following way:

$$\begin{aligned} x_1 &= 1 \text{ if } x \text{ has value } A_1, \text{ and } x_1 = 0 \text{ otherwise,} \\ x_2 &= 1 \text{ if } x \text{ has value } A_2, \text{ and } x_2 = 0 \text{ otherwise,} \end{aligned}$$

and let y_1 and y_2 be defined in the corresponding way with respect to the values B_1 and B_2. Denote the observed $n \times 4$ data matrix as $\mathbf{U} = (\mathbf{x}_1 : \mathbf{x}_2 : \mathbf{y}_1 : \mathbf{y}_2) = (\mathbf{X} : \mathbf{Y})$. We are interested in the statistical dependence of the variables x and y and hence we prepare the following frequency table (contingency table):

		y		
		B_1	B_2	total
x	A_1	f_{11}	f_{12}	r_1
	A_2	f_{21}	f_{22}	r_2
total		c_1	c_2	n

Let e_{ij} denote the expected frequency (for the usual χ^2-statistic for testing the independence) of the cell (i,j) and

$$e_{ij} = \frac{r_i c_j}{n}\,, \quad \mathbf{E} = (\mathbf{e}_1 : \mathbf{e}_2)\,, \quad \mathbf{F} = \begin{pmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \end{pmatrix} = (\mathbf{f}_1 : \mathbf{f}_2)\,,$$

and $\mathbf{c} = \binom{c_1}{c_2}$, $\mathbf{r} = \binom{r_1}{r_2}$. We may assume that all elements of $\mathbf{c}$ and $\mathbf{r}$ are nonzero. Confirm the following:

(a) $\operatorname{cor_d}(\mathbf{x}_1, \mathbf{x}_2) = -1$, $\operatorname{rank}(\mathbf{X} : \mathbf{Y}) \leq 3$, $\operatorname{rank}[\operatorname{cor_d}(\mathbf{X} : \mathbf{Y})] \leq 2$,

(b) $\mathbf{X}'\mathbf{1}_n = \mathbf{r}$, $\mathbf{Y}'\mathbf{1}_n = \mathbf{r}$, $\mathbf{X}'\mathbf{X} = \operatorname{diag}(\mathbf{r}) = \mathbf{D_r}$, $\mathbf{Y}'\mathbf{Y} = \operatorname{diag}(\mathbf{c}) = \mathbf{D_c}$,

(c) $\mathbf{X}'\mathbf{Y} = \mathbf{F}$, $\mathbf{E} = \mathbf{rc}'/n = \mathbf{X}'\mathbf{1}_n\mathbf{1}_n'\mathbf{Y}/n = \mathbf{X}'\mathbf{J}\mathbf{Y}$.

(d) The columns of $\mathbf{X}'\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}$ represent the conditional relative frequencies (distributions) of x.

(e) $\mathbf{F} - \mathbf{E} = \mathbf{X}'\mathbf{C}\mathbf{Y}$, where $\mathbf{C}$ is the centering matrix, and hence $\frac{1}{n-1}(\mathbf{F} - \mathbf{E})$ is the sample (cross)covariance matrix between the x- and y-variables.

0.8 (Continued . . .).

(a) Show that $\mathbf{X}'\mathbf{C}\mathbf{Y}$, $\mathbf{X}'\mathbf{C}\mathbf{X}$, and $\mathbf{Y}'\mathbf{C}\mathbf{Y}$ are double-centered; $\mathbf{A}_{n\times p}$ is said to be double-centered if $\mathbf{A}\mathbf{1}_p = \mathbf{0}_n$ and $\mathbf{A}'\mathbf{1}_n = \mathbf{0}_p$.

(b) Prove that $\mathbf{1}_n \in \mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{Y})$ and that it is possible that $\dim \mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{Y}) > 1$.

(c) Show, using the rule $\operatorname{rk}(\mathbf{CY}) = \operatorname{rk}(\mathbf{Y}) - \dim \mathscr{C}(\mathbf{Y}) \cap \mathscr{C}(\mathbf{C})^{\perp}$, see Theorem 5 (p. 121), that

$$\operatorname{rk}(\mathbf{Y}'\mathbf{C}\mathbf{Y}) = \operatorname{rk}(\mathbf{CY}) = c - 1 \quad \text{and} \quad \operatorname{rk}(\mathbf{X}'\mathbf{C}\mathbf{X}) = \operatorname{rk}(\mathbf{CX}) = r - 1\,,$$

where c and r refer to the number of categories of y and x, respectively; see Exercise 19.12 (p. 412) In this situation of course $c = r = 2$.

(d) Confirm that $(\mathbf{Y}'\mathbf{Y})^{-1}$ is a generalized inverse of $\mathbf{Y}'\mathbf{C}\mathbf{Y}$, i.e.,

$$\mathbf{Y}'\mathbf{C}\mathbf{Y} \cdot (\mathbf{Y}'\mathbf{Y})^{-1} \cdot \mathbf{Y}'\mathbf{C}\mathbf{Y} = \mathbf{Y}'\mathbf{C}\mathbf{Y}\,,$$
$$(\mathbf{D_c} - \tfrac{1}{n}\mathbf{cc}') \cdot \mathbf{D_c}^{-1} \cdot (\mathbf{D_c} - \tfrac{1}{n}\mathbf{cc}') = \mathbf{D_c} - \tfrac{1}{n}\mathbf{cc}'.$$

See also part (b) of Exercise 0.11 (p. 49), Exercise 4.9 (p. 120), and Exercise 19.12 (p. 412).

0.9 (Continued . . .).

(a) What is the interpretation of the matrix

$$\mathbf{G} = \sqrt{n}(\mathbf{X}'\mathbf{X})^{-1/2}\mathbf{X}'\mathbf{C}\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1/2} = \sqrt{n}\,\mathbf{D_r}^{-1/2}(\mathbf{F} - \mathbf{E})\mathbf{D_c}^{-1/2}\,?$$

(b) Convince yourself that the matrix

$$\mathbf{G}_* = \mathbf{D_r}^{-1/2}(\mathbf{F} - \mathbf{E})\mathbf{D_c}^{-1/2}$$

remains invariant if instead of frequencies we consider proportions so that the matrix $\mathbf{F}$ is replaced with $\frac{1}{n}\mathbf{F}$ and the matrices $\mathbf{E}$, $\mathbf{D_r}$ and $\mathbf{D_c}$ are calculated accordingly.

(c) Show that the χ^2-statistic for testing the independence of x and y can be written as

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(f_{ij}-e_{ij})^2}{e_{ij}} = \|\mathbf{G}\|_F^2 = \operatorname{tr}(\mathbf{G}'\mathbf{G}) = n\operatorname{tr}(\mathbf{P_X}\mathbf{C}\mathbf{P_Y}\mathbf{C})\,.$$

See also Exercise 19.13 (p. 413).

(d) Show that the contribution of the ith column of $\mathbf{F}$ on the χ^2, $\chi^2(\mathbf{f}_i)$, say, can be expressed as (a kind of squared Mahalanobis distance)

$$\chi^2(\mathbf{f}_i) = (\mathbf{f}_i - \mathbf{e}_i)'\mathbf{D}^{-1}(\mathbf{f}_i - \mathbf{e}_i)\,,$$

where

$$\mathbf{D} = \operatorname{diag}(\mathbf{e}_i) = \begin{pmatrix} e_{i1} & 0 \\ 0 & e_{i2} \end{pmatrix} = \begin{pmatrix} r_1 c_i/n & 0 \\ 0 & r_2 c_i/n \end{pmatrix} = c_i \begin{pmatrix} r_1/n & 0 \\ 0 & r_2/n \end{pmatrix}.$$

0.10 (Multinomial distribution). Consider the random vectors (for simplicity only three-dimensional)

$$\mathbf{z}_1 = \begin{pmatrix} z_{11} \\ z_{21} \\ z_{31} \end{pmatrix}, \dots, \mathbf{z}_m = \begin{pmatrix} z_{1m} \\ z_{2m} \\ z_{3m} \end{pmatrix}, \quad \mathbf{x} = \mathbf{z}_1 + \dots + \mathbf{z}_m\,,$$

where $\mathbf{z}_i$ are identically and independently distributed random vectors so that each $\mathbf{z}_i$ is defined so that only one element gets value 1 the rest being 0. Let $\mathrm{P}(z_{i1} = 1) = p_1$, $\mathrm{P}(z_{i2} = 1) = p_2$, and $\mathrm{P}(z_{i3} = 1) = p_3$ for $i = 1, \dots, m$; $p_1 + p_2 + p_3 = 1$, each $p_i > 0$, and denote $\mathbf{p} = (p_1, p_2, p_3)'$. Show that

$$\mathrm{E}(\mathbf{z}_i) = (p_1, p_2, p_3)' = \mathbf{p}\,, \quad \mathrm{E}(\mathbf{x}) = m\mathbf{p}\,,$$

and

$$\begin{aligned}\operatorname{cov}(\mathbf{z}_i) &= \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & -p_1p_3 \\ -p_2p_1 & p_2(1-p_2) & -p_2p_3 \\ -p_3p_1 & -p_3p_2 & p_3(1-p_3) \end{pmatrix} = \begin{pmatrix} p_1 & 0 & 0 \\ 0 & p_2 & 0 \\ 0 & 0 & p_3 \end{pmatrix} - \mathbf{p}\mathbf{p}' \\ &:= \mathbf{D_p} - \mathbf{p}\mathbf{p}' := \mathbf{\Sigma}\,, \quad \operatorname{cov}(\mathbf{x}) = m\mathbf{\Sigma}\,.\end{aligned}$$

Then $\mathbf{x}$ follows a multinomial distribution with parameters m and $\mathbf{p}$: $\mathbf{x} \sim \mathrm{Mult}(m, \mathbf{p})$.

0.11 (Continued ...). Confirm:

(a) $\boldsymbol{\Sigma}$ is double-centered (row and column sums are zero), singular and has rank 2.

(b) $\boldsymbol{\Sigma}\mathbf{D}_{\mathbf{p}}^{-1}\boldsymbol{\Sigma} = \boldsymbol{\Sigma}$, i.e., $\mathbf{D}_{\mathbf{p}}^{-1}$ is a generalized inverse of $\boldsymbol{\Sigma}$. Confirm that $\mathbf{D}_{\mathbf{p}}^{-1}$ does not necessarily satisfy any other Moore–Penrose conditions. See also Exercises 0.8 (p. 48) and 4.9 (p. 120).

(c) We can think (confirm ...) that the columns (or rows if we wish) of a contingency table are realizations of a multinomial random variable. Let $\mathbf{x}_1$ and $\mathbf{x}_2$ represent two columns (two observations) from such a random variable and assume that instead of the frequencies we consider proportions $\mathbf{y}_1 = \mathbf{x}_1/c_1$ and $\mathbf{y}_2 = \mathbf{x}_2/c_2$. Then $\mathbf{x}_i \sim \mathrm{Mult}(c_i, \mathbf{p})$ and $\mathrm{cov}(\mathbf{y}_i) = \frac{1}{c_i}\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma} = \mathbf{D}_{\mathbf{p}} - \mathbf{pp}'$, and the squared Mahalanobis distance, say M, between the vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ can be defined as follows:

$$\begin{aligned} M &= c_1c_2(c_1+c_2)^{-1}(\mathbf{x}_1-\mathbf{x}_2)'\boldsymbol{\Sigma}^{-}(\mathbf{x}_1-\mathbf{x}_2) \\ &= c_1c_2(c_1+c_2)^{-1}(\mathbf{x}_1-\mathbf{x}_2)'\mathbf{D}_{\mathbf{p}}^{-1}(\mathbf{x}_1-\mathbf{x}_2)\,. \end{aligned}$$

Neudecker (1997), Puntanen, Styan & Subak-Sharpe (1998), Greenacre (2007, p. 270).

0.12. Let $\mathbf{P}_{n\times n}$ be an idempotent matrix. Show that

$$\mathscr{C}(\mathbf{P}) \cap \mathscr{C}(\mathbf{I}_n - \mathbf{P}) = \{\mathbf{0}\} \quad \text{and} \quad \mathscr{C}(\mathbf{I}_n - \mathbf{P}) = \mathscr{N}(\mathbf{P})\,.$$

0.13. Confirm: $\quad \mathbf{A} \geq_{\mathsf{L}} \mathbf{B} \quad \text{and} \quad \mathbf{B} \geq_{\mathsf{L}} \mathbf{C} \quad \Longrightarrow \quad \mathbf{A} \geq_{\mathsf{L}} \mathbf{C}\,.$

0.14. Suppose that $\mathbf{x}$ and $\mathbf{y}$ are p-dimensional random vectors. Confirm:

(a) $\mathrm{cov}(\mathbf{x}+\mathbf{y}) = \mathrm{cov}(\mathbf{x}) + \mathrm{cov}(\mathbf{y}) \iff \mathrm{cov}(\mathbf{x},\mathbf{y}) = -\,\mathrm{cov}(\mathbf{y},\mathbf{x})$; if $\mathbf{A} = -\mathbf{A}'$, $\mathbf{A}$ is said to be skew-symmetric.

(b) $\mathrm{cov}(\mathbf{x}-\mathbf{y}) = \mathrm{cov}(\mathbf{x}) - \mathrm{cov}(\mathbf{y}) \iff \mathrm{cov}(\mathbf{x},\mathbf{y}) + \mathrm{cov}(\mathbf{y},\mathbf{x}) = 2\,\mathrm{cov}(\mathbf{y})$.

0.15. Let $\mathbf{Gy}$ and $\mathbf{Hy}$ be the BLUE and OLSE of $\mathbf{X}\boldsymbol{\beta}$, respectively, under $\{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \mathbf{V}\}$. Using (0.220) (p. 43), confirm the following:

$$\begin{aligned} \mathrm{cov}(\mathbf{Hy}-\mathbf{Gy}) &= \mathrm{cov}(\mathbf{Hy}) - \mathrm{cov}(\mathbf{Gy}) = \mathbf{HVM}(\mathbf{MVM})^{-}\mathbf{MVH}\,, \\ \mathrm{cov}(\mathbf{y}-\mathbf{Gy}) &= \mathrm{cov}(\tilde{\boldsymbol{\varepsilon}}) = \mathrm{cov}(\mathbf{y}) - \mathrm{cov}(\mathbf{Gy}) = \mathbf{VM}(\mathbf{MVM})^{-}\mathbf{MV}\,. \end{aligned}$$

0.16. Consider the model $\mathscr{M} = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{V}\}$, where $\mathbf{V}$ is positive definite and $\mathbf{X}$ has full column rank. Premultiplying the equation $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ by $\mathbf{V}^{-1/2}$, we get the model

$$\mathscr{M}_* = \{\mathbf{y}_*,\, \mathbf{X}_*\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\}\,, \quad \text{where } \mathbf{y}_* = \mathbf{V}^{-1/2}\mathbf{y}\,,\ \mathbf{X}_* = \mathbf{V}^{-1/2}\mathbf{X}\,.$$

Confirm that in view of (0.201) (p. 41),

$$\begin{aligned}\mathrm{OLSE}(\boldsymbol{\beta} \mid \mathscr{M}_*) &= (\mathbf{X}_*'\mathbf{X}_*)^{-1}\mathbf{X}_*'\mathbf{y}_* \\ &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} := \mathbf{G}\mathbf{y} = \mathrm{BLUE}(\boldsymbol{\beta} \mid \mathscr{M}_*)\,.\end{aligned}$$

Clearly $\mathbf{G}\mathbf{y}$ is an unbiased estimator of $\boldsymbol{\beta}$ under $\mathscr{M}$. Let $\mathbf{B}\mathbf{y}$ be another unbiased estimator of $\boldsymbol{\beta}$ under $\mathscr{M}$, i.e., $\mathbf{B}\mathbf{X} = \mathbf{I}_p$. Show that then $\mathbf{B}\mathbf{y} = \mathbf{B}\mathbf{V}^{1/2}\mathbf{V}^{-1/2}\mathbf{y} = \mathbf{B}\mathbf{V}^{1/2}\mathbf{y}_*$, and in light of (0.201),

$$\begin{aligned}\mathrm{cov}[(\mathbf{X}_*'\mathbf{X}_*)^{-1}\mathbf{X}_*'\mathbf{y}_*] &\leq_{\mathsf{L}} \mathrm{cov}(\mathbf{B}\mathbf{V}^{1/2}\mathbf{y}_*)\,, \\ (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} &\leq_{\mathsf{L}} \mathbf{B}\mathbf{V}\mathbf{B}' \quad \text{for all } \mathbf{B}\colon \mathbf{B}\mathbf{X} = \mathbf{I}_p\,,\end{aligned}$$

so that $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$ is the BLUE for $\boldsymbol{\beta}$ under $\mathscr{M}$. See also Exercise 10.8 (p. 261).

0.17 (Continued . . .). Show that the result of the previous Exercise 0.16 can be expressed as the equivalence of the following statements:

(a) $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \leq_{\mathsf{L}} \mathbf{K}\mathbf{K}'$ for all $\mathbf{K}\colon \mathbf{K}\mathbf{V}^{-1/2}\mathbf{X} = \mathbf{I}_p$,
(b) $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \leq_{\mathsf{L}} \mathbf{L}\mathbf{V}\mathbf{L}'$ for all $\mathbf{L}\colon \mathbf{L}\mathbf{X} = \mathbf{I}_p$.

0.18 (Continued . . .). Confirm that the result of Exercise 0.16 can be introduced also as follows: Denote $\mathbf{G}\mathbf{y} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$ and let $\mathbf{B}\mathbf{y}$ be an arbitrary linear unbiased estimator of $\boldsymbol{\beta}$. Then $\mathbf{G}\mathbf{X} = \mathbf{B}\mathbf{X} = \mathbf{I}_p$, so that $(\mathbf{G} - \mathbf{B})\mathbf{X} = \mathbf{0}$, i.e., $\mathbf{X}'(\mathbf{G} - \mathbf{B})' = \mathbf{0}$, which implies that $\mathscr{C}(\mathbf{G}' - \mathbf{B}') \subset \mathscr{N}(\mathbf{X}') = \mathscr{C}(\mathbf{X})^{\perp} = \mathscr{C}(\mathbf{M})$, where $\mathbf{M} = \mathbf{I}_n - \mathbf{P}_{\mathbf{X}}$. Hence

$$\mathbf{G}' - \mathbf{B}' = \mathbf{M}\mathbf{U} \quad \text{for some matrix } \mathbf{U}.$$

Premultiplying the above equation by $\mathbf{G}\mathbf{V}$ yields $\mathbf{G}\mathbf{V}(\mathbf{G}' - \mathbf{B}') = \mathbf{0}$, and thereby $\mathbf{G}\mathbf{V}\mathbf{G}' = \mathbf{G}\mathbf{V}\mathbf{B}' = \mathbf{B}\mathbf{V}\mathbf{G}'$, which further implies

$$(\mathbf{B} - \mathbf{G})\mathbf{V}(\mathbf{B} - \mathbf{G})' = \mathbf{B}\mathbf{V}\mathbf{B}' - \mathbf{G}\mathbf{V}\mathbf{G}' \geq_{\mathsf{L}} \mathbf{0}\,.$$

Toutenburg & Shalabh (2009, p. 572).

0.19. Consider a random matrix $\mathbf{U}_{n\times d} = (\mathbf{u}_1 : \ldots : \mathbf{u}_d) = (\mathbf{u}_{(1)} : \ldots : \mathbf{u}_{(n)})'$, where $\mathrm{cov}(\mathbf{u}_j, \mathbf{u}_k) = \sigma_{jk}\mathbf{V}$, $j, k = 1, \ldots, d$. Show that

$$\mathrm{cov}(\mathbf{u}_{(r)}, \mathbf{u}_{(s)}) = v_{rs}\boldsymbol{\Sigma}\,, \quad r, s = 1, \ldots, n\,.$$

0.20 (Block design). Using the notation of Example 0.3 (p. 31), create the matrix $\mathbf{T}$ in the following block designs:

$$d_{\#} = \begin{pmatrix} A & B & C \\ C & B & A \end{pmatrix}, \quad d_* = \begin{pmatrix} A & B & C \\ A & C & B \end{pmatrix}.$$

0.21. Consider the block model (0.140) (p. 31). Show that

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} \mathbf{1}_n'\mathbf{1}_n & \mathbf{1}_n'\mathbf{T} & \mathbf{1}_n'\mathbf{B} \\ \mathbf{T}'\mathbf{1}_n & \mathbf{T}'\mathbf{T} & \mathbf{T}'\mathbf{B} \\ \mathbf{B}'\mathbf{1}_n & \mathbf{B}'\mathbf{T} & \mathbf{B}'\mathbf{B} \end{pmatrix} = \begin{pmatrix} n & \mathbf{r}' & \mathbf{c}' \\ \mathbf{r} & \mathbf{D}_r & \mathbf{N} \\ \mathbf{c} & \mathbf{N}' & \mathbf{D}_c \end{pmatrix}$$

$$= \begin{pmatrix} n & r_1 & r_2 & \dots & r_t & c_1 & c_2 & \dots & c_b \\ r_1 & r_1 & 0 & \dots & 0 & & & & \\ r_2 & 0 & r_2 & \dots & 0 & & \mathbf{N} & & \\ \vdots & \vdots & & \ddots & & & & & \\ r_t & 0 & 0 & \dots & r_t & & & & \\ c_1 & & & & & c_1 & 0 & \dots & 0 \\ c_2 & & \mathbf{N}' & & & 0 & c_2 & \dots & 0 \\ \vdots & & & & & \vdots & \vdots & \ddots & \\ c_b & & & & & 0 & 0 & \dots & c_b \end{pmatrix},$$

where $\mathbf{D}_r = \operatorname{diag}(\mathbf{r})$, $\mathbf{D}_c = \operatorname{diag}(\mathbf{c})$, and $\mathbf{N} = \{n_{ij}\} \in \mathbb{R}^{t\times b}$ is a so-called incidence matrix; n_{ij} = the number of units receiving the ith treatment in the jth block. Thus $\mathbf{N}$ is a contingency table like $\mathbf{F}$ in Exercise 0.7 (p. 47); now x = the treatment and y = the block.

Bapat (2000, pp. 99-103).

0.22 (Orthogonal rotation). Denote

$$\mathbf{A}_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \qquad \mathbf{B}_\theta = \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix}.$$

Show that any 2×2 orthogonal matrix $\mathbf{Q}$ is $\mathbf{A}_\theta$ or $\mathbf{B}_\theta$ for some θ. Transformation $\mathbf{A}_\theta\mathbf{u}_{(i)}$ rotates the observation $\mathbf{u}_{(i)}$ by the angle θ in the counter-clockwise direction, and $\mathbf{B}_\theta\mathbf{u}_{(i)}$ makes the reflection of the observation $\mathbf{u}_{(i)}$ with respect to the line $y = \tan\left(\frac{\theta}{2}\right)x$. For a graphic illustration of a rotation, see Figure 19.3 (p. 404).

Horn & Johnson (1990, p. 68).

0.23 (Kronecker product). Confirm some of the following important (particularly in multivariate linear models) properties of the Kronecker product and the vec-operation:

(a) $\mathbf{a}'\otimes\mathbf{b} = \mathbf{b}\mathbf{a}' = \mathbf{b}\otimes\mathbf{a}'$,
(b) $(\mathbf{A}\otimes\mathbf{B})' = \mathbf{A}'\otimes\mathbf{B}'$,
(c) $(\mathbf{A}\otimes\mathbf{B})^{-1} = \mathbf{A}^{-1}\otimes\mathbf{B}^{-1}$, if $\mathbf{A}$ and $\mathbf{B}$ are nonsingular,
(d) $(\mathbf{A}\otimes\mathbf{B})^{-} = \mathbf{A}^{-}\otimes\mathbf{B}^{-}$,
(e) $\mathbf{A}\otimes\mathbf{B} \neq \mathbf{B}\otimes\mathbf{A}$, in general,
(f) $(\mathbf{A}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{D}) = \mathbf{AC}\otimes\mathbf{BD}$,
(g) $\operatorname{vec}(\mathbf{A}_{n\times p}\mathbf{B}_{p\times q}) = (\mathbf{B}'\otimes\mathbf{I}_n)\operatorname{vec}(\mathbf{A}) = (\mathbf{I}_q\otimes\mathbf{A})\operatorname{vec}(\mathbf{B})$,
(h) $\operatorname{vec}(\mathbf{a}\mathbf{b}') = \mathbf{b}\otimes\mathbf{a}$,
(i) $\operatorname{vec}(\mathbf{ABC}) = (\mathbf{C}'\otimes\mathbf{A})\operatorname{vec}(\mathbf{B})$; see Exercise 11.1 (p. 280),

(j) $\mathbf{P}_{\mathbf{A}\otimes\mathbf{B}} = \mathbf{P}_{\mathbf{A}} \otimes \mathbf{P}_{\mathbf{B}}$.

Henderson & Searle (1979, 1981b), Harville (1997, Ch. 16), Abadir & Magnus (2005, pp. 273–284), Seber (2008, §11.1).

0.24. Consider the set of numbers $\mathcal{A} = \{1, 2, \dots, N\}$ and let $x_1, x_2, \dots, x_p$ denote a random sample selected without a replacement from $\mathcal{A}$. Denote $y = x_1 + x_2 + \dots + x_p = \mathbf{1}_p'\mathbf{x}$. Confirm the following:

(a) $\operatorname{var}(x_i) = \frac{N^2-1}{12}\,, \quad \operatorname{cor}(x_i, x_j) = -\frac{1}{N-1} = \varrho\,, \quad i, j = 1, \dots, p\,,$

(b) $\operatorname{cor}^2(x_1, y) = \operatorname{cor}^2(x_1, x_1 + \dots + x_p) = \frac{1}{p} + \big(1 - \frac{1}{p}\big)\varrho\,.$

See also Section 10.6 (p. 234).

0.25 (Hotelling's T^2). Let $\mathbf{U}_1'$ and $\mathbf{U}_2'$ be independent random samples from $\mathrm{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $\mathrm{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, respectively. Denote $\mathbf{T}_i = \mathbf{U}_i'(\mathbf{I}_{n_i} - \mathbf{J}_{n_i})\mathbf{U}_i$, and $\mathbf{S}_* = \frac{1}{f}(\mathbf{T}_1 + \mathbf{T}_2)$, where $f = n_1 + n_2 - 2$. Confirm that

$$T^2 = \tfrac{n_1 n_2}{n_1+n_2}(\bar{\mathbf{u}}_1 - \bar{\mathbf{u}}_2)'\mathbf{S}_*^{-1}(\bar{\mathbf{u}}_1 - \bar{\mathbf{u}}_2) \sim \mathrm{T}^2(p, n_1 + n_2 - 2)\,,$$

where $\mathrm{T}^2(a, b)$ refers to the Hotelling's T^2 distribution; see (0.128) (p. 26). It can be shown that if $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, then

$$\tfrac{n_1+n_2-p-1}{(n_1+n_2-2)p}\, T^2 \sim \mathrm{F}(p, n_1 + n_2 - p - 1).$$

0.26 (Continued ...). Show that if $n_1 = 1$, then Hotelling's T^2 becomes

$$T^2 = \tfrac{n_2}{n_2+1}(\mathbf{u}_{(1)} - \bar{\mathbf{u}}_2)'\mathbf{S}_2^{-1}(\mathbf{u}_{(1)} - \bar{\mathbf{u}}_2)\,.$$

0.27 (Finiteness matters). Throughout this book, we assume that the expectations, variances and covariances that we are dealing with are finite. Then, for example, independence of the random variables x and y implies that $\operatorname{cor}(x, y) = 0$. The reader is encouraged to have a look at the paper by Mukhopadhyay (2010), which provides counterexamples to this implication in situations when the finiteness is not holding.

0.28 (Magic square). Confirm that the matrix

$$\mathbf{A} = \begin{pmatrix} 16 & 3 & 2 & 13 \\ 5 & 10 & 11 & 8 \\ 9 & 6 & 7 & 12 \\ 4 & 15 & 14 & 1 \end{pmatrix},$$

appearing in Philatelic Item 0.1 (p. 54), is a *magic square*, i.e., a $k \times k$ array such that the numbers in every row, column and in each of the two main diagonals add up to the same *magic sum*, 34 in this case; here $k = 4$. The matrix $\mathbf{A}$ here defines a *classic magic square* since the entries in $\mathbf{A}$ are the consecutive integers $1, 2, \dots, k^2$.

Show that the Moore–Penrose inverse

$$\mathbf{A}^{+} = \frac{1}{34 \cdot 80}\begin{pmatrix} 275 & -201 & -167 & 173 \\ 37 & -31 & -65 & 139 \\ -99 & 105 & 71 & 3 \\ -133 & 207 & 241 & -235 \end{pmatrix}$$

is also a magic square (though not a classic magic square) and that its magic sum is 1/34.

There are many patterns of numbers in **A** and in $\mathbf{A}^{+}$ which add up to the magic sum (34 for **A** and 1/34 for $\mathbf{A}^{+}$); see Trenkler & Trenkler (2001). For more about "magic generalized inverses", see Chu, Drury, Styan & Trenkler (2010).

Philatelic Item 0.1 The magic square defined by the matrix **A** in Exercise 0.28 appears in Albrecht Dürer's copper-plate engraving *Melencolia I* which is shown on philatelic sheetlets from Aitutaki (Cook Islands) 1986, *Scott* 391, and Mongolia 1978, *Scott* 1039.

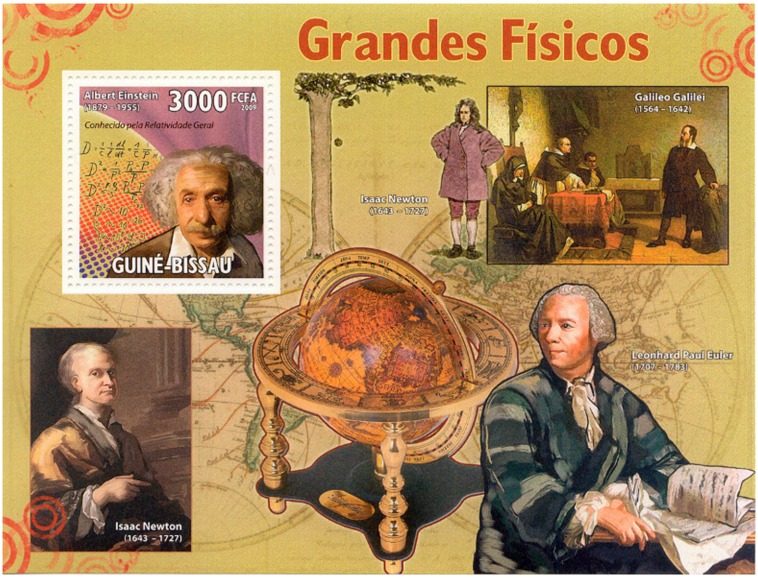
0.29 (Magic square, continued …). Confirm that the matrix

$$\mathbf{B} = \begin{pmatrix} 44 & 54 & 55 & 41 \\ 49 & 47 & 46 & 52 \\ 45 & 51 & 50 & 48 \\ 56 & 42 & 43 & 53 \end{pmatrix}$$

is a magic square. Find a scalar c so that $\mathbf{B} = \mathbf{PAQ} + c\mathbf{E}$, where **A** is the Dürer magic square in Exercise 0.28, while **E** is the 4×4 matrix with every element equal to 1 and the matrices $\mathbf{P}$ and $\mathbf{Q}$ are permutation matrices. The matrix $\mathbf{P}_{n\times n}$ is a *permutation matrix* if it is obtained from $\mathbf{I}_n$ by rearranging its columns.

The magic square represented by the matrix $\mathbf{B}$ here is associated with a porcelain plate made in Jingdezhen, China, ca. 1775, and presented to Queen Mary (1867–1953) in 1906; the plate is now in the Victoria and Albert Museum in London. For further details see Cheng (1984).

For more about magic squares we recommend the books by Benson & Jacoby (1976) and Pickover (2002).



Philatelic Item 0.2 Philatelic sheetlet from Guinea-Bissau 2009 depicting Albert Einstein (1879–1955), Leonhard Paul Euler (1707–1783), Galileo Galilei (1564–1642), and Isaac Newton (1643–1727). In addition to contributions involving magic squares and Latin squares, see Styan (2007), Euler made important discoveries in fields as diverse as infinitesimal calculus and graph theory; he is also renowned for his work in mechanics, fluid dynamics, optics, and astronomy.

Opening Figures: Continued—see page 2

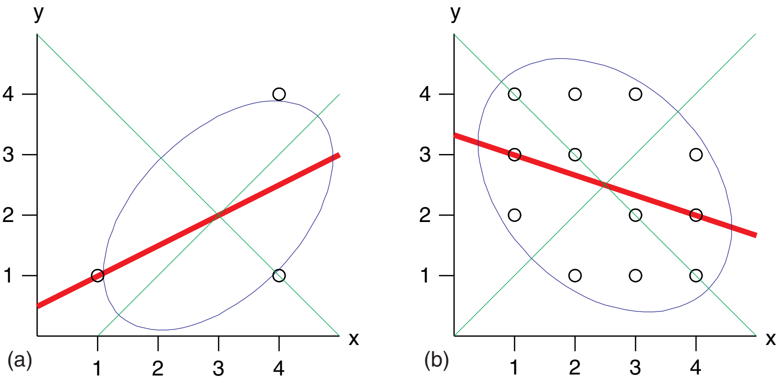


Figure 0.8 Completed Figures 0.1a and 0.1b. Regression line a bit thicker. The confidence ellipses (constant-density contours) are based on the assumption that samples are from $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are equal to their sample values. The first major axis of the ellipse will be along the 135° (slope -1) line through $\boldsymbol{\mu}$, and it is the line satisfying the minimization Task 2 on page 1. See also Figure 19.3 (p. 404).

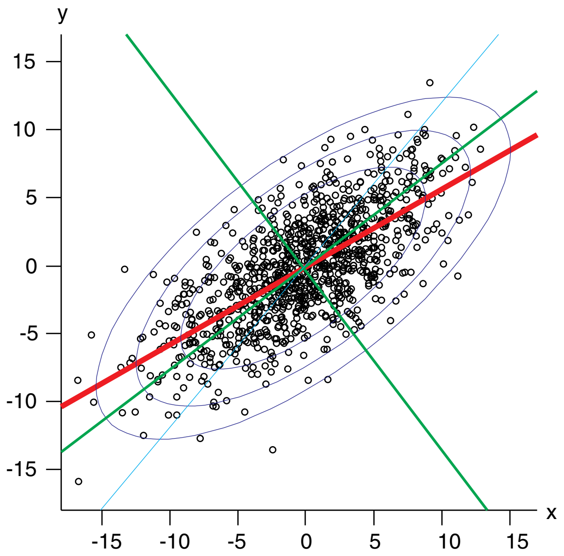


Figure 0.9 Completed Figure 0.2: Observations from $N_2(\mathbf{0}, \boldsymbol{\Sigma})$; $\sigma_x = 5$, $\sigma_y = 4$, $\varrho_{xy} = 0.7$. Regression line has the slope $\hat{\beta}_1 \approx \varrho_{xy}\sigma_y/\sigma_x$. Also the regression line of x on y is drawn. The direction of the first major axis of the contour ellipse is determined by $\mathbf{t}_1$, the first eigenvector of $\boldsymbol{\Sigma}$.

Chapter 1
Easy Column Space Tricks

> *If you've heard this story before, don't stop me, because I'd like to hear it again.*
>
> Groucho Marx

The column space of an $n \times m$ matrix $\mathbf{A}$,

$$\mathscr{C}(\mathbf{A}_{n\times m}) = \{\, \mathbf{y} \in \mathbb{R}^n : \text{there exists } \mathbf{x} \in \mathbb{R}^m \text{ such that } \mathbf{y} = \mathbf{A}\mathbf{x} \,\}\,, \tag{1.1}$$

and, correspondingly, the null space of $\mathbf{A}$,

$$\mathscr{N}(\mathbf{A}_{n\times m}) = \{\, \mathbf{x} \in \mathbb{R}^m : \mathbf{A}\mathbf{x} = \mathbf{0} \,\}\,, \tag{1.2}$$

are in every-day use throughout this book. In this chapter we take a good look at some of their properties, most of them rather elementary. Our experience is that decent steps in linear models are slow to take unless a reasonable set of column space tricks is in the immediate access. Some of the results that we go through are already likely to be in the reader's toolbox. However, there cannot be harm in repeating these helpful rules and going through their proofs.

Theorem 1 (Column space properties). *For conformable matrices* $\mathbf{A}$ *and* $\mathbf{B}$*, the following statements hold:*

(a) $\mathscr{N}(\mathbf{A}) = \mathscr{N}(\mathbf{A}'\mathbf{A})$,

(b) $\mathscr{C}(\mathbf{A})^{\perp} = \mathscr{C}(\mathbf{A}^{\perp}) = \mathscr{N}(\mathbf{A}')$,

(c) $\mathscr{C}(\mathbf{A}) \subset \mathscr{C}(\mathbf{B}) \iff \mathscr{C}(\mathbf{B})^{\perp} \subset \mathscr{C}(\mathbf{A})^{\perp}$,

(d) $\mathscr{C}(\mathbf{A}) = \mathscr{C}(\mathbf{A}\mathbf{A}')$,

(e) $\operatorname{rank}(\mathbf{A}) = \operatorname{rank}(\mathbf{A}') = \operatorname{rank}(\mathbf{A}\mathbf{A}') = \operatorname{rank}(\mathbf{A}'\mathbf{A})$,

(f) $\mathbb{R}^n = \mathscr{C}(\mathbf{A}_{n\times m}) \boxplus \mathscr{C}(\mathbf{A}_{n\times m})^{\perp} = \mathscr{C}(\mathbf{A}) \boxplus \mathscr{N}(\mathbf{A}')$,

(g) $\operatorname{rank}(\mathbf{A}_{n\times m}) = n - \operatorname{rank}(\mathbf{A}^{\perp}) = n - \dim \mathscr{N}(\mathbf{A}')$,

(h) $\mathscr{C}(\mathbf{A} : \mathbf{B}) = \mathscr{C}(\mathbf{A}) + \mathscr{C}(\mathbf{B})$,

(i) $\operatorname{rank}(\mathbf{A} : \mathbf{B}) = \operatorname{rank}(\mathbf{A}) + \operatorname{rank}(\mathbf{B}) - \dim \mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B})$,

(j) $\operatorname{rank}\begin{pmatrix}\mathbf{A}\\ \mathbf{B}\end{pmatrix} = \operatorname{rank}(\mathbf{A}) + \operatorname{rank}(\mathbf{B}) - \dim \mathscr{C}(\mathbf{A}') \cap \mathscr{C}(\mathbf{B}')$,

S. Puntanen et al., *Matrix Tricks for Linear Statistical Models: Our Personal Top Twenty*,
DOI 10.1007/978-3-642-10473-2_2,

(k) $\mathscr{C}(\mathbf{A}+\mathbf{B}) \subset \mathscr{C}(\mathbf{A}) + \mathscr{C}(\mathbf{B}) = \mathscr{C}(\mathbf{A} : \mathbf{B})$,

(l) $\operatorname{rank}(\mathbf{A}+\mathbf{B}) \leq \operatorname{rank}(\mathbf{A} : \mathbf{B}) \leq \operatorname{rank}(\mathbf{A}) + \operatorname{rank}(\mathbf{B})$,

(m) $\mathscr{C}(\mathbf{A} : \mathbf{B})^{\perp} = \mathscr{C}(\mathbf{A})^{\perp} \cap \mathscr{C}(\mathbf{B})^{\perp}$.

Proof. (a) First,

$$\mathbf{x} \in \mathscr{N}(\mathbf{A}) \implies \mathbf{A}\mathbf{x} = \mathbf{0} \implies \mathbf{A}'\mathbf{A}\mathbf{x} = \mathbf{0} \implies \mathbf{x} \in \mathscr{N}(\mathbf{A}'\mathbf{A})\,. \tag{1.3a}$$

To go the other way,

$$\begin{aligned}\mathbf{x} \in \mathscr{N}(\mathbf{A}'\mathbf{A}) &\implies \mathbf{A}'\mathbf{A}\mathbf{x} = \mathbf{0} \implies \mathbf{x}'\mathbf{A}'\mathbf{A}\mathbf{x} = 0 \\ &\implies \mathbf{A}\mathbf{x} = \mathbf{0} \implies \mathbf{x} \in \mathscr{N}(\mathbf{A})\,.\end{aligned} \tag{1.3b}$$

(b) comes from

$$\begin{aligned}\mathscr{C}(\mathbf{A}_{n\times m})^{\perp} &= \{\, \mathbf{y} \in \mathbb{R}^n : \mathbf{y}'\mathbf{A}\mathbf{z} = 0 \text{ for all } \mathbf{z} \in \mathbb{R}^m \,\} \\ &= \{\, \mathbf{y} \in \mathbb{R}^n : \mathbf{y}'\mathbf{A} = \mathbf{0}' \,\} = \mathscr{N}(\mathbf{A}')\,.\end{aligned} \tag{1.4}$$

(c) We see that

$$\begin{aligned}\mathscr{C}(\mathbf{A}) = \mathscr{C}(\mathbf{a}_1 : \ldots : \mathbf{a}_m) \subset \mathscr{C}(\mathbf{B}) &\implies \exists\, \mathbf{f}_i \colon \mathbf{a}_i = \mathbf{B}\mathbf{f}_i,\ i = 1,\ldots,m \\ &\implies \mathbf{A} = \mathbf{B}\mathbf{F} \text{ for some } \mathbf{F} \implies \mathbf{A}' = \mathbf{F}'\mathbf{B}'.\end{aligned} \tag{1.5a}$$

Hence $\mathbf{x} \in \mathscr{N}(\mathbf{B}') = \mathscr{C}(\mathbf{B})^{\perp} \implies \mathbf{x} \in \mathscr{N}(\mathbf{A}') = \mathscr{C}(\mathbf{A})^{\perp}$, and thereby we have shown that

$$\mathscr{C}(\mathbf{A}) \subset \mathscr{C}(\mathbf{B}) \implies \mathscr{C}(\mathbf{B})^{\perp} \subset \mathscr{C}(\mathbf{A})^{\perp}. \tag{1.5b}$$

The reverse relation follows by changing the roles of $\mathbf{A}$ and $\mathbf{A}^{\perp}$, and $\mathbf{B}$ and $\mathbf{B}^{\perp}$.

(d) follows from

$$\begin{aligned}\mathscr{N}(\mathbf{A}') = \mathscr{N}(\mathbf{A}\mathbf{A}') &\iff [\mathscr{N}(\mathbf{A}')]^{\perp} = [\mathscr{N}(\mathbf{A}\mathbf{A}')]^{\perp} \\ &\iff \mathscr{C}(\mathbf{A}) = \mathscr{C}(\mathbf{A}\mathbf{A}')\,.\end{aligned} \tag{1.6}$$

(e) It is obvious that $\operatorname{rank}(\mathbf{A}) = \operatorname{rank}(\mathbf{A}\mathbf{A}')$ and $\operatorname{rank}(\mathbf{A}') = \operatorname{rank}(\mathbf{A}'\mathbf{A})$ but we still have to show that $\mathbf{A}$ and $\mathbf{A}'$ have the same maximal number of linearly independent columns. We omit the proof which can be done easily using the full-rank decomposition of the matrix $\mathbf{A}$; see Theorem 17 (p. 349).

(f) Because $\mathbb{R}^n = \mathcal{U} \boxplus \mathcal{U}^{\perp}$ for any subspace $\mathcal{U} \subset \mathbb{R}^n$, we get (f).

(g) In view of (f), we have

$$n = \operatorname{rank}(\mathbf{A}) + \operatorname{rank}(\mathbf{A}^{\perp})\,, \tag{1.7a}$$

i.e.,

$$\operatorname{rank}(\mathbf{A}^{\perp}) = (\text{number of rows of } \mathbf{A}) - \operatorname{rank}(\mathbf{A})\,, \tag{1.7b}$$

which implies (g).

(h) comes from

$$\begin{aligned}\mathscr{C}(\mathbf{A} : \mathbf{B}) &= \{\, \mathbf{y} : \mathbf{y} = (\mathbf{A} : \mathbf{B})\binom{\mathbf{x}}{\mathbf{z}} \text{ for some } \mathbf{x},\, \mathbf{z} \,\} \\ &= \{\, \mathbf{y} : \mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} \text{ for some } \mathbf{x},\, \mathbf{z} \,\} \\ &= \{\, \mathbf{y} : \mathbf{y} = \mathbf{y}_1 + \mathbf{y}_2 \text{ for some } \mathbf{y}_1 \in \mathscr{C}(\mathbf{A}),\, \mathbf{y}_2 \in \mathscr{C}(\mathbf{B}) \,\} \\ &= \mathscr{C}(\mathbf{A}) + \mathscr{C}(\mathbf{B})\,. \end{aligned} \tag{1.8}$$

(i) From (h) it follows that

$$\begin{aligned}\operatorname{rank}(\mathbf{A} : \mathbf{B}) = \dim \mathscr{C}(\mathbf{A} : \mathbf{B}) &= \dim\,[\mathscr{C}(\mathbf{A}) + \mathscr{C}(\mathbf{B})] \\ &= \dim \mathscr{C}(\mathbf{A}) + \dim \mathscr{C}(\mathbf{B}) - \dim \mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B}) \\ &= \operatorname{rank}(\mathbf{A}) + \operatorname{rank}(\mathbf{B}) - \dim \mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B})\,. \end{aligned} \tag{1.9}$$

(j) This comes by noting that $\operatorname{rank}\binom{\mathbf{A}}{\mathbf{B}} = \operatorname{rank}(\mathbf{A}' : \mathbf{B}')$.

(k) Writing

$$\mathbf{A} + \mathbf{B} = (\mathbf{A} : \mathbf{B})\binom{\mathbf{I}}{\mathbf{I}}\,, \tag{1.10a}$$

we see that

$$\mathscr{C}(\mathbf{A} + \mathbf{B}) = \mathscr{C}\big[(\mathbf{A} : \mathbf{B})\binom{\mathbf{I}}{\mathbf{I}}\big] \subset \mathscr{C}(\mathbf{A} : \mathbf{B})\,, \tag{1.10b}$$

where the last inclusion comes from the fact

$$\mathscr{C}(\mathbf{U}\mathbf{F}) \subset \mathscr{C}(\mathbf{U}) \quad \text{for any conformable matrices } \mathbf{U},\, \mathbf{F}\,. \tag{1.10c}$$

(l) is a consequence of (k).

(m) comes from the following:

$$\begin{aligned}\mathbf{y} \in \mathscr{C}(\mathbf{A} : \mathbf{B})^{\perp} &\iff \mathbf{y}'(\mathbf{A} : \mathbf{B}) = \mathbf{0}' \iff \mathbf{A}'\mathbf{y} = \mathbf{0},\ \mathbf{B}'\mathbf{y} = \mathbf{0} \\ &\iff \mathbf{y} \in \mathscr{C}(\mathbf{A})^{\perp} \cap \mathscr{C}(\mathbf{B})^{\perp}. \end{aligned} \tag{1.11}$$

□

We may almost confess that originally we planned to omit to represent the results of Theorem 1, because the reader is almost expected to be aware of them. However, these simple properties are so essential that we ..., so, needless to say. They may look sterile but please do not underestimate them.

1.1 Equality $\mathscr{C}(\mathbf{X}'\mathbf{V}\mathbf{X}) = \mathscr{C}(\mathbf{X}'\mathbf{V})$ when V Is Nonnegative Definite

Suppose that $\mathbf{X}$ is an $n \times p$ matrix and $\mathbf{V}$ is an $n \times n$ symmetric nonnegative definite matrix. Then the following simple result is frequently used:

$$\mathscr{C}(\mathbf{X}'\mathbf{V}\mathbf{X}) = \mathscr{C}(\mathbf{X}'\mathbf{V})\,. \tag{1.12}$$

The proof of (1.12) comes by noting that $\mathbf{V}$ must be of the form $\mathbf{V} = \mathbf{L}'\mathbf{L}$ for some matrix $\mathbf{L}$, and writing

$$\begin{aligned}\mathscr{C}(\mathbf{X}'\mathbf{V}\mathbf{X}) &= \mathscr{C}(\mathbf{X}'\mathbf{L}'\mathbf{L}\mathbf{X}) \subset \mathscr{C}(\mathbf{X}'\mathbf{L}'\mathbf{L}) \\ &= \mathscr{C}(\mathbf{X}'\mathbf{V}) \subset \mathscr{C}(\mathbf{X}'\mathbf{L}') = \mathscr{C}[\mathbf{X}'\mathbf{L}'(\mathbf{X}'\mathbf{L}')'] \\ &= \mathscr{C}(\mathbf{X}'\mathbf{L}'\mathbf{L}\mathbf{X}) = \mathscr{C}(\mathbf{X}'\mathbf{V}\mathbf{X})\,. \end{aligned} \tag{1.13}$$

Now clearly (1.13) implies (1.12). Trivially we have also

$$\operatorname{rank}(\mathbf{X}'\mathbf{V}\mathbf{X}) = \operatorname{rank}(\mathbf{X}'\mathbf{V}) = \operatorname{rank}(\mathbf{V}\mathbf{X})\,, \tag{1.14}$$

and

$$\mathbf{X}'\mathbf{V}\mathbf{X} = \mathbf{0} \iff \mathbf{X}'\mathbf{V} = \mathbf{0}\,. \tag{1.15}$$

1.2 Estimability in a Simple ANOVA

Consider the situation of Example 0.2 (p. 30) where we had the following models:

$$\mathscr{M}_*\colon\quad \mathbf{y} = \begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{31} \\ y_{32} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} + \boldsymbol{\varepsilon} = \mathbf{X}_*\boldsymbol{\beta}_* + \boldsymbol{\varepsilon}\,, \tag{1.16a}$$

$$\mathscr{M}_\#\colon\quad \mathbf{y} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{pmatrix} + \boldsymbol{\varepsilon} = \mathbf{X}_\#\boldsymbol{\beta}_\# + \boldsymbol{\varepsilon}\,. \tag{1.16b}$$

In a more general situation, when we have g groups (or treatments) and n_i observations in the ith group, we can write these models as

$$\mathscr{M}_*\colon\quad \mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_g \end{pmatrix} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_2} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{1}_{n_g} \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_g \end{pmatrix} + \boldsymbol{\varepsilon}$$

$$= \mathbf{X}_* \boldsymbol{\beta}_* + \boldsymbol{\varepsilon}\,, \tag{1.17a}$$

$$\mathscr{M}_{\#}\colon \quad \mathbf{y} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{1}_{n_2} & \mathbf{0} & \mathbf{1}_{n_2} & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{1}_{n_g} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{1}_{n_g} \end{pmatrix} \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \vdots \\ \tau_g \end{pmatrix} + \boldsymbol{\varepsilon} = \mathbf{X}_{\#}\boldsymbol{\beta}_{\#} + \boldsymbol{\varepsilon}\,, \tag{1.17b}$$

where $n = n_1 + \dots + n_g$. We see immediately that $\boldsymbol{\beta}_*$ is estimable under $\mathscr{M}_*$ because $\mathbf{X}_*$ has full column rank. On the other hand, $\boldsymbol{\beta}_{\#}$ is not estimable under $\mathscr{M}_{\#}$. Which parametric functions of $\boldsymbol{\beta}_{\#}$ are estimable? We answer to this question by using some properties of Theorem 1.

To simplify the notation, let us drop the subscript # and denote the model (1.17b) as

$$\mathscr{M}\colon \quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = (\mathbf{1}_n : \mathbf{X}_0) \begin{pmatrix} \mu \\ \boldsymbol{\tau} \end{pmatrix} + \boldsymbol{\varepsilon}\,. \tag{1.18}$$

Now the parametric function $\mathbf{k}'\boldsymbol{\beta}$ is estimable under $\mathscr{M}$ if and only if

$$\mathbf{k} \in \mathscr{C}(\mathbf{X}') = \mathscr{C}\begin{pmatrix} \mathbf{1}'_n \\ \mathbf{X}'_0 \end{pmatrix} = \mathscr{C}\begin{pmatrix} \mathbf{1}'_{n_1} & \mathbf{1}'_{n_2} & \dots & \mathbf{1}'_{n_g} \\ \mathbf{1}'_{n_1} & \mathbf{0}' & \dots & \mathbf{0}' \\ \vdots & \vdots & & \vdots \\ \mathbf{0}' & \mathbf{0}' & \dots & \mathbf{1}'_{n_g} \end{pmatrix} = \mathscr{C}\begin{pmatrix} \mathbf{1}'_g \\ \mathbf{I}_g \end{pmatrix}. \tag{1.19}$$

Moreover, because

$$\operatorname{rank}\begin{pmatrix} \mathbf{1}'_g \\ \mathbf{I}_g \end{pmatrix} = \operatorname{rank}(\mathbf{1}_g : \mathbf{I}_g) = g\,, \tag{1.20}$$

we have, in view of part (g) of Theorem 1,

$$\operatorname{rank}\begin{pmatrix} \mathbf{1}'_g \\ \mathbf{I}_g \end{pmatrix}^{\perp} = (g+1) - g = 1\,. \tag{1.21}$$

We show next that one choice for $\begin{pmatrix} \mathbf{1}'_g \\ \mathbf{I}_g \end{pmatrix}^{\perp}$ is $\begin{pmatrix} -1 \\ \mathbf{1}_g \end{pmatrix}$, i.e.,

$$\begin{pmatrix} -1 \\ \mathbf{1}_g \end{pmatrix} := \mathbf{u} \in \left\{ \begin{pmatrix} \mathbf{1}'_g \\ \mathbf{I}_g \end{pmatrix}^{\perp} \right\}. \tag{1.22}$$

The result (1.22) follows from the facts

$$\mathbf{u}' \begin{pmatrix} \mathbf{1}'_g \\ \mathbf{I}_g \end{pmatrix} = \mathbf{0}'\,, \quad \operatorname{rank}(\mathbf{u}) = \operatorname{rank}\begin{pmatrix} \mathbf{1}'_g \\ \mathbf{I}_g \end{pmatrix}^{\perp} = 1\,. \tag{1.23}$$

According to (1.19), the parametric function $\mathbf{k}'\boldsymbol{\beta}$ is estimable if and only if

$$\mathbf{k} \in \left(\mathscr{C}\begin{pmatrix}\mathbf{1}_g' \\ \mathbf{I}_g\end{pmatrix}^{\perp}\right)^{\perp} = \mathscr{C}(\mathbf{u})^{\perp}, \tag{1.24}$$

i.e.,

$$\mathbf{k}'\mathbf{u} = 0, \quad \text{where} \quad \mathbf{u} = \begin{pmatrix}-1 \\ \mathbf{1}_g\end{pmatrix}. \tag{1.25}$$

We can also study the estimability of a parametric function of $\tau_1, \dots, \tau_g$ (dropping off the parameter μ); denote this function as $\boldsymbol{\ell}'\boldsymbol{\tau}$. Then

$$(0, \boldsymbol{\ell}')\begin{pmatrix}\mu \\ \boldsymbol{\tau}\end{pmatrix} = \boldsymbol{\ell}'\boldsymbol{\tau}\,, \tag{1.26}$$

and on account of (1.25), the estimability condition for $\boldsymbol{\ell}'\boldsymbol{\tau}$ becomes $\boldsymbol{\ell}'\mathbf{1}_g = 0$. We may summarize our findings as follows.

Proposition 1.1. *Consider the one-way* ANOVA *model* (1.18). *Then* $\mathbf{k}'\boldsymbol{\beta}$ *is estimable if and only if*

$$\mathbf{k}'\begin{pmatrix}-1 \\ \mathbf{1}_g\end{pmatrix} = 0\,, \tag{1.27}$$

and $\boldsymbol{\ell}'\boldsymbol{\tau}$ *is estimable if and only if*

$$\boldsymbol{\ell}'\mathbf{1}_g = 0\,. \tag{1.28}$$

The parametric function $\boldsymbol{\ell}'\boldsymbol{\tau}$ is called a *contrast* if $\ell_1 + \cdots + \ell_g = 0$, i.e., (1.28) holds. Contrasts of the form $\tau_i - \tau_j$, $i \neq j$, are called elementary contrasts.

For the estimability in the block model, see Exercise 16.2 (p. 348).

In passing we may mention that (1.22) is a special case of the following general result:

Proposition 1.2. *One choice for a matrix* $\begin{pmatrix}\mathbf{B}_{n\times q} \\ \mathbf{I}_q\end{pmatrix}^{\perp}$ *is* $\begin{pmatrix}\mathbf{I}_n \\ -\mathbf{B}'\end{pmatrix}$, *i.e.,*

$$\begin{pmatrix}\mathbf{I}_n \\ -\mathbf{B}'\end{pmatrix} \in \left\{\begin{pmatrix}\mathbf{B}_{n\times q} \\ \mathbf{I}_q\end{pmatrix}^{\perp}\right\}. \tag{1.29}$$

1.3 Column Space of the Covariance Matrix

Consider the partitioned $(p+q)$-dimensional random vector $\mathbf{z} = \binom{\mathbf{x}}{\mathbf{y}}$ whose covariance matrix is

$$\operatorname{cov}(\mathbf{z}) = \operatorname{cov}\begin{pmatrix}\mathbf{x} \\ \mathbf{y}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\Sigma}_{\mathbf{xx}} & \boldsymbol{\Sigma}_{\mathbf{xy}} \\ \boldsymbol{\Sigma}_{\mathbf{yx}} & \boldsymbol{\Sigma}_{\mathbf{yy}}\end{pmatrix} = \boldsymbol{\Sigma}\,. \tag{1.30}$$

Because every covariance matrix is nonnegative definite, there exists a matrix $\mathbf{L} = (\mathbf{L}_1 : \mathbf{L}_2)$ such that $\mathbf{\Sigma} = \mathbf{L}'\mathbf{L}$:

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{L}_1' \\ \mathbf{L}_2' \end{pmatrix} (\mathbf{L}_1 : \mathbf{L}_2) = \begin{pmatrix} \mathbf{L}_1'\mathbf{L}_1 & \mathbf{L}_1'\mathbf{L}_2 \\ \mathbf{L}_2'\mathbf{L}_1 & \mathbf{L}_2'\mathbf{L}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{\Sigma_{xx}} & \mathbf{\Sigma_{xy}} \\ \mathbf{\Sigma_{yx}} & \mathbf{\Sigma_{yy}} \end{pmatrix}. \tag{1.31}$$

In view of

$$\mathscr{C}(\mathbf{L}_1'\mathbf{L}_2) \subset \mathscr{C}(\mathbf{L}_1') = \mathscr{C}(\mathbf{L}_1'\mathbf{L}_1), \tag{1.32}$$

we can immediately conclude the following simple but important facts:

$$\mathscr{C}(\mathbf{\Sigma_{xy}}) \subset \mathscr{C}(\mathbf{\Sigma_{xx}}), \quad \mathscr{C}(\mathbf{\Sigma_{yx}}) \subset \mathscr{C}(\mathbf{\Sigma_{yy}}). \tag{1.33}$$

From (1.33) we can also conclude that

$$\mathbf{\Sigma_{yy}} = \mathbf{0} \implies \mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma_{xx}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \tag{1.34}$$

and in particular

$$\operatorname{cov}\begin{pmatrix} \mathbf{x} \\ y \end{pmatrix} = \mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma_{xx}} & \boldsymbol{\sigma}_{\mathbf{x}y} \\ \boldsymbol{\sigma}_{\mathbf{x}y}' & \sigma_y^2 = 0 \end{pmatrix} \implies \mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma_{xx}} & \mathbf{0}_p \\ \mathbf{0}_p' & 0 \end{pmatrix}. \tag{1.35}$$

Moreover, let us denote $\mathrm{E}(\mathbf{x}) = \boldsymbol{\mu}_{\mathbf{x}} \in \mathbb{R}^p$ and consider the random vector

$$\mathbf{u} = \mathbf{Qx}, \quad \text{where } \mathbf{Q} = \mathbf{I}_p - \mathbf{P}_{(\mathbf{\Sigma_{xx}} : \boldsymbol{\mu}_{\mathbf{x}})}. \tag{1.36}$$

Now

$$\mathrm{E}(\mathbf{Qx}) = (\mathbf{I}_p - \mathbf{P}_{(\mathbf{\Sigma_{xx}} : \boldsymbol{\mu}_{\mathbf{x}})})\boldsymbol{\mu}_{\mathbf{x}} = \mathbf{0}, \tag{1.37a}$$

$$\operatorname{cov}(\mathbf{Qx}) = \mathbf{Q\Sigma_{xx}Q} = \mathbf{0}, \tag{1.37b}$$

and hence $(\mathbf{I}_p - \mathbf{P}_{(\mathbf{\Sigma_{xx}} : \boldsymbol{\mu}_{\mathbf{x}})})\mathbf{x} = \mathbf{0}$ with probability 1, and so

$$\mathbf{x} \in \mathscr{C}(\mathbf{\Sigma_{xx}} : \boldsymbol{\mu}_{\mathbf{x}}) \quad \text{with probability } 1. \tag{1.38}$$

For the corresponding condition in a linear model, see Section 5.2 (p. 125).

Notice that for $\mathbf{Q}_{\mathbf{\Sigma_{xx}}} = \mathbf{I}_p - \mathbf{P}_{\mathbf{\Sigma_{xx}}}$,

$$\mathrm{E}[\mathbf{Q}_{\mathbf{\Sigma_{xx}}}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})] = (\mathbf{I}_p - \mathbf{P}_{\mathbf{\Sigma_{xx}}})(\boldsymbol{\mu}_{\mathbf{x}} - \boldsymbol{\mu}_{\mathbf{x}}) = \mathbf{0}, \tag{1.39a}$$

$$\operatorname{cov}[\mathbf{Q}_{\mathbf{\Sigma_{xx}}}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})] = \mathbf{Q}_{\mathbf{\Sigma_{xx}}}\mathbf{\Sigma_{xx}}\mathbf{Q}_{\mathbf{\Sigma_{xx}}} = \mathbf{0}, \tag{1.39b}$$

and hence

$$\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}} \in \mathscr{C}(\mathbf{\Sigma_{xx}}) \quad \text{with probability } 1. \tag{1.40}$$

For a clarity, let us write the following proposition:

Proposition 1.3. *Using the above notation,*

(a) $\mathbf{x} \in \mathscr{C}(\mathbf{\Sigma_{xx}} : \boldsymbol{\mu}_{\mathbf{x}})$ *with probability* 1,

(b) $\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}} \in \mathscr{C}(\boldsymbol{\Sigma}_{\mathbf{xx}})$ *with probability* 1*.*

There is one particular consequence from (1.33) that deserves attention:

$$\text{there exists a matrix } \mathbf{B} \text{ such that } \boldsymbol{\Sigma}_{\mathbf{xx}}\mathbf{B} = \boldsymbol{\Sigma}_{\mathbf{xy}}\,, \tag{1.41}$$

i.e.,

$$\mathbf{L}_1'\mathbf{L}_1\mathbf{B} = \mathbf{L}_1'\mathbf{L}_2 \quad \text{for some } \mathbf{B} \in \mathbb{R}^{p\times q}\,. \tag{1.42}$$

This is indeed evident because $\mathscr{C}(\mathbf{L}_1'\mathbf{L}_2) \subset \mathscr{C}(\mathbf{L}_1') = \mathscr{C}(\mathbf{L}_1'\mathbf{L}_1)$. One solution to $\boldsymbol{\Sigma}_{\mathbf{xx}}\mathbf{B} = \boldsymbol{\Sigma}_{\mathbf{xy}}$ is

$$\mathbf{B} = \boldsymbol{\Sigma}_{\mathbf{xx}}^{-}\boldsymbol{\Sigma}_{\mathbf{xy}}\,, \tag{1.43}$$

where $\boldsymbol{\Sigma}_{\mathbf{xx}}^{-}$ is an arbitrary generalized inverse of $\boldsymbol{\Sigma}_{\mathbf{xx}}$. By varying $\boldsymbol{\Sigma}_{\mathbf{xx}}^{-}$ through all generalized inverses of $\boldsymbol{\Sigma}_{\mathbf{xx}}$ we can generate all solutions to $\boldsymbol{\Sigma}_{\mathbf{xx}}\mathbf{B} = \boldsymbol{\Sigma}_{\mathbf{xy}}$. According to Theorem 11 (p. 267), an alternative way to express all solutions is

$$\mathbf{B} = \boldsymbol{\Sigma}_{\mathbf{xx}}^{-}\boldsymbol{\Sigma}_{\mathbf{xy}} + (\mathbf{I}_p - \boldsymbol{\Sigma}_{\mathbf{xx}}^{-}\boldsymbol{\Sigma}_{\mathbf{xx}})\mathbf{Z}\,, \tag{1.44}$$

where $\boldsymbol{\Sigma}_{\mathbf{xx}}^{-}$ is an arbitrary (but fixed) generalized inverse of $\boldsymbol{\Sigma}_{\mathbf{xx}}$ and $\mathbf{Z} \in \mathbb{R}^{p\times q}$ is free to vary.

The fact (1.41) is a crucial one in several connections. For example, there are occasions when it is extremely useful to find a matrix $\mathbf{A} \in \mathbb{R}^{q\times p}$ so that the random vectors $\mathbf{x}$ and $\mathbf{y} - \mathbf{Ax}$ are uncorrelated, i.e.,

$$\operatorname{cov}(\mathbf{x}, \mathbf{y} - \mathbf{Ax}) = \boldsymbol{\Sigma}_{\mathbf{xy}} - \boldsymbol{\Sigma}_{\mathbf{xx}}\mathbf{A}' = \mathbf{0}\,. \tag{1.45}$$

Now (1.45) holds if and only if $\mathbf{A}'$ is a solution to

$$\boldsymbol{\Sigma}_{\mathbf{xx}}\mathbf{A}' = \boldsymbol{\Sigma}_{\mathbf{xy}}\,, \tag{1.46}$$

which indeed, in view if (1.41), is consistent, the general solution being

$$\mathbf{A}' = \boldsymbol{\Sigma}_{\mathbf{xx}}^{-}\boldsymbol{\Sigma}_{\mathbf{xy}} + (\mathbf{I}_p - \boldsymbol{\Sigma}_{\mathbf{xx}}^{-}\boldsymbol{\Sigma}_{\mathbf{xx}})\mathbf{Z}\,, \tag{1.47}$$

where $\mathbf{Z} \in \mathbb{R}^{p\times q}$ is free to vary. Transposing (1.47) yields

$$\mathbf{A} = \boldsymbol{\Sigma}_{\mathbf{yx}}(\boldsymbol{\Sigma}_{\mathbf{xx}}^{-})' + \mathbf{Z}'[\mathbf{I}_p - \boldsymbol{\Sigma}_{\mathbf{xx}}(\boldsymbol{\Sigma}_{\mathbf{xx}}^{-})']\,. \tag{1.48}$$

Choosing $\boldsymbol{\Sigma}_{\mathbf{xx}}^{-}$ as $(\boldsymbol{\Sigma}_{\mathbf{xx}}^{-})'$ (this is indeed possible, please confirm!), the general expression for $\mathbf{A}$ can be written as

$$\mathbf{A} = \boldsymbol{\Sigma}_{\mathbf{xy}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-} + \mathbf{Z}'(\mathbf{I}_p - \boldsymbol{\Sigma}_{\mathbf{xx}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-})\,. \tag{1.49}$$

For the above expression, see also (11.16) (p. 269).

We might extend the above simple but helpful results as a proposition.

Proposition 1.4. *Using the above notation, the following statements are equivalent:*

(a) $\operatorname{cov}(\mathbf{x}, \mathbf{y} - \mathbf{Ax}) = \mathbf{0}\,.$

(b) *The matrix* $\mathbf{A}$ *is a solution to* $\mathbf{A\Sigma_{xx}} = \mathbf{\Sigma_{yx}}$.

(c) *The matrix* $\mathbf{A}$ *is of the form*

$$\mathbf{A} = \mathbf{\Sigma_{yx}\Sigma_{xx}^-} + \mathbf{Z}'(\mathbf{I}_p - \mathbf{\Sigma_{xx}\Sigma_{xx}^-})\,, \tag{1.50}$$

where $\mathbf{Z} \in \mathbb{R}^{p\times q}$ *is free to vary.*

(d) $\mathbf{A}(\mathbf{x} - \boldsymbol{\mu}_\mathbf{x}) = \mathbf{\Sigma_{yx}\Sigma_{xx}^-}(\mathbf{x} - \boldsymbol{\mu}_\mathbf{x})$ *with probability* 1.

Proof. The equivalence of (a), (b) and (c) has already been shown. In part (d), it is essential to observe that

$$\mathscr{C}(\mathbf{\Sigma_{yx}'}) = \mathscr{C}(\mathbf{\Sigma_{xy}}) \subset \mathscr{C}(\mathbf{\Sigma_{xx}}) \tag{1.51}$$

and

$$\mathbf{x} - \boldsymbol{\mu}_\mathbf{x} \in \mathscr{C}(\mathbf{\Sigma_{xx}}) \quad \text{with probability 1.} \tag{1.52}$$

Hence, in view of Theorem 12 (p. 283), the random vector $\mathbf{\Sigma_{yx}\Sigma_{xx}^-}(\mathbf{x} - \boldsymbol{\mu}_\mathbf{x})$ is invariant with respect to the choice of $\mathbf{\Sigma_{xx}^-}$ with probability 1. Choosing $\mathbf{\Sigma_{xx}^-}$ as $\mathbf{\Sigma_{xx}^+}$ we see at once that (d) implies that

$$\mathbf{Ax} = \mathbf{\Sigma_{yx}\Sigma_{xx}^+x} + \mathbf{b} \quad \text{for some } \mathbf{b} \in \mathbb{R}^q, \tag{1.53}$$

and so

$$\begin{aligned}
\operatorname{cov}(\mathbf{x}, \mathbf{y} - \mathbf{Ax}) &= \operatorname{cov}(\mathbf{x}, \mathbf{y} - \mathbf{\Sigma_{yx}\Sigma_{xx}^+x}) \\
&= \operatorname{cov}(\mathbf{x}, \mathbf{y}) - \operatorname{cov}(\mathbf{x}, \mathbf{\Sigma_{yx}\Sigma_{xx}^+x}) \\
&= \mathbf{\Sigma_{xy}} - \mathbf{\Sigma_{xx}\Sigma_{xx}^+\Sigma_{xy}} \\
&= \mathbf{\Sigma_{xy}} - \mathbf{\Sigma_{xy}} = \mathbf{0}\,,
\end{aligned} \tag{1.54}$$

where we have used the fact

$$\mathbf{BB^-C} = \mathbf{C} \iff \mathscr{C}(\mathbf{C}) \subset \mathscr{C}(\mathbf{B})\,. \tag{1.55}$$

Thus we have shown that (d) implies (a) and thereby (b) and (c). To go the other way, postmultiplying (1.50) by $\mathbf{x} - \boldsymbol{\mu}_\mathbf{x}$ and using (1.51) and (1.52) gives

$$\begin{aligned}
\mathbf{A}(\mathbf{x} - \boldsymbol{\mu}_\mathbf{x}) &= \mathbf{\Sigma_{yx}\Sigma_{xx}^-}(\mathbf{x} - \boldsymbol{\mu}_\mathbf{x}) + \mathbf{Z}'(\mathbf{I}_p - \mathbf{\Sigma_{xx}\Sigma_{xx}^-})(\mathbf{x} - \boldsymbol{\mu}_\mathbf{x}) \\
&= \mathbf{\Sigma_{yx}\Sigma_{xx}^-}(\mathbf{x} - \boldsymbol{\mu}_\mathbf{x})\,.
\end{aligned} \tag{1.56}$$

□

For a reference to Proposition 1.4, see Sengupta & Jammalamadaka (2003, p. 56).

The squared population Mahalanobis distance between $\mathbf{x}$ and $\boldsymbol{\mu}_\mathbf{x}$ is defined for a positive definite $\mathbf{\Sigma_{xx}}$ as

$$\mathrm{MHLN}^2(\mathbf{x},\, \boldsymbol{\mu}_\mathbf{x},\, \mathbf{\Sigma_{xx}}) = (\mathbf{x} - \boldsymbol{\mu}_\mathbf{x})'\mathbf{\Sigma_{xx}^{-1}}(\mathbf{x} - \boldsymbol{\mu}_\mathbf{x})\,. \tag{1.57}$$

For a singular $\mathbf{\Sigma}$ we can define

$$\mathrm{MHLN}^2(\mathbf{x},\, \boldsymbol{\mu}_{\mathbf{x}},\, \mathbf{\Sigma}_{\mathbf{xx}}) = (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})'\mathbf{\Sigma}_{\mathbf{xx}}^{-}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})\,, \tag{1.58}$$

and it is noteworthy that (1.58) it is invariant (with probability 1) with respect to the choice of $\mathbf{\Sigma}_{\mathbf{xx}}^{-}$. This follows immediately from the fact that $\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}} \in \mathscr{C}(\mathbf{\Sigma}_{\mathbf{xx}})$ with probability 1.

For the sample Mahalanobis distance we have the corresponding property:

$$\mathrm{MHLN}^2(\mathbf{x}_{(i)},\, \bar{\mathbf{z}},\, \mathbf{S}_{\mathbf{xx}}) = (\mathbf{x}_{(i)} - \bar{\mathbf{x}})'\mathbf{S}_{\mathbf{xx}}^{-}(\mathbf{x}_{(i)} - \bar{\mathbf{x}}) = \tilde{\mathbf{x}}_{(i)}'\mathbf{S}_{\mathbf{xx}}^{-}\tilde{\mathbf{x}}_{(i)}\,, \tag{1.59}$$

which remains invariant for any choice of $\mathbf{S}_{\mathbf{xx}}^{-}$. To confirm this, we recall that in (1.59)

$$\tilde{\mathbf{x}}_{(i)} = \mathbf{x}_{(i)} - \bar{\mathbf{x}}: \quad i\text{th centered observation,} \tag{1.60a}$$

$$\mathbf{X} = (\mathbf{x}_{(1)} : \ldots : \mathbf{x}_{(n)})': \quad \text{data matrix,} \tag{1.60b}$$

$$\widetilde{\mathbf{X}} = (\tilde{\mathbf{x}}_{(1)} : \ldots : \tilde{\mathbf{x}}_{(n)})': \quad \text{centered data matrix,} \tag{1.60c}$$

$$\mathbf{S}_{\mathbf{xx}} = \tfrac{1}{n-1}\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}} = \tfrac{1}{n-1}\sum_{i=1}^{n}\tilde{\mathbf{x}}_{(i)}\tilde{\mathbf{x}}_{(i)}': \quad \text{sample covariance matrix,} \tag{1.60d}$$

and $\bar{\mathbf{x}} = (\bar{x}_1, \ldots, \bar{x}_p)'$. Now of course we have $\tilde{\mathbf{x}}_{(i)} \in \mathscr{C}(\widetilde{\mathbf{X}}') = \mathscr{C}(\mathbf{S}_{\mathbf{xx}})$, which confirms the invariance of the sample Mahalanobis distance (1.59).

1.4 Exercises

1.1. Consider the matrix $\mathbf{A} = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$. Let the elements of $\mathbf{A}$ be independent observations from the random variable x. What is your guess for the rank of $\mathbf{A}$ if x follows

(a) uniform discrete distribution, values being the digits $0, 1, \ldots, 9$,
(b) uniform continuous distribution, values being from the interval $[0, 1]$,
(c) Bernoulli distribution: $\mathrm{P}(x = 1) = p$, $\mathrm{P}(x = 0) = 1 - p = q$.

1.2. Consider the nonnull vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$. Prove:

$$\mathbf{P}_{\mathbf{u}}\mathbf{P}_{\mathbf{v}} = \mathbf{P}_{\mathbf{v}}\mathbf{P}_{\mathbf{u}} \iff \mathbf{u} = \lambda\mathbf{v} \text{ for some } \lambda \in \mathbb{R} \text{ or } \mathbf{u}'\mathbf{v} = 0\,.$$

1.3. Prove Proposition 1.2 (p. 62): $\left(\begin{smallmatrix} \mathbf{I}_n \\ -\mathbf{B}' \end{smallmatrix}\right) \in \left\{\left(\begin{smallmatrix} \mathbf{B}_{n\times q} \\ \mathbf{I}_q \end{smallmatrix}\right)^{\perp}\right\}$.

1.4. Show that the matrices $\mathbf{A}$ and $\mathbf{A}'$ have the same maximal number of linearly independent columns.
Hint: Utilize the full rank decomposition of $\mathbf{A}$, Theorem 17 (p. 349).

1.5. Consider an $n \times p$ matrix $\mathbf{A}$ which has full column rank. Confirm that the columns of $\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1/2}$ form an orthonormal basis for $\mathscr{C}(\mathbf{A})$.
Hint: Recall that $(\mathbf{A}'\mathbf{A})^{-1/2} = \mathbf{T}\boldsymbol{\Lambda}^{-1/2}\mathbf{T}'$, where $\mathbf{A}'\mathbf{A} = \mathbf{T}\boldsymbol{\Lambda}\mathbf{T}'$ is the eigenvalue decomposition of $\mathbf{A}'\mathbf{A}$.

1.6. Let $\operatorname{rank}(\mathbf{A}_{n\times p}) = a$ and $\operatorname{rank}(\mathbf{B}_{n\times q}) = b$. Show that the statements

$$\text{(a) } \mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B}) = \{\mathbf{0}\} \quad \text{and} \quad \text{(b) } \mathscr{C}(\mathbf{A})^{\perp} \cap \mathscr{C}(\mathbf{B})^{\perp} = \{\mathbf{0}\}$$

are equivalent if and only if $\operatorname{rank}(\mathbf{B}) = \operatorname{rank}(\mathbf{A}^{\perp}) = n - \operatorname{rank}(\mathbf{A})$.

1.7. Show that for conformable matrices $\mathbf{A}, \mathbf{B}$ and $\mathbf{Z}$ the following holds:

$$\mathscr{C}(\mathbf{A}) \subset \mathscr{C}(\mathbf{B} : \mathbf{Z}) \;\&\; \mathscr{C}(\mathbf{B}) \subset \mathscr{C}(\mathbf{A} : \mathbf{Z}) \iff \mathscr{C}(\mathbf{A} : \mathbf{Z}) = \mathscr{C}(\mathbf{B} : \mathbf{Z})\,.$$

1.8. Consider the two Latin square matrices

$$\mathbf{L}_1 = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \\ 3 & 4 & 1 & 2 \\ 4 & 3 & 2 & 1 \end{pmatrix}, \qquad \mathbf{L}_2 = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 1 & 4 & 2 \\ 2 & 4 & 1 & 3 \\ 4 & 3 & 2 & 1 \end{pmatrix}.$$

Show that both $\mathbf{L}_1$ and $\mathbf{L}_2$ are singular, and find their ranks. Construct some choices for $\mathbf{L}_i^{\perp}, i = 1, 2$.

In general, a Latin square of size $k \times k$ is a square array of k symbols in k rows and k columns such that every symbol occurs once in each row and once in each column. A 4×4 Latin square is in *standard-form* whenever $1, 2, 3, 4$ is in the top row. Show that there are 24 standard-form Latin squares of order 4.

A Latin square is said tobe *criss-cross* whenever all the elements in the main forwards diagonal are equal and all the elements in the main backwards diagonal are equal. Show that the matrices $\mathbf{L}_1$ and $\mathbf{L}_2$ are the only two 4×4 criss-cross Latin squares in standard form.

The earliest application of a Latin square in Europe seems to be the one represented by this matrix $\mathbf{L}_1$ for the “First elemental figure” in *Ars Demonstrativa* by the Majorcan writer and philosopher Ramon Llull (1232–1316) first published in 1283, see Plate XIII in and Styan, Boyer & Chu (2009). Recently surfaced manuscripts show that Ramon Llull anticipated by several centuries prominent work on electoral systems, see, e.g., Hägele & Pukelsheim (2001). The stamps shown here are from Spain 1963 (*Scott* 1197), and Andorra/Spanish Administration 1987 (*Scott* 182). Depicted on the Andorran stamp is the detail “De Nativitat” from the Catalan manuscript *Llibre de Doctrina Pueril;* for a French translation see “De la nativité” in *Raymond Llulle: Doctrine d’Enfant* by Llinarès (1969), pp. 102–103.

The Latin square matrix $\mathbf{L}_2$ has been used in printing postage stamps as noted by Loly & Styan (2010a). Depicted on the sheetlet of stamps

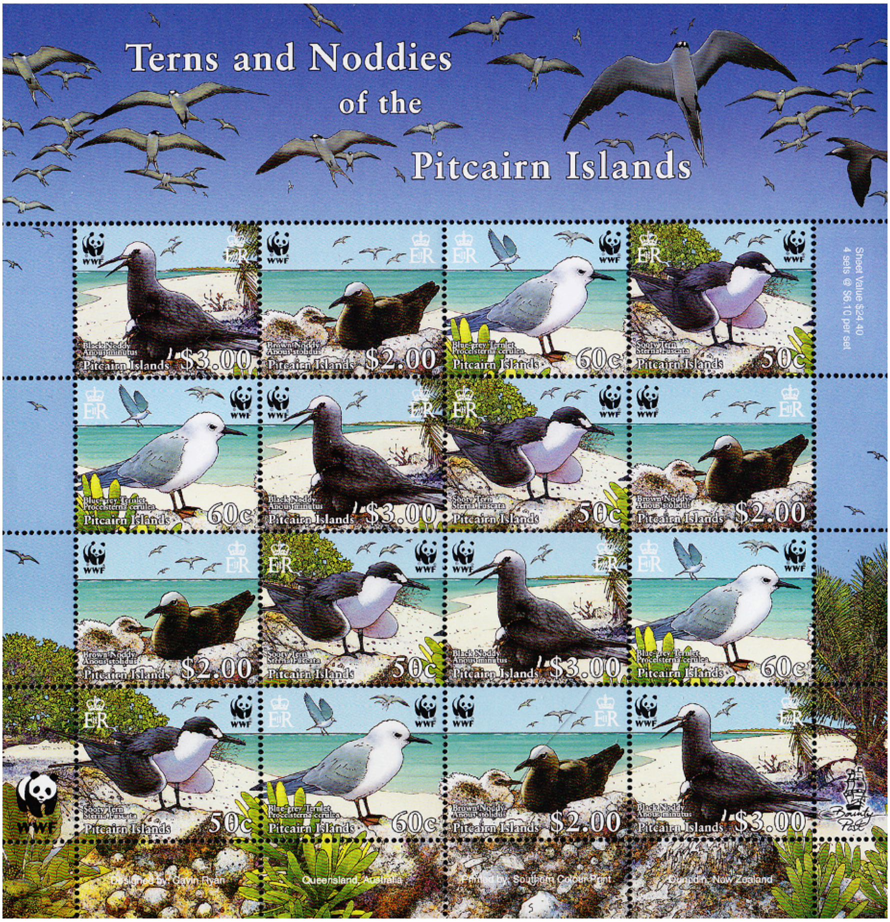
Philatelic Item 1.1 Stamps for Ramon Llull (1232–1316).

(Philatelic Item 1.2) from the Pitcairn Islands 2007, *Scott* 647, are the black noddy, the closely related brown noddy, the blue-grey ternlet, and the sooty tern.

As indicated by Chu, Puntanen & Styan (2009) and by Loly & Styan (2010b), 5 × 5 Latin squares have also been used in printing postage stamps. Depicted on the sheetlet of stamps (Philatelic Item 1.3) from Pakistan 2004, *Scott* 1045, are five kinds of tropical fish: neon tetra, striped gourami, black widow, yellow dwarf cichlid, and tiger barb. This Latin square has several interesting properties. In particular it is of the type known as a “knight’s move” or “Knut Vik” design, as described by Preece (1990):

> “Of Latin squares used for crop experiments with a single set of treatments, the earliest examples (published in 1924) are 5 × 5 squares of the systematic type known as Knut or “knight’s move” designs (Knut Vik being a [Norwegian] person, not a Scandinavian translation of “knight’s move”!), see Vik (1924); these are Latin squares where all cells containing any one of the treatments can be visited by a succession of knight’s moves (as in chess) and where no two diagonally adjacent cells have the same treatment.”

For more about Latin squares we recommend the books by Dénes & Keedwell (1974) and Laywine & Mullen (1998).



Philatelic Item 1.2 A 4 × 4 philatelic Latin square.

Philatelic Item 1.3 A 5 × 5 philatelic Latin square.

Chapter 2
Easy Projector Tricks

> *Poorly conceived abstraction:*
> *aerodynamics in a viscous fluid.*
>
> P. C. Mahalanobis

In this chapter we go through some basic properties of the orthogonal projectors. We are great fans of projectors and believe that their heavy use simplifies many considerations and helps to understand what is going on particularly in various minimization problems. We first consider the orthogonal projections under the standard inner product, and then under the more general cases. We believe that it is instructive to proceed in this order instead of first going to the general case.

Theorem 2 (Orthogonal projector). *Let* $\mathbf{A}$ *be an* $n \times m$ *matrix,* $\operatorname{rank}(\mathbf{A}) = r > 0$*, and let the inner product and the corresponding norm be defined as* $\langle \mathbf{t}, \mathbf{u} \rangle = \mathbf{t}'\mathbf{u}$*, and* $\|\mathbf{t}\| = \sqrt{\langle \mathbf{t}, \mathbf{t} \rangle}$*, respectively. Further, let*

- $\mathbf{A}^{\perp} \in \mathbb{R}^{n \times q}$ *be a matrix spanning* $\mathscr{C}(\mathbf{A})^{\perp} = \mathscr{N}(\mathbf{A}')$*,*
- *the columns of the matrices* $\mathbf{A}_b \in \mathbb{R}^{n \times r}$ *and* $\mathbf{A}_b^{\perp} \in \mathbb{R}^{n \times (n-r)}$ *form bases for* $\mathscr{C}(\mathbf{A})$ *and* $\mathscr{C}(\mathbf{A})^{\perp}$*, respectively.*

Then the following conditions are equivalent ways to define the unique matrix $\mathbf{P}$:

(a) *The matrix* $\mathbf{P}$ *transforms every* $\mathbf{y} \in \mathbb{R}^n$,

$$\mathbf{y} = \mathbf{y}_A + \mathbf{y}_{A^{\perp}}, \quad \mathbf{y}_A \in \mathscr{C}(\mathbf{A}),\ \mathbf{y}_{A^{\perp}} \in \mathscr{C}(\mathbf{A})^{\perp}, \tag{2.1}$$

into its projection onto $\mathscr{C}(\mathbf{A})$ *along* $\mathscr{C}(\mathbf{A})^{\perp}$*; that is, for each* $\mathbf{y}$ *in* (2.1)*, the multiplication* $\mathbf{P}\mathbf{y}$ *gives the projection* $\mathbf{y}_A$:

$$\mathbf{P}\mathbf{y} = \mathbf{y}_A\,. \tag{2.2}$$

(b) $\mathbf{P}(\mathbf{A}\mathbf{b} + \mathbf{A}^{\perp}\mathbf{c}) = \mathbf{A}\mathbf{b}$ *for all* $\mathbf{b} \in \mathbb{R}^m$, $\mathbf{c} \in \mathbb{R}^q$.

(c) $\mathbf{P}(\mathbf{A} : \mathbf{A}^{\perp}) = (\mathbf{A} : \mathbf{0})$.

(d) $\mathbf{P}(\mathbf{A}_b : \mathbf{A}_b^{\perp}) = (\mathbf{A}_b : \mathbf{0})$.

(e) $\mathscr{C}(\mathbf{P}) \subset \mathscr{C}(\mathbf{A})$, $\quad \min_{\mathbf{b}} \|\mathbf{y} - \mathbf{A}\mathbf{b}\|^2 = \|\mathbf{y} - \mathbf{P}\mathbf{y}\|^2$ *for all* $\mathbf{y} \in \mathbb{R}^n$.

S. Puntanen et al., *Matrix Tricks for Linear Statistical Models: Our Personal Top Twenty*,
DOI 10.1007/978-3-642-10473-2_3, © Springer-Verlag Berlin Heidelberg 2011

(f) $\mathscr{C}(\mathbf{P}) \subset \mathscr{C}(\mathbf{A}), \quad \mathbf{P}'\mathbf{A} = \mathbf{A}$.

(g) $\mathscr{C}(\mathbf{P}) = \mathscr{C}(\mathbf{A}), \quad \mathbf{P}'\mathbf{P} = \mathbf{P}$.

(h) $\mathscr{C}(\mathbf{P}) = \mathscr{C}(\mathbf{A}), \quad \mathbf{P}^2 = \mathbf{P}, \quad \mathbf{P}' = \mathbf{P}$.

(i) $\mathscr{C}(\mathbf{P}) = \mathscr{C}(\mathbf{A}), \quad \mathbb{R}^n = \mathscr{C}(\mathbf{P}) \boxplus \mathscr{C}(\mathbf{I}_n - \mathbf{P})$.

(j) $\mathbf{P} = \mathbf{A}(\mathbf{A}'\mathbf{A})^{-}\mathbf{A}' = \mathbf{A}\mathbf{A}^{+}$.

(k) $\mathbf{P} = \mathbf{A}_b(\mathbf{A}_b'\mathbf{A}_b)^{-1}\mathbf{A}_b'$.

(l) $\mathbf{P} = \mathbf{A}_o\mathbf{A}_o' = \mathbf{U}\left(\begin{smallmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{smallmatrix}\right)\mathbf{U}'$, *where* $\mathbf{U} = (\mathbf{A}_o : \mathbf{A}_o^{\perp})$, *the columns of* $\mathbf{A}_o$ *and* $\mathbf{A}_o^{\perp}$ *forming orthonormal bases for* $\mathscr{C}(\mathbf{A})$ *and* $\mathscr{C}(\mathbf{A})^{\perp}$, *respectively.*

The matrix $\mathbf{P} = \mathbf{P}_{\mathbf{A}}$ *is the orthogonal projector onto the column space* $\mathscr{C}(\mathbf{A})$ *with respect to the standard inner product* $\langle \mathbf{t}, \mathbf{u} \rangle = \mathbf{t}'\mathbf{u}$. *Correspondingly, the matrix* $\mathbf{I}_n - \mathbf{P}_{\mathbf{A}}$ *is the orthogonal projector onto the column space* $\mathscr{C}(\mathbf{A}^{\perp})$:

$$\mathbf{I}_n - \mathbf{P}_{\mathbf{A}} = \mathbf{P}_{\mathbf{A}^{\perp}} \,. \tag{2.3}$$

Proof. Consider the statement (a) and let $\mathbf{y} \in \mathbb{R}^n$ have a decomposition

$$\mathbf{y} = \mathbf{A}\mathbf{b} + \mathbf{A}^{\perp}\mathbf{c} \,, \tag{2.4}$$

for some vectors $\mathbf{b}$ and $\mathbf{c}$ which are not necessarily unique—even though $\mathbf{A}\mathbf{b}$ and $\mathbf{A}^{\perp}\mathbf{c}$ are unique. Premultiplying (2.4) by $\mathbf{P}$ yields

$$\mathbf{P}\mathbf{y} = \mathbf{P}(\mathbf{A}\mathbf{b} + \mathbf{A}^{\perp}\mathbf{c}) = \mathbf{A}\mathbf{b} \,. \tag{2.5}$$

Hence (b) is simply (a) represented in a shorter form. Equation (2.5) must now hold for all $\mathbf{b}$ and $\mathbf{c}$ (of conformable dimension), that is, we have

$$\mathbf{P}(\mathbf{A} : \mathbf{A}^{\perp}) \begin{pmatrix} \mathbf{b} \\ \mathbf{c} \end{pmatrix} = (\mathbf{A} : \mathbf{0}) \begin{pmatrix} \mathbf{b} \\ \mathbf{c} \end{pmatrix} \quad \text{for all } \mathbf{b}, \, \mathbf{c} \,. \tag{2.6}$$

From (2.6) we immediately conclude the equivalence of (b) and (c). Moreover, if the columns of $\mathbf{A}_b$ form a basis for $\mathscr{C}(\mathbf{A})$, then according to Theorem 17 (p. 349), the matrix $\mathbf{A}$ has a full rank decomposition

$$\mathbf{A} = \mathbf{A}_b\mathbf{L}' \quad \text{for some } \mathbf{L} \in \mathbb{R}^{m \times r}, \tag{2.7}$$

where $\mathbf{L}$ has full column rank. Hence, postmultiplying

$$\mathbf{P}\mathbf{A}_b\mathbf{L}' = \mathbf{A}_b\mathbf{L}' \tag{2.8}$$

by $\mathbf{L}(\mathbf{L}'\mathbf{L})^{-1}$ yields

$$\mathbf{P}\mathbf{A}_b = \mathbf{A}_b \,. \tag{2.9}$$

Trivially (2.9) implies (2.8). Similarly we can conclude that $\mathbf{P}\mathbf{A}^{\perp} = \mathbf{0}$ if and only if $\mathbf{P}\mathbf{A}_b^{\perp} = \mathbf{0}$ Thus we have confirmed that (a), (b), (c), and (d) are equivalent.

What about the explicit expression for $\mathbf{P}$? We may notice that utilizing some particular properties of generalized inverses, we could at once conclude that

$$\mathbf{A}(\mathbf{A}'\mathbf{A})^{-}\mathbf{A}'(\mathbf{A} : \mathbf{A}^{\perp}) = (\mathbf{A} : \mathbf{0})\,, \tag{2.10}$$

and thereby confirm that $\mathbf{A}(\mathbf{A}'\mathbf{A})^{-}\mathbf{A}'$ is an expression for $\mathbf{P}$. However, below we will *introduce*, using a longer route, expressions for $\mathbf{P}$. Before that we may recall that $\mathbf{P}$ is unique—once the column space $\mathscr{C}(\mathbf{A})$ is given. Namely, if $\mathbf{Q}$ is another matrix satisfying $\mathbf{Q}\mathbf{y} = \mathbf{P}\mathbf{y}$ for every $\mathbf{y}$, then obviously $\mathbf{Q} = \mathbf{P}$.

Denote, for simplicity, $\mathbf{A}_b^{\perp} = \mathbf{Z}$. Then it is straightforward to check that

$$(\mathbf{A}_b : \mathbf{A}_b^{\perp})^{-1} = (\mathbf{A}_b : \mathbf{Z})^{-1} = \begin{pmatrix} (\mathbf{A}_b'\mathbf{A}_b)^{-1}\mathbf{A}_b' \\ (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' \end{pmatrix}. \tag{2.11}$$

Postmultiplying $\mathbf{P}(\mathbf{A}_b : \mathbf{A}_b^{\perp}) = (\mathbf{A}_b : \mathbf{0})$ by $(\mathbf{A}_b : \mathbf{A}_b^{\perp})^{-1}$, yields (k):

$$\mathbf{P} = (\mathbf{A}_b : \mathbf{0}) \begin{pmatrix} (\mathbf{A}_b'\mathbf{A}_b)^{-1}\mathbf{A}_b' \\ (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' \end{pmatrix} = \mathbf{A}_b(\mathbf{A}_b'\mathbf{A}_b)^{-1}\mathbf{A}_b'\,. \tag{2.12}$$

Let us next introduce the expression $\mathbf{A}(\mathbf{A}'\mathbf{A})^{-}\mathbf{A}'$ for $\mathbf{P}$. Premultiplying (2.4) by $\mathbf{A}'$ yields the *normal equation*

$$\mathbf{A}'\mathbf{A}\mathbf{b} = \mathbf{A}'\mathbf{y}\,, \tag{2.13}$$

i.e.,

$$\mathbf{A}'(\mathbf{y} - \mathbf{A}\mathbf{b}) = \mathbf{0}\,. \tag{2.14}$$

From (2.14) we see that if $\mathbf{b}_*$ is a solution to the normal equation (2.13), then $\mathbf{y} - \mathbf{A}\mathbf{b}_*$ is orthogonal to $\mathscr{C}(\mathbf{A})$. Now, because

$$\mathbf{y} = \mathbf{A}\mathbf{b}_* + (\mathbf{y} - \mathbf{A}\mathbf{b}_*), \quad \text{where } \mathbf{A}\mathbf{b}_* \in \mathscr{C}(\mathbf{A}),\ \mathbf{y} - \mathbf{A}\mathbf{b}_* \in \mathscr{C}(\mathbf{A})^{\perp}, \tag{2.15}$$

the vector $\mathbf{A}\mathbf{b}_*$ is the projection of $\mathbf{y}$ onto $\mathscr{C}(\mathbf{A})$. Notice that equation (2.13) is trivially consistent because $\mathbf{A}'\mathbf{y} \in \mathscr{C}(\mathbf{A}'\mathbf{A}) = \mathscr{C}(\mathbf{A}')$. The set of all vectors $\mathbf{b}_*$ satisfying (2.13) can be expressed as

$$\mathbf{b}_* = (\mathbf{A}'\mathbf{A})^{-}\mathbf{A}'\mathbf{y} + [\mathbf{I}_m - (\mathbf{A}'\mathbf{A})^{-}\mathbf{A}'\mathbf{A}]\mathbf{t}\,, \tag{2.16}$$

where $\mathbf{t}$ is free to vary. Premultiplying (2.16) by $\mathbf{A}$, and using

$$\mathbf{A}(\mathbf{A}'\mathbf{A})^{-}\mathbf{A}'\mathbf{A} = \mathbf{A}\,, \tag{2.17}$$

gives

$$\mathbf{A}\mathbf{b}_* = \mathbf{A}(\mathbf{A}'\mathbf{A})^{-}\mathbf{A}'\mathbf{y}\,, \tag{2.18}$$

and thus we have obtained the expression (j) for matrix $\mathbf{P}$.

In passing we would like to emphasize the importance of (2.17); this result frequently appears in this book. Without using the knowledge that the matrix $\mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{A}'$ is an orthogonal projector, (2.17) can be proved using the rank cancellation rule, see Section 6.1 (p. 146), or by using the fact that $\mathbf{L}(\mathbf{A}'\mathbf{A})^-\mathbf{A}'\mathbf{A} = \mathbf{L}$ if and only if $\mathscr{C}(\mathbf{L}') \subset \mathscr{C}(\mathbf{A}'\mathbf{A})$; see Theorem 12 (p. 283).

Another way to introduce the expression (j) is the following; see, e.g., Rao (1974, Lemma 2.6). Suppose that $\mathbf{P}(\mathbf{A} : \mathbf{A}^\perp) = (\mathbf{A} : \mathbf{0})$ holds. Then

$$\mathbf{PA}^\perp = \mathbf{0} \iff \mathscr{C}(\mathbf{A}^\perp) \subset \mathscr{N}(\mathbf{P}) = \mathscr{C}(\mathbf{P}')^\perp \iff \mathscr{C}(\mathbf{P}') \subset \mathscr{C}(\mathbf{A})\,, \tag{2.19}$$

and hence

$$\mathbf{P} = \mathbf{LA}' \quad \text{for some } \mathbf{L}\,. \tag{2.20}$$

Substituting (2.20) into $\mathbf{PA} = \mathbf{A}$ yields

$$\mathbf{LA}'\mathbf{A} = \mathbf{A}\,. \tag{2.21}$$

One solution for the consistent equation (2.21) is $\mathbf{L} = \mathbf{A}(\mathbf{A}'\mathbf{A})^-$ and hence we get

$$\mathbf{P} = \mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{A}'. \tag{2.22}$$

Next we consider the connection between (e) and (g). We can write

$$\begin{aligned}\|\mathbf{y} - \mathbf{Ab}\|^2 &= \|\mathbf{y} - \mathbf{Py} + (\mathbf{Py} - \mathbf{Ab})\|^2 \\ &= \|\mathbf{y} - \mathbf{Py}\|^2 + \|\mathbf{Py} - \mathbf{Ab}\|^2 + 2\delta\,, \end{aligned} \tag{2.23}$$

where

$$\delta = \mathbf{y}'(\mathbf{I}_n - \mathbf{P})'(\mathbf{Py} - \mathbf{Ab})\,. \tag{2.24}$$

It is now clear that if $\delta = 0$, then (e) holds:

$$\|\mathbf{y} - \mathbf{Ab}\|^2 \geq \|\mathbf{y} - \mathbf{Py}\|^2 \quad \text{for all } \mathbf{b} \in \mathbb{R}^m,\ \mathbf{y} \in \mathbb{R}^n, \tag{2.25}$$

where the equality holds if and only if $\mathbf{Py} = \mathbf{Ab}$. Condition $\mathscr{C}(\mathbf{P}) = \mathscr{C}(\mathbf{A})$ implies that $\mathbf{Py} - \mathbf{Ab} = \mathbf{Py} - \mathbf{Pt} = \mathbf{Pu}$ for some $\mathbf{t}$ and $\mathbf{u}$, which together with $\mathbf{P}'\mathbf{P} = \mathbf{P}$ means that $\delta = 0$; this confirms that (g) implies (e). Note that

$$\mathbf{P}'\mathbf{P} = \mathbf{P} \iff \mathbf{P}' = \mathbf{P} \text{ and } \mathbf{P} = \mathbf{P}^2, \tag{2.26}$$

which means that (g) and (h) are equivalent.

It becomes a bit trickier to confirm that (e) implies (f). Assume now that (e) holds. Then $\mathscr{C}(\mathbf{P}) \subset \mathscr{C}(\mathbf{A})$ implies that $\mathbf{Py} - \mathbf{Ab} = \mathbf{At}$ for some $\mathbf{t}$. According to (e), we must now have

$$\|\mathbf{y} - \mathbf{Py} + \mathbf{Ab}\|^2 = \|\mathbf{y} - \mathbf{At}\|^2 \geq \|\mathbf{y} - \mathbf{Py}\|^2 \quad \text{for all } \mathbf{b},\ \mathbf{y}\,. \tag{2.27}$$

In view of Proposition 2.1 below (p. 76), (2.27) is equivalent to

$$\mathbf{y}'(\mathbf{I}_n - \mathbf{P})'\mathbf{Ab} = 0 \quad \text{for all } \mathbf{b},\ \mathbf{y}\,, \tag{2.28}$$

i.e., $\mathbf{P}'\mathbf{A} = \mathbf{A}$. Hence we have proved that (e) is equivalent to (f):

$$\text{(i) } \mathscr{C}(\mathbf{P}) \subset \mathscr{C}(\mathbf{A})\,, \qquad \text{(ii) } \mathbf{P}'\mathbf{A} = \mathbf{A}\,. \tag{f}$$

What about the connection between (f) and (g)? Condition (i) in (f) means that there exists a matrix $\mathbf{L}$ such that $\mathbf{P} = \mathbf{AL}$. Postmultiplying (ii) by $\mathbf{L}$ yields $\mathbf{P}'\mathbf{P} = \mathbf{P}$ which further implies that $\mathbf{P}' = \mathbf{P}$, and so (ii) becomes $\mathbf{PA} = \mathbf{A}$. This confirms that (f) implies the following three conditions:

$$\text{(iii) } \mathscr{C}(\mathbf{P}) \subset \mathscr{C}(\mathbf{A})\,, \quad \text{(iv) } \mathbf{P}'\mathbf{P} = \mathbf{P}\,, \quad \text{(v) } \mathbf{PA} = \mathbf{A}\,, \tag{2.29}$$

or, equivalently,

$$\text{(iii) } \mathscr{C}(\mathbf{P}) \subset \mathscr{C}(\mathbf{A})\,, \quad \text{(iv}'\text{) } \mathbf{P}' = \mathbf{P},\ \mathbf{P}^2 = \mathbf{P}\,, \quad \text{(v) } \mathbf{PA} = \mathbf{A}\,. \tag{2.30}$$

Condition (v) above implies the inclusion $\mathscr{C}(\mathbf{A}) \subset \mathscr{C}(\mathbf{P})$ which together with (iii) implies the equality $\mathscr{C}(\mathbf{P}) = \mathscr{C}(\mathbf{A})$. Moreover, $\mathscr{C}(\mathbf{P}) = \mathscr{C}(\mathbf{A})$ means that $\mathbf{A} = \mathbf{PK}$ for some $\mathbf{K}$, which together with the idempotency of $\mathbf{P}$ guarantees that

$$\mathbf{P} \cdot \mathbf{A} = \mathbf{P} \cdot \mathbf{PK} = \mathbf{PK} = \mathbf{A}\,. \tag{2.31}$$

Thus (2.30) can be replaced with

$$\text{(vi) } \mathscr{C}(\mathbf{P}) = \mathscr{C}(\mathbf{A})\,, \qquad \text{(vii) } \mathbf{P}' = \mathbf{P},\ \mathbf{P}^2 = \mathbf{P}\,, \tag{h}$$

or equivalently with

$$\text{(viii) } \mathscr{C}(\mathbf{P}) = \mathscr{C}(\mathbf{A})\,, \qquad \text{(ix) } \mathbf{P}'\mathbf{P} = \mathbf{P}\,. \tag{g}$$

Hence we have proved that (f) indeed implies (g). For an alternative proof, see Mitra & Rao (1974, Lemma 2.1). □

In the simple situation when we project $\mathbf{y} \in \mathbb{R}^n$ onto the line $\mathscr{C}(\mathbf{a})$, where $\mathbf{0} \neq \mathbf{a} \in \mathbb{R}^n$, Theorem 2 gives the projection

$$\mathbf{P_a y} = \mathbf{a}(\mathbf{a}'\mathbf{a})^{-1}\mathbf{a}'\mathbf{y} = \frac{\mathbf{a}'\mathbf{y}}{\mathbf{a}'\mathbf{a}}\,\mathbf{a} := \alpha_* \mathbf{a}\,. \tag{2.32}$$

We can now cite Steele (2004, p. 57):

> "Anyone who is already familiar with a proof of this formula [$\mathbf{P_a y}$ is the point in $\mathscr{C}(\mathbf{a})$ which is closest to $\mathbf{y}$] should rise to this challenge by looking for a new proof. In fact, the projection formula is wonderfully provable, and successful derivations may be obtained by calculus, by algebra, or even by direct arguments which require nothing more than a clever guess and Cauchy's inequality."

The following result, see Rao (1973a, p. 73, Ex. 19.4), was used in the proof of Theorem 2. We could have survived without it but it appears to be useful in several connections and hence it may be appropriate to represent it here.

Proposition 2.1. *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be* $n \times m$ *matrices and* $\mathbf{V}$ *be an* $m \times m$ *nonnegative definite (semi)inner product matrix:* $\langle \mathbf{a}, \mathbf{b} \rangle_{\mathbf{V}} = \langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a}'\mathbf{V}\mathbf{b}$. *Then the following statements are equivalent:*

(a) $\|\mathbf{A}\boldsymbol{\alpha}\|_{\mathbf{V}} \leq \|\mathbf{A}\boldsymbol{\alpha} + \mathbf{B}\boldsymbol{\beta}\|_{\mathbf{V}}$ *for all* $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^m$,

(b) $\|\mathbf{a}\|_{\mathbf{V}} \leq \|\mathbf{a} + \mathbf{b}\|_{\mathbf{V}}$ *for all* $\mathbf{a} \in \mathscr{C}(\mathbf{A})$, $\mathbf{b} \in \mathscr{C}(\mathbf{B})$,

(c) $\langle \mathbf{a}, \mathbf{b} \rangle_{\mathbf{V}} = 0$ *for all* $\mathbf{a} \in \mathscr{C}(\mathbf{A})$, $\mathbf{b} \in \mathscr{C}(\mathbf{B})$,

(d) $\mathbf{A}'\mathbf{V}\mathbf{B} = \mathbf{0}$.

Proof. Of course, (b) is simply another way to express (a), and similarly (d) is an alternative way to write (c). Assume now that (c) holds. Then (dropping off the subscript $\mathbf{V}$) $\|\mathbf{a}+\mathbf{b}\|^2 = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2$, which implies (b). Assume then that (b) holds. Then we also have

$$\|\mathbf{a}\|^2 \leq \|\mathbf{a} + \lambda\mathbf{b}\|^2 \quad \text{for all } \mathbf{a} \in \mathscr{C}(\mathbf{A}),\ \mathbf{b} \in \mathscr{C}(\mathbf{B}),\ \lambda \in \mathbb{R}, \tag{2.33}$$

i.e., $\|\mathbf{a}\|^2 \leq \|\mathbf{a}\|^2 + \lambda^2\|\mathbf{b}\|^2 + 2\lambda\langle \mathbf{a}, \mathbf{b} \rangle$, that is,

$$\lambda^2\|\mathbf{b}\|^2 + 2\lambda\langle \mathbf{a}, \mathbf{b} \rangle := u^2\lambda^2 + 2v\lambda := f(\lambda) \geq 0 \quad \text{for all } \lambda \in \mathbb{R}. \tag{2.34}$$

In order for (2.34) to hold, the term $v = \langle \mathbf{a}, \mathbf{b} \rangle$ must (why?) be equal to 0. Hence our claim "(b) $\implies$ (c)" is proved. □

For a clarity, we may re-express one crucial finding from Theorem 2:

Proposition 2.2 (Normal equations). *Let* $\mathbf{b}_*$ *be an arbitrary solution for the normal equation* $\mathbf{A}'\mathbf{A}\mathbf{b} = \mathbf{A}'\mathbf{y}$. *Then* $\mathbf{A}\mathbf{b}_*$ *is the orthogonal projection of* $\mathbf{y}$ *onto* $\mathscr{C}(\mathbf{A})$, *i.e.,* $\mathbf{A}\mathbf{b}_* = \mathbf{P}_{\mathbf{A}}\mathbf{y}$, *and*

$$\min_{\mathbf{b}} \|\mathbf{y} - \mathbf{A}\mathbf{b}\|^2 = \|\mathbf{y} - \mathbf{A}\mathbf{b}_*\|^2 = \mathbf{y}'(\mathbf{I}_n - \mathbf{P}_{\mathbf{A}})\mathbf{y}. \tag{2.35}$$

Proof. Let $\mathbf{b}_*$ be a solution to the normal equation

$$\mathbf{A}'\mathbf{A}\mathbf{b} = \mathbf{A}'\mathbf{y}. \tag{2.36}$$

Then

$$\begin{aligned}
\|\mathbf{y} - \mathbf{A}\mathbf{b}\|^2 &= \|\mathbf{y} - \mathbf{A}\mathbf{b}_* + (\mathbf{A}\mathbf{b}_* - \mathbf{A}\mathbf{b})\|^2 \\
&= \|\mathbf{y} - \mathbf{A}\mathbf{b}_*\|^2 + \|\mathbf{A}\mathbf{b}_* - \mathbf{A}\mathbf{b}\|^2 + 2(\mathbf{A}\mathbf{b}_* - \mathbf{A}\mathbf{b})'(\mathbf{y} - \mathbf{A}\mathbf{b}_*) \\
&= \|\mathbf{y} - \mathbf{A}\mathbf{b}_*\|^2 + \|\mathbf{A}(\mathbf{b}_* - \mathbf{b})\|^2 + 2\mathbf{z}'\mathbf{A}'(\mathbf{y} - \mathbf{A}\mathbf{b}_*) \\
&= \|\mathbf{y} - \mathbf{A}\mathbf{b}_*\|^2 + \|\mathbf{A}(\mathbf{b}_* - \mathbf{b})\|^2,
\end{aligned} \tag{2.37}$$

where $\mathbf{z} = \mathbf{b}_* - \mathbf{b}$. The last equality above follows from (2.36). Hence we have, for any $\mathbf{b}_*$ satisfying the normal equation (2.36),

$$\|\mathbf{y} - \mathbf{A}\mathbf{b}\|^2 \geq \|\mathbf{y} - \mathbf{A}\mathbf{b}_*\|^2 \quad \text{for all } \mathbf{b},\, \mathbf{y}. \tag{2.38}$$

2.1 Minimal Angle

The following result is almost obvious.

Proposition 2.3 (Maximal cosine and orthogonal projector). *The maximal cosine (minimal angle) between the nonzero vectors* $\mathbf{y}$ *and* $\mathbf{Ab}$ *(for variable* $\mathbf{b}$*) is obtained when* $\mathbf{Ab} = a\mathbf{P_A y}$ *for some* $a \in \mathbb{R}$:

$$\max_{\mathbf{Ab}\neq\mathbf{0}} \cos^2(\mathbf{y}, \mathbf{Ab}) = \max_{\mathbf{Ab}\neq\mathbf{0}} \frac{(\mathbf{y}'\mathbf{Ab})^2}{\mathbf{y}'\mathbf{y}\cdot\mathbf{b}'\mathbf{A}'\mathbf{Ab}} = \frac{\mathbf{y}'\mathbf{P_A y}}{\mathbf{y}'\mathbf{y}}\,. \tag{2.39}$$

Proof. In view of the Cauchy–Schwarz inequality, Theorem 20 (p. 415), we get

$$(\mathbf{y}'\mathbf{Ab})^2 = (\mathbf{y}'\mathbf{P_A}\cdot\mathbf{Ab})^2 \leq \mathbf{y}'\mathbf{P_A y}\cdot\mathbf{b}'\mathbf{A}'\mathbf{Ab}\,, \tag{2.40}$$

and so, for $\mathbf{y} \neq \mathbf{0}$, $\mathbf{Ab} \neq \mathbf{0}$,

$$\cos^2(\mathbf{y}, \mathbf{Ab}) = \frac{(\mathbf{y}'\mathbf{Ab})^2}{\mathbf{y}'\mathbf{y}\cdot\mathbf{b}'\mathbf{A}'\mathbf{Ab}} \leq \frac{\mathbf{y}'\mathbf{P_A y}\cdot\mathbf{b}'\mathbf{A}'\mathbf{Ab}}{\mathbf{y}'\mathbf{y}\cdot\mathbf{b}'\mathbf{A}'\mathbf{Ab}} = \frac{\mathbf{y}'\mathbf{P_A y}}{\mathbf{y}'\mathbf{y}}\,. \tag{2.41}$$

An alternative simple proof goes as follows. First, it is clear that

$$\cos^2(\mathbf{y}, \mathbf{Ab}) = \frac{\|\mathbf{P_{Ab}y}\|^2}{\|\mathbf{y}\|^2} = \frac{\mathbf{y}'\mathbf{P_{Ab}y}}{\mathbf{y}'\mathbf{y}}\,. \tag{2.42}$$

On account of Proposition 7.1 (p. 152), there is the Löwner ordering

$$\mathbf{P_{Ab}} \leq_{\mathsf{L}} \mathbf{P_A} \quad \text{for all } \mathbf{b} \in \mathbb{R}^m, \tag{2.43}$$

i.e.,

$$\|\mathbf{P_{Ab}y}\|^2 \leq \|\mathbf{P_A y}\|^2 \quad \text{for all } \mathbf{b} \in \mathbb{R}^m,\ \mathbf{y} \in \mathbb{R}^n, \tag{2.44}$$

which proves (2.39). □

The minimal angle between the nonzero vectors $\mathbf{Ab}$ and $\mathbf{y}$ can be seen as the minimal angle between the subspaces $\mathcal{A} = \mathscr{C}(\mathbf{A})$ and $\mathcal{B} = \mathscr{C}(\mathbf{y})$:

$$\max_{\substack{\mathbf{u}\in\mathcal{A},\,\mathbf{v}\in\mathcal{B}\\ \mathbf{u}\neq\mathbf{0},\,\mathbf{v}\neq\mathbf{0}}} \cos^2(\mathbf{u}, \mathbf{v}) = \cos^2 \min_{\substack{\mathbf{u}\in\mathcal{A},\,\mathbf{v}\in\mathcal{B}\\ \mathbf{u}\neq\mathbf{0},\,\mathbf{v}\neq\mathbf{0}}} \angle(\mathbf{u}, \mathbf{v})\,. \tag{2.45}$$

The minimal angle between the subspaces $\mathcal{A} = \mathscr{C}(\mathbf{A})$ and $\mathcal{B} = \mathscr{C}(\mathbf{B})$ is defined to be the number $0 \leq \theta_{\min} \leq \pi/2$ for which

$$\begin{aligned}\cos^2\theta_{min} = \max_{\substack{\mathbf{u}\in\mathcal{A},\,\mathbf{v}\in\mathcal{B}\\ \mathbf{u}\neq\mathbf{0},\,\mathbf{v}\neq\mathbf{0}}} \cos^2(\mathbf{u}, \mathbf{v}) &= \max_{\substack{\mathbf{A}\boldsymbol{\alpha}\neq\mathbf{0}\\ \mathbf{B}\boldsymbol{\beta}\neq\mathbf{0}}} \frac{(\boldsymbol{\alpha}'\mathbf{A}'\mathbf{B}\boldsymbol{\beta})^2}{\boldsymbol{\alpha}'\mathbf{A}'\mathbf{A}\boldsymbol{\alpha}\cdot\boldsymbol{\beta}'\mathbf{B}'\mathbf{B}\boldsymbol{\beta}} \\ &:= \max_{\substack{\mathbf{A}\boldsymbol{\alpha}\neq\mathbf{0}\\ \mathbf{B}\boldsymbol{\beta}\neq\mathbf{0}}} f(\boldsymbol{\alpha}, \boldsymbol{\beta})\,. \end{aligned} \tag{2.46}$$

The above task in the case when $\mathbf{A}$ and $\mathbf{B}$ are columnwise orthonormal is considered via the singular value decomposition on page 134. Here we use a different approach. Notice that obviously $\cos^2\theta_{min} = 0$ if and only if $\mathscr{C}(\mathbf{A})$ and $\mathscr{C}(\mathbf{B})$ are orthogonal to each other, i.e., $\mathbf{A}'\mathbf{B} = \mathbf{0}$. Similarly $\cos^2\theta_{min} = 1$ if and only if $\mathscr{C}(\mathbf{A})$ and $\mathscr{C}(\mathbf{B})$ are not disjoint, i.e., $\mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B}) \neq \{\mathbf{0}\}$.

Let us first keep the vector $\boldsymbol{\alpha}$ given and let $\boldsymbol{\beta}$ vary. Then, in light of Proposition 2.3, for $\mathbf{A}\boldsymbol{\alpha} \neq \mathbf{0}$ we have

$$\max_{\mathbf{B}\boldsymbol{\beta}\neq\mathbf{0}} \cos^2(\mathbf{A}\boldsymbol{\alpha}, \mathbf{B}\boldsymbol{\beta}) = \max_{\mathbf{B}\boldsymbol{\beta}\neq\mathbf{0}} f(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{\boldsymbol{\alpha}'\mathbf{A}'\mathbf{P}_{\mathbf{B}}\mathbf{A}\boldsymbol{\alpha}}{\boldsymbol{\alpha}'\mathbf{A}'\mathbf{A}\boldsymbol{\alpha}} := g(\boldsymbol{\alpha}), \tag{2.47}$$

and the maximum is attained if and only if $\mathbf{B}\boldsymbol{\beta}$ is a multiple of $\mathbf{P}_{\mathbf{B}}\mathbf{A}\boldsymbol{\alpha}$ i.e.,

$$\mathbf{B}\boldsymbol{\beta} = a\mathbf{P}_{\mathbf{B}}\mathbf{A}\boldsymbol{\alpha} \quad \text{for some nonzero } a \in \mathbb{R}. \tag{2.48}$$

In view of Proposition 18.9 (p. 377),

$$\begin{aligned}\max_{\mathbf{A}\boldsymbol{\alpha}\neq\mathbf{0}} \frac{\boldsymbol{\alpha}'\mathbf{A}'\mathbf{P}_{\mathbf{B}}\mathbf{A}\boldsymbol{\alpha}}{\boldsymbol{\alpha}'\mathbf{A}'\mathbf{A}\boldsymbol{\alpha}} &= \frac{\boldsymbol{\alpha}_1'\mathbf{A}'\mathbf{P}_{\mathbf{B}}\mathbf{A}\boldsymbol{\alpha}_1}{\boldsymbol{\alpha}_1'\mathbf{A}'\mathbf{A}\boldsymbol{\alpha}_1} \\ &= \mathrm{ch}_1[(\mathbf{A}'\mathbf{A})^+\mathbf{A}'\mathbf{P}_{\mathbf{B}}\mathbf{A})] \\ &= \mathrm{ch}_1[\mathbf{A}(\mathbf{A}'\mathbf{A})^+\mathbf{A}'\mathbf{P}_{\mathbf{B}}] \\ &= \mathrm{ch}_1(\mathbf{P}_{\mathbf{A}}\mathbf{P}_{\mathbf{B}}) = \lambda_1^2. \end{aligned} \tag{2.49}$$

It is noteworthy that according to Proposition 18.9 (p. 377), λ_1^2 in (2.49) is also the largest proper eigenvalue of $\mathbf{A}'\mathbf{P}_{\mathbf{B}}\mathbf{A}$ with respect to $\mathbf{A}'\mathbf{A}$ and $\boldsymbol{\alpha}_1$ is the corresponding proper eigenvector satisfying

$$\mathbf{A}'\mathbf{P}_{\mathbf{B}}\mathbf{A}\boldsymbol{\alpha}_1 = \lambda_1^2\mathbf{A}'\mathbf{A}\boldsymbol{\alpha}_1, \quad \mathbf{A}\boldsymbol{\alpha}_1 \neq \mathbf{0}. \tag{2.50}$$

Premultiplying (2.50) by $\mathbf{A}(\mathbf{A}'\mathbf{A})^+$ yields

$$\mathbf{P}_{\mathbf{A}}\mathbf{P}_{\mathbf{B}}\mathbf{A}\boldsymbol{\alpha}_1 = \lambda_1^2\mathbf{A}\boldsymbol{\alpha}_1, \tag{2.51}$$

showing that $\mathbf{A}\boldsymbol{\alpha}_1$ is an eigenvector of $\mathbf{P}_{\mathbf{A}}\mathbf{P}_{\mathbf{B}}$.

Having now characterized $\boldsymbol{\alpha}_1$ so that it maximizes $g(\boldsymbol{\alpha})$, we should find $\boldsymbol{\beta}_1$ so that $f(\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1)$ attains its maximum; this happens if and only if (2.48) holds. Premultiplying (2.51) by $\mathbf{P}_{\mathbf{B}}$ and using (2.48) we can conclude that

$$\mathbf{P}_{\mathbf{B}}\mathbf{P}_{\mathbf{A}}\mathbf{P}_{\mathbf{B}}\mathbf{A}\boldsymbol{\alpha}_1 = \lambda_1^2\mathbf{P}_{\mathbf{B}}\mathbf{A}\boldsymbol{\alpha}_1, \tag{2.52a}$$

$$\mathbf{P}_{\mathbf{B}}\mathbf{P}_{\mathbf{A}}\mathbf{B}\boldsymbol{\beta}_1 = \lambda_1^2\mathbf{B}\boldsymbol{\beta}_1, \tag{2.52b}$$

and

$$\mathbf{B}'\mathbf{P}_{\mathbf{A}}\mathbf{B}\boldsymbol{\beta}_1 = \lambda_1^2\mathbf{B}'\mathbf{B}\boldsymbol{\beta}_1. \tag{2.53}$$

We can summarize our conclusions in the following proposition.

Proposition 2.4. *Let* $\mathbf{A}_{n\times a}$ *and* $\mathbf{B}_{n\times b}$ *be given matrices. Then*

$$\max_{\substack{\mathbf{A}\boldsymbol{\alpha}\neq\mathbf{0},\\ \mathbf{B}\boldsymbol{\beta}\neq\mathbf{0}}} \frac{(\boldsymbol{\alpha}'\mathbf{A}'\mathbf{B}\boldsymbol{\beta})^2}{\boldsymbol{\alpha}'\mathbf{A}'\mathbf{A}\boldsymbol{\alpha}\cdot\boldsymbol{\beta}'\mathbf{B}'\mathbf{B}\boldsymbol{\beta}} = \frac{(\boldsymbol{\alpha}_1'\mathbf{A}'\mathbf{B}\boldsymbol{\beta}_1)^2}{\boldsymbol{\alpha}_1'\mathbf{A}'\mathbf{A}\boldsymbol{\alpha}_1\cdot\boldsymbol{\beta}_1'\mathbf{B}'\mathbf{B}\boldsymbol{\beta}_2}$$

$$= \frac{\boldsymbol{\alpha}_1'\mathbf{A}'\mathbf{P_B}\mathbf{A}\boldsymbol{\alpha}_1}{\boldsymbol{\alpha}_1'\mathbf{A}'\mathbf{A}\boldsymbol{\alpha}_1} = \mathrm{ch}_1(\mathbf{P_A}\mathbf{P_B}) = \lambda_1^2\,, \tag{2.54}$$

where λ_1^2 *is the largest proper eigenvalue of* $\mathbf{A}'\mathbf{P_B}\mathbf{A}$ *with respect to* $\mathbf{A}'\mathbf{A}$ *and* $\boldsymbol{\alpha}_1$ *is the corresponding proper eigenvector satisfying*

$$\mathbf{A}'\mathbf{P_B}\mathbf{A}\boldsymbol{\alpha}_1 = \lambda_1^2\mathbf{A}'\mathbf{A}\boldsymbol{\alpha}_1\,, \quad \mathbf{A}\boldsymbol{\alpha}_1 \neq \mathbf{0}\,. \tag{2.55}$$

The vector $\boldsymbol{\beta}_1$ *is the proper eigenvector satisfying*

$$\mathbf{B}'\mathbf{P_A}\mathbf{B}\boldsymbol{\beta}_1 = \lambda_1^2\mathbf{B}'\mathbf{B}\boldsymbol{\beta}_1\,, \quad \mathbf{B}\boldsymbol{\beta}_1 \neq \mathbf{0}\,. \tag{2.56}$$

Remark 2.1. Notice that trivially

$$\max_{\mathbf{b}} \cos^2(\mathbf{y},\mathbf{A}\mathbf{b}) = 1 - \min_{\mathbf{b}} \sin^2(\mathbf{y},\mathbf{A}\mathbf{b})\,, \tag{2.57}$$

and that

$$\sin^2(\mathbf{y},\mathbf{A}\mathbf{b}) = \frac{\|(\mathbf{I}_n-\mathbf{P}_{\mathbf{Ab}})\mathbf{y}\|^2}{\|\mathbf{y}\|^2} \geq \frac{\|(\mathbf{I}_n-\mathbf{P_A})\mathbf{y}\|^2}{\|\mathbf{y}\|^2} = \sin^2(\mathbf{y},\mathbf{P_A}\mathbf{y})\,. \tag{2.58}$$

□

Remark 2.2. It is worth emphasizing that the tasks of solving $\mathbf{b}$ from

$$\min_{\mathbf{b}} \|\mathbf{y}-\mathbf{A}\mathbf{b}\|^2 \quad \text{and} \quad \max_{\mathbf{b}} \cos^2(\mathbf{y},\mathbf{A}\mathbf{b})\,, \tag{2.59}$$

yield essentially the same solutions $\mathbf{A}\mathbf{b}_* = \mathbf{P_A}\mathbf{y}$. Obviously $\cos^2(\mathbf{y},\mathbf{A}\mathbf{b}) = \cos^2(\mathbf{y},\lambda\mathbf{A}\mathbf{b})$ for any $\lambda \neq 0$. The equivalence of the tasks in (2.59) will be utilized in many connections. □

2.2 Minimizing $\mathrm{var_d}(\mathbf{y}-\mathbf{X}_0\mathbf{b})$

As an example of the usefulness of Remark 2.2, consider the following tasks:

$$\text{(i)}\ \min_{\mathbf{b}} \mathrm{var_d}(\mathbf{y}-\mathbf{X}_0\mathbf{b}) \quad \text{and} \quad \text{(ii)}\ \max_{\mathbf{b}} \mathrm{cor}_{\mathrm{d}}^2(\mathbf{y},\mathbf{X}_0\mathbf{b})\,, \tag{2.60}$$

where $\mathbf{U} = (\mathbf{X}_0 : \mathbf{y}) = (\mathbf{x}_1 : \ldots : \mathbf{x}_k : \mathbf{y})$ is a given $n \times p$ data matrix, $p = k+1$. Let $\mathbf{C}$ be the centering matrix and denote the centered $\mathbf{U}$ as $\mathbf{C}\mathbf{U} = \widetilde{\mathbf{U}} = (\widetilde{\mathbf{X}}_0 : \tilde{\mathbf{y}})$. Then

$$\mathrm{var_d}(\mathbf{y} - \mathbf{X}_0\mathbf{b}) = \tfrac{1}{n-1}\|\tilde{\mathbf{y}} - \widetilde{\mathbf{X}}_0\mathbf{b}\|^2, \tag{2.61}$$

$$\mathrm{cor_d^2}(\mathbf{y}, \mathbf{X}_0\mathbf{b}) = \cos^2(\tilde{\mathbf{y}}, \widetilde{\mathbf{X}}_0\mathbf{b})\,. \tag{2.62}$$

In view of Remark 2.2, the solution $\mathbf{b}_*$ to the above problems is

$$\mathbf{b}_* = (\widetilde{\mathbf{X}}_0'\widetilde{\mathbf{X}}_0)^{-1}\widetilde{\mathbf{X}}_0'\tilde{\mathbf{y}} = \mathbf{T}_{\mathbf{xx}}^{-1}\mathbf{t}_{\mathbf{x}y} = \mathbf{S}_{\mathbf{xx}}^{-1}\mathbf{s}_{\mathbf{x}y} := \hat{\boldsymbol{\beta}}_{\mathbf{x}}\,, \tag{2.63}$$

where $\widetilde{\mathbf{X}}_0$ is assumed to have full column rank and

$$\mathbf{T} = \widetilde{\mathbf{U}}'\widetilde{\mathbf{U}} = \begin{pmatrix} \widetilde{\mathbf{X}}_0'\widetilde{\mathbf{X}}_0 & \widetilde{\mathbf{X}}_0'\tilde{\mathbf{y}} \\ \tilde{\mathbf{y}}'\widetilde{\mathbf{X}}_0 & \tilde{\mathbf{y}}'\tilde{\mathbf{y}} \end{pmatrix} = \begin{pmatrix} \mathbf{T}_{\mathbf{xx}} & \mathbf{t}_{\mathbf{x}y} \\ \mathbf{t}_{\mathbf{x}y}' & t_{yy} \end{pmatrix}, \tag{2.64a}$$

$$\mathbf{S} = \tfrac{1}{n-1}\mathbf{T}\,, \quad \mathbf{S}_{\mathbf{xx}} = \mathrm{cov_d}(\mathbf{X}_0)\,, \quad \mathbf{s}_{\mathbf{x}y} = \mathrm{cov_d}(\mathbf{X}_0, \mathbf{y})\,. \tag{2.64b}$$

Then

$$\begin{aligned} \min_{\mathbf{b}} \mathrm{var_d}(\mathbf{y} - \mathbf{X}_0\mathbf{b}) &= \tfrac{1}{n-1}\,\tilde{\mathbf{y}}'(\mathbf{I}_n - \mathbf{P}_{\widetilde{\mathbf{X}}_0})\tilde{\mathbf{y}} = \tfrac{1}{n-1}(t_{yy} - \mathbf{t}_{\mathbf{x}y}'\mathbf{T}_{\mathbf{xx}}^{-1}\mathbf{t}_{\mathbf{x}y}) \\ &= s_y^2 - \mathbf{s}_{\mathbf{x}y}'\mathbf{S}_{\mathbf{xx}}^{-1}\mathbf{s}_{\mathbf{x}y} = \tfrac{1}{n-1}\,\mathbf{y}'(\mathbf{I}_n - \mathbf{H})\mathbf{y}\,, \end{aligned} \tag{2.65}$$

where the last equality follows from the decomposition

$$\mathbf{I}_n - \mathbf{H} = \mathbf{I}_n - \mathbf{P}_{(\mathbf{1}\,:\,\mathbf{X}_0)} = \mathbf{C} - \mathbf{P}_{\mathbf{CX}_0} = \mathbf{C} - \mathbf{CX}_0\mathbf{T}_{\mathbf{xx}}^{-1}\mathbf{X}_0'\mathbf{C}\,, \tag{2.66}$$

see (8.108) (p. 170). Moreover,

$$\max_{\mathbf{b}} \mathrm{cor_d^2}(\mathbf{y}, \mathbf{X}_0\mathbf{b}) = \frac{\mathbf{s}_{\mathbf{x}y}'\mathbf{S}_{\mathbf{xx}}^{-1}\mathbf{s}_{\mathbf{x}y}}{s_y^2} = \frac{\mathbf{t}_{\mathbf{x}y}'\mathbf{T}_{\mathbf{xx}}^{-1}\mathbf{t}_{\mathbf{x}y}}{t_{yy}} = R_{y\cdot\mathbf{x}}^2\,, \tag{2.67}$$

where $R_{y\cdot\mathbf{x}}^2$ refers to the sample multiple correlation coefficient squared.

2.3 Minimizing $(\mathbf{Y} - \mathbf{XB})'(\mathbf{Y} - \mathbf{XB})$

For a given matrix $\mathbf{X} \in \mathbb{R}^{n\times m}$ (notation $\mathbf{X}$ having now no relation to the model matrix) we have seen that for all $\mathbf{b} \in \mathbb{R}^m$

$$(\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb}) \geq (\mathbf{y} - \mathbf{P_X y})'(\mathbf{y} - \mathbf{P_X y}) = \mathbf{y}'(\mathbf{I}_n - \mathbf{P_X})\mathbf{y}\,, \tag{2.68}$$

where the equality is attained if $\mathbf{Xb} = \mathbf{P_X y}$. What about if $\mathbf{y}$ and $\mathbf{b}$ are replaced by $\mathbf{Y} \in \mathbb{R}^{n\times p}$ and $\mathbf{B} \in \mathbb{R}^{m\times p}$? Then our object for minimization would be the $p \times p$ matrix $(\mathbf{Y} - \mathbf{XB})'(\mathbf{Y} - \mathbf{XB})$. We might make a guess that corresponding to (2.68), for all $\mathbf{B} \in \mathbb{R}^{m\times p}$ we have the following Löwner ordering:

$$(\mathbf{Y} - \mathbf{XB})'(\mathbf{Y} - \mathbf{XB}) \geq_{\mathsf{L}} (\mathbf{Y} - \mathbf{P_X Y})'(\mathbf{Y} - \mathbf{P_X Y}) = \mathbf{Y}'\mathbf{Q_X Y}, \tag{2.69}$$

where $\mathbf{Q_X} = \mathbf{I}_n - \mathbf{P_X}$. This is indeed true as can be easily seen. Namely,

$$\begin{aligned}
(\mathbf{Y} - \mathbf{XB})'(\mathbf{Y} - \mathbf{XB}) \\
&= (\mathbf{Y} - \mathbf{P_X Y} + \mathbf{P_X Y} - \mathbf{XB})'(\mathbf{Y} - \mathbf{P_X Y} + \mathbf{P_X Y} - \mathbf{XB}) \\
&:= (\mathbf{Q_X Y} + \mathbf{XL})'(\mathbf{Q_X Y} + \mathbf{XL}) \qquad [\mathbf{XL} := \mathbf{P_X Y} - \mathbf{XB}] \\
&= \mathbf{Y'Q_X Y} + \mathbf{L'X'XL} \\
&\geq_{\mathsf{L}} \mathbf{Y'Q_X Y}. \qquad (2.70)
\end{aligned}$$

Let us write a proposition about the above result.

Proposition 2.5. *Suppose that* $\mathbf{X} \in \mathbb{R}^{n\times m}$ *is a given matrix. Then for any* $\mathbf{Y} \in \mathbb{R}^{n\times p}$ *and* $\mathbf{B} \in \mathbb{R}^{m\times p}$ *we have the Löwner partial ordering*

$$(\mathbf{Y} - \mathbf{XB})'(\mathbf{Y} - \mathbf{XB}) \geq_{\mathsf{L}} (\mathbf{Y} - \mathbf{P_X Y})'(\mathbf{Y} - \mathbf{P_X Y}) = \mathbf{Y'Q_X Y}. \qquad (2.71)$$

The equality in (2.71) *holds if and only if* $\mathbf{XB} := \mathbf{XB}_* = \mathbf{P_X Y}$. *The matrix* $\mathbf{B}_*$ *can be written as*

$$\mathbf{B}_* = (\mathbf{X'X})^- \mathbf{X'Y}. \qquad (2.72)$$

2.4 Orthogonal Projector with Respect to V

Proposition 2.6 (Orthogonal projector with respect to the inner product matrix V). *Let* $\mathbf{A}$ *be an* $n \times m$ *matrix,* $\operatorname{rank}(\mathbf{A}) = r > 0$, *and let the inner product* (*and the corresponding norm*) *be defined as* $\langle \mathbf{t}, \mathbf{u} \rangle_{\mathbf{V}} = \mathbf{t'Vu}$ *where* $\mathbf{V}$ *is a positive definite symmetric matrix. Further, let* $\mathbf{A}^{\perp}_{\mathbf{V}}$ *be an* $n \times q$ *matrix spanning*

$$\mathscr{C}(\mathbf{A})^{\perp}_{\mathbf{V}} = \mathscr{N}(\mathbf{A'V}) = \mathscr{C}(\mathbf{VA})^{\perp} = \mathscr{C}(\mathbf{V}^{-1}\mathbf{A}^{\perp}). \qquad (2.73)$$

Then the following conditions are equivalent ways to define the unique matrix $\mathbf{P}_*$:

(a) $\mathbf{P}_*(\mathbf{Ab} + \mathbf{A}^{\perp}_{\mathbf{V}}\mathbf{c}) = \mathbf{Ab}$ *for all* $\mathbf{b} \in \mathbb{R}^m$, $\mathbf{c} \in \mathbb{R}^q$.

(b) $\mathbf{P}_*(\mathbf{A} : \mathbf{A}^{\perp}_{\mathbf{V}}) = \mathbf{P}_*(\mathbf{A} : \mathbf{V}^{-1}\mathbf{A}^{\perp}) = (\mathbf{A} : \mathbf{0})$.

(c) $\mathscr{C}(\mathbf{P}_*) \subset \mathscr{C}(\mathbf{A})$, $\quad \min_{\mathbf{b}} \|\mathbf{y} - \mathbf{Ab}\|^2_{\mathbf{V}} = \|\mathbf{y} - \mathbf{P}_*\mathbf{y}\|^2_{\mathbf{V}}$ *for all* $\mathbf{y} \in \mathbb{R}^n$.

(d) $\mathscr{C}(\mathbf{P}_*) \subset \mathscr{C}(\mathbf{A})$, $\quad \mathbf{P}'_*\mathbf{VA} = \mathbf{VA}$.

(e) $\mathbf{P}'_*(\mathbf{VA} : \mathbf{A}^{\perp}) = (\mathbf{VA} : \mathbf{0})$.

(f) $\mathscr{C}(\mathbf{P}_*) = \mathscr{C}(\mathbf{A})$, $\quad \mathbf{P}'_*\mathbf{V}(\mathbf{I}_n - \mathbf{P}_*) = \mathbf{0}$.

(g) $\mathscr{C}(\mathbf{P}_*) = \mathscr{C}(\mathbf{A})$, $\quad \mathbf{P}^2_* = \mathbf{P}_*$, $\quad (\mathbf{VP}_*)' = \mathbf{VP}_*$.

(h) $\mathscr{C}(\mathbf{P}_*) = \mathscr{C}(\mathbf{A})$, $\quad \mathbb{R}^n = \mathscr{C}(\mathbf{P}_*) \boxplus \mathscr{C}(\mathbf{I}_n - \mathbf{P}_*)$; *here* $\boxplus$ *refers to the orthogonality with respect to the given inner product.*

(i) $\mathbf{P}_* = \mathbf{A}(\mathbf{A}'\mathbf{V}\mathbf{A})^-\mathbf{A}'\mathbf{V}$, *which is invariant for any choice of* $(\mathbf{A}'\mathbf{V}\mathbf{A})^-$, *and thereby* $\mathbf{P}_* = \mathbf{A}(\mathbf{A}'\mathbf{V}\mathbf{A})^+\mathbf{A}'\mathbf{V}$.

The matrix $\mathbf{P}_* = \mathbf{P}_{\mathbf{A};\mathbf{V}}$ *is the orthogonal projector onto the column space* $\mathscr{C}(\mathbf{A})$ *with respect to the inner product* $\langle \mathbf{t}, \mathbf{u}\rangle_{\mathbf{V}} = \mathbf{t}'\mathbf{V}\mathbf{u}$. *Correspondingly, the matrix* $\mathbf{I}_n - \mathbf{P}_{\mathbf{A};\mathbf{V}}$ *is the orthogonal projector onto the column space* $\mathscr{C}(\mathbf{A}^\perp_{\mathbf{V}})$:

$$\begin{aligned}\mathbf{I}_n - \mathbf{P}_{\mathbf{A};\mathbf{V}} &= \mathbf{P}_{\mathbf{A}^\perp_{\mathbf{V}};\mathbf{V}} = \mathbf{V}^{-1}\mathbf{Z}(\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z})^-\mathbf{Z}' \\ &= (\mathbf{P}_{\mathbf{Z};\mathbf{V}^{-1}})' = \mathbf{P}'_{\mathbf{Z};\mathbf{V}^{-1}} = \mathbf{P}_{\mathbf{V}^{-1}\mathbf{Z};\mathbf{V}}\,, \end{aligned} \tag{2.74}$$

where $\mathbf{Z} \in \{\mathbf{A}^\perp\}$, *i.e.,*

$$\mathbf{P}_{\mathbf{A};\mathbf{V}} = \mathbf{I}_n - \mathbf{P}'_{\mathbf{A}^\perp;\mathbf{V}^{-1}} = \mathbf{I}_n - \mathbf{P}_{\mathbf{V}^{-1}\mathbf{A}^\perp;\mathbf{V}}\,. \tag{2.75}$$

Proof. As was mentioned earlier, it might be more natural to put Proposition 2.6 as Theorem 2 and then consider the simpler situation but we believe it is more instructive to start with a simpler case. In any event, the proof below is very much parallel to that of Theorem 2.

Let us first confirm that (2.73) indeed holds. Just as in (0.58) (p. 13), we have

$$\begin{aligned}\mathscr{C}(\mathbf{A})^\perp_{\mathbf{V}} &= \{\,\mathbf{y} \in \mathbb{R}^n : \mathbf{z}'\mathbf{A}'\mathbf{V}\mathbf{y} = 0 \text{ for all } \mathbf{z} \in \mathbb{R}^m\,\} \\ &= \{\,\mathbf{y} \in \mathbb{R}^n : \mathbf{A}'\mathbf{V}\mathbf{y} = \mathbf{0}\,\} \\ &= \mathscr{N}(\mathbf{A}'\mathbf{V}) = \mathscr{C}(\mathbf{V}\mathbf{A})^\perp = \mathscr{C}(\mathbf{V}^{-1}\mathbf{A}^\perp)\,, \end{aligned} \tag{2.76}$$

where the last equality can be concluded from

$$\mathbf{A}'\mathbf{V} \cdot \mathbf{V}^{-1}\mathbf{A}^\perp = \mathbf{0} \implies \mathscr{C}(\mathbf{V}^{-1}\mathbf{A}^\perp) \subset \mathscr{N}(\mathbf{A}'\mathbf{V})\,, \tag{2.77}$$

and

$$\operatorname{rank}(\mathbf{V}^{-1}\mathbf{A}^\perp) = \operatorname{rank}(\mathbf{A}^\perp) = n - \operatorname{rank}(\mathbf{A})\,, \tag{2.78a}$$

$$\dim \mathscr{C}(\mathbf{V}\mathbf{A})^\perp = n - \operatorname{rank}(\mathbf{V}\mathbf{A}) = n - \operatorname{rank}(\mathbf{A})\,. \tag{2.78b}$$

Now the first equality in (2.74) is obvious (or almost obvious) while the other representations follow by simple substitutions in view of (2.76). The formula (2.75) is very useful and it is utilized in several places in this book.

Corresponding to (2.4), for every $\mathbf{y} \in \mathbb{R}^n$ we have the decomposition

$$\mathbf{y} = \mathbf{A}\mathbf{b} + \mathbf{A}^\perp_{\mathbf{V}}\mathbf{c} = \mathbf{A}\mathbf{b} + \mathbf{V}^{-1}\mathbf{A}^\perp\mathbf{c}\,, \tag{2.79}$$

for some vectors $\mathbf{b}$ and $\mathbf{c}$. Premultiplying (2.79) by $\mathbf{A}'\mathbf{V}$ yields the *generalized normal equation*

$$\mathbf{A}'\mathbf{V}\mathbf{A}\mathbf{b} = \mathbf{A}'\mathbf{V}\mathbf{y}\,, \tag{2.80}$$

i.e.,

$$\mathbf{A}'\mathbf{V}(\mathbf{y} - \mathbf{A}\mathbf{b}) = \mathbf{0}\,. \tag{2.81}$$

Hence, if $\mathbf{b}_*$ is a solution to the generalized normal equation (2.80), then

$$\mathbf{y} - \mathbf{A}\mathbf{b}_* \in \mathscr{C}(\mathbf{A})^{\perp}_{\mathbf{V}}\,, \tag{2.82}$$

and $\mathbf{A}\mathbf{b}_*$ is the orthogonal projection of $\mathbf{y}$ onto $\mathscr{C}(\mathbf{A})$ (with respect to the given inner product). Now the vector

$$\mathbf{b}_* = (\mathbf{A}'\mathbf{V}\mathbf{A})^{-}\mathbf{A}'\mathbf{V}\mathbf{y} + [\mathbf{I}_m - (\mathbf{A}'\mathbf{V}\mathbf{A})^{-}(\mathbf{A}'\mathbf{V}\mathbf{A})]\mathbf{t}\,, \tag{2.83}$$

where $\mathbf{t}$ is free to vary, is the general solution to the (consistent) equation (2.80). Premultiplying (2.83) by $\mathbf{A}$ gives

$$\mathbf{A}\mathbf{b}_* = \mathbf{A}(\mathbf{A}'\mathbf{V}\mathbf{A})^{-}\mathbf{A}'\mathbf{V}\mathbf{y}\,, \tag{2.84}$$

and thus we have obtained the expression (i) for the matrix $\mathbf{P}_*$. Note that

$$\mathbf{A}(\mathbf{A}'\mathbf{V}\mathbf{A})^{-}\mathbf{A}'\mathbf{V}\mathbf{A} = \mathbf{A} \tag{2.85}$$

when $\mathbf{V}$ is positive definite; see (6.13) (p. 146).

The proof of the equivalence of statements (f), (g) and (h) is straightforward and we omit it here.

Let us consider (c). Then, just as in (2.23),

$$\begin{aligned}\|\mathbf{y} - \mathbf{A}\mathbf{b}\|^2_{\mathbf{V}} &= \|\mathbf{y} - \mathbf{P}_*\mathbf{y} + (\mathbf{P}_*\mathbf{y} - \mathbf{A}\mathbf{b})\|^2_{\mathbf{V}} \\ &= \|\mathbf{y} - \mathbf{P}_*\mathbf{y}\|^2_{\mathbf{V}} + \|\mathbf{P}_*\mathbf{y} - \mathbf{A}\mathbf{b}\|^2_{\mathbf{V}} + 2\delta\,, \end{aligned}\tag{2.86}$$

where

$$\delta = \mathbf{y}'(\mathbf{I}_n - \mathbf{P}_*)'\mathbf{V}(\mathbf{P}_*\mathbf{y} - \mathbf{A}\mathbf{b})\,. \tag{2.87}$$

It is easy to confirm that $\delta = 0$ if any of the statements (f), (g) and (h) holds.

To prove that (c) implies (d), we can proceed just as in the proof of Theorem 2. Corresponding to (2.27) and Proposition 2.1 (p. 76), the equation (2.28) becomes

$$\mathbf{y}'(\mathbf{I}_n - \mathbf{P}_*)'\mathbf{V}\mathbf{A}\mathbf{b} = 0 \quad \text{for all } \mathbf{b},\, \mathbf{y}\,, \tag{2.88}$$

i.e.,

$$\mathbf{P}_*'\mathbf{V}\mathbf{A} = \mathbf{V}\mathbf{A}\,. \tag{2.89}$$

Hence we know that (c) is equivalent to conditions

$$\text{(i) } \mathscr{C}(\mathbf{P}_*) \subset \mathscr{C}(\mathbf{A})\,, \quad \text{(ii) } \mathbf{P}_*'\mathbf{V}\mathbf{A} = \mathbf{V}\mathbf{A}\,, \tag{d}$$

which can be equivalently expressed as

$$\mathbf{P}_*'(\mathbf{V}\mathbf{A} : \mathbf{A}^{\perp}) = (\mathbf{V}\mathbf{A} : \mathbf{0})\,. \tag{e}$$

From (i) in (d) above it follows that $\mathbf{P}_* = \mathbf{AL}$ for some matrix $\mathbf{L}$. Postmultiplying (ii) by $\mathbf{L}$ yields

$$\mathbf{P}_*'\mathbf{VP}_* = \mathbf{VP}_* \,, \tag{2.90}$$

and thereby $\mathbf{VP}_* = \mathbf{P}_*'\mathbf{V}$. Substituting this into (ii) gives

$$\mathbf{VP}_*\mathbf{A} = \mathbf{VA} \,. \tag{2.91}$$

Thus we have shown that conditions (i) and (ii) in (d) imply the following three conditions:

$$\text{(iii) } \mathscr{C}(\mathbf{P}_*) \subset \mathscr{C}(\mathbf{A}) \,, \quad \text{(iv) } \mathbf{P}_*'\mathbf{VP}_* = \mathbf{VP}_* \,, \quad \text{(v) } \mathbf{VP}_*\mathbf{A} = \mathbf{VA} \,. \tag{2.92}$$

It is obvious that the reverse relation "(2.92) $\implies$ (d)" holds as well and hence we can conclude that (c) and (2.92) are equivalent. Notice also that

$$\mathbf{P}_*'\mathbf{VP}_* = \mathbf{VP}_* \iff (\mathbf{VP}_*)' = \mathbf{VP}_* \text{ and } \mathbf{VP}_*^2 = \mathbf{VP}_* \,. \tag{2.93}$$

Here it is time to take a breath and notice that while proving that "(2.92) $\iff$ (d)" we have *not* assumed that the inner product matrix $\mathbf{V}$ is positive definite; it could be a singular nonnegative definite matrix. This fact will be utilized when dealing with the generalized orthogonal projector in the next section.

However, if $\mathbf{V}$ is positive definite, then (v) becomes $\mathbf{P}_*\mathbf{A} = \mathbf{A}$, and so $\mathscr{C}(\mathbf{A}) \subset \mathscr{C}(\mathbf{P}_*)$ which together with (iii) means that $\mathscr{C}(\mathbf{P}_*) = \mathscr{C}(\mathbf{A})$. Hence in the positive definite case, (2.92) can be replaced with

$$\text{(iii}*\text{) } \mathscr{C}(\mathbf{P}_*) = \mathscr{C}(\mathbf{A}) \,, \quad \text{(iv}*\text{) } \mathbf{P}_*'\mathbf{VP}_* = \mathbf{VP}_* \,, \quad \text{(v}*\text{) } \mathbf{P}_*\mathbf{A} = \mathbf{A} \,. \tag{2.94}$$

Now

$$\mathbf{P}_*'\mathbf{VP}_* = \mathbf{VP}_* \implies \mathbf{VP}_*^2 = \mathbf{VP}_* \implies \mathbf{P}_*^2 = \mathbf{P}_* \,. \tag{2.95}$$

The idempotency of $\mathbf{P}_*$ and $\mathscr{C}(\mathbf{P}_*) = \mathscr{C}(\mathbf{A})$ imply immediately that $\mathbf{P}_*\mathbf{A} = \mathbf{A}$ holds. Hence we can delete (v$*$) from (2.94) and thereby obtain the conditions appearing in (f)–(h). □

We will return several times to the important matrix

$$\mathbf{P}_* = \mathbf{A}(\mathbf{A}'\mathbf{VA})^-\mathbf{A}'\mathbf{V} \,, \tag{2.96}$$

where $\mathbf{V}$ is positive definite, but let us devote a minute to it right now. First, its invariance with respect to $(\mathbf{A}'\mathbf{VA})^-$ can be shown easily using Theorem 12 (p. 283) or by the rank cancellation rule, see (6.13) (p. 146). Moreover, Proposition 2.6 tells that $\mathbf{P}_*$ has the following properties (to some of which we will return to later on):

$$\mathbf{A}(\mathbf{A}'\mathbf{VA})^-\mathbf{A}'\mathbf{V} \text{ is idempotent} \,, \tag{2.97a}$$

$$\operatorname{rank}[\mathbf{A}(\mathbf{A}'\mathbf{VA})^-\mathbf{A}'\mathbf{V}] = \operatorname{rank}(\mathbf{A}) \,, \tag{2.97b}$$

$$\mathscr{C}[\mathbf{A}(\mathbf{A}'\mathbf{V}\mathbf{A})^{-}\mathbf{A}'\mathbf{V}] = \mathscr{C}(\mathbf{A})\,, \tag{2.97c}$$

$$\mathbf{A}(\mathbf{A}'\mathbf{V}\mathbf{A})^{-}\mathbf{A}'\mathbf{V}\mathbf{A} = \mathbf{A}\,. \tag{2.97d}$$

2.5 Generalized Orthogonal Projector

Let $\mathbf{V}$ be a singular nonnegative definite matrix. Then $\langle \mathbf{t}, \mathbf{u} \rangle_{\mathbf{V}} = \mathbf{t}'\mathbf{V}\mathbf{u}$ is a semi-inner product and the corresponding seminorm (squared) is $\|\mathbf{t}\|_{\mathbf{V}}^2 = \mathbf{t}'\mathbf{V}\mathbf{t}$. Now when $\mathbf{V}$ is singular, there are nonzero vectors $\mathbf{t}$ whose norm is $\|\mathbf{t}\|_{\mathbf{V}} = 0$. Then $\mathbf{t} \in \mathscr{N}(\mathbf{V}) = \mathscr{C}(\mathbf{V})^{\perp}$. Moreover, in view of (2.76) (p. 82), we have

$$\begin{aligned} \mathscr{C}(\mathbf{A})^{\perp}_{\mathbf{V}} &= \{\, \mathbf{y} \in \mathbb{R}^n : \mathbf{z}'\mathbf{A}'\mathbf{V}\mathbf{y} = 0 \text{ for all } \mathbf{z} \in \mathbb{R}^m \,\} \\ &= \mathscr{N}(\mathbf{A}'\mathbf{V}) = \mathscr{C}(\mathbf{V}\mathbf{A})^{\perp}. \end{aligned} \tag{2.98}$$

As discussed earlier, when $\mathbf{V}$ is positive definite, we have

$$\mathscr{C}(\mathbf{V}\mathbf{A})^{\perp} = \mathscr{C}(\mathbf{V}^{-1}\mathbf{A}^{\perp})\,, \quad \dim \mathscr{C}(\mathbf{V}\mathbf{A})^{\perp} = n - \operatorname{rank}(\mathbf{A})\,, \tag{2.99}$$

and

$$\mathbf{A}^{\perp}_{\mathbf{V}} = \mathbf{V}^{-1}\mathbf{A}^{\perp}, \quad \mathbb{R}^n = \mathscr{C}(\mathbf{A}) \oplus \mathscr{C}(\mathbf{A}^{\perp}_{\mathbf{V}})\,, \tag{2.100}$$

and for every $\mathbf{y} \in \mathbb{R}^n$ we have a unique decomposition

$$\mathbf{y} = \mathbf{y}_A + \mathbf{y}_{A^{\perp}} \quad \text{for some } \mathbf{y}_A \in \mathscr{C}(\mathbf{A}),\ \mathbf{y}_{A^{\perp}} \in \mathscr{C}(\mathbf{A}^{\perp}_{\mathbf{V}})\,. \tag{2.101}$$

For a singular nonnegative definite matrix $\mathbf{V}$ we can define the matrix $\mathbf{A}^{\perp}_{\mathbf{V}}$ again as any matrix spanning $\mathscr{C}(\mathbf{A})^{\perp}_{\mathbf{V}}$, and so, in view of (2.98),

$$\mathscr{C}(\mathbf{A}^{\perp}_{\mathbf{V}}) = \mathscr{C}(\mathbf{A})^{\perp}_{\mathbf{V}} = \mathscr{N}(\mathbf{A}'\mathbf{V}) = \mathscr{C}(\mathbf{V}\mathbf{A})^{\perp}. \tag{2.102}$$

Let $\mathbf{b}_*$ be a solution to the generalized normal equation

$$\mathbf{A}'\mathbf{V}\mathbf{A}\mathbf{b} = \mathbf{A}'\mathbf{V}\mathbf{y}\,, \quad \text{i.e.,} \quad \mathbf{A}'\mathbf{V}(\mathbf{y} - \mathbf{A}\mathbf{b}) = \mathbf{0}\,, \tag{2.103}$$

in which case $\mathbf{y} - \mathbf{A}\mathbf{b}_* \in \mathscr{C}(\mathbf{A}^{\perp}_{\mathbf{V}})$. Then

$$\mathbf{y} = \mathbf{A}\mathbf{b}_* + (\mathbf{y} - \mathbf{A}\mathbf{b}_*)\,, \quad \text{where } \mathbf{A}\mathbf{b}_* \in \mathscr{C}(\mathbf{A}),\ \mathbf{y} - \mathbf{A}\mathbf{b}_* \in \mathscr{C}(\mathbf{A}^{\perp}_{\mathbf{V}})\,, \tag{2.104}$$

which means that for (even) a singular $\mathbf{V}$ we do have the decomposition

$$\mathbb{R}^n = \mathscr{C}(\mathbf{A}) + \mathscr{C}(\mathbf{A}^{\perp}_{\mathbf{V}}) = \mathscr{C}(\mathbf{A}) + \mathscr{C}(\mathbf{V}\mathbf{A})^{\perp}. \tag{2.105}$$

However, the above decomposition is not necessarily a direct sum, as is easily demonstrated by considering the matrix $\mathbf{L}$ satisfying

$$\mathscr{C}(\mathbf{L}) = \mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{V})^{\perp} = \mathscr{C}(\mathbf{A}) \cap \mathscr{N}(\mathbf{V}) . \tag{2.106}$$

Then $\mathbf{A}'\mathbf{V}\mathbf{L}\mathbf{u} = \mathbf{0}$ for any vector $\mathbf{u}$, and so $\mathbf{L}\mathbf{u} \in \mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{A}^{\perp}_{\mathbf{V}})$. Note that the vector $\mathbf{L}\mathbf{u}$ has now zero length: $\|\mathbf{L}\mathbf{u}\|^2_{\mathbf{V}} = \mathbf{u}'\mathbf{L}'\mathbf{V}\mathbf{L}\mathbf{u} = 0$.

For any nonnegative definite $\mathbf{V}$ we have, on account of Theorem 5 (p. 121),

$$\begin{aligned} \dim \mathscr{C}(\mathbf{V}\mathbf{A})^{\perp} &= n - \operatorname{rank}(\mathbf{V}\mathbf{A}) \\ &= [n - \operatorname{rank}(\mathbf{A})] + \dim \mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{V})^{\perp}, \end{aligned} \tag{2.107}$$

which means that (2.105) becomes a direct sum decomposition if and only if

$$\mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{V})^{\perp} = \{\mathbf{0}\} . \tag{2.108}$$

As mentioned by Rao & Rao (1998, p. 81), the nonuniqueness of (2.105) makes it difficult to define the projection of $\mathbf{y}$ onto $\mathscr{C}(\mathbf{A})$. However, according to Rao & Rao (1998, p. 81): “We need not worry about it. We can get around this difficulty.”

Let us consider the minimization of $\|\mathbf{y} - \mathbf{A}\mathbf{b}\|^2_{\mathbf{V}}$. If $\mathbf{b}_*$ is a solution to the generalized normal equation (2.103), we have

$$\begin{aligned} \|\mathbf{y} - \mathbf{A}\mathbf{b}\|^2_{\mathbf{V}} &= \|\mathbf{y} - \mathbf{A}\mathbf{b}_* + (\mathbf{A}\mathbf{b}_* - \mathbf{A}\mathbf{b})\|^2_{\mathbf{V}} \\ &= \|\mathbf{y} - \mathbf{A}\mathbf{b}_*\|^2_{\mathbf{V}} + \|\mathbf{A}(\mathbf{b}_* - \mathbf{b})\|^2_{\mathbf{V}} \\ &\geq \|\mathbf{y} - \mathbf{A}\mathbf{b}_*\|^2_{\mathbf{V}} , \end{aligned} \tag{2.109}$$

and so the minimum of $\|\mathbf{y} - \mathbf{A}\mathbf{b}\|^2_{\mathbf{V}}$ is attained when $\mathbf{b}_*$ is a solution to (2.103). However, in this situation the vector $\mathbf{A}\mathbf{b}_*$ may not be unique. The general solution to (2.103) can be expressed as

$$\mathbf{b}_* = (\mathbf{A}'\mathbf{V}\mathbf{A})^{-}\mathbf{A}'\mathbf{V}\mathbf{y} + [\mathbf{I}_m - (\mathbf{A}'\mathbf{V}\mathbf{A})^{-}\mathbf{A}'\mathbf{V}\mathbf{A}]\mathbf{u} , \tag{2.110}$$

where $\mathbf{u}$ is free to vary. Premultiplying (2.110) by $\mathbf{A}$ yields

$$\mathbf{A}\mathbf{b}_* = \mathbf{A}(\mathbf{A}'\mathbf{V}\mathbf{A})^{-}\mathbf{A}'\mathbf{V}\mathbf{y} + \mathbf{A}[\mathbf{I}_m - (\mathbf{A}'\mathbf{V}\mathbf{A})^{-}\mathbf{A}'\mathbf{V}\mathbf{A}]\mathbf{u} . \tag{2.111}$$

On account of Theorem 12 (p. 283) the last term above vanishes for all $\mathbf{u}$, i.e.,

$$\mathbf{A}(\mathbf{A}'\mathbf{V}\mathbf{A})^{-}\mathbf{A}'\mathbf{V}\mathbf{A} = \mathbf{A} , \tag{2.112}$$

if and only if $\mathscr{C}(\mathbf{A}') \subset \mathscr{C}(\mathbf{A}'\mathbf{V}\mathbf{A}) = \mathscr{C}(\mathbf{A}'\mathbf{V})$, i.e., if and only if (2.108) holds:

$$\mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{V})^{\perp} = \{\mathbf{0}\} ; \tag{2.113}$$

see also Proposition 5.4 (p. 132). Notice that if

$$\mathbf{t} = \mathbf{A}[\mathbf{I}_m - (\mathbf{A}'\mathbf{V}\mathbf{A})^{-}\mathbf{A}'\mathbf{V}\mathbf{A}]\mathbf{u} \in \mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{V})^{\perp}, \tag{2.114}$$

then $\|\mathbf{t}\|_{\mathbf{V}} = 0$. Actually, see (5.118) (p. 138),

$$\mathscr{C}\big[\mathbf{A}[\mathbf{I}_m - (\mathbf{A}'\mathbf{V}\mathbf{A})^{-}\mathbf{A}'\mathbf{V}\mathbf{A}]\big] = \mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{V})^{\perp}. \tag{2.115}$$

Now, in view of the proof of Proposition 2.4, in particular in light of (2.92) (p. 84), we get the following characterization for the generalized orthogonal projector; see Mitra & Rao (1974). Some related considerations appear also in Harville (1997, §14.12.i).

Proposition 2.7. *Let* $\mathbf{A}$ *be an* $n \times m$ *matrix,* $\operatorname{rank}(\mathbf{A}) = r > 0$*, and let* $\mathbf{V}$ *be a symmetric nonnegative definite matrix (possibly singular), and let the semi-inner product (and the corresponding seminorm) be defined as* $\langle \mathbf{t}, \mathbf{u} \rangle_{\mathbf{V}} = \mathbf{t}'\mathbf{V}\mathbf{u}$*. Then the following conditions are equivalent ways to define the (possibly nonunique) matrix* $\mathbf{P}$:

(a) (i) $\mathscr{C}(\mathbf{P}) \subset \mathscr{C}(\mathbf{A})$, (ii) $\|\mathbf{y} - \mathbf{A}\mathbf{b}\|_{\mathbf{V}}^2 \geq \|\mathbf{y} - \mathbf{P}\mathbf{y}\|_{\mathbf{V}}^2$ *for all* $\mathbf{b}$, $\mathbf{y}$.

(b) (i) $\mathscr{C}(\mathbf{P}) \subset \mathscr{C}(\mathbf{A})$, (ii) $\mathbf{P}'\mathbf{V}\mathbf{A} = \mathbf{V}\mathbf{A}$.

(c) $\mathbf{P}'(\mathbf{V}\mathbf{A} : \mathbf{A}^{\perp}) = (\mathbf{V}\mathbf{A} : \mathbf{0})$.

(d) (i) $\mathscr{C}(\mathbf{P}) \subset \mathscr{C}(\mathbf{A})$, (ii) $\mathbf{P}'\mathbf{V}\mathbf{P} = \mathbf{V}\mathbf{P}$, (iii) $\mathbf{V}\mathbf{P}\mathbf{A} = \mathbf{V}\mathbf{A}$.

(e) (i) $\mathscr{C}(\mathbf{P}) \subset \mathscr{C}(\mathbf{A})$, (ii) $(\mathbf{V}\mathbf{P})' = \mathbf{V}\mathbf{P}$, (iii) $\mathbf{V}\mathbf{P}\mathbf{A} = \mathbf{V}\mathbf{A}$.

(f) (i) $\mathscr{C}(\mathbf{P}) \subset \mathscr{C}(\mathbf{A})$, (ii) $\mathbf{P}'\mathbf{V}\mathbf{P} = \mathbf{V}\mathbf{P}$, (iii) $\operatorname{rank}(\mathbf{V}\mathbf{P}) = \operatorname{rank}(\mathbf{V}\mathbf{A})$.

(g) $\mathbf{P} = \mathbf{A}(\mathbf{A}'\mathbf{V}\mathbf{A})^{-}\mathbf{A}'\mathbf{V} + \mathbf{A}[\mathbf{I}_m - (\mathbf{A}'\mathbf{V}\mathbf{A})^{-}\mathbf{A}'\mathbf{V}\mathbf{A}]\mathbf{U}$, *where* $\mathbf{U}$ *is arbitrary.*

(h) $\mathbf{P} = \mathbf{A}(\mathbf{A}'\mathbf{V}\mathbf{A})^{-}\mathbf{A}'\mathbf{V} + \mathbf{A}(\mathbf{I}_m - \mathbf{P}_{\mathbf{A}'\mathbf{V}})\mathbf{U}$, *where* $\mathbf{U}$ *is arbitrary.*

The matrix $\mathbf{P} = \mathbf{P}_{\mathbf{A};\mathbf{V}}$ *is the generalized orthogonal projector onto the column space* $\mathscr{C}(\mathbf{A})$ *with respect to the semi-inner product* $\langle \mathbf{t}, \mathbf{u} \rangle_{\mathbf{V}} = \mathbf{t}'\mathbf{V}\mathbf{u}$*. The matrix* $\mathbf{P}_{\mathbf{A};\mathbf{V}}$ *is unique if and only if*

$$\mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{V})^{\perp} = \{\mathbf{0}\}, \quad \textit{i.e.,} \quad \operatorname{rank}(\mathbf{A}) = \operatorname{rank}(\mathbf{V}\mathbf{A}). \tag{2.116}$$

For related considerations, see also Rao (1974), Tian (2007, 2009a, 2010), Tian & Tian (2010), Tian & Wiens (2006), Tian & Takane (2008b,a, 2009b).

We will complete this section by a quick comment regarding the projector denoted as $\mathbf{P}_{\mathbf{A}|\mathbf{B}}$; see Rao (1974). As was mentioned on page 6 of the Introduction, $\mathbf{P}_{\mathbf{A}|\mathbf{B}}$ is a notation for a matrix satisfying

$$\mathbf{P}_{\mathbf{A}|\mathbf{B}}(\mathbf{A} : \mathbf{B}) = (\mathbf{A} : \mathbf{0}), \tag{2.117}$$

where it is assumed that $\mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B}) = \{\mathbf{0}\}$, but $\mathscr{C}(\mathbf{A}) \oplus \mathscr{C}(\mathbf{B})$ does not necessarily coincide with the entire $\mathbb{R}^n$. Then (2.117) defines the generalized projector $\mathbf{P}_{\mathbf{A}|\mathbf{B}}$ onto $\mathscr{C}(\mathbf{A})$ along

Photograph 2.1 Yongge Tian (New Delhi, 1992).

$\mathscr{C}(\mathbf{B})$. However, in this situation the projector $\mathbf{P}_{\mathbf{A}|\mathbf{B}}$ need not be unique and idempotent as is the case when $\mathscr{C}(\mathbf{A}) \oplus \mathscr{C}(\mathbf{B}) = \mathbb{R}^n$.

There is one crucial difference between the singular and nonsingular $\mathbf{V}$: for a positive definite $\mathbf{V}$ we can characterize the orthogonal projector $\mathbf{P}_{\mathbf{A};\mathbf{V}}$ by equation

$$\mathbf{P}_{\mathbf{A};\mathbf{V}}(\mathbf{A} : \mathbf{V}^{-1}\mathbf{A}^{\perp}) = (\mathbf{A} : \mathbf{0})\,, \tag{2.118}$$

but for a singular $\mathbf{V}$ equation (2.118) of course does not work. In view of Proposition 2.7 we can define the generalized orthogonal projector $\mathbf{P}_{\mathbf{A};\mathbf{V}}$ by equation

$$\mathbf{P}'_{\mathbf{A};\mathbf{V}}(\mathbf{V}\mathbf{A} : \mathbf{A}^{\perp}) = (\mathbf{V}\mathbf{A} : \mathbf{0})\,. \tag{2.119}$$

Hence we have

$$\{\mathbf{P}'_{\mathbf{A};\mathbf{V}}\} = \{\mathbf{P}_{\mathbf{V}\mathbf{A}|\mathbf{A}^{\perp}}\}\,. \tag{2.120}$$

It is left to the reader to confirm the following identities; see Rao (1974):

$$(\mathbf{P}_{\mathbf{V}\mathbf{A}|\mathbf{A}^{\perp}} + \mathbf{P}_{\mathbf{A}^{\perp}|\mathbf{V}\mathbf{A}})\mathbf{z} = \mathbf{z}\,, \tag{2.121a}$$

$$(\mathbf{P}_{\mathbf{V}\mathbf{A}^{\perp}|\mathbf{A}} + \mathbf{P}_{\mathbf{A}|\mathbf{V}\mathbf{A}^{\perp}})\mathbf{y} = \mathbf{y}\,, \tag{2.121b}$$

$$\mathbf{P}_{\mathbf{A}|\mathbf{V}\mathbf{A}^{\perp}}\mathbf{y} = (\mathbf{I}_n - \mathbf{P}'_{\mathbf{A}^{\perp};\mathbf{V}})\mathbf{y}\,, \tag{2.121c}$$

for all $\mathbf{z} \in \mathscr{C}(\mathbf{A}^{\perp} : \mathbf{V}\mathbf{A}) = \mathscr{C}(\mathbf{A}^{\perp} : \mathbf{V})$ and $\mathbf{y} \in \mathscr{C}(\mathbf{A} : \mathbf{V}\mathbf{A}^{\perp}) = \mathscr{C}(\mathbf{A} : \mathbf{V})$.

Moreover, let

$$\mathbf{W} = \mathbf{V} + \mathbf{X}\mathbf{L}\mathbf{L}'\mathbf{X}', \quad \text{where } \mathscr{C}(\mathbf{W}) = \mathscr{C}(\mathbf{X} : \mathbf{V})\,, \tag{2.122}$$

and denote $\mathbf{P}_{\mathbf{X};\mathbf{W}^-} = \mathbf{X}(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}'\mathbf{W}^-$, where $\mathbf{W}^-$ is an arbitrary nonnegative definite generalized inverse of $\mathbf{W}$. Then it appears, see, e.g., Proposition 10.5 (p. 228), that

$$\mathbf{P}_{\mathbf{X}|\mathbf{V}\mathbf{X}^{\perp}}\mathbf{y} = \mathbf{P}_{\mathbf{X};\mathbf{W}^-}\mathbf{y} \quad \text{for all } \mathbf{y} \in \mathscr{C}(\mathbf{X} : \mathbf{V})\,. \tag{2.123}$$

The matrices $\mathbf{P}_{\mathbf{X}|\mathbf{V}\mathbf{X}^{\perp}}$, $\mathbf{I}_n - \mathbf{P}'_{\mathbf{X}^{\perp};\mathbf{V}}$, and $\mathbf{P}_{\mathbf{X};\mathbf{W}^-}$ obviously yield the BLUE of $\mathbf{X}\boldsymbol{\beta}$ under $\{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$ and hence they are of great importance in linear models.

2.6 Exercises

2.1. Prove the equivalence of statements (f), (g) and (h) of Proposition 2.6 (p. 81).

2.2. Prove the statements in (2.121) (p. 88).

2.3. [Equality of two projectors under different inner products] Suppose that $\mathbf{V}_1$ and $\mathbf{V}_2$ are positive definite $n \times n$ matrices and $\operatorname{rank}(\mathbf{X}_{n\times p}) = r$. Prove the equivalence of the following statements:

(a) $\mathbf{P}_{\mathbf{X};\mathbf{V}_1^{-1}} = \mathbf{P}_{\mathbf{X};\mathbf{V}_2^{-1}}\,,$ (b) $\mathbf{X}'\mathbf{V}_2^{-1}\mathbf{P}_{\mathbf{X};\mathbf{V}_1^{-1}} = \mathbf{X}'\mathbf{V}_2^{-1}\,,$

(c) $\mathbf{P}'_{\mathbf{X};\mathbf{V}_1^{-1}}\mathbf{V}_2^{-1}\mathbf{P}_{\mathbf{X};\mathbf{V}_1^{-1}} = \mathbf{V}_2^{-1}\mathbf{P}_{\mathbf{X};\mathbf{V}_1^{-1}}\,,$ (d) $\mathbf{V}_2^{-1}\mathbf{P}_{\mathbf{X};\mathbf{V}_1^{-1}}$ is symmetric,

(e) $\mathscr{C}(\mathbf{V}_1^{-1}\mathbf{X}) = \mathscr{C}(\mathbf{V}_2^{-1}\mathbf{X})\,,$ (f) $\mathscr{C}(\mathbf{V}_1\mathbf{X}^{\perp}) = \mathscr{C}(\mathbf{V}_2\mathbf{X}^{\perp})\,,$

(g) $\mathscr{C}(\mathbf{V}_2\mathbf{V}_1^{-1}\mathbf{X}) = \mathscr{C}(\mathbf{X})\,,$ (h) $\mathbf{X}'\mathbf{V}_1^{-1}\mathbf{V}_2\mathbf{M} = \mathbf{0}\,,$

(i) $\mathscr{C}(\mathbf{V}_1^{-1/2}\mathbf{V}_2\mathbf{V}_1^{-1/2}\cdot\mathbf{V}_1^{-1/2}\mathbf{X}) = \mathscr{C}(\mathbf{V}_1^{-1/2}\mathbf{X})\,,$

(j) $\mathscr{C}(\mathbf{V}_1^{-1/2}\mathbf{X})$ has a basis $\mathbf{U} = (\mathbf{u}_1 : \ldots : \mathbf{u}_r)$ comprising a set of r eigenvectors of $\mathbf{V}_1^{-1/2}\mathbf{V}_2\mathbf{V}_1^{-1/2}$,

(k) $\mathbf{V}_1^{-1/2}\mathbf{X} = \mathbf{U}\mathbf{A}$ for some $\mathbf{A}_{r\times p}$, $\operatorname{rank}(\mathbf{A}) = r$,

(l) $\mathbf{X} = \mathbf{V}_1^{1/2}\mathbf{U}\mathbf{A}$; the columns of $\mathbf{V}_1^{1/2}\mathbf{U}$ are r eigenvectors of $\mathbf{V}_2\mathbf{V}_1^{-1}$,

(m) $\mathscr{C}(\mathbf{X})$ has a basis comprising a set of r eigenvectors of $\mathbf{V}_2\mathbf{V}_1^{-1}$.

Some statements above can be conveniently proved using Proposition 10.1 (p. 218). See also Section 11.1 (p. 269), Exercise 11.9 (p. 281), and (18.81) (p. 368).

Thomas (1968), Harville (1997, p. 265), Tian & Takane (2008b, 2009b), Hauke, Markiewicz & Puntanen (2011).

2.4 (Continued ...). Show that $\mathscr{C}(\mathbf{V}_1\mathbf{X}) = \mathscr{C}(\mathbf{V}_2\mathbf{X}) \not\Rightarrow \mathscr{C}(\mathbf{V}_1\mathbf{X}^{\perp}) = \mathscr{C}(\mathbf{V}_2\mathbf{X}^{\perp})$, but the following statements are equivalent:

(a) $\mathscr{C}(\mathbf{V}_1\mathbf{X}) = \mathscr{C}(\mathbf{X})\,,$ (b) $\mathscr{C}(\mathbf{V}_1\mathbf{X}^{\perp}) = \mathscr{C}(\mathbf{X}^{\perp})\,,$

(c) $\mathscr{C}(\mathbf{V}_1^{-1}\mathbf{X}) = \mathscr{C}(\mathbf{X})\,,$ (d) $\mathscr{C}(\mathbf{V}_1^{-1}\mathbf{X}^{\perp}) = \mathscr{C}(\mathbf{X}^{\perp})\,.$

2.5. Suppose that $\mathbf{V}$ is positive definite $n\times n$ matrix and $\mathbf{X} \in \mathbb{R}^{n\times p}$. Confirm:

$$\begin{aligned}
\mathbf{P}_{\mathbf{X};\mathbf{V}^{-1}} &= \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1} \\
&= \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1/2}\mathbf{V}^{-1/2}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1/2}\mathbf{V}^{-1/2} \\
&= \mathbf{V}^{1/2}\mathbf{P}_{\mathbf{V}^{-1/2}\mathbf{X}}\mathbf{V}^{-1/2} \\
&= \mathbf{V}^{1/2}(\mathbf{I}_n - \mathbf{P}_{(\mathbf{V}^{-1/2}\mathbf{X})^{\perp}})\mathbf{V}^{-1/2} \\
&= \mathbf{V}^{1/2}(\mathbf{I}_n - \mathbf{P}_{\mathbf{V}^{1/2}\mathbf{M}})\mathbf{V}^{-1/2} \\
&= \mathbf{I}_n - \mathbf{V}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^{-}\mathbf{M} = \mathbf{I}_n - \mathbf{P}'_{\mathbf{M};\mathbf{V}}\,.
\end{aligned}$$

See also (2.74) (p. 82), Proposition 5.9 (p. 140), and part (i) of Theorem 15 (p. 318).

2.6. Suppose that $\mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B}) = \{\mathbf{0}_n\} = \mathscr{C}(\mathbf{C}) \cap \mathscr{C}(\mathbf{D})$. Show that then

$$\{\,\mathbf{P}_{\mathbf{C}|\mathbf{D}}\,\} \subset \{\,\mathbf{P}_{\mathbf{A}|\mathbf{B}}\,\} \quad\iff\quad \mathscr{C}(\mathbf{A}) \subset \mathscr{C}(\mathbf{C}) \;\text{ and }\; \mathscr{C}(\mathbf{B}) \subset \mathscr{C}(\mathbf{D})\,.$$

Kala (1981, Lemma 2.5).

Philatelic Item 2.1 Prasanta Chandra Mahalanobis (1893–1972) was an Indian scientist and applied statistician. In 1936 he introduced what is now called the Mahalanobis distance (see the formula in red in the First Day Cover above and pages 24–25). In 1931 Mahalanobis founded the Indian Statistical Institute in Kolkata (shown on the stamp, Kolkata was then known as Calcutta), and contributed to the design of large-scale sample surveys. The second five-year plan (1956–1961) focused on industry, especially heavy industry; domestic production of industrial products was encouraged, particularly in the development of the public sector. The plan followed the Mahalanobis model, an economic development model developed by Mahalanobis in 1953. The stamp (*Scott* 1449) was issued by India in 1993, the 100th anniversary of the year in which Mahalanobis was born.

Chapter 3
Easy Correlation Tricks

> *In conclusion I must repeat what was said before, that it is impossible to go deeper into this subject without using very technical language and dealing freely with conceptions that are, unhappily, quite unfamiliar to the large majority of educated men. I can only say that there is a vast field of topics that fall under the laws of correlation, which lies quite open to the research of any competent person who cares to investigate it.*
>
> Francis Galton (1890): *Kinship and Correlation*

In this chapter we remind the reader about one fundamental fact: the geometric interpretation of the sample correlation coefficient. It's hardly news for the reader but because of its importance it is worth its own chapter.

Let the $n \times 2$ data matrix $\mathbf{U}$ be partitioned as

$$\mathbf{U} = (\mathbf{x} : \mathbf{y}) = \begin{pmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_n & y_n \end{pmatrix} = \begin{pmatrix} \mathbf{u}'_{(1)} \\ \mathbf{u}'_{(2)} \\ \vdots \\ \mathbf{u}'_{(n)} \end{pmatrix}. \tag{3.1}$$

Here $\mathbf{u}_{(i)} = \binom{x_i}{y_i} \in \mathbb{R}^2$ represents the ith case or the ith observation in the observation space, and the vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ represent the two variables in the variable space. Let $\bar{\mathbf{u}} = \binom{\bar{x}}{\bar{y}} \in \mathbb{R}^2$ denote the mean vector of x- and y-variables and $\mathbf{S}$ the sample covariance matrix:

$$\bar{\mathbf{u}} = \tfrac{1}{n}(\mathbf{u}_{(1)} + \mathbf{u}_{(2)} + \cdots + \mathbf{u}_{(n)}) = \tfrac{1}{n}\mathbf{U}'\mathbf{1}_n = \begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix}, \tag{3.2a}$$

$$\mathbf{S} = \tfrac{1}{n-1}\mathbf{U}'\mathbf{C}\mathbf{U} = \tfrac{1}{n-1}\mathbf{T} = \tfrac{1}{n-1}\sum_{i=1}^{n}(\mathbf{u}_{(i)} - \bar{\mathbf{u}})(\mathbf{u}_{(i)} - \bar{\mathbf{u}})', \tag{3.2b}$$

where $\mathbf{C}$ is the $n \times n$ centering matrix: $\mathbf{C} = \mathbf{I}_n - \mathbf{J} = \mathbf{I}_n - \mathbf{P}_{\mathbf{1}}$. In what follows, we most of the time drop off the subscript from the vector $\mathbf{1}_n$; from the context its dimension should be obvious.

Theorem 3 (Correlation coefficient geometrically). *For conformable matrices, the following statements hold:*

(a) *The vector* $\bar{\bar{\mathbf{y}}} = \bar{y}\mathbf{1}$ *is the orthogonal projection of the variable vector* $\mathbf{y}$ *onto the column space* $\mathscr{C}(\mathbf{1})$*:*

$$\bar{\bar{\mathbf{y}}} = \bar{y}\mathbf{1} = \mathbf{J}\mathbf{y} = \mathbf{P}_{\mathbf{1}}\mathbf{y}\,. \tag{3.3}$$

S. Puntanen et al., *Matrix Tricks for Linear Statistical Models: Our Personal Top Twenty*,
DOI 10.1007/978-3-642-10473-2_4, © Springer-Verlag Berlin Heidelberg 2011

(b) *The centered variable vector* $\tilde{\mathbf{y}}$ *is the orthogonal projection of* $\mathbf{y}$ *onto the column space* $\mathscr{C}(\mathbf{1})^{\perp}$:

$$\tilde{\mathbf{y}} = \mathbf{y} - \mathbf{J}\mathbf{y} = \mathbf{C}\mathbf{y} = (\mathbf{I} - \mathbf{P}_{\mathbf{1}})\mathbf{y} = \mathbf{P}_{\mathbf{1}^{\perp}}\mathbf{y} = \mathbf{P}_{\mathbf{C}}\mathbf{y}\,. \tag{3.4}$$

(c) *Let the variances of the variables* x *and* y *be nonzero, i.e.,* $\mathbf{x} \notin \mathscr{C}(\mathbf{1})$ *and* $\mathbf{y} \notin \mathscr{C}(\mathbf{1})$*. Then the sample correlation coefficient* r_{xy} *is the cosine of the angle between the centered variable vectors:*

$$\begin{aligned} r_{xy} &= \operatorname{cor_s}(x,y) = \operatorname{cor_d}(\mathbf{x},\mathbf{y}) \\ &= \cos(\mathbf{C}\mathbf{x},\mathbf{C}\mathbf{y}) = \cos(\tilde{\mathbf{x}},\tilde{\mathbf{y}}) \\ &= \frac{\mathbf{x}'\mathbf{C}\mathbf{y}}{\sqrt{\mathbf{x}'\mathbf{C}\mathbf{x}\cdot\mathbf{y}'\mathbf{C}\mathbf{y}}}\,. \end{aligned} \tag{3.5}$$

(d) $\mathbf{y}$ *is centered* $\iff \mathbf{y} \in \mathscr{C}(\mathbf{1})^{\perp} = \mathscr{C}(\mathbf{C}) = \mathscr{N}(\mathbf{1}')$.

Proof. All proofs are straightforward consequences from the definitions of the projectors $\mathbf{J}$ and $\mathbf{C}$. See also Figure 3.1 (p. 93). We might, for completeness, consider the observed values $y_1, y_2, \dots, y_n$ of y and the task to find $\alpha \in \mathbb{R}$ such that it minimizes

$$f(\alpha) = \sum_{i=1}^{n}(y_i - \alpha)^2. \tag{3.6}$$

The solution comes from

$$\begin{aligned} \min_{\alpha} f(\alpha) &= \min_{\alpha}\sum_{i=1}^{n}(y_i-\alpha)^2 = \min_{\alpha}\|\mathbf{y}-\alpha\mathbf{1}\|^2 = \min_{\mathbf{u}\in\mathscr{C}(\mathbf{1})}\|\mathbf{y}-\mathbf{u}\|^2 \\ &= \|\mathbf{y}-\mathbf{P}_{\mathbf{1}}\mathbf{y}\|^2 = \|\mathbf{y}-\mathbf{J}\mathbf{y}\|^2 = \|\mathbf{y}-\bar{y}\mathbf{1}\|^2 = \sum_{i=1}^{n}(y_i-\bar{y})^2. \end{aligned} \tag{3.7}$$

□

3.1 Orthogonality and Uncorrelatedness

It is essential to recall that orthogonality and uncorrelatedness are not necessarily identical concepts. The following proposition summarizes some related features.

Proposition 3.1. *Let* $\mathbf{x}$ *and* $\mathbf{y}$ *be nonzero* n*-dimensional variable vectors. Then it is possible that*

(a) $\cos(\mathbf{x},\mathbf{y})$ *is high, but* $\operatorname{cor_d}(\mathbf{x},\mathbf{y}) = 0$,

(b) $\cos(\mathbf{x},\mathbf{y}) = 0$, *but* $\operatorname{cor_d}(\mathbf{x},\mathbf{y}) = 1$.

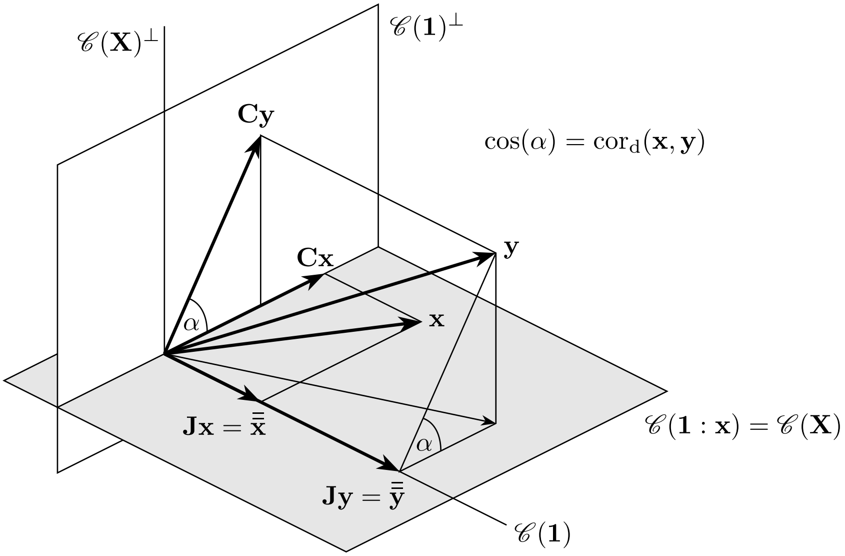


Figure 3.1 Correlation between $\mathbf{x}$ and $\mathbf{y}$ geometrically.

Moreover, let $\mathbf{x} \notin \mathscr{C}(\mathbf{1})$ *and* $\mathbf{y} \notin \mathscr{C}(\mathbf{1})$. *Then*

$$\operatorname{cor_d}(\mathbf{x}, \mathbf{y}) = 0 \iff \mathbf{y} \in \mathscr{C}(\mathbf{Cx})^\perp = \mathscr{C}(\mathbf{1} : \mathbf{x})^\perp \boxplus \mathscr{C}(\mathbf{1})\,. \tag{3.8}$$

Proof. For the situation (a), Belsley (1991, p. 20) has an illustrative example where correlation is zero but the cosine is very close to 1. For this purpose, consider the data matrix

$$(\mathbf{x} : \mathbf{y}) = \begin{pmatrix} 1 & 1 \\ -1 & 1 \\ 0 & -2 \end{pmatrix}. \tag{3.9}$$

Here $\mathbf{x}, \mathbf{y} \in \mathscr{C}(\mathbf{C}) = \mathscr{N}(\mathbf{1}')$ and $\mathbf{x}'\mathbf{y} = 0$, i.e., $\mathbf{x}$ and $\mathbf{y}$ are centered and uncorrelated: $\operatorname{cor_d}(\mathbf{x}, \mathbf{y}) = 0$. Define $\mathbf{t}$ and $\mathbf{u}$ so that

$$\mathbf{t} = \mathbf{x} + \alpha\mathbf{1}\,, \quad \mathbf{u} = \mathbf{y} + \alpha\mathbf{1}\,, \tag{3.10}$$

where α is a nonzero real number. Then $\operatorname{cor_d}(\mathbf{t}, \mathbf{u}) = 0$ but

$$\cos(\mathbf{t}, \mathbf{u}) = \frac{3\alpha^2}{\sqrt{(\mathbf{x}'\mathbf{x} + 3\alpha^2)(\mathbf{y}'\mathbf{y} + 3\alpha^2)}} = \frac{3}{\sqrt{\left(\frac{\mathbf{x}'\mathbf{x}}{\alpha^2} + 3\right)\left(\frac{\mathbf{y}'\mathbf{y}}{\alpha^2} + 3\right)}}\,, \tag{3.11}$$

which can be made arbitrarily close to 1 by choosing α very large. We could also consider variables

$$\mathbf{v} = \alpha\mathbf{x} + \mathbf{1}\,, \quad \mathbf{w} = \alpha\mathbf{y} + \mathbf{1}\,, \quad \text{where } \alpha \neq 0\,, \tag{3.12}$$

in which case $\operatorname{cor}_{\mathrm{d}}(\mathbf{v}, \mathbf{w}) = 0$ but

$$\cos(\mathbf{v}, \mathbf{w}) = \frac{3}{\sqrt{(\mathbf{x}'\mathbf{x}\alpha^2 + 3)(\mathbf{y}'\mathbf{y}\alpha^2 + 3)}}\,. \tag{3.13}$$

Now $\cos(\mathbf{v}, \mathbf{w})$ approaches 1 when α tends to 0.

For (b) the following data gives an example:

$$(\mathbf{x} : \mathbf{y}) = \begin{pmatrix} 0 & -1 \\ 0 & -1 \\ 1 & 0 \end{pmatrix}. \tag{3.14}$$

Consider then (3.8). It is obvious that

$$\mathbf{x}'\mathbf{C}\mathbf{y} = 0 \iff \mathbf{y} \in \mathscr{N}(\mathbf{x}'\mathbf{C}) = \mathscr{C}(\mathbf{C}\mathbf{x})^{\perp}. \tag{3.15}$$

We next prove that

$$\mathscr{C}(\mathbf{C}\mathbf{x}) = \mathscr{C}(\mathbf{1} : \mathbf{x}) \cap \mathscr{C}(\mathbf{1})^{\perp} = \mathscr{C}(\mathbf{1} : \mathbf{x}) \cap \mathscr{C}(\mathbf{C})\,. \tag{3.16}$$

Now

$$\mathbf{C}\mathbf{x} = \mathbf{x} - \bar{x}\mathbf{1} = (\mathbf{1} : \mathbf{x}) \begin{pmatrix} -\bar{x} \\ 1 \end{pmatrix}, \tag{3.17a}$$

$$(\mathbf{1} : \mathbf{C}\mathbf{x}) = (\mathbf{1} : \mathbf{x}) \begin{pmatrix} 1 & -\bar{x} \\ 0 & 1 \end{pmatrix} := (\mathbf{1} : \mathbf{x})\mathbf{A}\,, \tag{3.17b}$$

where $\mathbf{A} = \left(\begin{smallmatrix} 1 & -\bar{x} \\ 0 & 1 \end{smallmatrix}\right)$ is nonsingular and hence we get (a very useful fact to be used also later on)

$$\mathscr{C}(\mathbf{1} : \mathbf{x}) = \mathscr{C}(\mathbf{1} : \mathbf{C}\mathbf{x})\,. \tag{3.18}$$

Now it is clear that $\mathbf{C}\mathbf{x} \in \mathscr{C}(\mathbf{1} : \mathbf{C}\mathbf{x}) \cap \mathscr{C}(\mathbf{C})$. To show the reverse inclusion, consider

$$\mathbf{u} \in \mathscr{C}(\mathbf{1} : \mathbf{C}\mathbf{x}) \cap \mathscr{C}(\mathbf{C})\,, \tag{3.19}$$

which implies that

$$\mathbf{u} = a\mathbf{1} + b\mathbf{C}\mathbf{x} = \mathbf{C}\mathbf{c} \quad \text{for some } a, b \in \mathbb{R},\ \mathbf{c} \in \mathbb{R}^n. \tag{3.20}$$

Premultiplying (3.20) by $\mathbf{C}$ shows that $\mathbf{u} = b\mathbf{C}\mathbf{x} \in \mathscr{C}(\mathbf{C}\mathbf{x})$, and thus (3.16) is proved. The claim

$$\mathscr{C}(\mathbf{C}\mathbf{x})^{\perp} = \mathscr{C}(\mathbf{1} : \mathbf{x})^{\perp} \boxplus \mathscr{C}(\mathbf{1}) \tag{3.21}$$

follows by taking orthocomplements of each side of (3.16). □

We may note in passing that (3.18) means that

$$\begin{aligned}\mathbf{P}_{(\mathbf{1}\,:\,\mathbf{x})} = \mathbf{P}_{(\mathbf{1}\,:\,\mathbf{Cx})} &= (\mathbf{1} \,:\, \mathbf{Cx}) \begin{pmatrix} n^{-1} & 0 \\ 0 & (\mathbf{x}'\mathbf{Cx})^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{1} \\ \mathbf{Cx} \end{pmatrix} \\ &= \mathbf{J} + \mathbf{Cxx}'\mathbf{C}/\mathbf{x}'\mathbf{Cx} = \mathbf{J} + \mathbf{P}_{\mathbf{Cx}}\,, \qquad (3.22a) \\ \mathbf{I} - \mathbf{P}_{(\mathbf{1}\,:\,\mathbf{x})} &= \mathbf{C} - \mathbf{P}_{\mathbf{Cx}}\,. \qquad (3.22b)\end{aligned}$$

The decomposition (3.22) follows also at once from Theorem 8 (p. 155).

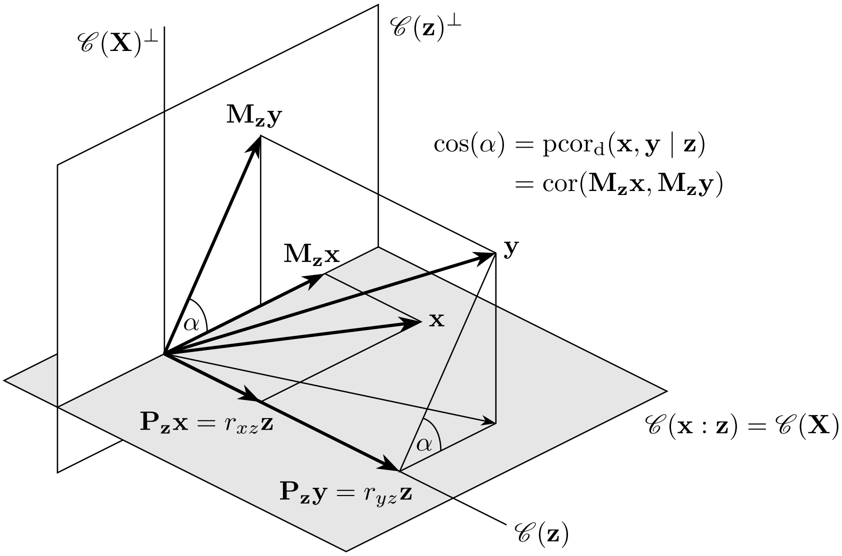


Figure 3.2 Partial correlation $r_{xy\cdot z}$ geometrically.

3.2 On the Role of the Vector 1

In this section we discuss some features of the constant term in linear regression, that is, on the role of vector **1**; see Isotalo, Puntanen & Styan (2006).

So, consider a very simple linear model

$$\mathscr{M}_0 = \{\mathbf{y},\, \mathbf{1}\beta_0,\, \sigma^2\mathbf{I}\}\,. \qquad (3.23)$$

Then obviously

- the vector $\mathbf{Jy} = \mathrm{OLSE}(\mathbf{1}\beta_0) = \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'\mathbf{y} = \bar{y}\mathbf{1}$ is the vector of the fitted values,
- the vector $\mathbf{Cy} = \mathbf{y} - \mathbf{Jy} = \tilde{\mathbf{y}}$ is the vector of the residuals,
- the residual sum of squares under $\mathscr{M}_0$ is

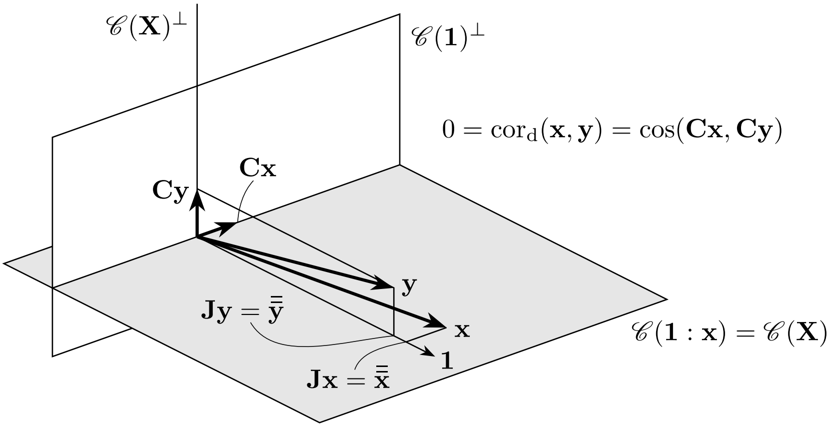


Figure 3.3 High cosine between $\mathbf{x}$ and $\mathbf{y}$ but no correlation. Vectors $\mathbf{x}$ and $\mathbf{y}$ very close to $\mathbf{1}$.

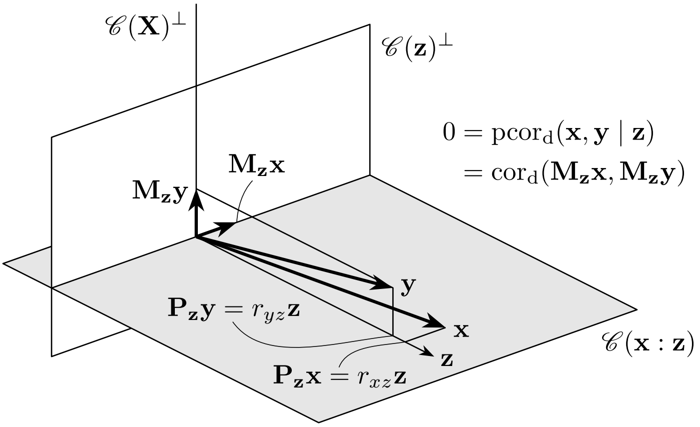


Figure 3.4 Partial correlation $r_{xy\cdot z} = 0$. Note that if $\mathbf{y}$ is explained by $\mathbf{x}$ and $\mathbf{z}$, then the $\mathbf{x}$'s regression coefficient is 0.

$$\mathrm{SSE}(\mathscr{M}_0) = \mathrm{SSE}_0 = \mathbf{y}'(\mathbf{I} - \mathbf{J})\mathbf{y} = \mathbf{y}'\mathbf{C}\mathbf{y} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \mathrm{SST}\,. \quad (3.24)$$

What happens here is that we "eliminate" the effect of the column vector of ones, i.e., $\mathbf{1}$, from $\mathbf{y}$, and, what is left is just the residual which in this case is the centered $\mathbf{y}$, i.e., $\tilde{\mathbf{y}}$.

Consider now the model

$$\mathscr{M}_{xy} = \{\mathbf{y},\, \mathbf{x}\beta,\, \sigma^2\mathbf{I}\}\,. \quad (3.25)$$

Then

- The vector $\hat{\mathbf{y}} = \mathrm{OLSE}(\mathbf{x}\beta)$ is the vector of the fitted values:

$$\hat{\mathbf{y}} = \mathbf{P}_{\mathbf{x}}\mathbf{y} = \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}' \cdot \mathbf{y} = \hat{\beta}\mathbf{x}\,, \qquad \text{where } \hat{\beta} = \frac{\mathbf{x}'\mathbf{y}}{\mathbf{x}'\mathbf{x}}\,. \tag{3.26}$$

- The vector $(\mathbf{I} - \mathbf{P}_{\mathbf{x}})\mathbf{y}$ is the vector of the residuals.
- The residual sum of squares under $\mathscr{M}_{xy}$ is

$$\mathrm{SSE}(\mathscr{M}_{xy}) = \mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{x}})\mathbf{y} = \|(\mathbf{I} - \mathbf{P}_{\mathbf{x}})\mathbf{y}\|^2. \tag{3.27}$$

A very natural measure for the "goodness" of this model would be

$$R^2_{xy} = \frac{\|\mathbf{P}_{\mathbf{x}}\mathbf{y}\|^2}{\|\mathbf{y}\|^2} = \frac{\|\hat{\mathbf{y}}\|^2}{\|\mathbf{y}\|^2} = \frac{\mathbf{y}' \cdot \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}' \cdot \mathbf{y}}{\mathbf{y}'\mathbf{y}} = \cos^2(\mathbf{x}, \mathbf{y}). \tag{3.28}$$

Note that the goodness measure above can also be expressed as

$$R^2_{xy} = 1 - \sin^2(\mathbf{x}, \mathbf{y}) = 1 - \frac{\|(\mathbf{I} - \mathbf{P}_{\mathbf{x}})\mathbf{y}\|^2}{\|\mathbf{y}\|^2} = 1 - \frac{\mathrm{SSE}(\mathscr{M}_{xy})}{\|\mathbf{y}\|^2}\,. \tag{3.29}$$

The quantity R^2_{xy} defined here is called the *coefficient of determination* under $\mathscr{M}_{xy}$. It is essential to notice that there is no vector $\mathbf{1}$ in the model $\mathscr{M}_{xy}$ and hence R^2_{xy} does not necessarily equal to r^2_{xy}, which is the coefficient of determination in the model $\mathscr{M}_{12} = \{\mathbf{y},\, \beta_0\mathbf{1} + \beta_1\mathbf{x},\, \sigma^2\mathbf{I}\}$.

If we now "eliminate the effect of $\mathbf{1}$ from $\mathbf{y}$ and $\mathbf{x}$" (as the phraseology goes), i.e., we center them, then we can consider the model

$$\mathscr{M}_{xy\cdot 1} = \{\tilde{\mathbf{y}},\, \tilde{\mathbf{x}}\beta,\, \#\}\,, \tag{3.30}$$

where we have deliberately left the covariance matrix unnotated. Notice that the model $\mathscr{M}_{xy\cdot 1}$ can be obtained by premultiplying the model $\mathscr{M}_{xy}$ by the centering matrix $\mathbf{C}$. Now a natural measure for the goodness of the model $\mathscr{M}_{xy\cdot 1}$ would be

$$R^2_{xy\cdot 1} = \frac{\|\mathbf{P}_{\tilde{\mathbf{x}}}\tilde{\mathbf{y}}\|^2}{\|\tilde{\mathbf{y}}\|^2} = \frac{\tilde{\mathbf{y}}' \cdot \tilde{\mathbf{x}}(\tilde{\mathbf{x}}'\tilde{\mathbf{x}})^{-1}\tilde{\mathbf{x}}' \cdot \tilde{\mathbf{y}}}{\tilde{\mathbf{y}}'\tilde{\mathbf{y}}} = \cos^2(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})\,. \tag{3.31}$$

Obviously $R^2_{xy\cdot 1} = \mathrm{cor_d}^2(\mathbf{x}, \mathbf{y}) = r^2_{xy}$, and hence we have shown that r^2_{xy} can be interpreted as a measure of goodness when $\mathbf{y}$ is regressed on $\mathbf{x}$ after the effect of $\mathbf{1}$ has been eliminated from both $\mathbf{x}$ and $\mathbf{y}$.

Above we have ended up to some well-known concepts in linear regression. We have used only natural and intuitive ways for our considerations. Let us next take another more formal look at the situation.

Consider the following four models:

$$\mathscr{M}_{12} = \{\mathbf{y},\, \beta_0\mathbf{1} + \beta_1\mathbf{x},\, \sigma^2\mathbf{I}\}\,, \qquad \mathscr{M}_{12\cdot 1} = \{\mathbf{C}\mathbf{y},\, \beta_1\mathbf{C}\mathbf{x},\, \sigma^2\mathbf{C}\}\,, \tag{3.32a}$$

$$\mathscr{M}_{c} = \{\mathbf{y},\, \beta_0\mathbf{1} + \beta_1\mathbf{C}\mathbf{x},\, \sigma^2\mathbf{I}\}\,, \qquad \mathscr{M}_{r} = \{\mathbf{y},\, \beta_1\mathbf{C}\mathbf{x},\, \sigma^2\mathbf{I}\}\,. \tag{3.32b}$$

The model $\mathscr{M}_{12}$ is the full model and all other models are various versions of it. Obviously $\mathscr{M}_{12\cdot 1}$ corresponds to $\mathscr{M}_{xy\cdot 1}$ which was considered above. In all these versions we have done something related to the constant term: we have centered something. The above models frequently appear in practice (and in teaching regression in statistics courses).

Taking a look at the four models in (3.32), we immediately observe an interesting feature which we may state as a proposition:

Proposition 3.2. *Consider the models defined in* (3.32), *and let* $\mathbf{X} = (\mathbf{1} : \mathbf{x})$ *have full column rank. Then* $\hat{\beta}_1$ *is the same in each model, i.e.,*

$$\hat{\beta}_1(\mathscr{M}_{12}) = \hat{\beta}_1(\mathscr{M}_{12\cdot 1}) = \hat{\beta}_1(\mathscr{M}_r) = \hat{\beta}_1(\mathscr{M}_c)\,. \tag{3.33}$$

Moreover, the residuals under the models $\mathscr{M}_{12}$, $\mathscr{M}_{12\cdot 1}$, *and* $\mathscr{M}_c$ *are identical.*

Proof. The first two equalities in (3.33) are obvious. Model $\mathscr{M}_c$ is a reparameterization of $\mathscr{M}_{12}$. This is seen from $(\mathbf{1} : \mathbf{Cx}) = (\mathbf{1} : \mathbf{x})\mathbf{A}$, where

$$\mathbf{A} = \begin{pmatrix} 1 & -\bar{x} \\ 0 & 1 \end{pmatrix}. \tag{3.34}$$

It is easy to confirm that $\hat{\beta}_0(\mathscr{M}_c) = \bar{y}$ and $\hat{\beta}_1(\mathscr{M}_c) = \hat{\beta}_1(\mathscr{M}_{12})$. In view of (3.18), we have

$$\mathscr{C}(\mathbf{1} : \mathbf{Cx}) = \mathscr{C}(\mathbf{1} : \mathbf{x}), \quad \mathbf{P}_{(\mathbf{1} : \mathbf{Cx})} = \mathbf{P}_{(\mathbf{1} : \mathbf{x})}\,, \tag{3.35}$$

and hence the residual vectors under $\mathscr{M}_c$ and $\mathscr{M}_{12}$ are identical:

$$\operatorname{res}(\mathscr{M}_c) = \operatorname{res}(\mathscr{M}_{12}) = \mathbf{y} - \mathbf{Hy} = \mathbf{My}\,. \tag{3.36}$$

The residual vector under $\mathscr{M}_{12\cdot 1}$ becomes

$$\begin{aligned}
\operatorname{res}(\mathscr{M}_{12\cdot 1}) &= \mathbf{Cy} - \mathbf{P}_{\mathbf{Cx}}\mathbf{Cy} \\
&= \mathbf{y} - [\mathbf{Jy} + \mathbf{Cx}(\mathbf{x}'\mathbf{Cx})^{-1}\mathbf{x}'\mathbf{Cy}] \\
&= \mathbf{y} - [\bar{y}\mathbf{1} + (\mathbf{x} - \bar{x}\mathbf{1})\hat{\beta}_1] \\
&= \mathbf{y} - [(\bar{y} - \bar{x}\hat{\beta}_1)\mathbf{1} + \hat{\beta}_1\mathbf{x}] \\
&= \mathbf{y} - (\hat{\beta}_0\mathbf{1} + \hat{\beta}_1\mathbf{x}) \\
&= \mathbf{y} - (\mathbf{1} : \mathbf{x})\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} \\
&= \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \operatorname{res}(\mathscr{M}_{12})\,,
\end{aligned} \tag{3.37}$$

where

$$\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1 \quad \text{and} \quad \hat{\beta}_1 = \mathbf{x}'\mathbf{Cy}/\mathbf{x}'\mathbf{Cx}\,. \tag{3.38}$$

We shall later give an alternative proof for (3.37) using a decomposition of the orthogonal projector $\mathbf{H}$; see (8.72) (p. 164). □

The first equality in (3.33) is known as (a special case of) the Frisch–Waugh–Lovell theorem, see Section 8.4 (p. 163).

We may end up to R^2 via different routes. One very natural approach in which the vector $\mathbf{1}$ has an important role is to consider the simple basic model where the only explanatory variable is a constant: $\mathscr{M}_0 = \{\mathbf{y},\, \mathbf{1}\beta_0,\, \sigma^2\mathbf{I}\}$. Under $\mathscr{M}_0$ we have $\mathrm{OLSE}(\mathbf{1}\beta_0) = \mathbf{J}\mathbf{y} = \bar{y}\mathbf{1}$, while the residual vector is the centered $\mathbf{y}$, that is, $\mathbf{C}\mathbf{y}$, and hence the residual sum of squares under $\mathscr{M}_0$ is

$$\mathrm{SSE}_0 = \mathbf{y}'(\mathbf{I}-\mathbf{J})\mathbf{y} = \mathbf{y}'\mathbf{C}\mathbf{y} = \|\mathbf{C}\mathbf{y}\|^2 = t_{yy} = \sum_{i=1}^{n}(y_i-\bar{y})^2 = \mathrm{SST}\,. \tag{3.39}$$

We may want to compare the full model $\mathscr{M}_{12} = \{\mathbf{y},\, \mathbf{1}\beta_0 + \mathbf{x}\beta_1,\, \sigma^2\mathbf{I}\}$ and the simple basic model $\mathscr{M}_0$ by means of the residual sum of squares: how much benefit is gained in the residual sum of squares when using also the x-variable as an explanatory variable. The residual sum of squares under $\mathscr{M}_{12}$ is

$$\begin{aligned}\mathrm{SSE} &= \mathbf{y}'(\mathbf{I}-\mathbf{H})\mathbf{y} = \|(\mathbf{I}-\mathbf{H})\mathbf{y}\|^2 \\ &= \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 = \sum_{i=1}^{n}\big[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\big]^2.\end{aligned} \tag{3.40}$$

The change in SSE when moving from $\mathscr{M}_0$ to $\mathscr{M}_{12}$ is

$$\mathrm{SSE}_0 - \mathrm{SSE} = \mathbf{y}'(\mathbf{I}-\mathbf{J})\mathbf{y} - \mathbf{y}'(\mathbf{I}-\mathbf{H})\mathbf{y} = \mathbf{y}'(\mathbf{H}-\mathbf{J})\mathbf{y} = \mathrm{SSR}\,, \tag{3.41}$$

which is called "sum of squares due to regression". In this situation the matrix $\mathbf{H}-\mathbf{J}$ is symmetric and idempotent, see Proposition 7.1 (p. 152) and (8.119) (p. 171):

$$\mathbf{H}-\mathbf{J} = \mathbf{P}_{\mathbf{C}\mathbf{x}} = \mathbf{P}_{\tilde{\mathbf{x}}} = \mathbf{P}_{\mathscr{C}(\mathbf{X})\cap\mathscr{C}(\mathbf{1})^\perp}\,. \tag{3.42}$$

Hence

$$\mathrm{SSR} = \|(\mathbf{H}-\mathbf{J})\mathbf{y}\|^2 = \sum_{i=1}^{n}(\hat{y}_i-\bar{y}_i)^2. \tag{3.43}$$

The value of SSR tells the reduction in SSE when using $\mathscr{M}_{12}$ instead $\mathscr{M}_0$, but it is definitely more informative to study the *relative* reduction in SSE, that is, we have reasons to calculate the ratio

$$\frac{\mathrm{SSE}_0-\mathrm{SSE}}{\mathrm{SSE}_0} = \frac{\mathrm{SST}-\mathrm{SSE}}{\mathrm{SST}} = \frac{\mathrm{SSR}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}} = R^2. \tag{3.44}$$

One fundamental property of R^2 defined above is that it equals the square of the correlation coefficient between the $\mathbf{y}$ and $\mathbf{H}\mathbf{y}$. We return to this result later on but we state it here since it nicely illustrates the important role of $\mathbf{1}$.

Proposition 3.3. *Consider the model* $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}\}$, *where* $\mathbf{1} \in \mathscr{C}(\mathbf{X})$ *and let* R^2 *be defined as in* (3.44). *Then*

$$R = \operatorname{cor_d}(\mathbf{y}, \mathbf{Hy}) = \cos(\mathbf{Cy}, \mathbf{CHy})\,. \tag{3.45}$$

Proof. We'll return to the proof later on, see Proposition 8.6 (p. 172). □

Taking a look at the $n \times (k+1)$ model matrix $\mathbf{X} = (\mathbf{1} : \mathbf{x}_1 : \ldots : \mathbf{x}_k)$, the first column $\mathbf{1}$ looks just as any other column but there is one big difference: all other variables represented in the model matrix $\mathbf{X}$ have a nonzero sample variance. This is so because of course we can request that there are no multiples of $\mathbf{1}$ among $\mathbf{x}_1, \ldots, \mathbf{x}_k$. Belsley (1991, p. 196) writes:

> "Much confusion surrounding centering arises because of some commonly held misconceptions about the 'constant term'. This section [6.8] aims at several of these issues with the goal of showing that, for most of the part, despite much practice to the contrary, the constant is most reasonably viewed as just another element in a regression analysis that plays no role different from any other 'variate'."

The geometry behind decomposition SST = SSR + SSE is illustrated in Figure 3.5.

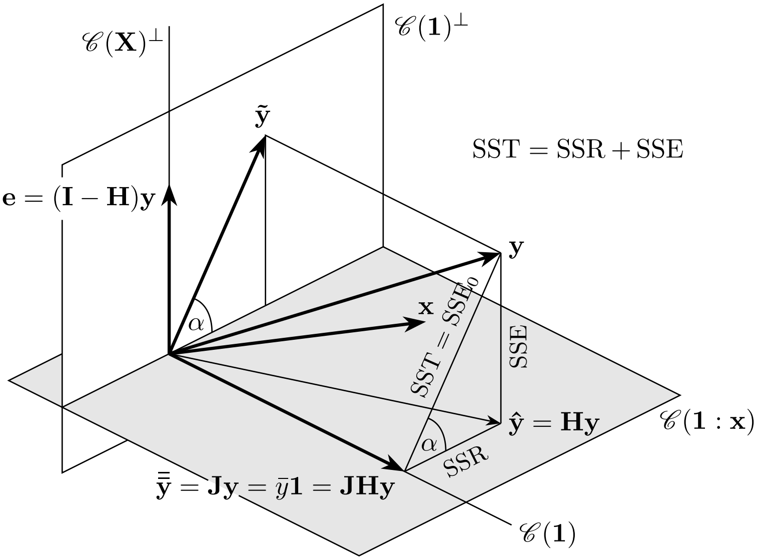


Figure 3.5 Illustration of SST = SSR + SSE.

3.3 Partial Correlation Geometrically

Here we very shortly consider sample partial correlations and remind the readers that they are just ordinary correlations between specific residuals.

Consider two linear models:

$$\mathscr{M}_x = \{\mathbf{x},\, (\mathbf{1} : \mathbf{z})\boldsymbol{\beta},\, \sigma^2\mathbf{I}\}\,, \quad \mathscr{M}_y = \{\mathbf{y},\, (\mathbf{1} : \mathbf{z})\boldsymbol{\beta},\, \sigma^2\mathbf{I}\}\,. \tag{3.46}$$

The residuals in these models are

$$\mathbf{u} = \mathbf{e}_{\mathbf{x}\cdot\mathbf{Z}} = \operatorname{res}(\mathbf{x}; \mathbf{Z}) = (\mathbf{I} - \mathbf{P}_{\mathbf{Z}})\mathbf{x}\,, \tag{3.47a}$$

$$\mathbf{v} = \mathbf{e}_{\mathbf{y}\cdot\mathbf{Z}} = \operatorname{res}(\mathbf{y}; \mathbf{Z}) = (\mathbf{I} - \mathbf{P}_{\mathbf{Z}})\mathbf{y}\,, \tag{3.47b}$$

where $\mathbf{Z} = (\mathbf{1} : \mathbf{z})$. Premultiplying the models in (3.46) by the centering matrix $\mathbf{C}$ we get the models which have the same residuals as in (3.47). To simplify the notation, let us assume that vectors $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$ are all centered and denote the corresponding models as

$$\mathscr{M}_{cx} = \{\mathbf{x},\, \mathbf{z}\beta,\, \#\}\,, \quad \mathscr{M}_{cy} = \{\mathbf{y},\, \mathbf{z}\beta,\, \#\}\,. \tag{3.48}$$

Moreover, since we are now interested in the correlation between these two residuals, we can assume that the vectors $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$ in (3.48) are not only centered but they also have unit length:

$$\mathbf{1}_n'(\mathbf{x} : \mathbf{y} : \mathbf{z}) = \mathbf{0}\,, \quad \mathbf{x}'\mathbf{x} = \mathbf{y}'\mathbf{y} = \mathbf{z}'\mathbf{z} = 1\,. \tag{3.49}$$

Now

$$\mathbf{x}'\mathbf{y} = r_{xy}\,, \quad \mathbf{x}'\mathbf{z} = r_{xz}\,, \quad \mathbf{y}'\mathbf{z} = r_{yz}\,, \tag{3.50a}$$

$$\mathbf{u} = \mathbf{e}_{\mathbf{x}\cdot\mathbf{z}} = \operatorname{res}(\mathbf{x}; \mathbf{z}) = (\mathbf{I} - \mathbf{P}_{\mathbf{z}})\mathbf{x} = (\mathbf{I} - \mathbf{z}\mathbf{z}')\mathbf{x} = \mathbf{x} - r_{xz}\mathbf{z}\,, \tag{3.50b}$$

$$\mathbf{v} = \mathbf{e}_{\mathbf{y}\cdot\mathbf{z}} = \operatorname{res}(\mathbf{y}; \mathbf{z}) = (\mathbf{I} - \mathbf{P}_{\mathbf{z}})\mathbf{y} = (\mathbf{I} - \mathbf{z}\mathbf{z}')\mathbf{y} = \mathbf{y} - r_{yz}\mathbf{z}\,, \tag{3.50c}$$

and

$$\mathbf{u}'\mathbf{v} = \mathbf{x}'(\mathbf{I} - \mathbf{z}\mathbf{z}')\mathbf{y} = r_{xy} - r_{xz}r_{yz}\,, \tag{3.51a}$$

$$\mathbf{u}'\mathbf{u} = \mathbf{x}'(\mathbf{I} - \mathbf{z}\mathbf{z}')\mathbf{x} = 1 - r_{xz}^2\,, \tag{3.51b}$$

$$\mathbf{v}'\mathbf{v} = \mathbf{y}'(\mathbf{I} - \mathbf{z}\mathbf{z}')\mathbf{y} = 1 - r_{yz}^2\,, \tag{3.51c}$$

and so we obtain the well-known formula for the partial correlation coefficient:

$$r_{xy\cdot z} = \operatorname{pcor_d}(\mathbf{x}, \mathbf{y} \mid \mathbf{z}) = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}}\,. \tag{3.52}$$

Obviously $r_{xy\cdot z}$ is well-defined only if $r_{xz}^2 \neq 1$ and f $r_{yz}^2 \neq 1$. In matrix terms we have

$$\begin{aligned}(\mathbf{u} : \mathbf{v})'(\mathbf{u} : \mathbf{v}) &= (\mathbf{x} : \mathbf{y})'(\mathbf{I} - \mathbf{P}_{\mathbf{z}})(\mathbf{x} : \mathbf{y}) \\ &= (\mathbf{x} : \mathbf{y})'(\mathbf{x} : \mathbf{y}) - (\mathbf{x} : \mathbf{y})'\mathbf{z}\mathbf{z}'(\mathbf{x} : \mathbf{y}) \\ &= \mathbf{R}_{11} - \mathbf{r}_{12}\mathbf{r}_{12}' := \mathbf{R}_{11\cdot 2}\,, \end{aligned} \tag{3.53}$$

where

$$\operatorname{cor_d}(\mathbf{x} : \mathbf{y} : \mathbf{z}) = \mathbf{R} = \begin{pmatrix} \mathbf{R}_{11} & \mathbf{r}_{12} \\ \mathbf{r}_{12}' & 1 \end{pmatrix}. \tag{3.54}$$

Hence the matrix of the partial correlations can be expressed as

$$\operatorname{pcor_d}[(\mathbf{x} : \mathbf{y}) \mid \mathbf{z}] = [\operatorname{diag}(\mathbf{R}_{11\cdot 2})]^{-1/2}\mathbf{R}_{11\cdot 2}[\operatorname{diag}(\mathbf{R}_{11\cdot 2})]^{-1/2}. \tag{3.55}$$

In terms of the original variables, we have

$$r_{xy\cdot z} = \cos[\operatorname{res}(\mathbf{x}; \mathbf{Z}), \operatorname{res}(\mathbf{y}; \mathbf{Z})] = \cos[(\mathbf{I} - \mathbf{P}_{\mathbf{Z}})\mathbf{x}, (\mathbf{I} - \mathbf{P}_{\mathbf{Z}})\mathbf{y}]\,, \tag{3.56}$$

where $\mathbf{Z} = (\mathbf{1} : \mathbf{z})$. Denoting

$$\mathbf{U} = (\mathbf{x} : \mathbf{y} : \mathbf{z}) = (\mathbf{U}_1 : \mathbf{z}) \tag{3.57}$$

we get the matrix of the residual vectors

$$\mathbf{E}_{\mathbf{U}_1\cdot\mathbf{z}} = (\mathbf{I} - \mathbf{P}_{(\mathbf{1}\,:\,\mathbf{z})})\mathbf{U}_1 = (\mathbf{e}_{\mathbf{x}\cdot\mathbf{z}} : \mathbf{e}_{\mathbf{y}\cdot\mathbf{z}})\,. \tag{3.58}$$

Because, see (3.22) (p. 94), $\mathbf{I} - \mathbf{P}_{(\mathbf{1}\,:\,\mathbf{z})} = \mathbf{C} - \mathbf{P}_{\mathbf{Cz}}$, where $\mathbf{C}$ is the centering matrix, we can write $\mathbf{E}_{\mathbf{U}_1\cdot\mathbf{z}} = (\mathbf{C} - \mathbf{P}_{\mathbf{Cz}})\mathbf{U}_1$. Moreover, denoting

$$\mathbf{T}_{11\cdot 2} := \mathbf{E}_{\mathbf{U}_1\cdot\mathbf{z}}'\mathbf{E}_{\mathbf{U}_1\cdot\mathbf{z}} = \mathbf{U}_1'(\mathbf{C} - \mathbf{P}_{\mathbf{Cz}})\mathbf{U}_1 = \mathbf{U}_1'\mathbf{C}\mathbf{U}_1 - \mathbf{U}_1'\mathbf{P}_{\mathbf{Cz}}\mathbf{U}_1\,, \tag{3.59}$$

the matrix of partial correlations can be expressed as

$$\operatorname{pcor_d}(\mathbf{U}_1 \mid \mathbf{z}) = [\operatorname{diag}(\mathbf{T}_{11\cdot 2})]^{-1/2}\mathbf{T}_{11\cdot 2}[\operatorname{diag}(\mathbf{T}_{11\cdot 2})]^{-1/2}. \tag{3.60}$$

Notice, in view of the rules for inverting a partitioned matrix, see Section 13.1 (p. 294), that $\mathbf{T}_{11\cdot 2}^{-1}$ is upper-left block of the inverse of

$$\mathbf{T} = \mathbf{U}'\mathbf{C}\mathbf{U} = \begin{pmatrix} \mathbf{U}_1'\mathbf{C}\mathbf{U}_1 & \mathbf{U}_1'\mathbf{C}\mathbf{z} \\ \mathbf{z}'\mathbf{C}\mathbf{U}_1 & \mathbf{z}'\mathbf{C}\mathbf{z} \end{pmatrix} = \begin{pmatrix} \mathbf{T}_{11} & \mathbf{t}_{12} \\ \mathbf{t}_{12}' & t_{zz} \end{pmatrix}. \tag{3.61}$$

3.4 Exercises

3.1. Let $\mathbf{U} = (\mathbf{1} : \mathbf{x} : \mathbf{y})$ be a 100×3 matrix where $\mathbf{x}$ and $\mathbf{y}$ comprise the observed values of variables x and y, and assume that $s_x^2 > 0$ and $s_y^2 > 0$.

(a) Suppose that the columns of $\mathbf{U}$ are orthogonal. Show that $r_{xy} = 0$.

(b) Confirm that if $\operatorname{rank}(\mathbf{U}) = 2$, then $r_{xy}^2 = 1$.

3.2. Let $\mathbf{x}$ and $\mathbf{y}$ be n-dimensional variable vectors (comprising the observed values of x an y). Show that

$$\operatorname{cor_d}(a\mathbf{1} + b\mathbf{x}, c\mathbf{1} + d\mathbf{y}) = \begin{cases} \operatorname{cor_d}(\mathbf{x}, \mathbf{y}) & \text{if } bd > 0\,, \\ -\operatorname{cor_d}(\mathbf{x}, \mathbf{y}) & \text{if } bd < 0\,, \end{cases}$$

or in other notation,

$$\operatorname{cor_d}\left[(\mathbf{1} : \mathbf{x})\left(\begin{smallmatrix} a \\ b \end{smallmatrix}\right), (\mathbf{1} : \mathbf{y})\left(\begin{smallmatrix} c \\ d \end{smallmatrix}\right)\right] = \begin{cases} r_{xy} & \text{if } bd > 0\,, \\ -r_{xy} & \text{if } bd < 0\,. \end{cases}$$

3.3. Consider the 3×3 data matrix

$$\mathbf{U} = (\mathbf{1}_3 : \mathbf{x} : \mathbf{y}) = \begin{pmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{pmatrix}.$$

Confirm that the area of the triangle with vertices (x_i, y_i), $i = 1, 2, 3$, is $\frac{1}{2}\det(\mathbf{U})$. Show that $\det(\mathbf{U}) = 0 \iff$ the three data points lie in the same line, and hence, assuming that x and y have nonzero variances, we have $r_{xy}^2 = 1 \iff \det(\mathbf{U}) = 0$.

3.4. Consider the variables x_1 and x_2 and their sum $y = x_1 + x_2$. Let the corresponding variable vectors be $\mathbf{y}$, $\mathbf{x}_1$, and $\mathbf{x}_2$. Assume that $\mathbf{x}_1$ and $\mathbf{x}_2$ are centered and of equal length so that $\mathbf{x}_1'\mathbf{x}_1 = \mathbf{x}_2'\mathbf{x}_2 = d^2$. Show that

$$\operatorname{var_s}(y) = \tfrac{1}{n-1} 2d^2(1 + r_{12})\,, \quad \operatorname{cor_s^2}(x_1, y) = \tfrac{1}{2}(1 + r_{12})\,.$$

Moreover, let $\mathbf{u} = \mathbf{x}_1 - \mathbf{P_y}\mathbf{x}_1$ and $\mathbf{v} = \mathbf{x}_2 - \mathbf{P_y}\mathbf{x}_2$. Show that $\mathbf{u} = -\mathbf{v}$ and thereby $r_{12 \cdot y} = \operatorname{cor_d}(\mathbf{u}, \mathbf{v}) = -1$.

3.5. Let the variable vectors $\mathbf{x}$ and $\mathbf{y}$ be centered and of equal length.

(a) What is $\operatorname{cor_s}(x - y, x + y)$?

(b) Draw a figure from which you can conclude the result above.

(c) What about if $\mathbf{x}'\mathbf{x} = \mathbf{y}'\mathbf{y} + 0.001$?

3.6. Consider the variable vectors $\mathbf{x}$ and $\mathbf{y}$ which are centered and of unit length. Let $\mathbf{u}$ be the residual when $\mathbf{y}$ is explained by $\mathbf{x}$ and $\mathbf{v}$ be the residual when $\mathbf{x}$ is explained by $\mathbf{y}$. What is the $\operatorname{cor_d}(\mathbf{u}, \mathbf{v})$? Draw a figure about the situation.

3.7. Consider the variable vectors $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$ which are centered and of unit length and whose correlation matrix is

$$\mathbf{R} = \begin{pmatrix} 1 & 0 & -1/\sqrt{2} \\ 0 & 1 & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} & 1 \end{pmatrix}.$$

Show that $\mathbf{x} - \mathbf{y} + \sqrt{2}\mathbf{z} = \mathbf{0}$.

3.8. Consider the 17-dimensional variable vectors $\mathbf{x} = (1, \dots, 1, 1, 4)'$ and $\mathbf{y} = (-1, \dots, -1, -1 + 0.001, 4)'$. Find $\operatorname{cor}_{\mathrm{d}}(\mathbf{x}, \mathbf{y})$ and $\cos(\mathbf{x}, \mathbf{y})$.

3.9. Suppose that $\mathbf{U} = (\mathbf{x} : \mathbf{y} : \mathbf{z}) \in \mathbb{R}^{100 \times 3}$ is a data matrix, where the x-values represent the results of throwing a dice and y-values are observations from the normal distribution $\mathrm{N}(0, 1)$, and $\mathbf{z} = \mathbf{x} + \mathbf{y}$. What is your guess for the length of the variable vector $\mathbf{z}$?

3.10. Consider a 3×3 centering matrix

$$\mathbf{C} = \begin{pmatrix} 2/3 & -1/3 & -1/3 \\ -1/3 & 2/3 & -1/3 \\ -1/3 & -1/3 & 2/3 \end{pmatrix}.$$

Find the following:

(a) a basis for $\mathscr{C}(\mathbf{C})$,
(b) orthonormal bases for $\mathscr{C}(\mathbf{C})$ and $\mathscr{C}(\mathbf{C})^{\perp}$,
(c) eigenvalues and orthonormal eigenvectors of $\mathbf{C}$,
(d) $\mathbf{C}^{+}$, and $\mathbf{C}^{-}$ which has (i) rank 2, (ii) rank 3.

3.11. Suppose that the variable vectors $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$ are centered and of unit length. Show that corresponding to (3.8) (p. 93),

$$r_{xy\cdot z} = 0 \iff \mathbf{y} \in \mathscr{C}(\mathbf{Q}_{\mathbf{z}}\mathbf{x})^{\perp} = \mathscr{C}(\mathbf{x} : \mathbf{z})^{\perp} \boxplus \mathscr{C}(\mathbf{z}),$$

where $\mathbf{Q}_{\mathbf{z}} = \mathbf{I}_n - \mathbf{P}_{\mathbf{z}}$. See also Exercise 8.7 (p. 185).

Chapter 4
Generalized Inverses in a Nutshell

> *Any philosophy that can be put in a nutshell belongs there.*
>
> Sydney J. Harris

Let $\mathbf{A}$ be a given $n \times m$ matrix and $\mathbf{y}$ a given $n \times 1$ vector. Consider the linear equation

$$\mathbf{Ab} = \mathbf{y}\,. \tag{4.1}$$

Our task is to find such a vector $\mathbf{b} \in \mathbb{R}^m$ that $\mathbf{Ab} = \mathbf{y}$ is satisfied. It is well known that (4.1) has

(a) no solution at all, or

(b) a unique solution, or

(c) infinite number of solutions.

Moreover, it is easy to conclude the following:

- The equation $\mathbf{Ab} = \mathbf{y}$ has a solution, i.e., it is consistent, if and only if $\mathbf{y}$ belongs to $\mathscr{C}(\mathbf{A})$.

If a solution exists, it is unique if and only if the columns of $\mathbf{A}$ are linearly independent, i.e., $\mathbf{A}$ has full column rank. Then $(\mathbf{A}'\mathbf{A})^{-1}$ exists and premultiplying (4.1) by $(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$ yields

$$\mathbf{b}_0 = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{y} := \mathbf{G}_1\mathbf{y}\,. \tag{4.2}$$

It is obvious that the unique solution to (4.1) can be expressed also as

$$\mathbf{b}_0 = (\mathbf{A}'\mathbf{N}\mathbf{A})^{-1}\mathbf{A}'\mathbf{N}\mathbf{y} := \mathbf{G}_2\mathbf{y}\,, \tag{4.3}$$

where $\mathbf{N}$ an arbitrary $n \times n$ matrix satisfying condition $\operatorname{rank}(\mathbf{A}'\mathbf{N}\mathbf{A}) = m$. Note that it is possible that $\mathbf{G}_1\mathbf{y} = \mathbf{G}_2\mathbf{y}$ but $\mathbf{G}_1 \neq \mathbf{G}_2$. If $\mathbf{A}$ is a nonsingular square matrix, then of course $\mathbf{b}_0 = \mathbf{A}^{-1}\mathbf{y}$.

We can now try to find such an $m \times n$ matrix $\mathbf{G}$, which would behave like $\mathbf{A}^{-1}$ as much as possible; for example, we might wish that $\mathbf{G}$ has such a property that if (4.1) is consistent, then $\mathbf{Gy}$ is one solution. Then $\mathbf{G}$ would work like $\mathbf{A}^{-1}$ while solving a linear equation; corresponding to $\mathbf{A}^{-1}\mathbf{y}$ we have $\mathbf{Gy}$ as a solution to $\mathbf{Ab} = \mathbf{y}$.

S. Puntanen et al., *Matrix Tricks for Linear Statistical Models: Our Personal Top Twenty*,
DOI 10.1007/978-3-642-10473-2_5,

We might emphasize (and repeat) that $\mathbf{Ab} = \mathbf{y}$ does not necessarily have a solution in which case $\mathbf{G}$ does not help in finding a nonexistent solution.

We can define the generalized inverse $\mathbf{G}_{m\times n}$ of matrix $\mathbf{A}_{n\times m}$ in three equivalent ways, see Rao & Mitra (1971b, pp. 20–21):

Theorem 4. *The matrix* $\mathbf{G}_{m\times n}$ *is a generalized inverse of* $\mathbf{A}_{n\times m}$ *if any of the following equivalent conditions holds:*

(a) *The vector* $\mathbf{Gy}$ *is a solution to* $\mathbf{Ab} = \mathbf{y}$ *always when this equation is consistent, i.e., always when* $\mathbf{y} \in \mathscr{C}(\mathbf{A})$.

(b) (b1) $\mathbf{GA}$ *is idempotent and* $\operatorname{rank}(\mathbf{GA}) = \operatorname{rank}(\mathbf{A})$, *or equivalently*
(b2) $\mathbf{AG}$ *is idempotent and* $\operatorname{rank}(\mathbf{AG}) = \operatorname{rank}(\mathbf{A})$.

(c) $\mathbf{AGA} = \mathbf{A}$. (mp1)

Proof. We first show that that (c) and (b2) are equivalent, i.e,

$$\mathbf{AGA} = \mathbf{A} \iff \mathbf{AG} \text{ is idempotent and } \operatorname{rank}(\mathbf{AG}) = \operatorname{rank}(\mathbf{A})\,. \tag{4.4}$$

Assume first that $\mathbf{AGA} = \mathbf{A}$. Then of course $\mathbf{AGAG} = \mathbf{AG}$, so that $\mathbf{AG}$ is idempotent. Because

$$\operatorname{rank}(\mathbf{A}) = \operatorname{rank}(\mathbf{AGA}) \le \operatorname{rank}(\mathbf{AG}) \le \operatorname{rank}(\mathbf{A})\,, \tag{4.5}$$

we necessarily have $\operatorname{rank}(\mathbf{AG}) = \operatorname{rank}(\mathbf{A})$. To go the other way round, assume that $\mathbf{AG}$ is idempotent and $\operatorname{rank}(\mathbf{AG}) = \operatorname{rank}(\mathbf{A})$. Then, in view of the rank cancellation rule, Theorem 6 (p. 145), we can cancel the underlined terms from the following equation:

$$\mathbf{AGA\underline{G}} = \mathbf{A\underline{G}}\,. \tag{4.6}$$

Hence (c) implies (b2) and "(c) $\iff$ (b2)" is confirmed. The proof of "(c) $\iff$ (b1)" is quite analogous.

To prove "(a) $\iff$ (c)", let the vector $\mathbf{Gy}$ be a solution to (4.1) always when (4.1) is consistent, i.e.,

$$\mathbf{AGy} = \mathbf{y} \quad \text{for all } \mathbf{y} \in \mathscr{C}(\mathbf{A})\,, \tag{4.7}$$

i.e.,

$$\mathbf{AGAu} = \mathbf{Au} \quad \text{for all } \mathbf{u} \in \mathbb{R}^m, \tag{4.8}$$

which obviously is equivalent to (c). □

A generalized inverse is not necessarily unique—it always exists but it is unique if and only if $\mathbf{A}^{-1}$ exists. If $\mathbf{A}^{-1}$ exists, then pre- and postmultiplying (mp1) by $\mathbf{A}^{-1}$ we get $\mathbf{G} = \mathbf{A}^{-1}$.

The set of all generalized inverses of $\mathbf{A}$ is denoted as $\{\mathbf{A}^-\}$:

$$\{\mathbf{A}^-\} = \{\, \mathbf{G} : \mathbf{AGA} = \mathbf{A} \,\}\,. \tag{4.9}$$

Ben-Israel & Greville (2003) use the notation

$$\mathbf{A}\{1\} = \{\, \mathbf{G} : \mathbf{AGA} = \mathbf{A} \,\}, \tag{4.10}$$

and denote by $\mathbf{A}^{(1)}$ any member in $\mathbf{A}\{1\}$. Matrix $\mathbf{A}^{(1)}$ is also called a $\{1\}$-inverse or inner inverse.

We may emphasize the difference between the uniqueness of solution to $\mathbf{Ab} = \mathbf{y}$ and the uniqueness of $\mathbf{A}^-$:

- solution is unique $\iff$ $\mathbf{A}$ has full column rank,
- $\mathbf{A}^-$ is unique $\iff$ $\mathbf{A}$ is a nonsingular square matrix.

The existence of the generalized inverse can be easily proved using the full rank decomposition of $\mathbf{A}$; see Theorem 17 (p. 349):

$$\mathbf{A} = \mathbf{UV}'. \tag{FRD}$$

In (FRD) we have

$$\mathbf{U} \in \mathbb{R}^{n\times r},\ \mathbf{V} \in \mathbb{R}^{m\times r}, \quad \operatorname{rank}(\mathbf{U}) = \operatorname{rank}(\mathbf{V}) = r = \operatorname{rank}(\mathbf{A})\,. \tag{4.11}$$

Now

$$\mathbf{UV}' \cdot \mathbf{V}(\mathbf{V}'\mathbf{V})^{-1}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}' \cdot \mathbf{UV}' = \mathbf{UV}', \tag{4.12}$$

i.e.,

$$\mathbf{G} := \mathbf{V}(\mathbf{V}'\mathbf{V})^{-1}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}' \in \{\mathbf{A}^-\}\,. \tag{4.13}$$

It is easy to confirm that $\mathbf{G}$ defined in (4.13) satisfies, in addition to

$$\mathbf{AGA} = \mathbf{A}\,, \tag{mp1}$$

also the following three conditions:

$$\text{(mp2) } \mathbf{GAG} = \mathbf{G}\,, \quad \text{(mp3) } (\mathbf{AG})' = \mathbf{AG}\,, \quad \text{(mp4) } (\mathbf{GA})' = \mathbf{GA}\,. \tag{4.14}$$

Matrix $\mathbf{G}$ satisfying all four (mp$i$) conditions is called the *Moore–Penrose inverse* and it is denoted as $\mathbf{A}^+$. The matrix $\mathbf{A}^+$ appears to be unique.

Eliakim Hastings Moore defined a unique inverse or "general reciprocal" in his work which was published posthumously in 1935; the results were announced in 1920 in an abstract. Moore's[1] work appears to have been largely unnoticed until a resurgence of interest in the early 1950s centered on the use of generalized inverses in least-squares problems, which were not considered

---

[1] Here is a theorem from Moore (1935, p. 20):

$$\mathfrak{U}^C \mathfrak{B}^1\ {}^{\mathrm{II}}\mathfrak{B}^2\ {}^{\mathrm{II}}\kappa^{12} \cdot)\cdot$$

$$\exists \,|\, \lambda^{21\ \text{type}\ \mathfrak{M}^2_{\kappa^*}\bar{\mathfrak{M}}^1_{\kappa}} \ni \cdot S^2\kappa^{12}\lambda^{21} = \delta^{11}_{\mathfrak{M}^1_{\kappa}} \cdot S^1\lambda^{21}\kappa^{12} = \delta^{22}_{\mathfrak{M}^2_{\kappa^*}}\,.$$

by Moore. In 1955 Roger Penrose showed that Moore's generalized inverse is the unique matrix $\mathbf{G}$ that satisfies the four equations. For more about Moore, see Ben-Israel (2002) and Ben-Israel & Greville (2003, Appendix A).

As a simple example, consider the 2×2 matrix $\mathbf{A} = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right) = \mathbf{1}\mathbf{1}'$. Then we immediately see that

$$\mathbf{AGA} = \mathbf{1}\mathbf{1}'\mathbf{G}\mathbf{1}\mathbf{1}' = \mathbf{1}\mathbf{1}' \iff \mathbf{1}'\mathbf{G}\mathbf{1} = 1 \tag{4.15a}$$

$$\iff$$

$$\mathbf{G} = \begin{pmatrix} a & b \\ c & 1-a-b-c \end{pmatrix}, \quad \text{where } a,\ b,\ c \text{ are free to vary.} \tag{4.15b}$$

Moreover, the Moore–Penrose is $\mathbf{A}^{+} = \frac{1}{4}\left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right) = \frac{1}{4}\mathbf{A}$.

If $\mathbf{G}$ satisfies the condition (mp$i$), we denote $\mathbf{G} = \mathbf{A}_{i}^{-}$, or $\mathbf{G} \in \{\mathbf{A}_{i}^{-}\}$, and if it satisfies the conditions i and j, we denote $\mathbf{G} = \mathbf{A}_{ij}^{-}$ or $\mathbf{G} \in \{\mathbf{A}_{ij}^{-}\}$ and it may be called an $\{ij\}$-inverse.

Let us prove the following:

$$\mathbf{G} \in \{\mathbf{A}_{13}^{-}\} \implies \mathbf{AG} \text{ is unique,} \tag{4.16a}$$

$$\mathbf{G} \in \{\mathbf{A}_{14}^{-}\} \implies \mathbf{GA} \text{ is unique.} \tag{4.16b}$$

Assume that $\mathbf{G}, \mathbf{F} \in \{\mathbf{A}_{13}^{-}\}$, and so

$$\mathbf{AGA} = \mathbf{AFA} = \mathbf{A}\,, \quad \mathbf{AG} = \mathbf{G}'\mathbf{A}', \quad \mathbf{AF} = \mathbf{F}'\mathbf{A}'. \tag{4.16c}$$

Hence

$$\begin{aligned} \mathbf{AG} &= (\mathbf{AFA})\mathbf{G} = (\mathbf{AF})(\mathbf{AG}) \\ &= \mathbf{F}'\mathbf{A}'\mathbf{G}'\mathbf{A}' = \mathbf{F}'(\mathbf{AGA})' = \mathbf{F}'\mathbf{A}' = \mathbf{AF}\,, \end{aligned} \tag{4.16d}$$

and so (4.16a) is proved. The proof of (4.16b) goes in the corresponding way.

If the columns of $\mathbf{A}$ are linearly independent, we immediately observe that

$$(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}' \in \{\mathbf{A}^{-}\}\,. \tag{4.17}$$

Actually $(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$ is the Moore–Penrose inverse of $\mathbf{A}$:

$$\mathbf{A}^{+} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}', \quad \text{when } \operatorname{rank}(\mathbf{A}_{n\times m}) = m\,. \tag{4.18}$$

The matrix $\mathbf{A}_{L} := (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$ is the left inverse of $\mathbf{A}$ in the sense that it satisfies the equation $\mathbf{A}_{L}\mathbf{A} = \mathbf{I}_{m}$. If the rows of $\mathbf{A}$ are linearly independent, then

$$\mathbf{A}^{+} = \mathbf{A}'(\mathbf{A}\mathbf{A}')^{-1}, \quad \text{when } \operatorname{rank}(\mathbf{A}_{n\times m}) = n\,. \tag{4.19}$$

Then $\mathbf{A}_{R} := \mathbf{A}'(\mathbf{A}\mathbf{A}')^{-1}$ is the right inverse of $\mathbf{A}$. It is also easy to confirm that

$$(\mathbf{A}'\mathbf{A})^{+}\mathbf{A}' = \mathbf{A}'(\mathbf{A}\mathbf{A}')^{+} = \mathbf{A}^{+}. \tag{4.20}$$

In particular, if $\mathbf{A}$ is a nonzero column vector $\mathbf{a}$, then

$$\mathbf{a}^+ = (\mathbf{a}'\mathbf{a})^{-1}\mathbf{a}' = \frac{1}{\mathbf{a}'\mathbf{a}}\mathbf{a}' \text{ (= row vector)}, \quad \mathbf{a} \in \mathbb{R}^n, \quad \mathbf{a} \neq \mathbf{0}\,. \tag{4.21}$$

Correspondingly, if $\mathbf{A}$ is a row vector $\mathbf{a}'$, then

$$(\mathbf{a}')^+ = \mathbf{a}(\mathbf{a}'\mathbf{a})^{-1} = \frac{1}{\mathbf{a}'\mathbf{a}}\mathbf{a} \text{ (= column vector)}, \quad \mathbf{a} \in \mathbb{R}^n, \quad \mathbf{a} \neq \mathbf{0}\,. \tag{4.22}$$

What about the generalized inverse of the null matrix? It is obvious that any conformable matrix is appropriate for $\mathbf{0}^-$, and almost equally obvious that $\mathbf{0}^+ = \mathbf{0}$. Hence for a real number λ, $\lambda^+ = 1/\lambda$ if $\lambda \neq 0$ and $\lambda^+ = 0$ if $\lambda = 0$. It is also obvious that for the diagonal matrix we have

$$\begin{pmatrix} d_1 & 0 & \dots & 0 \\ 0 & d_2 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & d_n \end{pmatrix}^+ = \begin{pmatrix} d_1^+ & 0 & \dots & 0 \\ 0 & d_2^+ & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & d_n^+ \end{pmatrix}. \tag{4.23}$$

We observe that the rank of $\mathbf{A}^-$ is always greater than or equal to the rank of $\mathbf{A}$:

$$\operatorname{rank}(\mathbf{A}^-) \geq \operatorname{rank}(\mathbf{A})\,. \tag{4.24}$$

This follows from the following:

$$\operatorname{rank}(\mathbf{A}^-) \geq \operatorname{rank}(\mathbf{A}\mathbf{A}^-) \geq \operatorname{rank}(\mathbf{A}\mathbf{A}^-\mathbf{A}) = \operatorname{rank}(\mathbf{A})\,. \tag{4.25}$$

4.1 Generalized Inverse & the Singular Value Decomposition

We shall return to the connection between generalized inverse and singular value decomposition in Chapter 19 (p. 391), but it is appropriate to have a quick look at this useful relation right here.

Let $\mathbf{A}_{n\times m}$ have a singular value decomposition

$$\mathbf{A} = \mathbf{U}\begin{pmatrix} \mathbf{\Delta}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{V}', \tag{4.26}$$

where $\mathbf{U}_{n\times n}$ and $\mathbf{V}_{m\times m}$ are orthogonal and $\mathbf{\Delta}_1 = \operatorname{diag}(\delta_1, \dots, \delta_r)$, δ_i being the ith largest singular value of $\mathbf{A}$ and $\operatorname{rank}(\mathbf{A}) = r$. Then the following statements hold:

$$\mathbf{G} \in \{\mathbf{A}^-\} \iff \mathbf{G} = \mathbf{V}\begin{pmatrix} \mathbf{\Delta}_1^{-1} & \mathbf{K} \\ \mathbf{L} & \mathbf{N} \end{pmatrix}\mathbf{U}', \tag{4.27}$$

$$\mathbf{G} = \mathbf{A}^+ \iff \mathbf{G} = \mathbf{V}\begin{pmatrix} \mathbf{\Delta}_1^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{U}' = \mathbf{V}_1\mathbf{\Delta}_1^{-1}\mathbf{U}_1', \tag{4.28}$$

where **K**, **L**, and **N** are arbitrary matrices (of conformable dimensions). Hence by varying the elements of **K**, **L** and **N** we can generate all generalized inverses of **A**.

The proof of (4.28) goes straightforwardly by checking all four Moore–Penrose conditions. It is worth noting that (4.28) immediately implies a couple of very useful results. First,

$$\mathscr{C}(\mathbf{A}^+) = \mathscr{C}(\mathbf{V}_1) = \mathscr{C}(\mathbf{A}'), \tag{4.29}$$

and so the columns of $\mathbf{V}_1$ form an orthonormal basis for $\mathscr{C}(\mathbf{A}')$; recall that columns of $\mathbf{U}_1$ form an orthonormal basis for $\mathscr{C}(\mathbf{A})$. Secondly, we get the extremely useful expressions for orthogonal projectors $\mathbf{P}_{\mathbf{A}}$ and $\mathbf{P}_{\mathbf{A}'}$:

$$\mathbf{A}\mathbf{A}^+ = \mathbf{U}_1\mathbf{U}_1' = \mathbf{P}_{\mathbf{A}}, \quad \mathbf{A}^+\mathbf{A} = \mathbf{V}_1\mathbf{V}_1' = \mathbf{P}_{\mathbf{A}'}. \tag{4.30}$$

To prove (4.27), assume first that **G** can be expressed as

$$\mathbf{G} = \mathbf{V}\begin{pmatrix} \mathbf{\Delta}_1^{-1} & \mathbf{K} \\ \mathbf{L} & \mathbf{N} \end{pmatrix}\mathbf{U}'. \tag{4.31}$$

Then

$$\begin{aligned} \mathbf{AGA} &= \mathbf{U\Delta V}' \cdot \mathbf{V}\begin{pmatrix} \mathbf{\Delta}_1^{-1} & \mathbf{K} \\ \mathbf{L} & \mathbf{N} \end{pmatrix}\mathbf{U}' \cdot \mathbf{U\Delta V}' \\ &= \mathbf{U}\begin{pmatrix} \mathbf{\Delta}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\begin{pmatrix} \mathbf{\Delta}_1^{-1} & \mathbf{K} \\ \mathbf{L} & \mathbf{N} \end{pmatrix}\begin{pmatrix} \mathbf{\Delta}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{V}' = \mathbf{U}\begin{pmatrix} \mathbf{\Delta}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{V}' = \mathbf{A}. \end{aligned} \tag{4.32}$$

Hence the part "$\Longleftarrow$" of the proof of (4.27) is done. Assume then that there exists a matrix **G** such that $\mathbf{AGA} = \mathbf{A}$, i.e.,

$$\mathbf{U\Delta V}' \cdot \mathbf{G} \cdot \mathbf{U\Delta V}' = \mathbf{U\Delta V}'. \tag{4.33}$$

Premultipying (4.33) by $\mathbf{U}'$ and postmultiplying it by **V** yields

$$\mathbf{\Delta V}'\mathbf{GU\Delta} = \mathbf{\Delta}, \tag{4.34}$$

i.e.,

$$\begin{pmatrix} \mathbf{\Delta}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\begin{pmatrix} \mathbf{F}_{11} & \mathbf{F}_{12} \\ \mathbf{F}_{21} & \mathbf{F}_{22} \end{pmatrix}\begin{pmatrix} \mathbf{\Delta}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{\Delta}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \tag{4.35}$$

where

$$\mathbf{V}'\mathbf{GU} := \begin{pmatrix} \mathbf{F}_{11} & \mathbf{F}_{12} \\ \mathbf{F}_{21} & \mathbf{F}_{22} \end{pmatrix}. \tag{4.36}$$

Equation (4.35) implies that $\mathbf{\Delta}_1\mathbf{F}_{11}\mathbf{\Delta}_1 = \mathbf{\Delta}_1$, i.e., $\mathbf{F}_{11} = \mathbf{\Delta}_1^{-1}$. Premultiplying

$$\mathbf{V}'\mathbf{G}\mathbf{U} = \begin{pmatrix} \mathbf{\Delta}_1^{-1} & \mathbf{F}_{12} \\ \mathbf{F}_{21} & \mathbf{F}_{22} \end{pmatrix} \tag{4.37}$$

by $\mathbf{V}$ and postmultiplying it by $\mathbf{U}'$, leads to

$$\mathbf{G} = \mathbf{V} \begin{pmatrix} \mathbf{\Delta}_1^{-1} & \mathbf{F}_{12} \\ \mathbf{F}_{21} & \mathbf{F}_{22} \end{pmatrix} \mathbf{U}'. \tag{4.38}$$

So: since the equation $\mathbf{AGA} = \mathbf{A}$ implies that $\mathbf{G}$ has the representation (4.38), the result (4.27) is proved.

The result (4.27) describes concretely the lack of uniqueness of the generalized inverse: the matrices $\mathbf{K}$, $\mathbf{L}$ and $\mathbf{N}$ have $nm - r^2$ elements and they can be chosen arbitrarily. It is easy to believe, e.g., that we can choose these elements in such a manner that $\mathbf{A}^-$ has a given rank k:

$$\operatorname{rank}(\mathbf{A}) \leq k = \operatorname{rank}(\mathbf{A}^-) \leq \min(n, m)\,. \tag{4.39}$$

In addition to (4.27), we can give the following general representations of a generalized inverse:

Proposition 4.1. *Two alternative representations of a general solution to a generalized inverse of* $\mathbf{A}$ *are*

(a) $\mathbf{G} = \mathbf{A}^- + \mathbf{U} - \mathbf{A}^-\mathbf{AUAA}^-$,

(b) $\mathbf{G} = \mathbf{A}^- + \mathbf{V}(\mathbf{I}_n - \mathbf{AA}^-) + (\mathbf{I}_m - \mathbf{A}^-\mathbf{A})\mathbf{W}$,

where $\mathbf{A}^-$ *is a particular generalized inverse and* $\mathbf{U}$, $\mathbf{V}$, $\mathbf{W}$ *are free to vary. In particular, choosing* $\mathbf{A}^- = \mathbf{A}^+$*, the general representations can be expressed as*

(c) $\mathbf{G} = \mathbf{A}^+ + \mathbf{U} - \mathbf{P}_{\mathbf{A}'}\mathbf{U}\mathbf{P}_{\mathbf{A}}$,

(d) $\mathbf{G} = \mathbf{A}^+ + \mathbf{V}(\mathbf{I}_n - \mathbf{P}_{\mathbf{A}}) + (\mathbf{I}_m - \mathbf{P}_{\mathbf{A}'})\mathbf{W}$,

(e) $\mathbf{G} = \mathbf{A}^+ + \mathbf{V}\mathbf{A}^{\perp} + (\mathbf{A}')^{\perp}\mathbf{W}$.

For the proof of Proposition 4.1, see Theorem 11 (p. 267), Rao & Mitra (1971b, Th. 2.4.1), and Ben-Israel & Greville (2003, p. 52).

If $\mathbf{A}_{n\times n}$ is a symmetric nonnegative definite matrix and $r = \operatorname{rank}(\mathbf{A})$, then its eigenvalue decomposition is the same as its singular value decomposition:

$$\begin{aligned} \mathbf{A} = \mathbf{T}\mathbf{\Lambda}\mathbf{T}' = (\mathbf{T}_1 : \mathbf{T}_0) \begin{pmatrix} \mathbf{\Lambda}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{T}_1' \\ \mathbf{T}_0' \end{pmatrix} = \mathbf{T}_1\mathbf{\Lambda}_1\mathbf{T}_1' \\ = \lambda_1\mathbf{t}_1\mathbf{t}_1' + \cdots + \lambda_r\mathbf{t}_r\mathbf{t}_r'\,, \end{aligned} \tag{EVD}$$

where $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$, $\mathbf{\Lambda}_1 = \operatorname{diag}(\lambda_1, \dots, \lambda_r)$, and $\lambda_1 \geq \cdots \geq \lambda_r > 0 = \lambda_{r+1} = \cdots = \lambda_n$ are the ordered eigenvalues of $\mathbf{A}$. The columns of $\mathbf{T}$ are the corresponding orthonormal eigenvectors of $\mathbf{A}$. Then

$$\mathbf{A}^+ = \mathbf{T}\boldsymbol{\Lambda}^+\mathbf{T}' = (\mathbf{T}_1 : \mathbf{T}_0)\begin{pmatrix}\boldsymbol{\Lambda}_1^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{T}_1'\\ \mathbf{T}_0'\end{pmatrix} = \mathbf{T}_1\boldsymbol{\Lambda}_1^{-1}\mathbf{T}_1'$$
$$= \frac{1}{\lambda_1}\mathbf{t}_1\mathbf{t}_1' + \cdots + \frac{1}{\lambda_r}\mathbf{t}_r\mathbf{t}_r' \,. \tag{4.40}$$

Photograph 4.1 Götz Trenkler (Tampere, 2007).

If $\mathbf{A}$ is symmetric but not nonnegative definite, then some eigenvalues λ_i are negative but still (4.40) is the Moore–Penrose inverse of $\mathbf{A}$. However, the (EVD) is not anymore the singular value decomposition of $\mathbf{A}$. What small changes you have then to do in (EVD) to get an SVD of $\mathbf{A}$?

We complete this section with a matrix decomposition of Hartwig & Spindelböck (1984), heavily used in a series of papers by Oskar Maria Baksalary and Götz Trenkler, see, e.g., Baksalary, Styan & Trenkler (2009), and Trenkler (2006).

Proposition 4.2. *Let* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *be of rank* r. *Then there exists an orthogonal* $\mathbf{U} \in \mathbb{R}^{n\times n}$ *such that*

$$\mathbf{A} = \mathbf{U}\begin{pmatrix}\boldsymbol{\Delta}\mathbf{K} & \boldsymbol{\Delta}\mathbf{L}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{U}', \tag{4.41}$$

where $\boldsymbol{\Delta} = \operatorname{diag}(\delta_1\mathbf{I}_{r_1}, \dots, \delta_t\mathbf{I}_{r_t})$ *is the diagonal matrix of singular values of* $\mathbf{A}$, $\delta_1 > \delta_2 > \cdots > \delta_t > 0$, $r_1 + r_2 + \cdots + r_t = r$, *and* $\mathbf{K} \in \mathbb{R}^{r\times r}$, $\mathbf{L} \in \mathbb{R}^{r\times(n-r)}$ *satisfy*

$$\mathbf{K}\mathbf{K}' + \mathbf{L}\mathbf{L}' = \mathbf{I}_r \,. \tag{4.42}$$

4.2 Generalized Inverse & Oblique Projector

We recall that the idempotent matrix $\mathbf{P}$ has the following property:

$$\mathbf{P}^2 = \mathbf{P} \implies \mathbf{P} \text{ is the oblique projector onto } \mathscr{C}(\mathbf{P}) \text{ along } \mathscr{N}(\mathbf{P})\,, \tag{4.43}$$

where the direction space can be also written as

$$\mathscr{N}(\mathbf{P}) = \mathscr{C}(\mathbf{I} - \mathbf{P})\,. \tag{4.44}$$

The matrix product $\mathbf{A}\mathbf{A}^-$ is always (for any choice of $\mathbf{A}^-$) idempotent and hence $\mathbf{A}\mathbf{A}^-$ is the oblique projector onto $\mathscr{C}(\mathbf{A}\mathbf{A}^-)$ along $\mathscr{N}(\mathbf{A}\mathbf{A}^-)$. Since

$$\mathscr{C}(\mathbf{A}\mathbf{A}^-) \subset \mathscr{C}(\mathbf{A}) \text{ and } \operatorname{rank}(\mathbf{A}\mathbf{A}^-) = \operatorname{rank}(\mathbf{A})\,, \tag{4.45}$$

we have

$$\mathscr{C}(\mathbf{A}\mathbf{A}^-) = \mathscr{C}(\mathbf{A})\,, \tag{4.46}$$

and hence

$$\mathbf{A}\mathbf{A}^- \text{ is the oblique projector onto } \mathscr{C}(\mathbf{A}) \text{ along } \mathscr{N}(\mathbf{A}\mathbf{A}^-)\,, \tag{4.47}$$

where the direction space can be written as

$$\mathscr{N}(\mathbf{A}\mathbf{A}^-) = \mathscr{C}(\mathbf{I}_n - \mathbf{A}\mathbf{A}^-)\,. \tag{4.48}$$

Correspondingly,

$$\mathbf{A}^-\mathbf{A} \text{ is the oblique projector onto } \mathscr{C}(\mathbf{A}^-\mathbf{A}) \text{ along } \mathscr{N}(\mathbf{A}^-\mathbf{A})\,. \tag{4.49}$$

The space $\mathscr{N}(\mathbf{A}^-\mathbf{A})$ can be expressed also as

$$\mathscr{N}(\mathbf{A}^-\mathbf{A}) = \mathscr{C}(\mathbf{I}_m - \mathbf{A}^-\mathbf{A})\,. \tag{4.50}$$

Because

$$\mathscr{N}(\mathbf{A}) \subset \mathscr{N}(\mathbf{A}^-\mathbf{A}) \subset \mathscr{N}(\mathbf{A}\mathbf{A}^-\mathbf{A}) = \mathscr{N}(\mathbf{A})\,, \tag{4.51}$$

the nullspace $\mathscr{N}(\mathbf{A})$ can be expressed as

$$\mathscr{N}(\mathbf{A}_{n\times m}) = \mathscr{C}(\mathbf{I}_m - \mathbf{A}^-\mathbf{A})\,. \tag{4.52}$$

The symmetry of $\mathbf{A}\mathbf{A}^+$ and $\mathbf{A}^+\mathbf{A}$ implies the fundamental results

$$\mathbf{A}\mathbf{A}^+ = \mathbf{P}_{\mathbf{A}}\,, \quad \mathbf{A}^+\mathbf{A} = \mathbf{P}_{\mathbf{A}'}\,. \tag{4.53}$$

Note that we can express $\mathscr{C}(\mathbf{A})^\perp$ in terms of $\mathbf{A}^-$:

$$\mathscr{C}(\mathbf{A})^\perp = \mathscr{C}[\mathbf{I}_n - (\mathbf{A}')^-\mathbf{A}'] = \mathscr{C}[\mathbf{I}_n - (\mathbf{A}^-)'\mathbf{A}']\,. \tag{4.54}$$

This comes from

$$\mathscr{C}(\mathbf{A})^\perp = \mathscr{N}(\mathbf{A}')\,. \tag{4.55}$$

Using (4.52) and (4.55) we obtain the first equality in (4.54) at once. The second one follows from the fact

$$\{(\mathbf{A}^-)'\} = \{(\mathbf{A}')^-\}\,. \tag{4.56}$$

A gentle warning may take a place regarding (4.56): it is not fully correct to write $(\mathbf{A}^-)' = (\mathbf{A}')^-$ because (4.56) means the equality between two *sets*. However, we do have always

$$(\mathbf{A}^+)' = (\mathbf{A}')^+\,. \tag{4.57}$$

We have denoted

$$\mathbf{A}^{\perp} = \text{ any matrix such that } \mathscr{C}(\mathbf{A}^{\perp}) = \mathscr{N}(\mathbf{A}') \,. \tag{4.58}$$

We have, for example, the following choices for $\mathbf{A}^{\perp}$:

$$\mathbf{I}_n - (\mathbf{A}')^{-}\mathbf{A}' \in \{\mathbf{A}^{\perp}\} \,, \qquad \mathbf{I}_n - (\mathbf{A}^{-})'\mathbf{A}' \in \{\mathbf{A}^{\perp}\} \,. \tag{4.59}$$

Replacing $(\mathbf{A}')^{-}$ with $(\mathbf{A}')^{+}$ in (4.58) and using

$$(\mathbf{A}')^{+} = (\mathbf{A}^{+})', \quad (\mathbf{A}')^{+}\mathbf{A}' = \mathbf{A}\mathbf{A}^{+}, \tag{4.60}$$

we get

$$\mathbf{I}_n - \mathbf{A}\mathbf{A}^{+} = \mathbf{I}_n - \mathbf{P}_{\mathbf{A}} \in \{\mathbf{A}^{\perp}\} \,. \tag{4.61}$$

For the geometry of generalized inverses, the reader is referred e.g., to Kruskal (1975), Rao & Yanai (1985a), and Rao (1985b).

4.3 Generalized Inverse & Solutions to a Linear Equation

As stated earlier, if the equation $\mathbf{Ab} = \mathbf{y}$ is consistent, then one solution is $\mathbf{A}^{-}\mathbf{y}$. According to (0.11) (p. 5), the general solution $\mathbf{b}_0$, say, to a consistent $\mathbf{Ab} = \mathbf{y}$ can be expressed as

$$\mathbf{b}_0 = \{\text{one solution to } \mathbf{Ab} = \mathbf{y}\} + \{\text{the general solution to } \mathbf{Ab} = \mathbf{0}\} \,, \tag{4.62}$$

i.e.,

$$\mathbf{b}_0 = \mathbf{A}^{-}\mathbf{y} + \mathbf{t} \,, \tag{4.63}$$

where $\mathbf{t}$ is an arbitrary solution to $\mathbf{At} = \mathbf{0}$, i.e., $\mathbf{t} \in \mathscr{N}(\mathbf{A}) = \mathscr{C}(\mathbf{I} - \mathbf{A}^{-}\mathbf{A})$. Hence the general solution is expressible as

$$\mathbf{b}_0 = \mathbf{A}^{-}\mathbf{y} + (\mathbf{I} - \mathbf{A}^{-}\mathbf{A})\mathbf{z} \,, \tag{4.64}$$

where $\mathbf{A}^{-}$ is a (fixed) generalized inverse of $\mathbf{A}$ and $\mathbf{z}$ is an arbitrary vector. When we vary $\mathbf{z}$, we obtain all solutions to $\mathbf{Ab} = \mathbf{y}$. On the other hand, the following result holds; see Rao & Mitra (1971b, p. 27).

Proposition 4.3. *The class of all solutions to a consistent equation* $\mathbf{Ab} = \mathbf{y}$ *is* $\{\mathbf{Gy}\}$*, where* $\mathbf{G}$ *varies through all generalized inverses of* $\mathbf{A}$.

We will return to the properties of the solutions to linear equations later on more thoroughly, see Theorem 11 (p. 267), but let us take a quick tour right now here.

The solution $\mathbf{A}^{+}\mathbf{y}$ has one particularly interesting feature: it is the solution which has minimal norm, i.e., it is the shortest solution. This is seen from the following. Let $\mathbf{b}_{*}$ be an arbitrary solution to $\mathbf{Ab} = \mathbf{y}$. Then

$$\begin{aligned}\|\mathbf{b}_*\|^2 &= \|\mathbf{A}^+\mathbf{y} + (\mathbf{I}_m - \mathbf{A}^+\mathbf{A})\mathbf{z}\|^2 \\ &= \|\mathbf{A}^+\mathbf{y}\|^2 + \|(\mathbf{I}_m - \mathbf{A}^+\mathbf{A})\mathbf{z}\|^2 + 2(\mathbf{A}^+\mathbf{y})'(\mathbf{I}_m - \mathbf{A}^+\mathbf{A})\mathbf{z} \\ &= \|\mathbf{A}^+\mathbf{y}\|^2 + \|(\mathbf{I}_m - \mathbf{A}^+\mathbf{A})\mathbf{z}\|^2 \\ &\geq \|\mathbf{A}^+\mathbf{y}\|^2 \qquad \text{for all } \mathbf{z} \in \mathbb{R}^m, \end{aligned} \tag{4.65}$$

which follows from the orthogonality of the vectors

$$\mathbf{A}^+\mathbf{y} := \mathbf{A}'\mathbf{u} \in \mathscr{C}(\mathbf{A}') \quad (\text{for some } \mathbf{u}) \tag{4.66}$$

and

$$(\mathbf{I}_m - \mathbf{A}^+\mathbf{A})\mathbf{z} = (\mathbf{I}_m - \mathbf{P}_{\mathbf{A}'})\mathbf{z} \in \mathscr{C}(\mathbf{A}')^\perp = \mathscr{N}(\mathbf{A})\,. \tag{4.67}$$

If $\mathbf{G}$ is such a generalized inverse which gives the minimum norm solution (using the standard Euclidean norm) to a consistent equation

$$\mathbf{A}\mathbf{b} = \mathbf{y} \tag{4.68}$$

for any $\mathbf{y}$ [which makes (4.68) consistent], then $\mathbf{G}$ is a *minimum norm g-inverse* of $\mathbf{A}$, and is denoted as $\mathbf{A}_m^-$. Let us denote such a generalized inverse as $\mathbf{G}$. Then $\mathbf{G} \in \{\mathbf{A}_m^-\}$ if and only if it satisfies the following inequality:

$$\|\mathbf{G}\mathbf{y}\| \leq \|\mathbf{G}\mathbf{y} + (\mathbf{I}_m - \mathbf{G}\mathbf{A})\mathbf{z}\| \quad \text{for all } \mathbf{y} \in \mathscr{C}(\mathbf{A}),\ \mathbf{z} \in \mathbb{R}^m, \tag{4.69}$$

i.e.,

$$\|\mathbf{G}\mathbf{A}\mathbf{t}\| \leq \|\mathbf{G}\mathbf{A}\mathbf{t} + (\mathbf{I}_m - \mathbf{G}\mathbf{A})\mathbf{z}\| \quad \text{for all } \mathbf{t} \in \mathbb{R}^n,\ \mathbf{z} \in \mathbb{R}^m, \tag{4.70}$$

In light of Proposition 2.1 (p. 76), the statement (4.70) is equivalent to

$$(\mathbf{G}\mathbf{A})' = (\mathbf{G}\mathbf{A})'\mathbf{G}\mathbf{A}\,, \tag{4.71}$$

which immediately implies the symmetry of $\mathbf{G}\mathbf{A}$:

$$(\mathbf{G}\mathbf{A})' = \mathbf{G}\mathbf{A}\,. \tag{4.72}$$

On the other hand, postmultiplying (4.72) by the idempotent matrix $\mathbf{G}\mathbf{A}$ yields (4.71). Hence we have shown that $\mathbf{G}$ is a minimum norm g-inverse of $\mathbf{A}$ if and only if

$$\mathbf{A}\mathbf{G}\mathbf{A} = \mathbf{A} \quad \text{and} \quad (\mathbf{G}\mathbf{A})' = \mathbf{G}\mathbf{A}\,, \tag{4.73}$$

i.e., $\mathbf{G}$ is a $\{14\}$-inverse of $\mathbf{A}$.

If the norm is defined as $\|\mathbf{x}\|_\mathbf{N}^2 = \mathbf{x}'\mathbf{N}\mathbf{x}$, where $\mathbf{N}$ is a nonnegative definite (possibly singular) matrix, then our task is to find such a generalized inverse $\mathbf{G}$ of $\mathbf{A}$ which satisfies

$$\|\mathbf{G}\mathbf{y}\|_\mathbf{N} = \min_{\mathbf{A}\mathbf{b}=\mathbf{y}} \|\mathbf{b}\|_\mathbf{N} \quad \text{for all } \mathbf{y} \in \mathscr{C}(\mathbf{A})\,, \tag{4.74a}$$

$$\mathbf{y}'\mathbf{G}'\mathbf{N}\mathbf{G}\mathbf{y} = \min_{\mathbf{A}\mathbf{b}=\mathbf{y}} \mathbf{b}'\mathbf{N}\mathbf{b} \quad \text{for all } \mathbf{y} \in \mathscr{C}(\mathbf{A})\,. \tag{4.74b}$$

In this situation, the condition (4.71) comes

$$(\mathbf{G}\mathbf{A})'\mathbf{N} = (\mathbf{G}\mathbf{A})'\mathbf{N}\mathbf{G}\mathbf{A}\,, \tag{4.75}$$

while (4.73) becomes

$$\mathbf{A}\mathbf{G}\mathbf{A} = \mathbf{A} \quad \text{and} \quad (\mathbf{G}\mathbf{A})'\mathbf{N} = \mathbf{N}\mathbf{G}\mathbf{A}\,, \tag{4.76}$$

or, equivalently,

$$(\mathbf{G}\mathbf{A})'\mathbf{A}' = \mathbf{A}' \quad \text{and} \quad (\mathbf{G}\mathbf{A})'\mathbf{N}(\mathbf{I}_m - \mathbf{G}\mathbf{A}) = \mathbf{0}\,. \tag{4.77}$$

Because $\mathscr{C}(\mathbf{I}_m - \mathbf{A}^-\mathbf{A}) = \mathscr{N}(\mathbf{A}) = \mathscr{C}(\mathbf{A}')^\perp$, we can rewrite (4.77) as

$$(\mathbf{G}\mathbf{A})'\mathbf{A}' = \mathbf{A}' \quad \text{and} \quad (\mathbf{G}\mathbf{A})'\mathbf{N}(\mathbf{A}')^\perp = \mathbf{0}\,, \tag{4.78}$$

which shows that

$$(\mathbf{G}\mathbf{A})' \in \{\mathbf{P}_{\mathbf{A}'|\mathbf{N}(\mathbf{A}')^\perp}\}\,. \tag{4.79}$$

Moreover, in view of (2.123) (p. 88),

$$(\mathbf{G}\mathbf{A})' \in \{\mathbf{P}_{\mathbf{A}';(\mathbf{N}+\mathbf{A}'\mathbf{A})^-}\}\,, \tag{4.80}$$

where $(\mathbf{N}+\mathbf{A}'\mathbf{A})^-$ is an arbitrary nonnegative definite generalized inverse of $\mathbf{N}+\mathbf{A}'\mathbf{A}$.

We may denote a matrix $\mathbf{G}$ satisfying (4.76) as $\mathbf{A}^-_{m(\mathbf{N})}$. If $\mathbf{N}$ is positive definite, then (4.79) shows that $\mathbf{Q} := (\mathbf{G}\mathbf{A})' = \mathbf{A}'\mathbf{G}'$ is the orthogonal projector onto $\mathscr{C}(\mathbf{A}')$ with respect to the inner product matrix $\mathbf{N}^{-1}$:

$$\mathbf{A}'\mathbf{G}' = \mathbf{P}_{\mathbf{A}';\mathbf{N}^{-1}} = \mathbf{A}'(\mathbf{A}\mathbf{N}^{-1}\mathbf{A}')^-\mathbf{A}\mathbf{N}^{-1}, \tag{4.81}$$

which is invariant with respect to the choice of $(\mathbf{A}\mathbf{N}^{-1}\mathbf{A}')^-$. One expression for $\mathbf{G}'$ is

$$\mathbf{G}' = (\mathbf{A}\mathbf{N}^{-1}\mathbf{A}')^-\mathbf{A}\mathbf{N}^{-1}, \tag{4.82}$$

and so one choice for $\mathbf{G} = \mathbf{A}^-_{m(\mathbf{N})}$ is, choosing $(\mathbf{A}\mathbf{N}^{-1}\mathbf{A}')^-$ symmetric,

$$\mathbf{G} = \mathbf{A}^-_{m(\mathbf{N})} = \mathbf{N}^{-1}\mathbf{A}'(\mathbf{A}\mathbf{N}^{-1}\mathbf{A}')^-. \tag{4.83}$$

If $\mathbf{N}$ is singular, then one choice for $\mathbf{A}^-_{m(\mathbf{N})}$ is

$$\mathbf{A}^-_{m(\mathbf{N})} = (\mathbf{N}+\mathbf{A}'\mathbf{A})^-\mathbf{A}'[\mathbf{A}(\mathbf{N}+\mathbf{A}'\mathbf{A})^-\mathbf{A}']^-. \tag{4.84}$$

This can be concluded from (4.80). Rao & Mitra (1971b, pp. 46–47) introduce (4.84) by first showing that the minimum of $\mathbf{b}'\mathbf{N}\mathbf{b}$ under the condition $\mathbf{A}\mathbf{b} =$

$\mathbf{y}$ is obtained for $\mathbf{b} = \mathbf{b}_*$ if and only if there exists a vector $\boldsymbol{\lambda}$ such that $\mathbf{b}_*$ is a solution to

$$\begin{pmatrix} \mathbf{N} & \mathbf{A}' \\ \mathbf{A} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{b}_* \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{y} \end{pmatrix} . \tag{4.85}$$

Solving $\mathbf{b}_*$ from (4.85) they arrive at (4.84).

Consider then the equation

$$\mathbf{A}\mathbf{b} = \mathbf{y} , \tag{4.86}$$

which has *no* solution. Let $\mathbf{G}$ be such a matrix that $\mathbf{G}\mathbf{y}$ is a least squares solution to (4.86) (for any $\mathbf{y}$), i.e.,

$$\|\mathbf{y} - \mathbf{A}\mathbf{G}\mathbf{y}\|^2 \leq \|\mathbf{y} - \mathbf{A}\mathbf{b}\|^2 \quad \text{for all } \mathbf{b} \in \mathbb{R}^m,\ \mathbf{y} \in \mathbb{R}^n. \tag{4.87}$$

Then $\mathbf{G}$ is a least squares g-inverse (it indeed appears to be a generalized inverse), and it is denoted as $\mathbf{A}^-_\ell$. If the inner product matrix in $\mathbb{R}^n$ is $\mathbf{V}$, then we use notation $\mathbf{A}^-_{\ell(\mathbf{V})}$.

The minimization problem in (4.87) is now fully exploited in Propositions 2.6 (p. 81) and 2.7 (p. 87), because the matrix $\mathbf{A}\mathbf{G}$ in (4.87) is nothing but the orthogonal projector onto $\mathscr{C}(\mathbf{A})$ when the inner product matrix is $\mathbf{V}$:

$$\mathbf{A}\mathbf{G} = \mathbf{P}_{\mathbf{A};\mathbf{V}} . \tag{4.88}$$

Hence one choice for $\mathbf{A}^-_{\ell(\mathbf{V})}$ is

$$\mathbf{A}^-_{\ell(\mathbf{V})} = (\mathbf{A}'\mathbf{V}\mathbf{A})^-\mathbf{A}'\mathbf{V}. \tag{4.89}$$

It is easy to confirm that for a positive definite $\mathbf{V}$ we have

$$\big\{[(\mathbf{A}')^-_{m(\mathbf{V})}]'\big\} = \big\{(\mathbf{A})^-_{\ell(\mathbf{V}^{-1})}\big\} . \tag{4.90}$$

For example, consider the estimable parametric function $\mathbf{k}'\boldsymbol{\beta}$ under the model $\{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \mathbf{V}\}$. Then there exists a vector $\mathbf{a} \in \mathbb{R}^n$ such that

$$\mathbf{X}'\mathbf{a} = \mathbf{k} . \tag{4.91}$$

In order to find the BLUE for $\mathbf{k}'\boldsymbol{\beta}$, we have to find the minimum norm solution (in $\mathbf{a}$) for (4.91). The solution is $\mathbf{a}_* := (\mathbf{X}')^-_{m(\mathbf{V})}\mathbf{k} := \mathbf{G}\mathbf{k}$, where

$$\mathbf{G} = (\mathbf{X}')^-_{m(\mathbf{V})} = (\mathbf{V} + \mathbf{X}\mathbf{X}')^-\mathbf{X}[\mathbf{X}'(\mathbf{V} + \mathbf{X}\mathbf{X}')^-\mathbf{X}]^-. \tag{4.92}$$

Furthermore,

$$\begin{aligned} \mathrm{BLUE}(\mathbf{k}'\boldsymbol{\beta}) &= \mathbf{a}_*'\mathbf{y} = \mathbf{k}'\mathbf{G}'\mathbf{y} = \mathbf{k}'[(\mathbf{X}')^-_{m(\mathbf{V})}]'\mathbf{y} \\ &= \mathbf{k}'[\mathbf{X}'(\mathbf{V} + \mathbf{X}\mathbf{X}')^-\mathbf{X}]^-\mathbf{X}'(\mathbf{V} + \mathbf{X}\mathbf{X}')^-\mathbf{y} . \end{aligned} \tag{4.93}$$

4.4 A Simple Example

As a simple example, let us consider the linear equation

$$b_2 = b_1 + 1\,, \tag{4.94}$$

which in matrix form is

$$\mathbf{A}\mathbf{b} = \mathbf{y}\,, \quad \mathbf{A} = \mathbf{a}' = (1, -1)\,, \quad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}, \quad \mathbf{y} = 1\,. \tag{4.95}$$

Let us denote the set of all solutions to (4.95) as

$$\mathcal{V} = \left\{ \mathbf{b} \in \mathbb{R}^2 : (1, -1) \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = 1 \right\} = \{ \mathbf{b} \in \mathbb{R}^2 : b_2 = b_1 - 1 \}\,. \tag{4.96}$$

The only nonzero singular value of $\mathbf{a}'$ is the square root of the eigenvalue of $\mathbf{a}'\mathbf{a}$, i.e., $\operatorname{sg}_1(\mathbf{a}') = \sqrt{2}$. Hence it is easy to conclude that the singular value decomposition of $\mathbf{A} = \mathbf{a}'$ is

$$\mathbf{a}' = (1, -1) = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}' = 1 \cdot (\sqrt{2}, 0) \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}. \tag{4.97}$$

Now the general expression for all the generalized inverses of $\mathbf{a}'$ is

$$(\mathbf{a}')^- = \mathbf{V} \begin{pmatrix} 1/\sqrt{2} \\ \alpha \end{pmatrix} = \begin{pmatrix} 1/2 & 1/2 \\ -1/2 & 1/2 \end{pmatrix} \begin{pmatrix} 1 \\ \alpha \end{pmatrix} = \begin{pmatrix} 1/2 \\ -1/2 \end{pmatrix} + \alpha \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}, \tag{4.98}$$

where α can be any real number.

Choosing $\alpha = 3$ in (4.96) yields the generalized inverse

$$(\mathbf{a}')^- = (1, -1)^- = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \tag{4.99}$$

and the corresponding solution is $\mathbf{b}_0 = \binom{2}{1} \cdot 1 = \binom{2}{1}$. Putting $\alpha = 0$ gives the Moore–Penrose inverse

$$(\mathbf{a}')^+ = \begin{pmatrix} 1/2 \\ -1/2 \end{pmatrix} = \mathbf{a}(\mathbf{a}'\mathbf{a})^{-1}, \tag{4.100}$$

which yields the solution $\mathbf{b}_1$ in the Figure 4.1: $\mathbf{b}_1 = \mathbf{A}^+\mathbf{y}$.

The nullspace of $\mathbf{A} = \mathbf{a}'$ is

$$\mathscr{N}(\mathbf{A}) = \{ \mathbf{b} \in \mathbb{R}^2 : \mathbf{A}\mathbf{b} = \mathbf{0} \} = \{ \mathbf{b} \in \mathbb{R}^2 : b_2 = -b_1 \} = \mathscr{C} \begin{pmatrix} 1 \\ -1 \end{pmatrix}. \tag{4.101}$$

Hence the set of all solutions to $\mathbf{A}\mathbf{b} = \mathbf{y}$ is

$$\mathcal{V} = \mathbf{b}_0 + \mathscr{N}(\mathbf{A}), \tag{4.102}$$

which is exactly the same conclusion as done above using the SVD.

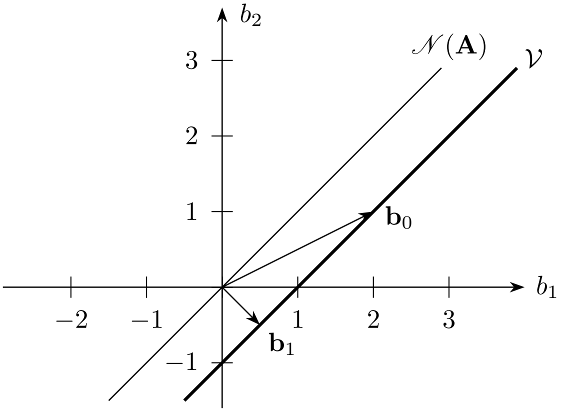


Figure 4.1 Set $\mathcal{V}$ is the set of all solutions to $\mathbf{Ab} = \mathbf{y}$. The vector $\mathbf{b}_1$ is the solution with the shortest norm: $\mathbf{b}_1 = \mathbf{A}^+\mathbf{y}$.

4.5 Exercises

4.1. Find all generalized inverses of $\mathbf{A} = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$. What is $\mathbf{A}^+$? What about the reflexive generalized inverse of $\mathbf{A}$? Characterize the set of symmetric nonnegative definite generalized inverses of $\mathbf{A}$.

4.2. Denote

$$\mathbf{A}_{n\times m} = \begin{pmatrix} \mathbf{B} & \mathbf{C} \\ \mathbf{D} & \mathbf{E} \end{pmatrix}, \quad \mathbf{G}_{m\times n} = \begin{pmatrix} \mathbf{B}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix},$$

where $\operatorname{rank}(\mathbf{B}_{r\times r}) = r = \operatorname{rank}(\mathbf{A})$. Show that $\mathbf{G} \in \{\mathbf{A}^-\}$.

4.3. Show that $(\mathbf{A}'\mathbf{A})^-\mathbf{A}'$ is a $\{123\}$-inverse of $\mathbf{A}$.

4.4. Prove: $\mathbf{AGA} = \mathbf{A}$ & $\operatorname{rank}(\mathbf{G}) = \operatorname{rank}(\mathbf{A}) \implies \mathbf{G} \in \left\{\mathbf{A}^-_{12}\right\}$, i.e., $\mathbf{G}$ is a reflexive generalized inverse of $\mathbf{A}$.

4.5. Confirm:

(a) $\begin{pmatrix} \mathbf{0}_{p\times n} \\ \mathbf{A}^- \end{pmatrix} \in \{(\mathbf{0}_{n\times p} : \mathbf{A}_{n\times m})^-\}$,

(b) $\det(\mathbf{P}) \neq 0, \det(\mathbf{Q}) \neq 0 \implies \mathbf{Q}^{-1}\mathbf{A}^-\mathbf{P}^{-1} \in \{(\mathbf{PAQ})^-\}, \mathbf{Q}^{-1}\mathbf{A}^+\mathbf{P}^{-1} = (\mathbf{PAQ})^+$.

4.6. Prove Propositions 4.1 (p. 111), 4.2 (p. 112), and 4.3 (p. 114).

4.7. Let $\mathbf{A}$ be symmetric but not nonnegative definite so that some of its eigenvalues are negative. Then (4.40) is the Moore–Penrose inverse of $\mathbf{A}$. However, the (EVD) is not anymore the singular value decomposition of $\mathbf{A}$. What changes you have then to do in (EVD) to get an SVD of $\mathbf{A}$?

4.8. Prove:

$$\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{V}\mathbf{X})^{+} = \mathbf{X}'\mathbf{V}(\mathbf{X}'\mathbf{V})^{+} = (\mathbf{X}'\mathbf{V}\mathbf{X})^{+}(\mathbf{X}'\mathbf{V}\mathbf{X}) = (\mathbf{V}\mathbf{X})^{+}\mathbf{V}\mathbf{X}\,.$$

4.9 (Multinomial distribution). Consider the vector $\mathbf{p} = (p_1, \dots, p_n)'$, where $p_i > 0$, $i = 1, \dots, n$, and $\mathbf{1}_n'\mathbf{p} = 1$, and denote

$$\boldsymbol{\Sigma} = \mathbf{D}_{\mathbf{p}} - \mathbf{p}\mathbf{p}', \quad \text{where } \mathbf{D}_{\mathbf{p}} = \operatorname{diag}(\mathbf{p})\,,$$

which is a double-centered matrix; $\boldsymbol{\Sigma}\mathbf{1}_n = \boldsymbol{\Sigma}'\mathbf{1}_n = \mathbf{0}$; see Exercises 0.8–0.11 (pp. 48–49). Confirm the following:

(a) $\boldsymbol{\Sigma} \in \mathrm{NND}_n$. Hint: write $\mathbf{A} = \left(\begin{smallmatrix}\mathbf{D}_{\mathbf{p}} & \mathbf{p} \\ \mathbf{p}' & 1\end{smallmatrix}\right)$ and use, e.g., (14.8) (p. 306).

(b) $\operatorname{rank}(\boldsymbol{\Sigma}) = n - 1$, $\mathscr{N}(\boldsymbol{\Sigma}) = \mathscr{C}(\mathbf{1}_n)$

(c) $\mathbf{D}_{\mathbf{p}}^{-1} \in \{\boldsymbol{\Sigma}^{-}\}$, and the general representation for $\boldsymbol{\Sigma}^{-}$ is

$$\boldsymbol{\Sigma}^{-} = \mathbf{D}_{\mathbf{p}}^{-1} + \mathbf{t}\mathbf{1}_n' + \mathbf{1}_n\mathbf{u}', \quad \boldsymbol{\Sigma}^{-} = \boldsymbol{\Sigma}^{+} + \mathbf{v}\mathbf{1}_n' + \mathbf{1}_n\mathbf{w}',$$

where the vectors $\mathbf{t}, \mathbf{u}, \mathbf{v}$ and $\mathbf{w}$ are free to vary. Show also that $\boldsymbol{\Sigma}^{+} = \mathbf{C}\boldsymbol{\Sigma}^{-}\mathbf{C}$, where $\mathbf{C}$ is the centering matrix and $\boldsymbol{\Sigma}^{-}$ is an arbitrary generalized inverse of $\boldsymbol{\Sigma}$.

Puntanen, Styan & Subak-Sharpe (1998), Styan & Subak-Sharpe (1997), Tanabe & Sagae (1992), Watson (1996b).

4.10. Prove that the following statements are equivalent:

(a) $\mathbf{A}\mathbf{Y} = \mathbf{0}$,

(b) $\mathscr{C}(\mathbf{Y}) \subset \mathscr{N}(\mathbf{A}) = \mathscr{C}(\mathbf{A}')^{\perp} = \mathscr{C}(\mathbf{I} - \mathbf{A}^{-}\mathbf{A})$,

(c) $\mathbf{Y} = (\mathbf{I} - \mathbf{A}^{-}\mathbf{A})\mathbf{L}$ for some $\mathbf{L}$,

and correspondingly, show the equivalence of the following statements:

(d) $\mathbf{X}\mathbf{A} = \mathbf{0}$,

(e) $\mathscr{C}(\mathbf{X}') \subset \mathscr{N}(\mathbf{A}') = \mathscr{C}(\mathbf{A})^{\perp} = \mathscr{C}[\mathbf{I} - (\mathbf{A}^{-})'\mathbf{A}']$,

(f) $\mathbf{X}' = [\mathbf{I} - (\mathbf{A}^{-})'\mathbf{A}']\mathbf{K}'$ for some $\mathbf{K}'$,

and hence the general solutions to $\mathbf{A}\mathbf{Y} = \mathbf{0}$ and $\mathbf{X}\mathbf{A} = \mathbf{0}$ are, respectively,

$$\mathbf{Y} = (\mathbf{I} - \mathbf{A}^{-}\mathbf{A})\mathbf{L} = (\mathbf{A}')^{\perp}\mathbf{L}\,, \quad \mathbf{X} = \mathbf{K}(\mathbf{I} - \mathbf{A}\mathbf{A}^{-}) = \mathbf{K}(\mathbf{A}^{\perp})',$$

where $\mathbf{L}$ and $\mathbf{K}$ are free to vary.

Chapter 5
Rank of the Partitioned Matrix and the Matrix Product

> *More than any other time in history, mankind faces a crossroads. One path leads to despair and utter hopelessness. The other, to total extinction. Let us pray we have wisdom to choose correctly.*
>
> Woody Allen: *My Speech to the Graduates*

There are numerous situations in the world of linear models and multivariate analysis when we need to find some appropriate expressions for the rank of the matrix product $\mathbf{AB}$, or of the partitioned matrix $(\mathbf{A} : \mathbf{B})$, for conformable matrices $\mathbf{A}$ and $\mathbf{B}$. Our favourite expressions are represented in the following theorem.

Theorem 5 (Ranks of $(\mathbf{A} : \mathbf{B})$ and $\mathbf{AB}$). *The rank of the partitioned matrix*

$$(\mathbf{A}_{n\times a} : \mathbf{B}_{n\times b}) \tag{5.1}$$

can be expressed, for any choice of generalized inverse $\mathbf{A}^-$, *as*

$$\begin{aligned} \operatorname{rank}(\mathbf{A} : \mathbf{B}) &= \operatorname{rank}[\mathbf{A} : (\mathbf{I}_n - \mathbf{A}\mathbf{A}^-)\mathbf{B}] \\ &= \operatorname{rank}(\mathbf{A}) + \operatorname{rank}[(\mathbf{I}_n - \mathbf{A}\mathbf{A}^-)\mathbf{B}] \\ &= \operatorname{rank}(\mathbf{A}) + \operatorname{rank}[(\mathbf{I}_n - \mathbf{A}\mathbf{A}^+)\mathbf{B}] \\ &= \operatorname{rank}(\mathbf{A}) + \operatorname{rank}[(\mathbf{I}_n - \mathbf{P}_{\mathbf{A}})\mathbf{B}] \,, \end{aligned} \tag{5.2}$$

while the rank of the matrix product $\mathbf{A}_{n\times a}\mathbf{B}_{a\times m}$ *is*

$$\begin{aligned} \operatorname{rank}(\mathbf{AB}) &= \operatorname{rank}(\mathbf{A}) - \dim \mathscr{C}(\mathbf{A}') \cap \mathscr{C}(\mathbf{B}^\perp) \\ &= \operatorname{rank}(\mathbf{A}) - \dim \mathscr{C}(\mathbf{A}') \cap \mathscr{N}(\mathbf{B}') \\ &= \operatorname{rank}(\mathbf{A}) - \dim \mathscr{C}(\mathbf{A}') \cap \mathscr{C}(\mathbf{I}_a - \mathbf{P}_{\mathbf{B}}) \,. \end{aligned} \tag{5.3}$$

Proof. Consider the matrices $\mathbf{A}_{n\times a}$ and $\mathbf{B}_{n\times b}$ and let $\mathbf{A}^- \in \mathbb{R}^{a\times n}$ be an arbitrary (but fixed) generalized inverse of $\mathbf{A}$. Denoting

$$\mathbf{L} = \begin{pmatrix} \mathbf{I}_a & -\mathbf{A}^-\mathbf{B} \\ \mathbf{0} & \mathbf{I}_b \end{pmatrix}, \tag{5.4}$$

we obtain

S. Puntanen et al., *Matrix Tricks for Linear Statistical Models: Our Personal Top Twenty*,
DOI 10.1007/978-3-642-10473-2_6,

$$\begin{aligned}(\mathbf{A} : \mathbf{B})\mathbf{L} &= (\mathbf{A} : \mathbf{B}) \begin{pmatrix} \mathbf{I}_a & -\mathbf{A}^{-}\mathbf{B} \\ \mathbf{0} & \mathbf{I}_b \end{pmatrix} \\ &= (\mathbf{A} : -\mathbf{A}\mathbf{A}^{-}\mathbf{B} + \mathbf{B}) = [\mathbf{A} : (\mathbf{I}_n - \mathbf{A}\mathbf{A}^{-})\mathbf{B}] \,. \end{aligned} \tag{5.5}$$

The matrix $\mathbf{L}$ is nonsingular and hence $(\mathbf{A} : \mathbf{B})$ and $(\mathbf{A} : \mathbf{B})\mathbf{L}$ have the same ranks:

$$\begin{aligned}\operatorname{rank}(\mathbf{A} : \mathbf{B}) &= \operatorname{rank}[(\mathbf{A} : \mathbf{B})\mathbf{L}] = \operatorname{rank}[\mathbf{A} : (\mathbf{I}_n - \mathbf{A}\mathbf{A}^{-})\mathbf{B}] \\ &= \operatorname{rank}(\mathbf{A}) + \operatorname{rank}[(\mathbf{I}_n - \mathbf{A}\mathbf{A}^{-})\mathbf{B}] \\ &\qquad - \dim \mathscr{C}(\mathbf{A}) \cap \mathscr{C}[(\mathbf{I}_n - \mathbf{A}\mathbf{A}^{-})\mathbf{B}] \,. \end{aligned} \tag{5.6}$$

We shall next show that the column spaces $\mathscr{C}(\mathbf{A})$ and $\mathscr{C}[(\mathbf{I}_n - \mathbf{A}\mathbf{A}^{-})\mathbf{B}]$ are disjoint, and thus (5.6) implies (5.2). If

$$\mathbf{u} \in \mathscr{C}(\mathbf{A}) \cap \mathscr{C}[(\mathbf{I}_n - \mathbf{A}\mathbf{A}^{-})\mathbf{B}] \,, \tag{5.7}$$

then $\mathbf{u}$ can be expressed as

$$\mathbf{u} = \mathbf{A}\mathbf{a} = (\mathbf{I}_n - \mathbf{A}\mathbf{A}^{-})\mathbf{B}\mathbf{b} \quad \text{for some } \mathbf{a} \in \mathbb{R}^a,\ \mathbf{b} \in \mathbb{R}^b. \tag{5.8}$$

Premultiplying (5.8) by $\mathbf{A}\mathbf{A}^{-}$ yields

$$\begin{aligned}\mathbf{A}\mathbf{A}^{-}\mathbf{u} &= \mathbf{A}\mathbf{A}^{-}\mathbf{A}\mathbf{a} = \mathbf{u} \\ &= \mathbf{A}\mathbf{A}^{-}(\mathbf{I}_n - \mathbf{A}\mathbf{A}^{-})\mathbf{B}\mathbf{b} = \mathbf{0} \,, \end{aligned} \tag{5.9}$$

and so indeed $\mathscr{C}(\mathbf{A})$ and $\mathscr{C}[(\mathbf{I}_n - \mathbf{A}\mathbf{A}^{-})\mathbf{B}]$ are disjoint.

Consider then the matrices $\mathbf{A}_{n\times a}$ and $\mathbf{B}_{a\times m}$ (which guarantees that the product $\mathbf{AB}$ is properly defined). To prove (5.3), we first write the obvious equation

$$\operatorname{rank}(\mathbf{B}^{\perp} : \mathbf{A}') = \operatorname{rank}(\mathbf{A}) + \operatorname{rank}(\mathbf{B}^{\perp}) - \dim \mathscr{C}(\mathbf{A}') \cap \mathscr{C}(\mathbf{B}^{\perp}) \,. \tag{5.10}$$

On the other hand, in view of (5.2), we have

$$\begin{aligned}\operatorname{rank}(\mathbf{B}^{\perp} : \mathbf{A}') &= \operatorname{rank}(\mathbf{B}^{\perp}) + \operatorname{rank}[(\mathbf{I}_a - \mathbf{P}_{\mathbf{B}^{\perp}})\mathbf{A}'] \\ &= \operatorname{rank}(\mathbf{B}^{\perp}) + \operatorname{rank}(\mathbf{P}_{\mathbf{B}}\mathbf{A}') \\ &= \operatorname{rank}(\mathbf{B}^{\perp}) + \operatorname{rank}(\mathbf{B}'\mathbf{A}') \\ &= \operatorname{rank}(\mathbf{B}^{\perp}) + \operatorname{rank}(\mathbf{AB}) \,, \end{aligned} \tag{5.11}$$

where we have used the fact that $\operatorname{rank}(\mathbf{P}_{\mathbf{B}}\mathbf{A}') = \operatorname{rank}(\mathbf{B}'\mathbf{A}')$; this becomes from

$$\operatorname{rank}(\mathbf{B}'\mathbf{A}') \geq \operatorname{rank}(\mathbf{P}_{\mathbf{B}}\mathbf{A}') \geq \operatorname{rank}(\mathbf{B}'\mathbf{P}_{\mathbf{B}}\mathbf{A}') = \operatorname{rank}(\mathbf{B}'\mathbf{A}') \,; \tag{5.12}$$

see also (5.38) (p. 127). Combining (5.10) and (5.11) yields (5.3). □

Notice that (5.2) obviously implies that

$$\operatorname{rank}\begin{pmatrix}\mathbf{A}_{n\times p}\\ \mathbf{B}_{m\times p}\end{pmatrix} = \operatorname{rank}(\mathbf{A}' : \mathbf{B}') = \operatorname{rank}(\mathbf{A}) + \operatorname{rank}[(\mathbf{I}_p - \mathbf{P}_{\mathbf{A}'})\mathbf{B}'] \\ = \operatorname{rank}\begin{pmatrix}\mathbf{A}\\ \mathbf{B}(\mathbf{I}_p - \mathbf{P}_{\mathbf{A}'})\end{pmatrix}. \tag{5.13}$$

Moreover, because

$$\operatorname{rank}(\mathbf{A}_{n\times a} : \mathbf{B}_{n\times b}) = \operatorname{rank}(\mathbf{A}) + \operatorname{rank}[(\mathbf{I}_n - \mathbf{P}_{\mathbf{A}})\mathbf{B}] \\ = \operatorname{rank}(\mathbf{A}) + \operatorname{rank}(\mathbf{B}) - \dim \mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B}), \tag{5.14}$$

the following statements concerning the disjointness of two column spaces are equivalent:

$$\mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B}) = \{\mathbf{0}\}, \tag{5.15a}$$

$$\operatorname{rank}(\mathbf{B}) = \operatorname{rank}[(\mathbf{I}_n - \mathbf{P}_{\mathbf{A}})\mathbf{B}], \tag{5.15b}$$

$$\mathscr{C}(\mathbf{B}') = \mathscr{C}[\mathbf{B}'(\mathbf{I}_n - \mathbf{P}_{\mathbf{A}})]. \tag{5.15c}$$

For a collection of the disjointness conditions, see Theorem 16 (p. 343). For the so-called matrix rank method introduced by Yongge Tian (dealing with the elementary block matrix operations), see, e.g., Tian (2007, 2009a), Tian & Takane (2008a,b, 2009b), and Tian & Wiens (2006).

References

(5.2): Marsaglia & Styan (1974a, Th. 4),

(5.3): Marsaglia & Styan (1974a, Cor. 6.2),
Rao (1973a, p. 28; First Edition 1965, p. 27),
Zyskind & Martin (1969, p. 1194),
Rao & Bhimasankaram (2000, Th. 3.5.11),
Rao & Rao (1998, p. 426).

5.1 Decomposition $\mathscr{C}(\mathbf{X} : \mathbf{V}) = \mathscr{C}(\mathbf{X} : \mathbf{V}\mathbf{X}^{\perp})$

When dealing with linear models we frequently need the column space decomposition represented in the following proposition:

Proposition 5.1. *Consider the linear model* $\{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$ *and denote* $\mathbf{M} = \mathbf{I}_n - \mathbf{H} = \mathbf{I}_n - \mathbf{P}_{\mathbf{X}}$. *Then*

$$\mathscr{C}(\mathbf{X} : \mathbf{V}) = \mathscr{C}(\mathbf{X} : \mathbf{V}\mathbf{M}) = \mathscr{C}(\mathbf{X} : \mathbf{V}\mathbf{X}^{\perp}) = \mathscr{C}(\mathbf{X}) \oplus \mathscr{C}(\mathbf{V}\mathbf{M}). \tag{5.16}$$

Proof. The proof of (5.16) is very easy using Theorem 5; cf. Rao (1974, Lemma 2.1). First, it is clear that

$$\mathscr{C}(\mathbf{X} : \mathbf{VM}) \subset \mathscr{C}(\mathbf{X} : \mathbf{V})\,. \tag{5.17}$$

The disjointness

$$\mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{VM}) = \{\mathbf{0}\} \tag{5.18}$$

can be proved by considering $\mathbf{u} \in \mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{VM})$, which implies

$$\mathbf{u} = \mathbf{Xa} = \mathbf{VMb} \quad \text{for some } \mathbf{a} \in \mathbb{R}^p,\ \mathbf{b} \in \mathbb{R}^n. \tag{5.19}$$

Premultiplying (5.19) by $\mathbf{M}$ yields

$$\mathbf{Mu} = \mathbf{MXa} = \mathbf{0} = \mathbf{MVMb}\,. \tag{5.20}$$

Obviously

$$\begin{aligned}\mathbf{MVMb} = \mathbf{0} &\implies \mathbf{b}'\mathbf{MVMb} = 0\\ &\implies \mathbf{V}^{1/2}\mathbf{Mb} = \mathbf{0} \implies \mathbf{VMb}\ (= \mathbf{u}) = \mathbf{0}\,,\end{aligned} \tag{5.21}$$

and hence the disjointness (5.18) is proved. In view of the disjointness and (5.2) we have

$$\begin{aligned}\operatorname{rank}(\mathbf{X} : \mathbf{V}) &= \operatorname{rank}(\mathbf{X}) + \operatorname{rank}(\mathbf{MV}) = \operatorname{rank}(\mathbf{X}) + \operatorname{rank}[(\mathbf{MV})']\\ &= \operatorname{rank}(\mathbf{X}) + \operatorname{rank}(\mathbf{VM}) = \operatorname{rank}(\mathbf{X} : \mathbf{VM})\,.\end{aligned} \tag{5.22}$$

Now (5.17), (5.18), and (5.22) imply our claim (5.16). □

As a related task we may show that

$$\begin{aligned}\operatorname{rank}\begin{pmatrix}\mathbf{V} & \mathbf{X}\\ \mathbf{X}' & \mathbf{0}\end{pmatrix} &= \operatorname{rank}\begin{pmatrix}\mathbf{V}\\ \mathbf{X}'\end{pmatrix} + \operatorname{rank}\begin{pmatrix}\mathbf{X}\\ \mathbf{0}\end{pmatrix}\\ &= \operatorname{rank}(\mathbf{X} : \mathbf{V}) + \operatorname{rank}(\mathbf{X})\,,\end{aligned} \tag{5.23}$$

and thereby

$$\mathscr{C}\begin{pmatrix}\mathbf{V}\\ \mathbf{X}'\end{pmatrix} \cap \mathscr{C}\begin{pmatrix}\mathbf{X}\\ \mathbf{0}\end{pmatrix} = \{\mathbf{0}\}\,. \tag{5.24}$$

Namely, according to (5.2) (p. 121),

$$\begin{aligned}\operatorname{rank}\begin{pmatrix}\mathbf{V} & \mathbf{X}\\ \mathbf{X}' & \mathbf{0}\end{pmatrix} &= \operatorname{rank}\begin{pmatrix}\mathbf{X}\\ \mathbf{0}\end{pmatrix} + \operatorname{rank}\left((\mathbf{I}_{n+p} - \mathbf{P}_{\left(\begin{smallmatrix}\mathbf{X}\\ \mathbf{0}\end{smallmatrix}\right)})\begin{pmatrix}\mathbf{V} & \mathbf{X}\\ \mathbf{X}' & \mathbf{0}\end{pmatrix}\right)\\ &= \operatorname{rank}(\mathbf{X}) + \operatorname{rank}\left(\begin{pmatrix}\mathbf{M} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_p\end{pmatrix}\begin{pmatrix}\mathbf{V} & \mathbf{X}\\ \mathbf{X}' & \mathbf{0}\end{pmatrix}\right)\\ &= \operatorname{rank}(\mathbf{X}) + \operatorname{rank}\begin{pmatrix}\mathbf{VM}\\ \mathbf{X}'\end{pmatrix}\end{aligned}$$

$$= \operatorname{rank}(\mathbf{X}) + \operatorname{rank}(\mathbf{X} : \mathbf{VM})\,. \tag{5.25}$$

Of course an alternative simple proof for (5.24) would be to choose $\mathbf{u}$ from the intersection $\mathscr{C}\big(\begin{smallmatrix}\mathbf{V}\\ \mathbf{X}'\end{smallmatrix}\big) \cap \mathscr{C}\big(\begin{smallmatrix}\mathbf{X}\\ \mathbf{0}\end{smallmatrix}\big)$ which implies that

$$\mathbf{u} := \begin{pmatrix}\mathbf{u}_1\\ \mathbf{u}_2\end{pmatrix} = \begin{pmatrix}\mathbf{Va}\\ \mathbf{X}'\mathbf{a}\end{pmatrix} = \begin{pmatrix}\mathbf{Xb}\\ \mathbf{0}_p\end{pmatrix} \quad \text{for some } \mathbf{a} \in \mathbb{R}^n,\ \mathbf{b} \in \mathbb{R}^p. \tag{5.26}$$

Hence $\mathbf{u}_2 = \mathbf{0}_p$ and now $\mathbf{a}'\mathbf{Va} = \mathbf{a}'\mathbf{Xb} = 0$ implies that $\mathbf{u}_1 = \mathbf{Va} = \mathbf{0}_n$.

5.2 Consistency Condition $\mathbf{y} \in \mathscr{C}(\mathbf{X} : \mathbf{V})$

We recall that the consistency condition (0.222) (p. 43) of the linear model $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$ can be expressed as

$$\mathbf{y} \in \mathscr{C}(\mathbf{X} : \mathbf{V}) = \mathscr{C}(\mathbf{X} : \mathbf{VM})\,. \tag{5.27}$$

The condition (5.27) can be obtained by considering the random variable

$$\mathbf{Q}_{(\mathbf{X}:\mathbf{V})}\mathbf{y}\,, \quad \text{where } \mathbf{Q}_{(\mathbf{X}:\mathbf{V})} = \mathbf{I}_n - \mathbf{P}_{(\mathbf{X}:\mathbf{V})}\,. \tag{5.28}$$

Now

$$\mathrm{E}(\mathbf{Q}_{(\mathbf{X}:\mathbf{V})}\mathbf{y}) = (\mathbf{I}_n - \mathbf{P}_{(\mathbf{X}:\mathbf{V})})\mathbf{X} = \mathbf{0}\,, \tag{5.29a}$$

$$\operatorname{cov}(\mathbf{Q}_{(\mathbf{X}:\mathbf{V})}\mathbf{y}) = \mathbf{Q}_{(\mathbf{X}:\mathbf{V})}\mathbf{V}\mathbf{Q}_{(\mathbf{X}:\mathbf{V})} = \mathbf{0}\,, \tag{5.29b}$$

and hence, cf. (1.38) (p. 63), $(\mathbf{I}_n - \mathbf{P}_{(\mathbf{X}:\mathbf{V})})\mathbf{y} = \mathbf{0}$ with probability 1, and so (5.27) holds with probability 1.

It is noteworthy that if the model $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$ is correct, i.e., it really holds (so to say), then necessarily (5.27) holds with probability 1. Hence the phrase "assume that (5.27) holds with probability 1" is actually unnecessary if we assume that the model $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$ is correct. However, it is a common practice in the literature of linear models to use phrases like "consider the model $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$ and assume that it is consistent". This is best understood in such a manner that the observed (realized) value of the random vector $\mathbf{y}$ satisfies (5.27). For a discussion concerning the consistency concept, see, e.g., Puntanen & Styan (1990), Christensen (1990), Farebrother (1990), Harville (1990b), and Baksalary, Rao & Markiewicz (1992).

As stated on p. 43 of the Introduction, the consistency condition means, for example, that whenever we have some statements where the random vector $\mathbf{y}$ is involved, these statements need to hold only for those values of $\mathbf{y}$ which belong to $\mathscr{C}(\mathbf{X} : \mathbf{V})$. Thus, two estimators $\mathbf{Gy}$ and $\mathbf{Fy}$ are equal if and only if

$$\mathbf{Gy} = \mathbf{Fy} \quad \text{for all } \mathbf{y} \in \mathscr{C}(\mathbf{X} : \mathbf{V})\,, \tag{5.30a}$$

that is,

$$\mathbf{G}(\mathbf{X}\mathbf{a} + \mathbf{V}\mathbf{M}\mathbf{b}) = \mathbf{F}(\mathbf{X}\mathbf{a} + \mathbf{V}\mathbf{M}\mathbf{b}) \quad \text{for all } \mathbf{a} \in \mathbb{R}^p,\ \mathbf{b} \in \mathbb{R}^n. \tag{5.30b}$$

5.3 Rank of a Partitioned Nonnegative Definite Matrix

Consider a nonnegative definite matrix $\mathbf{A} = \mathbf{L}'\mathbf{L}$ such that

$$\mathbf{A} = \begin{pmatrix} \mathbf{L}_1' \\ \mathbf{L}_2' \end{pmatrix} (\mathbf{L}_1 : \mathbf{L}_2) = \begin{pmatrix} \mathbf{L}_1'\mathbf{L}_1 & \mathbf{L}_1'\mathbf{L}_2 \\ \mathbf{L}_2'\mathbf{L}_1 & \mathbf{L}_2'\mathbf{L}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}. \tag{5.31}$$

Now, in light of (5.2), we can express the rank of $\mathbf{A}$ as follows:

$$\begin{aligned} \operatorname{rank}(\mathbf{A}) &= \operatorname{rank}(\mathbf{L}_1 : \mathbf{L}_2) \\ &= \operatorname{rank}(\mathbf{L}_1) + \operatorname{rank}[(\mathbf{I} - \mathbf{P}_{\mathbf{L}_1})\mathbf{L}_2] \\ &= \operatorname{rank}(\mathbf{L}_1'\mathbf{L}_1) + \operatorname{rank}[\mathbf{L}_2'(\mathbf{I} - \mathbf{P}_{\mathbf{L}_1})\mathbf{L}_2] \\ &= \operatorname{rank}(\mathbf{A}_{11}) + \operatorname{rank}(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-}\mathbf{A}_{12}) \\ &= \operatorname{rank}(\mathbf{A}_{11}) + \operatorname{rank}(\mathbf{A}_{22\cdot 1}) \,. \end{aligned} \tag{5.32}$$

The fact (5.32) is one useful property of the Schur complement: for a nonnegative definite $\mathbf{A}$, the rank is additive on the Schur complement; see (13.61) (p. 299).

5.4 Equality of $\mathscr{C}(\mathbf{AF})$ and $\mathscr{C}(\mathbf{AG})$

Let us prove the following simple result:

$$\mathscr{C}(\mathbf{G}) = \mathscr{C}(\mathbf{F}) \implies \operatorname{rank}(\mathbf{AG}) = \operatorname{rank}(\mathbf{AF}) \,. \tag{5.33}$$

The proof comes at once from

$$\begin{aligned} \operatorname{rank}(\mathbf{AG}) &= \operatorname{rank}(\mathbf{A}) - \dim \mathscr{C}(\mathbf{A}') \cap \mathscr{C}(\mathbf{G}^{\perp}) \\ &= \operatorname{rank}(\mathbf{A}) - \dim \mathscr{C}(\mathbf{A}') \cap \mathscr{C}(\mathbf{F}^{\perp}) = \operatorname{rank}(\mathbf{AF}) \,. \end{aligned} \tag{5.34}$$

Actually, $\mathscr{C}(\mathbf{G}) = \mathscr{C}(\mathbf{F})$ implies that $\mathbf{G} = \mathbf{FK}$ for some $\mathbf{K}$ and hence $\mathscr{C}(\mathbf{AG}) = \mathscr{C}(\mathbf{AFK}) \subset \mathscr{C}(\mathbf{AF})$ which together with (5.34) means that

$$\mathscr{C}(\mathbf{G}) = \mathscr{C}(\mathbf{F}) \implies \mathscr{C}(\mathbf{AG}) = \mathscr{C}(\mathbf{AF}) \,. \tag{5.35}$$

In particular, because $\mathscr{C}(\mathbf{LL}') = \mathscr{C}(\mathbf{L})$, we always have

$$\mathscr{C}(\mathbf{ALL}') = \mathscr{C}(\mathbf{AL}) \,. \tag{5.36}$$

The rank rule for the product yields immediately also the following result:

Proposition 5.2. *The following statements are equivalent:*

(a) $\mathscr{C}(\mathbf{A}_{n\times m}) = \mathscr{C}(\mathbf{B}_{n\times p})$.

(b) *There exists* $\mathbf{F}\colon \mathbf{A}_{n\times m} = \mathbf{B}_{n\times p}\mathbf{F}_{p\times m}$, *where* $\mathscr{C}(\mathbf{B}') \cap \mathscr{C}(\mathbf{F}^{\perp}) = \{\mathbf{0}\}$.

Moreover, if $\mathbf{B}_{n\times p}$ *has full column rank, then* (a) *is equivalent to the following:*

(c) *There exists* $\mathbf{F}\colon \mathbf{A}_{n\times m} = \mathbf{B}_{n\times p}\mathbf{F}_{p\times m}$, *where* $\operatorname{rank}(\mathbf{F}) = p$.

As regards part (c), notice that if $\mathbf{B}_{n\times p}$ has full column rank, then $\mathscr{C}(\mathbf{B}') = \mathbb{R}^p$, and hence $\mathscr{C}(\mathbf{B}') \cap \mathscr{C}(\mathbf{F}^{\perp}) = \{\mathbf{0}\}$ if and only if $\mathscr{C}(\mathbf{F}^{\perp}) = \{\mathbf{0}\}$, i.e., $\mathscr{C}(\mathbf{F}) = \mathbb{R}^p$, which means that $\mathbf{F}_{p\times m}$ has full row rank; cf. Theorem 17 (p. 349) concerning the full rank decomposition.

Note that trivially

$$\mathscr{C}(\mathbf{A}\mathbf{P}_{\mathbf{F}}) = \mathscr{C}(\mathbf{A}\mathbf{F}), \tag{5.37}$$

and

$$\begin{aligned} \operatorname{rank}(\mathbf{A}\mathbf{F}) &= \operatorname{rank}(\mathbf{A}\mathbf{P}_{\mathbf{F}}) = \operatorname{rank}(\mathbf{P}_{\mathbf{F}}\mathbf{A}') \\ &= \operatorname{rank}(\mathbf{F}'\mathbf{A}') = \operatorname{rank}(\mathbf{P}_{\mathbf{F}}\mathbf{P}_{\mathbf{A}'}) = \operatorname{rank}(\mathbf{P}_{\mathbf{A}'}\mathbf{P}_{\mathbf{F}}) . \end{aligned} \tag{5.38}$$

5.5 Rank of the Model Matrix

Let the $n \times p$ model matrix $\mathbf{X}$ be partitioned as

$$\mathbf{X} = (\mathbf{1} : \mathbf{x}_1 : \ldots : \mathbf{x}_k) = (\mathbf{1} : \mathbf{X}_0) , \tag{5.39}$$

and so $p = k + 1$. The sample covariance matrix of the x-variables is

$$\operatorname{cov}_{\mathrm{d}}(\mathbf{X}_0) = \tfrac{1}{n-1}\mathbf{X}_0'\mathbf{C}\mathbf{X}_0 = \tfrac{1}{n-1}\mathbf{T}_{\mathbf{xx}} = \mathbf{S}_{\mathbf{xx}} , \tag{5.40}$$

and the sample correlation matrix is

$$\operatorname{cor}_{\mathrm{d}}(\mathbf{X}_0) = [\operatorname{diag}(\mathbf{T}_{\mathbf{xx}})]^{-1/2}\mathbf{T}_{\mathbf{xx}}[\operatorname{diag}(\mathbf{T}_{\mathbf{xx}})]^{-1/2} = \mathbf{R}_{\mathbf{xx}} . \tag{5.41}$$

While calculating the correlations, we assume that all x-variables have nonzero variances, that is, the matrix $\operatorname{diag}(\mathbf{T}_{\mathbf{xx}})$ is positive definite, or in other words:

$$\mathbf{x}_i \notin \mathscr{C}(\mathbf{1}) , \quad i = 1, \ldots, k . \tag{5.42}$$

Using Theorem 5, we can now prove the following result:

Proposition 5.3. *The rank of the model matrix* $\mathbf{X} = (\mathbf{1} : \mathbf{X}_0)$ *can be expressed as*

$$\begin{aligned} \operatorname{rank}(\mathbf{X}) &= 1 + \operatorname{rank}(\mathbf{X}_0) - \dim \mathscr{C}(\mathbf{1}) \cap \mathscr{C}(\mathbf{X}_0) && (5.43a) \\ &= \operatorname{rank}(\mathbf{1} : \mathbf{C}\mathbf{X}_0) = 1 + \operatorname{rank}(\mathbf{C}\mathbf{X}_0) && (5.43b) \\ &= 1 + \operatorname{rank}(\mathbf{T}_{\mathbf{xx}}) && (5.43c) \\ &= 1 + \operatorname{rank}(\mathbf{S}_{\mathbf{xx}}) \,, && (5.43d) \end{aligned}$$

and thereby

$$\begin{aligned} \operatorname{rank}(\mathbf{S}_{\mathbf{xx}}) &= \operatorname{rank}(\mathbf{X}) - 1 = \operatorname{rank}(\mathbf{C}\mathbf{X}_0) \\ &= \operatorname{rank}(\mathbf{X}_0) - \dim \mathscr{C}(\mathbf{1}) \cap \mathscr{C}(\mathbf{X}_0) \,. \end{aligned} \tag{5.44}$$

If all x*-variables have nonzero variances, i.e., the correlation matrix* $\mathbf{R}_{\mathbf{xx}}$ *is properly defined, then* $\operatorname{rank}(\mathbf{R}_{\mathbf{xx}}) = \operatorname{rank}(\mathbf{S}_{\mathbf{xx}})$*. Moreover, the following statements are equivalent:*

(a) $\det(\mathbf{S}_{\mathbf{xx}}) \neq 0$,

(b) $\operatorname{rank}(\mathbf{X}) = k + 1$,

(c) $\operatorname{rank}(\mathbf{X}_0) = k$ *and* $\mathbf{1} \notin \mathscr{C}(\mathbf{X}_0)$.

Proof. Equation (5.43a) is obvious, and (5.43b) follows immediately from (5.2). Clearly (5.43c) holds since

$$\operatorname{rank}(\mathbf{C}\mathbf{X}_0) = \operatorname{rank}(\mathbf{X}_0'\mathbf{C}'\mathbf{C}\mathbf{X}_0) = \operatorname{rank}(\mathbf{X}_0'\mathbf{C}\mathbf{X}_0) = \operatorname{rank}(\mathbf{T}_{\mathbf{xx}}) \,. \tag{5.45}$$

Now (5.43) implies the equivalence between (a), (b), and (c) of at once, and thereby our claims are proved. □

We may note that

$$\mathscr{C}(\mathbf{X}) = \mathscr{C}(\mathbf{1} : \mathbf{C}\mathbf{X}_0), \quad (\mathbf{1} : \mathbf{C}\mathbf{X}_0) = (\mathbf{1} : \mathbf{X}_0)\mathbf{L} \,, \tag{5.46}$$

where, see also (3.17a) (p. 94),

$$\mathbf{L} = \begin{pmatrix} 1 & -\mathbf{1}^{+}\mathbf{X}_0 \\ \mathbf{0} & \mathbf{I}_k \end{pmatrix} = \begin{pmatrix} 1 & -\frac{1}{n}\mathbf{1}'\mathbf{X}_0 \\ \mathbf{0} & \mathbf{I}_k \end{pmatrix} = \begin{pmatrix} 1 & -\bar{\mathbf{x}} \\ \mathbf{0} & \mathbf{I}_k \end{pmatrix}, \quad \bar{\mathbf{x}} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_k \end{pmatrix}. \tag{5.47}$$

It is noteworthy that

$$r_{ij} = 1 \text{ for some } i \neq j \implies \det(\mathbf{R}_{\mathbf{xx}}) = 0 \quad (\text{but not vice versa}) \,. \tag{5.48}$$

For example, if

$$\mathbf{S}_{\mathbf{xx}} = \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}, \quad \mathbf{R}_{\mathbf{xx}} = \begin{pmatrix} 1 & -0.5 & -0.5 \\ -0.5 & 1 & -0.5 \\ -0.5 & -0.5 & 1 \end{pmatrix}, \tag{5.49}$$

then

$$\det(\mathbf{R}_{\mathbf{xx}}) = \det(\mathbf{S}_{\mathbf{xx}}) = 0, \quad \operatorname{rank}(\mathbf{R}_{\mathbf{xx}}) = \operatorname{rank}(\mathbf{S}_{\mathbf{xx}}) = 2\,. \tag{5.50}$$

The linear dependence of the columns of $\mathbf{R}_{\mathbf{xx}}$ can be also concluded from

$$\mathbf{R}_{\mathbf{xx}}\mathbf{1}_3 = \mathbf{0}\,. \tag{5.51}$$

If $\mathbf{Z} = (\mathbf{z}_1 : \mathbf{z}_2 : \mathbf{z}_3)$ represents the data matrix based on centered and standardized (length of each column is 1) values yielding (5.49), then $\mathbf{R}_{\mathbf{xx}} = \mathbf{Z}'\mathbf{Z}$. Now $\mathbf{Z}'\mathbf{Z}\mathbf{1}_3 = \mathbf{0}$ implies that $\mathbf{Z}\mathbf{1}_3 = \mathbf{0}$, i.e., $\mathbf{z}_3 = -\mathbf{z}_1 - \mathbf{z}_2$.

We see that the covariance matrix $\mathbf{S}_{\mathbf{xx}}$ is singular if and only if $\operatorname{rank}(\mathbf{X}) < k+1$, i.e., there exists a nonzero vector $\mathbf{a} \in \mathbb{R}^{k+1}$ such that

$$\mathbf{X}\mathbf{a} = a_0\mathbf{1} + a_1\mathbf{x}_1 + \cdots + a_k\mathbf{x}_k = \mathbf{0}\,. \tag{5.52}$$

Suppose, for notational simplicity, that $a_k \neq 0$. Then (5.52) implies that $\mathbf{x}_k$ can be expressed as

$$\mathbf{x}_k = b_0\mathbf{1} + b_1\mathbf{x}_1 + \cdots + b_{k-1}\mathbf{x}_{k-1} = (\mathbf{1} : \mathbf{x}_1 : \ldots : \mathbf{x}_{k-1})\mathbf{b} := \mathbf{X}_1\mathbf{b}\,, \tag{5.53}$$

for an appropriate $\mathbf{b} \in \mathbb{R}^k$. Condition (5.53) means that $\mathbf{x}_k \in \mathscr{C}(\mathbf{X}_1)$, i.e.,

$$\mathbf{P}_{\mathbf{X}_1}\mathbf{x}_k = \mathbf{x}_k\,, \quad \text{where } \mathbf{X}_1 = (\mathbf{1} : \mathbf{x}_1 : \ldots : \mathbf{x}_{k-1})\,. \tag{5.54}$$

If we consider, as in Section 8.3 (p. 161), the model $\mathscr{M}_k$ where

$$\mathscr{M}_k = \{\mathbf{x}_k, \mathbf{X}_1\boldsymbol{\gamma}, \sigma^2\mathbf{I}\}\,, \tag{5.55}$$

then the residual sum of squares $\operatorname{SSE}(k)$, say, is zero if (5.54) holds. In particular, if $\mathbf{x}_k \notin \mathscr{C}(\mathbf{1})$ so that the variance of x_k is nonzero, then (5.54) means that

$$R^2_{k\cdot 1\ldots k-1} = 1\,, \tag{5.56}$$

where $R^2_{k\cdot 1\ldots k-1}$ is the coefficient of determination when x_k is explained by all other x_i's (plus a constant).

For example, if $\mathbf{X} = (\mathbf{1} : \mathbf{x} : \mathbf{z})$ where $s_x^2 > 0$ and $s_z^2 > 0$, then $\operatorname{rank}(\mathbf{X}) = 2$ if and only if $r_{xz}^2 = 1$; see Exercise 3.1 (p. 102).

As references to nonsingularity of covariance matrix, we may mention Das Gupta (1971) and Trenkler (1995).

5.6 The Rank of the Extended Model Matrix

Consider the model matrix which is partitioned row-wise as

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_{(i)} \\ \mathbf{x}'_{(i)} \end{pmatrix}, \tag{5.57}$$

where, for the notational convenience, $\mathbf{x}'_{(i)}$ is the last row of $\mathbf{X}$ and $\mathbf{X}_{(i)}$ comprises the rest of the rows of $\mathbf{X}$. Let $\mathbf{u}$ denote the last column of $\mathbf{I}_n$. Then

$$\mathbf{I}_n - \mathbf{P_u} = \mathbf{I}_n - \begin{pmatrix} \mathbf{0}_{(n-1)\times(n-1)} & \mathbf{0}_{n-1} \\ \mathbf{0}'_{n-1} & 1 \end{pmatrix} = \begin{pmatrix} \mathbf{I}_{n-1} & \mathbf{0}_{n-1} \\ \mathbf{0}'_{n-1} & 0 \end{pmatrix}, \tag{5.58}$$

and

$$\begin{aligned} \operatorname{rank}(\mathbf{X} : \mathbf{u}) &= 1 + \operatorname{rank}(\mathbf{X}) - \dim \mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{u}) \\ &= 1 + \operatorname{rank}[(\mathbf{I}_n - \mathbf{P_u})\mathbf{X}] \\ &= 1 + \operatorname{rank}(\mathbf{X}_{(i)}) . \end{aligned} \tag{5.59}$$

Moreover, we have

$$\dim \mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{u}) = \operatorname{rank}(\mathbf{X}) - \operatorname{rank}(\mathbf{X}_{(i)}) , \tag{5.60}$$

and obviously

$$\operatorname{rank}(\mathbf{X}) = \operatorname{rank}(\mathbf{X}_{(i)}) \iff \mathbf{x}_{(i)} \in \mathscr{C}(\mathbf{X}'_{(i)}) , \tag{5.61}$$

and so

$$\begin{aligned} \mathbf{x}_{(i)} \in \mathscr{C}(\mathbf{X}'_{(i)}) &\iff \mathbf{u} \notin \mathscr{C}(\mathbf{X}) \\ &\iff \mathbf{Hu} \neq \mathbf{u} \iff \mathbf{u}'\mathbf{Hu} \neq 1 \iff h_{ii} \neq 1 \\ &\iff m_{ii} = 1 - h_{ii} > 0 . \end{aligned} \tag{5.62}$$

If we consider the extended model, see Section 8.12 (p. 180):

$$\mathscr{M}_Z = \{\mathbf{y},\, \mathbf{Z}\boldsymbol{\gamma},\, \sigma^2\mathbf{I}\} = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta} + \mathbf{i}_i\delta,\, \sigma^2\mathbf{I}\} , \tag{5.63}$$

where

$$\mathbf{Z} = (\mathbf{X} : \mathbf{i}_i) , \quad \mathbf{i}_i = \text{the } i\text{th column of } \mathbf{I}_n , \quad \boldsymbol{\gamma} = \begin{pmatrix} \boldsymbol{\beta} \\ \delta \end{pmatrix}, \tag{5.64}$$

then δ is estimable under $\mathscr{M}_Z$ if and only if $\mathbf{i}_i \notin \mathscr{C}(\mathbf{X})$; see (8.35) (p. 160). Thus we can summarize our findings as follows:

$$\begin{aligned} m_{ii} > 0 &\iff \mathbf{M}\mathbf{i}_i \neq \mathbf{0} \iff \mathbf{i}_i \notin \mathscr{C}(\mathbf{X}) \\ &\iff \mathbf{x}_{(i)} \in \mathscr{C}(\mathbf{X}'_{(i)}) \\ &\iff \operatorname{rank}(\mathbf{X}_{(i)}) = \operatorname{rank}(\mathbf{X}) \\ &\iff \delta \text{ is estimable under } \mathscr{M}_Z = \{\mathbf{y},\, \mathbf{Z}\boldsymbol{\gamma},\, \sigma^2\mathbf{I}\}\,. \end{aligned} \tag{5.65}$$

5.7 Matrix Volume and the Sample Generalized Variance

Consider the centered data matrix $\widetilde{\mathbf{U}} = (\tilde{\mathbf{x}} : \tilde{\mathbf{y}})$ whose covariance matrix is $\mathbf{S}$. Then the determinant of $\mathbf{S}$, i.e., the generalized variance of this data set, is

$$\det(\mathbf{S}) = s_x^2 s_y^2 - s_{xy}^2 = s_x^2 s_y^2\Big(1 - \frac{s_{xy}^2}{s_x^2 s_y^2}\Big) = s_x^2 s_y^2(1 - r_{xy}^2)\,, \tag{5.66}$$

and denoting $\mathbf{T} = \widetilde{\mathbf{U}}'\widetilde{\mathbf{U}}$, we have

$$\det(\widetilde{\mathbf{U}}'\widetilde{\mathbf{U}}) = \det(\mathbf{T}) = t_{xx}t_{yy}(1 - r_{xy}^2)\,. \tag{5.67}$$

There is an interesting connection between $\det(\mathbf{S})$ and the area of the parallelogram generated by the centered variable vectors $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$; see Figure 5.1. The squared area of this parallelogram is called the squared volume of the matrix $\widetilde{\mathbf{U}}$, and it is

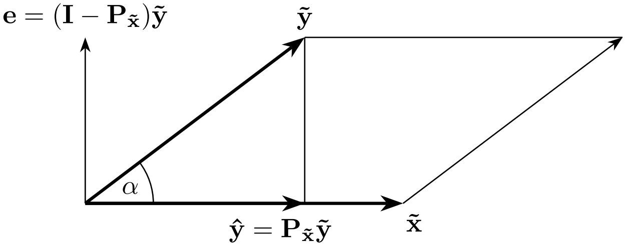


Figure 5.1 A parallelogram defined by the centered variable vectors $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$.

$$(\text{area})^2 := \operatorname{vol}^2(\widetilde{\mathbf{U}}) = \|\tilde{\mathbf{x}}\|^2\|\mathbf{e}\|^2, \tag{5.68}$$

where $\mathbf{e} = \tilde{\mathbf{y}} - \mathbf{P}_{\tilde{\mathbf{x}}}\tilde{\mathbf{y}}$. Now we have

$$\begin{aligned} \|\mathbf{e}\|^2 &= \tilde{\mathbf{y}}'(\mathbf{I}_n - \mathbf{P}_{\tilde{\mathbf{x}}})\tilde{\mathbf{y}} = \tilde{\mathbf{y}}'[\mathbf{I}_n - \tilde{\mathbf{x}}(\tilde{\mathbf{x}}'\tilde{\mathbf{x}})^{-1}\tilde{\mathbf{x}}']\tilde{\mathbf{y}} \\ &= t_{yy} - \frac{t_{xy}^2}{t_{xx}} = t_{yy}(1 - r_{xy}^2)\,, \end{aligned} \tag{5.69}$$

and hence

$$\operatorname{vol}^2(\widetilde{\mathbf{U}}) = t_{xx}t_{yy}(1 - r_{xy}^2) = \det(\mathbf{T}) = \det(\widetilde{\mathbf{U}}'\widetilde{\mathbf{U}}) = (n-1)^2 \det(\mathbf{S}) \,, \tag{5.70}$$

and

$$\det(\mathbf{S}) = \operatorname{vol}^2(\widetilde{\mathbf{U}})/(n-1)^2. \tag{5.71}$$

In the case of p variables, i.e., when $\widetilde{\mathbf{U}} = (\tilde{\mathbf{u}}_1 : \ldots : \tilde{\mathbf{u}}_p)$,

$$\det(\mathbf{S}_{p\times p}) = \operatorname{vol}^2(\widetilde{\mathbf{U}}_{n\times p})/(n-1)^p, \tag{5.72}$$

where the "volume" refers to the parallelepiped spanned by the columns of $\widetilde{\mathbf{U}}$. Three-dimensional parallelepiped is a skewed rectangular box.

So: for a fixed data set, the generalized sample variance is proportional to the square of the volume generated by the centered variable vectors $\tilde{\mathbf{u}}_1$, $\ldots$, $\tilde{\mathbf{u}}_p$. For further reading regarding the concept of volume, see Anderson (2003, Section 7.5) and Johnson & Wichern (2007, §3.4). Ben-Israel & Greville (2003, p. 29) consider also the r-dimensional volume, $\operatorname{vol}_r(\mathbf{A})$, of a matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$, where $\operatorname{rank}(\mathbf{A}) = r$. They define it as a square root of the sum of the squared determinants of all $r \times r$ submatrices, which yields

$$\operatorname{vol}_r(\mathbf{A}) = \operatorname{sg}_1(\mathbf{A}) \cdots \operatorname{sg}_r(\mathbf{A}) \,, \tag{5.73}$$

where $\operatorname{sg}_i(\mathbf{A})$ refers to the ith largest singular value of $\mathbf{A}$. See also Ben-Israel (1992) and Miao & Ben-Israel (1992).

5.8 Properties of $\mathbf{X}'\mathbf{V}\mathbf{X}$

Proposition 5.4. *Let* $\mathbf{V}$ *be a symmetric nonnegative definite* $n \times n$ *matrix and let* $\mathbf{X}$ *be an* $n \times p$ *matrix. Then the following statements are equivalent:*

(a) $\operatorname{rank}(\mathbf{X}'\mathbf{V}\mathbf{X}) = \operatorname{rank}(\mathbf{X})$,

(b) $\operatorname{rank}(\mathbf{X}'\mathbf{V}^+\mathbf{X}) = \operatorname{rank}(\mathbf{X})$,

(c) $\mathscr{C}(\mathbf{X}'\mathbf{V}\mathbf{X}) = \mathscr{C}(\mathbf{X}')$,

(d) $\mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{V}^\perp) = \{\mathbf{0}\}$,

(e) $\mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{V}\mathbf{X})^\perp = \{\mathbf{0}\}$,

(f) $\mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{V}^+\mathbf{X})^\perp = \{\mathbf{0}\}$.

Moreover,

$$\operatorname{rank}(\mathbf{X}'\mathbf{V}\mathbf{X}) = p \iff \operatorname{rank}(\mathbf{X}) = p \text{ and } \mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{V}^\perp) = \{\mathbf{0}\} \,. \tag{5.74}$$

Proof. We first note that in view of (5.3) and (1.14) (p. 60), we have

$$\operatorname{rank}(\mathbf{X}'\mathbf{V}\mathbf{X}) = \operatorname{rank}(\mathbf{X}'\mathbf{V}) = \operatorname{rank}(\mathbf{X}) - \dim \mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{V})^{\perp}, \tag{5.75a}$$

$$\operatorname{rank}(\mathbf{X}'\mathbf{V}^{+}\mathbf{X}) = \operatorname{rank}(\mathbf{X}'\mathbf{V}^{+}) = \operatorname{rank}(\mathbf{X}) - \dim \mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{V}^{+})^{\perp}. \tag{5.75b}$$

In light of (4.29) (p. 110), we have for any matrix $\mathbf{A}$, $\mathscr{C}(\mathbf{A}^{+}) = \mathscr{C}(\mathbf{A}')$, and so $\mathscr{C}(\mathbf{V}) = \mathscr{C}(\mathbf{V}^{+})$ and thereby (a) and (b) are equivalent. The other equivalences follow at once from (5.75a) and (5.75b). A necessary and sufficient condition for $\mathbf{X}'\mathbf{V}\mathbf{X}$ (and hence $\mathbf{X}'\mathbf{V}^{+}\mathbf{X}$) to be invertible follows at once from (5.75a). □

5.9 Number of Unit Eigenvalues of $\mathbf{P_A P_B}$

Let $\mathbf{A} \in \mathbb{R}^{n \times a}$, $\mathbf{B} \in \mathbb{R}^{n \times b}$, and

$$\mathbf{A}'\mathbf{A} = \mathbf{I}_a\,, \quad \mathbf{B}'\mathbf{B} = \mathbf{I}_b\,, \tag{5.76}$$

which means that the columns of $\mathbf{A}$ and $\mathbf{B}$ form orthonormal bases for the column spaces $\mathscr{C}(\mathbf{A})$ and $\mathscr{C}(\mathbf{B})$, respectively, and $\mathbf{P_A} = \mathbf{A}\mathbf{A}'$ and $\mathbf{P_B} = \mathbf{B}\mathbf{B}'$. Our question is now: What is the number of unit eigenvalues (those eigenvalues that equal to one), say u, of the product $\mathbf{P_A P_B}$? The product of two orthogonal projectors is frequently met in statistics and hence it is useful to know some of its properties.

First, because, see Proposition (13.2) (p. 299),

$$\mathbf{UV} \text{ and } \mathbf{VU} \text{ have the same nonzero eigenvalues,} \tag{5.77}$$

the matrices $\mathbf{P_A P_B}$ and $\mathbf{P_A P_B P_A}$ have the same nonzero eigenvalues. The symmetry and nonnegative definiteness of $\mathbf{P_A P_B P_A}$ guarantees that these eigenvalues are all real and nonnegative. Moreover, because

$$\|\mathbf{P_B P_A x}\| \le \|\mathbf{P_A x}\| \le \|\mathbf{x}\| \quad \text{for all } \mathbf{x} \in \mathbb{R}^n, \tag{5.78}$$

all eigenvalues of $\mathbf{P_A P_B}$ are less than or equal to 1:

$$\operatorname{ch}_1(\mathbf{P_A P_B P_A}) = \max_{\mathbf{x} \neq \mathbf{0}} \frac{\mathbf{x}'\mathbf{P_A P_B P_A x}}{\mathbf{x}'\mathbf{x}} \le 1\,. \tag{5.79}$$

Similarly, the matrices $\mathbf{A}\mathbf{A}'\mathbf{B}\mathbf{B}' = \mathbf{P_A P_B}$ and $\mathbf{A}'\mathbf{B}\mathbf{B}'\mathbf{A}$ have the same nonzero eigenvalues. Hence our task is to find how many times, u, say, 1 is repeated as a root of the characteristic polynomial

$$\det(\mathbf{A}'\mathbf{B}\mathbf{B}'\mathbf{A} - \lambda \mathbf{I}_a)\,. \tag{5.80}$$

Now, according to (18.17) (p. 359), u is the dimension of the eigenspace of $\mathbf{A}'\mathbf{B}\mathbf{B}'\mathbf{A}$ with respect to $\lambda = 1$, i.e.,

$$u = \dim \mathscr{N}(\mathbf{A}'\mathbf{B}\mathbf{B}'\mathbf{A} - \mathbf{I}_a)\,. \tag{5.81}$$

In view of $\mathbf{A}'\mathbf{B}\mathbf{B}'\mathbf{A} - \mathbf{I}_a = \mathbf{A}'(\mathbf{P_B} - \mathbf{I}_n)\mathbf{A}$, we have

$$\begin{aligned} u &= a - \operatorname{rank}[\mathbf{A}'(\mathbf{I}_n - \mathbf{P_B})\mathbf{A}] \\ &= a - \operatorname{rank}[\mathbf{A}'(\mathbf{I}_n - \mathbf{P_B})] \\ &= a - [\operatorname{rank}(\mathbf{A}) - \dim \mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B})] \\ &= \dim \mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B})\,, \end{aligned} \tag{5.82}$$

that is,

$$\#\{\operatorname{ch}(\mathbf{P_A}\mathbf{P_B}) = 1\} = \dim \mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B})\,. \tag{5.83}$$

Notice that the squared singular values of $\mathbf{A}'\mathbf{B}$ are precisely the eigenvalues of $\mathbf{A}'\mathbf{B}\mathbf{B}'\mathbf{A}$ and hence

$$\operatorname{sg}_1^2(\mathbf{A}'\mathbf{B}) = \operatorname{ch}_1(\mathbf{A}'\mathbf{B}\mathbf{B}'\mathbf{A}) = \operatorname{ch}_1(\mathbf{P_A}\mathbf{P_B}) = \|\mathbf{P_A}\mathbf{P_B}\|_2\,, \tag{5.84}$$

where $\|\cdot\|_2$ refers to the matrix 2-norm (spectral norm) of the matrix argument. Using $\mathbf{A}'\mathbf{A} = \mathbf{I}_a$, $\mathbf{B}'\mathbf{B} = \mathbf{I}_b$, and the maximal property of the largest singular value, see Section 19.1 (p. 395), we can write

$$\begin{aligned} \operatorname{sg}_1^2(\mathbf{A}'\mathbf{B}) &= \max_{\boldsymbol{\alpha} \neq \mathbf{0},\, \boldsymbol{\beta} \neq \mathbf{0}} \frac{(\boldsymbol{\alpha}'\mathbf{A}'\mathbf{B}\boldsymbol{\beta})^2}{\boldsymbol{\alpha}'\boldsymbol{\alpha} \cdot \boldsymbol{\beta}'\boldsymbol{\beta}} = \max_{\|\boldsymbol{\alpha}\| = \|\boldsymbol{\beta}\| = 1} (\boldsymbol{\alpha}'\mathbf{A}'\mathbf{B}\boldsymbol{\beta})^2 \\ &= \max_{\boldsymbol{\alpha} \neq \mathbf{0},\, \boldsymbol{\beta} \neq \mathbf{0}} \frac{(\boldsymbol{\alpha}'\mathbf{A}'\mathbf{B}\boldsymbol{\beta})^2}{\boldsymbol{\alpha}'\mathbf{A}'\mathbf{A}\boldsymbol{\alpha} \cdot \boldsymbol{\beta}'\mathbf{B}'\mathbf{B}\boldsymbol{\beta}} \\ &= \max_{\mathbf{u} \in \mathscr{C}(\mathbf{A}),\, \mathbf{v} \in \mathscr{C}(\mathbf{B})} \cos^2(\mathbf{u}, \mathbf{v}) \\ &= \cos^2 \theta_{\min}\,. \end{aligned} \tag{5.85}$$

The angle $\theta_{\min}$, $0 \leq \theta_{\min} \leq \pi/2$, yielding the maximal cosine is called the minimal angle or the first principal angle between the subspaces $\mathscr{C}(\mathbf{A})$ and $\mathscr{C}(\mathbf{B})$; see also (2.46) (p. 77). The other singular values of $\mathbf{A}'\mathbf{B}$ are the cosines of the other principal angles. We will return to these concepts in the context of the canonical correlations. For references, see, e.g., Afriat (1957), Baksalary & Trenkler (2009a), Ben-Israel & Greville (2003, p. 230), Galántai (2008), Ipsen & Meyer (1995), Meyer (2000, §5.15), and Miao & Ben-Israel (1992, 1996).

5.10 Properties of $\mathbf{X}'\mathbf{V}\mathbf{X} - \mathbf{X}'\mathbf{V}\mathbf{Z}(\mathbf{Z}'\mathbf{V}\mathbf{Z})^{-}\mathbf{Z}'\mathbf{V}\mathbf{X}$

Let $\mathbf{Z}$ be a matrix with property

$$\mathscr{C}(\mathbf{Z}) = \mathscr{C}(\mathbf{X})^{\perp} = \mathscr{C}(\mathbf{M})\,. \tag{5.86}$$

Denote

$$u = \dim \mathscr{C}(\mathbf{VH}) \cap \mathscr{C}(\mathbf{VM}) = \dim \mathscr{C}(\mathbf{VX}) \cap \mathscr{C}(\mathbf{VX}^{\perp}) \,, \tag{5.87}$$

and

$$\mathbf{\Sigma}_{11\cdot 2} = \mathbf{X'VX} - \mathbf{X'VZ(Z'VZ)^{-}Z'VX} = \mathbf{\Sigma}/\mathbf{Z'VZ} \,, \tag{5.88}$$

where $\mathbf{\Sigma}_{11\cdot 2} = \mathbf{\Sigma}/\mathbf{Z'VZ}$ refers to the Schur complement of $\mathbf{Z'VZ}$ in $\mathbf{\Sigma}$:

$$\mathbf{\Sigma} = (\mathbf{X} : \mathbf{Z})'\mathbf{V}(\mathbf{X} : \mathbf{Z}) = \begin{pmatrix} \mathbf{X'VX} & \mathbf{X'VZ} \\ \mathbf{Z'VX} & \mathbf{Z'VZ} \end{pmatrix} . \tag{5.89}$$

The following result appears to be useful.

Proposition 5.5. *With the above notation, the dimension* $u = \dim \mathscr{C}(\mathbf{VH}) \cap \mathscr{C}(\mathbf{VM})$ *can be expressed in the following alternative forms:*

(a) $u = \dim \mathscr{C}(\mathbf{V}^{1/2}\mathbf{X}) \cap \mathscr{C}(\mathbf{V}^{1/2}\mathbf{Z})$,

(b) $u = \operatorname{rank}(\mathbf{V}) - \dim \mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{V}) - \dim \mathscr{C}(\mathbf{Z}) \cap \mathscr{C}(\mathbf{V})$,

(c) $u = \operatorname{rank}(\mathbf{X}'\mathbf{P}_{\mathbf{V}}\mathbf{Z}) = \operatorname{rank}(\mathbf{H}\mathbf{P}_{\mathbf{V}}\mathbf{M})$,

(d) $u = \#\{\operatorname{ch}(\mathbf{P}_{\mathbf{V}^{1/2}\mathbf{H}}\mathbf{P}_{\mathbf{V}^{1/2}\mathbf{M}}) = 1\}$.

Moreover,

$$\begin{aligned} \operatorname{rank}(\mathbf{\Sigma}_{11\cdot 2}) &= \operatorname{rank}(\mathbf{VX}) - u \\ &= \operatorname{rank}(\mathbf{V}) - \operatorname{rank}(\mathbf{VZ}) = \dim \mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{V}) \,. \end{aligned} \tag{5.90}$$

Proof. The proof of (a) and (b) goes easily by noting that

$$\begin{aligned} \operatorname{rank}[\mathbf{V}(\mathbf{X} : \mathbf{Z})] &= \operatorname{rank}(\mathbf{V}) \\ &= \operatorname{rank}(\mathbf{VX}) + \operatorname{rank}(\mathbf{VZ}) - \dim \mathscr{C}(\mathbf{VX}) \cap \mathscr{C}(\mathbf{VZ}) \\ &= \operatorname{rank}(\mathbf{V}^{1/2}) = \operatorname{rank}[\mathbf{V}^{1/2}(\mathbf{X} : \mathbf{Z})] \\ &= \operatorname{rank}(\mathbf{V}^{1/2}\mathbf{X}) + \operatorname{rank}(\mathbf{V}^{1/2}\mathbf{Z}) - \dim \mathscr{C}(\mathbf{V}^{1/2}\mathbf{X}) \cap \mathscr{C}(\mathbf{V}^{1/2}\mathbf{Z}) \\ &= \operatorname{rank}(\mathbf{VX}) + \operatorname{rank}(\mathbf{VZ}) - \dim \mathscr{C}(\mathbf{V}^{1/2}\mathbf{X}) \cap \mathscr{C}(\mathbf{V}^{1/2}\mathbf{Z}) \,. \end{aligned} \tag{5.91}$$

To prove (c), we first note that

$$\begin{aligned} \operatorname{rank}(\mathbf{H}\mathbf{P}_{\mathbf{V}}\mathbf{M}) &= \operatorname{rank}(\mathbf{H}\mathbf{P}_{\mathbf{V}}) - \dim \mathscr{C}(\mathbf{P}_{\mathbf{V}}\mathbf{H}) \cap \mathscr{C}(\mathbf{H}) \\ &= \operatorname{rank}(\mathbf{HV}) - \dim \mathscr{C}(\mathbf{P}_{\mathbf{V}}\mathbf{X}) \cap \mathscr{C}(\mathbf{X}) \,. \end{aligned} \tag{5.92}$$

Next we prove the following:

$$\mathscr{C}(\mathbf{P}_{\mathbf{V}}\mathbf{X}) \cap \mathscr{C}(\mathbf{X}) = \mathscr{C}(\mathbf{V}) \cap \mathscr{C}(\mathbf{X}) \,. \tag{5.93}$$

It is clear that $\mathscr{C}(\mathbf{P}_{\mathbf{V}}\mathbf{X}) \cap \mathscr{C}(\mathbf{X}) \subset \mathscr{C}(\mathbf{V}) \cap \mathscr{C}(\mathbf{X})$. Let then $\mathbf{a} \in \mathscr{C}(\mathbf{V}) \cap \mathscr{C}(\mathbf{X})$ and so

$$\mathbf{a} = \mathbf{V}\boldsymbol{\alpha} = \mathbf{X}\boldsymbol{\beta} \quad \text{for some } \boldsymbol{\alpha},\ \boldsymbol{\beta}\,. \tag{5.94}$$

Premultiplying (5.94) by $\mathbf{P_V}$ yields

$$\mathbf{P_V a} = \mathbf{P_V V}\boldsymbol{\alpha} = \mathbf{V}\boldsymbol{\alpha} = \mathbf{a} = \mathbf{P_V X}\boldsymbol{\beta} \in \mathscr{C}(\mathbf{P_V X})\,, \tag{5.95}$$

and hence (5.93) indeed holds. Substituting (5.93) into (5.92) gives

$$\begin{aligned}\operatorname{rank}(\mathbf{HP_V M}) &= \operatorname{rank}(\mathbf{HV}) - \dim \mathscr{C}(\mathbf{H}) \cap \mathscr{C}(\mathbf{V}) \\ &= \operatorname{rank}(\mathbf{V}) - \dim \mathscr{C}(\mathbf{M}) \cap \mathscr{C}(\mathbf{V}) - \dim \mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{V}) \\ &= u\,. \end{aligned} \tag{5.96}$$

The result (c) can also be proved by using (5.3) repeatedly; see Puntanen (1987, Th. 3.4.1) and Drury, Liu, Lu, Puntanen et al. (2002, Lemma 2).

Part (d) comes from (5.83) at once.

Applying the rank additivity rule (5.32) (p. 126) on the partitioned matrix $\mathbf{\Sigma}$ in (5.89), we observe that

$$\operatorname{rank}(\mathbf{\Sigma}) = \operatorname{rank}(\mathbf{V}) = \operatorname{rank}(\mathbf{Z'VZ}) + \operatorname{rank}(\mathbf{\Sigma}_{11\cdot 2})\,, \tag{5.97}$$

and so

$$\begin{aligned}\operatorname{rank}(\mathbf{\Sigma}_{11\cdot 2}) &= \operatorname{rank}(\mathbf{V}) - \operatorname{rank}(\mathbf{VZ}) \\ &= \operatorname{rank}(\mathbf{V}) - [\operatorname{rank}(\mathbf{V}) - \dim \mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{V})] \\ &= \dim \mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{V})\,. \end{aligned} \tag{5.98}$$

It is easy to confirm that also $\operatorname{rank}(\mathbf{\Sigma}_{11\cdot 2}) = \operatorname{rank}(\mathbf{VX}) - u$ and thereby (5.90) is proved. □

The figure u appears to be the number of unit canonical correlations between random vectors $\mathbf{Hy}$ and $\mathbf{My}$ when $\operatorname{cov}(\mathbf{y}) = \mathbf{V}$, or in short,

$$u = \#\text{ of unit canonical corr's between } \mathbf{Hy} \text{ and } \mathbf{My}\,. \tag{5.99}$$

Namely, if

$$\operatorname{cov}\begin{pmatrix}\mathbf{Hy}\\ \mathbf{My}\end{pmatrix} = \begin{pmatrix}\mathbf{HVH} & \mathbf{HVM}\\ \mathbf{MVH} & \mathbf{MVM}\end{pmatrix}, \tag{5.100}$$

then, according Proposition 18.11 (p. 381) and (18.180) (p. 383), the nonzero canonical correlations between $\mathbf{Hy}$ and $\mathbf{My}$ are the nonzero eigenvalues of matrix

$$\mathbf{U} := (\mathbf{HVH})^{-}\mathbf{HVM}(\mathbf{MVM})^{-}\mathbf{MVH}\,, \tag{5.101}$$

and the nonzero eigenvalues of $\mathbf{U}$ are the same as those of $\mathbf{P}_{\mathbf{V}^{1/2}\mathbf{H}}\mathbf{P}_{\mathbf{V}^{1/2}\mathbf{M}}$.

It is noteworthy that the general expression for the covariance matrix of the BLUE for $\mathbf{X}\boldsymbol{\beta}$ can be written as

$$\operatorname{cov}[\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta})] = \mathbf{HVH} - \mathbf{HVM}(\mathbf{MVM})^{-}\mathbf{MVH}\,, \tag{5.102}$$

and hence, in light of (5.98),

$$\operatorname{rank} \operatorname{cov}[\operatorname{BLUE}(\mathbf{X}\boldsymbol{\beta})] = \dim \mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{V}) \,. \tag{5.103}$$

Moreover, because obviously $\mathscr{C}[\operatorname{cov}(\mathbf{X}\tilde{\boldsymbol{\beta}})] \subset \mathscr{C}(\mathbf{X})$ and $\mathscr{C}[\operatorname{cov}(\mathbf{X}\tilde{\boldsymbol{\beta}})] \subset \mathscr{C}(\mathbf{V})$ (why does this hold?) we get $\mathscr{C}[\operatorname{cov}(\mathbf{X}\tilde{\boldsymbol{\beta}})] = \mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{V})$. Recalling from (0.232) (p. 45) that the BLUE's residual

$$\tilde{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}} = \mathbf{V}\dot{\mathbf{M}}\mathbf{y} = \mathbf{V}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^{-}\mathbf{M}\mathbf{y} \tag{5.104}$$

has the covariance matrix

$$\operatorname{cov}(\tilde{\boldsymbol{\varepsilon}}) = \mathbf{V}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^{-}\mathbf{M}\mathbf{V} = \mathbf{V}\dot{\mathbf{M}}\mathbf{V} \,, \tag{5.105}$$

an eager reader can prove the following proposition.

Proposition 5.6. *Consider the model* $\{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$ *and let* $\tilde{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}$ *be the* BLUE*'s residual. Then*

$$\mathscr{C}[\operatorname{cov}(\mathbf{X}\tilde{\boldsymbol{\beta}})] = \mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{V}) \,, \quad \text{and} \quad \mathscr{C}[\operatorname{cov}(\tilde{\boldsymbol{\varepsilon}})] = \mathscr{C}(\mathbf{V}\mathbf{M}) \,. \tag{5.106}$$

If $\mathbf{X}$ has full column rank, then

$$\begin{aligned} \operatorname{cov}(\tilde{\boldsymbol{\beta}}) &= (\mathbf{X}'\mathbf{X})^{-1}[\mathbf{X}'\mathbf{V}\mathbf{X} - \mathbf{X}'\mathbf{V}\mathbf{Z}(\mathbf{Z}'\mathbf{V}\mathbf{Z})^{-}\mathbf{Z}'\mathbf{V}\mathbf{X}](\mathbf{X}'\mathbf{X})^{-1} \\ &= \operatorname{cov}(\hat{\boldsymbol{\beta}}) - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{Z}(\mathbf{Z}'\mathbf{V}\mathbf{Z})^{-}\mathbf{Z}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned} \tag{5.107}$$

Hence

$$\operatorname{rank}[\operatorname{cov}(\tilde{\boldsymbol{\beta}})] = \dim \mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{V}) \le p \,, \tag{5.108}$$

and thereby (when $\mathbf{X}$ has full column rank)

$$\det[\operatorname{cov}(\tilde{\boldsymbol{\beta}})] \neq 0 \iff \mathscr{C}(\mathbf{X}) \subset \mathscr{C}(\mathbf{V}) \,. \tag{5.109}$$

5.11 Intersection $\mathscr{C}(\mathrm{A}) \cap \mathscr{C}(\mathrm{B})$

The following result gives an explicit expression for the intersection $\mathscr{C}(\mathbf{A})$ and $\mathscr{C}(\mathbf{B})$ which can sometimes be very helpful. For a reference, see Rao & Mitra (1971b, Complement 7, p. 118).

Proposition 5.7. *Consider the matrices* $\mathbf{A}_{n\times a}$ *and* $\mathbf{B}_{n\times b}$ *and denote* $\mathbf{Q}_{\mathbf{B}} = \mathbf{I}_n - \mathbf{P}_{\mathbf{B}}$. *Then*

$$\begin{aligned} \mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B}) &= \mathscr{C}\big[\mathbf{A}(\mathbf{A}'\mathbf{B}^{\perp})^{\perp}\big] \\ &= \mathscr{C}\big[\mathbf{A}(\mathbf{A}'\mathbf{Q}_{\mathbf{B}})^{\perp}\big] = \mathscr{C}\big[\mathbf{A}(\mathbf{I}_a - \mathbf{P}_{\mathbf{A}'\mathbf{Q}_{\mathbf{B}}})\big] \\ &= \mathscr{C}\big(\mathbf{A}[\mathbf{I}_a - (\mathbf{A}'\mathbf{Q}_{\mathbf{B}}\mathbf{A})^{-}\mathbf{A}'\mathbf{Q}_{\mathbf{B}}\mathbf{A}]\big) \,. \end{aligned} \tag{5.110}$$

Proof. Denote

$$\mathbf{F} = \mathbf{A}(\mathbf{A}'\mathbf{B}^{\perp})^{\perp} = \mathbf{A}(\mathbf{A}'\mathbf{Q}_{\mathbf{B}})^{\perp}. \tag{5.111}$$

Then $\mathscr{C}(\mathbf{F}) \subset \mathscr{C}(\mathbf{A})$. Moreover,

$$\mathbf{Q}_{\mathbf{B}}\mathbf{F} = \mathbf{Q}_{\mathbf{B}}\mathbf{A}(\mathbf{A}'\mathbf{Q}_{\mathbf{B}})^{\perp} = (\mathbf{A}\mathbf{Q}_{\mathbf{B}})'(\mathbf{A}'\mathbf{Q}_{\mathbf{B}})^{\perp} = \mathbf{0}\,, \tag{5.112}$$

i.e., $(\mathbf{I}_n - \mathbf{P}_{\mathbf{B}})\mathbf{F} = \mathbf{0}$ and thereby $\mathbf{P}_{\mathbf{B}}\mathbf{F} = \mathbf{F}$, and so $\mathscr{C}(\mathbf{F}) \subset \mathscr{C}(\mathbf{B})$ and

$$\mathscr{C}(\mathbf{F}) \subset \mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B})\,. \tag{5.113}$$

Using the rank rule of the product, we get

$$\begin{aligned} \operatorname{rank}(\mathbf{F}) = \operatorname{rank}\big[\mathbf{A}(\mathbf{A}'\mathbf{Q}_{\mathbf{B}})^{\perp}\big] &= \operatorname{rank}(\mathbf{A}) - \dim \mathscr{C}(\mathbf{A}') \cap \mathscr{C}(\mathbf{A}'\mathbf{Q}_{\mathbf{B}}) \\ &= \operatorname{rank}(\mathbf{A}) - \dim \mathscr{C}(\mathbf{A}'\mathbf{Q}_{\mathbf{B}}) \\ &= \operatorname{rank}(\mathbf{A}) - \operatorname{rank}(\mathbf{A}'\mathbf{Q}_{\mathbf{B}}) \\ &= \dim \mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B})\,. \end{aligned} \tag{5.114}$$

Because $\mathscr{C}(\mathbf{A}'\mathbf{Q}_{\mathbf{B}}) = \mathscr{C}(\mathbf{A}'\mathbf{Q}_{\mathbf{B}}\mathbf{A})$, we have

$$\mathscr{C}(\mathbf{A}'\mathbf{Q}_{\mathbf{B}})^{\perp} = \mathscr{C}(\mathbf{A}'\mathbf{Q}_{\mathbf{B}}\mathbf{A})^{\perp} = \mathscr{C}\big[\mathbf{I}_a - (\mathbf{A}'\mathbf{Q}_{\mathbf{B}}\mathbf{A})^{-}\mathbf{A}'\mathbf{Q}_{\mathbf{B}}\mathbf{A}\big] \tag{5.115}$$

for any $(\mathbf{A}'\mathbf{Q}_{\mathbf{B}}\mathbf{A})^{-}$, and hence

$$\mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B}) = \mathscr{C}\big(\mathbf{A}[\mathbf{I}_a - (\mathbf{A}'\mathbf{Q}_{\mathbf{B}}\mathbf{A})^{-}\mathbf{A}'\mathbf{Q}_{\mathbf{B}}\mathbf{A}]\big)\,. \tag{5.116}$$

□

It is obvious that

$$\mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B})^{\perp} = \mathscr{C}\big[\mathbf{A}(\mathbf{A}'\mathbf{B})^{\perp}\big] = \mathscr{C}\big[\mathbf{A}(\mathbf{I}_a - \mathbf{P}_{\mathbf{A}'\mathbf{B}})\big]\,. \tag{5.117}$$

In particular, if $\mathbf{X} \in \mathbb{R}^{n\times p}$ and $\mathbf{V}_{n\times n}$ is nonnegative definite, then

$$\begin{aligned} \mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{V})^{\perp} &= \mathscr{C}\big[\mathbf{X}(\mathbf{X}'\mathbf{V})^{\perp}\big] = \mathscr{C}\big[\mathbf{X}(\mathbf{X}'\mathbf{V}\mathbf{X})^{\perp}\big] \\ &= \mathscr{C}\big(\mathbf{X}[\mathbf{I}_p - (\mathbf{X}'\mathbf{V}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}\mathbf{X}]\big)\,, \end{aligned} \tag{5.118}$$

and denoting $\mathbf{M} = \mathbf{I}_n - \mathbf{P}_{\mathbf{X}}$,

$$\begin{aligned} \mathscr{C}(\mathbf{X})^{\perp} \cap \mathscr{C}(\mathbf{V})^{\perp} &= \mathscr{C}(\mathbf{X} : \mathbf{V})^{\perp} \\ &= \mathscr{C}\big(\mathbf{M}[\mathbf{I}_n - (\mathbf{M}\mathbf{V}\mathbf{M})^{-}\mathbf{M}\mathbf{V}\mathbf{M}]\big)\,. \end{aligned} \tag{5.119a}$$

Moreover, it is straightforward to confirm (please do so) that

$$\mathscr{C}(\mathbf{X} : \mathbf{V}) = \mathscr{N}\big([\mathbf{I}_n - \mathbf{M}\mathbf{V}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^{-}]\mathbf{M}\big)\,. \tag{5.119b}$$

5.12 Solution (in Y) to Y(A : B) = (0 : B)

In this example we consider the equation

$$\mathbf{Y}(\mathbf{A} : \mathbf{B}) = (\mathbf{0} : \mathbf{B}) , \quad \text{i.e.,} \quad \begin{pmatrix} \mathbf{A}' \\ \mathbf{B}' \end{pmatrix} \mathbf{Y}' = \begin{pmatrix} \mathbf{0} \\ \mathbf{B}' \end{pmatrix} . \tag{5.120}$$

We know that there exists a solution $\mathbf{Y}$ for (5.120) if and only if

$$\mathscr{C} \begin{pmatrix} \mathbf{0} \\ \mathbf{B}' \end{pmatrix} \subset \mathscr{C} \begin{pmatrix} \mathbf{A}' \\ \mathbf{B}' \end{pmatrix} . \tag{5.121}$$

The following result gives an equivalent (and very handy) characterization for the consistency of (5.120); see also Theorem 16 (p. 343) and Kala (1981, Lemma 2.2).

Proposition 5.8. *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be given* $n \times a$ *and* $n \times b$ *matrices. Then the equation*

$$\mathbf{Y}(\mathbf{A} : \mathbf{B}) = (\mathbf{0} : \mathbf{B}) , \tag{5.122}$$

i.e.,

$$\begin{pmatrix} \mathbf{A}' \\ \mathbf{B}' \end{pmatrix} \mathbf{Y}' = \begin{pmatrix} \mathbf{0} \\ \mathbf{B}' \end{pmatrix} , \tag{5.123}$$

has a solution (for $\mathbf{Y} \in \mathbb{R}^{n \times n}$*) if and only if*

$$\mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B}) = \{\mathbf{0}\} . \tag{5.124}$$

The solution is unique if and only if $\mathscr{C}(\mathbf{A} : \mathbf{B}) = \mathbb{R}^n$.

Proof. From (5.123) we conclude that (5.122) has a unique solution if and only if the columns of $\binom{\mathbf{A}'}{\mathbf{B}'} \in \mathbb{R}^{(a+b) \times n}$ are linearly independent, i.e., $\operatorname{rank}(\mathbf{A} : \mathbf{B})' = \operatorname{rank}(\mathbf{A} : \mathbf{B}) = n$ which is trivially equivalent to $\mathscr{C}(\mathbf{A} : \mathbf{B}) = \mathbb{R}^n$.

Assume then that (5.122) has a solution $\mathbf{Y}$. Then $\mathbf{YA} = \mathbf{0}$ implies that

$$\mathscr{C}(\mathbf{A}) \subset \mathscr{N}(\mathbf{Y}) , \tag{5.125}$$

while $\mathbf{YB} = \mathbf{B}$ implies that $\operatorname{rank}(\mathbf{B}'\mathbf{Y}') = \operatorname{rank}(\mathbf{B})$, and hence

$$\operatorname{rank}(\mathbf{B}) - \dim \mathscr{C}(\mathbf{B}) \cap \mathscr{C}(\mathbf{Y}')^{\perp} = \operatorname{rank}(\mathbf{B}) , \tag{5.126}$$

i.e.,

$$\mathscr{C}(\mathbf{B}) \cap \mathscr{N}(\mathbf{Y}) = \{\mathbf{0}\} . \tag{5.127}$$

Now (5.125) and (5.127) together imply the disjointness condition (5.124).

To go the other way, assume that the disjointness condition (5.124) holds. Then, in light of (5.15) (p. 123),

$$\mathscr{C}(\mathbf{B}') = \mathscr{C}[\mathbf{B}'(\mathbf{I}_n - \mathbf{P_A})]\,, \tag{5.128}$$

which implies the existence of a matrix **F** such that

$$\mathbf{B}' = \mathbf{B}'(\mathbf{I}_n - \mathbf{P_A})\mathbf{F}'. \tag{5.129}$$

The choice $\mathbf{Y} = \mathbf{F}(\mathbf{I}_n - \mathbf{P_A})$ clearly satisfies (5.122), and hence the proof is completed. □

We recall that the existence of a matrix **G** satisfying the fundamental BLUE equation (0.202) (p. 41), i.e.,

$$\mathbf{G}(\mathbf{X} : \mathbf{VM}) = (\mathbf{X} : \mathbf{0})\,, \tag{5.130}$$

is now guaranteed by Proposition 5.8.

5.13 Some Properties of $\mathscr{C}(\mathbf{X})^{\perp}_{\mathbf{V}^{-1}}$

Proposition 5.9. *Consider the linear model* $\{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \mathbf{V}\}$ *and let* $\mathscr{C}(\mathbf{X})^{\perp}_{\mathbf{V}^{-1}}$ *denote the set of vectors which are orthogonal to every vector in* $\mathscr{C}(\mathbf{X})$ *with respect to the inner product matrix* $\mathbf{V}^{-1}$. *Then the following sets are identical:*

(a) $\mathscr{C}(\mathbf{X})^{\perp}_{\mathbf{V}^{-1}}$, (b) $\mathscr{C}(\mathbf{V}\mathbf{X}^{\perp})$, (c) $\mathscr{N}(\mathbf{X}'\mathbf{V}^{-1})$,

(d) $\mathscr{C}(\mathbf{V}^{-1}\mathbf{X})^{\perp}$, (e) $\mathscr{N}(\mathbf{P}_{\mathbf{X};\mathbf{V}^{-1}})$, (f) $\mathscr{C}(\mathbf{I}_n - \mathbf{P}_{\mathbf{X};\mathbf{V}^{-1}})$.

Denote

$$\mathbf{W} = \mathbf{V} + \mathbf{XUX}', \quad \textit{where } \mathscr{C}(\mathbf{W}) = \mathscr{C}(\mathbf{X} : \mathbf{V})\,. \tag{5.131}$$

Then

$$\mathscr{C}(\mathbf{V}\mathbf{X}^{\perp}) = \mathscr{C}(\mathbf{W}^{-}\mathbf{X} : \mathbf{I}_n - \mathbf{W}^{-}\mathbf{W})^{\perp}, \tag{5.132}$$

where $\mathbf{W}^{-}$ *is an arbitrary (but fixed) generalized inverse of* **W**. *The column space* $\mathscr{C}(\mathbf{V}\mathbf{X}^{\perp})$ *can be expressed also as*

$$\mathscr{C}(\mathbf{V}\mathbf{X}^{\perp}) = \mathscr{C}\big[(\mathbf{W}^{-})'\mathbf{X} : \mathbf{I}_n - (\mathbf{W}^{-})'\mathbf{W}'\big]^{\perp}. \tag{5.133}$$

Moreover, let **V** *be possibly singular and assume that* $\mathscr{C}(\mathbf{X}) \subset \mathscr{C}(\mathbf{V})$. *Then*

$$\mathscr{C}(\mathbf{V}\mathbf{X}^{\perp}) = \mathscr{C}(\mathbf{V}^{-}\mathbf{X} : \mathbf{I}_n - \mathbf{V}^{-}\mathbf{V})^{\perp} \subset \mathscr{C}(\mathbf{V}^{-}\mathbf{X})^{\perp}, \tag{5.134}$$

where the inclusion becomes equality if and only if **V** *is positive definite.*

Proof. The parts (a)–(f) are merely simple repetitions from (2.73) in Proposition 2.6 (p. 81). Notice that the idempotency of $\mathbf{P}_{\mathbf{X};\mathbf{V}^{-1}}$ implies immediately the equality of (e) and (f). Obviously $\mathscr{C}(\mathbf{VM}) \subset \mathscr{N}(\mathbf{P}_{\mathbf{X};\mathbf{V}^{-1}})$ and $\operatorname{rank}(\mathbf{VM}) = \operatorname{rank}(\mathbf{M})$, while

$$\dim \mathscr{N}(\mathbf{P}_{\mathbf{X};\mathbf{V}^{-1}}) = n - \operatorname{rank}(\mathbf{P}_{\mathbf{X};\mathbf{V}^{-1}}) = n - \operatorname{rank}(\mathbf{X}) = \operatorname{rank}(\mathbf{M}) , \tag{5.135}$$

and thereby $\mathscr{C}(\mathbf{VM}) = \mathscr{N}(\mathbf{P}_{\mathbf{X};\mathbf{V}^{-1}})$.

Let us next prove (5.132), i.e.,

$$\mathscr{C}(\mathbf{VM})^{\perp} = \mathscr{N}(\mathbf{MV}) = \mathscr{C}(\mathbf{W}^{-}\mathbf{X} : \mathbf{I}_n - \mathbf{W}^{-}\mathbf{W}) . \tag{5.136}$$

Because $\mathscr{C}(\mathbf{X}) \subset \mathscr{C}(\mathbf{W})$, there exists a matrix $\mathbf{A}$ such that $\mathbf{X} = \mathbf{WA}$ and hence

$$\mathbf{WW}^{-}\mathbf{X} = \mathbf{WW}^{-}\mathbf{WA} = \mathbf{WA} = \mathbf{X} , \tag{5.137}$$

and so

$$\begin{aligned} \mathbf{MVW}^{-}\mathbf{X} &= \mathbf{M}(\mathbf{V} + \mathbf{XUX}')\mathbf{W}^{-}\mathbf{X} \\ &= \mathbf{MWW}^{-}\mathbf{X} = \mathbf{MX} = \mathbf{0} . \end{aligned} \tag{5.138}$$

Similarly,

$$\mathbf{MV}(\mathbf{I}_n - \mathbf{W}^{-}\mathbf{W}) = \mathbf{MW}(\mathbf{I}_n - \mathbf{W}^{-}\mathbf{W}) = \mathbf{0} , \tag{5.139}$$

which together with (5.138) implies that

$$\mathscr{C}(\mathbf{W}^{-}\mathbf{X} : \mathbf{I}_n - \mathbf{W}^{-}\mathbf{W}) \subset \mathscr{N}(\mathbf{MV}) = \mathscr{C}(\mathbf{VM})^{\perp} . \tag{5.140}$$

The disjointness

$$\mathscr{C}(\mathbf{W}^{-}\mathbf{X}) \cap \mathscr{C}(\mathbf{I}_n - \mathbf{W}^{-}\mathbf{W}) = \{\mathbf{0}\} \tag{5.141}$$

is seen by taking

$$\mathbf{u} = \mathbf{W}^{-}\mathbf{Xa} = (\mathbf{I}_n - \mathbf{W}^{-}\mathbf{W})\mathbf{b} \quad \text{for some } \mathbf{a}, \mathbf{b} , \tag{5.142}$$

and premultiplying (5.142) by $\mathbf{W}$. It remains to prove that

$$\dim \mathscr{C}(\mathbf{VM})^{\perp} = \operatorname{rank}(\mathbf{W}^{-}\mathbf{X}) + \operatorname{rank}(\mathbf{I}_n - \mathbf{W}^{-}\mathbf{W}) . \tag{5.143}$$

Now $\dim \mathscr{C}(\mathbf{VM})^{\perp} = n - \operatorname{rank}(\mathbf{VM})$, and

$$\begin{aligned} \operatorname{rank}(\mathbf{I}_n - \mathbf{W}^{-}\mathbf{W}) &= n - \operatorname{rank}(\mathbf{W}) = n - \operatorname{rank}(\mathbf{X} : \mathbf{V}) \\ &= n - [\operatorname{rank}(\mathbf{X}) + \operatorname{rank}(\mathbf{VM})] , && (5.144\text{a}) \\ \operatorname{rank}(\mathbf{W}^{-}\mathbf{X}) &= \operatorname{rank}(\mathbf{X}) , && (5.144\text{b}) \end{aligned}$$

where the last equality comes from

$$\operatorname{rank}(\mathbf{X}) \geq \operatorname{rank}(\mathbf{W}^{-}\mathbf{X}) \geq \operatorname{rank}(\mathbf{WW}^{-}\mathbf{X}) = \operatorname{rank}(\mathbf{X}) . \tag{5.145}$$

Summing up (5.144) proves (5.132), the result due to Rao (1973b, p. 278). Result (5.134) follows from (5.132).

To prove the representation (5.133), we utilize Proposition 12.1 (p. 286), which says, for example, that for $\mathbf{W} = \mathbf{V} + \mathbf{XUX}'$ the following holds:

$$\mathscr{C}(\mathbf{X} : \mathbf{V}) = \mathscr{C}(\mathbf{W}) \iff \mathscr{C}(\mathbf{X} : \mathbf{V}) = \mathscr{C}(\mathbf{W}') \,. \tag{5.146}$$

Using (5.146), we get immediately

$$\mathscr{C}(\mathbf{VX}^{\perp})^{\perp} = \mathscr{C}\big[(\mathbf{W}')^{-}\mathbf{X} : \mathbf{I}_n - (\mathbf{W}')^{-}\mathbf{W}'\big] \,. \tag{5.147}$$

Noting that one choice for $(\mathbf{W}')^{-}$ is $(\mathbf{W}^{-})'$, we obtain (5.133). □

Notice that using $\mathscr{C}(\mathbf{X}^{\perp}_{\mathbf{V}^{-1}}) = \mathscr{C}(\mathbf{VM})$, (2.74) and (2.75) (p. 82):

$$\begin{aligned}\mathbf{I}_n - \mathbf{P}_{\mathbf{X};\mathbf{V}^{-1}} &= \mathbf{P}_{\mathbf{X}^{\perp}_{\mathbf{V}^{-1}};\mathbf{V}^{-1}} = \mathbf{P}_{\mathbf{VM};\mathbf{V}^{-1}} \\ &= \mathbf{VM}(\mathbf{MVM})^{-}\mathbf{M} = \mathbf{P}'_{\mathbf{M};\mathbf{V}} \,, \end{aligned} \tag{5.148}$$

i.e.,

$$\mathbf{P}_{\mathbf{X};\mathbf{V}^{-1}} = \mathbf{I}_n - \mathbf{P}_{\mathbf{VM};\mathbf{V}^{-1}} = \mathbf{I}_n - \mathbf{P}'_{\mathbf{M};\mathbf{V}} \,. \tag{5.149}$$

It is of interest to note that the “perp” symbol $\perp$ falls down, so to say, very “nicely” when $\mathbf{V}$ is positive definite:

$$\mathscr{C}(\mathbf{VX}^{\perp})^{\perp} = \mathscr{C}(\mathbf{V}^{-1}\mathbf{X}) \,, \tag{5.150}$$

but when $\mathbf{V}$ is singular we have to use a much more complicated rule to fall down the $\perp$ symbol:

$$\mathscr{C}(\mathbf{VX}^{\perp})^{\perp} = \mathscr{C}(\mathbf{W}^{-}\mathbf{X} : \mathbf{I}_n - \mathbf{W}^{-}\mathbf{W}) \,, \tag{5.151}$$

where $\mathbf{W}$ is defined as in (5.131).

In passing we note that according to Proposition 5.6 (p. 137) the $\mathscr{C}(\mathbf{VX}^{\perp})$ is precisely the column space of the covariance matrix of the BLUE’s residual:

$$\mathscr{C}(\mathbf{VX}^{\perp}) = \mathscr{C}[\operatorname{cov}(\tilde{\boldsymbol{\varepsilon}})] = \mathscr{C}[\operatorname{cov}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})] \,. \tag{5.152}$$

The result (5.151) offers us one way to introduce one particular representation for the matrix $\mathbf{G} \in \{\mathbf{P}_{\mathbf{X}|\mathbf{VM}}\}$, i.e.,

$$\mathbf{G}(\mathbf{X} : \mathbf{VM}) = (\mathbf{X} : \mathbf{0}) \,. \tag{5.153}$$

Now (5.153) implies that $(\mathbf{VM})'\mathbf{G}' = \mathbf{0}$ so that

$$\mathscr{C}(\mathbf{G}') \subset \mathscr{C}(\mathbf{VM})^{\perp} = \mathscr{C}\big[(\mathbf{W}^{-})'\mathbf{X} : \mathbf{I}_n - (\mathbf{W}^{-})'\mathbf{W}'\big] \,, \tag{5.154a}$$

$$\mathbf{G}' = (\mathbf{W}^{-})'\mathbf{XA}' + \big[\mathbf{I}_n - (\mathbf{W}^{-})'\mathbf{W}'\big]\mathbf{B}' \quad \text{for some } \mathbf{A}, \mathbf{B} \,, \tag{5.154b}$$

$$\mathbf{G} = \mathbf{AX}'\mathbf{W}^{-} + \mathbf{B}(\mathbf{I}_n - \mathbf{WW}^{-}) \quad \text{for some } \mathbf{A}, \mathbf{B} \,. \tag{5.154c}$$

Requesting $\mathbf{GX} = \mathbf{X}$ yields

$$\mathbf{AX'W^-X} + \mathbf{B}(\mathbf{I}_n - \mathbf{WW^-})\mathbf{X} = \mathbf{AX'W^-X} = \mathbf{X}\,, \tag{5.155}$$

because $\mathbf{WW^-X} = \mathbf{X}$. From (5.155) we can solve $\mathbf{A}$, one obvious solution being $\mathbf{A} = \mathbf{X}(\mathbf{X'W^-X})^-$; see Proposition 12.1 (p. 286), and thus

$$\mathbf{G} = \mathbf{X}(\mathbf{X'W^-X})^-\mathbf{X'W^-} \in \{\mathbf{P}_{\mathbf{X}|\mathbf{VM}}\}\,. \tag{5.156}$$

5.14 Exercises

5.1. Show that for conformable matrices we have

$$\operatorname{rank}\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{0} \end{pmatrix} = \operatorname{rank}(\mathbf{B}) + \operatorname{rank}(\mathbf{C}) + \operatorname{rank}\big[(\mathbf{I} - \mathbf{P_B})\mathbf{A}(\mathbf{I} - \mathbf{P_{C'}})\big]\,.$$

Marsaglia & Styan (1974a, Th. 19).

5.2. Show that the result of Proposition 5.6 (p. 137) can be written as

$$\mathscr{C}(\mathbf{GV}) = \mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{V})\,, \quad \text{and} \quad \mathscr{C}[(\mathbf{I}_n - \mathbf{G})\mathbf{V}] = \mathscr{C}(\mathbf{VM})\,,$$

where $\mathbf{Gy}$ is the BLUE of $\mathbf{X}\boldsymbol{\beta}$.

5.3. Denote, as in Section 5.6 (p. 130), $\mathbf{X} = \begin{pmatrix} \mathbf{X}_{(i)} \\ \mathbf{x}'_{(i)} \end{pmatrix}$, $\mathbf{H} = \mathbf{P_X}$, $\mathbf{i}_n$ = last column of $\mathbf{I}_n$. Show that $\mathbf{Hi}_n = \mathbf{i}_n$ if and only if $\mathbf{x}_{(i)}$ is a nonzero vector satisfying $\mathbf{x}_{(i)} \notin \mathscr{C}(\mathbf{X}'_{(i)})$. See also Theorem 16 (p. 343).

5.4. Consider the linear model $\{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$, where $\mathscr{C}(\mathbf{X}) \subset \mathscr{C}(\mathbf{V})$, and $\mathbf{X}$ has full column rank p. Confirm that

(a) $\mathbf{X'V^-X} = \mathbf{X'V^+X}$ for any $\mathbf{V}^-$,

(b) $\operatorname{rank}(\mathbf{X'V^-X}) = p$ for any $\mathbf{V}^-$, and so $\mathbf{X'V^-X}$ is invertible for any $\mathbf{V}^-$.

5.5. Using $\operatorname{rank}(\mathbf{AB}) = \operatorname{rank}(\mathbf{A}) - \dim \mathscr{C}(\mathbf{A}') \cap \mathscr{C}(\mathbf{B}^\perp)$ prove the Sylvester's inequality:

$$\operatorname{rank}(\mathbf{A}_{n\times p}\mathbf{B}_{p\times m}) \geq \operatorname{rank}(\mathbf{A}) + \operatorname{rank}(\mathbf{B}) - p\,,$$

and confirm that the equality holds if and only if $\mathscr{N}(\mathbf{A}) \subset \mathscr{C}(\mathbf{B})$.

5.6. Consider the linear model $\{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$ and denote $\mathbf{W} = \mathbf{V} + \mathbf{XUX}'$, where $\mathscr{C}(\mathbf{W}) = \mathscr{C}(\mathbf{X} : \mathbf{V})$, and let $\mathbf{W}^-$ be an arbitrary generalized inverse of $\mathbf{W}$. Confirm that then

$$\begin{gathered} \mathscr{C}(\mathbf{W^-X}) \oplus \mathscr{C}(\mathbf{X})^\perp = \mathbb{R}^n, \quad \mathscr{C}(\mathbf{W^-X})^\perp \oplus \mathscr{C}(\mathbf{X}) = \mathbb{R}^n, \\ \mathscr{C}[(\mathbf{W}^-)'\mathbf{X}] \oplus \mathscr{C}(\mathbf{X})^\perp = \mathbb{R}^n, \quad \mathscr{C}[(\mathbf{W}^-)'\mathbf{X}]^\perp \oplus \mathscr{C}(\mathbf{X}) = \mathbb{R}^n. \end{gathered}$$

Mitra & Moore (1973, p. 140).

5.7 (Continued ...). Prove the following:

$$\mathscr{C}(\mathbf{VM}) = \mathscr{C}(\mathbf{W}^{-}\mathbf{X})^{\perp} \iff \mathscr{C}(\mathbf{X} : \mathbf{V}) = \mathbb{R}^n.$$

5.8. Let $\mathbf{A} \in \mathbb{R}^{n\times a}$ and $\mathbf{B} \in \mathbb{R}^{n\times b}$, and denote $\mathbf{Q_A} = \mathbf{I}_n - \mathbf{P_A}$, $\mathbf{Q_B} = \mathbf{I}_n - \mathbf{P_B}$. Show that

$$\begin{aligned}\operatorname{rk}(\mathbf{P_A P_B Q_A}) &= \operatorname{rk}(\mathbf{P_A P_B}) + \operatorname{rk}(\mathbf{P_A} : \mathbf{P_B}) - \operatorname{rk}(\mathbf{A}) - \operatorname{rk}(\mathbf{B}) \\ &= \operatorname{rk}(\mathbf{P_A P_B}) + \operatorname{rk}(\mathbf{Q_A P_B}) - \operatorname{rk}(\mathbf{B}) \\ &= \operatorname{rk}(\mathbf{P_B P_A}) + \operatorname{rk}(\mathbf{Q_B P_A}) - \operatorname{rk}(\mathbf{A}) \\ &= \operatorname{rk}(\mathbf{P_B P_A Q_B}) .\end{aligned}$$

Drury, Liu, Lu, Puntanen et al. (2002, p. 463).

5.9. Let us denote, as in (5.57) (p. 130), $\mathbf{X} = \begin{pmatrix} \mathbf{X}_{(i)} \\ \mathbf{x}'_{(i)} \end{pmatrix}$, $\mathbf{i}_i$ = the last column of $\mathbf{I}_n$. Confirm the equivalence of the following statements; see also part (f) of Exercise 8.17 (p. 188), and Theorem 16 (p. 343).

(a) $\mathbf{x}_{(i)} \notin \mathscr{C}(\mathbf{X}'_{(i)})$, (b) $\mathbf{i}_i \in \mathscr{C}(\mathbf{X})$, (c) $\mathbf{x}'_{(i)}(\mathbf{X}'\mathbf{X})^{-}\mathbf{x}_{(i)} = 1$

(d) $\mathbf{H} = \mathbf{P_X} = \begin{pmatrix} \mathbf{P}_{\mathbf{X}_{(i)}} & \mathbf{0}_{n-1} \\ \mathbf{0}'_{n-1} & 1 \end{pmatrix}$, (e) $\mathbf{x}'_{(i)}(\mathbf{X}'\mathbf{X})^{-}\mathbf{x}_{(i)} = 1$.

Chapter 6
Rank Cancellation Rule

> *There's no sense in being precise*
> *when you don't even know what you're talking about.*
>
> John von Neumann

If $a \in \mathbb{R}$ and $y \in \mathbb{R}$ have property $ay \neq 0$, then trivially

$$\ell ay = may \implies \ell a = ma\,, \tag{6.1}$$

that is, we can cancel y from (6.1) (as well as a). For matrices, the corresponding cancellation does not work. However, there is a very handy trick, the rank cancellation rule, which allows cancellations for matrices in the style of (6.1). It seems, according to our experience, that this simple rule has not received so much appreciation in statistical literature as it earns.

Theorem 6 (Rank cancellation rule: RCR). *For any conformable matrices involved,*

$$\mathbf{LAY} = \mathbf{MAY} \textit{ and } \operatorname{rank}(\mathbf{AY}) = \operatorname{rank}(\mathbf{A}) \implies \mathbf{LA} = \mathbf{MA}\,, \tag{6.2}$$

$$\mathbf{DAM} = \mathbf{DAN} \textit{ and } \operatorname{rank}(\mathbf{DA}) = \operatorname{rank}(\mathbf{A}) \implies \mathbf{AM} = \mathbf{AN}\,. \tag{6.3}$$

Furthermore,

$$\operatorname{rank}(\mathbf{AY}) = \operatorname{rank}(\mathbf{A}) \implies \operatorname{rank}(\mathbf{KAY}) = \operatorname{rank}(\mathbf{KA}) \textit{ for all } \mathbf{K}\,. \tag{6.4}$$

Proof. In view of the rank rule (5.3) (p. 121) of the matrix product, we have

$$\operatorname{rank}(\mathbf{AY}) = \operatorname{rank}(\mathbf{A}) - \dim \mathscr{C}(\mathbf{A}') \cap \mathscr{N}(\mathbf{Y}')\,, \tag{6.5}$$

and so, assuming that $\operatorname{rank}(\mathbf{AY}) = \operatorname{rank}(\mathbf{A})$,

$$\mathscr{C}(\mathbf{A}') \cap \mathscr{N}(\mathbf{Y}') = \{\mathbf{0}\}\,. \tag{6.6}$$

Assumption $\mathbf{LAY} = \mathbf{MAY}$ can, of course, be rewritten as

$$\mathbf{Y}'(\mathbf{A}'\mathbf{M}' - \mathbf{A}'\mathbf{L}') = \mathbf{0}\,, \tag{6.7}$$

which is equivalent to

S. Puntanen et al., *Matrix Tricks for Linear Statistical Models: Our Personal Top Twenty*,
DOI 10.1007/978-3-642-10473-2_7, © Springer-Verlag Berlin Heidelberg 2011

$$\mathscr{C}(\mathbf{A}'\mathbf{M}' - \mathbf{A}'\mathbf{L}') \subset \mathscr{N}(\mathbf{Y}') \,. \tag{6.8}$$

On the other hand, obviously $\mathscr{C}(\mathbf{A}'\mathbf{M}' - \mathbf{A}'\mathbf{L}') \subset \mathscr{C}(\mathbf{A}')$, which, together with (6.8), implies that

$$\mathscr{C}(\mathbf{A}'\mathbf{M}' - \mathbf{A}'\mathbf{L}') \subset \mathscr{C}(\mathbf{A}') \cap \mathscr{N}(\mathbf{Y}') \,. \tag{6.9}$$

Using (6.6), our claim (6.2) is proved. The proofs of the other parts of the theorem go along the same lines. For a different proof of the RCR, using a full rank decomposition, see Section 17.4 (p. 354). □

The rank cancellation rule was introduced by Marsaglia & Styan (1974a, Th. 2). See also Rao & Bhimasankaram (2000, Th. 3.5.7).

6.1 Simple Properties of the Hat Matrix H

We consider two (well-known) results giving simple proofs based on the rank cancellation rule. Consider the following:

$$\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X} = \mathbf{X} \,, \quad \text{i.e.,} \quad \mathbf{H}\mathbf{X} = \mathbf{X} \,, \tag{6.10a}$$

$$\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}' \text{ is invariant for any choice of } (\mathbf{X}'\mathbf{X})^{-}. \tag{6.10b}$$

Of course, it would be trivial to prove (6.10a), if we simply were to use the fact that $\mathbf{H}$ is the orthogonal projector onto $\mathscr{C}(\mathbf{X})$. On the other hand, it is illustrative to consider first the identity

$$\underline{\mathbf{X}}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X} = \underline{\mathbf{X}}'\mathbf{X} \,. \tag{6.11}$$

Because $\operatorname{rank}(\mathbf{X}'\mathbf{X}) = \operatorname{rank}(\mathbf{X})$, we can, in light of (6.3), cancel the underlined terms, and hence (6.10a) follows. Similarly, consider the equation

$$\mathbf{X}'\mathbf{X} = \underline{\mathbf{X}}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\underline{\underline{\mathbf{X}}} = \underline{\mathbf{X}}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{\sim}\mathbf{X}'\underline{\underline{\mathbf{X}}} = \mathbf{X}'\mathbf{X} \,, \tag{6.12}$$

where $(\mathbf{X}'\mathbf{X})^{-}$ and $(\mathbf{X}'\mathbf{X})^{\sim}$ are two arbitrary generalized inverses. The underlined matrices can now be cancelled and thus (6.10b) is proved.

Let $\mathbf{V}$ be a positive definite matrix. Then, corresponding to (6.10), we have

$$\mathbf{X}(\mathbf{X}'\mathbf{V}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}\mathbf{X} = \mathbf{X} \,, \quad \text{i.e.,} \quad \mathbf{P}_{\mathbf{X};\mathbf{V}}\mathbf{X} = \mathbf{X} \,, \tag{6.13a}$$

$$\mathbf{X}(\mathbf{X}'\mathbf{V}\mathbf{X})^{-}\mathbf{X}'\mathbf{V} \text{ is invariant for any choice of } (\mathbf{X}'\mathbf{V}\mathbf{X})^{-}. \tag{6.13b}$$

Because $\operatorname{rank}(\mathbf{X}'\mathbf{V}\mathbf{X}) = \operatorname{rank}(\mathbf{X})$ for a positive definite $\mathbf{V}$, property (6.13a) comes from

$$\underline{\mathbf{X}'\mathbf{V}}\mathbf{X}(\mathbf{X}'\mathbf{V}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}\mathbf{X} = \underline{\mathbf{X}'\mathbf{V}}\mathbf{X} \tag{6.14}$$

by deleting the underlined terms.

Moreover, if $\mathbf{V}$ is nonnegative definite (possibly singular), then the RCR gives

$$\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{V}\mathbf{X})^-\mathbf{X}'\mathbf{V} = \mathbf{X}'\mathbf{V}\,, \tag{6.15a}$$

$$\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{V}\mathbf{X})^-\mathbf{X}'\mathbf{V} \text{ is invariant for any choice of } (\mathbf{X}'\mathbf{V}\mathbf{X})^-. \tag{6.15b}$$

Notice also that in view of $\operatorname{rank}(\mathbf{B}\mathbf{B}') = \operatorname{rank}(\mathbf{B})$, (6.4) implies that for any (conformable) matrix $\mathbf{Z}$ the following holds:

$$\operatorname{rank}(\mathbf{A}\mathbf{A}'\mathbf{Z}\mathbf{B}\mathbf{B}') = \operatorname{rank}(\mathbf{A}\mathbf{A}'\mathbf{Z}\mathbf{B})\,. \tag{6.16}$$

Similarly,

$$\operatorname{rank}(\mathbf{L}\mathbf{B}) \geq \operatorname{rank}(\mathbf{L}\mathbf{B}\mathbf{B}') \geq \operatorname{rank}(\mathbf{L}\mathbf{B}\mathbf{B}'\mathbf{L}') = \operatorname{rank}(\mathbf{L}\mathbf{B}) \tag{6.17}$$

for any $\mathbf{L}$, and hence

$$\begin{aligned}\operatorname{rank}(\mathbf{A}\mathbf{A}'\mathbf{Z}\mathbf{B}\mathbf{B}') &= \operatorname{rank}(\mathbf{A}'\mathbf{Z}\mathbf{B}\mathbf{B}') \\ &= \operatorname{rank}(\mathbf{A}'\mathbf{Z}\mathbf{B}) = \operatorname{rank}(\mathbf{A}\mathbf{A}'\mathbf{Z}\mathbf{B})\,. \end{aligned}\tag{6.18}$$

In particular, for a nonnegative definite $\mathbf{V} = \mathbf{L}\mathbf{L}'$ we have

$$\begin{aligned}\operatorname{rank}(\mathbf{A}\mathbf{A}'\mathbf{V}\mathbf{A}\mathbf{A}') &= \operatorname{rank}(\mathbf{A}'\mathbf{V}\mathbf{A}) = \operatorname{rank}(\mathbf{A}'\mathbf{L}\mathbf{L}'\mathbf{A}) \\ &= \operatorname{rank}(\mathbf{L}'\mathbf{A}) = \operatorname{rank}(\mathbf{L}\mathbf{L}'\mathbf{A}) = \operatorname{rank}(\mathbf{V}\mathbf{A})\,. \end{aligned}\tag{6.19}$$

6.2 Properties of $\mathbf{X}(\mathbf{X}'\mathbf{V}^+\mathbf{X})^-\mathbf{X}'\mathbf{V}^+\mathbf{y}$

Consider the linear model $\{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{V}\}$, and denote

$$\mathbf{G} = \mathbf{X}(\mathbf{X}'\mathbf{V}^+\mathbf{X})^-\mathbf{X}'\mathbf{V}^+ = \mathbf{X}\mathbf{L}\,, \tag{6.20}$$

where $(\mathbf{X}'\mathbf{V}^+\mathbf{X})^-$ is an arbitrary (but fixed) generalized inverse of $\mathbf{X}'\mathbf{V}^+\mathbf{X}$, and $\mathbf{L} = (\mathbf{X}'\mathbf{V}^+\mathbf{X})^-\mathbf{X}'\mathbf{V}^+$. The estimator $\mathbf{G}\mathbf{y}$ may be called a weighted least squares estimator (WLSE) of $\mathbf{X}\boldsymbol{\beta}$.

Proposition 6.1. *Let* $\mathbf{G}$ *be defined as in* (6.20)*, i.e.,*

$$\mathbf{G} = \mathbf{X}(\mathbf{X}'\mathbf{V}^+\mathbf{X})^-\mathbf{X}'\mathbf{V}^+ = \mathbf{X}\mathbf{L}\,. \tag{6.21}$$

Then the following five statements are equivalent:

(a) $\operatorname{rank}(\mathbf{X}'\mathbf{V}\mathbf{X}) = \operatorname{rank}(\mathbf{X})$,

(b) $\mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{V}^\perp) = \{\mathbf{0}\}$,

(c) $\mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{V}\mathbf{X})^\perp = \{\mathbf{0}\}$,

(d) $\mathbf{L} = (\mathbf{X}'\mathbf{V}^+\mathbf{X})^-\mathbf{X}'\mathbf{V}^+ \in \{\mathbf{X}^-\}$,

(e) $\mathbf{Gy} = \mathbf{X}(\mathbf{X}'\mathbf{V}^+\mathbf{X})^-\mathbf{X}'\mathbf{V}^+\mathbf{y}$ *is unbiased for* $\mathbf{X}\boldsymbol{\beta}$.

Moreover, the following three statements are equivalent:

(f) $\mathscr{C}(\mathbf{X}) \subset \mathscr{C}(\mathbf{V})$,

(g) $\mathbf{G} = \mathbf{X}(\mathbf{X}'\mathbf{V}^+\mathbf{X})^-\mathbf{X}'\mathbf{V}^+$ *satisfies the equation*

$$\mathbf{G}(\mathbf{X} : \mathbf{VM}) = (\mathbf{X} : \mathbf{0}), \tag{6.22}$$

that is, $\mathbf{Gy} = \mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta})$,

(h) $\mathbf{Gy} = \mathbf{X}(\mathbf{X}'\mathbf{V}^+\mathbf{X})^-\mathbf{X}'\mathbf{V}^+\mathbf{y}$ *is unbiased for* $\mathbf{X}\boldsymbol{\beta}$, *i.e.,*

$$\mathbf{GX} = \mathbf{X}, \tag{6.23}$$

and

$$\mathbf{X}'\mathbf{P}_{\mathbf{V}}\mathbf{M} = \mathbf{0}. \tag{6.24}$$

Proof. The equivalence of (a), (b) and (c) was demonstrated already in Proposition 5.4 (p. 132).

Next we prove that "(a) $\iff$ (e)". Statement (e) is equivalent to

$$\mathbf{X}(\mathbf{X}'\mathbf{V}^+\mathbf{X})^-\mathbf{X}'\mathbf{V}^+\mathbf{X} = \mathbf{X}, \tag{e}$$

which trivially is equivalent to (d). Consider the identity

$$\underline{\mathbf{X}'\mathbf{V}^+}\mathbf{X}(\mathbf{X}'\mathbf{V}^+\mathbf{X})^-\mathbf{X}'\mathbf{V}^+\mathbf{X} = \underline{\mathbf{X}'\mathbf{V}^+}\mathbf{X}. \tag{6.25}$$

If (a) holds we can cancel the underlined terms and hence (a) implies (e). To go the other way, (e) implies that

$$\mathrm{rank}(\mathbf{X}) \geq \mathrm{rank}(\mathbf{X}'\mathbf{V}^+) \geq \mathrm{rank}[\mathbf{X}(\mathbf{X}'\mathbf{V}^+\mathbf{X})^-\mathbf{X}'\mathbf{V}^+\mathbf{X}] = \mathrm{rank}(\mathbf{X}), \tag{6.26}$$

and hence $\mathrm{rank}(\mathbf{X}) = \mathrm{rank}(\mathbf{X}'\mathbf{V}^+) = \mathrm{rank}(\mathbf{X}'\mathbf{V})$, i.e., (a) holds.

To prove "(g) $\iff$ (h)", assume first that (g) holds and rewrite (6.22) as

$$\mathbf{X}(\mathbf{X}'\mathbf{V}^+\mathbf{X})^-\mathbf{X}'\mathbf{V}^+\mathbf{X} = \mathbf{X}, \tag{6.27a}$$

$$\mathbf{X}(\mathbf{X}'\mathbf{V}^+\mathbf{X})^-\mathbf{X}'\mathbf{V}^+\mathbf{VM} = \mathbf{0}. \tag{6.27b}$$

Because (why?)

$$\mathrm{rank}[\mathbf{X}(\mathbf{X}'\mathbf{V}^+\mathbf{X})^-\mathbf{X}'\mathbf{V}^+] = \mathrm{rank}(\mathbf{X}'\mathbf{V}^+), \tag{6.28}$$

we can cancel $\mathbf{X}(\mathbf{X}'\mathbf{V}^+\mathbf{X})^-$ from (6.27b) and write (6.27b) equivalently as

$$\mathbf{X}'\mathbf{V}^+\mathbf{VM} = \mathbf{X}'\mathbf{P}_{\mathbf{V}}\mathbf{M} = \mathbf{0}, \tag{6.29}$$

and thus we have proved that (g) and (h) are equivalent.

It remains to show that that (f) and (h) are equivalent. If (f) holds, then also (b) holds and thereby (6.27a) must hold. Furthermore, (f) clearly implies (6.27b), thus confirming that (f) implies (h). We still have to prove that (h), i.e., (6.27a) & (6.29), implies (f). To see this, note first that (6.29) can be equivalently expressed as

$$\mathscr{C}(\mathbf{P}_{\mathbf{V}}\mathbf{X}) \subset \mathscr{C}(\mathbf{X})\,. \tag{6.30}$$

Condition (6.27a) means that

$$\operatorname{rank}(\mathbf{X}) = \operatorname{rank}(\mathbf{X}'\mathbf{V}^{+}) = \operatorname{rank}(\mathbf{X}'\mathbf{V}) = \operatorname{rank}(\mathbf{X}'\mathbf{P}_{\mathbf{V}})\,, \tag{6.31}$$

which, together with (6.30), implies

$$\mathscr{C}(\mathbf{P}_{\mathbf{V}}\mathbf{X}) = \mathscr{C}(\mathbf{X})\,. \tag{6.32}$$

Moreover, it is obvious that $\mathscr{C}(\mathbf{P}_{\mathbf{V}}\mathbf{X}) \subset \mathscr{C}(\mathbf{V})$, which, together with (6.32) yields (f). □

The linear model

$$\mathscr{M} = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{V}\}\,, \quad \text{where } \mathscr{C}(\mathbf{X}) \subset \mathscr{C}(\mathbf{V})\,, \tag{6.33}$$

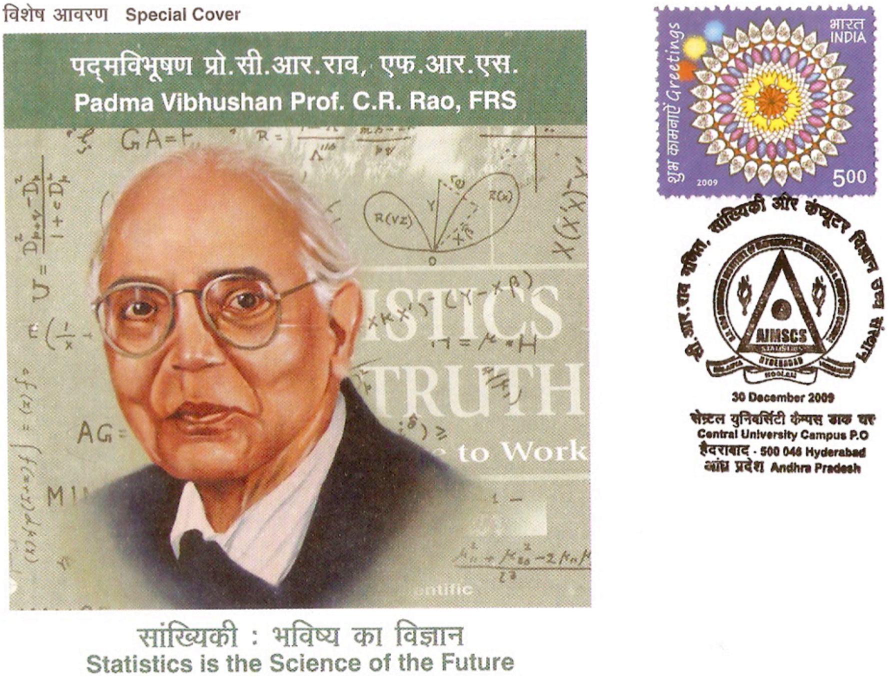
is often called a *weakly singular* linear model. In view of the invariance Theorem 12 (p. 283) and the consistency condition $\mathbf{y} \in \mathscr{C}(\mathbf{V})$, it can be shown that under a weakly singular linear model $\{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{V}\}$, the estimator

$$\mathbf{X}(\mathbf{X}'\mathbf{V}^{-}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-}\mathbf{y} \quad \text{is invariant for all g-inverses involved.} \tag{6.34}$$

For comparison of the OLSE, BLUE, and WLSE, the reader is referred to Baksalary & Kala (1983a), and Lowerre (1974).

Further References

Zyskind & Martin (1969),
Baksalary & Kala (1983a),
Baksalary & Puntanen (1989),
Baksalary, Puntanen & Styan (1990b),
Puntanen & Scott (1996),
Searle (1994),
Yanai (1990).



Philatelic Item 6.1 Calyampudi Radhakrishna Rao, FRS, (b. 1920) is a world-renowned Indian-born statistician and National Medal of Science Awardee. *The Times of India* chose C. R. Rao as one of the top 10 Indian Scientists of all time. Shown here is a Special Cover from India Post: Hyderabad, 30 December 2009.

Chapter 7
Sum of Orthogonal Projectors

> *I speak, by the way, not with any sense of futility, but with a panicky conviction of the absolute meaninglessness of existence which could easily be misinterpreted as pessimism.*
>
> Woody Allen: *My Speech to the Graduates*

In this chapter we give a well-known result regarding the sum of two orthogonal projectors. It is a short chapter, but nevertheless, we believe that it is worthwhile to keep this result in the gallery of active tools when dealing with matrices in statistics. This chapter is strongly related to the next chapter on a decomposition of orthogonal projector, beginning on page 155.

Just as in Chapter 2, we first consider a simpler situation, i.e., the projectors under the standard inner product, and only after that take a look at the more general case in Proposition 7.2 (p. 151).

Theorem 7 (Sum of orthogonal projectors). *Let* $\mathbf{P_A}$ *and* $\mathbf{P_B}$ *be orthogonal projectors (with respect to the standard inner product) onto* $\mathscr{C}(\mathbf{A})$ *and* $\mathscr{C}(\mathbf{B})$*, respectively. Then*

$$\mathbf{P_A} + \mathbf{P_B} \textit{ is an orthogonal projector} \iff \mathbf{A}'\mathbf{B} = \mathbf{0}\,, \tag{7.1}$$

in which case

$$\mathbf{P_A} + \mathbf{P_B} = \mathbf{P}_{(\mathbf{A}\,:\,\mathbf{B})}\,. \tag{7.2}$$

Proof. The matrix

$$\mathbf{P} := \mathbf{P_A} + \mathbf{P_B} \tag{7.3}$$

is an orthogonal projector if and only if it is idempotent and symmetric. Since $\mathbf{P_A}$ and $\mathbf{P_B}$ are symmetric, $\mathbf{P}$ is trivially symmetric. Idempotency holds if

$$\mathbf{P}^2 = \mathbf{P_A} + \mathbf{P_B} + \mathbf{P_A}\mathbf{P_B} + \mathbf{P_B}\mathbf{P_A} = \mathbf{P_A} + \mathbf{P_B} = \mathbf{P}\,, \tag{7.4}$$

that is, $\mathbf{P}$ is idempotent (and thereby an orthogonal projector) if and only if

$$\mathbf{P_A}\mathbf{P_B} = -\mathbf{P_B}\mathbf{P_A}\,. \tag{7.5}$$

Postmultiplying (7.5) by $\mathbf{P_B}$ yields $\mathbf{P_A}\mathbf{P_B} = -\mathbf{P_B}\mathbf{P_A}\mathbf{P_B}$ which shows that $\mathbf{P_A}\mathbf{P_B}$ is symmetric: $(\mathbf{P_A}\mathbf{P_B})' = \mathbf{P_A}\mathbf{P_B}$, i.e., $\mathbf{P_B}\mathbf{P_A} = \mathbf{P_A}\mathbf{P_B}$. Hence we must have

S. Puntanen et al., *Matrix Tricks for Linear Statistical Models: Our Personal Top Twenty*,
DOI 10.1007/978-3-642-10473-2_8,

$$\mathbf{P_A P_B} = -\mathbf{P_A P_B}\,, \tag{7.6}$$

which means that

$$\mathbf{P_A P_B} = \mathbf{0}\,. \tag{7.7}$$

Premultiplying (7.7) by $\mathbf{A}'$ and postmultiplying by $\mathbf{B}$ yields $\mathbf{A}'\mathbf{B} = \mathbf{0}$, and thus we have proved that (7.5) implies $\mathbf{A}'\mathbf{B} = \mathbf{0}$. The proof of the reverse implication "$\mathbf{A}'\mathbf{B} = \mathbf{0} \implies$ (7.5)" is obvious.

To prove (7.2), we have to show that

$$\mathscr{C}(\mathbf{P_A} + \mathbf{P_B}) = \mathscr{C}(\mathbf{A} : \mathbf{B})\,, \tag{7.8}$$

which comes directly from

$$\mathscr{C}(\mathbf{P_A} + \mathbf{P_B}) = \mathscr{C}\left[(\mathbf{P_A} : \mathbf{P_B})\begin{pmatrix}\mathbf{P_A}\\ \mathbf{P_B}\end{pmatrix}\right] = \mathscr{C}(\mathbf{P_A} : \mathbf{P_B}) = \mathscr{C}(\mathbf{A} : \mathbf{B})\,. \tag{7.9}$$

□

Proposition 7.1 (Difference of orthogonal projectors). *Let* $\mathbf{P_A}$ *and* $\mathbf{P_B}$ *be orthogonal projectors (with respect to the standard inner product) onto* $\mathscr{C}(\mathbf{A})$ *and* $\mathscr{C}(\mathbf{B})$*, respectively. Then the following five statements are equivalent:*

(a) $\mathbf{P_A} - \mathbf{P_B}$ *is an orthogonal projector,*

(b) $\mathbf{P_A P_B} = \mathbf{P_B P_A} = \mathbf{P_B}$,

(c) $\mathscr{C}(\mathbf{B}) \subset \mathscr{C}(\mathbf{A})$,

(d) $\|\mathbf{P_A x}\| \geq \|\mathbf{P_B x}\|$ *for all* $\mathbf{x}$,

(e) $\mathbf{P_A} - \mathbf{P_B} \geq_{\mathsf{L}} \mathbf{0}$.

If any of the above conditions hold, then

$$\mathbf{P_A} - \mathbf{P_B} = \mathbf{P}_{(\mathbf{I}-\mathbf{P_B})\mathbf{A}} = \mathbf{P}_{\mathscr{C}(\mathbf{A})\cap\mathscr{C}(\mathbf{B})^{\perp}}\,. \tag{7.10}$$

Proof. It is easy to confirm that (a), ..., (e) are equivalent, the proof is left as an exercise. It remains to show that if any of them holds then

$$\mathscr{C}(\mathbf{P_A} - \mathbf{P_B}) = \mathscr{C}[(\mathbf{I} - \mathbf{P_B})\mathbf{P_A}] = \mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B}^{\perp})\,. \tag{7.11}$$

Because

$$\mathbf{P_A} - \mathbf{P_B} = (\mathbf{I} - \mathbf{P_B})\mathbf{P_A} = \mathbf{P_A}(\mathbf{I} - \mathbf{P_B})\,, \tag{7.12}$$

we can conclude at once that the first equality in (7.11) holds. Moreover, (7.12) implies that

$$\mathscr{C}(\mathbf{P_A} - \mathbf{P_B}) \subset \mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B}^{\perp})\,. \tag{7.13}$$

In view of (c) and (5.3) (p. 121), we have

$$\begin{aligned}\dim \mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B}^{\perp}) &= \operatorname{rank}(\mathbf{A}) - \operatorname{rank}(\mathbf{A}'\mathbf{B}) \\ &= \operatorname{rank}(\mathbf{A}) - \operatorname{rank}(\mathbf{P_A}\mathbf{B}) \\ &= \operatorname{rank}(\mathbf{A}) - \operatorname{rank}(\mathbf{B}) \,. \end{aligned} \tag{7.14}$$

In light of $\mathscr{C}(\mathbf{B}) \subset \mathscr{C}(\mathbf{A})$, we get

$$\begin{aligned}\operatorname{rank}(\mathbf{P_A} - \mathbf{P_B}) &= \operatorname{rank}[\mathbf{P_A}(\mathbf{I} - \mathbf{P_B})] \\ &= \operatorname{rank}(\mathbf{A}) - \dim \mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B}) \\ &= \operatorname{rank}(\mathbf{A}) - \operatorname{rank}(\mathbf{B}) \,, \end{aligned} \tag{7.15}$$

and thus the second equality in (7.11) is proved. □

Let us next consider the following generalization of Theorem 7.

Proposition 7.2 (Sum of orthogonal projectors with respect to the inner product matrix V). *Let* $\mathbf{P}_{\mathbf{A};\mathbf{V}}$ *and* $\mathbf{P}_{\mathbf{B};\mathbf{V}}$ *be orthogonal projectors with respect to the positive definite inner product matrix* $\mathbf{V}$ *onto* $\mathscr{C}(\mathbf{A})$ *and* $\mathscr{C}(\mathbf{B})$*, respectively. Then*

$$\mathbf{P}_{\mathbf{A};\mathbf{V}} + \mathbf{P}_{\mathbf{B};\mathbf{V}} \textit{ is an orthogonal projector } \iff \mathbf{A}'\mathbf{V}\mathbf{B} = \mathbf{0} \,, \tag{7.16}$$

in which case $\mathbf{P}_{\mathbf{A};\mathbf{V}} + \mathbf{P}_{\mathbf{B};\mathbf{V}} = \mathbf{P}_{(\mathbf{A}\,:\,\mathbf{B});\mathbf{V}}$.

Proof. Let us simplify the notation so that $\mathbf{P_A} = \mathbf{P}_{\mathbf{A};\mathbf{V}}$ and $\mathbf{P_B} = \mathbf{P}_{\mathbf{B};\mathbf{V}}$. Then, according to Proposition 2.6 (p. 81), $\mathbf{P} = \mathbf{P_A} + \mathbf{P_B}$ is orthogonal projector with respect to $\mathbf{V}$ if and only if

$$\mathbf{P}^2 = \mathbf{P} \quad \text{and} \quad \mathbf{V}\mathbf{P} = \mathbf{P}'\mathbf{V}. \tag{7.17}$$

The only essential difference with the proof of Theorem 7 concerns now the treatment of the idempotency condition of (7.17), which yields

$$\mathbf{P_A}\mathbf{P_B} = -\mathbf{P_B}\mathbf{P_A} \,. \tag{7.18}$$

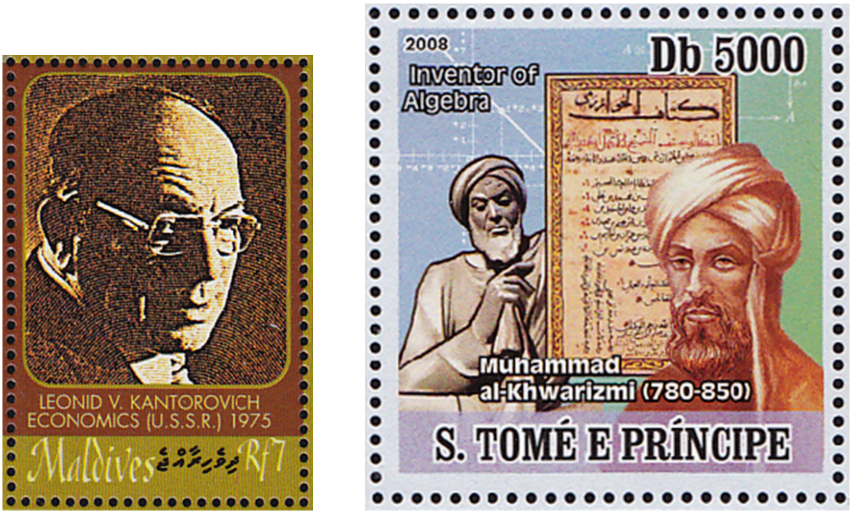
Premultiplying (7.18) by $\mathbf{V}$ gives $\mathbf{V}\mathbf{P_A}\mathbf{P_B} = -\mathbf{V}\mathbf{P_B}\mathbf{P_A}$, i.e.,

$$\mathbf{P}_\mathbf{A}'\mathbf{V}\mathbf{P_B} = -\mathbf{P}_\mathbf{B}'\mathbf{V}\mathbf{P_A} \,, \tag{7.19}$$

where we have used $\mathbf{V}\mathbf{P_A} = \mathbf{P}_\mathbf{A}'\mathbf{V}$ and $\mathbf{V}\mathbf{P_B} = \mathbf{P}_\mathbf{B}'\mathbf{V}$. Pre- and postmultiplying (7.19) by $\mathbf{P}_\mathbf{A}'$ and $\mathbf{P_B}$, respectively, yields

$$\mathbf{P}_\mathbf{A}'\mathbf{V}\mathbf{P_B} = -\mathbf{P}_\mathbf{A}'\mathbf{P}_\mathbf{B}'\mathbf{V}\mathbf{P_A}\mathbf{P_B} = \mathbf{P}_\mathbf{A}'\mathbf{P}_\mathbf{B}'\mathbf{V}\mathbf{P_B}\mathbf{P_A} \,, \tag{7.20}$$

where we have used (7.18). From (7.20) we see that $\mathbf{P}_\mathbf{A}'\mathbf{V}\mathbf{P_B}$ is symmetric. Hence, on account of (7.19), $\mathbf{P}_\mathbf{A}'\mathbf{V}\mathbf{P_B} = -\mathbf{P}_\mathbf{A}'\mathbf{V}\mathbf{P_B}$ which implies that $\mathbf{P}_\mathbf{A}'\mathbf{V}\mathbf{P_B} = \mathbf{0}$, i.e., $\mathbf{A}'\mathbf{V}\mathbf{B} = \mathbf{0}$. □



Philatelic Item 7.1 Shown on the stamp (left panel) from the Republic of Maldives issued in 2001 (*Scott* 2571e), Leonid Vitaliyevich Kantorovich (1912–1986) was a Soviet/Russian mathematician and economist, known for his theory and development of techniques for the optimal allocation of resources. He was the winner of the Nobel Memorial Prize in Economic Sciences in 1975. For the Kantorovich inequality see Section 20.3 (p. 418). Shown on the stamp (right panel) from the Democratic Republic of São Tomé and Príncipe issued in 2008 is Abū Abdallāh Muḥammad ibn Mūsā al-Khwārizmī (c. 780–c. 850), a Persian mathematician, astronomer and geographer and scholar in the House of Wisdom in Baghdad. He is considered the founder of algebra, a credit he shares with Diophantus of Alexandria.

Philatelic Item 7.2 Pierre-Simon marquis de Laplace (1749–1827) was a French mathematician and astronomer whose work was pivotal to the development of mathematical astronomy and statistics. For the Laplace expansion of a determinant, see page 429. The stamp from France (left panel) was issued in 1983 (*Scott* 2565) and the stamp (right panel) from the Republic of Guinea was issued in 2008.

Chapter 8
A Decomposition of the Orthogonal Projector

> *Many a young man starts in life with a natural gift for exaggeration which, if nurtured in congenial and sympathetic surroundings, or by the imitation of the best models, might grow into something really great and wonderful. But, as a rule, he comes to nothing. – He either falls into careless habits of accuracy or takes to frequenting the society of the aged and the well-informed.*
>
> Oscar Wilde: *The Decay of Lying*

In this chapter we consider a decomposition of the orthogonal projector onto the column space of the partitioned matrix $(\mathbf{A} : \mathbf{B})$ and demonstrate its usefulness through several examples. The decomposition introduced is a consequence of Theorem 7 (p. 151).

Theorem 8 (A decomposition of orthogonal projector). *The orthogonal projector* (*with respect to the standard inner product*) *onto*

$$\mathscr{C}(\mathbf{A}_{n\times a} : \mathbf{B}_{n\times b}) \tag{8.1}$$

can be decomposed as

$$\mathbf{P}_{(\mathbf{A}\,:\,\mathbf{B})} = \mathbf{P}_{\mathbf{A}} + \mathbf{P}_{(\mathbf{I}_n - \mathbf{P}_{\mathbf{A}})\mathbf{B}} = \mathbf{P}_{\mathbf{A}} + \mathbf{P}_{\mathscr{C}(\mathbf{A}\,:\,\mathbf{B})\cap\mathscr{C}(\mathbf{A})^{\perp}}\,. \tag{8.2}$$

Proof. We see at once that the sum $\mathbf{P}_{\mathbf{A}} + \mathbf{P}_{(\mathbf{I}-\mathbf{P}_{\mathbf{A}})\mathbf{B}}$ is an orthogonal projector since $\mathbf{A}'(\mathbf{I} - \mathbf{P}_{\mathbf{A}})\mathbf{B} = \mathbf{0}$. It remains to show that

$$\mathscr{C}(\mathbf{A} : \mathbf{B}) = \mathscr{C}[\mathbf{A} : (\mathbf{I}_n - \mathbf{P}_{\mathbf{A}})\mathbf{B}]\,, \tag{8.3}$$

which follows from the obvious column space inclusion

$$\mathscr{C}[\mathbf{A} : (\mathbf{I}_n - \mathbf{P}_{\mathbf{A}})\mathbf{B}] \subset \mathscr{C}(\mathbf{A} : \mathbf{B})\,, \tag{8.4}$$

and the rank equality (5.2) (p. 121). We might recall that according to (5.5),

$$(\mathbf{A} : \mathbf{B})\mathbf{L} = (\mathbf{A} : \mathbf{B}) \begin{pmatrix} \mathbf{I}_a & -\mathbf{A}^{+}\mathbf{B} \\ \mathbf{0} & \mathbf{I}_b \end{pmatrix} = [\mathbf{A} : (\mathbf{I}_n - \mathbf{P}_{\mathbf{A}})\mathbf{B}]\,. \tag{8.5}$$

The matrix $\mathbf{L}$ is nonsingular and hence (8.5) would immediately imply (8.3).

The second equality in (8.2) follows by applying (7.10) (p. 152) to the difference $\mathbf{P}_{(\mathbf{A}\,:\,\mathbf{B})} - \mathbf{P}_{\mathbf{A}}$. It could also be proved by confirming that

$$\mathscr{C}[(\mathbf{I} - \mathbf{P}_{\mathbf{A}})\mathbf{B}] = \mathscr{C}(\mathbf{A} : \mathbf{B}) \cap \mathscr{C}(\mathbf{A})^{\perp}. \tag{8.6}$$

S. Puntanen et al., *Matrix Tricks for Linear Statistical Models: Our Personal Top Twenty*,
DOI 10.1007/978-3-642-10473-2_9, © Springer-Verlag Berlin Heidelberg 2011

Notice that denoting $\mathbf{Q_A} = \mathbf{I} - \mathbf{P_A}$, we get immediately

$$\mathbf{I} - \mathbf{P}_{(\mathbf{A}\,:\,\mathbf{B})} = \mathbf{Q_A} - \mathbf{P}_{\mathbf{Q_A}\mathbf{B}} = \mathbf{Q_A}(\mathbf{I} - \mathbf{P}_{\mathbf{Q_A}\mathbf{B}})\,. \tag{8.7}$$

Proposition 8.1 (A decomposition of orthogonal projector with respect to the inner product matrix V). *Let* $\mathbf{V}$ *be a symmetric positive definite matrix* (*of appropriate dimension*). *Then*

$$\mathbf{P}_{(\mathbf{A}\,:\,\mathbf{B});\mathbf{V}} = \mathbf{P}_{\mathbf{A};\mathbf{V}} + \mathbf{P}_{(\mathbf{I}-\mathbf{P}_{\mathbf{A};\mathbf{V}})\mathbf{B};\mathbf{V}}\,. \tag{8.8}$$

Proof. The proof is parallel to Theorem 8 by first noting that

$$\mathbf{A}'\mathbf{V}(\mathbf{I} - \mathbf{P}_{\mathbf{A};\mathbf{V}}) = \mathbf{0}\,, \quad \text{i.e.,} \quad \mathbf{P}'_{\mathbf{A};\mathbf{V}}\mathbf{V}\mathbf{A} = \mathbf{V}\mathbf{A}\,. \tag{8.9}$$

That (8.9) indeed holds, see Proposition 2.6 (p. 81). □

Let us denote $\mathbf{Q}_{\mathbf{A};\mathbf{V}} = \mathbf{I} - \mathbf{P}_{\mathbf{A};\mathbf{V}} = \mathbf{P}'_{\mathbf{Z};\mathbf{V}^{-1}}\,, \mathbf{Z} \in \{\mathbf{A}^{\perp}\}$, where in view of (2.75) (p. 82), or part (i) of Theorem 15 (p. 319),

$$\mathbf{P}_{\mathbf{A};\mathbf{V}} = \mathbf{A}(\mathbf{A}'\mathbf{V}\mathbf{A})^{-}\mathbf{A}'\mathbf{V} = \mathbf{I} - \mathbf{V}^{-1}\mathbf{Z}(\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z})^{-}\mathbf{Z}' = \mathbf{I} - \mathbf{P}'_{\mathbf{Z};\mathbf{V}^{-1}}\,. \tag{8.10}$$

Then we can introduce the following explicit representation for $\mathbf{P}_{(\mathbf{I}-\mathbf{P}_{\mathbf{A};\mathbf{V}})\mathbf{B};\mathbf{V}}$:

$$\begin{aligned}
\mathbf{P}_{(\mathbf{I}-\mathbf{P}_{\mathbf{A};\mathbf{V}})\mathbf{B};\mathbf{V}} &= \mathbf{Q}_{\mathbf{A};\mathbf{V}}\mathbf{B}(\mathbf{B}'\mathbf{Q}'_{\mathbf{A};\mathbf{V}}\mathbf{V}\mathbf{Q}_{\mathbf{A};\mathbf{V}}\mathbf{B})^{-}\mathbf{B}'\mathbf{Q}'_{\mathbf{A};\mathbf{V}}\mathbf{V} \\
&= \mathbf{Q}_{\mathbf{A};\mathbf{V}}\mathbf{B}(\mathbf{B}'\mathbf{V}\mathbf{Q}_{\mathbf{A};\mathbf{V}}\mathbf{B})^{-}\mathbf{B}'\mathbf{Q}'_{\mathbf{A};\mathbf{V}}\mathbf{V} \\
&= \mathbf{P}'_{\mathbf{Z};\mathbf{V}^{-1}}\mathbf{B}(\mathbf{B}'\mathbf{V}\mathbf{P}'_{\mathbf{Z};\mathbf{V}^{-1}}\mathbf{B})^{-}\mathbf{B}'\mathbf{P}_{\mathbf{Z};\mathbf{V}^{-1}}\mathbf{V} \\
&= \mathbf{P}'_{\mathbf{Z};\mathbf{V}^{-1}}\mathbf{B}[\mathbf{B}'\mathbf{Z}(\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z})^{-}\mathbf{Z}'\mathbf{B}]^{-}\mathbf{B}'\mathbf{Z}(\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z})^{-}\mathbf{Z}' \\
&= \mathbf{V}^{-1}\dot{\mathbf{Z}}\mathbf{B}(\mathbf{B}'\dot{\mathbf{Z}}\mathbf{B})^{-}\mathbf{B}'\dot{\mathbf{Z}}\,,
\end{aligned} \tag{8.11}$$

where $\dot{\mathbf{Z}} = \mathbf{Z}(\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z})^{-}\mathbf{Z}'$.

References to Projectors

Baksalary (1987),
Baksalary & Kala (1979),
Baksalary & Trenkler (2009b),
Ben-Israel & Greville (2003, §2.7),
Bryant (1984),
Chipman & Rao (1964),
Galántai (2004, 2008),
Groß & Trenkler (1998),
Halmos (1951, 1958),
Kala (1981, 2008),
Kala & Pordzik (2006),
Mitra (1993),
Mitra & Rao (1974),
Mäkeläinen (1970b),
Rao (1974),
Rao & Mitra (1971b, §5.1),
Rao & Yanai (1979),
Seber (1980), (1984, Appendix B),
Seber & Lee (2003, Appendix B),
Sibuya (1970),
Takane & Yanai (1999),
Takeuchi, Yanai & Mukherjee (1982, §1.3),
Trenkler (1994, 2006).

8.1 Mahalanobis Distance & the Hat Matrix H

Let the $n \times p$ model matrix $\mathbf{X}$ be partitioned as

$$\mathbf{X} = (\mathbf{1} : \mathbf{x}_1 : \ldots : \mathbf{x}_k) = (\mathbf{1} : \mathbf{X}_0) = \begin{pmatrix} 1 & \mathbf{x}'_{(1)} \\ 1 & \mathbf{x}'_{(2)} \\ \vdots & \vdots \\ 1 & \mathbf{x}'_{(n)} \end{pmatrix}; \quad p = k + 1\,. \tag{8.12}$$

Here $\mathbf{x}_{(i)} \in \mathbb{R}^k$ represents the ith case or the ith observation in the observation space, and $\mathbf{x}_i \in \mathbb{R}^n$ represents the ith variable in the variable space. Let $\bar{\mathbf{x}} \in \mathbb{R}^k$ denote the mean vector of the x-variables and $\mathbf{S}_{\mathbf{xx}}$ the sample covariance matrix:

$$\bar{\mathbf{x}} = \tfrac{1}{n}(\mathbf{x}_{(1)} + \mathbf{x}_{(2)} + \cdots + \mathbf{x}_{(n)}) = \tfrac{1}{n}\mathbf{X}'_0\mathbf{1} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_k \end{pmatrix}, \tag{8.13a}$$

$$\mathbf{S}_{\mathbf{xx}} = \tfrac{1}{n-1}\mathbf{X}'_0\mathbf{C}\mathbf{X}_0 = \tfrac{1}{n-1}\mathbf{T}_{\mathbf{xx}} = \tfrac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_{(i)} - \bar{\mathbf{x}})(\mathbf{x}_{(i)} - \bar{\mathbf{x}})', \tag{8.13b}$$

where $\mathbf{C} = \mathbf{I}_n - \mathbf{P}_{\mathbf{1}}$ is the centering matrix. Then the squared Mahalanobis distance of the observation $\mathbf{x}_{(i)}$ from the mean vector $\bar{\mathbf{x}}$ is defined as

$$\mathrm{MHLN}^2(\mathbf{x}_{(i)}, \bar{\mathbf{x}}, \mathbf{S}_{\mathbf{xx}}) = (\mathbf{x}_{(i)} - \bar{\mathbf{x}})'\mathbf{S}_{\mathbf{xx}}^{-1}(\mathbf{x}_{(i)} - \bar{\mathbf{x}})\,. \tag{8.14}$$

We will show that the ith element of the hat matrix $\mathbf{H}$ has the following interesting connection to the Mahalanobis distance:

Proposition 8.2. *The ith element of the hat matrix $\mathbf{H}$ can be expressed as*

$$h_{ii} = \tfrac{1}{n} + \tfrac{1}{n-1}\,\mathrm{MHLN}^2(\mathbf{x}_{(i)}, \bar{\mathbf{x}}, \mathbf{S}_{\mathbf{xx}})\,. \tag{8.15}$$

Moreover,

$$\mathrm{MHLN}^2(\mathbf{x}_{(i)}, \bar{\mathbf{x}}, \mathbf{S}_{\mathbf{xx}}) = (n-1)\tilde{h}_{ii}\,, \quad \textit{where } \widetilde{\mathbf{H}} = \mathbf{P}_{\mathbf{C}\mathbf{X}_0}\,. \tag{8.16}$$

Proof. Using the decomposition (8.2), we can write

$$\begin{aligned} \mathbf{H} &= \mathbf{P}_{(\mathbf{1}\,:\,\mathbf{X}_0)} = \mathbf{P}_{\mathbf{1}} + \mathbf{P}_{\mathbf{C}\mathbf{X}_0} \\ &= \mathbf{J} + \mathbf{C}\mathbf{X}_0(\mathbf{X}'_0\mathbf{C}\mathbf{X}_0)^{-1}\mathbf{X}'_0\mathbf{C} \\ &= \mathbf{J} + \mathbf{C}\mathbf{X}_0\mathbf{T}_{\mathbf{xx}}^{-1}\mathbf{X}'_0\mathbf{C} \\ &= \mathbf{J} + \tfrac{1}{n-1}\mathbf{C}\mathbf{X}_0\mathbf{S}_{\mathbf{xx}}^{-1}\mathbf{X}'_0\mathbf{C} \\ &= \mathbf{J} + \tfrac{1}{n-1}\widetilde{\mathbf{X}}_0\mathbf{S}_{\mathbf{xx}}^{-1}\widetilde{\mathbf{X}}'_0\,, \end{aligned} \tag{8.17}$$

where $\widetilde{\mathbf{X}}_0 = \mathbf{C}\mathbf{X}_0$ is the centered $\mathbf{X}_0$:

$$\widetilde{\mathbf{X}}_0 = \mathbf{C}\mathbf{X}_0 = \begin{pmatrix} (\mathbf{x}_{(1)} - \bar{\mathbf{x}})' \\ (\mathbf{x}_{(2)} - \bar{\mathbf{x}})' \\ \vdots \\ (\mathbf{x}_{(n)} - \bar{\mathbf{x}})' \end{pmatrix} . \tag{8.18}$$

Pre- and postmultiplying (8.17) by $\mathbf{i}_i'$ and $\mathbf{i}_i$, respectively, yields

$$h_{ii} = \tfrac{1}{n} + \tfrac{1}{n-1}(\mathbf{x}_{(i)} - \bar{\mathbf{x}})'\mathbf{S}_{\mathbf{xx}}^{-1}(\mathbf{x}_{(i)} - \bar{\mathbf{x}}) . \tag{8.19}$$

In view of (8.17), we have $\widetilde{\mathbf{X}}_0\mathbf{S}_{\mathbf{xx}}^{-1}\widetilde{\mathbf{X}}_0' = (n-1)\mathbf{P}_{\mathbf{C}\mathbf{X}_0}$, and hence the ith squared Mahalanobis distance is $(n-1)\tilde{h}_{ii}$, where $\tilde{h}_{ii}$ is the ith diagonal element of the orthogonal projector $\mathbf{P}_{\mathbf{C}\mathbf{X}_0} = \widetilde{\mathbf{H}}$; see also Section 20.5 (p. 420). □

In (8.14) the covariance matrix $\mathbf{S}_{\mathbf{xx}}$ is positive definite, but if $\mathbf{S}_{\mathbf{xx}}$ is singular, then we can define the squared Mahalanobis distance as

$$\mathrm{MHLN}^2(\mathbf{x}_{(i)}, \bar{\mathbf{x}}, \mathbf{S}_{\mathbf{xx}}) = (\mathbf{x}_{(i)} - \bar{\mathbf{x}})'\mathbf{S}_{\mathbf{xx}}^{+}(\mathbf{x}_{(i)} - \bar{\mathbf{x}}) , \tag{8.20}$$

which is actually invariant for the choice of the $\mathbf{S}_{\mathbf{xx}}^{-}$; see page 66.

For the use of hat matrix in statistics, see, e.g., Morgenthaler & Clerc-Bérod (1997), Chatterjee & Hadi (1986, 1988), and Hoaglin & Welsch (1978).

8.2 Subvector of $\hat{\boldsymbol{\beta}}$

Let us consider the partitioned linear model

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} , \tag{8.21}$$

or in other notation,

$$\mathscr{M}_{12} = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{V}\} = \{\mathbf{y},\, \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2,\, \sigma^2\mathbf{V}\} , \tag{8.22}$$

where $\mathbf{X}_1$ is an $n \times p_1$ matrix and $\mathbf{X}_2$ is an $n \times p_2$; $p = p_1 + p_2$. We also denote

$$\mathscr{M}_1 = \{\mathbf{y},\, \mathbf{X}_1\boldsymbol{\beta},\, \sigma^2\mathbf{V}\} , \quad \mathscr{M}_2 = \{\mathbf{y},\, \mathbf{X}_2\boldsymbol{\beta},\, \sigma^2\mathbf{V}\} . \tag{8.23}$$

Models $\mathscr{M}_1$ and $\mathscr{M}_2$ can be called small models while $\mathscr{M}_{12}$ is a full model. When $\operatorname{rank}(\mathbf{X}) = p$, the ordinary least squares estimator of $\boldsymbol{\beta}$ under $\mathscr{M}_{12}$ is

$$\mathrm{OLSE}(\boldsymbol{\beta}) = \hat{\boldsymbol{\beta}}(\mathscr{M}_{12}) = \hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{X}^{+}\mathbf{y} , \tag{8.24}$$

where the notation $\hat{\boldsymbol\beta}(\mathscr{M}_{12})$ emphasizes that the estimator is calculated under the model $\mathscr{M}_{12}$. We repeat again that if $\mathbf{X}$ has no full column rank, then $\hat{\boldsymbol\beta}$ simply denotes an arbitrary solution to the normal equation $\mathbf{X}'\mathbf{X}\boldsymbol\beta = \mathbf{X}'\mathbf{y}$ and may not be considered as a proper estimator. If $\mathbf{X}$ has no full column rank, we still use the notation

$$\mathrm{OLSE}(\boldsymbol\beta) = \hat{\boldsymbol\beta} = \begin{pmatrix} \hat{\boldsymbol\beta}_1 \\ \hat{\boldsymbol\beta}_2 \end{pmatrix} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}\,. \tag{8.25}$$

Varying $(\mathbf{X}'\mathbf{X})^{-}$ we can generate all expressions for $\hat{\boldsymbol\beta}$. Note that $\mathrm{OLSE}(\mathbf{X}\boldsymbol\beta) = \mathbf{H}\mathbf{y}$ is unique (once $\mathbf{y}$ is given) and

$$\mathrm{OLSE}(\mathbf{X}\boldsymbol\beta) = \mathbf{H}\mathbf{y} = \mathbf{X}_1\hat{\boldsymbol\beta}_1 + \mathbf{X}_2\hat{\boldsymbol\beta}_2\,, \tag{8.26}$$

where $\mathbf{X}_i\hat{\boldsymbol\beta}_i$ and $\hat{\boldsymbol\beta}_i$ are not necessarily unique.

We may recall that if $\mathbf{K}$ is a $p\times q$ matrix, then the OLSE of $\mathbf{K}'\boldsymbol\beta$ is defined as

$$\mathrm{OLSE}(\mathbf{K}'\boldsymbol\beta) = \mathbf{K}'\cdot\mathrm{OLSE}(\boldsymbol\beta) = \mathbf{K}'\hat{\boldsymbol\beta}\,. \tag{8.27}$$

We might use the notation $\mathrm{OLSE}(\mathbf{K}'\boldsymbol\beta) = \widehat{\mathbf{K}'\boldsymbol\beta}$, but we prefer that one used in (8.27).

When dealing with regression models, it is often very helpful to have the explicit expression for

$$\hat{\boldsymbol\beta}_2 \quad\text{or}\quad \mathbf{X}_2\hat{\boldsymbol\beta}_2 \quad\text{or}\quad \mathbf{K}_2\hat{\boldsymbol\beta}_2,\ \text{where } \mathbf{K}_2 \in \mathbb{R}^{q\times p_2}. \tag{8.28}$$

However, such expressions are not necessarily unique. According to Section 12.2 (p. 284), and Proposition 16.1 (p. 345), $\mathbf{X}_2\hat{\boldsymbol\beta}_2$ is invariant with respect to the choice of $\hat{\boldsymbol\beta}_2$ if and only if $\mathbf{X}_2\boldsymbol\beta_2$ is estimable. Moreover, $\mathbf{K}_2\hat{\boldsymbol\beta}_2$ is unique if and only if $\mathbf{K}_2\boldsymbol\beta_2$ is estimable. It can be seen that

$$\mathbf{K}_2\boldsymbol\beta_2 \text{ is estimable} \Leftrightarrow \mathscr{C}(\mathbf{K}_2') \subset \mathscr{C}(\mathbf{X}_2'\mathbf{M}_1) \Leftrightarrow \exists\,\mathbf{F}\colon \mathbf{K}_2 = \mathbf{F}\mathbf{M}_1\mathbf{X}_2\,, \tag{8.29}$$

where

$$\mathbf{M}_1 = \mathbf{I}_n - \mathbf{P}_{\mathbf{X}_1}\,,\quad \mathbf{P}_{\mathbf{X}_1} = \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-}\mathbf{X}_1'\,. \tag{8.30}$$

Now

$$\{\,\mathbf{K}_2\boldsymbol\beta_2 : \mathbf{K}_2\boldsymbol\beta_2 \text{ is estimable}\,\} = \{\,\mathbf{F}\mathbf{M}_1\mathbf{X}_2\boldsymbol\beta_2 : \mathbf{F} \in \mathbb{R}^{q\times n}\,\}\,. \tag{8.31}$$

Trivially $\mathbf{M}_1\mathbf{X}_2\boldsymbol\beta_2$ is estimable and thereby $\mathbf{M}_1\mathbf{X}_2\hat{\boldsymbol\beta}_2$ is unique for any choice of $\hat{\boldsymbol\beta}_2$.

The estimability concept will be needed in the following proposition, where we will give short & simple proofs for the expressions $\mathbf{M}_1\mathbf{X}_2\hat{\boldsymbol\beta}_2$, $\mathbf{X}_2\hat{\boldsymbol\beta}_2$ and $\hat{\boldsymbol\beta}_2$; see also Proposition 16.1 (p. 345).

Proposition 8.3. *Consider the partitioned linear model* $\mathscr{M}_{12}$. *Then* $\mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2$ *is estimable and its* OLSE *is*

$$\mathbf{M}_1\mathbf{X}_2\hat{\boldsymbol{\beta}}_2 = \mathbf{M}_1\mathbf{X}_2(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^-\mathbf{X}_2'\mathbf{M}_1\mathbf{y} = \mathbf{M}_1\mathbf{X}_2\hat{\boldsymbol{\beta}}_2(\mathscr{M}_{12})\,. \tag{8.32}$$

Let $\mathbf{X}_2\boldsymbol{\beta}_2$ *be estimable under the model* $\mathscr{M}_{12}$, *i.e.,*

$$\operatorname{rank}(\mathbf{M}_1\mathbf{X}_2) = \operatorname{rank}(\mathbf{X}_2)\,, \ \textit{or equivalently}\ \mathscr{C}(\mathbf{X}_1)\cap\mathscr{C}(\mathbf{X}_2) = \{\mathbf{0}\}\,. \tag{8.33}$$

Then the OLSE *of* $\mathbf{X}_2\boldsymbol{\beta}_2$ *is*

$$\mathbf{X}_2\hat{\boldsymbol{\beta}}_2 = \mathbf{X}_2(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^-\mathbf{X}_2'\mathbf{M}_1\mathbf{y} = \mathbf{X}_2\hat{\boldsymbol{\beta}}_2(\mathscr{M}_{12})\,. \tag{8.34}$$

Moreover, if $\boldsymbol{\beta}_2$ *is estimable under the model* $\mathscr{M}_{12}$, *i.e.,*

$$\operatorname{rank}(\mathbf{M}_1\mathbf{X}_2) = \operatorname{rank}(\mathbf{X}_2) = p_2\,, \tag{8.35}$$

then

$$\hat{\boldsymbol{\beta}}_2 = (\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{M}_1\mathbf{y} = \hat{\boldsymbol{\beta}}_2(\mathscr{M}_{12})\,. \tag{8.36}$$

Proof. We may first note that, in view of (5.3) (p. 121),

$$\begin{aligned}\operatorname{rank}(\mathbf{M}_1\mathbf{X}_2) &= \operatorname{rank}(\mathbf{X}_2) - \dim\mathscr{C}(\mathbf{X}_2)\cap\mathscr{C}(\mathbf{M}_1^\perp)\\ &= \operatorname{rank}(\mathbf{X}_2) - \dim\mathscr{C}(\mathbf{X}_2)\cap\mathscr{C}(\mathbf{X}_1)\,,\end{aligned} \tag{8.37}$$

and hence indeed

$$\operatorname{rank}(\mathbf{M}_1\mathbf{X}_2) = \operatorname{rank}(\mathbf{X}_2) \iff \mathscr{C}(\mathbf{X}_1)\cap\mathscr{C}(\mathbf{X}_2) = \{\mathbf{0}\}\,. \tag{8.38}$$

Note also that the expression (8.34) is invariant with respect to the choice of $(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^-$; see Theorem 12 (p. 283) and Proposition 16.1 (p. 345).

To introduce (8.32), (8.34) and (8.36), we first write, in light of the projector decomposition (8.2) (p. 155),

$$\mathbf{H} = \mathbf{P}_{(\mathbf{X}_1\,:\,\mathbf{X}_2)} = \mathbf{P}_{\mathbf{X}_1} + \mathbf{P}_{\mathbf{M}_1\mathbf{X}_2} = \mathbf{P}_{\mathbf{X}_1} + \mathbf{P}_{\mathbf{M}_1\mathbf{X}_2}\,, \tag{8.39a}$$

$$\mathbf{H}\mathbf{y} = \mathbf{X}_1\hat{\boldsymbol{\beta}}_1 + \mathbf{X}_2\hat{\boldsymbol{\beta}}_2 = \mathbf{P}_{\mathbf{X}_1}\mathbf{y} + \mathbf{P}_{\mathbf{M}_1\mathbf{X}_2}\mathbf{y}\,. \tag{8.39b}$$

Premultiplying (8.39b) by $\mathbf{M}_1$ gives

$$\mathbf{M}_1\mathbf{X}_2\hat{\boldsymbol{\beta}}_2 = \mathbf{P}_{\mathbf{M}_1\mathbf{X}_2}\mathbf{y} = \mathbf{M}_1\mathbf{X}_2(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^-\mathbf{X}_2'\mathbf{M}_1\mathbf{y}\,. \tag{8.40}$$

Now if $\operatorname{rank}(\mathbf{M}_1\mathbf{X}_2) = \operatorname{rank}(\mathbf{X}_2)$ holds, we can cancel, in view of the rank cancellation rule (p. 145), the left-most $\mathbf{M}_1$ and thus (8.34) follows. Similarly, condition $\operatorname{rank}(\mathbf{M}_1\mathbf{X}_2) = p_2$ implies (8.36), and thus our claims are proved. □

Note that if the condition $\operatorname{rank}(\mathbf{M}_1\mathbf{X}_2) = \operatorname{rank}(\mathbf{X}_2)$ holds, then $\mathbf{H}\mathbf{y}$ can be expressed as

$$\mathbf{Hy} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}_1(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^-\mathbf{X}_1'\mathbf{M}_2\mathbf{y} + \mathbf{X}_2(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^-\mathbf{X}_2'\mathbf{M}_1\mathbf{y}\,. \tag{8.41}$$

Next we show the following related result.

Proposition 8.4. *If* $\operatorname{rank}(\mathbf{M}_1\mathbf{X}_2) = \operatorname{rank}(\mathbf{X}_2)$, *then*

$$\mathbf{X}_1\hat{\boldsymbol{\beta}}_1(\mathscr{M}_{12}) = \mathbf{X}_1\hat{\boldsymbol{\beta}}_1(\mathscr{M}_1) - \mathbf{P}_{\mathbf{X}_1}\mathbf{X}_2\hat{\boldsymbol{\beta}}_2(\mathscr{M}_{12})\,, \tag{8.42}$$

and if, in addition to $\operatorname{rank}(\mathbf{M}_1\mathbf{X}_2) = \operatorname{rank}(\mathbf{X}_2)$, *the matrix* $\mathbf{X}_1$ *has full column rank, then*

$$\hat{\boldsymbol{\beta}}_1(\mathscr{M}_{12}) = \hat{\boldsymbol{\beta}}_1(\mathscr{M}_1) - (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\hat{\boldsymbol{\beta}}_2(\mathscr{M}_{12})\,. \tag{8.43}$$

The expressions (8.42) and (8.43) offer a possibility to check the effect of additional columns of $\mathbf{X}$ (i.e., additional explanatory variables) on the regression coefficients.

Proof (of Proposition 8.4). To prove (8.42), we rewrite (8.39b)

$$\mathbf{X}_1\hat{\boldsymbol{\beta}}_1(\mathscr{M}_{12}) + \mathbf{X}_2\hat{\boldsymbol{\beta}}_2(\mathscr{M}_{12}) = \mathbf{X}_1\hat{\boldsymbol{\beta}}_1(\mathscr{M}_1) + \mathbf{M}_1\mathbf{X}_2\hat{\boldsymbol{\beta}}_2(\mathscr{M}_{12})\,, \tag{8.44}$$

which can be rearranged as

$$\mathbf{X}_1\hat{\boldsymbol{\beta}}_1(\mathscr{M}_{12}) = \mathbf{X}_1\hat{\boldsymbol{\beta}}_1(\mathscr{M}_1) - \mathbf{P}_{\mathbf{X}_1}\mathbf{X}_2\hat{\boldsymbol{\beta}}_2(\mathscr{M}_{12})\,, \tag{8.45}$$

and thus (8.42) is obtained. Clearly we get (8.43) if—in addition to disjointness condition (8.33)—the matrix $\mathbf{X}_1$ has full column rank. Note that $\mathbf{X}_2$ does not necessarily need to have full column rank. □

Notice that one way express the decomposition (8.39b) is the following:

$$\operatorname{OLSE}(\mathbf{X}\boldsymbol{\beta} \mid \mathscr{M}_{12}) = \operatorname{OLSE}(\mathbf{X}_1\boldsymbol{\beta}_1 \mid \mathscr{M}_1) + \operatorname{OLSE}(\mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2 \mid \mathscr{M}_{12})\,. \tag{8.46}$$

We may also note that the covariance matrix of $\hat{\boldsymbol{\beta}}_2$ under $\{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{V}\}$ is

$$\operatorname{cov}(\hat{\boldsymbol{\beta}}_2 \mid \mathscr{M}_{12}) = \sigma^2(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{M}_1\mathbf{V}\mathbf{M}_1\mathbf{X}_2(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}, \tag{8.47}$$

and hence

$$\begin{aligned}\mathbf{V} = \mathbf{I}_n \implies \operatorname{cov}(\hat{\boldsymbol{\beta}}_2 \mid \mathscr{M}_{12}) &= \sigma^2(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}\\ &\geq_{\mathsf{L}} \sigma^2(\mathbf{X}_2'\mathbf{X}_2)^{-1} = \operatorname{cov}(\hat{\boldsymbol{\beta}}_2 \mid \mathscr{M}_2)\,. \end{aligned}\tag{8.48}$$

8.3 A Single Regression Coefficient

The formula for a single regression coefficient is now easily available. Let $\mathbf{X}$ be partitioned as

$$\mathbf{X} = (\mathbf{1} : \mathbf{x}_1 : \ldots : \mathbf{x}_{k-1} : \mathbf{x}_k) := (\mathbf{X}_1 : \mathbf{x}_k)\,, \tag{8.49}$$

where

$$\mathbf{x}_k \notin \mathscr{C}(\mathbf{X}_1)\,, \tag{8.50}$$

which is a necessary and sufficient condition for the estimability of β_k. Then

$$\hat{\beta}_k = \frac{\mathbf{x}_k'\mathbf{M}_1\mathbf{y}}{\mathbf{x}_k'\mathbf{M}_1\mathbf{x}_k}\,. \tag{8.51}$$

If $\mathbf{V} = \mathbf{I}$, we have

$$\operatorname{var}(\hat{\beta}_k) = \frac{\sigma^2}{\mathbf{x}_k'\mathbf{M}_1\mathbf{x}_k} = \frac{\sigma^2}{\|(\mathbf{I} - \mathbf{P}_{\mathbf{X}_1})\mathbf{x}_k\|^2}\,, \tag{8.52}$$

and so we may expect problems in estimation if the column $\mathbf{x}_k$ is “almost” in the column space of $\mathbf{X}_1$. If we consider such a model $\mathscr{M}_k$ where $\mathbf{x}_k$ is explained by all other x variables (plus a constant), that is, the model matrix is $\mathbf{X}_1$,

$$\mathscr{M}_k = \{\mathbf{x}_k,\, \mathbf{X}_1\boldsymbol{\gamma},\, \sigma^2\mathbf{I}\}\,, \tag{8.53}$$

then the residual sum of squares and the total sum of squares are

$$\operatorname{SSE}(k) = \mathbf{x}_k'\mathbf{M}_1\mathbf{x}_k\,, \quad \operatorname{SST}(k) = \mathbf{x}_k'\mathbf{C}\mathbf{x}_k = t_{kk}\,, \tag{8.54}$$

and so (assuming that $t_{kk} > 0$)

$$\begin{aligned} \operatorname{SSE}(k) &= \operatorname{SST}(k) - \operatorname{SSR}(k) = \operatorname{SST}(k)\Big(1 - \frac{\operatorname{SSR}(k)}{\operatorname{SST}(k)}\Big) \\ &= \operatorname{SST}(k)(1 - R_k^2) = t_{kk}(1 - R_k^2)\,, \end{aligned} \tag{8.55}$$

where $R_k^2 = R_{k\cdot 1\ldots k-1}^2$ is the coefficient of determination when x_k is explained by all other x_i’s. Hence the variance of $\hat{\beta}_k$ can be expressed as

$$\operatorname{var}(\hat{\beta}_k) = \frac{1}{1 - R_k^2} \cdot \frac{1}{t_{kk}} \cdot \sigma^2 := \frac{\operatorname{VIF}_k}{t_{kk}} \cdot \sigma^2, \tag{8.56}$$

where

$$\operatorname{VIF}_k = \frac{1}{1 - R_k^2} \tag{8.57}$$

refers to the so-called *variance inflation factor*. As Belsley (1991, p. 29) states, much interesting work on collinearity diagnostics has been done with VIF-like measures, although the concept masquerades under various names; see, e.g., Stewart (1987).

8.4 The Reduced Model & the Frisch–Waugh–Lovell Theorem

Let us consider the full rank partitioned linear model

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}\,, \tag{8.58}$$

or in other notation, $\mathscr{M}_{12} = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}\}$. Now we know that

$$\hat{\boldsymbol{\beta}}_2(\mathscr{M}_{12}) = (\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{M}_1\mathbf{y}\,. \tag{8.59}$$

Premultiplying (8.58) by the orthogonal projector $\mathbf{M}_1$ yields the model

$$\mathscr{M}_{12\cdot 1} = \{\mathbf{M}_1\mathbf{y},\, \mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2,\, \sigma^2\mathbf{M}_1\}\,. \tag{8.60}$$

Taking a look at the models, we can immediately make an important conclusion: the OLS estimators of $\boldsymbol{\beta}_2$ under the models $\mathscr{M}_{12}$ and $\mathscr{M}_{12\cdot 1}$ coincide:

$$\hat{\boldsymbol{\beta}}_2(\mathscr{M}_{12}) = \hat{\boldsymbol{\beta}}_2(\mathscr{M}_{12\cdot 1}) = (\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{M}_1\mathbf{y}\,. \tag{8.61}$$

The equality (8.61) is just the result that Davidson & MacKinnon (1993, 2004, §1.4, §2.4) call the *Frisch–Waugh–Lovell theorem* "…since those papers seem to have introduced, and then reintroduced, it to econometricians"; see Frisch & Waugh (1933), Lovell (1963, 2008), and Baltagi (2000) and (2008, §7.3). Geometric illustration is in Figure 8.6 (p. 184). For the generalization of the Frisch–Waugh–Lovell theorem, see Section 15.4 (p. 328).

Representation (8.61) is valid if (and only if)

$$\operatorname{rank}(\mathbf{M}_1\mathbf{X}_2) = p_2\,, \tag{8.62}$$

which is the estimability condition for $\boldsymbol{\beta}_2$ under $\mathscr{M}_{12}$ and also under $\mathscr{M}_{12\cdot 1}$. However, in both models the parametric function $\mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2$ is always estimable and

$$\mathbf{M}_1\mathbf{X}_2\hat{\boldsymbol{\beta}}_2(\mathscr{M}_{12}) = \mathbf{M}_1\mathbf{X}_2\hat{\boldsymbol{\beta}}_2(\mathscr{M}_{12\cdot 1})\,. \tag{8.63}$$

Notice that according to (8.46), we have the decomposition

$$\operatorname{OLSE}(\mathbf{X}\boldsymbol{\beta} \mid \mathscr{M}_{12}) = \operatorname{OLSE}(\mathbf{X}_1\boldsymbol{\beta}_1 \mid \mathscr{M}_1) + \operatorname{OLSE}(\mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2 \mid \mathscr{M}_{12})\,. \tag{8.64}$$

This can now be expressed also as

$$\operatorname{OLSE}(\mathbf{X}\boldsymbol{\beta} \mid \mathscr{M}_{12}) = \operatorname{OLSE}(\mathbf{X}_1\boldsymbol{\beta}_1 \mid \mathscr{M}_1) + \operatorname{OLSE}(\mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2 \mid \mathscr{M}_{12\cdot 1})\,. \tag{8.65}$$

We call $\mathscr{M}_{12\cdot 1}$ a *reduced model*; Groß & Puntanen (2000a) call it a correctly reduced model. In the reduced model,

- the response variable

$$\mathbf{M}_1\mathbf{y} = \operatorname{res}(\mathbf{y}; \mathbf{X}_1) \tag{8.66}$$

is the residual when $\mathbf{y}$ is explained by the variables represented by $\mathbf{X}_1$,

- the explanatory variables

$$\mathbf{M}_1\mathbf{X}_2 = \operatorname{res}(\mathbf{X}_2; \mathbf{X}_1) \tag{8.67}$$

are the residuals of the variables represented by $\mathbf{X}_2$ "after elimination of $\mathbf{X}_1$".

The covariance matrix in model $\{\mathbf{M}_1\mathbf{y},\ \mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2,\ \sigma^2\mathbf{M}_1\}$ is singular and hence there is a good reason to worry whether the $\operatorname{OLSE}(\mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2)$ equals $\operatorname{BLUE}(\mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2)$ under that model. The answer is positive because clearly the column space inclusion

$$\mathscr{C}(\mathbf{M}_1 \cdot \mathbf{M}_1\mathbf{X}_2) \subset \mathscr{C}(\mathbf{M}_1\mathbf{X}_2) \tag{8.68}$$

holds; (8.68) corresponds to condition (vi) in Theorem 10.1 (p. 218).

When $\mathbf{X}_2 = \mathbf{x}_k$, the reduced model is

$$\mathscr{M}_{12\cdot 1} = \{\mathbf{M}_1\mathbf{y},\ \mathbf{M}_1\mathbf{x}_k\beta_k,\ \sigma^2\mathbf{M}_1\}\,, \tag{8.69}$$

and the coefficient of determination in this reduced model is the squared correlation between $\mathbf{M}_1\mathbf{y}$ and $\mathbf{M}_1\mathbf{x}_k$, which is precisely the squared partial correlation between $\mathbf{y}$ and $\mathbf{x}_k$ after elimination of all other $\mathbf{x}$-variables:

$$\begin{aligned} R^2(\mathscr{M}_{12\cdot 1}) &= \operatorname{cor_d}^2(\mathbf{M}_1\mathbf{y},\ \mathbf{M}_1\mathbf{x}_k) \\ &= \operatorname{pcor_d}^2(\mathbf{y},\ \mathbf{x}_k \mid \mathbf{X}_1) = r^2_{yk\cdot 12\ldots k-1}\,. \end{aligned} \tag{8.70}$$

The plots of residuals $\mathbf{M}_1\mathbf{y}$ and $\mathbf{M}_1\mathbf{x}_k$ are called added variable plots; see, e.g., Weisberg (2005, p. 49).

Assuming that $\mathbf{1} \in \mathscr{C}(\mathbf{X}_1)$, the columns $\mathbf{M}_1\mathbf{y}$ and $\mathbf{M}_1\mathbf{x}_k$ are centered and $R^2(\mathscr{M}_{12\cdot 1})$ can be expressed as

$$\begin{aligned} R^2(\mathscr{M}_{12\cdot 1}) = \operatorname{cor_d}^2(\mathbf{M}_1\mathbf{y},\ \mathbf{M}_1\mathbf{x}_k) &= \frac{(\mathbf{y}'\mathbf{M}_1\mathbf{x}_k)^2}{\mathbf{y}'\mathbf{M}_1\mathbf{y} \cdot \mathbf{x}_k'\mathbf{M}_1\mathbf{x}_k} \\ &= \frac{\mathbf{y}'\mathbf{M}_1\mathbf{x}_k\mathbf{x}_k'\mathbf{M}_1\mathbf{y}}{\mathbf{y}'\mathbf{M}_1\mathbf{y} \cdot \mathbf{x}_k'\mathbf{M}_1\mathbf{x}_k} = \frac{\mathbf{y}'\mathbf{P}_{\mathbf{M}_1\mathbf{x}_k}\mathbf{y}}{\mathbf{y}'\mathbf{M}_1\mathbf{y}}\,. \end{aligned} \tag{8.71}$$

It is also interesting to observe that the OLSE's residual under the reduced model $\mathscr{M}_{12\cdot 1}$ equals the OLSE's residual under the full model $\mathscr{M}_{12}$, in short,

$$\operatorname{res}(\mathscr{M}_{12\cdot 1}) = \operatorname{res}(\mathscr{M}_{12})\,, \quad \text{i.e.,} \quad \operatorname{res}(\mathbf{M}_1\mathbf{y}; \mathbf{M}_1\mathbf{X}_2) = \operatorname{res}(\mathbf{y}; \mathbf{X})\,. \tag{8.72}$$

This was already stated in Proposition 3.2 (p. 98) when the model matrix was simply $(\mathbf{1} : \mathbf{x})$. In the general case, the residual vector under $\mathscr{M}_{12\cdot 1}$ becomes, in view of (8.39a),

$$\begin{aligned}\operatorname{res}(\mathscr{M}_{12\cdot 1}) &= \mathbf{M}_1\mathbf{y} - \mathbf{P}_{\mathbf{M}_1\mathbf{X}_2}\mathbf{M}_1\mathbf{y} \\ &= \mathbf{y} - (\mathbf{P}_{\mathbf{X}_1}\mathbf{y} + \mathbf{P}_{\mathbf{M}_1\mathbf{X}_2}\mathbf{y}) \\ &= \mathbf{y} - \mathbf{P}_{(\mathbf{X}_1 : \mathbf{X}_1)}\mathbf{y} = \mathbf{M}\mathbf{y} = \operatorname{res}(\mathscr{M}_{12})\,. \end{aligned} \tag{8.73}$$

Notice that obviously

$$\mathbf{P}_{\mathbf{M}_1\mathbf{X}_2}\mathbf{M}_1\mathbf{y} = \mathbf{P}_{\mathbf{M}_1\mathbf{X}_2}\mathbf{y}\,. \tag{8.74}$$

Trivially (8.72) implies that the residual sum of squares are identical under $\mathscr{M}_{12\cdot 1}$ and $\mathscr{M}_{12}$:

$$\operatorname{SSE}(\mathscr{M}_{12\cdot 1}) = \operatorname{SSE}(\mathscr{M}_{12})\,. \tag{8.75}$$

The total sum of squares under $\mathscr{M}_{12\cdot 1}$ is

$$\operatorname{SST}(\mathscr{M}_{12\cdot 1}) = \mathbf{y}'\mathbf{M}_1\mathbf{y} = \operatorname{SSE}(\mathscr{M}_1)\,, \tag{8.76}$$

and

$$\begin{aligned}\operatorname{SSR}(\mathscr{M}_{12\cdot 1}) &= \operatorname{SST}(\mathscr{M}_{12\cdot 1}) - \operatorname{SSE}(\mathscr{M}_{12\cdot 1}) \\ &= \operatorname{SSE}(\mathscr{M}_1) - \operatorname{SSE}(\mathscr{M}_{12}) = \mathbf{y}'\mathbf{P}_{\mathbf{M}_1\mathbf{X}_2}\mathbf{y}\,. \end{aligned} \tag{8.77}$$

In the reduced model, the regression sum of squares $\operatorname{SSR}(\mathscr{M}_{12\cdot 1})$ is therefore the difference between SSE's of the full model $\mathscr{M}_{12}$ and the small model $\mathscr{M}_1$. Thus we obtain, corresponding to (8.71),

$$\begin{aligned}R^2(\mathscr{M}_{12\cdot 1}) &= \frac{\operatorname{SSR}(\mathscr{M}_{12\cdot 1})}{\operatorname{SST}(\mathscr{M}_{12\cdot 1})} = \frac{\mathbf{y}'\mathbf{P}_{\mathbf{M}_1\mathbf{X}_2}\mathbf{y}}{\mathbf{y}'\mathbf{M}_1\mathbf{y}} \\ &= \frac{\mathbf{y}'\mathbf{M}_1\mathbf{y} - \mathbf{y}'\mathbf{M}\mathbf{y}}{\mathbf{y}'\mathbf{M}_1\mathbf{y}} = 1 - \frac{\mathbf{y}'\mathbf{M}\mathbf{y}}{\mathbf{y}'\mathbf{M}_1\mathbf{y}}\,, \end{aligned} \tag{8.78}$$

and thereby $1 - R^2(\mathscr{M}_{12\cdot 1}) = \frac{\mathbf{y}'\mathbf{M}\mathbf{y}}{\mathbf{y}'\mathbf{M}_1\mathbf{y}}$. Because $1 - R^2(\mathscr{M}_{12}) = \frac{\mathbf{y}'\mathbf{M}\mathbf{y}}{\mathbf{y}'\mathbf{C}\mathbf{y}}$, we also have

$$1 - R^2(\mathscr{M}_{12}) = \frac{\mathbf{y}'\mathbf{M}_1\mathbf{y}}{\mathbf{y}'\mathbf{C}\mathbf{y}} \cdot \frac{\mathbf{y}'\mathbf{M}\mathbf{y}}{\mathbf{y}'\mathbf{M}_1\mathbf{y}} = [1 - R^2(\mathscr{M}_1)][1 - R^2(\mathscr{M}_{12\cdot 1})]\,. \tag{8.79}$$

If $\mathbf{X}_2 = \mathbf{x}_k$, then

$$1 - r^2_{yk\cdot 1\ldots k-1} = \frac{\mathbf{y}'\mathbf{M}\mathbf{y}}{\mathbf{y}'\mathbf{M}_1\mathbf{y}}\,, \tag{8.80}$$

and

$$\begin{aligned}1 - R^2(\mathscr{M}_{12}) &= [1 - R^2(\mathscr{M}_1)](1 - r^2_{yk\cdot 1\ldots k-1}) \\ &= (1 - r^2_{y1})(1 - r^2_{y2\cdot 1}) \cdots (1 - r^2_{yk\cdot 1\ldots k-1})\,. \end{aligned} \tag{8.81}$$

8.5 The Subvectors $\hat{\boldsymbol{\beta}}_0$ and $\hat{\boldsymbol{\beta}}_{\mathbf{x}}$

Let us consider the partitioned linear model

$$\mathbf{y} = \mathbf{1}\beta_0 + \mathbf{X}_0\boldsymbol{\beta}_{\mathbf{x}} + \boldsymbol{\varepsilon}\,, \tag{8.82}$$

or in other notation,

$$\mathscr{M}_{12} = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{V}\} = \{\mathbf{y},\, \mathbf{1}\beta_0 + \mathbf{X}_0\boldsymbol{\beta}_{\mathbf{x}},\, \sigma^2\mathbf{V}\}\,, \tag{8.83}$$

where $\mathbf{X}_0$ is an $n \times k$ matrix:

$$\mathbf{X} = (\mathbf{1} : \mathbf{x}_1 : \ldots : \mathbf{x}_k) = (\mathbf{1} : \mathbf{X}_0)\,; \quad p = k + 1 = \operatorname{rank}(\mathbf{X})\,. \tag{8.84}$$

Using Proposition 8.3 (p. 159) we get at once

$$\hat{\boldsymbol{\beta}}_{\mathbf{x}} = (\mathbf{X}_0'\mathbf{C}\mathbf{X}_0)^{-1}\mathbf{X}_0'\mathbf{C}\mathbf{y} = \mathbf{T}_{\mathbf{xx}}^{-1}\mathbf{t}_{\mathbf{x}y} = \mathbf{S}_{\mathbf{xx}}^{-1}\mathbf{s}_{\mathbf{x}y}\,, \tag{8.85}$$

where

$$\mathbf{T} = (\mathbf{X}_0 : \mathbf{y})'\mathbf{C}(\mathbf{X}_0 : \mathbf{y}) = \begin{pmatrix} \mathbf{X}_0'\mathbf{C}\mathbf{X}_0 & \mathbf{X}_0'\mathbf{C}\mathbf{y} \\ \mathbf{y}'\mathbf{C}\mathbf{X}_0 & \mathbf{y}'\mathbf{C}\mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{T}_{\mathbf{xx}} & \mathbf{t}_{\mathbf{x}y} \\ \mathbf{t}_{\mathbf{x}y}' & t_{yy} \end{pmatrix}, \tag{8.86}$$

and the sample covariance and correlation matrices of variables x_1, x_2, …, x_k, y are

$$\mathbf{S} = \tfrac{1}{n-1}\mathbf{T} = \begin{pmatrix} \mathbf{S}_{\mathbf{xx}} & \mathbf{s}_{\mathbf{x}y} \\ \mathbf{s}_{\mathbf{x}y}' & s_y^2 \end{pmatrix}, \quad \mathbf{R} = \begin{pmatrix} \mathbf{R}_{\mathbf{xx}} & \mathbf{r}_{\mathbf{x}y} \\ \mathbf{r}_{\mathbf{x}y}' & 1 \end{pmatrix}. \tag{8.87}$$

To solve for the intercept $\hat{\beta}_0$, we first write the equality

$$\mathbf{H}\mathbf{y} = \mathbf{1}\hat{\beta}_0 + \mathbf{X}_0\hat{\boldsymbol{\beta}}_{\mathbf{x}} = \mathbf{P}_{\mathbf{1}}\mathbf{y} + \mathbf{P}_{\mathbf{C}\mathbf{X}_0}\mathbf{y} = \mathbf{J}\mathbf{y} + \mathbf{C}\mathbf{X}_0\hat{\boldsymbol{\beta}}_{\mathbf{x}}\,, \tag{8.88}$$

i.e., $\mathbf{1}\hat{\beta}_0 = \mathbf{J}\mathbf{y} - \mathbf{J}\mathbf{X}_0\hat{\boldsymbol{\beta}}_{\mathbf{x}}$, from which it immediately follows that

$$\hat{\beta}_0 = \bar{y} - \bar{\mathbf{x}}'\hat{\boldsymbol{\beta}}_{\mathbf{x}} = \bar{y} - (\bar{x}_1\hat{\beta}_1 + \cdots + \bar{x}_k\hat{\beta}_k)\,. \tag{8.89}$$

If the columns of the matrix $(\mathbf{X}_0 : \mathbf{y})$ are centered and of equal length, then it is easy to confirm that

$$\hat{\beta}_0 = 0\,, \qquad \hat{\boldsymbol{\beta}}_{\mathbf{x}} = \mathbf{R}_{\mathbf{xx}}^{-1}\mathbf{r}_{\mathbf{x}y} := \hat{\boldsymbol{\alpha}}\,. \tag{8.90}$$

The vector $\hat{\boldsymbol{\alpha}}$ comprises the so-called standardized regression coefficients.

Note also that if $\mathbf{V} = \mathbf{I}$ (and $\operatorname{rank}(\mathbf{X}) = p$), then

$$\operatorname{cov}(\hat{\boldsymbol{\beta}}_{\mathbf{x}}) = \sigma^2(\mathbf{X}_0'\mathbf{C}\mathbf{X}_0)^{-1} = \sigma^2\mathbf{T}_{\mathbf{xx}}^{-1}\,. \tag{8.91}$$

In particular, if $\mathscr{M}_{12} = \{\mathbf{y},\, \mathbf{1}\beta_0 + \mathbf{x}\beta_1,\, \sigma^2\mathbf{I}\}$, then $\mathbf{M}_1 = \mathbf{I} - \mathbf{J} = \mathbf{C}$, and

$$\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1 , \quad \hat{\beta}_1 = (\mathbf{x}'\mathbf{C}\mathbf{x})^{-1}\mathbf{x}'\mathbf{C}\mathbf{y} = \frac{\mathbf{x}'\mathbf{C}\mathbf{y}}{\mathbf{x}'\mathbf{C}\mathbf{x}} , \quad \operatorname{var}(\hat{\beta}_1) = \frac{\sigma^2}{\mathbf{x}'\mathbf{C}\mathbf{x}} . \tag{8.92}$$

Consider the model $\{\mathbf{y},\, (\mathbf{1} : \mathbf{x}_1 : \mathbf{x}_2)\boldsymbol{\beta},\, \sigma^2\mathbf{I}\}$. Then the correlation between $\hat{\beta}_1$ and $\hat{\beta}_2$ is $-r_{12}$, the negative of the correlation between x_1 and x_2:

$$\operatorname{cor}(\hat{\beta}_1, \hat{\beta}_2) = -\operatorname{cor}_{\mathrm{d}}(\mathbf{x}_1, \mathbf{x}_2) = -r_{12} . \tag{8.93}$$

This comes from the following:

$$\begin{aligned}
\operatorname{cov}(\hat{\boldsymbol{\beta}}_{(2)}) &= \operatorname{cov}\begin{pmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{pmatrix} = \sigma^2(\mathbf{X}_0'\mathbf{C}\mathbf{X}_0)^{-1} = \sigma^2\mathbf{T}_{\mathbf{xx}}^{-1} = \sigma^2\begin{pmatrix}t_{11} & t_{12}\\ t_{12} & t_{22}\end{pmatrix}^{-1}\\
&= \frac{\sigma^2}{t_{11}t_{22} - t_{12}^2}\begin{pmatrix}t_{22} & -t_{12}\\ -t_{12} & t_{11}\end{pmatrix} \implies \operatorname{cor}(\hat{\beta}_1, \hat{\beta}_2) = \frac{-t_{12}}{\sqrt{t_{11}\cdot t_{22}}} = -r_{12} .
\end{aligned} \tag{8.94}$$

It is also very useful to observe that $\bar{y}$ and $\hat{\boldsymbol{\beta}}_{\mathbf{x}}$ are uncorrelated under $\{\mathbf{y},\, (\mathbf{1} : \mathbf{X}_0)\boldsymbol{\beta},\, \sigma^2\mathbf{I}\}$. This is seen from

$$\operatorname{cov}(\hat{\boldsymbol{\beta}}_{\mathbf{x}}, \bar{y}) = \operatorname{cov}(\mathbf{T}_{\mathbf{xx}}^{-1}\mathbf{X}_0'\mathbf{C}\mathbf{y},\, \tfrac{1}{n}\mathbf{1}'\mathbf{y}) = \mathbf{T}_{\mathbf{xx}}^{-1}\mathbf{X}_0'\mathbf{C}\cdot\sigma^2\mathbf{I}\cdot\tfrac{1}{n}\mathbf{1} = \mathbf{0} . \tag{8.95}$$

Consider the full rank model $\mathscr{M} = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}\}$ and suppose that our interest is to predict a new future value of random variable y_* when the x-variables have values $(1, x_1^*, \dots, x_k^*)' = \left(\begin{smallmatrix}1\\ \mathbf{x}_*\end{smallmatrix}\right) = \mathbf{x}_\#$. We assume that y_* follows the same model as the components of $\mathbf{y}$ in $\mathscr{M}$:

$$y_* = \mathbf{x}_\#'\boldsymbol{\beta} + \varepsilon_* = \beta_0 + \mathbf{x}_*'\boldsymbol{\beta}_{\mathbf{x}} + \varepsilon_* , \quad \operatorname{var}(\varepsilon_*) = \sigma^2. \tag{8.96}$$

In addition we assume that $\operatorname{cov}(\mathbf{y}, y_*) = \mathbf{0}$. A good linear predictor (actually the best, BLUP, see Section 10.9, page 245) for y_* on the basis of $\mathbf{y}$ is naturally

$$\begin{aligned}
\hat{y}_* &= \mathbf{x}_\#'\hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}_{\mathbf{x}}'\mathbf{x}_*\\
&= (\bar{y} - \hat{\boldsymbol{\beta}}_{\mathbf{x}}'\bar{\mathbf{x}}) + \hat{\boldsymbol{\beta}}_{\mathbf{x}}'\mathbf{x}_* = \bar{y} + \hat{\boldsymbol{\beta}}_{\mathbf{x}}'(\mathbf{x}_* - \bar{\mathbf{x}}) .
\end{aligned} \tag{8.97}$$

Then $\hat{y}_*$ is obviously an unbiased predictor for y_* in the sense that

$$\operatorname{E}(\hat{y}_* - y_*) = 0 , \text{ where } \hat{y}_* - y_* = e_* = \text{prediction error.} \tag{8.98}$$

The variance of $\hat{y}_*$ is

$$\operatorname{var}(\hat{y}_*) = \operatorname{var}(\mathbf{x}_\#'\hat{\boldsymbol{\beta}}) = \sigma^2\mathbf{x}_\#'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_\# , \tag{8.99}$$

which, using $\operatorname{cov}(\hat{\boldsymbol{\beta}}_{\mathbf{x}}, \bar{y}) = \mathbf{0}$, can be expressed as

$$\begin{aligned}
\operatorname{var}(\hat{y}_*) &= \operatorname{var}\big[\bar{y} + \hat{\boldsymbol{\beta}}_{\mathbf{x}}'(\mathbf{x}_* - \bar{\mathbf{x}})\big]\\
&= \operatorname{var}(\bar{y}) + \operatorname{var}\big[\hat{\boldsymbol{\beta}}_{\mathbf{x}}'(\mathbf{x}_* - \bar{\mathbf{x}})\big]
\end{aligned}$$

$$\begin{aligned} &= \sigma^2\big[\tfrac{1}{n} + (\mathbf{x}_* - \bar{\mathbf{x}})'\mathbf{T}_{\mathbf{xx}}^{-1}(\mathbf{x}_* - \bar{\mathbf{x}})\big] \\ &= \sigma^2\big[\tfrac{1}{n} + \tfrac{1}{n-1}(\mathbf{x}_* - \bar{\mathbf{x}})'\mathbf{S}_{\mathbf{xx}}^{-1}(\mathbf{x}_* - \bar{\mathbf{x}})\big] \\ &= \sigma^2\big[\tfrac{1}{n} + \tfrac{1}{n-1}\,\mathrm{MHLN}^2(\mathbf{x}_*,\, \bar{\mathbf{x}},\, \mathbf{S}_{\mathbf{xx}})\big]\,. \end{aligned} \tag{8.100}$$

In view of the assumption $\operatorname{cov}(\mathbf{y}, y_*) = \mathbf{0}$, which implies that y_* and $\hat{y}_*$ are uncorrelated, the variance of the prediction error becomes

$$\begin{aligned} \operatorname{var}(e_*) &= \operatorname{var}(y_* - \hat{y}_*) = \operatorname{var}(y_*) + \operatorname{var}(\hat{y}_*) \\ &= \sigma^2[1 + \mathbf{x}_\#'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_\#]\,. \end{aligned} \tag{8.101}$$

When $k = 1$ we get

$$\operatorname{var}(\hat{y}_*) = \sigma^2\left[\frac{1}{n} + \frac{(x_* - \bar{x})^2}{\mathrm{SS}_x}\right], \quad \operatorname{var}(e_*) = \sigma^2\left[1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{\mathrm{SS}_x}\right]. \tag{8.102}$$

The formulas for $\operatorname{var}(\hat{y}_*)$ and $\operatorname{var}(e_*)$, replacing σ^2 with $\hat{\sigma}^2$, give basis for the confidence interval for $\mathrm{E}(y_*)$ and the prediction interval for y_*; see, e.g., Seber & Lee (2003, Ch. 5).

8.6 The Decomposition SST = SSR + SSE

Let us consider the models

$$\mathscr{M}_{12} = \{\mathbf{y},\, \mathbf{1}\beta_0 + \mathbf{X}_0\boldsymbol{\beta}_{\mathbf{x}},\, \sigma^2\mathbf{I}\}\,, \quad \text{and} \quad \mathscr{M}_0 = \{\mathbf{y},\, \mathbf{1}\beta_0,\, \sigma^2\mathbf{I}\}\,. \tag{8.103}$$

The residual sum of squares under $\mathscr{M}_{12}$ is

$$\mathrm{SSE} = \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y} = \|(\mathbf{I} - \mathbf{H})\mathbf{y}\|^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{8.104}$$

Under the simple basic model $\mathscr{M}_0$, we have $\mathrm{OLSE}(\mathbf{1}\beta_0) = \mathbf{J}\mathbf{y} = \bar{y}\mathbf{1}$, and the residual vector is the centered $\mathbf{y}$, $\mathbf{C}\mathbf{y}$, and hence the residual sum of squares under $\mathscr{M}_0$ is

$$\mathrm{SSE}_0 = \mathbf{y}'(\mathbf{I} - \mathbf{J})\mathbf{y} = \mathbf{y}'\mathbf{C}\mathbf{y} = \|\mathbf{C}\mathbf{y}\|^2 = t_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \mathrm{SST}\,, \tag{8.105}$$

which is called the total sum of squares.

The change in SSE when moving from $\mathscr{M}_0$ to $\mathscr{M}_{12}$ is

$$\begin{aligned} \mathrm{SSE}_0 - \mathrm{SSE} &= \mathrm{SST} - \mathrm{SSE} \\ &= \mathbf{y}'(\mathbf{I} - \mathbf{J})\mathbf{y} - \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y} \\ &= \mathbf{y}'(\mathbf{H} - \mathbf{J})\mathbf{y} = \mathrm{SSR}\,, \end{aligned} \tag{8.106}$$

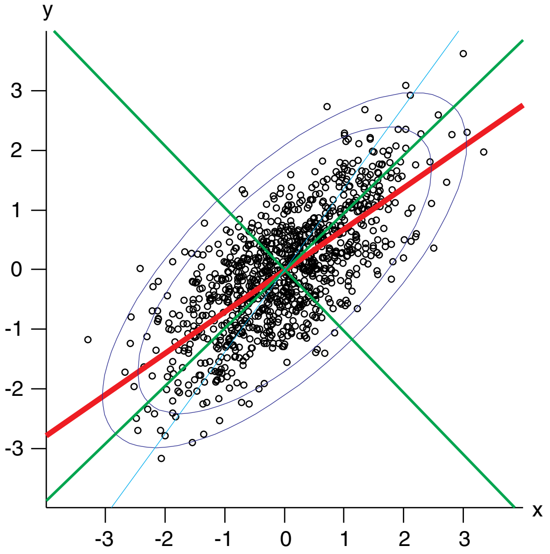


Figure 8.1 1000 observations from $\mathrm{N}_2(\mathbf{0}, \boldsymbol{\Sigma})$; $\sigma_x = \sigma_y = 1$, $\varrho_{xy} = 0.7$.

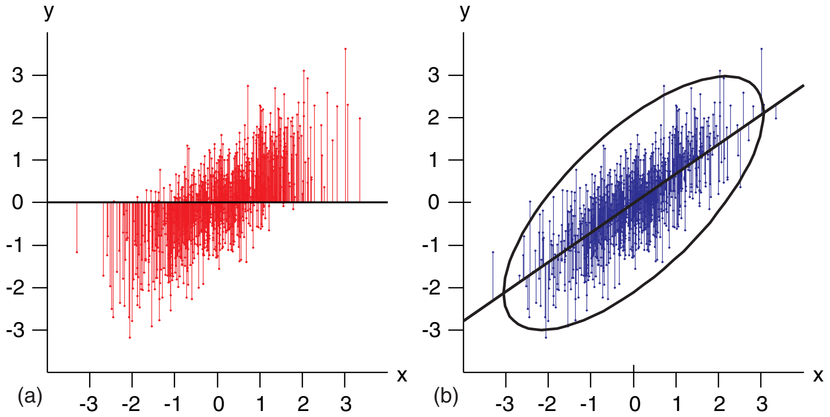


Figure 8.2 Fitting the linear model into the scatterplot of Figure 8.1. (a) Residuals from the model $\mathscr{M}_0 = \{\mathbf{y}, \mathbf{1}\beta_0, \sigma^2\mathbf{I}\}$, (b) residuals from $\mathscr{M}_{12} = \{\mathbf{y}, \mathbf{1}\beta_0 + \mathbf{x}\beta_1, \sigma^2\mathbf{I}\}$.

which is the "sum of squares due to regression"; see also (3.41) (p. 99). The *relative reduction* in SSE when using $\mathscr{M}_{12}$ instead of $\mathscr{M}_0$ is

$$R^2 = \frac{\mathrm{SSE}_0 - \mathrm{SSE}}{\mathrm{SSE}_0} = \frac{\mathrm{SST} - \mathrm{SSE}}{\mathrm{SST}} = \frac{\mathrm{SSR}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}\,, \tag{8.107}$$

which is the coefficient of determination, i.e., the multiple correlation coefficient squared.

Notice that when $\mathbf{X} = (\mathbf{1} : \mathbf{X}_0)$, we have, corresponding to (3.42) (p. 99),

$$\mathbf{H} = \mathbf{P}_{(\mathbf{1} : \mathbf{X}_0)} = \mathbf{J} + \mathbf{P}_{\mathbf{C}\mathbf{X}_0} = \mathbf{J} + \mathbf{P}_{\tilde{\mathbf{X}}_0} = \mathbf{J} + \mathbf{P}_{\mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{1})^\perp}, \tag{8.108}$$

and hence

$$\mathbf{H} - \mathbf{J} = \mathbf{P}_{\mathbf{C}\mathbf{X}_0} = \mathbf{P}_{\tilde{\mathbf{X}}_0} = \mathbf{P}_{\mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{1})^\perp} \,. \tag{8.109}$$

As an orthogonal projector, $\mathbf{H} - \mathbf{J}$ is idempotent (and symmetric) and so

$$\mathrm{SSR} = \mathbf{y}'(\mathbf{H} - \mathbf{J})\mathbf{y} = \|(\mathbf{H} - \mathbf{J})\mathbf{y}\|^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y}_i)^2. \tag{8.110}$$

In the above definition of the R^2 it was essential to have the column $\mathbf{1}$ in the model matrix $\mathbf{X}$. Let us us now relax from this assumption for a while and "redefine" the above sums of squares as

$$\mathrm{SST} = \sum_{i=1}^n (y_i - \bar{y})^2 = \|(\mathbf{I} - \mathbf{J})\mathbf{y}\|^2, \tag{8.111a}$$

$$\mathrm{SSR} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \|(\mathbf{H} - \mathbf{J})\mathbf{y}\|^2, \tag{8.111b}$$

$$\mathrm{SSE} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \|(\mathbf{I} - \mathbf{H})\mathbf{y}\|^2. \tag{8.111c}$$

Note that the only "new" thing concerns SSR. Of course we trivially have

$$(\mathbf{I} - \mathbf{J})\mathbf{y} = (\mathbf{H} - \mathbf{J})\mathbf{y} + (\mathbf{I} - \mathbf{H})\mathbf{y}\,, \tag{8.112}$$

and equally trivially we *always* have

$$\mathbf{y}'(\mathbf{I} - \mathbf{J})\mathbf{y} = \mathbf{y}'(\mathbf{H} - \mathbf{J})\mathbf{y} + \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}\,. \tag{8.113}$$

However, the equation

$$\|(\mathbf{I} - \mathbf{J})\mathbf{y}\|^2 = \|(\mathbf{H} - \mathbf{J})\mathbf{y}\|^2 + \|(\mathbf{I} - \mathbf{H})\mathbf{y}\|^2 \tag{8.114}$$

holds if and only if $(\mathbf{H} - \mathbf{J})\mathbf{y}$ and $(\mathbf{I} - \mathbf{H})\mathbf{y}$ are orthogonal, i.e.,

$$\mathbf{y}'(\mathbf{H} - \mathbf{J})(\mathbf{I} - \mathbf{H})\mathbf{y} = \mathbf{y}'(\mathbf{J}\mathbf{H} - \mathbf{J})\mathbf{y} = 0. \tag{8.115}$$

Now (8.115) holds (for all $\mathbf{y}$) if and only if $\mathbf{J} = \mathbf{J}\mathbf{H}$, and thereby $\mathbf{J} = \mathbf{J}\mathbf{H} = \mathbf{H}\mathbf{J}$. Proposition 7.1 (p. 152) immediately gives the following Proposition:

Proposition 8.5. *Consider the model* $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}\}$. *Then the decomposition* $\mathrm{SST} = \mathrm{SSR} + \mathrm{SSE}$, *i.e.,*

$$\|(\mathbf{I} - \mathbf{J})\mathbf{y}\|^2 = \|(\mathbf{H} - \mathbf{J})\mathbf{y}\|^2 + \|(\mathbf{I} - \mathbf{H})\mathbf{y}\|^2 \tag{8.116}$$

holds for all $\mathbf{y}$ *if and only if* $\mathbf{1} \in \mathscr{C}(\mathbf{X})$.

Note that for $\mathrm{SST} = \mathrm{SSR} + \mathrm{SSE}$ to hold, it is not necessary that the vector $\mathbf{1}$ is explicitly a column of $\mathbf{X}$; it is enough if $\mathbf{1} \in \mathscr{C}(\mathbf{X})$. The decomposition is also valid in particular if $\mathbf{y}$ is centered or $\mathbf{y} \in \mathscr{C}(\mathbf{X})$.

When $\mathbf{1} \notin \mathscr{C}(\mathbf{X})$, it makes no sense to use R^2 as a statistic to describe how well OLS fits the data. When $\mathbf{1} \notin \mathscr{C}(\mathbf{X})$, R^2 as defined as $1 - \mathrm{SSE} \,/\, \mathrm{SST}$, can be even negative! In the no-intercept model, it is natural (see, e.g., Searle 1982, p. 379) to consider the decomposition

$$\mathbf{y}'\mathbf{y} = \mathbf{y}'\mathbf{H}\mathbf{y} + \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}\,, \quad \mathrm{SST}_* = \mathrm{SSR}_* + \mathrm{SSE}\,. \tag{8.117}$$

Then the coefficient of determination can be defined as

$$R_*^2 = \frac{\mathrm{SSR}_*}{\mathrm{SST}_*} = \frac{\mathbf{y}'\mathbf{H}\mathbf{y}}{\mathbf{y}'\mathbf{y}} = 1 - \frac{\mathrm{SSE}}{\mathbf{y}'\mathbf{y}} = \cos^2(\mathbf{y}, \mathbf{H}\mathbf{y})\,. \tag{8.118}$$

See also (3.29) (p. 97).

8.7 Representations of R^2

The coefficient of determination has various equivalent representations, like for example:

$$\begin{aligned}
R^2 &= \frac{\mathbf{y}'(\mathbf{I} - \mathbf{J})\mathbf{y} - \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}}{\mathbf{y}'(\mathbf{I} - \mathbf{J})\mathbf{y}} && \text{(8.119a)}\\
&= \frac{\mathbf{y}'(\mathbf{H} - \mathbf{J})\mathbf{y}}{\mathbf{y}'(\mathbf{I} - \mathbf{J})\mathbf{y}} && \text{(8.119b)}\\
&= \frac{\mathbf{y}'\mathbf{P}_{\mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{1})^\perp}\mathbf{y}}{\mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{1}})\mathbf{y}} && \text{(8.119c)}\\
&= \frac{\mathbf{y}'\mathbf{P}_{\mathbf{C}\mathbf{X}_0}\mathbf{y}}{\mathbf{y}'\mathbf{C}\mathbf{y}} && \text{(8.119d)}\\
&= \frac{\mathbf{y}'\mathbf{C}\mathbf{X}_0(\mathbf{X}_0'\mathbf{C}\mathbf{X}_0)^{-1}\mathbf{X}_0'\mathbf{C}\mathbf{y}}{\mathbf{y}'\mathbf{C}\mathbf{y}} && \text{(8.119e)}\\
&= \frac{\mathbf{t}_{\mathbf{x}y}'\mathbf{T}_{\mathbf{xx}}^{-1}\mathbf{t}_{\mathbf{x}y}}{t_{yy}} = \frac{\mathbf{s}_{\mathbf{x}y}'\mathbf{S}_{\mathbf{xx}}^{-1}\mathbf{s}_{\mathbf{x}y}}{s_y^2} = \mathbf{r}_{\mathbf{x}y}'\mathbf{R}_{\mathbf{xx}}^{-1}\mathbf{r}_{\mathbf{x}y} && \text{(8.119f)}\\
&= \mathbf{r}_{\mathbf{x}y}'\hat{\boldsymbol{\alpha}} = r_{1y}\hat{\alpha}_1 + \cdots + r_{ky}\hat{\alpha}_k\,, && \text{(8.119g)}
\end{aligned}$$

where $\hat{\boldsymbol{\alpha}}$ is the vector of standardized regression coefficients. Above we have used the notation coming from the partitioned matrix of the corrected sums of squares and cross products of x's and y:

$$\begin{aligned}\mathbf{T} &= \begin{pmatrix}\mathbf{X}_0'\mathbf{C}\mathbf{X}_0 & \mathbf{X}_0'\mathbf{C}\mathbf{y}\\ \mathbf{y}'\mathbf{C}\mathbf{X}_0 & \mathbf{y}'\mathbf{C}\mathbf{y}\end{pmatrix} = \begin{pmatrix}\mathbf{T}_{\mathbf{xx}} & \mathbf{t}_{\mathbf{x}y}\\ \mathbf{t}_{\mathbf{x}y}' & t_{yy}\end{pmatrix}\\ &= \begin{pmatrix} t_{11} & t_{12} & \dots & t_{1k} & t_{1y}\\ \vdots & \vdots & & \vdots & \vdots\\ t_{k1} & t_{k2} & \dots & t_{kk} & t_{ky}\\ t_{y1} & t_{y2} & \dots & t_{yk} & t_{yy}\end{pmatrix}.\end{aligned} \tag{8.120}$$

The presentations in (8.119) can be easily proved using properties of orthogonal projectors. The true inverses can be replaced with generalized inverses whenever needed. For example, in view of

$$\mathbf{H} = \mathbf{P}_{(\mathbf{1}\,:\,\mathbf{X}_0)} = \mathbf{J} + \mathbf{P}_{\mathbf{C}\mathbf{X}_0}\,, \tag{8.121}$$

we get

$$\mathbf{I} - \mathbf{H} = \mathbf{C} - \mathbf{C}\mathbf{X}_0(\mathbf{X}_0'\mathbf{C}\mathbf{X}_0)^{-1}\mathbf{X}_0'\mathbf{C}\,, \tag{8.122}$$

and hence

$$\mathrm{SSE} = \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y} = t_{yy} - \mathbf{t}_{\mathbf{x}y}'\mathbf{T}_{\mathbf{xx}}^{-1}\mathbf{t}_{\mathbf{x}y}\,. \tag{8.123}$$

Here SSE is the Schur complement of $\mathbf{T}_{\mathbf{xx}}$ in $\mathbf{T}$:

$$\mathrm{SSE} = t_{yy\cdot\mathbf{x}} = \frac{1}{t^{yy}} = \mathbf{T}/\mathbf{T}_{\mathbf{xx}}\,, \tag{8.124}$$

where the notation t^{yy} refers to the last diagonal element of the matrix $\mathbf{T}^{-1}$. Moreover, it is easy to confirm the following:

$$\frac{1}{1 - R^2} = t_{yy}t^{yy} = r^{yy}, \tag{8.125}$$

where r^{yy} is the last diagonal element of the correlation matrix $\mathbf{R}^{-1}$. We may note that, corresponding to (8.125), the ith diagonal element of $\mathbf{R}_{\mathbf{xx}}^{-1}$ can be expressed as

$$r^{ii} = \frac{1}{1 - R^2_{i\cdot 1\dots i-1,i+1,\dots k}} = \mathrm{VIF}_i\,, \tag{8.126}$$

while

$$R^2_{i\cdot 1\dots i-1,i+1,\dots k} = 1 - \frac{1}{r^{ii}}\,. \tag{8.127}$$

For the use of $\mathbf{R}^{-1}$ in statistics, see, e.g., Raveh (1985).

8.8 *R* as a Maximal Correlation

One fundamental (and well-known) property of R defined above is that it equals the correlation coefficient between the $\mathbf{y}$ and $\mathbf{Hy}$:

Proposition 8.6. *Consider the model* $\mathscr{M} = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}\}$, *where* $\mathbf{1} \in \mathscr{C}(\mathbf{X})$ *and let* R^2 *be defined as in* (8.107). *Then*

$$R = \operatorname{cor}_{\mathrm{d}}(\mathbf{y}, \mathbf{Hy}) = \cos\bigl[(\mathbf{I} - \mathbf{J})\mathbf{y}, (\mathbf{I} - \mathbf{J})\mathbf{Hy}\bigr] = \cos(\mathbf{Cy}, \mathbf{CHy})\,. \tag{8.128}$$

Proof. The crucial property needed in (8.128) is the following:

$$\mathbf{J} = \mathbf{HJ} = \mathbf{JH}\,. \tag{8.129}$$

To confirm the important equality (8.129), we note that in view of

$$\mathbf{H1} = \mathbf{1} \iff \mathbf{1} \in \mathscr{C}(\mathbf{X})\,, \tag{8.130}$$

we indeed have

$$\mathbf{HJ} = \mathbf{H11}'\tfrac{1}{n} = \mathbf{11}'\tfrac{1}{n} = \mathbf{J}\,, \tag{8.131}$$

from which (8.129) follows.

To prove (8.128), note first that in view of (8.129), $(\mathbf{I} - \mathbf{J})\mathbf{H} = \mathbf{H} - \mathbf{J}$. Hence our claim is

$$R = \operatorname{cor}_{\mathrm{d}}(\mathbf{y}, \mathbf{Hy}) = \cos[\mathbf{Cy}, (\mathbf{H} - \mathbf{J})\mathbf{y}] := \frac{a}{\sqrt{b \cdot c}}\,, \tag{8.132}$$

which indeed is true because

$$a = \mathbf{y}'(\mathbf{I} - \mathbf{J})\mathbf{Hy} = \mathbf{y}'(\mathbf{H} - \mathbf{J})\mathbf{y} = \mathrm{SSR}\,, \tag{8.133a}$$

$$b = \mathbf{y}'(\mathbf{I} - \mathbf{J})\mathbf{y} = \mathrm{SST}\,, \tag{8.133b}$$

$$c = \mathbf{y}'\mathbf{H}(\mathbf{I} - \mathbf{J})\mathbf{Hy} = \mathrm{SSR}\ . \tag{8.133c}$$

□

We may recall that if the vector $\mathbf{y}$ is projected onto the line $\mathscr{C}(\mathbf{1})$, we get the mean vector

$$\mathbf{Jy} = \bar{y}\mathbf{1}\,. \tag{8.134}$$

But now, in view of (8.129), also

$$\mathbf{J}\hat{\mathbf{y}} = \mathbf{J}(\mathbf{Hy}) = \mathbf{Jy}\,. \tag{8.135}$$

As a result we obtain a right-angled triangle whose corners are at the tops of the vectors $\mathbf{y}$, $\mathbf{Hy}$ $(= \hat{\mathbf{y}})$, and $\mathbf{Jy}$. Note that an identical triangle is formed by the tops of the vectors $\mathbf{Cy}$, $\mathbf{CHy}$, and $\mathbf{0}$; cf. Figure 8.4 (p. 183). The length (squared) of the hypotenuse of this triangle is SST, while the other sides have (squared) lengths SSE and SSR, and hence

$$\mathrm{SST} = \mathrm{SSR} + \mathrm{SSE}\,, \tag{8.136a}$$

$$\|\mathbf{y} - \mathbf{Jy}\|^2 = \|\mathbf{Jy} - \mathbf{Hy}\|^2 + \|\mathbf{y} - \mathbf{Hy}\|^2. \tag{8.136b}$$

It is also worth emphasizing that R^2 has the following property:

Proposition 8.7. *With the above notation,*

$$R^2 = \max_{\mathbf{b}} \operatorname{cor_d}^2(\mathbf{y}, \mathbf{X}_0\mathbf{b}) = \cos^2(\mathbf{Cy}, \mathbf{CX}_0\mathbf{b}_*)\,, \tag{8.137}$$

where

$$\mathbf{b}_* = (\mathbf{X}_0'\mathbf{C}\mathbf{X}_0)^{-1}\mathbf{X}_0'\mathbf{Cy} = \mathbf{T}_{\mathbf{xx}}^{-1}\mathbf{t}_{\mathbf{x}y} = \mathbf{S}_{\mathbf{xx}}^{-1}\mathbf{s}_{\mathbf{x}y} = \hat{\boldsymbol{\beta}}_{\mathbf{x}}\,. \tag{8.138}$$

Proof. The proof is essentially available in Remark 2.2 (p. 79), in (2.2) (p. 79), and in (9.19) (p. 193). Despite of that, let's have a quick look at the problem in hand.

For a simplicity, let us assume that the model matrix $\mathbf{X}$ has full column rank; if this is not the case then all inverses appearing in this context should be replaced by generalized inverses. Note also that in (8.137) we can replace $\mathbf{b}_*$ with $a\mathbf{b}_*$, for any nonzero scalar a.

It is obvious, see Proposition 2.3 (p. 77), that the cosine between a given vector $\mathbf{Cy}$ and a vector $\mathbf{CX}_0\mathbf{b}$ (with varying $\mathbf{b}$) is maximal when $\mathbf{CX}_0\mathbf{b}$ is the orthogonal projection of $\mathbf{Cy}$ onto $\mathbf{CX}_0$, i.e.,

$$\mathbf{CX}_0\mathbf{b}_* = \mathbf{P}_{\mathbf{CX}_0}\mathbf{Cy} = \mathbf{CX}_0\mathbf{T}_{\mathbf{xx}}^{-1}\mathbf{t}_{\mathbf{x}y}\,. \tag{8.139}$$

Hence we have

$$\begin{aligned} \max_{\mathbf{b}} \cos^2(\mathbf{Cy}, \mathbf{CX}_0\mathbf{b}) &= \cos^2(\mathbf{Cy}, \mathbf{P}_{\mathbf{CX}_0}\mathbf{Cy}) \\ &= \frac{\mathbf{y}'\mathbf{C}\mathbf{P}_{\mathbf{CX}_0}\mathbf{Cy}}{\mathbf{y}'\mathbf{Cy}} = \frac{\text{SSR}}{\text{SST}}\,. \end{aligned} \tag{8.140}$$

□

An alternative way to prove that R^2 has the maximal property (8.137), is to consider

$$\begin{aligned} \operatorname{cor_d}^2(\mathbf{y}, \mathbf{X}_0\mathbf{b}) &= \frac{(\mathbf{y}'\mathbf{CX}_0\mathbf{b})^2}{\mathbf{y}'\mathbf{Cy}\cdot\mathbf{b}'\mathbf{X}_0'\mathbf{CX}_0\mathbf{b}} = \frac{(\mathbf{t}_{\mathbf{x}y}'\mathbf{b})^2}{t_{yy}\cdot\mathbf{b}'\mathbf{T}_{\mathbf{xx}}\mathbf{b}} \\ &\le \frac{\mathbf{t}_2'\mathbf{T}_{\mathbf{xx}}^{-1}\mathbf{t}_{\mathbf{x}y}\cdot\mathbf{b}'\mathbf{T}_{\mathbf{xx}}\mathbf{b}}{t_{yy}\cdot\mathbf{b}'\mathbf{T}_{\mathbf{xx}}\mathbf{b}} = \frac{\mathbf{t}_{\mathbf{x}y}'\mathbf{T}_{\mathbf{xx}}^{-1}\mathbf{t}_{\mathbf{x}y}}{t_{yy}} = \frac{\text{SSR}}{\text{SST}}\,, \end{aligned} \tag{8.141}$$

where the inequality is based on the Cauchy–Schwarz inequality; see also Section 20.2 (p. 417).

Proposition 8.7 means that if we want to find such a linear combination of variables x_1, ..., x_k (whose observed values are in matrix $\mathbf{X}_0$) which has maximal correlation with the variable y (values in vector $\mathbf{y}$), then that linear combination is

$$\mathbf{X}_0\hat{\boldsymbol{\beta}}_{\mathbf{x}} = \hat{\beta}_1\mathbf{x}_1 + \cdots + \hat{\beta}_k\mathbf{x}_k\,, \tag{8.142}$$

where $\hat{\boldsymbol{\beta}}_{\mathbf{x}} = \mathbf{S}_{\mathbf{xx}}^{-1}\mathbf{s}_{\mathbf{x}y}$. Note also, see Remark 2.2 (p. 79), that if our task would be to find a vector $\mathbf{c}_*$ so that

$$\min_{\mathbf{c}} \operatorname{var}_{\mathrm{d}}(\mathbf{y} - \mathbf{X}_0\mathbf{c}) = \min_{\mathbf{c}} \tfrac{1}{n-1}\|\mathbf{C}\mathbf{y} - \mathbf{C}\mathbf{X}_0\mathbf{c}\|^2 = \operatorname{var}_{\mathrm{d}}(\mathbf{y} - \mathbf{X}_0\mathbf{c}_*)\,, \quad (8.143)$$

then the solution would again be $\hat{\boldsymbol{\beta}}_{\mathbf{x}}$:

$$\min_{\mathbf{c}} \operatorname{var}_{\mathrm{d}}(\mathbf{y} - \mathbf{X}_0\mathbf{c}) = \operatorname{var}_{\mathrm{d}}(\mathbf{y} - \mathbf{X}_0\hat{\boldsymbol{\beta}}_{\mathbf{x}})\,, \quad (8.144)$$

and

$$\begin{aligned} \operatorname{var}_{\mathrm{d}}(\mathbf{y} - \mathbf{X}_0\hat{\boldsymbol{\beta}}_{\mathbf{x}}) &= \tfrac{1}{n-1}\,\mathrm{SSE} = s_y^2 - \mathbf{s}'_{\mathbf{x}y}\mathbf{S}_{\mathbf{xx}}^{-1}\mathbf{s}_{\mathbf{x}y} \\ &= s_y^2\Big(1 - \frac{\mathbf{s}'_{\mathbf{x}y}\mathbf{S}_{\mathbf{xx}}^{-1}\mathbf{s}_{\mathbf{x}y}}{s_y^2}\Big) = s_y^2(1 - R^2)\,. \quad (8.145) \end{aligned}$$

8.9 Testing a Linear Hypothesis

In this section we simply present—without detailed explanation—some important formulas related to hypothesis testing using orthogonal projectors. We consider the full rank model $\mathscr{M}_{12} = \{\mathbf{y},\, \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2,\, \sigma^2\mathbf{I}\}$. The results below follow from the distributional properties mentioned on page 18.

First we note that in view of the decomposition (8.2) (p. 155), we have

$$\mathbf{H} - \mathbf{P}_{\mathbf{X}_1} = \mathbf{P}_{\mathbf{M}_1\mathbf{X}_2} = \mathbf{P}_{\mathscr{C}(\mathbf{X})\cap\mathscr{C}(\mathbf{M}_1)} = \mathbf{M}_1\mathbf{X}_2(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^-\mathbf{X}_2'\mathbf{M}_1\,, \quad (8.146)$$

where $\mathbf{X}_i \in \mathbb{R}^{n\times p_i}, i = 1, 2$. Then the following decompositions hold:

$$\begin{aligned} \mathrm{SSE} &= \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y} = \mathbf{y}'\mathbf{M}\mathbf{y} \\ &= \mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}_1} - \mathbf{P}_{\mathbf{M}_1\mathbf{X}_2})\mathbf{y} \\ &= \mathbf{y}'\mathbf{M}_1\mathbf{y} - \mathbf{y}'\mathbf{P}_{\mathbf{M}_1\mathbf{X}_2}\mathbf{y}\,, && (8.147a) \\ \mathrm{SSE}_H &= \mathrm{SSE}(\text{under the hypothesis } \boldsymbol{\beta}_2 = \mathbf{0}) \\ &= \mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}_1})\mathbf{y} = \mathbf{y}'\mathbf{M}_1\mathbf{y}\,, && (8.147b) \\ \mathrm{SSE}_H - \mathrm{SSE} &= \Delta\,\mathrm{SSE} \\ &= \text{change in SSE due to the hypothesis} \\ &= \mathbf{y}'(\mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{X}_1})\mathbf{y} = \mathbf{y}'\mathbf{P}_{\mathbf{M}_1\mathbf{X}_2}\mathbf{y} \\ &= \mathbf{y}'\mathbf{M}_1\mathbf{X}_2(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{M}_1\mathbf{y} \\ &= \hat{\boldsymbol{\beta}}_2'\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\hat{\boldsymbol{\beta}}_2 \\ &= \hat{\boldsymbol{\beta}}_2'\mathbf{T}_{22\cdot 1}\hat{\boldsymbol{\beta}}_2 \\ &= \hat{\boldsymbol{\beta}}_2'[\operatorname{cov}(\hat{\boldsymbol{\beta}}_2)]^{-1}\hat{\boldsymbol{\beta}}_2\sigma^2, && (8.147c) \end{aligned}$$

where

$$\begin{aligned}\mathbf{T}_{22\cdot 1} &= \mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2 = \mathbf{X}_2'(\mathbf{I} - \mathbf{P}_{\mathbf{X}_1})\mathbf{X}_2 \\ &= \text{the inverse of the lower-right } p_2 \times p_2 \text{ block of } (\mathbf{X}'\mathbf{X})^{-1} \\ &= \text{the Schur complement of } \mathbf{X}_1'\mathbf{X}_1 \text{ in } \mathbf{X}'\mathbf{X}\,. \qquad (8.147\text{d})\end{aligned}$$

If the hypothesis $\boldsymbol{\beta}_2 = \mathbf{0}$ is true and $\mathbf{y} \sim \mathrm{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, then

$$\begin{aligned}F &= \frac{(\mathrm{SSE}_H - \mathrm{SSE})/p_2}{\mathrm{SSE}\,/(n-p)} = \frac{\Delta\,\mathrm{SSE}\,/p_2}{\mathrm{SSE}\,/(n-p)} \\ &= \frac{\hat{\boldsymbol{\beta}}_2'\mathbf{T}_{22\cdot 1}\hat{\boldsymbol{\beta}}_2/p_2}{\hat{\sigma}^2} = \frac{\hat{\boldsymbol{\beta}}_2'[\mathrm{cov}(\hat{\boldsymbol{\beta}}_2)]^{-1}\hat{\boldsymbol{\beta}}_2\sigma^2/p_2}{\hat{\sigma}^2} \\ &= \hat{\boldsymbol{\beta}}_2'[\widehat{\mathrm{cov}}(\hat{\boldsymbol{\beta}}_2)]^{-1}\hat{\boldsymbol{\beta}}_2/p_2 \sim \mathrm{F}(p_2, n-p)\,, \qquad (8.148\text{a})\end{aligned}$$

where $\mathrm{F}(\cdot\,,\cdot)$ refers to the F-distribution and $\widehat{\mathrm{cov}}(\hat{\boldsymbol{\beta}}_2)$ to the estimated covariance matrix:

$$\widehat{\mathrm{cov}}(\hat{\boldsymbol{\beta}}_2) = \hat{\sigma}^2\mathbf{T}_{22\cdot 1}^{-1}\,. \qquad (8.148\text{b})$$

As a consequence, we can define the confidence ellipsoid for $\boldsymbol{\beta}_2$ as follows:

$$\frac{(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2)'\mathbf{T}_{22\cdot 1}(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2)/p_2}{\hat{\sigma}^2} \le F_{\alpha;p_2,n-p}\,, \qquad (8.149\text{a})$$

$$(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2)'[\widehat{\mathrm{cov}}(\hat{\boldsymbol{\beta}}_2)]^{-1}(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2)/p_2 \le F_{\alpha;p_2,n-p}\,, \qquad (8.149\text{b})$$

where $F_{\alpha;p_2,n-p}$ is the (upper) significance point corresponding to the α significance level.

If the hypothesis $\boldsymbol{\beta}_{\mathbf{x}} = \mathbf{0}$ is true, where $\boldsymbol{\beta} = \binom{\beta_0}{\boldsymbol{\beta}_{\mathbf{x}}}$ and $\mathbf{y} \sim \mathrm{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, then $\mathrm{SSE}_H = \mathrm{SST}$ and

$$\begin{aligned}F &= \frac{(\mathrm{SSE}_H - \mathrm{SSE})/k}{\mathrm{SSE}\,/(n-k-1)} = \frac{\mathrm{SSR}\,/k}{\mathrm{SSE}\,/(n-k-1)} \\ &= \frac{R^2/k}{(1-R^2)/(n-k-1)} \sim \mathrm{F}(k, n-k-1)\,. \qquad (8.150)\end{aligned}$$

Let us then have quick look at a more general hypothesis of the type

$$H\colon \mathbf{K}'\boldsymbol{\beta} = \mathbf{0}\,, \qquad (8.151)$$

where $\mathbf{K} \in \mathbb{R}^{p\times q}$ has full column rank and $\mathbf{K}'\boldsymbol{\beta}$ is estimable, i.e., $\mathbf{K} = \mathbf{X}'\mathbf{A}$ for some matrix $\mathbf{A}$. Denote

$$\begin{aligned}\mathrm{SSE}_H &= \mathrm{SSE}(\text{under the hypothesis } \mathbf{K}'\boldsymbol{\beta} = \mathbf{0}) \\ &= \min_H \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_r\|^2, \qquad (8.152)\end{aligned}$$

and let $\mathbf{L}$ be a matrix with the property $\mathscr{C}(\mathbf{L}) = \mathscr{N}(\mathbf{K}') = \mathscr{C}(\mathbf{K})^{\perp}$. Let further $\mathbf{X}_*$ be a matrix satisfying

$$\mathscr{C}(\mathbf{X}_*) = \{\mathbf{X}\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathscr{C}(\mathbf{L}) = \mathscr{N}(\mathbf{K}')\} = \mathscr{C}(\mathbf{XL}), \tag{8.153}$$

i.e.,

$$\mathscr{C}(\mathbf{X}_*) \text{ is the subspace of } \mathscr{C}(\mathbf{X}) \text{ where the hypothesis holds.} \tag{8.154}$$

Now we obviously have

$$\min_H \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{XL}\boldsymbol{\beta}\|^2 = \|\mathbf{y} - \mathbf{P}_{\mathbf{XL}}\mathbf{y}\|^2, \tag{8.155}$$

so that $\mathbf{X}\hat{\boldsymbol{\beta}}_r = \mathbf{P}_{\mathbf{XL}}\mathbf{y}$ and in particular if $\mathbf{L}$ and $\mathbf{X}$ have full column ranks,

$$\hat{\boldsymbol{\beta}}_r = \mathbf{L}(\mathbf{L}'\mathbf{X}'\mathbf{XL})^{-1}\mathbf{L}'\mathbf{X}'\mathbf{y}. \tag{8.156}$$

Suppose that $\mathbf{X}$ has full column rank and denote $\mathbf{N} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}$. Then $\mathscr{C}(\mathbf{X}_*)$ can be written as

$$\mathscr{C}(\mathbf{X}_*) = \{\mathbf{z} : \mathbf{z} \in \mathscr{C}(\mathbf{X}) \;\text{ and }\; \mathbf{N}'\mathbf{z} = \mathbf{0}\} = \mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{N})^{\perp}. \tag{8.157}$$

In view of (7.1) (p. 152), we have

$$\mathbf{P}_{\mathbf{X}} = \mathbf{P}_{\mathbf{N}} + \mathbf{P}_{\mathscr{C}(\mathbf{X})\cap\mathscr{C}(\mathbf{N})^{\perp}} = \mathbf{P}_{\mathbf{N}} + \mathbf{P}_{\mathbf{X}_*}, \tag{8.158a}$$

$$\begin{aligned}\mathbf{P}_{\mathbf{X}_*} &= \mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{N}} \\ &= \mathbf{H} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}]^{-1}\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}',\end{aligned} \tag{8.158b}$$

$$\hat{\boldsymbol{\beta}}_r = \hat{\boldsymbol{\beta}} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}]^{-1}\mathbf{K}'\hat{\boldsymbol{\beta}}, \tag{8.158c}$$

$$\mathrm{SSE}_H = \mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}_*})\mathbf{y} = \mathrm{SSE} + \hat{\boldsymbol{\beta}}'\mathbf{K}[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}]^{-1}\mathbf{K}'\hat{\boldsymbol{\beta}}. \tag{8.158d}$$

The estimator $\hat{\boldsymbol{\beta}}_r$ is the restricted OLSE of $\boldsymbol{\beta}$.

Under the model $\{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{V}\}$, where $\mathbf{V}$ is positive definite and $\mathbf{y} \sim \mathrm{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{V})$, the F-statistic for testing the hypothesis H: $\mathbf{K}'\boldsymbol{\beta} = \mathbf{d}$ appears to be

$$F(\mathbf{V}) = \frac{Q(\mathbf{V})/q}{\tilde{\sigma}^2} = \frac{[\mathrm{SSE}_H(\mathbf{V}) - \mathrm{SSE}(\mathbf{V})]/q}{\mathrm{SSE}(\mathbf{V})/(n-r)} \sim \mathrm{F}(q, n-r, \delta), \tag{8.159}$$

where $\mathbf{K}'\boldsymbol{\beta}$ is estimable, $\mathrm{rk}(\mathbf{K}_{p\times q}) = q$, $\tilde{\sigma}^2 = \mathrm{SSE}(\mathbf{V})/(n-r)$, $r = \mathrm{rk}(\mathbf{X})$, and

$$\mathrm{SSE}(\mathbf{V}) = \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2_{\mathbf{V}^{-1}} = (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}), \tag{8.160a}$$

$$\mathrm{SSE}_H(\mathbf{V}) = \min_H \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2_{\mathbf{V}^{-1}}, \tag{8.160b}$$

$$\begin{aligned}Q(\mathbf{V}) &= \mathrm{SSE}_H(\mathbf{V}) - \mathrm{SSE}(\mathbf{V}) = \Delta\,\mathrm{SSE}(\mathbf{V}) \\ &= \mathbf{y}'(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_*;\mathbf{V}^{-1}})\mathbf{y} - \mathbf{y}'(\mathbf{I}_n - \mathbf{P}_{\mathbf{X};\mathbf{V}^{-1}})\mathbf{y} \\ &= \mathbf{y}'(\mathbf{P}_{\mathbf{X};\mathbf{V}^{-1}} - \mathbf{P}_{\mathbf{X}_*;\mathbf{V}^{-1}})\mathbf{y}.\end{aligned} \tag{8.160c}$$

8.10 A Property of the Partial Correlation

Consider the last two regressors under the full rank model $\{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}\}$:

$$\mathbf{X} = (\mathbf{1} : \mathbf{x}_1 : \ldots : \mathbf{x}_{p-1} : \mathbf{x}_p) = (\mathbf{X}_1 : \mathbf{x}_{p-1} : \mathbf{x}_p) = (\mathbf{X}_1 : \mathbf{X}_2)\,. \tag{8.161}$$

Then

$$\begin{aligned}\operatorname{cov}(\hat{\boldsymbol{\beta}}_2) = \sigma^2(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1} &= \sigma^2 \begin{pmatrix} \mathbf{x}_{p-1}'\mathbf{M}_1\mathbf{x}_{p-1} & \mathbf{x}_{p-1}'\mathbf{M}_1\mathbf{x}_p \\ \mathbf{x}_p'\mathbf{M}_1\mathbf{x}_{p-1} & \mathbf{x}_p'\mathbf{M}_1\mathbf{x}_p \end{pmatrix}^{-1} \\ &= c \cdot \begin{pmatrix} \mathbf{x}_p'\mathbf{M}_1\mathbf{x}_p & -\mathbf{x}_{p-1}'\mathbf{M}_1\mathbf{x}_p \\ -\mathbf{x}_p'\mathbf{M}_1\mathbf{x}_{p-1} & \mathbf{x}_{p-1}'\mathbf{M}_1\mathbf{x}_{p-1} \end{pmatrix}, \end{aligned} \tag{8.162}$$

where c is a positive constant. Now the partial correlation between $\mathbf{x}_{p-1}$ and $\mathbf{x}_p$ when variables in $\mathbf{X}_1$ are held constant, is defined as the correlation between the residuals

$$\mathbf{e}_{p-1\cdot 1} = (\mathbf{I} - \mathbf{P}_{\mathbf{X}_1})\mathbf{x}_{p-1} = \mathbf{M}_1\mathbf{x}_{p-1}\,, \tag{8.163a}$$

$$\mathbf{e}_{p\cdot 1} = (\mathbf{I} - \mathbf{P}_{\mathbf{X}_1})\mathbf{x}_p = \mathbf{M}_1\mathbf{x}_p\,. \tag{8.163b}$$

Because these residuals are centered, their correlation is

$$\operatorname{cor}_{\mathrm{d}}(\mathbf{e}_{p-1\cdot 1}, \mathbf{e}_{p\cdot 1}) = \frac{\mathbf{x}_{p-1}'\mathbf{M}_1\mathbf{x}_p}{\sqrt{\mathbf{x}_{p-1}'\mathbf{M}_1\mathbf{x}_{p-1}}\,\sqrt{\mathbf{x}_p'\mathbf{M}_1\mathbf{x}_p}}\,. \tag{8.164}$$

Hence we immediately conclude the following result:

Proposition 8.8. *With the above notation,*

$$\operatorname{cor}(\hat{\beta}_{p-1}, \hat{\beta}_p) = -\operatorname{pcor}_{\mathrm{d}}(\mathbf{x}_{p-1}, \mathbf{x}_p \mid \mathbf{X}_1) = -r_{p-1,p\cdot 1\ldots p-2}\,. \tag{8.165}$$

Because the confidence ellipsoid for $\boldsymbol{\beta}_2$ can be defined by

$$\frac{(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2)'\mathbf{T}_{22\cdot 1}(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2)/p_2}{\hat{\sigma}^2} \le F_{\alpha;p_2,n-p}\,, \tag{8.166a}$$

$$(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2)'[\widehat{\operatorname{cov}}(\hat{\boldsymbol{\beta}}_2)]^{-1}(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2)/p_2 \le F_{\alpha;p_2,n-p}\,, \tag{8.166b}$$

we can conclude that the orientation of the confidence ellipse for

$$\boldsymbol{\beta}_2 = \begin{pmatrix} \beta_{p-1} \\ \beta_p \end{pmatrix} \tag{8.167}$$

is determined by (the negative of) the partial correlation between $\mathbf{x}_{p-1}$ and $\mathbf{x}_p$.

Regarding the result (8.165), we may refer to Belsley (1991, p. 33) who writes “This result, so well known in back-parlor gossip, seems strangely absent from statistics texts, so perhaps a formal demonstration is in order." The result (8.165) appears e.g. in Rao (1973a, p. 270) (with a missing minus sign), and in Kanto & Puntanen (1983, 1985).

8.11 Another Property of the Partial Correlation

There is one interesting (curious) feature of the partial correlation coefficient that may deserve attention; see Kanto & Puntanen (1983, 1985). To simplify notation, let the model matrix be partitioned as

$$\mathbf{X} = (\mathbf{x}_1 : \mathbf{x}_2 : \mathbf{1} : \mathbf{x}_3 : \ldots : \mathbf{x}_k) = (\mathbf{x}_1 : \mathbf{x}_2 : \mathbf{X}_3)\,, \tag{8.168}$$

and denote

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\boldsymbol{\beta}}_3 \end{pmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\,. \tag{8.169}$$

Proposition 8.9. *Consider two vectors of specific residuals:*

$$\mathbf{e}_{1\cdot 23} = (\mathbf{I} - \mathbf{P}_{(\mathbf{x}_2 \,:\, \mathbf{X}_3)})\mathbf{x}_1 = \mathbf{M}_{23}\mathbf{x}_1\,, \tag{8.170a}$$

$$\mathbf{e}_{2\cdot 13} = (\mathbf{I} - \mathbf{P}_{(\mathbf{x}_1 \,:\, \mathbf{X}_3)})\mathbf{x}_2 = \mathbf{M}_{13}\mathbf{x}_2\,, \tag{8.170b}$$

i.e., $\mathbf{e}_{1\cdot 23}$ *is the residual when* $\mathbf{x}_1$ *is explained by all other* $\mathbf{x}$*-variables. Then the following result holds:*

$$\operatorname{cor}_{\mathrm{d}}(\mathbf{e}_{1\cdot 23}, \mathbf{e}_{2\cdot 13}) = -\operatorname{pcor}_{\mathrm{d}}(\mathbf{x}_1, \mathbf{x}_2 \mid \mathbf{X}_3) = -r_{12\cdot 3\ldots k}\,. \tag{8.171}$$

Proof. To prove (8.171), we first note that

$$\operatorname{cor}_{\mathrm{d}}(\mathbf{e}_{1\cdot 23}, \mathbf{e}_{2\cdot 13}) = \frac{\mathbf{x}_1'\mathbf{M}_{23}\mathbf{M}_{13}\mathbf{x}_2}{\sqrt{\mathbf{x}_1'\mathbf{M}_{23}\mathbf{x}_1}\sqrt{\mathbf{x}_2'\mathbf{M}_{13}\mathbf{x}_2}}\,. \tag{8.172}$$

On the other hand,

$$\hat{\beta}_1 = (\mathbf{x}_1'\mathbf{M}_{23}\mathbf{x}_1)^{-1}\mathbf{x}_1'\mathbf{M}_{23}\mathbf{y}\,, \qquad \hat{\beta}_2 = (\mathbf{x}_2'\mathbf{M}_{13}\mathbf{x}_2)^{-1}\mathbf{x}_2'\mathbf{M}_{13}\mathbf{y}\,, \tag{8.173}$$

and, therefore, straightforward calculation shows that

$$\operatorname{cor}(\hat{\beta}_1, \hat{\beta}_2) = \frac{\mathbf{x}_1'\mathbf{M}_{23}\mathbf{M}_{13}\mathbf{x}_2}{\sqrt{\mathbf{x}_1'\mathbf{M}_{23}\mathbf{x}_1}\sqrt{\mathbf{x}_2'\mathbf{M}_{13}\mathbf{x}_2}}\,. \tag{8.174}$$

Now, combining (8.172), (8.174) and (8.165) yields our claim (8.171). □

8.12 Deleting an Observation: $\mathrm{cov}(\mathbf{y}) = \sigma^2\mathbf{I}$

Let us consider three linear models:

$$\mathscr{M} = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_n\}\,, \tag{8.175a}$$

$$\mathscr{M}_{(i)} = \{\mathbf{y}_{(i)},\, \mathbf{X}_{(i)}\boldsymbol{\beta},\, \sigma^2\mathbf{I}_{n-1}\}\,, \tag{8.175b}$$

$$\mathscr{M}_Z = \{\mathbf{y},\, \mathbf{Z}\gamma,\, \sigma^2\mathbf{I}_n\} = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta} + \mathbf{u}\delta,\, \sigma^2\mathbf{I}_n\}\,. \tag{8.175c}$$

By $\mathscr{M}_{(i)}$ we mean such a version of $\mathscr{M}$, where the ith case is deleted; thus $\mathbf{y}_{(i)}$ has $n-1$ elements and $\mathbf{X}_{(i)}$ has $n-1$ rows. For notational simplicity we delete the last observation. Model $\mathscr{M}_Z$ is an *extended* version of $\mathscr{M}$:

$$\mathbf{Z} = (\mathbf{X} : \mathbf{u})\,, \quad \mathbf{u} = (0, \dots, 0, 1)'\,, \quad \gamma = \begin{pmatrix} \boldsymbol{\beta} \\ \delta \end{pmatrix}. \tag{8.176}$$

We will use the following notation:

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\mathscr{M}) = \mathrm{OLSE}(\boldsymbol{\beta}) \text{ under } \mathscr{M}\,, \tag{8.177a}$$

$$\hat{\boldsymbol{\beta}}_Z = \hat{\boldsymbol{\beta}}(\mathscr{M}_Z) = \mathrm{OLSE}(\boldsymbol{\beta}) \text{ under } \mathscr{M}_Z\,, \tag{8.177b}$$

$$\hat{\delta} = \hat{\delta}(\mathscr{M}_Z) = \mathrm{OLSE}(\delta) \text{ under } \mathscr{M}_Z\,, \tag{8.177c}$$

$$\hat{\boldsymbol{\beta}}_{(i)} = \hat{\boldsymbol{\beta}}(\mathscr{M}_{(i)}) = \mathrm{OLSE}(\boldsymbol{\beta}) \text{ under } \mathscr{M}_{(i)}\,. \tag{8.177d}$$

Assume that $\mathbf{Z}$ has full column rank, which implies that

$$\mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{u}) = \{\mathbf{0}\}\,, \tag{8.178}$$

which means, see Section 5.6 (p. 130), that δ is estimable under the model $\mathscr{M}_Z$. Thereby $m_{ii} > 0$ and, in view of (8.36) (p. 160),

$$\hat{\boldsymbol{\beta}}_Z = [\mathbf{X}'(\mathbf{I} - \mathbf{u}\mathbf{u}')\mathbf{X}]^{-1}\mathbf{X}'(\mathbf{I} - \mathbf{u}\mathbf{u}')\mathbf{y}\,, \tag{8.179a}$$

$$\hat{\delta} = \frac{\mathbf{u}'\mathbf{M}\mathbf{y}}{\mathbf{u}'\mathbf{M}\mathbf{u}} = \frac{\hat{\varepsilon}_i}{m_{ii}}\,, \tag{8.179b}$$

where $\hat{\varepsilon}_i$ refers to the ith element of the residual vector $\hat{\boldsymbol{\varepsilon}} = \mathbf{M}\mathbf{y}$. Furthermore, from (8.179a) we see immediately that

$$\hat{\boldsymbol{\beta}}_Z = (\mathbf{X}_{(i)}'\mathbf{X}_{(i)})^{-1}\mathbf{X}_{(i)}'\mathbf{y}_{(i)} = \hat{\boldsymbol{\beta}}_{(i)}\,. \tag{8.180}$$

This result is actually a consequence of the Frisch–Waugh–Lovell theorem: model $\mathscr{M}_Z$ corresponds to the full model $\mathscr{M}_{12}$, model $\mathscr{M}_{(i)}$ to the reduced model $\mathscr{M}_{12\cdot1}$ obtained by premultiplying $\mathscr{M}_Z$ by $\mathbf{I}-\mathbf{u}\mathbf{u}'$, and $\mathscr{M}$ is the “small” model $\mathscr{M}_1$.

Now we can write

$$\mathbf{P}_{\mathbf{Z}}\mathbf{y} = \mathbf{Z}\hat{\boldsymbol{\gamma}} = (\mathbf{X} : \mathbf{u}) \begin{pmatrix} \hat{\boldsymbol{\beta}}_Z \\ \hat{\delta} \end{pmatrix} = \mathbf{X}\hat{\boldsymbol{\beta}}_{(i)} + \mathbf{u}\hat{\delta}\,. \tag{8.181}$$

On the other hand,

$$\begin{aligned} \mathbf{P}_{\mathbf{Z}}\mathbf{y} &= \mathbf{H}\mathbf{y} + \mathbf{P}_{\mathbf{M}\mathbf{u}}\mathbf{y} \\ &= \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{M}\mathbf{u}(\mathbf{u}'\mathbf{M}\mathbf{u})^{-1}\mathbf{u}'\mathbf{M}\mathbf{y} \\ &= \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{M}\mathbf{u}\frac{\hat{\varepsilon}_i}{m_{ii}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{M}\mathbf{u}\hat{\delta}\,. \end{aligned} \tag{8.182}$$

Equations (8.181) and (8.182) imply that

$$\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{(i)} = \mathbf{u}\hat{\delta} - \mathbf{M}\mathbf{u}\hat{\delta} = \mathbf{H}\mathbf{u}\hat{\delta}\,, \tag{8.183a}$$

$$\mathbf{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}) = \mathbf{H}\mathbf{u}\hat{\delta}\,, \tag{8.183b}$$

$$(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}) = h_{ii}\hat{\delta}^2. \tag{8.183c}$$

Premultiplying (8.183b) by $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ yields (cf. Belsley, Kuh & Welsch 1980, p. 13)

$$\mathrm{DFBETA}_i = \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\hat{\delta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\frac{\hat{\varepsilon}_i}{m_{ii}}\,. \tag{8.184}$$

Of course, we could have arrived at (8.184) directly by simply applying formula (8.43):

$$\hat{\boldsymbol{\beta}}_1(\mathscr{M}_{12}) = \hat{\boldsymbol{\beta}}_1(\mathscr{M}_1) - (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\hat{\boldsymbol{\beta}}_2(\mathscr{M}_{12})\,. \tag{8.185}$$

Note also that from (8.183c) we can obtain a representation for the Cook's distance (Cook, 1977):

$$\begin{aligned} \mathrm{COOK}_i^2 &= \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})}{p\hat{\sigma}^2} = \frac{h_{ii}\hat{\delta}^2}{p\hat{\sigma}^2} \\ &= \frac{h_{ii}}{m_{ii}}\,\frac{\hat{\varepsilon}_i^2}{m_{ii}\hat{\sigma}^2}\,\frac{1}{p} = \frac{h_{ii}}{m_{ii}}\,\frac{t_{i*}^2}{p}\,, \end{aligned} \tag{8.186}$$

where

$$t_{i*} = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{m_{ii}}} = \text{ the internally Studentized residual.} \tag{8.187}$$

Notice that we use the term "Cook's distance" from COOK_i^2—not from COOK_i; this seems to be a common practice even though there is a slight inconsistency with the terminology related to the Mahalanobis distance.

Furthermore, because

$$\mathbf{P_Z} = \mathbf{uu}' + \mathbf{P}_{(\mathbf{I}-\mathbf{uu}')\mathbf{X}} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0}' & 1 \end{pmatrix} + \begin{pmatrix} \mathbf{P}_{\mathbf{X}_{(i)}} & \mathbf{0} \\ \mathbf{0}' & 0 \end{pmatrix}, \tag{8.188}$$

we observe that the SSE under the extended model $\mathscr{M}_Z$ becomes

$$\mathrm{SSE}_Z = \mathbf{y}'(\mathbf{I}_n - \mathbf{P_Z})\mathbf{y} = \mathbf{y}_{(i)}'(\mathbf{I}_{n-1} - \mathbf{P}_{\mathbf{X}_{(i)}})\mathbf{y}_{(i)} = \mathrm{SSE}_{(i)}\,, \tag{8.189}$$

which further can be written as

$$\mathrm{SSE}_Z = \mathbf{y}'(\mathbf{I}_n - \mathbf{P_Z})\mathbf{y} = \mathbf{y}'(\mathbf{I}_n - \mathbf{H} - \mathbf{P_{Mu}})\mathbf{y} = \mathbf{y}'(\mathbf{M} - \mathbf{P_{Mu}})\mathbf{y}\,. \tag{8.190}$$

Hence the "usual" unbiased estimator of σ^2 under $\mathscr{M}_Z$ is

$$\tilde{\sigma}_{(i)}^2 = \frac{1}{n - \mathrm{rk}(\mathbf{Z})}\,\mathrm{SSE}_{(i)} = \frac{1}{n-p-1}\,\mathrm{SSE}_{(i)}\,. \tag{8.191}$$

Notice that (8.189) could have been obtained directly from (8.75) (p. 165), which says that the SSEs under the full model and the reduced model are equal. Now it is easy to conclude that the usual F-test statistic for testing the hypothesis $\delta = 0$ under the model $\mathscr{M}_Z$ becomes

$$\begin{aligned} t_i^2 &= \frac{\mathrm{SSE(under\ hypothesis)} - \mathrm{SSE}_Z}{\frac{1}{n-p-1}\,\mathrm{SSE}_Z} = \frac{\mathbf{y}'\mathbf{P_{Mu}}\mathbf{y}}{\frac{1}{n-p-1}\mathbf{y}'(\mathbf{M} - \mathbf{P_{Mu}})\mathbf{y}} \\ &= \frac{\hat{\varepsilon}_i^2}{\frac{1}{n-p-1}\,\mathrm{SSE}_{(i)}\,m_{ii}} = \frac{\hat{\varepsilon}_i^2}{\hat{\sigma}_{(i)}^2 m_{ii}}\,, \end{aligned} \tag{8.192}$$

which is the *externally Studentized residual* (squared).

We note, see (13.18) (p. 293), that applying the determinantal properties of the Schur complement to matrix

$$\mathbf{Z}'\mathbf{Z} = \begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{u} \\ \mathbf{u}'\mathbf{X} & 1 \end{pmatrix}, \tag{8.193}$$

we get

$$|\mathbf{Z}'\mathbf{Z}| = |\mathbf{X}'\mathbf{X}| \cdot |1 - \mathbf{u}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}| = |\mathbf{X}'\mathbf{X}| \cdot (1 - h_{ii}) \tag{8.194a}$$

$$= 1 \cdot |\mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{uu}'\mathbf{X}| = |\mathbf{X}_{(i)}'\mathbf{X}_{(i)}|\,, \tag{8.194b}$$

and hence

$$m_{ii} = 1 - h_{ii} = \frac{|\mathbf{Z}'\mathbf{Z}|}{|\mathbf{X}'\mathbf{X}|} = \frac{|\mathbf{X}_{(i)}'\mathbf{X}_{(i)}|}{|\mathbf{X}'\mathbf{X}|}\,. \tag{8.195}$$

For the consequences of deleting several observations, see Exercise 8.17 (p. 188). Some further generalizations for the considerations of this section appear in Section 15.8 (p. 337).

8.13 Geometrical Illustrations

This section comprises figures that illustrate the geometrical meanings of various concepts presented in this chapter. For geometric considerations in regression, see also Bring (1996), Bryant (1984), and Margolis (1979). For the history of the use of geometry in the linear model, see Herr (1980).

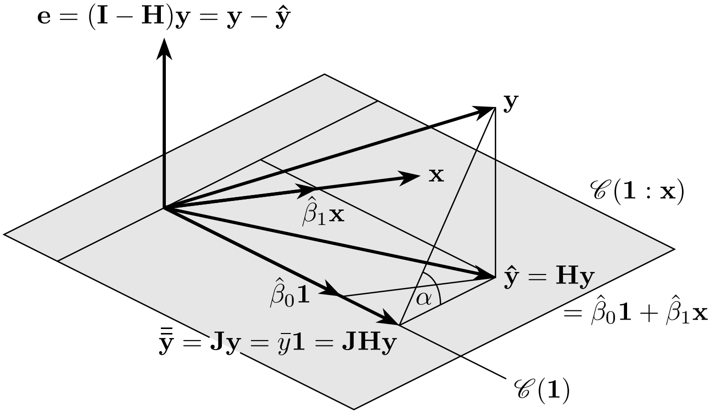


Figure 8.3 Projecting $\mathbf{y}$ onto $\mathscr{C}(\mathbf{1} : \mathbf{x})$.

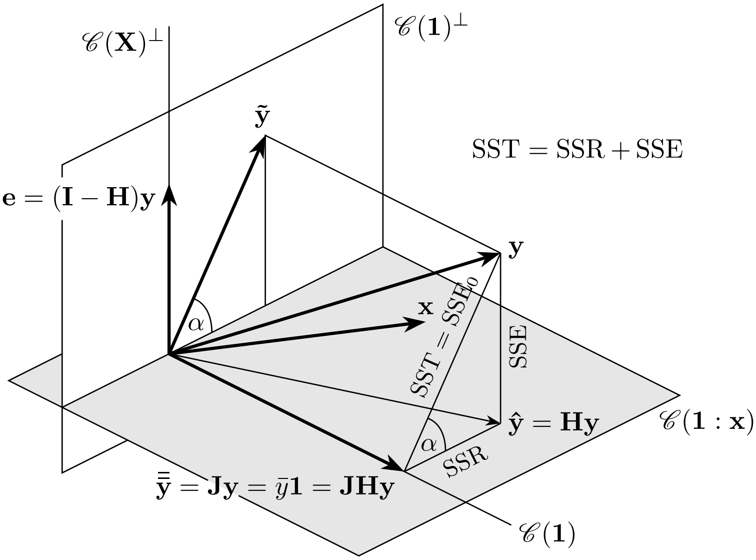


Figure 8.4 Illustration of SST = SSR + SSE.

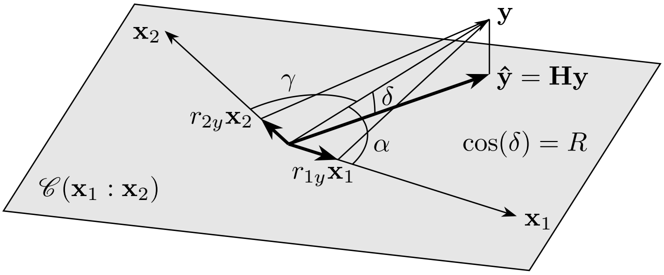


Figure 8.5 Large R^2 but r_{1y} and r_{2y} very small.

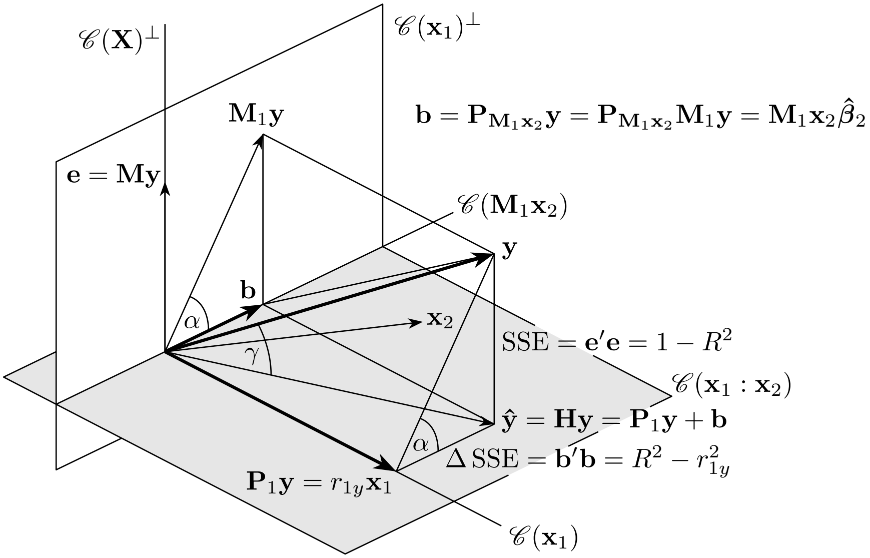


Figure 8.6 Illustration of the reduced model.

8.14 Exercises

8.1. Show that $\bar{y}$ and s_y^2 are independent when $\mathbf{y} \sim \mathrm{N}_n(\boldsymbol{\mu}, \sigma^2\mathbf{I}_n)$.
Hint: Write up $\bar{y}$ and s_y^2 in matrix forms and use the results of (e) on page 18. See also Anderson (2003, p. 77), and Zehna (1991).

8.2. Consider the models $\mathscr{M}_a = \{\mathbf{y}_a\,, \mathbf{X}\boldsymbol{\beta}_a\,, \sigma^2\mathbf{I}\}$ and $\mathscr{M}_b = \{\mathbf{y}_b\,, \mathbf{X}\boldsymbol{\beta}_b\,, \sigma^2\mathbf{I}\}$, where $\mathbf{X} = (\mathbf{1}_n : \mathbf{x}_1 : \mathbf{x}_2 : \mathbf{x}_3)$, $\operatorname{rank}(\mathbf{X}) = 4$, and $\mathbf{y}_b = \mathbf{y}_a - \mathbf{X}\mathbf{z}$, where $\mathbf{z} = (0, 0, 2, 3)'$.

(a) Express $\hat{\boldsymbol{\beta}}_b$ as a function of $\hat{\boldsymbol{\beta}}_a$.
(b) Compare the residuals under $\mathscr{M}_a$ and $\mathscr{M}_b$.
(c) $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{1}_n = ?$

8.3. Let $\mathbf{y} \sim \mathrm{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, where $\mathbf{X} = (\mathbf{1} : \mathbf{X}_0)$, $\boldsymbol{\beta} = \left(\begin{smallmatrix}\beta_0\\ \boldsymbol{\beta}_{\mathbf{x}}\end{smallmatrix}\right)$, and $\operatorname{rank}(\mathbf{X}) = k+1$. Prove the following:

(a) $\mathbf{y}'\mathbf{y}/\sigma^2 \sim \chi^2(n, \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}/\sigma^2)$,
(b) $\mathbf{y}'(\mathbf{I}-\mathbf{J})\mathbf{y}/\sigma^2 \sim \chi^2(n-1, \boldsymbol{\beta}'\mathbf{X}'\mathbf{C}\mathbf{X}\boldsymbol{\beta}/\sigma^2) = \chi^2(n-1, \boldsymbol{\beta}_{\mathbf{x}}'\mathbf{T}_{\mathbf{xx}}\boldsymbol{\beta}_{\mathbf{x}}/\sigma^2)$,
(c) $\mathbf{y}'(\mathbf{H}-\mathbf{J})\mathbf{y}/\sigma^2 = \mathrm{SSR}/\sigma^2 \sim \chi^2(k, \boldsymbol{\beta}_{\mathbf{x}}'\mathbf{T}_{\mathbf{xx}}\boldsymbol{\beta}_{\mathbf{x}}/\sigma^2)$,
(d) $\mathbf{y}'(\mathbf{I}-\mathbf{H})\mathbf{y}/\sigma^2 = \mathrm{SSE}/\sigma^2 \sim \chi^2(n-k-1)$.
(e) SSR and SSE are independent.

8.4. Prove: $(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1} \geq_{\mathsf{L}} (\mathbf{X}_2'\mathbf{X}_2)^{-1}$.

8.5. Consider the matrices $\mathbf{A} \in \mathbb{R}^{n\times a}$, $\mathbf{B} \in \mathbb{R}^{n\times b}$, and $\mathbf{T} \in \mathbb{R}^{n\times t}$, and assume that $\mathscr{C}(\mathbf{T}) \subset \mathscr{C}(\mathbf{A})$. Moreover, let $\mathbf{U}$ be any matrix satisfying $\mathscr{C}(\mathbf{U}) = \mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B})$ and denote $\mathbf{Q}_{\mathbf{T}} = \mathbf{I}_n - \mathbf{P}_{\mathbf{T}}$. Prove that

$$\mathscr{C}(\mathbf{Q}_{\mathbf{T}}\mathbf{U}) = \mathscr{C}(\mathbf{Q}_{\mathbf{T}}\mathbf{A}) \cap \mathscr{C}(\mathbf{Q}_{\mathbf{T}}\mathbf{B}).$$

Markiewicz & Puntanen (2009, Lemma 5).

8.6. Consider the model $\{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}\}$, where $\mathbf{1}_n \in \mathscr{C}(\mathbf{X})$ and denote the OLS residual as $\hat{\boldsymbol{\varepsilon}} = \mathbf{M}\mathbf{y}$. Prove the following:

(a) $\operatorname{cor}_{\mathrm{d}}(\mathbf{y}, \hat{\boldsymbol{\varepsilon}}) = (+)(1-R^2)^{1/2} \geq 0$, i.e., $\mathbf{y}$ and $\hat{\boldsymbol{\varepsilon}}$ cannot be negatively correlated,
(b) $\operatorname{cor}_{\mathrm{d}}(\hat{\mathbf{y}}, \hat{\boldsymbol{\varepsilon}}) = 0$, i.e., the fitted values and residuals are uncorrelated,
(c) $\operatorname{cor}_{\mathrm{d}}(\mathbf{x}_i, \hat{\boldsymbol{\varepsilon}}) = 0$, i.e., each x_i is uncorrelated with the residual,
(d) $\hat{\boldsymbol{\varepsilon}}'\mathbf{1}_n = \mathbf{y}'\mathbf{M}\mathbf{1}_n = 0$, i.e., $\sum_{i=1}^n \hat{\varepsilon}_i = 0 = \bar{\hat{\varepsilon}}$: $\hat{\boldsymbol{\varepsilon}}$ is centered,
(e) $\hat{\boldsymbol{\varepsilon}}'\hat{\mathbf{y}} = 0$, $\hat{\boldsymbol{\varepsilon}}'\mathbf{x}_i = 0$,
(f) $\hat{\mathbf{y}}'\mathbf{1}_n = \mathbf{y}'\mathbf{H}\mathbf{1}_n = \mathbf{y}'\mathbf{1}_n$, i.e., $\frac{1}{n}\sum_{i=1}^n \hat{y}_i = \bar{y}$,
(g) $\operatorname{cov}(\mathbf{y}, \mathbf{H}\mathbf{y}) = \sigma^2\mathbf{H} = \operatorname{cov}(\mathbf{H}\mathbf{y}), \operatorname{cov}(\mathbf{y}, \mathbf{M}\mathbf{y}) = \sigma^2\mathbf{M} = \operatorname{cov}(\mathbf{M}\mathbf{y})$; here $\mathbf{y}$ is a random vector.

8.7. Consider the model $\{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}\}$, where $\mathbf{X}$ has full column rank and

$$\mathbf{X} = (\mathbf{1} : \mathbf{x}_1 : \ldots : \mathbf{x}_q : \mathbf{x}_{q+1} : \ldots : \mathbf{x}_k) = (\mathbf{1} : \mathbf{X}_{01} : \mathbf{X}_2) = (\mathbf{X}_1 : \mathbf{X}_2),$$

and $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1', \hat{\boldsymbol{\beta}}_2')'$. Show that the following statements are equivalent:

(a) $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$,
(b) $\mathbf{X}_2'\mathbf{M}_1\mathbf{y} = \mathbf{0}$,
(c) $\mathbf{y} \in \mathscr{C}(\mathbf{M}_1\mathbf{X}_2)^{\perp} = \mathscr{C}(\mathbf{X})^{\perp} \boxplus \mathscr{C}(\mathbf{X}_1) = \mathscr{C}(\mathbf{X}_1 : \mathbf{X}_2)^{\perp} \boxplus \mathscr{C}(\mathbf{X}_1)$,
(d) $\operatorname{pcor}_{\mathrm{d}}(\mathbf{X}_2, \mathbf{y} \mid \mathbf{X}_{01}) = \mathbf{0}$ or $\mathbf{y} \in \mathscr{C}(\mathbf{X}_1)$.

See also Exercise 3.11 (p. 104).

8.8. Prove, using the notation of Section 8.3 (p. 161):

$$\hat{\beta}_k^2 = r^2_{yk\cdot 12\dots k-1} \cdot \frac{\mathbf{y}'\mathbf{M}_1\mathbf{y}}{\mathbf{x}_k'\mathbf{M}_1\mathbf{x}_k'} = r^2_{yk\cdot 12\dots k-1} \cdot t^{kk} \cdot \mathrm{SSE}(\mathbf{y};\mathbf{X}_1)\,.$$

8.9. Consider the linear model $\mathscr{M} = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}\}$, where $\mathbf{X} = (\mathbf{1} : \mathbf{x})\,, \mathbf{x} = \mathbf{i}_n$. Prove that the t-statistic for testing the hypothesis $\beta_1 = 0$ is

$$t = \frac{y_n - \bar{y}_{(n)}}{s_{(n)}/\sqrt{1-\frac{1}{n}}} = \frac{y_n - \bar{y}}{s_{(n)}\sqrt{1-\frac{1}{n}}}\,,$$

where $\bar{y}_{(n)}$ is the mean of $y_1, \dots, y_{n-1}$ and $s_{(n)}$ their standard deviation. Confirm that this test statistic is the externally Studentized residual; cf. (8.192) (p. 182).

8.10. Let $\mathbf{H} = \mathbf{P}_{\mathbf{X}}$ where $\mathbf{1} \in \mathscr{C}(\mathbf{X})$. Prove that $\frac{1}{n} \le h_{ii} \le \frac{1}{c}$, where $c =$ the number of those rows of $\mathbf{X}$ which are identical with $\mathbf{x}'_{(i)}$ $(c \ge 1)$. Show also that for each each off-diagonal element $h_{ij}^2 \le \frac{1}{4}$.

8.11. Consider the ANOVA model represented in (1.17a) (p. 61), when we have g groups (or treatments) and n_i observations in the ith group:

$$\mathscr{M}\colon \quad \mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_g \end{pmatrix} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_2} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{1}_{n_g} \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_g \end{pmatrix} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\,.$$

Confirm:

$$\mathbf{H} = \begin{pmatrix} \mathbf{J}_{n_1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_{n_2} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{J}_{n_g} \end{pmatrix}, \quad \mathbf{M} = \mathbf{I} - \mathbf{H} = \begin{pmatrix} \mathbf{C}_{n_1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{n_2} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{C}_{n_g} \end{pmatrix},$$

$$\mathbf{H}\mathbf{y} = \begin{pmatrix} \bar{y}_1\mathbf{1}_{n_1} \\ \bar{y}_2\mathbf{1}_{n_2} \\ \vdots \\ \bar{y}_g\mathbf{1}_{n_g} \end{pmatrix}, \quad \mathbf{M}\mathbf{y} = \begin{pmatrix} \mathbf{C}_{n_1}\mathbf{y}_1 \\ \mathbf{C}_{n_2}\mathbf{y}_2 \\ \vdots \\ \mathbf{C}_{n_g}\mathbf{y}_g \end{pmatrix}, \quad (\mathbf{H} - \mathbf{J}_n)\mathbf{y} = \begin{pmatrix} (\bar{y}_1 - \bar{y})\mathbf{1}_{n_1} \\ (\bar{y}_2 - \bar{y})\mathbf{1}_{n_2} \\ \vdots \\ (\bar{y}_g - \bar{y})\mathbf{1}_{n_g} \end{pmatrix},$$

$$\mathrm{SST} = \mathbf{y}'\mathbf{C}_n\mathbf{y} = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2, \quad \mathrm{SSE} = \mathbf{y}'\mathbf{M}\mathbf{y} = \mathrm{SS}_1 + \dots + \mathrm{SS}_g\,,$$

$$\mathrm{SSR} = \mathbf{y}'(\mathbf{H} - \mathbf{J}_n)\mathbf{y} = \sum_{i=1}^{g} n_i(\bar{y}_i - \bar{y})^2 := \mathrm{SSB}\,, \quad \mathrm{SST} = \mathrm{SSR} + \mathrm{SSE}\,.$$

Observe that under the hypothesis $H\colon \mu_1 = \dots = \mu_g$ we have $\mathrm{SSE}_H = \mathrm{SST}$ and hence $\mathrm{SSE}_H - \mathrm{SSE} = \mathrm{SSB}$, and, under normality, the F-test statistics for testing H is

$$F = \frac{\mathbf{y}'(\mathbf{H} - \mathbf{J}_n)\mathbf{y}}{\mathbf{y}'\mathbf{M}\mathbf{y}} = \frac{\mathrm{SSB}\,/(g-1)}{\mathrm{SSE}\,/(n-g)} \sim \mathrm{F}(g-1, n-g)\,.$$

In this context, a common practice is to use notation

$$\mathrm{SSR} = \mathrm{SSB} = \mathrm{SS}_{\text{Between}}\,, \quad \mathrm{SSE} = \mathrm{SS}_{\text{Within}}\,.$$

For the multivariate versions of the above sum of squares, see page 233.

8.12 (Continued ...). Show that in the case of two groups the F-statistics for the hypothesis $\mu_1 = \mu_2$ can be written as

$$\begin{aligned} F &= \frac{(\bar{y}_1 - \bar{y}_2)^2}{\dfrac{\mathrm{SS}_1 + \mathrm{SS}_2}{n-2}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)} = \frac{n_1(\bar{y}_1 - \bar{y})^2 + n_2(\bar{y}_2 - \bar{y})^2}{\dfrac{\mathrm{SS}_1 + \mathrm{SS}_2}{n-2}} \\ &= \frac{n_1 n_2}{n} \cdot (\bar{y}_1 - \bar{y}_2) \cdot \left(\frac{\mathrm{SS}_1 + \mathrm{SS}_2}{n-2}\right)^{-1} \cdot (\bar{y}_1 - \bar{y}_2) \\ &= \frac{(\bar{y}_1 - \bar{y}_2)^2}{\dfrac{\mathrm{SSE}}{n-2}\,\dfrac{n}{n_1 n_2}} \sim \mathrm{F}(1, n-2) = \mathrm{t}^2(n-2), \qquad n_1 + n_2 = n\,. \end{aligned}$$

Notice the connection between the above F and the Hotelling's T^2 in (10.115) (p. 234). See also Exercise 0.25 (p. 53).

8.13. Notice that in Figure 8.1 (p. 169) we generate observations from a two-dimensional distribution $\mathrm{N}_2(\mathbf{0}, \boldsymbol{\Sigma})$; $\sigma_x = \sigma_y = 1$, $\varrho_{xy} = 0.7$. Hence also the explanatory variable x is a random variable which is *not* the traditional situation when generating observations from a linear model: there the values of x are assumed to be fixed and only y is a random vector. Confirm that in the random-case (under multinormality) the hypothesis $\varrho_{xy} = 0$ can be tested in the "usual" way by calculating $F = \frac{r_{xy}^2}{(1-r_{xy}^2)/(n-2)}$. Moreover, convince yourself that in nonrandom case, i.e., in a standard linear model, there is no parameter corresponding to ϱ_{xy} and thereby r_{xy} is a summary statistics of the regression but not an estimate of the nonexisting "population correlation coefficient".

Weisberg (2005, pp. 80–82).

8.14. Show that in (8.159) (p. 177)

$$\begin{aligned} Q(\mathbf{V}) &= (\mathbf{K}'\tilde{\boldsymbol{\beta}} - \mathbf{d})'[\mathbf{K}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{K}]^{-1}(\mathbf{K}'\tilde{\boldsymbol{\beta}} - \mathbf{d}) \\ &= (\mathbf{K}'\tilde{\boldsymbol{\beta}} - \mathbf{d})'[\mathrm{cov}(\mathbf{K}'\tilde{\boldsymbol{\beta}})]^{-1}(\mathbf{K}'\tilde{\boldsymbol{\beta}} - \mathbf{d})\sigma^2, \end{aligned}$$

and the restricted BLUE for $\boldsymbol{\beta}$ (supposing $\mathbf{X}$ has full column rank) is

$$\tilde{\boldsymbol{\beta}}_r = \tilde{\boldsymbol{\beta}} - (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{K}[\mathbf{K}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{K}]^{-1}(\mathbf{K}'\tilde{\boldsymbol{\beta}} - \mathbf{d})\,.$$

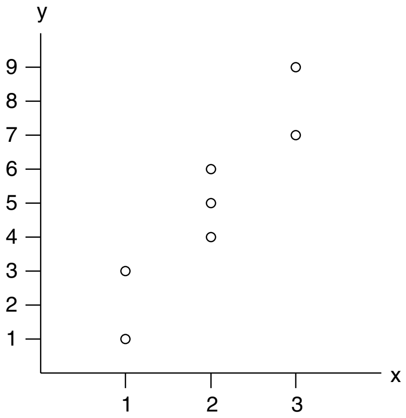


Figure 8.7 (Exercise 8.15) Three treatments, 1, 2 and 3; seven observations.

8.15. Consider the scatter plot of Figure 8.7, where we have seven data points and we have three treatments whose effect on y we wish to study using one-way ANOVA. Show that the F-statistic for testing the hypothesis $H\colon \mu_1 = \mu_2 = \mu_3$ has value $F = 12$ with $g - 1 = 2$ and $n - g = 4$ degrees of freedom, and

$$\mathrm{SST} = \mathrm{SS}_{\mathrm{Between}} + \mathrm{SS}_{\mathrm{Within}}\colon \quad 42 = 36 + 6\,.$$

8.16 (Continued ...). Suppose that the scatter plot of Figure 8.7 represents a two-dimensional discrete uniform distribution of random variables x and y. Let $\mathrm{E}(y \mid x)$ denote a random variable whose values are conditional means $\mathrm{E}(y \mid x = 1)$, $\mathrm{E}(y \mid x = 2)$, $\mathrm{E}(y \mid x = 3)$ with probabilities $\mathrm{P}(x = 1)$, $\mathrm{P}(x = 2)$, $\mathrm{P}(x = 3)$, respectively. Define the random variable $\mathrm{var}(y \mid x)$ in the corresponding way and show that

$$\mathrm{var}(y) = \mathrm{var}[\mathrm{E}(y \mid x)] + \mathrm{E}[\mathrm{var}(y \mid x)]\,. \tag{8.196}$$

Casella (2008, p. 9) calls (8.196) as “a most important equality” in his book on *Statistical Design*. Confirm the connection between (8.196) and the ANOVA decomposition $\mathrm{SST} = \mathrm{SSB} + \mathrm{SSE}$.

8.17 (Deleting the last q observations). Consider the models

$$\mathscr{M} = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}\} \text{ and its extended version } \mathscr{M}_Z = \{\mathbf{y},\, \mathbf{Z}\boldsymbol{\gamma},\, \sigma^2\mathbf{I}\},$$

where

$$\mathbf{Z} = (\mathbf{X} : \mathbf{U}), \quad \mathbf{X}_{n\times p} = \begin{pmatrix} \mathbf{X}_{(1)} \\ \mathbf{X}_{(2)} \end{pmatrix}, \quad \mathbf{X}_{(1)} \in \mathbb{R}^{(n-q)\times p}, \quad \mathbf{X}_{(2)} \in \mathbb{R}^{q\times p},$$

$$\mathbf{U}_{n\times q} = \begin{pmatrix} \mathbf{0} \\ \mathbf{I}_q \end{pmatrix}, \ \boldsymbol{\gamma} = \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\delta} \end{pmatrix}, \ \boldsymbol{\beta} \in \mathbb{R}^p, \ \boldsymbol{\delta} \in \mathbb{R}^q, \ \mathbf{H} = \begin{pmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} \\ \mathbf{H}_{21} & \mathbf{H}_{22} \end{pmatrix},$$

$$\mathbf{M} = \begin{pmatrix} \mathbf{I}_{n-q} - \mathbf{H}_{11} & -\mathbf{H}_{12} \\ -\mathbf{H}_{21} & \mathbf{I}_q - \mathbf{H}_{22} \end{pmatrix}.$$

Let $\mathscr{M}_{[2]}$ denote the version of model $\mathscr{M}$ where the last q observations have been deleted. Confirm the following statements:

(a) $\operatorname{rank}(\mathbf{M}_{22}) = q - \dim \mathscr{C}(\mathbf{U}) \cap \mathscr{C}(\mathbf{X}) = q - \operatorname{rank}(\mathbf{X}_{(2)}) + \dim \mathscr{C}(\mathbf{X}'_{(1)}) \cap \mathscr{C}(\mathbf{X}'_{(2)})$.

(b) $\mathbf{M}_{22}$ is positive definite $\iff \mathscr{C}(\mathbf{X}'_{(2)}) \subset \mathscr{C}(\mathbf{X}'_{(1)}) \iff \operatorname{rank}(\mathbf{X}_{(1)}) = \operatorname{rank}(\mathbf{X}) \iff \operatorname{ch}_1(\mathbf{U}'\mathbf{H}\mathbf{U}) = \operatorname{ch}_1(\mathbf{H}_{22}) < 1 \iff \mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{U}) = \{\mathbf{0}\} \iff \boldsymbol{\delta}$ is estimable under $\mathscr{M}_Z = \{\mathbf{y}, \mathbf{Z}\boldsymbol{\gamma}, \sigma^2\mathbf{I}\}$.

(c) If $\mathbf{X}_{(1)}$ has full column rank, then

$$\hat{\boldsymbol{\delta}} = (\mathbf{U}'\mathbf{M}\mathbf{U})^{-1}\mathbf{U}'\mathbf{M}\mathbf{y} = \mathbf{M}_{22}^{-1}\hat{\boldsymbol{\varepsilon}}_2 = \mathbf{M}_{22}^{-1}\mathbf{M}_{21}\mathbf{y}_1 + \mathbf{y}_2\,,$$

where $\hat{\boldsymbol{\varepsilon}}_2$ is the lower part of $\hat{\boldsymbol{\varepsilon}} = \mathbf{M}\mathbf{y}$, and $\mathbf{y}_2$ is the lower part of $\mathbf{y}$.

(d) $\mathrm{SSE}_Z = \mathbf{y}'(\mathbf{I} - \mathbf{P}_\mathbf{Z})\mathbf{y} = \mathrm{SSE}_{[2]} = \mathrm{SSE} - \mathbf{y}'\mathbf{P}_{\mathbf{MU}}\mathbf{y} = \mathrm{SSE} - \hat{\boldsymbol{\varepsilon}}_2'\mathbf{M}_{22}^{-1}\hat{\boldsymbol{\varepsilon}}_2 = \mathrm{SSE} - \hat{\boldsymbol{\delta}}'\mathbf{M}_{22}\hat{\boldsymbol{\delta}}$, where $\mathrm{SSE} = \mathbf{y}'\mathbf{M}\mathbf{y}$ and

$$\mathrm{SSE}_Z = \mathrm{SSE}_{[2]} = \mathbf{y}_1'(\mathbf{I}_{n-q} - \mathbf{P}_{\mathbf{X}_{(1)}})\mathbf{y}_1 \text{ is the SSE under } \mathscr{M}_{[2]}\,.$$

(e) $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{[2]} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{U}\mathbf{M}_{22}^{-1}\mathbf{U}'\mathbf{M}\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_{(2)}\mathbf{M}_{22}^{-1}\hat{\boldsymbol{\varepsilon}}_2$
$= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{U}\hat{\boldsymbol{\delta}}$, where $\hat{\boldsymbol{\beta}}_{[2]}$ is the OLSE of $\boldsymbol{\beta}$ under $\mathscr{M}_{[2]}$.

(f) $\mathscr{C}(\mathbf{X}'_{(1)}) \cap \mathscr{C}(\mathbf{X}'_{(2)}) = \{\mathbf{0}\} \iff \mathbf{H} = \begin{pmatrix} \mathbf{P}_{\mathbf{X}_{(1)}} & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_{\mathbf{X}_{(2)}} \end{pmatrix}$; see also Theorem 16 (p. 343).

(g) The F-test statistics for testing $H\colon \boldsymbol{\delta} = \mathbf{0}$ under $\mathscr{M}_Z$ can be written as

$$\begin{aligned} t^2_{[2]} &= \frac{\mathbf{y}'\mathbf{P}_{\mathbf{MU}}\mathbf{y}/q}{\mathbf{y}'(\mathbf{M} - \mathbf{P}_{\mathbf{MU}})\mathbf{y}/(n-p-q)} = \frac{\mathbf{y}'\mathbf{M}\mathbf{U}(\mathbf{U}'\mathbf{M}\mathbf{U})^{-1}\mathbf{U}'\mathbf{M}\mathbf{y}/q}{\mathrm{SSE}_{[2]}/(n-p-q)} \\ &= \frac{\hat{\boldsymbol{\varepsilon}}_2'\mathbf{M}_{22}^{-1}\hat{\boldsymbol{\varepsilon}}_2/q}{\mathrm{SSE}_{[2]}/(n-p-q)} \sim F(q, n-p-q, \boldsymbol{\delta}'\mathbf{M}_{22}\boldsymbol{\delta}/\sigma^2) \\ &= \text{the multiple case analogue of externally Studentized residual.} \end{aligned}$$

8.18. Consider the variables x_1, x_2 and y whose observed data matrix is $(\mathbf{x}_1 : \mathbf{x}_2 : \mathbf{y}) = (\mathbf{X}_0 : \mathbf{y})$ and the sample covariance matrix is

$$\text{(a) } \mathbf{S} = s^2 \begin{pmatrix} 1 & r & r \\ r & 1 & r \\ r & r & 1 \end{pmatrix} = \begin{pmatrix} \mathbf{S}_{xx} & \mathbf{s}_{\mathbf{x}y} \\ \mathbf{s}'_{\mathbf{x}y} & s^2 \end{pmatrix}, \quad \text{(b) } \mathbf{S} = s^2 \begin{pmatrix} 1 & r & r^2 \\ r & 1 & r \\ r^2 & r & 1 \end{pmatrix}.$$

Find $\mathbf{b}_* \in \mathbb{R}^2$ so that $\min_{\mathbf{b}} \operatorname{var}_{\mathrm{d}}(\mathbf{y} - \mathbf{X}_0\mathbf{b}) = \operatorname{var}_{\mathrm{d}}(\mathbf{y} - \mathbf{X}_0\mathbf{b}_*)$. Confirm that $\mathbf{b}_*$ maximizes $\operatorname{cor}_{\mathrm{d}}(\mathbf{y}, \mathbf{X}_0\mathbf{b})$.

8.19 (Properties of intersection $\mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B})$). Let $\mathbf{L}$ be a matrix which spans the intersection of $\mathscr{C}(\mathbf{A})$ and $\mathscr{C}(\mathbf{B})$, i.e., $\mathscr{C}(\mathbf{L}) = \mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B})$ Prove:

(a) $\mathscr{C}(\mathbf{A}) = \mathscr{C}[\mathbf{L} : (\mathbf{I} - \mathbf{P}_{\mathbf{L}})\mathbf{A}] = \mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B}) \boxplus \mathscr{C}[(\mathbf{I} - \mathbf{P}_{\mathbf{L}})\mathbf{A}]$,
(b) $\mathbf{P}_{\mathbf{A}} = \mathbf{P}_{\mathbf{L}} + \mathbf{P}_{(\mathbf{I}-\mathbf{P}_{\mathbf{L}})\mathbf{A}} = \mathbf{P}_{\mathbf{L}} + \mathbf{P}_{\mathscr{C}(\mathbf{A})\cap\mathscr{C}(\mathbf{L})^{\perp}}$,
(c) $\mathbf{P}_{\mathbf{A}}\mathbf{P}_{\mathbf{B}} = \mathbf{P}_{\mathbf{L}} + \mathbf{P}_{(\mathbf{I}-\mathbf{P}_{\mathbf{L}})\mathbf{A}}\mathbf{P}_{(\mathbf{I}-\mathbf{P}_{\mathbf{L}})\mathbf{B}}$,
(d) $(\mathbf{I} - \mathbf{P}_{\mathbf{L}})\mathbf{P}_{\mathbf{A}} = \mathbf{P}_{\mathbf{A}}(\mathbf{I} - \mathbf{P}_{\mathbf{L}}) = \mathbf{P}_{\mathbf{A}} - \mathbf{P}_{\mathbf{L}} = \mathbf{P}_{\mathscr{C}(\mathbf{A})\cap\mathscr{C}(\mathbf{L})^{\perp}}$,
(e) $\mathscr{C}[(\mathbf{I} - \mathbf{P}_{\mathbf{L}})\mathbf{A}] = \mathscr{C}[(\mathbf{I} - \mathbf{P}_{\mathbf{L}})\mathbf{P}_{\mathbf{A}}] = \mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{L})^{\perp}$,
(f) $\operatorname{rank}(\mathbf{A}) = \dim \mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B}) + \dim \mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{L})^{\perp}$.

8.20 (Commuting projectors). Denote $\mathscr{C}(\mathbf{L}) = \mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B})$. Prove that the following statements are equivalent; see Baksalary (1987):

(a) $\mathbf{P}_{\mathbf{A}}\mathbf{P}_{\mathbf{B}} = \mathbf{P}_{\mathbf{B}}\mathbf{P}_{\mathbf{A}}$,
(b) $\mathbf{P}_{\mathbf{A}}\mathbf{P}_{\mathbf{B}} = \mathbf{P}_{\mathscr{C}(\mathbf{A})\cap\mathscr{C}(\mathbf{B})} = \mathbf{P}_{\mathbf{L}}$,
(c) $\mathbf{P}_{(\mathbf{A}\,:\,\mathbf{B})} = \mathbf{P}_{\mathbf{A}} + \mathbf{P}_{\mathbf{B}} - \mathbf{P}_{\mathbf{A}}\mathbf{P}_{\mathbf{B}}$,
(d) $\mathscr{C}(\mathbf{A} : \mathbf{B}) \cap \mathscr{C}(\mathbf{B})^{\perp} = \mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B})^{\perp}$,
(e) $\mathscr{C}(\mathbf{A}) = \mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B}) \boxplus \mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B})^{\perp}$,
(f) $\operatorname{rank}(\mathbf{A}) = \dim \mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B}) + \dim \mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B})^{\perp}$,
(g) $\operatorname{rank}(\mathbf{A}'\mathbf{B}) = \dim \mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B})$,
(h) $\mathbf{P}_{(\mathbf{I}-\mathbf{P}_{\mathbf{A}})\mathbf{B}} = \mathbf{P}_{\mathbf{B}} - \mathbf{P}_{\mathbf{A}}\mathbf{P}_{\mathbf{B}}$,
(i) $\mathscr{C}(\mathbf{P}_{\mathbf{A}}\mathbf{B}) = \mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B})$,
(j) $\mathscr{C}(\mathbf{P}_{\mathbf{A}}\mathbf{B}) \subset \mathscr{C}(\mathbf{B})$,
(k) $\mathbf{P}_{(\mathbf{I}-\mathbf{P}_{\mathbf{L}})\mathbf{A}}\mathbf{P}_{(\mathbf{I}-\mathbf{P}_{\mathbf{L}})\mathbf{B}} = \mathbf{0}$.

8.21. Suppose that $\mathbf{P}_{\mathbf{A}}\mathbf{P}_{\mathbf{B}} = \mathbf{P}_{\mathbf{B}}\mathbf{P}_{\mathbf{A}}$. Show that then

$$\mathscr{C}(\mathbf{P}_{\mathbf{A}} : \mathbf{P}_{\mathbf{B}}) = \mathscr{C}(\mathbf{P}_{\mathbf{A}}\mathbf{P}_{\mathbf{B}}) \boxplus \mathscr{C}(\mathbf{P}_{\mathbf{A}}\mathbf{Q}_{\mathbf{B}}) \boxplus \mathscr{C}(\mathbf{P}_{\mathbf{B}}\mathbf{Q}_{\mathbf{A}}),$$

where $\mathbf{Q}_{\mathbf{A}} = \mathbf{I} - \mathbf{P}_{\mathbf{A}}$ and $\mathbf{Q}_{\mathbf{B}} = \mathbf{I} - \mathbf{P}_{\mathbf{B}}$.

Groß & Trenkler (1998), Kala (2009).

Chapter 9
Minimizing cov(y − Fx)

> *I don't play accurately—anyone can play accurately—but I play with a wonderful expression.*
>
> Oscar Wilde: *The Importance of Being Earnest*

In this chapter we consider the problem of finding the minimum—in the Löwner sense—for the covariance matrix of $\mathbf{y} - \mathbf{F}\mathbf{x}$ where $\mathbf{y}$ and $\mathbf{x}$ are given random vectors and the matrix $\mathbf{F}$ is free to vary. This is a fundamental task in linear models and multivariate analysis and the solution, utilized in several places in this book, is very much worth remembering.

Theorem 9 (Minimizing $\operatorname{cov}(\mathbf{y}-\mathbf{F}\mathbf{x})$). *Let $\mathbf{x}$ and $\mathbf{y}$ be p- and q-dimensional random vectors, respectively, and let $\mathbf{z} = \binom{\mathbf{x}}{\mathbf{y}}$ be a partitioned random vector with covariance matrix*

$$\operatorname{cov}(\mathbf{z}) = \operatorname{cov}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix} = \begin{pmatrix}\mathbf{\Sigma_{xx}} & \mathbf{\Sigma_{xy}}\\ \mathbf{\Sigma_{yx}} & \mathbf{\Sigma_{yy}}\end{pmatrix} = \mathbf{\Sigma}\,. \tag{9.1}$$

Then

$$\operatorname{cov}(\mathbf{y} - \mathbf{F}\mathbf{x}) \geq_{\mathsf{L}} \operatorname{cov}(\mathbf{y} - \mathbf{\Sigma_{yx}}\mathbf{\Sigma_{xx}^{-}}\mathbf{x}) \quad \textit{for all } \mathbf{F} \in \mathbb{R}^{q\times p}, \tag{9.2}$$

and the minimal covariance matrix is

$$\operatorname{cov}(\mathbf{y} - \mathbf{\Sigma_{yx}}\mathbf{\Sigma_{xx}^{-}}\mathbf{x}) = \mathbf{\Sigma_{yy}} - \mathbf{\Sigma_{yx}}\mathbf{\Sigma_{xx}^{-}}\mathbf{\Sigma_{xy}} = \mathbf{\Sigma_{yy\cdot x}} = \mathbf{\Sigma}/\mathbf{\Sigma_{xx}}\,, \tag{9.3}$$

the Schur complement of $\mathbf{\Sigma_{xx}}$ in $\mathbf{\Sigma}$.

Proof. The proof is very simple with the help of the Block Diagonalization Theorem 13 (p. 291). Let us denote

$$\mathbf{u} = \begin{pmatrix}\mathbf{u}_1\\ \mathbf{u}_2\end{pmatrix} = \mathbf{B}\mathbf{z} = \begin{pmatrix}\mathbf{I}_p & \mathbf{0}\\ -\mathbf{\Sigma_{yx}}\mathbf{\Sigma_{xx}^{-}} & \mathbf{I}_q\end{pmatrix}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix} = \begin{pmatrix}\mathbf{x}\\ \mathbf{y} - \mathbf{\Sigma_{yx}}\mathbf{\Sigma_{xx}^{-}}\mathbf{x}\end{pmatrix}. \tag{9.4}$$

Then

$$\operatorname{cov}(\mathbf{u}) = \mathbf{B}\mathbf{\Sigma}\mathbf{B}' = \begin{pmatrix}\mathbf{\Sigma_{xx}} & \mathbf{0}\\ \mathbf{0} & \mathbf{\Sigma_{yy}} - \mathbf{\Sigma_{yx}}\mathbf{\Sigma_{xx}^{-}}\mathbf{\Sigma_{xy}}\end{pmatrix}, \tag{9.5}$$

indicating that $\mathbf{u}_1$ and $\mathbf{u}_2$ are uncorrelated:

$$\operatorname{cov}(\mathbf{u}_1, \mathbf{u}_2) = \operatorname{cov}(\mathbf{x}, \mathbf{y} - \mathbf{\Sigma_{yx}}\mathbf{\Sigma_{xx}^{-}}\mathbf{x}) = \mathbf{0}\,. \tag{9.6}$$

S. Puntanen et al., *Matrix Tricks for Linear Statistical Models: Our Personal Top Twenty*,
DOI 10.1007/978-3-642-10473-2_10, © Springer-Verlag Berlin Heidelberg 2011

We can obviously write

$$\begin{aligned}\operatorname{cov}(\mathbf{y}-\mathbf{F}\mathbf{x}) &= \operatorname{cov}[(\mathbf{y}-\boldsymbol{\Sigma}_{\mathbf{yx}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-}\mathbf{x}) + (\boldsymbol{\Sigma}_{\mathbf{yx}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-}-\mathbf{F})\mathbf{x}] \\ &:= \operatorname{cov}(\mathbf{v}_1+\mathbf{v}_2)\,. \end{aligned} \tag{9.7}$$

In light of (9.6), the random vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ are uncorrelated and hence (9.7) can be written as

$$\begin{aligned}\operatorname{cov}(\mathbf{y}-\mathbf{F}\mathbf{x}) &= \operatorname{cov}(\mathbf{y}-\boldsymbol{\Sigma}_{\mathbf{yx}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-}\mathbf{x}) + \operatorname{cov}[(\boldsymbol{\Sigma}_{\mathbf{yx}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-}-\mathbf{F})\mathbf{x}] \\ &\geq_{\mathsf{L}} \operatorname{cov}(\mathbf{y}-\boldsymbol{\Sigma}_{\mathbf{yx}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-}\mathbf{x})\,. \end{aligned} \tag{9.8}$$

Thus our claim is proved. □

We can approach the above task also through projections. To do this, let us express the covariance matrix of $\binom{\mathbf{x}}{\mathbf{y}}$ as

$$\operatorname{cov}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\Sigma}_{\mathbf{xx}} & \boldsymbol{\Sigma}_{\mathbf{xy}}\\ \boldsymbol{\Sigma}_{\mathbf{yx}} & \boldsymbol{\Sigma}_{\mathbf{yy}}\end{pmatrix} = \begin{pmatrix}\mathbf{A}'\mathbf{A} & \mathbf{A}'\mathbf{B}\\ \mathbf{B}'\mathbf{A} & \mathbf{B}'\mathbf{B}\end{pmatrix}, \tag{9.9}$$

where $\mathbf{A} \in \mathbb{R}^{u\times p}$, $\mathbf{B} \in \mathbb{R}^{u\times q}$ (for some u). Then

$$\begin{aligned}\operatorname{cov}(\mathbf{y}-\mathbf{F}\mathbf{x}) &= \mathbf{B}'\mathbf{B} + \mathbf{F}\mathbf{A}'\mathbf{A}\mathbf{F}' - \mathbf{B}'\mathbf{A}\mathbf{F}' - \mathbf{F}\mathbf{A}'\mathbf{B} \\ &= (\mathbf{B}-\mathbf{A}\mathbf{F}')'(\mathbf{B}-\mathbf{A}\mathbf{F}')\,. \end{aligned} \tag{9.10}$$

In view of Proposition 2.5 (p. 81), we have

$$(\mathbf{B}-\mathbf{A}\mathbf{F}')'(\mathbf{B}-\mathbf{A}\mathbf{F}') \geq_{\mathsf{L}} \mathbf{B}'(\mathbf{I}_u - \mathbf{P}_{\mathbf{A}})\mathbf{B}\,, \tag{9.11}$$

where the equality is attained if and only if

$$\mathbf{A}\mathbf{F}' = \mathbf{P}_{\mathbf{A}}\mathbf{B} = \mathbf{A}\boldsymbol{\Sigma}_{\mathbf{xx}}^{+}\boldsymbol{\Sigma}_{\mathbf{xy}}\,, \tag{9.12}$$

in which case $\mathbf{F}'$ can be expressed as $\mathbf{F}'_{*} = (\mathbf{A}'\mathbf{A})^{+}\mathbf{A}'\mathbf{B} = \boldsymbol{\Sigma}_{\mathbf{xx}}^{+}\boldsymbol{\Sigma}_{\mathbf{xy}}$. Hence

$$\operatorname{cov}(\mathbf{y}-\mathbf{F}\mathbf{x}) \geq_{\mathsf{L}} \operatorname{cov}(\mathbf{y}-\mathbf{F}_{*}\mathbf{x}) = \operatorname{cov}(\mathbf{y}-\boldsymbol{\Sigma}_{\mathbf{yx}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{+}\mathbf{x})\,, \tag{9.13}$$

where $\boldsymbol{\Sigma}_{\mathbf{xx}}^{+}$ can be replaced with any $\boldsymbol{\Sigma}_{\mathbf{xx}}^{-}$.

In view of Proposition 1.4 (p. 64), we can "extend" Theorem 9 as follows.

Corollary 9.1. *The following statements are equivalent:*

(a) $\operatorname{cov}(\mathbf{y}-\mathbf{F}\mathbf{x}) \geq_{\mathsf{L}} \operatorname{cov}(\mathbf{y}-\mathbf{A}\mathbf{x})$ *for all* $\mathbf{F} \in \mathbb{R}^{q\times p}$.

(b) $\operatorname{cov}(\mathbf{x}, \mathbf{y}-\mathbf{A}\mathbf{x}) = \mathbf{0}$.

(c) *The matrix* $\mathbf{A}$ *is a solution to* $\mathbf{A}\boldsymbol{\Sigma}_{\mathbf{xx}} = \boldsymbol{\Sigma}_{\mathbf{yx}}$.

(d) *The matrix* $\mathbf{A}$ *is of the form*

$$\mathbf{A} = \boldsymbol{\Sigma}_{\mathbf{yx}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-} + \mathbf{L}(\mathbf{I}_p - \boldsymbol{\Sigma}_{\mathbf{xx}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-})\,, \tag{9.14}$$

where $\mathbf{L} \in \mathbb{R}^{q\times p}$ *is free to vary.*

(e) $\mathbf{A}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}}) = \boldsymbol{\Sigma}_{\mathbf{yx}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}})$ *with probability* 1.

In particular, if the random vector $\mathbf{z}$ is partitioned so that

$$\operatorname{cov}(\mathbf{z}) = \operatorname{cov}\begin{pmatrix}\mathbf{x}\\ y\end{pmatrix} = \begin{pmatrix}\boldsymbol{\Sigma}_{\mathbf{xx}} & \boldsymbol{\sigma}_{\mathbf{x}y}\\ \boldsymbol{\sigma}'_{\mathbf{x}y} & \sigma_y^2\end{pmatrix} = \boldsymbol{\Sigma}\,, \tag{9.15}$$

then

$$\begin{aligned}\min_{\mathbf{f}} \operatorname{var}(y-\mathbf{f}'\mathbf{x}) &= \operatorname{var}(y-\boldsymbol{\sigma}'_{\mathbf{x}y}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-}\mathbf{x}) = \sigma_y^2 - \boldsymbol{\sigma}'_{\mathbf{x}y}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-}\boldsymbol{\sigma}_{\mathbf{x}y}\\ &= \sigma_y^2\Big(1-\frac{\boldsymbol{\sigma}'_{\mathbf{x}y}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-}\boldsymbol{\sigma}_{\mathbf{x}y}}{\sigma_y^2}\Big) = \sigma_y^2(1-\varrho_{y\cdot\mathbf{x}}^2) = \sigma_{y\cdot\mathbf{x}}^2\,, \end{aligned}\tag{9.16}$$

where the minimum is obtained when $\mathbf{f}$ is

$$\mathbf{f}_* = \boldsymbol{\Sigma}_{\mathbf{xx}}^{-}\boldsymbol{\sigma}_{\mathbf{x}y}\,. \tag{9.17}$$

On the other hand, in view the Cauchy–Schwarz inequality, we have (assuming $\sigma_y^2 > 0$):

$$\begin{aligned}\operatorname{cor}^2(y,\mathbf{f}'\mathbf{x}) &= \frac{(\boldsymbol{\sigma}'_{\mathbf{x}y}\mathbf{f})^2}{\sigma_y^2\cdot\mathbf{f}'\boldsymbol{\Sigma}_{\mathbf{xx}}\mathbf{f}} \le \frac{\boldsymbol{\sigma}'_{\mathbf{x}y}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-}\boldsymbol{\sigma}_{\mathbf{x}y}\cdot\mathbf{f}'\boldsymbol{\Sigma}_{\mathbf{xx}}\mathbf{f}}{\sigma_y^2\cdot\mathbf{f}'\boldsymbol{\Sigma}_{\mathbf{xx}}\mathbf{f}}\\ &= \frac{\boldsymbol{\sigma}'_{\mathbf{x}y}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-}\boldsymbol{\sigma}_{\mathbf{x}y}}{\sigma_y^2} = \varrho_{y\cdot\mathbf{x}}^2 = \operatorname{cor}^2(y,\mathbf{f}'_*\mathbf{x})\,. \end{aligned}\tag{9.18}$$

Thus we have (again) confirmed that the tasks of solving $\mathbf{f}$ from

$$\min_{\mathbf{f}} \operatorname{var}(y-\mathbf{f}'\mathbf{x}) \quad\text{and}\quad \max_{\mathbf{f}} \operatorname{cor}^2(y,\mathbf{f}'\mathbf{x})\,, \tag{9.19}$$

yield essentially the same solutions $\mathbf{f}_* = \boldsymbol{\Sigma}_{\mathbf{xx}}^{-}\boldsymbol{\sigma}_{\mathbf{x}y}$. The corresponding properties concerning the sample variance and sample correlation are considered, e.g., in Section 2.2 (p. 79); see also Remark 2.2 (p. 79). The maximal correlation $\varrho_{y\cdot\mathbf{x}}^2$ is the squared *population multiple correlation coefficient.*

Of course, if $\sigma_y^2 = 0$, the multiple correlation coefficient $\varrho_{y\cdot\mathbf{x}}$ is not well-defined. However, in light of (1.35) (p. 63), $\sigma_y^2 = 0$ implies that $\boldsymbol{\sigma}_{\mathbf{x}y} = \mathbf{0}$ and so $\sigma_{y\cdot\mathbf{x}}^2 = 0$. Occasionally it might be worthwhile to make the following convention:

$$\varrho_{y\cdot\mathbf{x}} = 1 \quad\text{if}\quad \sigma_y^2 = 0\,. \tag{9.20}$$

Suppose that $\binom{\mathbf{x}}{\mathbf{y}} \sim \mathrm{N}_{p+q}(\boldsymbol{\mu},\boldsymbol{\Sigma})$, and denote $\mathbf{v} = \mathbf{y} - \boldsymbol{\Sigma}_{\mathbf{yx}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-}\mathbf{x}$ and so

$$\mathbf{y} = \mathbf{v} + \boldsymbol{\Sigma}_{\mathbf{yx}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-}\mathbf{x} = (\mathbf{y} - \boldsymbol{\Sigma}_{\mathbf{yx}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-}\mathbf{x}) + \boldsymbol{\Sigma}_{\mathbf{yx}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-}\mathbf{x}\,. \tag{9.21}$$

Now $\operatorname{cov}(\mathbf{v},\mathbf{x}) = \mathbf{0}$ implies, under multinormality, that $\mathbf{v}$ and $\mathbf{x}$ are (statistically) independent, and thereby the conditional distribution of $\mathbf{y}$ given that

$\mathbf{x}$ is held fixed at a selected value $\mathbf{x} = \underline{\mathbf{x}}$ is normal with

$$\mathrm{E}(\mathbf{y} \mid \mathbf{x} = \underline{\mathbf{x}}) = \mathrm{E}(\mathbf{v}) + \boldsymbol{\Sigma}_{\mathbf{yx}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-}\underline{\mathbf{x}} = \boldsymbol{\mu}_{\mathbf{y}} + \boldsymbol{\Sigma}_{\mathbf{yx}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-}(\underline{\mathbf{x}} - \boldsymbol{\mu}_{\mathbf{x}}) \,, \tag{9.22a}$$

$$\mathrm{cov}(\mathbf{y} \mid \mathbf{x} = \underline{\mathbf{x}}) = \mathrm{cov}(\mathbf{v}) = \boldsymbol{\Sigma}_{\mathbf{yy}} - \boldsymbol{\Sigma}_{\mathbf{yx}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-}\boldsymbol{\Sigma}_{\mathbf{xy}} = \boldsymbol{\Sigma}_{\mathbf{yy}\cdot\mathbf{x}} \,; \tag{9.22b}$$

see, e.g., Marsaglia (1964) and Muirhead (1982, p. 12).

9.1 Best Linear Predictor

Let $f(\mathbf{x})$ be a scalar-valued function of the random vector $\mathbf{x}$. The *mean squared error* of $f(\mathbf{x})$ with respect to y, y being a random variable or a fixed constant, is

$$\mathrm{MSE}[f(\mathbf{x}); y] = \mathrm{E}[y - f(\mathbf{x})]^2. \tag{9.23}$$

Correspondingly, if $\mathbf{y}$ is a vector (random or fixed) and $f(\mathbf{x})$ is a vector-valued random vector (function of random vector $\mathbf{x}$), the mean squared error matrix of $f(\mathbf{x})$ with respect to $\mathbf{y}$ is

$$\mathrm{MSEM}[f(\mathbf{x}); \mathbf{y}] = \mathrm{E}[\mathbf{y} - f(\mathbf{x})][\mathbf{y} - f(\mathbf{x})]'. \tag{9.24}$$

Let the random vector $\mathbf{z}$ (with $p + 1$ elements) be partitioned so that

$$\mathbf{z} = \begin{pmatrix} \mathbf{x} \\ y \end{pmatrix}, \quad \mathrm{E}(\mathbf{z}) = \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_{\mathbf{x}} \\ \mu_y \end{pmatrix}, \quad \mathrm{cov}(\mathbf{z}) = \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{xx}} & \boldsymbol{\sigma}_{\mathbf{x}y} \\ \boldsymbol{\sigma}'_{\mathbf{x}y} & \sigma_y^2 \end{pmatrix}. \tag{9.25}$$

We might be interested in predicting the random variable y on the basis of some function of the random vector $\mathbf{x}$; denote this function as $f(\mathbf{x})$. Then we can call $f(\mathbf{x})$ a predictor of y on the basis of $\mathbf{x}$. Naturally there are advantages to choose such a predictor $f(\mathbf{x})$ which minimizes the mean squared error $\mathrm{MSE}[f(\mathbf{x}); y] = \mathrm{E}[y - f(\mathbf{x})]^2$. Such a predictor is called the *best predictor* and we may denote it as $\mathrm{BP}(y; \mathbf{x})$. Then the $\mathrm{BP}(y; \mathbf{x})$ has the property

$$\min_{f(\mathbf{x})} \mathrm{E}[y - f(\mathbf{x})]^2 = \mathrm{E}[y - \mathrm{BP}(y; \mathbf{x})]^2. \tag{9.26}$$

It can be shown that

♠ the conditional expectation $\mathrm{E}(y \mid \mathbf{x})$ is the best predictor of y,

i.e.,

$$\mathrm{E}(y \mid \mathbf{x}) = \mathrm{BP}(y; \mathbf{x}) \,. \tag{9.27}$$

Here we have to consider $\mathrm{E}(y \mid \mathbf{x})$ as a random variable, not a real number. The conditional expectation $\mathrm{E}(y \mid \mathbf{x} = \underline{\mathbf{x}})$ is a function of $\underline{\mathbf{x}}$, and we may denote it as $\mathrm{E}(y \mid \mathbf{x} = \underline{\mathbf{x}}) := m(\underline{\mathbf{x}})$. Then replacing $\underline{\mathbf{x}}$ with $\mathbf{x}$, $m(\underline{\mathbf{x}})$ becomes a random variable $m(\mathbf{x})$, which is denoted as $\mathrm{E}(y \mid \mathbf{x})$. The proof of the important result (9.27) is left as an exercise; see, e.g., Christensen (2002, p. 132), Rao (1973a, p. 264), and Searle, Casella & McCulloch (1992, §7.2).

Consider next the *linear* (inhomogeneous) predictors:

$$\{\mathrm{LP}(y;\mathbf{x})\} = \{\, f(\mathbf{x}) : f(\mathbf{x}) = \mathbf{a}'\mathbf{x} + b,\ \mathbf{a} \in \mathbb{R}^p,\ b \in \mathbb{R}\,\}\,. \tag{9.28}$$

Then the *best linear predictor* (BLP) of y on the basis of $\mathbf{x}$ is defined as

$$\mathrm{BLP}(y;\mathbf{x}) = \mathbf{a}_*'\mathbf{x} + b_*\,, \tag{9.29}$$

if it minimizes the mean squared error:

$$\begin{aligned} \min_{\mathbf{a},b} \mathrm{MSE}(\mathbf{a}'\mathbf{x} + b; y) &= \min_{\mathbf{a},b} \mathrm{E}[y - (\mathbf{a}'\mathbf{x} + b)]^2 \\ &= \mathrm{E}[y - (\mathbf{a}_*'\mathbf{x} + b_*)]^2. \end{aligned} \tag{9.30}$$

It is easy to see that

$$\begin{aligned} \mathrm{E}[y - (\mathbf{a}'\mathbf{x} + b)]^2 &= \mathrm{E}[y - \mathbf{a}'\mathbf{x} - (\mu_y - \mathbf{a}'\boldsymbol{\mu}_{\mathbf{x}}) + (\mu_y - \mathbf{a}'\boldsymbol{\mu}_{\mathbf{x}} - b)]^2 \\ &= \mathrm{var}(y - \mathbf{a}'\mathbf{x}) + (\mu_y - \mathbf{a}'\boldsymbol{\mu}_{\mathbf{x}} - b)^2 \\ &= \text{“variance”} + \text{“bias}^2\text{”}\,. \end{aligned} \tag{9.31}$$

The second term above, i.e., the “bias^2”, becomes zero when b is chosen as

$$b_* = \mu_y - \mathbf{a}'\boldsymbol{\mu}_{\mathbf{x}}\,. \tag{9.32}$$

According to (9.16) (p. 193), the first term in (9.31), i.e., $\mathrm{var}(y - \mathbf{a}'\mathbf{x})$, attains its minimum if $\mathbf{a}$ is

$$\mathbf{a}_* = \boldsymbol{\Sigma}_{\mathbf{xx}}^{-}\boldsymbol{\sigma}_{\mathbf{x}y}\,, \tag{9.33}$$

and hence the formula for $\mathrm{BLP}(y;\mathbf{x})$ is

$$\begin{aligned} \mathrm{BLP}(y;\mathbf{x}) &= (\mu_y - \boldsymbol{\sigma}_{\mathbf{x}y}'\boldsymbol{\Sigma}_{\mathbf{xx}}^{-}\boldsymbol{\mu}_{\mathbf{x}}) + \boldsymbol{\sigma}_{\mathbf{x}y}'\boldsymbol{\Sigma}_{\mathbf{xx}}^{-}\mathbf{x} \\ &= \mu_y + \boldsymbol{\sigma}_{\mathbf{x}y}'\boldsymbol{\Sigma}_{\mathbf{xx}}^{-}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})\,, \end{aligned} \tag{9.34}$$

see Anderson (2003, p. 37), and Christensen (2002, Th. 6.3.2). In particular, if the conditional expectation $\mathrm{E}(y \mid \mathbf{x})$ is of the form

$$\mathrm{E}(y \mid \mathbf{x}) = \alpha + \boldsymbol{\beta}'\mathbf{x}\,, \tag{9.35}$$

then necessarily $\mathrm{BLP}(y;\mathbf{x}) = \mathrm{E}(y \mid \mathbf{x})$ and

$$\alpha = \mu_y - \boldsymbol{\beta}'\boldsymbol{\mu}_{\mathbf{x}}\,, \quad \text{and} \quad \boldsymbol{\beta} = \boldsymbol{\Sigma}_{\mathbf{xx}}^{-}\boldsymbol{\sigma}_{\mathbf{x}y}\,. \tag{9.36}$$

Corresponding to (0.168) (p. 36), we can state the following:

> ♠ whenever possible, the BLP “goes through” the conditional means. (9.37)

The *prediction error* is

$$e_{y\cdot\mathbf{x}} = y - \mathrm{BLP}(y;\mathbf{x}) = y - \mu_y - \boldsymbol{\sigma}'_{\mathbf{x}y}\boldsymbol{\Sigma}^{-}_{\mathbf{xx}}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})\,, \tag{9.38}$$

while the variance of the prediction error $e_{y\cdot\mathbf{x}}$ becomes

$$\begin{aligned}\mathrm{var}(e_{y\cdot\mathbf{x}}) &= \sigma_y^2 - \boldsymbol{\sigma}'_{\mathbf{x}y}\boldsymbol{\Sigma}^{-}_{\mathbf{xx}}\boldsymbol{\sigma}_{\mathbf{x}y} \\ &= \sigma_y^2\Big(1 - \frac{\boldsymbol{\sigma}'_{\mathbf{x}y}\boldsymbol{\Sigma}^{-}_{\mathbf{xx}}\boldsymbol{\sigma}_{\mathbf{x}y}}{\sigma_y^2}\Big) = \sigma_y^2(1 - \varrho^2_{y\cdot\mathbf{x}})\,,\end{aligned} \tag{9.39}$$

where $\varrho_{y\cdot\mathbf{x}}$ is the population multiple correlation coefficient:

$$\varrho^2_{y\cdot\mathbf{x}} = \mathrm{cor}^2[y, \mathrm{BLP}(y;\mathbf{x})] = \mathrm{cor}^2(y, \boldsymbol{\sigma}'_{\mathbf{x}y}\boldsymbol{\Sigma}^{-}_{\mathbf{xx}}\mathbf{x})\,. \tag{9.40}$$

Notice again that if $\sigma_y^2 = 0$, then $\mathrm{var}(e_{y\cdot\mathbf{x}}) = 0$ but the multiple correlation coefficient $\varrho_{y\cdot\mathbf{x}}$ is not well-defined.

Consider a more general case when

$$\mathbf{z} = \begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix}\boldsymbol{\mu}_{\mathbf{x}}\\ \boldsymbol{\mu}_{\mathbf{y}}\end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix}\boldsymbol{\Sigma}_{\mathbf{xx}} & \boldsymbol{\Sigma}_{\mathbf{xy}}\\ \boldsymbol{\Sigma}_{\mathbf{yx}} & \boldsymbol{\Sigma}_{\mathbf{yy}}\end{pmatrix}. \tag{9.41}$$

Then the linear predictor $\mathbf{Gx} + \mathbf{g}$ is the best linear predictor, BLP, for $\mathbf{y}$ on the basis of $\mathbf{x}$, if the Löwner ordering

$$\mathrm{MSEM}(\mathbf{Gx} + \mathbf{g}; \mathbf{y}) \leq_{\mathsf{L}} \mathrm{MSEM}(\mathbf{Fx} + \mathbf{f}; \mathbf{y}) \tag{9.42}$$

holds for every linear predictor $\mathbf{Fx} + \mathbf{f}$ of $\mathbf{y}$; above

$$\begin{aligned}\mathrm{MSEM}(\mathbf{Gx} + \mathbf{g}; \mathbf{y}) &= \mathrm{E}[\mathbf{y} - (\mathbf{Gx} + \mathbf{g})][\mathbf{y} - (\mathbf{Gx} + \mathbf{g})]' \\ &= \text{ mean squared error matrix.}\end{aligned} \tag{9.43}$$

In view of

$$\begin{aligned}&\mathrm{E}[\mathbf{y} - (\mathbf{Ax} + \mathbf{b})][\mathbf{y} - (\mathbf{Ax} + \mathbf{b})]' \\ &\qquad = \mathrm{cov}(\mathbf{y} - \mathbf{Ax}) + (\boldsymbol{\mu}_{\mathbf{y}} - \mathbf{A}\boldsymbol{\mu}_{\mathbf{x}} - \mathbf{b})(\boldsymbol{\mu}_{\mathbf{y}} - \mathbf{A}\boldsymbol{\mu}_{\mathbf{x}} - \mathbf{b})',\end{aligned} \tag{9.44}$$

it is easy to confirm that Theorem 9 (p. 191) gives

$$\begin{aligned}\mathrm{BLP}(\mathbf{y};\mathbf{x}) &= \boldsymbol{\mu}_{\mathbf{y}} - \boldsymbol{\Sigma}_{\mathbf{yx}}\boldsymbol{\Sigma}^{-}_{\mathbf{xx}}\boldsymbol{\mu}_{\mathbf{x}} + \boldsymbol{\Sigma}_{\mathbf{yx}}\boldsymbol{\Sigma}^{-}_{\mathbf{xx}}\mathbf{x} \\ &= \boldsymbol{\mu}_{\mathbf{y}} + \boldsymbol{\Sigma}_{\mathbf{yx}}\boldsymbol{\Sigma}^{-}_{\mathbf{xx}}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})\,.\end{aligned} \tag{9.45}$$

The difference $\mathbf{y} - \mathrm{BLP}(\mathbf{y};\mathbf{x})$ is the prediction error:

$$\mathbf{e}_{\mathbf{y}\cdot\mathbf{x}} = \mathbf{y} - \mathrm{BLP}(\mathbf{y};\mathbf{x}) = \mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}} - \boldsymbol{\Sigma}_{\mathbf{yx}}\boldsymbol{\Sigma}^{-}_{\mathbf{xx}}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})\,. \tag{9.46}$$

The random vector $\mathrm{BLP}(\mathbf{y};\mathbf{x})$ is also called the *regression of* $\mathbf{y}$ *on* $\mathbf{x}$ and $\mathbf{e}_{\mathbf{y}\cdot\mathbf{x}}$ the vector of *residuals of* $\mathbf{y}$ *from its regression on* $\mathbf{x}$; worth repeating:

$$\mathrm{BLP}(\mathbf{y};\mathbf{x}) = \text{the regression of } \mathbf{y} \text{ on } \mathbf{x}\,, \tag{9.47a}$$

$$\mathbf{e}_{\mathbf{y}\cdot\mathbf{x}} = \text{residuals of } \mathbf{y} \text{ from its regression on } \mathbf{x}\,. \tag{9.47b}$$

The matrix $\mathrm{cov}(\mathbf{e}_{\mathbf{y}\cdot\mathbf{x}}) = \mathbf{\Sigma}_{\mathbf{yy}\cdot\mathbf{x}}$ is the matrix of partial covariances of $\mathbf{y}$ (eliminating, so to say, the effect of $\mathbf{x}$). The ij-element of the matrix of *partial correlations* of $\mathbf{y}$ (eliminating $\mathbf{x}$) is

$$\varrho_{ij\cdot\mathbf{x}} = \frac{\sigma_{ij\cdot\mathbf{x}}}{\sqrt{\sigma_{ii\cdot\mathbf{x}}\sigma_{jj\cdot\mathbf{x}}}} = \mathrm{cor}(e_{y_i\cdot\mathbf{x}}, e_{y_j\cdot\mathbf{x}})\,, \tag{9.48}$$

where

$$e_{y_i\cdot\mathbf{x}} = y_i - \mu_{y_i} - \boldsymbol{\sigma}'_{\mathbf{x}y_i}\mathbf{\Sigma}^{-}_{\mathbf{xx}}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})\,, \quad \boldsymbol{\sigma}_{\mathbf{x}y_i} = \mathrm{cov}(\mathbf{x}, y_i)\,. \tag{9.49}$$

As mentioned in (9.21)-(9.22) (p. 193), one fundamental fact in multivariate analysis is that under multinormality, i.e., if $\mathbf{z} \sim \mathrm{N}_{p+q}(\boldsymbol{\mu}, \mathbf{\Sigma})$, where $\mathbf{z}$ is partitioned as in (9.41), the *conditional distribution* of $\mathbf{y}$ given that $\mathbf{x}$ is held fixed at a selected value $\mathbf{x} = \underline{\mathbf{x}}$ is normal with the mean

$$\mathrm{E}(\mathbf{y} \mid \mathbf{x} = \underline{\mathbf{x}}) = \boldsymbol{\mu}_{\mathbf{y}} + \mathbf{\Sigma}_{\mathbf{yx}}\mathbf{\Sigma}^{-1}_{\mathbf{xx}}(\underline{\mathbf{x}} - \boldsymbol{\mu}_{\mathbf{x}})\,, \tag{9.50}$$

and the covariance matrix (partial covariances)

$$\mathrm{cov}(\mathbf{y} \mid \mathbf{x} = \underline{\mathbf{x}}) = \mathbf{\Sigma}_{\mathbf{yy}} - \mathbf{\Sigma}_{\mathbf{yx}}\mathbf{\Sigma}^{-1}_{\mathbf{xx}}\mathbf{\Sigma}_{\mathbf{xy}} = \mathbf{\Sigma}_{\mathbf{yy}\cdot\mathbf{x}}\,. \tag{9.51}$$

Above (and what follows) we assume, for simplicity, that $\mathbf{\Sigma}$ is positive definite. Elements of $\mathbf{\Sigma}_{\mathbf{yy}\cdot\mathbf{x}}$ are, under multinormality, conditional variances and covariances, and the partial correlations are conditional correlations. In particular,

$$\mathrm{E}(y \mid \mathbf{x} = \underline{\mathbf{x}}) = \mu_y + \boldsymbol{\sigma}'_{\mathbf{x}y}\mathbf{\Sigma}^{-1}_{\mathbf{xx}}(\underline{\mathbf{x}} - \boldsymbol{\mu}_{\mathbf{x}})\,, \tag{9.52a}$$

and

$$\begin{aligned}\mathrm{var}(y \mid \mathbf{x} = \underline{\mathbf{x}}) &= \sigma^2_{y\cdot 12\ldots p} = \sigma^2_{y\cdot\mathbf{x}} = \sigma^2_y - \boldsymbol{\sigma}'_{\mathbf{x}y}\mathbf{\Sigma}^{-1}_{\mathbf{xx}}\boldsymbol{\sigma}_{\mathbf{x}y} \\ &= \sigma^2_y\Big(1 - \frac{\boldsymbol{\sigma}'_{\mathbf{x}y}\mathbf{\Sigma}^{-1}_{\mathbf{xx}}\boldsymbol{\sigma}_{\mathbf{x}y}}{\sigma^2_y}\Big) = \sigma^2_y(1 - \varrho^2_{y\cdot\mathbf{x}})\,. \end{aligned} \tag{9.52b}$$

When $n = 2$, and $\mathrm{cor}(x, y) = \varrho_{xy}$, we have

$$\begin{aligned}\mathrm{E}(y \mid x = \underline{x}) &= \mu_y + \frac{\sigma_{xy}}{\sigma^2_x}(\underline{x} - \mu_x) = \mu_y + \varrho_{xy}\frac{\sigma_y}{\sigma_x}(\underline{x} - \mu_x) \\ &= \Big(\mu_y - \frac{\sigma_{xy}}{\sigma^2_x}\mu_x\Big) + \frac{\sigma_{xy}}{\sigma^2_x}\,\underline{x}\,, \end{aligned} \tag{9.53a}$$

$$\mathrm{var}(y \mid x = \underline{x}) = \sigma^2_{y\cdot x} = \sigma^2_y(1 - \varrho^2_{xy}) \le \sigma^2_y = \mathrm{var}(y)\,. \tag{9.53b}$$

As was done earlier, we can consider the conditional expectation $\mathrm{E}(\mathbf{y} \mid \underline{\mathbf{x}})$ as a random vector by replacing $\underline{\mathbf{x}}$ by $\mathbf{x}$. Then, under multinormality,

$$\mathrm{E}(\mathbf{y} \mid \mathbf{x}) = \boldsymbol{\mu}_{\mathbf{y}} + \boldsymbol{\Sigma}_{\mathbf{yx}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}}) = \mathrm{BLP}(\mathbf{y};\mathbf{x})\,. \tag{9.54}$$

In view of (9.27) (p. 194), $\boldsymbol{\mu}_{\mathbf{y}} + \boldsymbol{\Sigma}_{\mathbf{yx}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})$ is (under multinormality) not only the best linear predictor but also the best predictor of $\mathbf{y}$.

It is worth emphasizing that if $\mathbf{z}$ is not multinormally distributed, then the matrix of the partial covariances $\boldsymbol{\Sigma}_{\mathbf{yy}\cdot\mathbf{x}}$ is not necessarily the covariance matrix of the conditional distribution, and correspondingly, the partial correlations are not necessarily conditional correlations.

We collect some properties of BLP in the following proposition.

Proposition 9.1. *Let* $\mathbf{Gx} + \mathbf{g}$ *be the best linear predictor for* $\mathbf{y}$ (*on the basis of* $\mathbf{x}$), *denoted by* $\mathrm{BLP}(\mathbf{y};\mathbf{x})$, *let* $\mathbf{e}_{\mathbf{y}\cdot\mathbf{x}}$ *be the corresponding residual* (*prediction error*). *Then*

(a) $\mathrm{BLP}(\mathbf{y};\mathbf{x}) = \boldsymbol{\mu}_{\mathbf{y}} + \boldsymbol{\Sigma}_{\mathbf{yx}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})$,

(b) $\mathrm{E}[\mathrm{BLP}(\mathbf{y};\mathbf{x})] = \boldsymbol{\mu}_{\mathbf{y}}$,

(c) $\mathbf{e}_{\mathbf{y}\cdot\mathbf{x}} = \mathbf{y} - \mathrm{BLP}(\mathbf{y};\mathbf{x}) = \mathbf{y} - [\boldsymbol{\mu}_{\mathbf{y}} + \boldsymbol{\Sigma}_{\mathbf{yx}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})]$,

(d) $\mathrm{cov}(\mathbf{e}_{\mathbf{y}\cdot\mathbf{x}}) = \boldsymbol{\Sigma}_{\mathbf{yy}} - \boldsymbol{\Sigma}_{\mathbf{yx}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-}\boldsymbol{\Sigma}_{\mathbf{xy}} = \mathrm{MSEM}[\mathrm{BLP}(\mathbf{y};\mathbf{x});\mathbf{y}]$,

(e) $\mathrm{cov}(\mathbf{e}_{\mathbf{y}\cdot\mathbf{x}}, \mathbf{x}) = \mathbf{0}$.

The random vector $\mathrm{BLP}(\mathbf{y};\mathbf{x}) = \boldsymbol{\mu}_{y} - \boldsymbol{\Sigma}_{\mathbf{yx}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})$ is sometimes called the linear expectation and denoted also as $\hat{\mathrm{E}}(\mathbf{y};\mathbf{x})$. In general,

- $\mathrm{BLP}(\mathbf{y};\mathbf{x})$ is neither the conditional expectation of $\mathbf{y}$ given $\mathbf{x}$
- nor an estimate of the conditional expectation.

9.2 What Is Regression?

Photograph 9.1 Bernard Flury (Auckland, 1995).

In (9.47) (p. 197) we defined what we mean by regression in the context of random variables: the regression of y on x is the $\mathrm{BLP}(y;x)$, the best linear predictor of y on the basis of x, which is not necessarily the same as the conditional mean $\mathrm{E}(y \mid x)$. For example Anderson (2003, p. 39), states that even when we have no multinormality—in which case $\mathrm{BLP}(y;x) = \mathrm{E}(y \mid x)$—we can define regression by $\mathrm{BLP}(y;x)$.

Flury (1997, p. 90) calls the conditional mean $\mathrm{E}(y \mid x = \underline{x})$ or its random version $\mathrm{E}(y \mid x)$ the regression of y on x. Notice that this definition does not require that $\mathrm{E}(y \mid x)$ is a linear function of x. Flury continues:

> ♠ "The student who has taken a regression course before may be somewhat surprised by this definition, because regression analysis is often viewed as a collection of techniques for fitting equations to observed data. Yet this definition expresses perfectly the central goal of regression analysis, namely, to estimate the (conditional) mean of a variable, knowing the values of one or several other variables."

Cook & Weisberg (1999, p. 37) state that

> ♠ "Regression is the study of how the conditional distribution of $y \mid x$ changes with the value of x."

The first sentence in Weisberg (2005) is:

> ♠ "Regression analysis answers questions about the dependence of a response variable on one or more predictors, including prediction of future values of a response, discovering which predictors are important, and estimating the impact of changing a predictor or a treatment on the value of the response."

The Preface of Ryan (2009) begins:

> ♠ "Contrary to the opinions that some people have expressed in the literature, regression analysis is a continuously evolving field. It is not already 'fully developed'."

Photograph 9.2 R. Dennis Cook (Tampere, 1990).

Graphics is of course an extremely important tool in regression-related matters, and every regression user should take a look at the classic scatter plots of Anscombe (1973); see Figure 9.1 (p. 200). [Cf. also papers by Chatterjee & Firat (2007) and Haslett & Govindaraju (2009).]

In the Preface of his book on *Regression Graphics*, Cook (1998) writes: "The original motivation came from wondering how far computer craphics could be pushed in a regression analysis. In the extreme, is it possible to conduct a regression analysis by using just graphics?" For related references to this rapidly developing new area (including concepts like "sufficient dimension reduction"), we may mention Li (1991), Cook (1998, 2007), Cook, Li & Chiaramonte (2007), and Cook & Forzani (2008).

The term *regression* itself has an exceptionally interesting history: see the excellent chapter entitled *Regression towards Mean* in Stigler (1999), where (on p. 177) he says that the story of Francis Galton's (1822–1911) discovery of regression is "an exciting one, involving science, experiment, mathematics,

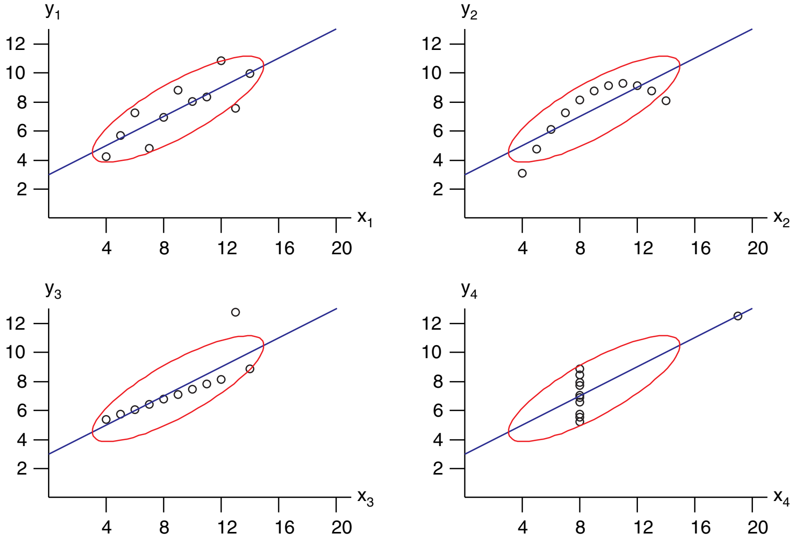


Figure 9.1 The four data sets of Anscombe (1973). Each pair yields identical numerical regression output. Lesson: plot, plot, plot!

simulation, and one of the great thought experiments of all time"[1]; see also Stigler (1986, Ch. 8).

Galton investigated the relationships between the heights of fathers and heights of sons. He found that tall fathers tend to have tall sons in general but he also found that very tall fathers tend to have sons who are in average shorter (in terms of standard deviations) than their fathers. This can be concluded well from a bivariate normal distribution:

$$\frac{\mathrm{E}(y \mid x = a) - \mu_y}{\sigma_y} = \varrho_{xy} \frac{a - \mu_x}{\sigma_x} \,. \tag{9.55}$$

Putting $\sigma_x = \sigma_y$ we have

$$\mathrm{E}(y \mid x = a) - \mu_y = \varrho_{xy}(a - \mu_x) \,. \tag{9.56}$$

In this father-son-height case $0 < \varrho_{xy} < 1$ and hence the son whose father has $x = a$ is expected to be closer to the average (of the sons) than his father is to the average of the fathers. Figures 9.2a and 9.2b, illustrating this situation,

---

[1] While telling about the regression fallacy, Stigler (1999, p. 169) refers to Harold Hotelling's (1934) criticism on some particular misunderstanding: "To 'prove' such a mathematical result by a costly and prolonged numerical study of many kinds of business profit and expense ratios is analogous to proving the multiplication table by arranging elephants in rows and columns, and then doing the same for numerous other kinds of animals. The performance, though perhaps entertaining, and having certain pedagogical value, is not important contribution to either zoölogy or to mathematics."

are based on classical data of Pearson & Lee (1903) and Galton (1886). See also a related papers by Wachsmuth, Wilkinson & Dallal (2003), and Smith (1997).

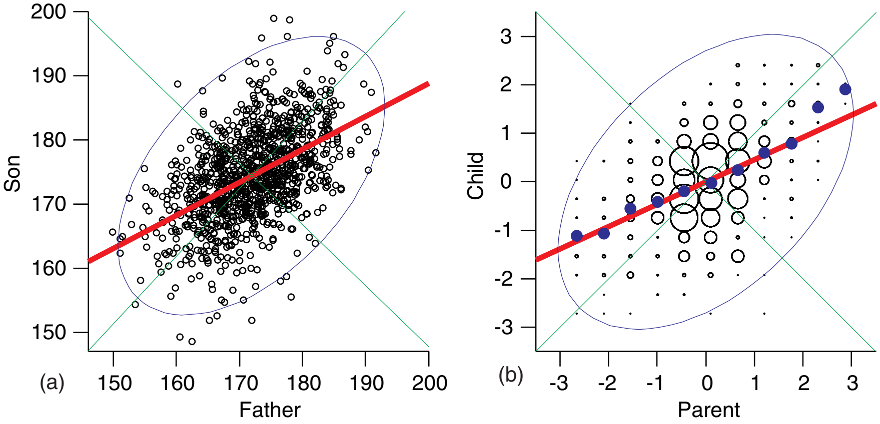


Figure 9.2 (a) Data of Pearson & Lee (1903): heights of fathers and sons, in centimeters; $n = 1078$. (b) Data of Galton (1886): height of "midparent" vs. height of adult child, $n = 928$. Measurements are centered and scaled so that variances are 1. The size of a circle is proportional to the number of observations in the corresponding cell of Galton's cross-tabulation. The conditional means are marked with filled circles.

9.3 Throwing Two Dice

Let us consider an extremely simple but illustrative example, see Flury (1997, Example 2.2.3). Let us roll (independently) two dice and denote the numbers obtained as random variables x_1 and x_2 and let y denote their sum:

$$y = x_1 + x_2 \,. \tag{9.57}$$

Let us consider the joint distribution of x_1 and y—see the Figure 9.3. Now we can trivially conclude that

$$\mathrm{E}(y \mid x_1 = \underline{x}_1) = 3.5 + \underline{x}_1, \quad \underline{x}_1 = 1, \dots, 6 \,, \tag{9.58a}$$

$$\mathrm{var}(y \mid x_1 = \underline{x}_1) = \mathrm{var}(x_2) = \sigma_2^2 \,. \tag{9.58b}$$

On the other hand

$$\mathrm{E}(x_i) = \mu_i = 3.5 \,, \quad \mathrm{var}(x_i) = \sigma_i^2 = \sigma^2 = 35/12, \quad i = 1, 2 \,, \tag{9.59a}$$

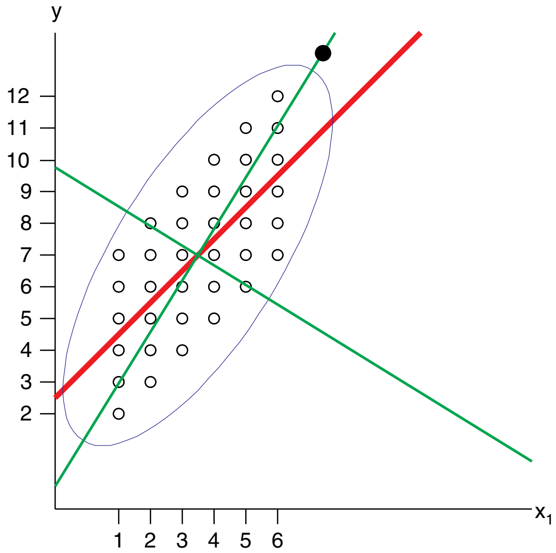


Figure 9.3 Throwing two dice. The random variable y is a sum of two dice, x_1 and x_2. The picture shows the joint distribution of x_1 and y. The thick line's equation is $y = 3.5 + x_1$, which is the BLP of y on the basis of x_1 when considering x_1 and y as random variables. The same line forms also the conditional expectation of y given that x_1 has a selected value $x_1 = \underline{x}_1$. The contour drawn is for the bivariate normal distribution $\mathrm{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are calculated from the joint distribution of x_1 and y. The big dot in the figure is the point $\boldsymbol{\mu} + a\mathbf{t}_1$, where $\mathbf{t}_1$ is the first eigenvector of $\boldsymbol{\Sigma}$ and a is an appropriate scalar.

$$\mathrm{E}(y) = \mu_y = 7\,, \quad \mathrm{var}(y) = \sigma_y^2 = \sigma^2 + \sigma^2 = 2\sigma^2, \tag{9.59b}$$

$$\mathrm{cov}(x_1, y) = \mathrm{cov}(x_1, x_1 + x_2) = \mathrm{var}(x_1) = \sigma^2, \tag{9.59c}$$

$$\mathrm{cor}^2(x_1, y) = \frac{[\mathrm{var}(x_1)]^2}{\mathrm{var}(y) \cdot \mathrm{var}(x_1)} = \frac{\mathrm{var}(x_1)}{\mathrm{var}(y)} = \frac{\sigma^2}{2\sigma^2} = \frac{1}{2}\,, \tag{9.59d}$$

$$\mathrm{E}\begin{pmatrix} x_1 \\ y \end{pmatrix} = \begin{pmatrix} 3.5 \\ 7 \end{pmatrix} = \boldsymbol{\mu}\,, \quad \mathrm{cov}\begin{pmatrix} x_1 \\ y \end{pmatrix} = \sigma^2 \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} = \boldsymbol{\Sigma}\,. \tag{9.59e}$$

Hence we get

$$\mathrm{BLP}(y; x_1) = \left(\mu_y - \mu_1 \frac{\mathrm{cov}(x_1, y)}{\mathrm{var}(x_1)}\right) + \frac{\mathrm{cov}(x_1, y)}{\mathrm{var}(x_1)}\, x_1 = 3.5 + x_1\,, \tag{9.60}$$

and so in this situation, even though the joint distribution is far away from the normal distribution, we have the equality $\mathrm{E}(y \mid x_1) = \mathrm{BLP}(y; x_1)$.

Figure 9.3 illustrates the situation. The thin lines are the principal axis of the contour ellipse for the bivariate normal distribution $\mathrm{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. They have the same directions as the eigenvectors of $\boldsymbol{\Sigma}$.

It is easy to conclude that if $y = x_1 + \cdots + x_p$, where x_i's refer to p independent dice then

$$\mathrm{E}(y \mid x_1) = 3.5(p-1) + x_1\,, \quad \mathrm{var}(y \mid x_1) = \sigma^2(p-1)\,, \tag{9.61a}$$

$$\mathrm{cor}^2(x_1, y) = \frac{\mathrm{var}(x_1)}{\mathrm{var}(y)} = \frac{\sigma^2}{p\sigma^2} = \frac{1}{p}\,. \tag{9.61b}$$

9.4 Predictive Approach to Principal Component Analysis

Next we consider somewhat different prediction problem: we are interested in predicting the d-dimensional random vector $\mathbf{z}$ on the basis of some linear combinations of $\mathbf{z}$, that is, we want to find

$$\mathrm{BLP}(\mathbf{z}; \mathbf{A}'\mathbf{z}) = \text{best linear predictor of } \mathbf{z} \text{ on the basis of } \mathbf{A}'\mathbf{z}\,, \tag{9.62}$$

where $\mathbf{A}' \in \mathbb{R}^{k \times d}$ is a given matrix. This task makes particular sense if we are able to find only few linear combinations of $\mathbf{z}$, which give a "reasonable" prediction for $\mathbf{z}$; this means that the number of the columns of $\mathbf{A}$, k, would be much smaller than d, the dimension of $\mathbf{z}$. The efficiency of prediction may be measured by the covariance matrix of the prediction error which we wish to be "small" in some appropriate sense. This technique can be seen as a predictive approach to the principal components; see, e.g., Rao (1973a, pp. 592–593), Rao (1964, §7) Christensen (2001, §3.2.1), and Jolliffe (2002, Ch. 2).

Let us denote

$$\mathrm{cov}\begin{pmatrix} \mathbf{A}'\mathbf{z} \\ \mathbf{z} \end{pmatrix} = \begin{pmatrix} \mathbf{A}'\mathbf{\Sigma}\mathbf{A} & \mathbf{A}'\mathbf{\Sigma} \\ \mathbf{\Sigma}\mathbf{A} & \mathbf{\Sigma} \end{pmatrix}. \tag{9.63}$$

Then the following result is easy to establish.

Proposition 9.2. *Consider the random vector* $\mathbf{z}$ *with* $\mathrm{E}(\mathbf{z}) = \boldsymbol{\mu}_{\mathbf{z}}$, $\mathrm{cov}(\mathbf{z}) = \mathbf{\Sigma}$, *and let the random vector* $\mathrm{BLP}(\mathbf{z}; \mathbf{A}'\mathbf{z})$ *denote the* BLP *of* $\mathbf{z}$ *based on* $\mathbf{A}'\mathbf{z}$. *Then*

$$\begin{aligned}
\mathrm{BLP}(\mathbf{z}; \mathbf{A}'\mathbf{z}) &= \boldsymbol{\mu}_{\mathbf{z}} + \mathrm{cov}(\mathbf{z}, \mathbf{A}'\mathbf{z})[\mathrm{cov}(\mathbf{A}'\mathbf{z})]^{-}[\mathbf{A}'\mathbf{z} - \mathrm{E}(\mathbf{A}'\mathbf{z})] \\
&= \boldsymbol{\mu}_{\mathbf{z}} + \mathbf{\Sigma}\mathbf{A}(\mathbf{A}'\mathbf{\Sigma}\mathbf{A})^{-}\mathbf{A}'(\mathbf{z} - \boldsymbol{\mu}_{\mathbf{z}}) \\
&= \boldsymbol{\mu}_{\mathbf{z}} + \mathbf{\Sigma}^{1/2}\mathbf{P}_{\mathbf{\Sigma}^{1/2}\mathbf{A}}(\mathbf{\Sigma}^{+})^{1/2}(\mathbf{z} - \boldsymbol{\mu}_{\mathbf{z}}) \\
&= \boldsymbol{\mu}_{\mathbf{z}} + (\mathbf{P}_{\mathbf{A};\mathbf{\Sigma}})'(\mathbf{z} - \boldsymbol{\mu}_{\mathbf{z}})\,,
\end{aligned} \tag{9.64}$$

where $(\mathbf{P}_{\mathbf{A};\mathbf{\Sigma}})' = \mathbf{\Sigma}\mathbf{A}(\mathbf{A}'\mathbf{\Sigma}\mathbf{A})^{-}\mathbf{A} = \mathbf{P}'_{\mathbf{A};\mathbf{\Sigma}}$, *and the covariance matrix of the prediction error* $\mathbf{e}_{\mathbf{z}\cdot\mathbf{A}'\mathbf{z}} = \mathbf{z} - \mathrm{BLP}(\mathbf{z}; \mathbf{A}'\mathbf{z})$ *is*

$$\begin{aligned}\operatorname{cov}[\mathbf{z} - \mathrm{BLP}(\mathbf{z}; \mathbf{A}'\mathbf{z})] &= \mathbf{\Sigma} - \mathbf{\Sigma}\mathbf{A}'(\mathbf{A}'\mathbf{\Sigma}\mathbf{A})^{-}\mathbf{A}'\mathbf{\Sigma} \\ &= \mathbf{\Sigma}(\mathbf{I}_d - \mathbf{P}_{\mathbf{A};\mathbf{\Sigma}}) = \mathbf{\Sigma}^{1/2}(\mathbf{I}_d - \mathbf{P}_{\mathbf{\Sigma}^{1/2}\mathbf{A}})\mathbf{\Sigma}^{1/2}. \quad (9.65)\end{aligned}$$

How to choose the matrix $\mathbf{A} \in \mathbb{R}^{d\times k}$? Proposing this question means that we have some relevant criterion in mind for constructing $\mathbf{A}$, and as said earlier, a natural measure of the goodness of the prediction is the “size” of the covariance matrix of the prediction error $\mathbf{e}_{\mathbf{z}\cdot\mathbf{A}'\mathbf{z}} = \mathbf{z} - \mathrm{BLP}(\mathbf{z}; \mathbf{A}'\mathbf{z})$. For simplicity we assume in what follows that $\operatorname{cov}(\mathbf{z}) = \mathbf{\Sigma}$ is positive definite.

Let us first consider the situation when $\mathbf{A}$ has only one column, $\mathbf{A} = \mathbf{a}$. Then the covariance matrix of the prediction error is

$$\operatorname{cov}[\mathbf{z} - \mathrm{BLP}(\mathbf{z}; \mathbf{a}'\mathbf{z})] = \mathbf{\Sigma} - \mathbf{\Sigma}\mathbf{a}(\mathbf{a}'\mathbf{\Sigma}\mathbf{a})^{-1}\mathbf{a}'\mathbf{\Sigma} := \mathbf{\Sigma}_{\mathbf{z}\cdot\mathbf{a}'\mathbf{z}}\,. \quad (9.66)$$

How should the vector $\mathbf{a}$ be chosen so that it would give the minimum, in some sense, for the (9.66)?

Minimizing the trace of $\mathbf{\Sigma}_{\mathbf{z}\cdot\mathbf{a}'\mathbf{z}}$, we have to maximize

$$\operatorname{tr}[\mathbf{\Sigma}\mathbf{a}(\mathbf{a}'\mathbf{\Sigma}\mathbf{a})^{-1}\mathbf{a}'\mathbf{\Sigma}] = \frac{\mathbf{a}'\mathbf{\Sigma}^2\mathbf{a}}{\mathbf{a}'\mathbf{\Sigma}\mathbf{a}}\,. \quad (9.67)$$

In view of Proposition 18.9 (p. 377),

$$\max_{\mathbf{a}\neq\mathbf{0}} \frac{\mathbf{a}'\mathbf{\Sigma}^2\mathbf{a}}{\mathbf{a}'\mathbf{\Sigma}\mathbf{a}} = \operatorname{ch}_1(\mathbf{\Sigma}^2\mathbf{\Sigma}^{-1}) = \operatorname{ch}_1(\mathbf{\Sigma}) = \lambda_1\,, \quad (9.68)$$

and the maximum is attained when $\mathbf{a}$ is chosen as $\mathbf{t}_1$, the eigenvector of $\mathbf{\Sigma}$ corresponding to the largest eigenvalue λ_1.

An alternative criterion to construct $\mathbf{a}$ is to minimize the Frobenius (Euclidean) norm of $\mathbf{\Sigma}_{\mathbf{z}\cdot\mathbf{a}'\mathbf{z}}$. This can be done using the Eckart–Young theorem, i.e., Proposition 19.5 (p. 401). For this purpose, let $\mathbf{\Sigma}$ have the eigenvalue decomposition

$$\mathbf{\Sigma} = \mathbf{T}\mathbf{\Lambda}\mathbf{T}' = \lambda_1\mathbf{t}_1\mathbf{t}_1' + \cdots + \lambda_d\mathbf{t}_d\mathbf{t}_d'\,. \quad (9.69)$$

Then

$$\min_{\operatorname{rk}(\mathbf{B})=k} \|\mathbf{\Sigma} - \mathbf{B}\|_F^2 = \min_{\operatorname{rk}(\mathbf{B})=k} \operatorname{tr}(\mathbf{\Sigma} - \mathbf{B})(\mathbf{\Sigma} - \mathbf{B})' = \lambda_{k+1}^2 + \cdots + \lambda_d^2\,, \quad (9.70)$$

and the minimum is attained when $\mathbf{B}$ is chosen as

$$\mathbf{B}_* = \mathbf{T}_{(k)}\mathbf{\Lambda}_{(k)}\mathbf{T}_{(k)}' = \lambda_1\mathbf{t}_1\mathbf{t}_1' + \cdots + \lambda_k\mathbf{t}_k\mathbf{t}_k'\,. \quad (9.71)$$

For $k = 1$ we have $\mathbf{B}_* = \lambda_1\mathbf{t}_1\mathbf{t}_1'$ and $\|\mathbf{\Sigma} - \mathbf{B}_*\|_F^2 = \operatorname{tr}(\mathbf{\Sigma}^2) - \lambda_1^2$. On the other hand, we obviously have

$$\min_{\mathbf{a}\neq\mathbf{0}} \|\mathbf{\Sigma} - \mathbf{\Sigma}\mathbf{a}(\mathbf{a}'\mathbf{\Sigma}\mathbf{a})^{-1}\mathbf{a}'\mathbf{\Sigma}\|_F \geq \min_{\operatorname{rk}(\mathbf{B})=1} \|\mathbf{\Sigma} - \mathbf{B}\|_F^2\,. \quad (9.72)$$

Putting $\mathbf{a} = \mathbf{t}_1$, we get

$$\boldsymbol{\Sigma}\mathbf{a}(\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a})^{-1}\mathbf{a}'\boldsymbol{\Sigma} = \frac{\lambda_1\mathbf{t}_1 \cdot \lambda_1\mathbf{t}_1'}{\lambda_1} = \lambda_1\mathbf{t}_1\mathbf{t}_1' \,. \tag{9.73}$$

Hence we can conclude that the following criteria yield the same results:

$$\min_{\mathbf{a}\neq\mathbf{0}} \operatorname{tr}[\boldsymbol{\Sigma} - \boldsymbol{\Sigma}\mathbf{a}(\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a})^{-1}\mathbf{a}'\boldsymbol{\Sigma}] \,, \qquad \min_{\mathbf{a}\neq\mathbf{0}} \|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}\mathbf{a}(\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a})^{-1}\mathbf{a}'\boldsymbol{\Sigma}\|_F \,. \tag{9.74}$$

The analogy appearing in (9.74) can be extended to a matrix $\mathbf{A}$ comprising k orthonormal columns. Minimizing the trace of $\boldsymbol{\Sigma}_{\mathbf{z}\cdot\mathbf{A}'\mathbf{z}}$, we have to maximize

$$\operatorname{tr}(\boldsymbol{\Sigma}^{1/2}\mathbf{P}_{\boldsymbol{\Sigma}^{1/2}\mathbf{A}}\boldsymbol{\Sigma}^{1/2}) = \operatorname{tr}(\mathbf{P}_{\boldsymbol{\Sigma}^{1/2}\mathbf{A}}\boldsymbol{\Sigma}) = \operatorname{tr}[\mathbf{A}'\boldsymbol{\Sigma}^2\mathbf{A}(\mathbf{A}'\boldsymbol{\Sigma}\mathbf{A})^{-1}] \,. \tag{9.75}$$

In view of Proposition 19.1 (p. 397),

$$\max_{\mathbf{L}'\mathbf{L}=\mathbf{I}_k} \operatorname{tr}(\mathbf{L}'\boldsymbol{\Sigma}\mathbf{L}) = \max_{\mathbf{L}'\mathbf{L}=\mathbf{I}_k} \operatorname{tr}(\mathbf{P}_{\mathbf{L}}\boldsymbol{\Sigma}) = \lambda_1 + \cdots + \lambda_k \,, \tag{9.76}$$

and thereby

$$\operatorname{tr}(\mathbf{P}_{\boldsymbol{\Sigma}^{1/2}\mathbf{A}}\boldsymbol{\Sigma}) \leq \lambda_1 + \cdots + \lambda_k \quad \text{for all } \mathbf{A} \in \mathbb{R}^{d\times k}. \tag{9.77}$$

Choosing $\mathbf{A} = \mathbf{T}_{(k)}$, we get

$$\begin{aligned} \operatorname{tr}(\mathbf{P}_{\boldsymbol{\Sigma}^{1/2}\mathbf{T}_{(k)}}\boldsymbol{\Sigma}) &= \operatorname{tr}[\mathbf{T}_{(k)}'\boldsymbol{\Sigma}^2\mathbf{T}_{(k)}(\mathbf{T}_{(k)}'\boldsymbol{\Sigma}\mathbf{T}_{(k)})^{-1}] \\ &= \lambda_1 + \cdots + \lambda_k \,, \end{aligned} \tag{9.78}$$

and hence the maximum of (9.75) is attained when $\mathbf{A}$ is chosen as $\mathbf{T}_{(k)}$. The corresponding result is obtained also when minimizing the Frobenius matrix norm $\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}\mathbf{A}(\mathbf{A}'\boldsymbol{\Sigma}\mathbf{A})^{-1}\mathbf{A}'\boldsymbol{\Sigma}\|_F$. For related considerations regarding the sample statistics, see Section 19.4 (p. 402).

Proposition 9.3. *Let* $\mathbf{A}$ *be a* $d \times k$ *matrix of orthonormal columns, and consider the best linear predictor of* $\mathbf{z}$ *on the basis of* $\mathbf{A}'\mathbf{z}$*. Then, the Euclidean norm of the covariance matrix of the prediction error* (*residual*)

$$\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}\mathbf{A}(\mathbf{A}'\boldsymbol{\Sigma}\mathbf{A})^{-1}\mathbf{A}'\boldsymbol{\Sigma}\|_F = \|\boldsymbol{\Sigma}^{1/2}(\mathbf{I}_d - \mathbf{P}_{\boldsymbol{\Sigma}^{1/2}\mathbf{A}})\boldsymbol{\Sigma}^{1/2}\|_F \,, \tag{9.79}$$

is a minimum when $\mathbf{A}$ *is chosen as* $\mathbf{T}_{(k)} = (\mathbf{t}_1 : \ldots : \mathbf{t}_k)$*, where* $\mathbf{t}_i$ *are the* k *first orthonormal eigenvectors of* $\boldsymbol{\Sigma}$*. Moreover, minimizing the trace of the covariance matrix of the prediction error, i.e., maximizing* $\operatorname{tr}(\mathbf{P}_{\boldsymbol{\Sigma}^{1/2}\mathbf{A}}\boldsymbol{\Sigma})$ *yields the same result.*

Consider a particular situation when $\mathbf{x}$ and $\mathbf{y}$ are p- and q-dimensional random vectors, and our task is to predict $\mathbf{z} = \binom{\mathbf{x}}{\mathbf{y}}$ on the basis of $\mathbf{x}$. Denote

$$\mathbf{z}=\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}, \quad \begin{pmatrix}\mathbf{x}\\ \mathbf{z}\end{pmatrix}=\begin{pmatrix}\mathbf{x}\\ \mathbf{x}\\ \mathbf{y}\end{pmatrix}=\begin{pmatrix}(\mathbf{I}_p : \mathbf{0})\mathbf{z}\\ \mathbf{z}\end{pmatrix}=\begin{pmatrix}\mathbf{A}'\mathbf{z}\\ \mathbf{z}\end{pmatrix}, \tag{9.80}$$

and

$$\operatorname{cov}\begin{pmatrix}\mathbf{A}'\mathbf{z}\\ \mathbf{z}\end{pmatrix}=\begin{pmatrix}\mathbf{A}'\boldsymbol{\Sigma}\mathbf{A} & \mathbf{A}'\boldsymbol{\Sigma}\\ \boldsymbol{\Sigma}\mathbf{A} & \boldsymbol{\Sigma}\end{pmatrix}=\begin{pmatrix}\boldsymbol{\Sigma}_{\mathbf{xx}} & \boldsymbol{\Sigma}_{\mathbf{xx}} & \boldsymbol{\Sigma}_{\mathbf{xy}}\\ \boldsymbol{\Sigma}_{\mathbf{xx}} & \boldsymbol{\Sigma}_{\mathbf{xx}} & \boldsymbol{\Sigma}_{\mathbf{xy}}\\ \boldsymbol{\Sigma}_{\mathbf{yx}} & \boldsymbol{\Sigma}_{\mathbf{yx}} & \boldsymbol{\Sigma}_{\mathbf{yy}}\end{pmatrix}. \tag{9.81}$$

Now we can prove the following.

Proposition 9.4 (Cook's trick). *The best linear predictor of* $\mathbf{z}=\binom{\mathbf{x}}{\mathbf{y}}$ *on the basis of* $\mathbf{x}$ *is*

$$\operatorname{BLP}(\mathbf{z};\mathbf{x})=\begin{pmatrix}\mathbf{x}\\ \boldsymbol{\mu}_{\mathbf{y}}+\boldsymbol{\Sigma}_{\mathbf{yx}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}})\end{pmatrix}=\begin{pmatrix}\mathbf{x}\\ \operatorname{BLP}(\mathbf{y};\mathbf{x})\end{pmatrix}. \tag{9.82}$$

Proof. We can write

$$\begin{aligned}\operatorname{BLP}(\mathbf{z};\mathbf{A}'\mathbf{z})&=\boldsymbol{\mu}_{\mathbf{z}}+\boldsymbol{\Sigma}\mathbf{A}(\mathbf{A}'\boldsymbol{\Sigma}\mathbf{A})^{-}\mathbf{A}'(\mathbf{z}-\boldsymbol{\mu}_{\mathbf{z}})\\ &=\begin{pmatrix}\boldsymbol{\mu}_{\mathbf{x}}\\ \boldsymbol{\mu}_{\mathbf{y}}\end{pmatrix}+\begin{pmatrix}\boldsymbol{\Sigma}_{\mathbf{xx}}\\ \boldsymbol{\Sigma}_{\mathbf{yx}}\end{pmatrix}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}})=\begin{pmatrix}\boldsymbol{\mu}_{\mathbf{x}}+\boldsymbol{\Sigma}_{\mathbf{xx}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}})\\ \boldsymbol{\mu}_{\mathbf{y}}+\boldsymbol{\Sigma}_{\mathbf{yx}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}})\end{pmatrix}\\ &=\begin{pmatrix}\mathbf{x}\\ \boldsymbol{\mu}_{\mathbf{y}}+\boldsymbol{\Sigma}_{\mathbf{yx}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}})\end{pmatrix}.\end{aligned} \tag{9.83}$$

Above we have used $\boldsymbol{\Sigma}_{\mathbf{xx}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}})=\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}}$, which follows from

$$\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}}\in\mathscr{C}(\boldsymbol{\Sigma}_{\mathbf{xx}}) \quad \text{with probability 1;} \tag{9.84}$$

see (1.40), page 63. □

We may end this section by an interesting citation from Christensen (2001, p. 123):

> "...principal component analysis in its modern form was originated by Hotelling (1933). It is curious that, although Hotelling was originally interested in a prediction problem (actually a factor analysis problem), he transformed his problem into one of finding vectors $\mathbf{a}_1,\dots,\mathbf{a}_r$ of fixed length such that $\mathbf{a}_i'\mathbf{y}$ has maximum variance subject to the condition that $\operatorname{cov}(\mathbf{a}_i'\mathbf{y},\mathbf{a}_j'\mathbf{y})=0$, $j=1,\dots,i-1$. This is the form in which principal component analysis is traditionally presented."

Interesting comments on the use/over-use of the principal component analysis are given by Speed (2008a).

9.5 A Property of the Prediction Errors

Let $\mathbf{z}$ be a d-dimensional random vector with $\operatorname{cov}(\mathbf{z}) = \mathbf{\Sigma}$ and denote $\mathbf{z}_{(i)} = (z_1, \dots, z_i)'$. Consider the following residuals (prediction errors)

$$e_{1\cdot 0} = z_1\,, \quad e_{2\cdot 1} = z_2 - \mathrm{BLP}(z_2; z_1)\,, \quad e_{3\cdot 12} = z_3 - \mathrm{BLP}(z_3; z_1, z_2)\,, \tag{9.85}$$

so that

$$e_{i\cdot 1\dots i-1} = z_i - \mathrm{BLP}(z_i; \mathbf{z}_{(i-1)})\,, \quad i = 2, \dots, d\,. \tag{9.86}$$

Let $\mathbf{e}$ be a d-dimensional random vector of these residuals:

$$\mathbf{e} = (e_{1\cdot 0}, e_{2\cdot 1}, \dots, e_{d\cdot 1\dots d-1})' = (e_1, e_2, \dots, e_d)'. \tag{9.87}$$

The vector $\mathbf{e}$ can be written as

$$\mathbf{e} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ f_{21} & 1 & 0 & \dots & 0 \\ f_{31} & f_{32} & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_{d1} & f_{d2} & f_{d3} & \dots & 1 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \\ \vdots \\ z_d \end{pmatrix} = \mathbf{F}\mathbf{z}\,, \tag{9.88}$$

for the relevant $\mathbf{F}$. In view of part (e) of Proposition 9.1 (p. 198), the residual e_i is uncorrelated with every element of $\mathbf{z}_{(i-1)}$. Because each e_i, $i = 1, \dots, i-1$, is a function of $\mathbf{z}_{(i-1)}$, the residual e_i is uncorrelated with each e_i, $i = 1, \dots, i-1$. Thereby we can conclude that $\operatorname{cov}(\mathbf{e}) = \mathbf{D}$ is a diagonal matrix, where the diagonal elements are the variances of the prediction errors:

$$\mathbf{D} = \operatorname{diag}(\sigma_1^2, \sigma_{2\cdot 1}^2, \dots, \sigma_{d\cdot 1\dots d-1}^2)\,, \tag{9.89}$$

with

$$\sigma_{2\cdot 1}^2 = \sigma_2^2 - \sigma_{12}^2/\sigma_1^2 = \sigma_2^2(1 - \varrho_{12}^2)\,, \tag{9.90a}$$

$$\begin{aligned} \sigma_{i\cdot \mathbf{x}_{(i-1)}}^2 &= \sigma_i^2 - [\operatorname{cov}(\mathbf{x}_{(i-1)}, x_i)]'[\operatorname{cov}(\mathbf{x}_{(i-1)})]^- \operatorname{cov}(\mathbf{x}_{(i-1)}, x_i) \\ &= \sigma_i^2(1 - \varrho_{i\cdot \mathbf{x}_{(i-1)}}^2)\,. \end{aligned} \tag{9.90b}$$

Here $\varrho_{i\cdot \mathbf{x}_{(i-1)}}$ refers to the population multiple correlation coefficient.

Because $\mathbf{F}$ is a lower triangular matrix with 1's in the diagonal, and $\operatorname{cov}(\mathbf{e}) = \mathbf{D} = \mathbf{F\Sigma F}'$, we have

$$\det(\mathbf{D}) = \det(\mathbf{F\Sigma F}') = \det(\mathbf{\Sigma}) = \sigma_1^2 \sigma_{2\cdot 1}^2 \cdots \sigma_{d\cdot 1\dots d-1}^2\,; \tag{9.91}$$

see also (13.45) (p. 298). Trivially $\mathbf{\Sigma}$ is positive definite if and only if all diagonal elements of $\mathbf{D}$ are nonzero (necessarily then positive). Note that $\mathbf{D} = \mathbf{F\Sigma F}'$ implies that

$$\mathbf{\Sigma} = \mathbf{F}^{-1}\mathbf{D}^{1/2}\mathbf{D}^{1/2}(\mathbf{F}')^{-1} := \mathbf{L}\mathbf{L}', \tag{9.92}$$

where $\mathbf{L} = \mathbf{F}^{-1}\mathbf{D}^{1/2}$ is a lower triangular matrix. As noted by Das Gupta (1997), this gives a statistical proof of the triangular factorization of a nonnegative definite matrix; see also Rao (1973a, p. 589). It is straightforward to confirm that all diagonal elements of $\mathbf{L}$ are nonnegative, and that if $\mathbf{\Sigma}$ is positive definite, then they are positive. If $\mathbf{\Sigma}$ is singular and thereby at least one diagonal element of $\mathbf{D}$ is zero, then the corresponding diagonal element of $\mathbf{L}$ is also zero as well as all elements of the corresponding column. The decomposition (9.92) is also known as the Cholesky decomposition.

We can do the above considerations also in the variable space using orthogonal projectors. To have a quick look at this approach, consider a centered $n \times d$ data matrix $\widetilde{\mathbf{X}} = (\tilde{\mathbf{x}}_1 : \ldots : \tilde{\mathbf{x}}_d)$ and the following residuals

$$\mathbf{e}_{1\cdot 0} = \tilde{\mathbf{x}}_1\,, \quad \mathbf{e}_{i\cdot 1\ldots i-1} = (\mathbf{I}_n - \mathbf{P}_{(\tilde{\mathbf{x}}_1 : \ldots : \tilde{\mathbf{x}}_{i-1})})\tilde{\mathbf{x}}_i\,, \tag{9.93}$$

and let $\mathbf{E}$ be an $n \times d$ matrix whose columns are these residuals:

$$\mathbf{E} = (\mathbf{e}_{1\cdot 0} : \mathbf{e}_{2\cdot 1} : \ldots : \mathbf{e}_{d\cdot 1\ldots d-1})\,. \tag{9.94}$$

Then $\mathbf{E}'\mathbf{E}$ appears to be a diagonal matrix where the ith diagonal element is

$$e_{ii} = \mathrm{SSE}(\tilde{\mathbf{x}}_i;\, \tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_{i-1})\,. \tag{9.95}$$

If the columns of $\widetilde{\mathbf{X}}$ have unit length, then

$$\det(\mathbf{E}'\mathbf{E}) = \det(\mathbf{R}) = (1 - r_{12}^2)(1 - R_{3\cdot 12}^2)\cdots(1 - R_{d\cdot 12\ldots d-1}^2)\,, \tag{9.96}$$

where $\mathbf{R}$ is the correlation matrix of from the data matrix $(\mathbf{x}_1 : \ldots : \mathbf{x}_d)$, and R^2 refers to the coefficient of determination.

9.6 Prediction in the Autocorrelation Structure

Consider, as on page 250, the p-dimensional random vector $\mathbf{y}$, which has the expectation $\mathrm{E}(\mathbf{y}) = \mathbf{0}$ and the covariance matrix:

$$\mathbf{\Sigma} = \sigma^2\{\varrho^{|i-j|}\} = \sigma^2 \begin{pmatrix} 1 & \varrho & \varrho^2 & \ldots & \varrho^{p-1} \\ \varrho & 1 & \varrho & \ldots & \varrho^{p-2} \\ \vdots & \vdots & \vdots & & \vdots \\ \varrho^{p-1} & \varrho^{p-2} & \varrho^{p-3} & \ldots & 1 \end{pmatrix} = \sigma^2 \begin{pmatrix} \mathbf{\Sigma}_{11} & \boldsymbol{\sigma}_{12} \\ \boldsymbol{\sigma}'_{12} & 1 \end{pmatrix}, \tag{9.97}$$

where $\varrho^2 < 1$, and

$$\boldsymbol{\sigma}_{12} = (\varrho^{p-1}, \varrho^{p-2}, \dots, \varrho^2, \varrho)' \in \mathbb{R}^{p-1}. \tag{9.98}$$

We observe that $\boldsymbol{\sigma}_{12}$ can be obtained by multiplying the last column of $\mathbf{\Sigma}_{11}$ by ϱ:

$$\boldsymbol{\sigma}_{12} = \varrho \mathbf{\Sigma}_{11} \mathbf{u}\,, \tag{9.99}$$

where $\mathbf{u}$ is the last column of $\mathbf{I}_{p-1}$. Therefore

$$\mathbf{\Sigma}_{11}^{-1} \boldsymbol{\sigma}_{12} = \mathbf{\Sigma}_{11}^{-1} \cdot \varrho \mathbf{\Sigma}_{11} \mathbf{u} = \varrho \mathbf{u} = (0, \dots, 0, \varrho)' := \mathbf{b}_* \in \mathbb{R}^{p-1}. \tag{9.100}$$

Denoting $\mathbf{y}_{(i-1)} = (y_1, \dots, y_{i-1})'$ for $i = 2, \dots, p$, we observe that the best linear predictor of y_p on the basis of $\mathbf{y}_{(p-1)}$ is

$$\mathrm{BLP}(y_p; \mathbf{y}_{(p-1)}) = \mathbf{b}_*' \mathbf{y} = \varrho y_{p-1}\,, \tag{9.101}$$

while the corresponding residual (prediction error) is

$$e_{y_p \cdot \mathbf{y}_{(p-1)}} = y_p - \varrho y_{p-1}\,. \tag{9.102}$$

For each y_i, $i = 2, \dots, p$, we can now consider the corresponding BLPs and residuals:

$$\mathrm{BLP}(y_i; \mathbf{y}_{(i-1)}) = \varrho y_{i-1}\,, \qquad e_{y_i \cdot \mathbf{y}_{(i-1)}} = y_i - \varrho y_{i-1} := e_i\,. \tag{9.103}$$

Putting $e_1 = y_1$, we can conclude, just as on page 207, that the elements of vector $\mathbf{e} = (e_1, \dots, e_p)'$ are uncorrelated and their covariance matrix is

$$\mathrm{cov}(\mathbf{e}) = \sigma^2 \begin{pmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & (1-\varrho^2)\mathbf{I}_{p-1} \end{pmatrix} := \sigma^2 \mathbf{D}\,. \tag{9.104}$$

Noting that

$$\mathbf{e} = \begin{pmatrix} y_1 \\ y_2 - \varrho y_1 \\ y_3 - \varrho y_2 \\ \vdots \\ y_p - \varrho y_{p-1} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 & 0 & 0 \\ -\varrho & 1 & 0 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -\varrho & 1 & 0 \\ 0 & 0 & 0 & \dots & 0 & -\varrho & 1 \end{pmatrix} \mathbf{y} = \mathbf{L}\mathbf{y}\,, \tag{9.105}$$

we can observe that $\mathrm{cov}(\mathbf{e}) = \mathrm{cov}(\mathbf{L}\mathbf{y}) = \sigma^2 \mathbf{L}\mathbf{\Sigma}\mathbf{L}'$, and hence

$$\mathbf{L}\mathbf{\Sigma}\mathbf{L}' = \mathbf{D}\,, \tag{9.106}$$

i.e., the matrix $\mathbf{L}$ diagonalizes $\mathbf{\Sigma}$. Because $\det(\mathbf{L}) = 1$, we have

$$\det(\mathbf{\Sigma}) = \det(\mathbf{D}) = (1-\varrho^2)^{p-1}, \tag{9.107}$$

while the inverse of $\mathbf{\Sigma}$ appears to be

$$\begin{aligned}\mathbf{\Sigma}^{-1} &= \mathbf{L}'\mathbf{D}^{-1}\mathbf{L} = \frac{1}{1-\varrho^2}\mathbf{K}'\mathbf{K} \\ &= \frac{1}{1-\varrho^2}\begin{pmatrix} 1 & -\varrho & 0 & \dots & 0 & 0 & 0 \\ -\varrho & 1+\varrho^2 & -\varrho & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -\varrho & 1+\varrho^2 & -\varrho \\ 0 & 0 & 0 & \dots & 0 & -\varrho & 1 \end{pmatrix},\end{aligned} \tag{9.108}$$

where

$$\mathbf{K} = \begin{pmatrix} \sqrt{1-\varrho^2} & 0 & 0 & \dots & 0 & 0 & 0 \\ -\varrho & 1 & 0 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -\varrho & 1 & 0 \\ 0 & 0 & 0 & \dots & 0 & -\varrho & 1 \end{pmatrix}. \tag{9.109}$$

9.7 Minimizing cov(Hy − FMy)

Consider the linear model $\{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$. Then $\mathbf{Gy}$ is unbiased for $\mathbf{X}\boldsymbol{\beta}$ whenever $\mathbf{G}$ satisfies the equation $\mathbf{GX} = \mathbf{X}$. Let $\mathcal{U}$ denote the set of all matrices $\mathbf{L}$ such that

$$\mathcal{U} = \{\,\mathbf{L} : \mathbf{Ly} = \mathrm{LUE}(\mathbf{X}\boldsymbol{\beta})\,\} = \{\,\mathbf{L} : \mathbf{LX} = \mathbf{X}\,\}. \tag{9.110}$$

In view of Theorem 11 (p. 267), the set $\mathcal{U}$ can be expressed in the following (equivalent) ways:

$$\mathcal{U}_1 = \{\,\mathbf{L} : \mathbf{L} = \mathbf{H} + \mathbf{F}_1(\mathbf{I}-\mathbf{H})\,\}, \quad \mathcal{U}_2 = \{\,\mathbf{L} : \mathbf{L} = \mathbf{I} + \mathbf{F}_2(\mathbf{I}-\mathbf{H})\,\}, \tag{9.111}$$

i.e.,

$$\mathcal{U}_1 = \{\,\mathbf{L} : \mathbf{L} = \mathbf{H} + \mathbf{F}_1\mathbf{M}\,\}, \quad \mathcal{U}_2 = \{\,\mathbf{L} : \mathbf{L} = \mathbf{I} + \mathbf{F}_2\mathbf{M}\,\}, \tag{9.112}$$

where $\mathbf{F}_1$ and $\mathbf{F}_2$ are free to vary. In other words, all unbiased linear estimators of $\mathbf{X}\boldsymbol{\beta}$ can be generated through

$$\mathbf{Hy} + \mathbf{F}_1\mathbf{My} \tag{9.113}$$

by varying $\mathbf{F}_1$. The joint covariance matrix of $\mathbf{Hy}$ and $\mathbf{My}$ is

$$\mathrm{cov}\begin{pmatrix}\mathbf{Hy}\\ \mathbf{My}\end{pmatrix} = \begin{pmatrix}\mathbf{HVH} & \mathbf{HVM}\\ \mathbf{MVH} & \mathbf{MVM}\end{pmatrix} := \mathbf{\Sigma}_{\mathrm{HM}}\,. \tag{9.114}$$

If we now apply the minimization result of Theorem 9 (p. 191) we obtain

$$\mathrm{cov}[\mathbf{Hy} - \mathbf{HVM}(\mathbf{MVM})^{-}\mathbf{My}] \leq_{\mathsf{L}} \mathrm{cov}(\mathbf{Hy} - \mathbf{FMy}) \quad \text{for all } \mathbf{F}\,. \tag{9.115}$$

This means that the minimal covariance matrix is the Schur complement

$$\mathbf{\Sigma}_{\mathrm{HM}}/\mathbf{MVM} = \mathbf{HVH} - \mathbf{HVM}(\mathbf{MVM})^{-}\mathbf{MVH}\,, \tag{9.116}$$

and that

$$\mathbf{Hy} - \mathbf{HVM}(\mathbf{MVM})^{-}\mathbf{My} = \mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta})\,. \tag{9.117}$$

In the same way it is possible to show that when minimizing

$$\operatorname{cov}(\mathbf{y} - \mathbf{F}_2\mathbf{My}) \tag{9.118}$$

with respect to $\mathbf{F}_2$, we end up to the representation

$$\mathbf{y} - \mathbf{VM}(\mathbf{MVM})^{-}\mathbf{My} = \mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta})\,. \tag{9.119}$$

Here we have

$$\operatorname{cov}\begin{pmatrix} \mathbf{y} \\ \mathbf{My} \end{pmatrix} = \begin{pmatrix} \mathbf{V} & \mathbf{VM} \\ \mathbf{MV} & \mathbf{MVM} \end{pmatrix} := \mathbf{\Sigma}_{\mathrm{IM}}\,. \tag{9.120}$$

So the BLUE's covariance matrix can be interpreted as a Schur complement:

$$\operatorname{cov}[\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta})] = \mathbf{\Sigma}_{\mathrm{HM}}/\mathbf{MVM} = \mathbf{\Sigma}_{\mathrm{IM}}/\mathbf{MVM}\,. \tag{9.121}$$

9.8 Exercises

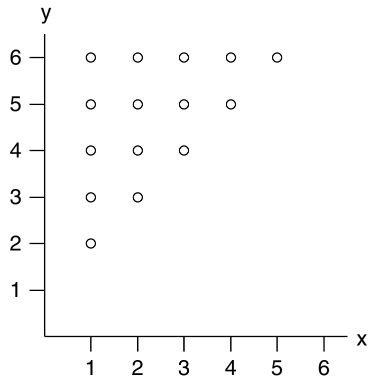


Figure 9.4 (See Exercise 9.1.) A two-dimensional discrete distribution, each of the 15 pairs appearing with the same probability.

9.1. Consider the Figure 9.4. Find the best predictor for y on the basis of x. What is $\operatorname{cor}(x, y)$? (Do your calculations without pen and paper.)

9.2. Prove that $\mathrm{E}(y \mid \mathbf{x})$ is the best predictor for y on the basis of $\mathbf{x}$, i.e., $\mathrm{E}(y \mid \mathbf{x}) = \mathrm{BP}(y; \mathbf{x})$.

Christensen (2002, p. 132), Rao (1973a, p. 264),
Searle, Casella & McCulloch (1992, §7.2).

9.3. Confirm that $\mathrm{E}[y - (\mathbf{a}'\mathbf{x} + b)]^2 = \mathrm{var}(y - \mathbf{a}'\mathbf{x}) + (\mu_y - \mathbf{a}'\boldsymbol{\mu}_{\mathbf{x}} - b)^2$.

9.4. Let $\binom{x}{y}$ be a random vector such that $\mathrm{E}(y \mid x = \underline{x}) = \alpha + \beta\underline{x}$. Show (in details) that then $\beta = \sigma_{xy}/\sigma_x^2$, and $\alpha = \mu_y - \beta \cdot \mu_x$. See also (9.35) and (9.36) (p. 195).

9.5. Prove Proposition 9.2 (p. 203).

9.6. Confirm that (9.96) (p. 208) holds:

$$\det(\mathbf{E}'\mathbf{E}) = \det(\mathbf{R}) = (1 - r_{12}^2)(1 - R_{3\cdot 12}^2) \cdots (1 - R_{d\cdot 12\ldots d-1}^2)\,.$$

9.7. Mustonen (1997) proposed the following statistics as a measure for total variability in multivariate normal distribution:

$$\mathrm{Mvar}(\boldsymbol{\Sigma}) = \max \sum_{i=1}^{d} \sigma_{i\cdot 12\ldots i-1}^2\,,$$

where the maximum is sought over all permutations of variables $z_1, \ldots, z_d$.

(a) Find $\mathrm{Mvar}(\boldsymbol{\Sigma})$ when $d = 2$.

(b) Confirm that $\mathrm{Mvar}(\boldsymbol{\Sigma}) = \mathrm{tr}(\mathbf{D}) = \mathrm{tr}(\mathbf{F}\boldsymbol{\Sigma}\mathbf{F}')$, where the lower diagonal matrix $\mathbf{F}$ is given as in (9.88) (p. 207).

(c) Show that $\det(\mathbf{D}) = \det(\mathbf{F}\boldsymbol{\Sigma}\mathbf{F}') = \det(\boldsymbol{\Sigma})$ for every ordering of the elements $\mathbf{y}$, but $\mathrm{Mvar}(\boldsymbol{\Sigma}) = \mathrm{tr}(\mathbf{D})$ can be dependent on the ordering of the elements of $\mathbf{y}$.

9.8. Let $\mathbf{z}$ be a partitioned random vector with positive definite covariance matrix

$$\mathrm{cov}(\mathbf{z}) = \mathrm{cov}\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{xx}} & \boldsymbol{\Sigma}_{\mathbf{xy}} \\ \boldsymbol{\Sigma}_{\mathbf{yx}} & \boldsymbol{\Sigma}_{\mathbf{yy}} \end{pmatrix} = \boldsymbol{\Sigma}\,,$$

where $\mathbf{x}$ has p elements and $\mathbf{y}$ has q elements. We can now consider the vector space $\mathcal{R} = \{\, u : u = \mathbf{a}'\mathbf{z} \text{ for some } \mathbf{a} \in \mathbb{R}^{p+q} \,\}$, where the inner product and norm are defined as

$$\langle u, v \rangle = \mathrm{cov}(u, v)\,, \quad \|u\|^2 = \mathrm{var}(u)\,.$$

Hence when $u = \mathbf{a}'\mathbf{z}$ and $v = \mathbf{b}'\mathbf{z}$, we have $\langle u, v \rangle = \mathbf{a}'\boldsymbol{\Sigma}\mathbf{b}$, $\|u\|^2 = \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}$. Orthogonality of the random variables u and v in $\mathcal{R}$ means that u and v are uncorrelated. Let $\mathcal{L}$ be a subspace of $\mathcal{R}$ defined as

$$\mathcal{L} = \{\, v : v = \mathbf{b}'\mathbf{x} \text{ for some } \mathbf{b} \in \mathbb{R}^p \,\} \subset \mathcal{R}\,.$$

Confirm that the orthocomplement of $\mathcal{L}$ is

$$\mathcal{L}^{\perp} = \{\, t \in \mathcal{R} : t = \mathbf{d}'(\mathbf{y} - \mathbf{\Sigma}_{\mathbf{yx}}\mathbf{\Sigma}_{\mathbf{xx}}^{-1}\mathbf{x}) \text{ for some } \mathbf{d} \in \mathbb{R}^q \,\} .$$

Note that

$$\mathbf{a}'\mathbf{y} = \mathbf{a}'\mathbf{\Sigma}_{\mathbf{yx}}\mathbf{\Sigma}_{\mathbf{xx}}^{-1}\mathbf{x} + \mathbf{a}'(\mathbf{y} - \mathbf{\Sigma}_{\mathbf{yx}}\mathbf{\Sigma}_{\mathbf{xx}}^{-1}\mathbf{x}) ,$$

and hence $\mathbf{a}'\mathbf{\Sigma}_{\mathbf{yx}}\mathbf{\Sigma}_{\mathbf{xx}}^{-1}\mathbf{x}$ can be considered as an orthogonal projection of $\mathbf{a}'\mathbf{y}$ onto the subspace $\mathcal{L}$. For the connection between geometry, statistics, and probability, see, e.g., Bryant (1984) and Margolis (1979).

9.9. Denote

$$\mathbf{\Sigma}^{+} = \begin{pmatrix} \mathbf{\Sigma}_{\mathbf{xx}} & \boldsymbol{\sigma}_{\mathbf{x}y} \\ \boldsymbol{\sigma}'_{\mathbf{x}y} & \sigma_y^2 \end{pmatrix}^{+} = \begin{pmatrix} \mathbf{A}'\mathbf{A} & \mathbf{A}'\mathbf{b} \\ \mathbf{b}'\mathbf{A} & \mathbf{b}'\mathbf{b} \end{pmatrix}^{+} := \begin{pmatrix} \cdot & \cdot \\ \cdot & \sigma^{yy} \end{pmatrix} .$$

Confirm:

(a) $\mathbf{b} \notin \mathscr{C}(\mathbf{A})$ and $\mathbf{b} \neq \mathbf{0} \implies \sigma^2_{y\cdot\mathbf{x}} > 0 , \quad \varrho^2_{y\cdot\mathbf{x}} < 1 , \quad \sigma^{yy} = 1/\sigma^2_{y\cdot\mathbf{x}} .$
(b) $\mathbf{b} \in \mathscr{C}(\mathbf{A})$ and $\mathbf{b} \neq \mathbf{0} \implies \sigma^2_{y\cdot\mathbf{x}} = 0 , \quad \varrho^2_{y\cdot\mathbf{x}} = 1 ,$
(c) $\mathbf{b} = \mathbf{0} \implies \sigma^2_{y\cdot\mathbf{x}} = 0 , \quad \sigma^{yy} = (\sigma^2_{y\cdot\mathbf{x}})^{+} = 0^{+} = 0 .$

The above results follow from Proposition 13.1 (p. 294). For related considerations, see also Rao (1981).

9.10. The model for factor analysis is $\mathbf{x} = \boldsymbol{\mu} + \mathbf{A}\mathbf{f} + \boldsymbol{\varepsilon}$, where $\mathbf{x}$ is an observable random vector of p components; $\mathrm{E}(\mathbf{x}) = \boldsymbol{\mu}$ and $\mathrm{cov}(\mathbf{x}) = \mathbf{\Sigma}$. Vector $\mathbf{f}$ is an m-dimensional random vector, $m \leq p$, whose elements are called (common) factors. The elements of $\boldsymbol{\varepsilon}$ are called specific or unique factors. The matrix $\mathbf{A}_{p\times m}$ is the unknown matrix of factor loadings. Moreover,

$$\begin{aligned} &\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0} , \quad \mathrm{cov}(\boldsymbol{\varepsilon}) = \mathbf{\Psi} = \mathrm{diag}(\psi_1^2, \dots, \psi_p^2) , \\ &\mathrm{E}(\mathbf{f}) = \mathbf{0} , \quad \mathrm{cov}(\mathbf{f}) = \mathbf{\Phi} = \mathbf{I}_m , \quad \mathrm{cov}(\mathbf{f}, \boldsymbol{\varepsilon}) = \mathbf{0} . \end{aligned}$$

The fundamental equation for factor analysis is $\mathrm{cov}(\mathbf{x}) = \mathrm{cov}(\mathbf{A}\mathbf{f}) + \mathrm{cov}(\boldsymbol{\varepsilon})$, i.e., $\mathbf{\Sigma} = \mathbf{A}\mathbf{A}' + \mathbf{\Psi}$. Show that

$$\mathrm{cov}\begin{pmatrix} \mathbf{x} \\ \mathbf{f} \end{pmatrix} = \begin{pmatrix} \mathbf{\Sigma} & \mathbf{A} \\ \mathbf{A}' & \mathbf{I}_m \end{pmatrix} \text{ and } \mathrm{cov}(\mathbf{f} - \mathbf{L}\mathbf{x}) \geq_{\mathsf{L}} \mathrm{cov}(\mathbf{f} - \mathbf{A}'\mathbf{\Sigma}^{-1}\mathbf{x}) \quad \text{for all } \mathbf{L} .$$

The matrix $\mathbf{L}_* = \mathbf{A}'\mathbf{\Sigma}^{-1}$ can be represented as

$$\mathbf{L}_* = \mathbf{A}'\mathbf{\Sigma}^{-1} = (\mathbf{A}'\mathbf{\Psi}^{-1}\mathbf{A} + \mathbf{I}_m)^{-1}\mathbf{A}'\mathbf{\Psi}^{-1},$$

and the individual factor scores can be obtained by $\mathbf{L}_*(\mathbf{x} - \boldsymbol{\mu})$. (In practice all quantities are replaced by their estimates.)

Anderson (2003, §14.7), Neudecker & Satorra (2003).

Philatelic Item 9.1 Johann Carl Friedrich Gauss (1777–1855) was a German mathematician and scientist who contributed significantly to many fields, including number theory, statistics, analysis, differential geometry, geodesy, geophysics, electrostatics, astronomy and optics. For the Gauss–Markov theorem, see page 41 and Theorem 10 (p. 216). The stamp (left panel) was issued by Germany (Deutsche Bundespost) in 1955 (*Scott* 725) and the stamp (right panel) was issued by Nicaragua in 1994 (*Scott* 1985i).

Chapter 10
BLUE

To make a name for learning
when other roads are barred,
take something very easy
and make it very hard.

Piet Hein

In this chapter we focus on the BLUE-related matters, so to say. The most important thing is the fundamental BLUE equation (10.4) (p. 216).

Over the years, one of our favorite research topics in linear models has been the equality between OLSE and BLUE of $\mathbf{X}\boldsymbol{\beta}$. In Proposition 10.1 (p. 218) we collect together some necessary and sufficient conditions for their equality. We find this collection very useful and we believe it includes several interesting linear algebraic problems.

While preparing this book, for a long time the title of this chapter was "OLSE vs. BLUE", but the simpler version "BLUE" describes better the matrix tricks under consideration. In the sections of this chapter we consider, for example, the best unbiased linear predictors, BLUPs, and mixed models. The equality of the BLUEs under two different linear models is treated in Section 11.1 (p. 269). A short summary of the best linear unbiased estimation appears in Puntanen & Styan (2011).

As noted by Puntanen & Styan (1989, p. 154), the first condition for the equality between OLSE and BLUE of $\mathbf{X}\boldsymbol{\beta}$ was obtained by Anderson (1948, p. 92):

> "Let $\mathbf{X}$ and $\mathbf{V}$ have full rank. If the p columns of the $n \times p$ matrix $\mathbf{X}$ are linear combinations of p of the eigenvectors of $\mathbf{V}$, then OLSE is BLUE."

Anderson's result was published in *Skandinavisk Aktuarietidskrift*, and as Anderson says in his interview in *Statistical Science* (DeGroot, 1986, p. 102): "As a result it did not get a great deal of attention … So from time to time people discover that paper."

There are numerous authors who have contributed in this area (for the history of OLSE vs. BLUE, see Puntanen & Styan 1989), but obviously the major breakthroughs were made by Rao (1967) and Zyskind (1967). For some further references from those years we may mention Kruskal (1968), Rao (1968), Mitra & Rao (1969), Watson (1967), Zyskind (1969, 1975), and Zyskind & Martin (1969). For the reviews of the OLSE vs. BLUE -related problems, see Puntanen & Styan (1989) [including comments by Kempthorne

S. Puntanen et al., *Matrix Tricks for Linear Statistical Models: Our Personal Top Twenty*,
DOI 10.1007/978-3-642-10473-2_11,

(1989) (see also Kempthorne 1976) and Searle (1989)], and Baksalary, Puntanen & Styan (1990a).

We represent here (again) the “fundamental BLUE equation” which guarantees that a particular linear estimator $\mathbf{Gy}$ is the BLUE of $\mathbf{X}\boldsymbol{\beta}$ under $\{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{V}\}$. These conditions—as well as the conditions for the equality between OLSE and BLUE—appear to be amazingly useful in various occasions; see e.g. Werner (1997). Notice that in these considerations σ^2 has no role and hence we may put $\sigma^2 = 1$.

Photograph 10.1 Hans Joachim Werner (right, Hyderabad, 2000).

Consider the eigenvalue decomposition of the covariance matrix $\mathbf{V}$:

$$\mathbf{V} = \mathbf{T}\boldsymbol{\Lambda}\mathbf{T}', \tag{10.1}$$

where $\boldsymbol{\Lambda}$ is an $n \times n$ diagonal matrix of the n eigenvalues λ_i of $\mathbf{V}$:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_v > 0 = \lambda_{v+1} = \cdots = \lambda_n\,, \quad v = \operatorname{rank}(\mathbf{V})\,. \tag{10.2}$$

The matrix $\mathbf{T}$ is an $n \times n$ matrix comprising the corresponding orthonormal eigenvectors $\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_n$. We also let $\lambda_{\{1\}} > \lambda_{\{2\}} > \cdots > \lambda_{\{t\}} \geq 0$ denote the t *distinct* eigenvalues of $\mathbf{V}$ with multiplicities $m_1, m_2, \dots, m_t$, and let $\mathbf{T}_{\{1\}}, \mathbf{T}_{\{2\}}, \dots, \mathbf{T}_{\{t\}}$ be matrices consisting of (the sets of) corresponding orthonormal eigenvectors so that

$$\mathbf{T} = (\mathbf{T}_{\{1\}} : \mathbf{T}_{\{2\}} : \ldots : \mathbf{T}_{\{t\}})\,, \quad \mathbf{T}_i'\mathbf{T}_i = \mathbf{I}_{m_i}\,, \quad i = 1, 2, \dots, t\,, \tag{10.3a}$$

$$\mathbf{V} = \lambda_{\{1\}}\mathbf{T}_{\{1\}}\mathbf{T}_{\{1\}}' + \lambda_{\{2\}}\mathbf{T}_{\{2\}}\mathbf{T}_{\{2\}}' + \cdots + \lambda_{\{t\}}\mathbf{T}_{\{t\}}\mathbf{T}_{\{t\}}'\,, \tag{10.3b}$$

$$m_1 + m_2 + \cdots + m_t = n\,. \tag{10.3c}$$

Theorem 10 (Fundamental BLUE **equation).** *Consider the general linear model* $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$. *Then the estimator* $\mathbf{Gy}$ *is the* BLUE *for* $\mathbf{X}\boldsymbol{\beta}$ *if and only if* $\mathbf{G}$ *satisfies the equation*

$$\mathbf{G}(\mathbf{X} : \mathbf{V}\mathbf{X}^{\perp}) = (\mathbf{X} : \mathbf{0})\,. \tag{10.4}$$

The corresponding condition for $\mathbf{Ay}$ *to be the* BLUE *of an estimable parametric function* $\mathbf{K}'\boldsymbol{\beta}$ *is*

$$\mathbf{A}(\mathbf{X} : \mathbf{V}\mathbf{X}^{\perp}) = (\mathbf{K}' : \mathbf{0})\,. \tag{10.5}$$

Proof. It is obvious that in condition $\mathbf{V}\mathbf{X}^{\perp} = \mathbf{0}$ we can have *any* $\mathbf{X}^{\perp}$. With the choice of $\mathbf{M} = \mathbf{I} - \mathbf{H}$ as $\mathbf{X}^{\perp}$ our claim becomes

$$\text{(a)}\ \ \mathbf{Gy} = \mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta}) \iff \text{(b)}\ \ \mathbf{G}(\mathbf{X} : \mathbf{VM}) = (\mathbf{X} : \mathbf{0})\,. \tag{10.6}$$

Notice that in view of Proposition 5.8 (p. 139), equation (b) indeed has a solution for $\mathbf{G}$ and that the solution is unique if and only if $\mathscr{C}(\mathbf{X} : \mathbf{VM}) = \mathbb{R}^n$. Trivially if $\mathbf{G}_1$ and $\mathbf{G}_2$ are two solutions to (b), then

$$\mathbf{G}_1(\mathbf{X} : \mathbf{VM}) = \mathbf{G}_2(\mathbf{X} : \mathbf{VM})\,. \tag{10.7}$$

Assume first that (b) holds for $\mathbf{G}$, and let $\mathbf{F}$ be an arbitrary matrix (of conformable dimension, of course) satisfying the unbiasedness condition $\mathbf{FX} = \mathbf{X}$. Then

$$(\mathbf{F} - \mathbf{G})\mathbf{X} = \mathbf{0}\,, \quad \text{i.e.,} \quad \mathscr{C}(\mathbf{F}' - \mathbf{G}') \subset \mathscr{C}(\mathbf{M})\,, \tag{10.8}$$

which implies that $\mathbf{F}' - \mathbf{G}' = \mathbf{ML}$ for some matrix $\mathbf{L}$. The assumption $\mathbf{GVM} = \mathbf{0}$ means that $\mathbf{GV}(\mathbf{F}' - \mathbf{G}') = \mathbf{GVML} = \mathbf{0}$, and thereby

$$\operatorname{cov}[\mathbf{Gy}, (\mathbf{F} - \mathbf{G})\mathbf{y}] = \mathbf{GV}(\mathbf{F} - \mathbf{G})' = \mathbf{0}\,. \tag{10.9}$$

Now the uncorrelatedness between $\mathbf{Gy}$ and $(\mathbf{F} - \mathbf{G})\mathbf{y}$ means that we can write

$$\begin{aligned} \operatorname{cov}(\mathbf{Fy}) &= \operatorname{cov}[(\mathbf{F} - \mathbf{G})\mathbf{y} + \mathbf{Gy}] \\ &= \operatorname{cov}[(\mathbf{F} - \mathbf{G})\mathbf{y}] + \operatorname{cov}(\mathbf{Gy}) \geq_{\mathsf{L}} \operatorname{cov}(\mathbf{Gy})\,, \end{aligned} \tag{10.10}$$

and so indeed (b) implies (a).

To go the other way round, we have to show that if $\mathbf{GX} = \mathbf{X}$ and

$$\mathbf{GVG}' \leq_{\mathsf{L}} \mathbf{FVF}' \quad \text{for all } \mathbf{F}\text{: } \mathbf{FX} = \mathbf{X}\,, \tag{10.11}$$

then

$$\mathbf{GVM} = \mathbf{0}\,. \tag{10.12}$$

The key in the proof is now choosing $\mathbf{F}$ in (10.11) so that it forces our claim to hold. Let us choose $\mathbf{F}$ so that it satisfies both $\mathbf{FX} = \mathbf{X}$ and $\mathbf{FVM} = \mathbf{0}$. (Convince yourself that such a choice is always possible!) Then

$$(\mathbf{G} - \mathbf{F})\mathbf{X} = \mathbf{0}\,, \quad \text{i.e.,} \quad \mathscr{C}(\mathbf{G}' - \mathbf{F}') \subset \mathscr{C}(\mathbf{M})\,, \tag{10.13}$$

which implies that $\mathbf{G}' - \mathbf{F}' = \mathbf{MK}$ for some matrix $\mathbf{K}$ and so

$$\mathbf{FV}(\mathbf{G}' - \mathbf{F}') = \mathbf{0}\,, \quad \text{i.e.,} \quad \operatorname{cov}[\mathbf{Fy}, (\mathbf{G} - \mathbf{F})\mathbf{y}] = \mathbf{0}\,. \tag{10.14}$$

Therefore,

$$\operatorname{cov}(\mathbf{Gy}) = \operatorname{cov}[(\mathbf{G} - \mathbf{F})\mathbf{y} + \mathbf{Fy}] = \operatorname{cov}[(\mathbf{G} - \mathbf{F})\mathbf{y}] + \operatorname{cov}(\mathbf{Fy})\,, \tag{10.15}$$

and so

$$\operatorname{cov}(\mathbf{Gy}) \geq_{\mathsf{L}} \operatorname{cov}(\mathbf{Fy})\,. \tag{10.16}$$

However, our assumption (10.11) implies that we must have the equality in (10.16) which holds if and only if

$$\operatorname{cov}[(\mathbf{G} - \mathbf{F})\mathbf{y}] = (\mathbf{G} - \mathbf{F})\mathbf{V}(\mathbf{G} - \mathbf{F})' = \mathbf{0}\,, \tag{10.17}$$

i.e., $(\mathbf{G} - \mathbf{F})\mathbf{V} = \mathbf{0}$, and so

$$\mathbf{GV} = \mathbf{FV}\,. \tag{10.18}$$

Postmultiplying (10.18) by $\mathbf{M}$ yields $\mathbf{GVM} = \mathbf{FVM} = \mathbf{0}$. This proves our claim (10.12).

If we replace the unbiasedness condition $\mathbf{GX} = \mathbf{X}$ with $\mathbf{AX} = \mathbf{K}'$ and then proceed just as in the above proof we obtain (10.5). □

The proof above follows Groß (2004). There is rich literature in this area; for recent proofs, see, for example, Puntanen, Styan & Werner (2000b), Baksalary (2004), and Baksalary & Trenkler (2009b).

Proposition 10.1 (OLSE **vs.** BLUE**).** *Consider the general linear model* $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$. *Then* $\mathrm{OLSE}(\mathbf{X}\boldsymbol{\beta}) = \mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta})$ *if and only if any of the following equivalent conditions holds.* (*Note:* $\mathbf{V}$ *is replaceable with* $\mathbf{V}^+$ *and* $\mathbf{H}$ *and* $\mathbf{M}$ *can be interchanged.*)

(i) $\mathbf{HV} = \mathbf{VH}$,

(ii) $\mathbf{HV} = \mathbf{HVH}$,

(iii) $\mathbf{HVM} = \mathbf{0}$,

(iv) $\mathbf{X}'\mathbf{VZ} = \mathbf{0}$, *where* $\mathscr{C}(\mathbf{Z}) = \mathscr{C}(\mathbf{M})$,

(v) $\mathscr{C}(\mathbf{VX}) \subset \mathscr{C}(\mathbf{X})$,

(vi) $\mathscr{C}(\mathbf{VX}) = \mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{V})$,

(vii) $\mathbf{HVH} \leq_{\mathsf{L}} \mathbf{V}$, *i.e.,* $\mathbf{V} - \mathbf{HVH}$ *is nonnegative definite,*

(viii) $\mathbf{HVH} \leq_{\mathsf{rs}} \mathbf{V}$, *i.e.,*

$$\operatorname{rank}(\mathbf{V} - \mathbf{HVH}) = \operatorname{rank}(\mathbf{V}) - \operatorname{rank}(\mathbf{HVH})\,,$$

i.e., $\mathbf{HVH}$ *and* $\mathbf{V}$ *are rank-subtractive, or in other words,* $\mathbf{HVH}$ *is below* $\mathbf{V}$ *with respect to the minus ordering,*

(ix) $\mathscr{C}(\mathbf{X})$ *has a basis consisting of* r *eigenvectors of* $\mathbf{V}$, *where* $r = \operatorname{rank}(\mathbf{X})$,

(x) $\operatorname{rank}(\mathbf{T}'_{\{1\}}\mathbf{X}) + \cdots + \operatorname{rank}(\mathbf{T}'_{\{t\}}\mathbf{X}) = \operatorname{rank}(\mathbf{X})$, *where* $\mathbf{T}_{\{i\}}$ *is a matrix consisting of the orthogonal eigenvectors corresponding to the* i*th largest eigenvalue* $\lambda_{\{i\}}$ *of* $\mathbf{V}$; $\lambda_{\{1\}} > \lambda_{\{2\}} > \cdots > \lambda_{\{t\}}$,

(xi) $\mathbf{T}'_{\{i\}}\mathbf{H}\mathbf{T}_{\{i\}} = (\mathbf{T}'_{\{i\}}\mathbf{H}\mathbf{T}_{\{i\}})^2$ *for all* $i = 1, 2, \dots, t$,

(xii) $\mathbf{T}'_{\{i\}}\mathbf{H}\mathbf{T}_{\{j\}} = \mathbf{0}$ *for all* $i, j = 1, 2, \dots, t$, $i \neq j$,

(xiii) *the squared nonzero canonical correlations between* $\mathbf{y}$ *and* $\mathbf{Hy}$ *are the nonzero eigenvalues of* $\mathbf{V}^-\mathbf{HVH}$ *for all* $\mathbf{V}^-$, *i.e.,*

$$\operatorname{cc}_+^2(\mathbf{y}, \mathbf{Hy}) = \operatorname{nzch}(\mathbf{V}^-\mathbf{HVH}) \quad \textit{for all } \mathbf{V}^-,$$

(xiv) $\mathbf{V}$ *can be expressed as* $\mathbf{V} = \mathbf{HAH} + \mathbf{MBM}$, *where* $\mathbf{A}$ *and* $\mathbf{B}$ *are symmetric nonnegative definite, i.e.,*

$$\mathbf{V} \in \mathcal{V}_1 = \{\, \mathbf{V} \geq_{\mathsf{L}} \mathbf{0} : \mathbf{V} = \mathbf{HAH} + \mathbf{MBM},\ \mathbf{A} \geq_{\mathsf{L}} \mathbf{0},\ \mathbf{B} \geq_{\mathsf{L}} \mathbf{0} \,\},$$

(xv) $\mathbf{V}$ *can be expressed as* $\mathbf{V} = \mathbf{XCX}' + \mathbf{ZDZ}'$, *where* $\mathbf{C}$ *and* $\mathbf{D}$ *are symmetric nonnegative definite, i.e.,*

$$\mathbf{V} \in \mathcal{V}_2 = \{\, \mathbf{V} \geq_{\mathsf{L}} \mathbf{0} : \mathbf{V} = \mathbf{XCX}' + \mathbf{ZDZ}',\ \mathbf{C} \geq_{\mathsf{L}} \mathbf{0},\ \mathbf{D} \geq_{\mathsf{L}} \mathbf{0} \,\},$$

(xvi) $\mathbf{V}$ *can be expressed as* $\mathbf{V} = \alpha\mathbf{I} + \mathbf{XKX}' + \mathbf{ZLZ}'$, *where* $\alpha \in \mathbb{R}$, *and* $\mathbf{K}$ *and* $\mathbf{L}$ *are symmetric, such that* $\mathbf{V}$ *is nonnegative definite, i.e.,*

$$\mathbf{V} \in \mathcal{V}_3 = \{\, \mathbf{V} \geq_{\mathsf{L}} \mathbf{0} : \mathbf{V} = \alpha\mathbf{I} + \mathbf{XKX}' + \mathbf{ZLZ}',\ \mathbf{K} = \mathbf{K}',\ \mathbf{L} = \mathbf{L}' \,\}.$$

Proof. We see first that $\mathbf{Hy}$ is the BLUE for $\mathbf{X}\boldsymbol{\beta}$ if and only if $\mathbf{H}$ satisfies the equation

$$\mathbf{H}(\mathbf{X} : \mathbf{VM}) = (\mathbf{X} : \mathbf{0}), \tag{10.19}$$

which holds if and only if $\mathbf{HVM} = \mathbf{0}$; this is condition (iii).

We next go through the proofs of some of the claims of Proposition 10.1. The reader is encouraged to complete the set of the proofs.

The equivalence between (i)–(iv) is easy to establish. Part (v) is equivalent to $\mathscr{C}(\mathbf{VH}) \subset \mathscr{C}(\mathbf{H})$, i.e.,

$$\mathbf{VH} = \mathbf{HL} \quad \text{for some } \mathbf{L}. \tag{10.20}$$

Premultiplying (10.20) by $\mathbf{H}$ yields $\mathbf{HVH} = \mathbf{HL} = \mathbf{VH}$, and hence $\mathbf{VH}$ is symmetric, i.e., (i) holds. It is obvious that (i) implies (iv).

Let us then consider the proof for "(iii) $\Longrightarrow$ (ix)"; see Zyskind (1967, Th. 1). Let the columns of $\mathbf{X}_*$ and $\mathbf{Z}$ form orthonormal bases of $\mathscr{C}(\mathbf{X})$ and $\mathscr{C}(\mathbf{X})^\perp$, respectively, so that $\mathbf{A}_{n\times n} = (\mathbf{X}_* : \mathbf{Z})$ is orthogonal. Then

$$\mathbf{HVM} = \mathbf{0} \iff \mathbf{X}_*'\mathbf{VZ} = \mathbf{0}, \tag{10.21}$$

and so $\mathbf{HVM} = \mathbf{0}$ implies that

$$\mathbf{A}'\mathbf{V}\mathbf{A} = \begin{pmatrix} \mathbf{X}_*' \\ \mathbf{Z}' \end{pmatrix} \mathbf{V}(\mathbf{X}_* : \mathbf{Z}) = \begin{pmatrix} \mathbf{X}_*'\mathbf{V}\mathbf{X}_* & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}'\mathbf{V}\mathbf{Z} \end{pmatrix}. \tag{10.22}$$

Let $\mathbf{R}_x \in \mathbb{R}^{r\times r}$ and $\mathbf{R}_z \in \mathbb{R}^{(n-r)\times(n-r)}$ be orthogonal matrices such that

$$\mathbf{R}_x'\mathbf{X}_*'\mathbf{V}\mathbf{X}_*\mathbf{R}_x = \mathbf{D}_x\,, \quad \mathbf{R}_z'\mathbf{Z}'\mathbf{V}\mathbf{Z}\mathbf{R}_z = \mathbf{D}_z\,, \tag{10.23}$$

where $\mathbf{D}_x$ and $\mathbf{D}_z$ are diagonal matrices; the columns of $\mathbf{R}_x$ and $\mathbf{R}_z$ comprise the orthonormal eigenvectors of $\mathbf{X}_*'\mathbf{V}\mathbf{X}_*$ and $\mathbf{Z}'\mathbf{V}\mathbf{Z}$, respectively. Then the matrix

$$\mathbf{S} = (\mathbf{X}_* : \mathbf{Z}) \begin{pmatrix} \mathbf{R}_x & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_z \end{pmatrix} = (\mathbf{X}_*\mathbf{R}_x : \mathbf{Z}\mathbf{R}_z) \in \mathbb{R}^{n\times n} \tag{10.24}$$

is an orthogonal matrix which diagonalizes $\mathbf{V}$:

$$\mathbf{S}'\mathbf{V}\mathbf{S} = \begin{pmatrix} \mathbf{R}_x' & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_z' \end{pmatrix} \begin{pmatrix} \mathbf{X}_*' \\ \mathbf{Z}' \end{pmatrix} \mathbf{V}(\mathbf{X}_* : \mathbf{Z}) \begin{pmatrix} \mathbf{R}_x & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_z \end{pmatrix} = \begin{pmatrix} \mathbf{D}_x & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_z \end{pmatrix}. \tag{10.25}$$

This implies that the columns of $\mathbf{S}$—and thereby the columns of $\mathbf{X}_*\mathbf{R}_x$—are orthonormal eigenvectors of $\mathbf{V}$. Because $\mathbf{R}_x$ is orthogonal, we obviously have $\mathscr{C}(\mathbf{X}_*\mathbf{R}_x) = \mathscr{C}(\mathbf{X})$ and so we have proved the implication "(iv) $\implies$ (ix)".

Let us then prove that "(ix) $\implies$ (i)". Let $\mathbf{T}_{(r)}$ denote the matrix whose columns are those orthonormal eigenvectors of $\mathbf{V}$ which have the property $\mathscr{C}(\mathbf{T}_{(r)}) = \mathscr{C}(\mathbf{X})$, and let $\mathbf{V}$ have the following eigenvalue decomposition (with the obvious notation):

$$\mathbf{V} = (\mathbf{T}_{(r)} : \mathbf{T}_{(-r)}) \begin{pmatrix} \mathbf{\Lambda}_{(r)} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Lambda}_{(-r)} \end{pmatrix} \begin{pmatrix} \mathbf{T}_{(r)}' \\ \mathbf{T}_{(-r)}' \end{pmatrix}. \tag{10.26}$$

Now we have $\mathbf{H} = \mathbf{T}_{(r)}\mathbf{T}_{(r)}'$, and hence

$$\begin{aligned} \mathbf{H}\mathbf{V} &= \mathbf{T}_{(r)}\mathbf{T}_{(r)}'\mathbf{V} = \mathbf{T}_{(r)}\mathbf{T}_{(r)}'(\mathbf{T}_{(r)} : \mathbf{T}_{(-r)})\mathbf{\Lambda}\mathbf{T}' \\ &= (\mathbf{T}_{(r)} : \mathbf{0}) \begin{pmatrix} \mathbf{\Lambda}_{(r)} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Lambda}_{(-r)} \end{pmatrix} \begin{pmatrix} \mathbf{T}_{(r)}' \\ \mathbf{T}_{(-r)}' \end{pmatrix} \\ &= \mathbf{T}_{(r)}\mathbf{\Lambda}_{(r)}\mathbf{T}_{(r)}' = (\mathbf{H}\mathbf{V})', \end{aligned} \tag{10.27}$$

and hence "(ix) $\implies$ (i)" is confirmed.

Consider then the claim (vii): $\mathbf{H}\mathbf{V}\mathbf{H} \leq_{\mathsf{L}} \mathbf{V}$. Assuming that $\mathbf{H}\mathbf{y}$ is the BLUE for $\mathbf{X}\boldsymbol{\beta}$, we necessarily have

$$\mathbf{H}\mathbf{V}\mathbf{H} = \operatorname{cov}(\mathbf{H}\mathbf{y}) \leq_{\mathsf{L}} \operatorname{cov}(\mathbf{y}) = \mathbf{V}\,, \tag{10.28}$$

because $\mathbf{y}$ is an unbiased estimator for $\mathbf{X}\boldsymbol{\beta}$. On the other hand, assume that (vii) holds, so that $\mathbf{V} - \mathbf{H}\mathbf{V}\mathbf{H} \geq_{\mathsf{L}} \mathbf{0}$. This implies that

$$\mathbf{U} := \begin{pmatrix}\mathbf{H}\\ \mathbf{M}\end{pmatrix}(\mathbf{V}-\mathbf{HVH})(\mathbf{H} : \mathbf{M}) = \begin{pmatrix}\mathbf{0} & \mathbf{HVM}\\ \mathbf{MVH} & \mathbf{MVM}\end{pmatrix} \geq_{\mathsf{L}} \mathbf{0}\,. \tag{10.29}$$

The nonnegative definiteness of $\mathbf{U}$ implies, see (1.33) (p. 63), that $\mathscr{C}(\mathbf{HVM}) \subset \mathscr{C}(\mathbf{0})$ and thereby $\mathbf{HVM} = \mathbf{0}$. Thus we have proved that (vii) implies (iii).

It is worth noting that the last conditions from (xiv) onwards are based on a slightly different approach than the previous ones: in the last conditions we characterize the set of nonnegative definite solutions $\mathbf{V}$ to the equation $\mathbf{HVM} = \mathbf{0}$. It is easily seen that if the covariance matrix $\mathbf{V}$ belongs to any of the classes $\mathcal{V}_i$, then indeed $\mathbf{HVM} = \mathbf{0}$, guaranteeing that $\mathbf{Hy}$ is the BLUE for $\mathbf{X}\boldsymbol{\beta}$. What remains to prove is that $\mathbf{HVM} = \mathbf{0}$ implies that $\mathbf{V}$ can be expressed as a member of $\mathcal{V}_i$.

Let us first show that that $\mathbf{HVM} = \mathbf{0}$ implies (xiv). To do this we note that $\mathbf{V}$ can be written as $\mathbf{V} = \mathbf{LL}'$ for some $\mathbf{L}$, and in view of $\mathbb{R}^n = \mathscr{C}(\mathbf{H}) \oplus \mathscr{C}(\mathbf{M})$, we can express $\mathbf{L}$ as $\mathbf{L} = \mathbf{HU}_1 + \mathbf{MU}_2$ for some $\mathbf{U}_1$ and $\mathbf{U}_2$. Hence

$$\mathbf{V} = \mathbf{HU}_1\mathbf{U}_1'\mathbf{H} + \mathbf{MU}_2\mathbf{U}_2'\mathbf{M} + \mathbf{HU}_1\mathbf{U}_2'\mathbf{M} + \mathbf{MU}_2\mathbf{U}_1'\mathbf{H}\,. \tag{10.30}$$

Premultiplying (10.30) by $\mathbf{H}$ and postmultiplying by $\mathbf{M}$ yields, in view of $\mathbf{HVM} = \mathbf{0}$,

$$\mathbf{HVM} = \mathbf{HU}_1\mathbf{U}_2'\mathbf{M} = \mathbf{0}\,, \tag{10.31}$$

and hence

$$\mathbf{V} = \mathbf{HU}_1\mathbf{U}_1'\mathbf{H} + \mathbf{MU}_2\mathbf{U}_2'\mathbf{M} := \mathbf{HAH} + \mathbf{MBM}\,, \tag{10.32}$$

where $\mathbf{A}$ and $\mathbf{B}$ are nonnegative definite. This confirms that (iii) implies (xiv).

We can conclude that $\mathbf{HVM} = \mathbf{0}$ implies (xv) by rewriting (10.32) as

$$\mathbf{V} = \mathbf{XX}^{+}\mathbf{A}(\mathbf{X}^{+})'\mathbf{X}' + \mathbf{ZZ}^{+}\mathbf{B}(\mathbf{Z}^{+})'\mathbf{Z}' := \mathbf{XCX}' + \mathbf{ZDZ}', \tag{10.33}$$

where $\mathbf{C}$ and $\mathbf{D}$ are nonnegative definite.

Let us next confirm that $\mathbf{HVM} = \mathbf{0}$ implies that $\mathbf{V}$ can be represented as a member of $\mathcal{V}_3$. So we assume that $\mathbf{V} = \mathbf{HAH} + \mathbf{MBM}$ for some $\mathbf{A}$ and $\mathbf{B}$. Noting that trivially $\mathbf{0} = \alpha(\mathbf{I} - \mathbf{H} - \mathbf{M})$ for any $\alpha \in \mathbb{R}$, we can write

$$\begin{aligned}\mathbf{V} &= \mathbf{HAH} + \mathbf{MBM} + \alpha(\mathbf{I}_n - \mathbf{H} - \mathbf{M})\\ &= \alpha\mathbf{I} + \mathbf{H}(\mathbf{A} - \alpha\mathbf{I})\mathbf{H} + \mathbf{M}(\mathbf{B} - \alpha\mathbf{I})\mathbf{M}\,,\end{aligned} \tag{10.34}$$

from which we can conclude that $\mathbf{V} \in \mathcal{V}_3$.

We leave the rest of the proofs to be done by the reader.

For the proofs of (viii) and (xiii), see Baksalary & Puntanen (1990a), and for the proof of (x), see T. W. Anderson (1971, Th. 10.2.2) and (1972, Th. 2) Baksalary & Kala (1977, p. 309), and Styan (1973b, Th. 2).

For coordinate-free and geometric considerations, see Kruskal (1968), Drygas (1970), Eaton (1970, 1978, 1983), Gnot, Klonecki & Zmyślony (1980), Zmyślony (1980), and Phillips (1990, 1992). □

Next we prove the following result.

Proposition 10.2. *Under the linear model* $\{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \mathbf{V}\}$ *the following statements are equivalent:*

(a) $\mathrm{OLSE}(\mathbf{X}\boldsymbol{\beta}) = \mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta})$,

(b) $\mathrm{OLSE}(\mathbf{K}'\boldsymbol{\beta}) = \mathrm{BLUE}(\mathbf{K}'\boldsymbol{\beta})$ *for every estimable parametric function* $\mathbf{K}'\boldsymbol{\beta}$.

Proof. If (b) holds then obviously (a) holds because $\mathbf{X}\boldsymbol{\beta}$ is estimable. Consider now an estimable parametric function $\mathbf{K}'\boldsymbol{\beta}$. Then necessarily $\mathbf{K}' = \mathbf{L}\mathbf{X}$ for some matrix $\mathbf{L}$, and so $\mathbf{K}'\boldsymbol{\beta} = \mathbf{L}\mathbf{X}\boldsymbol{\beta}$ and

$$\mathrm{OLSE}(\mathbf{K}'\boldsymbol{\beta}) = \mathrm{OLSE}(\mathbf{L}\mathbf{X}\boldsymbol{\beta}) = \mathbf{L}\mathbf{H}\mathbf{y} = \mathbf{L}\cdot\mathrm{OLSE}(\mathbf{X}\boldsymbol{\beta})\,. \tag{10.35}$$

Now (a) means that

$$\mathbf{H}(\mathbf{X} : \mathbf{V}\mathbf{M}) = (\mathbf{X} : \mathbf{0})\,. \tag{10.36}$$

Premultiplying (10.36) by $\mathbf{L}$ yields

$$\mathbf{L}\mathbf{H}(\mathbf{X} : \mathbf{V}\mathbf{M}) = (\mathbf{L}\mathbf{X} : \mathbf{0})\,, \tag{10.37}$$

which implies that $\mathbf{L}\mathbf{H}\mathbf{y} = \mathrm{BLUE}(\mathbf{L}\mathbf{X}\boldsymbol{\beta})$, i.e., $\mathrm{OLSE}(\mathbf{K}'\boldsymbol{\beta}) = \mathrm{BLUE}(\mathbf{K}'\boldsymbol{\beta})$ for every estimable $\mathbf{K}'\boldsymbol{\beta}$. □

Notice that if $\mathbf{X}$ has full column rank, then every parametric function $\mathbf{K}'\boldsymbol{\beta}$ is estimable and

$$\mathrm{OLSE}(\boldsymbol{\beta}) = \mathrm{BLUE}(\boldsymbol{\beta}) \iff \mathrm{OLSE}(\mathbf{X}\boldsymbol{\beta}) = \mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta})\,. \tag{10.38}$$

It is also useful to observe that premultiplying $\mathbf{G}(\mathbf{X} : \mathbf{V}\mathbf{M}) = (\mathbf{X} : \mathbf{0})$ by a given matrix $\mathbf{L}$ yields

$$\mathbf{L}\mathbf{G}(\mathbf{X} : \mathbf{V}\mathbf{M}) = (\mathbf{L}\mathbf{X} : \mathbf{0})\,. \tag{10.39}$$

This means that $\mathbf{L}\mathbf{G}\mathbf{y}$ is the BLUE for $\mathbf{L}\mathbf{X}\boldsymbol{\beta}$, or in other words

$$\mathrm{BLUE}(\mathbf{L}\mathbf{X}\boldsymbol{\beta}) = \mathbf{L}\cdot\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta})\,. \tag{10.40}$$

The structure of $\mathbf{V}$ for which the OLSE of $\mathbf{K}'\boldsymbol{\beta}$ equals the BLUE of $\mathbf{K}'\boldsymbol{\beta}$ is considered by Rao & Mitra (1971b, p. 159), Groß, Trenkler & Werner (2001, p. 123), and Isotalo & Puntanen (2009).

10.1 BLUE & Pandora's Box when $\mathbf{X} = \mathbf{1}_n$

Consider the linear model $\mathscr{M} = \{\mathbf{y},\, \mathbf{1}\beta,\, \mathbf{V}\}$, where $\mathbf{1} \in \mathbb{R}^n$. Let $\mathbf{g}'\mathbf{y}$ be a candidate for the $\mathrm{BLUE}(\beta)$. The estimator $\mathbf{g}'\mathbf{y}$ is unbiased for β if

$$\mathrm{E}(\mathbf{g}'\mathbf{y}) = \mathbf{g}'\mathbf{1}\beta = \beta \quad \text{for all } \beta \in \mathbb{R}\,, \qquad \text{i.e.,} \quad \mathbf{g}'\mathbf{1} = 1\,. \tag{10.41}$$

Therefore, for finding the BLUE(β), we have to

$$\text{minimize } \operatorname{var}(\mathbf{g}'\mathbf{y}) = \mathbf{g}'\mathbf{V}\mathbf{g} \quad \text{under the constraint } \mathbf{g}'\mathbf{1} = 1\,, \tag{10.42}$$

that is, we have to find such a solution $\mathbf{g}_*$ to the linear equation $\mathbf{g}'\mathbf{1} = 1$ which has the minimum norm $\|\mathbf{g}\| = \sqrt{\mathbf{g}'\mathbf{V}\mathbf{g}}$. One way to do this is to use the minimum-norm-g-inverse, see (4.93) (p. 117). The other way to proceed is to consider $\mathbf{g}_*$ as a solution to (10.42), i.e., $\mathbf{g}_*$ is a solution to the fundamental BLUE equation

$$\mathbf{g}'(\mathbf{1} : \mathbf{V}\mathbf{1}^{\perp}) = \mathbf{g}'(\mathbf{1} : \mathbf{V}\mathbf{C}) = (1 : \mathbf{0}')\,, \quad \text{i.e.,} \quad \begin{pmatrix} \mathbf{C}\mathbf{V} \\ \mathbf{1}' \end{pmatrix} \mathbf{g} = \begin{pmatrix} \mathbf{0} \\ 1 \end{pmatrix}, \tag{10.43}$$

where $\mathbf{C} = \mathbf{I} - \mathbf{P}_{\mathbf{1}}$ is the centering matrix; it is one choice for $\mathbf{1}^{\perp}$. Now

$$\mathbf{C}\mathbf{V}\mathbf{g} = \mathbf{0} \iff \mathbf{V}\mathbf{g} \in \mathscr{C}(\mathbf{1}) \iff \mathbf{V}\mathbf{g} = \gamma\mathbf{1} \quad \text{for some } \gamma \in \mathbb{R}. \tag{10.44}$$

Therefore we can conclude that $\mathbf{g}$ is a solution to (10.43) if and only if there exists a scalar γ such that $\mathbf{g}$ is a solution to

$$\begin{pmatrix} \mathbf{V} & \mathbf{1} \\ \mathbf{1}' & 0 \end{pmatrix} \begin{pmatrix} \mathbf{g} \\ \gamma \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ 1 \end{pmatrix}. \tag{10.45}$$

The equation (10.45) is often called the Pandora's Box for the linear model; see, e.g., Rao (1971c).

In a general situation we observe that

$$\mathbf{G}(\mathbf{X} : \mathbf{V}\mathbf{M}) = (\mathbf{X} : \mathbf{0}) \iff \begin{pmatrix} \mathbf{M}\mathbf{V} \\ \mathbf{X}' \end{pmatrix} \mathbf{G}' = \begin{pmatrix} \mathbf{0} \\ \mathbf{X}' \end{pmatrix}. \tag{10.46}$$

In view of

$$\mathbf{M}\mathbf{V}\mathbf{G}' = \mathbf{0} \iff \mathscr{C}(\mathbf{V}\mathbf{G}') \subset \mathscr{C}(\mathbf{X}) \iff \mathbf{V}\mathbf{G}' = \mathbf{X}\mathbf{L} \text{ for some } \mathbf{L}\,, \tag{10.47}$$

we can formulate the following proposition:

Proposition 10.3 (Pandora's Box). *Consider the general linear model* $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$*. Then the estimator* $\mathbf{G}\mathbf{y}$ *is the* BLUE *for* $\mathbf{X}\boldsymbol{\beta}$ *if and only if there exists a matrix* $\mathbf{L} \in \mathbb{R}^{p\times n}$ *so that* $\mathbf{G}$ *is a solution to*

$$\begin{pmatrix} \mathbf{V} & \mathbf{X} \\ \mathbf{X}' & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{G}' \\ \mathbf{L} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{X}' \end{pmatrix}. \tag{10.48}$$

For further properties of the Pandora's Box, see Exercise 10.18 (p. 264).

10.2 OLSE vs BLUE when $\mathbf{X} = \mathbf{1}_n$

Consider again the simple linear model $\mathscr{M} = \{\mathbf{y},\, \mathbf{1}\beta,\, \mathbf{V}\}$, where $\mathbf{1} \in \mathbb{R}^n$, and $\mathbf{V}$ is positive definite. Then we have

$$\operatorname{OLSE}(\beta) = \hat{\beta} = (\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'\mathbf{y} = \bar{y}\,, \qquad \operatorname{var}(\hat{\beta}) = \frac{1}{n^2}\mathbf{1}'\mathbf{V}\mathbf{1}\,, \tag{10.49a}$$

$$\operatorname{BLUE}(\beta) = \tilde{\beta} = (\mathbf{1}'\mathbf{V}^{-1}\mathbf{1})^{-1}\mathbf{1}'\mathbf{V}^{-1}\mathbf{y}\,, \quad \operatorname{var}(\tilde{\beta}) = (\mathbf{1}'\mathbf{V}^{-1}\mathbf{1})^{-1}. \tag{10.49b}$$

Because the "B" in BLUE means "best" in the sense of minimum variance, we must have

$$(\mathbf{1}'\mathbf{V}^{-1}\mathbf{1})^{-1} \leq \frac{1}{n^2}\mathbf{1}'\mathbf{V}\mathbf{1}\,. \tag{10.50}$$

We can rewrite (10.50) as

$$(\mathbf{1}'\mathbf{1})^2 \leq \mathbf{1}'\mathbf{V}^{-1}\mathbf{1} \cdot \mathbf{1}'\mathbf{V}\mathbf{1}\,, \tag{10.51}$$

which is a special case of the Cauchy–Schwarz inequality. The equality in (10.51) holds if and only if the vectors $\mathbf{V}^{1/2}\mathbf{1}$ and $\mathbf{V}^{-1/2}\mathbf{1}$ are linearly dependent, i.e., there exists a scalar $\lambda \in \mathbb{R}$ such that

$$\mathbf{V}^{1/2}\mathbf{1} = \lambda\mathbf{V}^{-1/2}\mathbf{1}\,, \tag{10.52}$$

which means that

$$\mathbf{V}\mathbf{1} = \lambda\mathbf{1}\,, \tag{10.53}$$

and hence $\mathbf{1}$ is an eigenvector of $\mathbf{V}$. In other words, the arithmetic mean is the BLUE whenever the covariance matrix $\mathbf{V}$ has all of its row totals equal; cf. also Blom (1976) and Zyskind (1969, p. 1361). Condition (10.53) is just a version of Anderson's original condition in 1948 for the equality of OLSE and BLUE.

Notice in passing that if $\mathbf{V}$ satisfies (10.53) and all its elements are nonnegative, then $\mathbf{V}/\lambda$ is a *doubly stochastic matrix*; both the row sums and column sums are all ones. If also the both diagonal sums equal to one, then $\mathbf{V}$ can be called a superstochastic matrix, see Gustafson & Styan (2009). For example, the matrix $\mathbf{A}/34$, where $\mathbf{A}$ is the magic square of Exercise 0.28 (p. 53) is a superstochastic matrix.

We may further point out that condition (10.53) is equivalent, for example, to each of the following conditions:

$$\mathbf{J}\mathbf{V} = \mathbf{V}\mathbf{J}\,, \quad \mathbf{J}\mathbf{V}\mathbf{J} = \mathbf{V}\mathbf{J}\,, \quad \mathbf{C}\mathbf{V} = \mathbf{V}\mathbf{C}\,, \tag{10.54}$$

where $\mathbf{J} = \mathbf{P}_{\mathbf{1}} = \frac{1}{n}\mathbf{1}\mathbf{1}'$, $\mathbf{C} = \mathbf{I} - \mathbf{J}$ is the centering matrix, and the covariance matrix $\mathbf{V}$ can be singular.

Let us consider a simple example when the model is $\{\mathbf{y},\, \mathbf{1}\beta,\, \mathbf{V}\}$ where

$$\mathbf{V} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix} \quad \text{and} \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}. \tag{10.55}$$

Then $\hat{\beta} = (\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'\mathbf{y} = \bar{y}$ and

$$\tilde{\beta} = (\mathbf{1}'\mathbf{V}^{-1}\mathbf{1})^{-1}\mathbf{1}'\mathbf{V}^{-1}\mathbf{y} = \frac{y_1 + \frac{1}{2}y_2 + \frac{1}{3}y_3}{\frac{11}{6}} = \frac{6y_1 + 3y_2 + 2y_3}{11}, \tag{10.56}$$

and

$$\operatorname{var}(\hat{\beta}) = \frac{1}{3^2}\mathbf{1}'\mathbf{V}\mathbf{1} = \frac{6}{9} = \frac{2}{3} = \frac{22}{33}, \tag{10.57a}$$

$$\operatorname{var}(\tilde{\beta}) = (\mathbf{1}'\mathbf{V}^{-1}\mathbf{1})^{-1} = \frac{6}{11} = \frac{18}{33}. \tag{10.57b}$$

Photograph 10.2 Roman Zmyślony (Smolenice Castle, Slovakia, 2009).

Photograph 10.3 Karl Gustafson (Auckland, 2005).

10.3 OLSE vs BLUE under the Intraclass Correlation Structure

Consider the linear model $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}\}$:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \text{where } \operatorname{cov}(\mathbf{y}) = \operatorname{cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}. \tag{10.58}$$

If we center the model $\mathscr{M}$, that is, if we premultiply (10.58) by the centering matrix $\mathbf{C}$, then we obtain the model

$$\mathscr{M}_c = \{\mathbf{Cy},\, \mathbf{CX}\boldsymbol{\beta},\, \sigma^2\mathbf{C}\}\,. \tag{10.59}$$

Here the covariance matrix is $\sigma^2\mathbf{C}$, which is singular, and we may wonder what happens to the BLUE. But now the condition (v) of Proposition 10.1 (p. 218), i.e., $\mathscr{C}(\mathbf{VX}) \subset \mathscr{C}(\mathbf{X})$ is satisfied since

$$\mathscr{C}(\mathbf{C}\cdot\mathbf{CX}) \subset \mathscr{C}(\mathbf{CX})\,. \tag{10.60}$$

Hence OLSE($\mathbf{X}\boldsymbol{\beta}$) equals BLUE($\mathbf{X}\boldsymbol{\beta}$) under $\mathscr{M}_c$.

Consider now $\mathbf{V}$ which has the intraclass correlation structure (of which the centering matrix $\mathbf{C}$ is an example), that is, $\mathbf{V}$ is of the type

$$\mathbf{V} = (1-\varrho)\mathbf{I}_n + \varrho\mathbf{1}_n\mathbf{1}_n' = \begin{pmatrix} 1 & \varrho & \dots & \varrho \\ \varrho & 1 & \dots & \varrho \\ \vdots & \vdots & \ddots & \vdots \\ \varrho & \varrho & \dots & 1 \end{pmatrix}, \tag{10.61}$$

where

$$-\frac{1}{n-1} \leq \varrho \leq 1\,. \tag{10.62}$$

Condition (10.62) guarantees the nonnegative definiteness of $\mathbf{V}$ because the eigenvalues of the intraclass correlation matrix are

$$\operatorname{ch}_1(\mathbf{V}) = 1 + (n-1)\varrho\,, \quad \operatorname{ch}_2(\mathbf{V}) = \dots = \operatorname{ch}_n(\mathbf{V}) = 1-\varrho\,; \tag{10.63}$$

see Proposition 18.2 (p. 366). In this situation

$$\mathbf{HV} = (1-\varrho)\mathbf{H} + \varrho\mathbf{H11}', \tag{10.64}$$

and thereby $\mathbf{HV} = \mathbf{VH}$ if and only if

$$\mathbf{H11}' = \mathbf{11}'\mathbf{H}\,, \quad \text{i.e.,} \quad \mathbf{HJ} = \mathbf{JH}\,, \quad \text{i.e.,} \quad \mathbf{P_X P_1} = \mathbf{P_1 P_X}\,. \tag{10.65}$$

We can now use the Exercise 8.20 (p. 190) to study the commutativity (10.65), or we can use Proposition 10.1 (p. 218) by replacing $\mathbf{V}$ with $\mathbf{H}$ and $\mathbf{H}$ with $\mathbf{J}$. Using the latter approach, we can conclude that (10.65) holds if and only if $\mathbf{1}$ is an eigenvector of $\mathbf{H}$, i.e.,

$$\mathbf{H1} = \lambda\mathbf{1}\,. \tag{10.66}$$

Because $\mathbf{H}$ is an orthogonal projector, its eigenvalues are 0 and 1, with multiplicities $n - \operatorname{rank}(\mathbf{X})$ and $\operatorname{rank}(\mathbf{X})$, respectively. Hence (10.66) holds if and only of

$$\mathbf{1} \in \mathscr{C}(\mathbf{X}) \quad \text{or} \quad \mathbf{1} \in \mathscr{C}(\mathbf{X})^{\perp}. \tag{10.67}$$

It is noteworthy that in Proposition 10.1 (p. 218), when characterizing the set of appropriate covariance matrices $\mathbf{V}$ that would yield the equality between $\mathrm{OLSE}(\mathbf{X}\boldsymbol{\beta})$ and $\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta})$, we keep the model matrix $\mathbf{X}$ fixed. There is, however, another approach, originating from McElroy (1967):

> *Problem:* Given a matrix $\mathbf{U}$ of $\mathrm{rank}(\mathbf{U}) = u \leq n-1$, characterize the set $\mathcal{W}$ of all nonnegative definite matrices $\mathbf{V}$ such that for every $\mathbf{X}$ satisfying
>
> $$\mathscr{C}(\mathbf{U}) \subset \mathscr{C}(\mathbf{X}), \tag{10.68}$$
>
> the equality $\mathrm{OLSE}(\mathbf{X}\boldsymbol{\beta}) = \mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta})$ holds uniformly with respect to $\mathbf{y} \in \mathscr{C}(\mathbf{X} : \mathbf{V})$.

Photograph 10.4 Thomas Mathew (Hyderabad, 2007).

McElroy considered the particular case when $\mathbf{U} = \mathbf{1}_n$, and showed that $\mathbf{V} \in \mathcal{W}$ if and only if

$$\mathbf{V} = a\mathbf{I}_n + b\mathbf{1}_n\mathbf{1}_n' \tag{10.69}$$

for such scalars a, b that $\mathbf{V}$ is nonnegative definite; in other words, $\mathbf{V}$ is a completely symmetric covariance matrix and thereby $\mathbf{V}$ has the intraclass correlation structure. A general condition was obtained by Zyskind (1969, Th. 1) and Balestra (1970, Th. 1). For a thorough discussion of this approach, see Baksalary & van Eijnsbergen (1988), who formulated the result as follows:

Proposition 10.4. *Under* $\{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{V}\}$, *let* $\mathbf{U}$ *be given. Then for every* $\mathbf{X}$ *satisfying* $\mathscr{C}(\mathbf{U}) \subset \mathscr{C}(\mathbf{X})$ *the equality* $\mathrm{OLSE}(\mathbf{X}\boldsymbol{\beta}) = \mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta})$ *holds for all* $\mathbf{y} \in \mathscr{C}(\mathbf{X} : \mathbf{V})$ *if and only if any of the following equivalent conditions holds:*

(a) $\mathbf{V}(\mathbf{I} - \mathbf{P}_{\mathbf{U}}) = \dfrac{\mathrm{tr}[\mathbf{V}(\mathbf{I} - \mathbf{P}_{\mathbf{U}})]}{n - \mathrm{rk}(\mathbf{U})}(\mathbf{I} - \mathbf{P}_{\mathbf{U}})$,

(b) $\mathbf{V}$ *can be expressed as* $\mathbf{V} = \alpha\mathbf{I} + \mathbf{U}\mathbf{A}\mathbf{U}'$, *where* $\alpha \in \mathbb{R}$, *and* $\mathbf{A}$ *is symmetric, such that* $\mathbf{V}$ *is nonnegative definite, i.e.,*

$$\mathbf{V} \in \mathcal{W} = \{\, \mathbf{V} \geq_{\mathsf{L}} \mathbf{0} : \mathbf{V} = \alpha\mathbf{I} + \mathbf{U}\mathbf{A}\mathbf{U}',\ \mathbf{A} = \mathbf{A}' \,\}\,.$$

Mathew (1983) considered the above problem in the context of two linear models $\{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}_1\}$ and $\{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}_2\}$. For interesting econometric-orienteered discussion on McElroy's result, see Larocca (2005) and Luskin (2008).

10.4 General Expressions for the BLUE

Using Theorem 11 (p. 267), we can get the following representations for the BLUE of $\mathbf{X}\boldsymbol{\beta}$.

Proposition 10.5. *The general solution for* $\mathbf{G}$ *satisfying*

$$\mathbf{G}(\mathbf{X} : \mathbf{VM}) = (\mathbf{X} : \mathbf{0}) \tag{10.70}$$

can be expressed, for example, in the following ways:

(a) $\mathbf{G}_1 = (\mathbf{X} : \mathbf{0})(\mathbf{X} : \mathbf{VM})^- + \mathbf{F}_1[\mathbf{I}_n - (\mathbf{X} : \mathbf{VM})(\mathbf{X} : \mathbf{VM})^-]$,

(b) $\mathbf{G}_2 = \mathbf{X}(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}'\mathbf{W}^- + \mathbf{F}_2(\mathbf{I}_n - \mathbf{WW}^-)$,

(c) $\mathbf{G}_3 = \mathbf{I}_n - \mathbf{VM}(\mathbf{MVM})^-\mathbf{M} + \mathbf{F}_3[\mathbf{I}_n - \mathbf{MVM}(\mathbf{MVM})^-]\mathbf{M}$,

(d) $\mathbf{G}_4 = \mathbf{H} - \mathbf{HVM}(\mathbf{MVM})^-\mathbf{M} + \mathbf{F}_4[\mathbf{I}_n - \mathbf{MVM}(\mathbf{MVM})^-]\mathbf{M}$,

where $\mathbf{F}_1$*,* $\mathbf{F}_2$*,* $\mathbf{F}_3$ *and* $\mathbf{F}_4$ *are arbitrary matrices,* $\mathbf{W} = \mathbf{V} + \mathbf{XUX}'$ *and* $\mathbf{U}$ *is any arbitrary matrix such that* $\mathscr{C}(\mathbf{W}) = \mathscr{C}(\mathbf{X} : \mathbf{V})$*.*

Proof. Above, each first term is a particular solution to (10.70) and the second term is the general solution to the corresponding homogeneous equation

$$\mathbf{G}(\mathbf{X} : \mathbf{VM}) = (\mathbf{0} : \mathbf{0})\,. \tag{10.71}$$

According to (11.5) (p. 267), the general solution to the consistent equation $\mathbf{YB} = \mathbf{C}$ is

$$\mathbf{Y} = \mathbf{CB}^- + \mathbf{Z}(\mathbf{I} - \mathbf{BB}^-)\,. \tag{10.72}$$

The representation (a) follows at once from (10.72).

The matrix $\mathbf{G}_2 = \mathbf{X}(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}'\mathbf{W}^-$ satisfies (10.70); see, e.g., (5.156) (p. 143), Proposition 12.1 (p. 286) and (12.34) (p. 288). It is easy to confirm that the first terms of $\mathbf{G}_3$ and $\mathbf{G}_4$ satisfy (10.70).

Because $\mathscr{C}(\mathbf{W}) = \mathscr{C}(\mathbf{X} : \mathbf{VM})$, (10.71) is equivalent to

$$\mathbf{GW} = \mathbf{0}\,, \tag{10.73}$$

for which the general solution is, cf. (11.10) (p. 268),

$$\mathbf{G} = \mathbf{F}(\mathbf{W}^{\perp})' = \mathbf{F}(\mathbf{I}_n - \mathbf{WW}^-)\,, \quad \text{where } \mathbf{F} \text{ is free to vary}\,, \tag{10.74}$$

or, equivalently,

$$\mathbf{G} = \mathbf{FT}\,, \quad \text{where } \mathscr{N}(\mathbf{T}) = \mathscr{C}(\mathbf{W}) \text{ and } \mathbf{F} \text{ is free to vary}\,. \tag{10.75}$$

In view of (5.119) (p. 138),

$$\mathscr{C}(\mathbf{X} : \mathbf{V})^{\perp} = \mathscr{C}\big(\mathbf{M}[\mathbf{I}_n - (\mathbf{MVM})^-\mathbf{MVM}]\big)\,, \tag{10.76a}$$

$$\mathscr{C}(\mathbf{X} : \mathbf{V}) = \mathscr{N}\big([\mathbf{I}_n - \mathbf{MVM}(\mathbf{MVM})^-]\mathbf{M}\big)\,, \tag{10.76b}$$

and hence we can choose $\mathbf{T}$ in (10.75) as

$$\mathbf{T} = [\mathbf{I}_n - \mathbf{MVM}(\mathbf{MVM})^-]\mathbf{M}\,. \tag{10.77}$$

□

For the relations between the representations of the BLUEs, see, e.g., Albert (1973), Baksalary & Kala (1978b), Pukelsheim (1977), Rao (1978), Rao & Mitra (1971a), and Watson (1972).

10.5 Multivariate Linear Model

Instead of one response variable y, consider d response variables $y_1, \dots, y_d$. Let the n observed values of these d variables, i.e., n observations from d-dimensional variable, be in the data matrix

$$\mathbf{Y} = (\mathbf{y}_1 : \mathbf{y}_2 : \dots : \mathbf{y}_d) = \begin{pmatrix} \mathbf{y}'_{(1)} \\ \vdots \\ \mathbf{y}'_{(n)} \end{pmatrix} \in \mathbb{R}^{n\times d}. \tag{10.78}$$

Denote

$$\mathbf{X} = (\mathbf{x}_1 : \mathbf{x}_2 : \dots : \mathbf{x}_p) = \begin{pmatrix} \mathbf{x}'_{(1)} \\ \vdots \\ \mathbf{x}'_{(n)} \end{pmatrix} \in \mathbb{R}^{n\times p}, \tag{10.79a}$$

$$\mathbb{B} = (\boldsymbol{\beta}_1 : \dots : \boldsymbol{\beta}_d) \in \mathbb{R}^{p\times d}. \tag{10.79b}$$

Assume now that

$$\mathbf{y}_j = \mathbf{X}\boldsymbol{\beta}_j + \boldsymbol{\varepsilon}_j\,, \quad j = 1, \dots, d\,, \tag{10.80a}$$

$$\operatorname{cov}(\mathbf{y}_j) = \operatorname{cov}(\boldsymbol{\varepsilon}_j) = \sigma_j^2\mathbf{I}_n\,, \quad \mathrm{E}(\boldsymbol{\varepsilon}_j) = \mathbf{0}\,, \quad j = 1, \dots, d\,, \tag{10.80b}$$

$$(\mathbf{y}_1 : \dots : \mathbf{y}_d) = \mathbf{X}(\boldsymbol{\beta}_1 : \dots : \boldsymbol{\beta}_d) + (\boldsymbol{\varepsilon}_1 : \dots : \boldsymbol{\varepsilon}_d)\,, \tag{10.80c}$$

$$\mathbf{Y} = \mathbf{X}\mathbb{B} + \boldsymbol{\mathcal{E}}, \quad \mathbf{Y} \in \mathbb{R}^{n\times d}, \quad \mathbf{X} \in \mathbb{R}^{n\times p}\,, \quad \mathbb{B} \in \mathbb{R}^{p\times d}. \tag{10.80d}$$

The columns of $\mathbf{Y}$, i.e., $\mathbf{y}_1, \dots, \mathbf{y}_d$, are n-dimensional random vectors such that

$$\mathbf{y}_j \sim (\mathbf{X}\boldsymbol{\beta}_i,\, \sigma_j^2\mathbf{I}_n)\,, \quad \operatorname{cov}(\mathbf{y}_j, \mathbf{y}_k) = \sigma_{jk}\mathbf{I}_n\,, \quad j, k = 1, \dots, d\,. \tag{10.81}$$

Transposing $\mathbf{Y} = \mathbf{X}\mathbb{B} + \boldsymbol{\mathcal{E}}$ we have

$$(\mathbf{y}_{(1)} : \dots : \mathbf{y}_{(n)}) = \mathbb{B}'(\mathbf{x}_{(1)} : \dots : \mathbf{x}_{(n)}) + (\boldsymbol{\varepsilon}_{(1)} : \dots : \boldsymbol{\varepsilon}_{(n)})\,, \tag{10.82}$$

where the (transposed) rows of $\mathbf{Y}$, i.e., $\mathbf{y}_{(1)}, \dots, \mathbf{y}_{(n)}$, are independent d-dimensional random vectors such that $\mathbf{y}_{(i)}$ has expectation $\mathbb{B}'\mathbf{x}_{(i)}$ and covariance matrix $\mathbf{\Sigma}$; in a short notation

$$\mathbf{y}_{(i)} \sim (\mathbb{B}'\mathbf{x}_{(i)}, \mathbf{\Sigma}), \quad \operatorname{cov}(\mathbf{y}_{(i)}, \mathbf{y}_{(j)}) = \mathbf{0}, \quad i \neq j. \tag{10.83}$$

Notice that in this setup

♠ rows (observations) are independent but columns (variables) may correlate. (10.84)

We can denote this model shortly as a triplet

$$\mathscr{A} = \{\mathbf{Y}, \mathbf{X}\mathbb{B}, \mathbf{\Sigma}\}. \tag{10.85}$$

Denoting

$$\mathbf{y}_* = \operatorname{vec}(\mathbf{Y}) = \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_d \end{pmatrix}, \ \operatorname{E}(\mathbf{y}_*) = \begin{pmatrix} \mathbf{X} & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{X} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_d \end{pmatrix} = (\mathbf{I}_d \otimes \mathbf{X}) \operatorname{vec}(\mathbb{B}), \tag{10.86}$$

we get

$$\operatorname{cov}(\mathbf{y}_*) = \begin{pmatrix} \sigma_1^2 \mathbf{I}_n & \sigma_{12}\mathbf{I}_n & \dots & \sigma_{1d}\mathbf{I}_n \\ \vdots & \vdots & & \vdots \\ \sigma_{d1}\mathbf{I}_n & \sigma_{d2}\mathbf{I}_n & \dots & \sigma_d^2 \mathbf{I}_n \end{pmatrix} = \mathbf{\Sigma} \otimes \mathbf{I}_n := \mathbf{V}_*. \tag{10.87}$$

The multivariate model can be rewritten as a univariate model

$$\mathscr{B} = \{\operatorname{vec}(\mathbf{Y}), (\mathbf{I}_d \otimes \mathbf{X}) \operatorname{vec}(\mathbb{B}), \mathbf{\Sigma} \otimes \mathbf{I}_n\} := \{\mathbf{y}_*, \mathbf{X}_*\boldsymbol{\beta}_*, \mathbf{V}_*\}; \tag{10.88}$$

see, e.g., Searle (1978), Seber (1984, §8.1), and Christensen (2001, pp. 3–5). By the analogy of the univariate model, we can estimate $\mathbf{X}\mathbb{B}$ by minimizing

$$\begin{aligned} \|\mathbf{y}_* - \mathbf{X}_*\boldsymbol{\beta}_*\|^2 &= \|\mathbf{y}_1 - \mathbf{X}\boldsymbol{\beta}_1\|^2 + \dots + \|\mathbf{y}_d - \mathbf{X}\boldsymbol{\beta}_d\|^2 \\ &= \operatorname{tr}(\mathbf{Y} - \mathbf{X}\mathbb{B})'(\mathbf{Y} - \mathbf{X}\mathbb{B}) = \|\mathbf{Y} - \mathbf{X}\mathbb{B}\|_F^2. \end{aligned} \tag{10.89}$$

The resulting $\mathbf{X}\hat{\mathbb{B}}$ is the OLS estimator of $\mathbf{X}\mathbb{B}$ which is $\mathbf{X}\hat{\mathbb{B}} = \mathbf{H}\mathbf{Y}$, so that the ith column of $\hat{\mathbb{B}}$ is $\hat{\boldsymbol{\beta}}_i = (\mathbf{X}'\mathbf{X})^- \mathbf{X}'\mathbf{y}_i$. Obviously we have

$$\mathbf{P}_{\mathbf{X}_*} = \mathbf{P}_{\mathbf{I}_d \otimes \mathbf{X}} = \mathbf{I}_d \otimes \mathbf{H} = \begin{pmatrix} \mathbf{H} & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{H} \end{pmatrix}. \tag{10.90}$$

We may recall from Proposition 2.5 (p. 81) that for any $\mathbf{Y} \in \mathbb{R}^{n\times d}$, $\mathbf{X} \in \mathbb{R}^{n\times p}$, and $\mathbb{B} \in \mathbb{R}^{p\times d}$ we have the Löwner partial ordering

$$(\mathbf{Y} - \mathbf{X}\mathbb{B})'(\mathbf{Y} - \mathbf{X}\mathbb{B}) \geq_{\mathsf{L}} (\mathbf{Y} - \mathbf{H}\mathbf{Y})'(\mathbf{Y} - \mathbf{H}\mathbf{Y}) = \mathbf{Y}'\mathbf{M}\mathbf{Y}, \tag{10.91}$$

where the equality holds if and only if $\mathbf{X}\mathbf{B} := \mathbf{X}\hat{\mathbb{B}} = \mathbf{P}_{\mathbf{X}}\mathbf{Y} = \mathbf{H}\mathbf{Y}$. The matrix

$$(\mathbf{Y} - \mathbf{X}\hat{\mathbb{B}})'(\mathbf{Y} - \mathbf{X}\hat{\mathbb{B}}) = \mathbf{Y}'\mathbf{M}\mathbf{Y} = \mathbf{E}_{\text{res}} \tag{10.92}$$

is often called a residual matrix.

Because $\mathbf{V}_*$ is not necessarily a multiple of the identity matrix it is not fully trivial that the OLSE of $\mathbf{X}_*\boldsymbol{\beta}_*$ would give the BLUE under the model $\mathscr{B} = \{\mathbf{y}_*, \mathbf{X}_*\boldsymbol{\beta}_*, \mathbf{V}_*\}$. However, the equality

$$\text{OLSE}(\mathbf{X}_*\boldsymbol{\beta}_*) = \text{BLUE}(\mathbf{X}_*\boldsymbol{\beta}_*) \tag{10.93}$$

follows from the commutativity (please confirm!)

$$\mathbf{V}_*\mathbf{P}_{\mathbf{X}_*} = \mathbf{P}_{\mathbf{X}_*}\mathbf{V}_* \,. \tag{10.94}$$

An alternative way to confirm (10.93) is to show, see Christensen (2001, p. 5), that

$$\mathscr{C}[(\boldsymbol{\Sigma} \otimes \mathbf{I}_n)(\mathbf{I}_d \otimes \mathbf{X})] = \mathscr{C}(\boldsymbol{\Sigma} \otimes \mathbf{X}) \subset \mathscr{C}(\mathbf{I}_d \otimes \mathbf{X}) \,. \tag{10.95}$$

We have observed that in multivariate model defined as above, the OLSE (being actually the BLUE) for $\mathbf{X}\mathbb{B}$ can be calculated simply as the OLSE from each individual model. Hence in that sense nothing is gained by this setup—for example the correlations between the y_i-variables have no effect on the estimates.

Under multinormality (and $\mathbf{X}$ having full column rank) the maximum likelihood estimates are

$$\text{MLE}(\mathbb{B}) = \hat{\mathbb{B}}, \quad \text{MLE}(\boldsymbol{\Sigma}) = \tfrac{1}{n}\mathbf{E}_{\text{res}} \,. \tag{10.96}$$

Under normality the random matrix $\mathbf{E}_{\text{res}} = \mathbf{Y}'\mathbf{M}\mathbf{Y}$ follows the Wishart distribution, $\mathbf{E}_{\text{res}} \sim \text{Wishart}_d(n-r, \boldsymbol{\Sigma})$, where $r = \text{rank}(\mathbf{X})$; see p. 26. We might be interested in testing a linear hypothesis $H\colon \mathbf{K}'\mathbb{B} = \mathbf{D}$, where $\mathbf{K} \in \mathbb{R}^{p\times q}$, $\text{rank}(\mathbf{K}) = q$, and $\mathbf{D} \in \mathbb{R}^{q\times d}$ are given matrices such that $\mathbf{K}'\mathbb{B}$ is estimable, i.e., $\mathscr{C}(\mathbf{K}) \subset \mathscr{C}(\mathbf{X}')$. Then it can be shown, cf. (8.158) (p. 177), that the minimum $\mathbf{E}_H$, say, of

$$(\mathbf{Y} - \mathbf{X}\mathbb{B})'(\mathbf{Y} - \mathbf{X}\mathbb{B}) \quad \text{subject to} \quad \mathbf{K}'\mathbb{B} = \mathbf{D} \tag{10.97}$$

occurs (in the Löwner sense) when $\mathbf{X}\mathbb{B}$ equals

$$\mathbf{X}\hat{\mathbb{B}}_H = \mathbf{X}\hat{\mathbb{B}} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{K}[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{K}]^{-1}(\mathbf{K}'\hat{\mathbb{B}} - \mathbf{D}) \,, \tag{10.98}$$

and so we may denote

$$\mathbf{E}_H - \mathbf{E}_{\text{res}} = (\mathbf{K}'\hat{\mathbb{B}} - \mathbf{D})'[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{K}]^{-1}(\mathbf{K}'\hat{\mathbb{B}} - \mathbf{D})\,. \tag{10.99}$$

As mentioned in Section 8.9 (p. 175), the standard test statistics for univariate models is proportional to the ratio

$$\frac{\text{SSE}_H - \text{SSE}}{\text{SSE}}\,, \tag{10.100}$$

where $\text{SSE} = \mathbf{y}'(\mathbf{I}_n - \mathbf{H})\mathbf{y}$, $\text{SSE}_H = \min_H \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$ and H is the hypothesis under consideration. As pointed out by Christensen (2001, p. 10), corresponding to (10.100), it is natural to base the hypothesis testing in multivariate model on some appropriate function of the matrix

$$(\mathbf{E}_H - \mathbf{E}_{\text{res}})\mathbf{E}_{\text{res}}^{-1}\,, \tag{10.101}$$

or on some closely related matrix. For further references, see, e.g., Christensen (2001, §1.2.1), Muirhead (1982, Ch. 10), and Seber (1984, §8.6).

As a simple example we may consider two independent samples $\mathbf{Y}_1' = (\mathbf{y}_{(11)} : \ldots : \mathbf{y}_{(1n_1)})$ and $\mathbf{Y}_2' = (\mathbf{y}_{(21)} : \ldots : \mathbf{y}_{(2n_2)})$ from d-dimensional normal distribution so that each $\mathbf{y}_{(1i)} \sim \text{N}_d(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $\mathbf{y}_{(2i)} \sim \text{N}_d(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$. This situation can be written in the form of a multivariate linear model as

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{y}'_{(11)} \\ \vdots \\ \mathbf{y}'_{(1n_1)} \\ \mathbf{y}'_{(21)} \\ \vdots \\ \mathbf{y}'_{(2n_2)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}'_1 \\ \vdots \\ \boldsymbol{\mu}'_1 \\ \boldsymbol{\mu}'_2 \\ \vdots \\ \boldsymbol{\mu}'_2 \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}'_{(11)} \\ \vdots \\ \boldsymbol{\varepsilon}'_{(1n_1)} \\ \boldsymbol{\varepsilon}'_{(21)} \\ \vdots \\ \boldsymbol{\varepsilon}'_{(2n_2)} \end{pmatrix}, \tag{10.102}$$

that is,

$$\mathbf{Y} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_2} \end{pmatrix} \begin{pmatrix} \boldsymbol{\mu}'_1 \\ \boldsymbol{\mu}'_2 \end{pmatrix} + \boldsymbol{\mathcal{E}} = \mathbf{X}\mathbb{B} + \boldsymbol{\mathcal{E}}\,. \tag{10.103}$$

Now

$$\mathbf{H} = \begin{pmatrix} \mathbf{J}_{n_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_{n_2} \end{pmatrix}, \quad \mathbf{M} = \mathbf{I}_n - \mathbf{H} = \begin{pmatrix} \mathbf{C}_{n_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{n_2} \end{pmatrix}, \tag{10.104}$$

$$\mathbf{HY} = \begin{pmatrix} \mathbf{J}_{n_1}\mathbf{Y}_1 \\ \mathbf{J}_{n_2}\mathbf{Y}_2 \end{pmatrix} = \begin{pmatrix} \bar{y}_{11}\mathbf{1}_{n_1} : \bar{y}_{12}\mathbf{1}_{n_1} : \ldots : \bar{y}_{1d}\mathbf{1}_{n_1} \\ \bar{y}_{21}\mathbf{1}_{n_2} : \bar{y}_{22}\mathbf{1}_{n_2} : \ldots : \bar{y}_{2d}\mathbf{1}_{n_2} \end{pmatrix} = \begin{pmatrix} \mathbf{1}_{n_1}\bar{\mathbf{y}}'_1 \\ \mathbf{1}_{n_2}\bar{\mathbf{y}}'_2 \end{pmatrix}, \tag{10.105}$$

and

$$\hat{\mathbb{B}} = \begin{pmatrix} \bar{\mathbf{y}}'_1 \\ \bar{\mathbf{y}}'_2 \end{pmatrix}, \quad \mathbf{E}_{\text{res}} = \mathbf{Y}'(\mathbf{I}_n - \mathbf{H})\mathbf{Y} = \mathbf{Y}'_1\mathbf{C}_{n_1}\mathbf{Y}_1 + \mathbf{Y}'_2\mathbf{C}_{n_2}\mathbf{Y}_2\,, \tag{10.106}$$

where $\mathbf{C}_{n_1}$ and $\mathbf{C}_{n_2}$ are centering matrices. We could be interested in testing the hypothesis $H\colon \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, i.e., $(1,-1)\mathbb{B} = \boldsymbol{\mu}_1' - \boldsymbol{\mu}_2' = \mathbf{0}_d'$. Notice that in Exercise 8.11 (p. 186) we considered the corresponding test for univariate situation. Because

$$\min\,(\mathbf{Y} - \mathbf{X}\mathbb{B})'(\mathbf{Y} - \mathbf{X}\mathbb{B}) \quad \text{subject to} \quad \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \tag{10.107}$$

occurs (in the Löwner sense) when $\mathbf{X}\mathbb{B}$ equals $\mathbf{J}_n\mathbf{Y}$, where

$$\mathbf{J}_n\mathbf{Y} = \mathbf{1}_n\bar{\mathbf{y}}' = \begin{pmatrix} \mathbf{1}_{n_1}\bar{\mathbf{y}}' \\ \mathbf{1}_{n_2}\bar{\mathbf{y}}' \end{pmatrix}, \tag{10.108}$$

we have (check the details)

$$\begin{aligned} \mathbf{E}_H - \mathbf{E}_{\text{res}} &= \mathbf{Y}'(\mathbf{I}_n - \mathbf{J}_n)\mathbf{Y} - \mathbf{Y}'(\mathbf{I}_n - \mathbf{H})\mathbf{Y} \\ &= \mathbf{Y}'(\mathbf{H} - \mathbf{J}_n)\mathbf{Y} = [(\mathbf{H} - \mathbf{J}_n)\mathbf{Y}]'(\mathbf{H} - \mathbf{J}_n)\mathbf{Y} \\ &= n_1(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}})(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}})' + n_2(\bar{\mathbf{y}}_2 - \bar{\mathbf{y}})(\bar{\mathbf{y}}_2 - \bar{\mathbf{y}})' \\ &= \frac{n_1 n_2}{n_1 + n_2}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)'. \end{aligned} \tag{10.109}$$

An alternative simple way to introduce (10.109) is to use (10.99) (p. 232). If we have g groups, we apparently have

$$\begin{aligned} \mathbf{E}_H - \mathbf{E}_{\text{res}} &= \mathbf{Y}'(\mathbf{I}_n - \mathbf{J}_n)\mathbf{Y} - \mathbf{Y}'(\mathbf{I}_n - \mathbf{H})\mathbf{Y} \\ &= \mathbf{Y}'(\mathbf{H} - \mathbf{J}_n)\mathbf{Y} \\ &= n_1(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}})(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}})' + \cdots + n_g(\bar{\mathbf{y}}_g - \bar{\mathbf{y}})(\bar{\mathbf{y}}_g - \bar{\mathbf{y}})'. \end{aligned} \tag{10.110}$$

Following the notation at the end of Exercise 8.11 (p. 186), we can denote

$$\mathbf{E}_H - \mathbf{E}_{\text{res}} = \mathbf{Y}'(\mathbf{H} - \mathbf{J}_n)\mathbf{Y} = \sum_{i=1}^{g} n_i(\bar{\mathbf{y}}_i - \bar{\mathbf{y}})(\bar{\mathbf{y}}_i - \bar{\mathbf{y}})' = \mathbf{E}_{\text{Between}}\,, \tag{10.111}$$

$$\mathbf{E}_{\text{res}} = \mathbf{Y}'(\mathbf{I}_n - \mathbf{H})\mathbf{Y} = \sum_{i=1}^{g} \mathbf{Y}_i'\mathbf{C}_{n_i}\mathbf{Y}_i = \mathbf{E}_{\text{Within}}\,. \tag{10.112}$$

The test statistics for the hypothesis $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ can be be based, e.g., on the

$$\begin{aligned} \operatorname{ch}_1\big[(\mathbf{E}_H - \mathbf{E}_{\text{res}})\mathbf{E}_{\text{res}}^{-1}\big] &= \operatorname{ch}_1\Big[\frac{n_1 n_2}{n_1 + n_2}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)'\mathbf{E}_{\text{res}}^{-1}\Big] \\ &= \alpha(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)'\mathbf{S}_{\#}^{-1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)\,, \end{aligned} \tag{10.113}$$

where $\mathbf{S}_{\#} = \frac{1}{n_1+n_2-2}\mathbf{E}_{\text{res}}$ and $\alpha = \frac{(n_1+n_2-2)n_1n_2}{n_1+n_2}$. Hence one appropriate test statistics appears to be a function of the squared Mahalanobis distance

$$\operatorname{MHLN}^2(\bar{\mathbf{y}}_1, \bar{\mathbf{y}}_2, \mathbf{S}_{\#}) = (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)'\mathbf{S}_{\#}^{-1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)\,; \tag{10.114}$$

see also (18.120) (p. 374). The statistics

$$\frac{n_1 n_2}{n_1 + n_2}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)'\mathbf{S}_{\#}^{-1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2) = T^2 \tag{10.115}$$

is the famous Hotelling's T^2, see page 26, and Exercise 0.25 (p. 53). For further details regarding the two-sample multivariate linear model, see, e.g., Seber (1984, §8.6.4) and Christensen (2001, §1.4). For the univariate version of (10.115) see Exercise 8.12 (p. 187).

A more general multivariate linear model where $\operatorname{cov}(\mathbf{y}_*) = \boldsymbol{\Sigma} \otimes \mathbf{V}$, with $\boldsymbol{\Sigma}$ and $\mathbf{V}$ possibly singular, is considered by Sengupta & Jammalamadaka (2003, Ch. 10).

10.6 Random Sample without Replacement

Let $y_1, \dots, y_n$ be a simple random sample from the digits $\{1, 2, \dots, N\}$, where $n \leq N$, (a) with replacement, (b) without a replacement. Express the observations as a linear model

$$y_i = \mu + \varepsilon_i\,, \quad i = 1, \dots, n\,. \tag{10.116}$$

Find the covariance matrix of $\mathbf{y}$. What is, of course, the true value of μ? Find OLSE(μ) and BLUE(μ). Prove that they are equal. What happens if the population is $\{a_1, a_2, \dots, a_N\}$?

The answers to the above questions follow from the fact that the random vector $\mathbf{y}$ has an intraclass correlation structure, and thereby OLSE equals BLUE. For a reference, see Zyskind (1969, p. 1361). See also Exercise 0.24 (p. 53),

10.7 Efficiency of the OLSE

Consider a simple linear model $\mathscr{M} = \{\mathbf{y},\, \mathbf{x}\beta,\, \sigma^2\mathbf{V}\}$, where $\mathbf{V}$ is positive definite, and, of course, $\mathbf{x} \neq \mathbf{0}$. Then

$$\hat{\beta} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y}\,, \qquad \operatorname{var}(\hat{\beta}) = (\mathbf{x}'\mathbf{x})^{-2}\mathbf{x}'\mathbf{V}\mathbf{x}\,, \tag{10.117a}$$

$$\tilde{\beta} = (\mathbf{x}'\mathbf{V}^{-1}\mathbf{x})^{-1}\mathbf{x}'\mathbf{V}^{-1}\mathbf{y}\,, \qquad \operatorname{var}(\tilde{\beta}) = (\mathbf{x}'\mathbf{V}^{-1}\mathbf{x})^{-1}, \tag{10.117b}$$

and $\operatorname{var}(\tilde{\beta}) \leq \operatorname{var}(\hat{\beta})$, i.e.,

$$(\mathbf{x}'\mathbf{V}^{-1}\mathbf{x})^{-1} \leq (\mathbf{x}'\mathbf{x})^{-2}\mathbf{x}'\mathbf{V}\mathbf{x} \quad \text{for all } \mathbf{x} \in \mathbb{R}^n. \tag{10.118}$$

Then we know that

$$\hat{\beta} = \tilde{\beta} \iff \mathbf{V}\mathbf{x} = \lambda\mathbf{x} \quad \text{for some } \mathbf{x} \in \mathbb{R}^n,\ \lambda \in \mathbb{R}\,, \tag{10.119}$$

which means that $\mathbf{x}$ is an eigenvector of $\mathbf{V}$. In this situation $\hat{\beta}$ is "as good as it can get", and

$$\operatorname{var}(\hat{\beta}) = \operatorname{var}(\tilde{\beta}) = (\mathbf{x}'\mathbf{V}^{-1}\mathbf{x})^{-1} = (\mathbf{x}'\mathbf{x})^{-2}\mathbf{x}'\mathbf{V}\mathbf{x}\,. \tag{10.120}$$

Notice in passing that

$$\operatorname{cov}(\hat{\beta}, \tilde{\beta}) = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}' \cdot \mathbf{V} \cdot \mathbf{V}^{-1}\mathbf{x}(\mathbf{x}'\mathbf{V}^{-1}\mathbf{x})^{-1} = \operatorname{var}(\tilde{\beta})\,, \tag{10.121}$$

and hence

$$\begin{aligned}\operatorname{cov}\begin{pmatrix}\hat{\beta}\\ \tilde{\beta}\end{pmatrix} &= \begin{pmatrix}\operatorname{var}(\hat{\beta}) & \operatorname{cov}(\hat{\beta},\tilde{\beta})\\ \operatorname{cov}(\tilde{\beta},\hat{\beta}) & \operatorname{var}(\tilde{\beta})\end{pmatrix}\\ &= \begin{pmatrix}(\mathbf{x}'\mathbf{x})^{-2}\mathbf{x}'\mathbf{V}\mathbf{x} & (\mathbf{x}'\mathbf{V}^{-1}\mathbf{x})^{-1}\\ (\mathbf{x}'\mathbf{V}^{-1}\mathbf{x})^{-1} & (\mathbf{x}'\mathbf{V}^{-1}\mathbf{x})^{-1}\end{pmatrix} = \begin{pmatrix}\operatorname{var}(\hat{\beta}) & \operatorname{var}(\tilde{\beta})\\ \operatorname{var}(\tilde{\beta}) & \operatorname{var}(\tilde{\beta})\end{pmatrix}, \end{aligned} \tag{10.122}$$

and thereby

$$\begin{aligned}\operatorname{var}(\hat{\beta} - \tilde{\beta}) &= \operatorname{var}(\hat{\beta}) + \operatorname{var}(\tilde{\beta}) - 2\operatorname{cov}(\hat{\beta},\tilde{\beta})\\ &= \operatorname{var}(\hat{\beta}) + \operatorname{var}(\tilde{\beta}) - 2\operatorname{var}(\tilde{\beta}) = \operatorname{var}(\hat{\beta}) - \operatorname{var}(\tilde{\beta})\,. \end{aligned} \tag{10.123}$$

It is natural to ask how "bad" the OLSE could be with respect to the BLUE. One obvious measure for this is the ratio of the variances:

$$\phi = \operatorname{eff}(\hat{\beta}) = \frac{\operatorname{var}(\tilde{\beta})}{\operatorname{var}(\hat{\beta})} = \frac{(\mathbf{x}'\mathbf{V}^{-1}\mathbf{x})^{-1}}{(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{V}\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}} = \frac{(\mathbf{x}'\mathbf{x})^2}{\mathbf{x}'\mathbf{V}\mathbf{x} \cdot \mathbf{x}'\mathbf{V}^{-1}\mathbf{x}}\,. \tag{10.124}$$

Clearly we have

$$0 < \operatorname{eff}(\hat{\beta}) \leq 1\,, \tag{10.125}$$

where the upper bound is obtained if and only if OLSE equals BLUE. The lower bound of ϕ is obtained from the Kantorovich inequality; see Section 20.3 (p. 418) and, e.g., Watson, Alpargu & Styan (1997), Alpargu, Drury & Styan (1997), Alpargu & Styan (2000):

$$\tau_1^2 := \frac{4\lambda_1\lambda_n}{(\lambda_1 + \lambda_n)^2} \leq \frac{(\mathbf{x}'\mathbf{x})^2}{\mathbf{x}'\mathbf{V}\mathbf{x} \cdot \mathbf{x}'\mathbf{V}^{-1}\mathbf{x}} = \operatorname{eff}(\hat{\beta}) = \phi\,, \tag{10.126}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$ are the eigenvalues of $\mathbf{V}$. The lower bound is obtained when $\mathbf{x}$ is proportional either to $\mathbf{t}_1 + \mathbf{t}_n$ or to $\mathbf{t}_1 - \mathbf{t}_n$; in short, $\mathbf{x}$ is proportional to

$$\mathbf{x}_{\text{bad}} = \mathbf{t}_1 \pm \mathbf{t}_n = \mathbf{T}(\mathbf{i}_1 \pm \mathbf{i}_n) = \mathbf{T}\begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ \pm 1 \end{pmatrix}, \tag{10.127}$$

where $\mathbf{i}_j$ is the jth column of $\mathbf{I}_n$ and $\mathbf{T} = (\mathbf{t}_1 : \mathbf{t}_2 : \ldots : \mathbf{t}_n)$ is the matrix with $\mathbf{t}_i$ being the orthonormal eigenvectors of $\mathbf{V}$ corresponding to the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$.

If we consider a weakly singular linear model, i.e.,

$$\mathscr{M} = \{\mathbf{y}, \mathbf{x}\beta, \mathbf{V}\}, \quad \text{where } \mathbf{x} \in \mathscr{C}(\mathbf{V}), \tag{10.128}$$

then, see, e.g., Proposition 6.1 (p. 147),

$$\tilde{\beta} = (\mathbf{x}'\mathbf{V}^{+}\mathbf{x})^{-1}\mathbf{x}'\mathbf{V}^{+}\mathbf{y}, \quad \operatorname{var}(\tilde{\beta}) = (\mathbf{x}'\mathbf{V}^{+}\mathbf{x})^{-1}, \tag{10.129}$$

and

$$\tau_1^2 = \frac{4\lambda_1\lambda_v}{(\lambda_1 + \lambda_v)^2} \leq \frac{(\mathbf{x}'\mathbf{x})^2}{\mathbf{x}'\mathbf{V}\mathbf{x} \cdot \mathbf{x}'\mathbf{V}^{+}\mathbf{x}} = \operatorname{eff}(\hat{\beta}) = \phi, \tag{10.130}$$

where $\operatorname{rank}(\mathbf{V}) = v$; see also Chu, Isotalo, Puntanen & Styan (2004). The lower bound is obtained when $\mathbf{x}$ is proportional to

$$\mathbf{x}_{\text{bad}} = \mathbf{t}_1 \pm \mathbf{t}_v . \tag{10.131}$$

There is one interesting interpretation of the efficiency that deserves special attention. Letting $\mathbf{V}^{1/2}$ denote the symmetric positive definite square root of $\mathbf{V}$, we observe that ϕ is a specific squared cosine:

$$\begin{aligned}\phi = \frac{(\mathbf{x}'\mathbf{x})^2}{\mathbf{x}'\mathbf{V}\mathbf{x} \cdot \mathbf{x}'\mathbf{V}^{-1}\mathbf{x}} &= \frac{(\mathbf{x}'\mathbf{V}^{-1/2} \cdot \mathbf{V} \cdot \mathbf{V}^{-1/2}\mathbf{x})^2}{\mathbf{x}'\mathbf{V}^{-1/2}\mathbf{V}^2\mathbf{V}^{-1/2}\mathbf{x} \cdot \mathbf{x}'\mathbf{V}^{-1/2}\mathbf{V}^{-1/2}\mathbf{x}} \\ &= \frac{(\mathbf{z}'\mathbf{V}\mathbf{z})^2}{\mathbf{z}'\mathbf{V}^2\mathbf{z} \cdot \mathbf{z}'\mathbf{z}} = \cos^2(\mathbf{V}\mathbf{z}, \mathbf{z}) := \cos^2\alpha, \end{aligned}\tag{10.132}$$

where

$$\mathbf{z} = \mathbf{V}^{-1/2}\mathbf{x}, \quad \mathbf{x} = \mathbf{V}^{1/2}\mathbf{z}. \tag{10.133}$$

It is clear that the upper bound for ϕ is 1, and it is obtained when the angle α between the vectors $\mathbf{V}\mathbf{z}$ and $\mathbf{z}$ is minimized (zero); this happens exactly when $\mathbf{z}$ is an eigenvector of $\mathbf{V}$, that is,

$$\mathbf{V}\mathbf{z} = \lambda\mathbf{z}, \quad \text{i.e.,} \quad \mathbf{V}\mathbf{x} = \lambda\mathbf{x} \quad \text{for some } \lambda \in \mathbb{R}. \tag{10.134}$$

On the other hand, we may look for a vector $\mathbf{z}$ such that $\phi = \cos^2(\mathbf{V}\mathbf{z}, \mathbf{z})$ is minimized, i.e., the angle α between the vectors $\mathbf{V}\mathbf{z}$ and $\mathbf{z}$ is maximized. Since an eigenvector of $\mathbf{V}$ solves the maximization of ϕ, a vector $\mathbf{z}$ solving the minimization of ϕ may be called an *antieigenvector* and the corresponding minimum an *antieigenvalue* (squared). Therefore, the first (smallest) antieigenvalue of $\mathbf{V}$ is

$$\tau_1 = \min_{\mathbf{z}\neq\mathbf{0}} \cos^2(\mathbf{Vz},\mathbf{z}) = \frac{2\sqrt{\lambda_1\lambda_n}}{\lambda_1+\lambda_n} = \frac{\sqrt{\lambda_1\lambda_n}}{(\lambda_1+\lambda_n)/2}\,; \tag{10.135}$$

the ratio of the geometric and arithmetic means of λ_1 and λ_n. The corresponding first antieigenvectors are

$$\begin{aligned}\mathbf{z}_1 &:= \mathbf{V}^{-1/2}\mathbf{x}_{\text{bad}} = \mathbf{T}\boldsymbol{\Lambda}^{-1/2}\mathbf{T}'\cdot\mathbf{T}(\mathbf{i}_1\pm\mathbf{i}_n)\\ &= \mathbf{T}\left(\frac{1}{\sqrt{\lambda_1}}\mathbf{i}_1 \pm \frac{1}{\sqrt{\lambda_n}}\mathbf{i}_n\right) = \frac{1}{\sqrt{\lambda_1}}\mathbf{t}_1 \pm \frac{1}{\sqrt{\lambda_n}}\mathbf{t}_n\,,\end{aligned} \tag{10.136}$$

which can be scaled to have unit lengths and so

$$\mathbf{z}_1 = \frac{\sqrt{\lambda_n}}{\sqrt{\lambda_1+\lambda_n}}\mathbf{t}_1 \pm \frac{\sqrt{\lambda_1}}{\sqrt{\lambda_1+\lambda_n}}\mathbf{t}_n\,. \tag{10.137}$$

The first (smallest) antieigenvalue of $\mathbf{V}$ is also called the cosine of the matrix $\mathbf{V}$, $\cos(\mathbf{V})$.

The second antieigenvalue of $\mathbf{V}$ is defined as

$$\tau_2 = \min_{\mathbf{z}\neq\mathbf{0},\,\mathbf{z}'\mathbf{z}_1=0} \cos^2(\mathbf{Vz},\mathbf{z}) = \frac{2\sqrt{\lambda_2\lambda_{n-1}}}{\lambda_2+\lambda_{n-1}} = \frac{\sqrt{\lambda_2\lambda_{n-1}}}{(\lambda_2+\lambda_{n-1})/2}\,, \tag{10.138}$$

and the corresponding antieigenvectors are

$$\mathbf{z}_2 = \mathbf{T}\left(\frac{1}{\sqrt{\lambda_2}}\mathbf{i}_2 \pm \frac{1}{\sqrt{\lambda_{n-1}}}\mathbf{i}_{n-1}\right) = \frac{1}{\sqrt{\lambda_2}}\mathbf{t}_2 \pm \frac{1}{\sqrt{\lambda_{n-1}}}\mathbf{t}_{n-1}\,. \tag{10.139}$$

The concept of antieigenvalue was introduced by Gustafson (1972). For recent papers in the antieigenvalues, see, e.g., Gustafson (1994, 1999, 2000, 2002, 2005, 2006, 2007, 2011), Gustafson & Rao (1997), Khattree (2001, 2002, 2003), Rao (2007), and Seddighin (2009).

The Kantorovich inequality yields also matrix versions. Such a possibility becomes natural when we note that if we assume that $\mathbf{x}'\mathbf{x} = 1$, then (10.126) (p. 235) becomes

$$\frac{4\lambda_1\lambda_n}{(\lambda_1+\lambda_n)^2}\mathbf{x}'\mathbf{V}\mathbf{x} \leq (\mathbf{x}'\mathbf{V}^{-1}\mathbf{x})^{-1} \leq \mathbf{x}'\mathbf{V}\mathbf{x}\,, \tag{10.140a}$$

$$(\mathbf{x}'\mathbf{V}^{-1}\mathbf{x})^{-1} \leq \mathbf{x}'\mathbf{V}\mathbf{x} \leq \frac{(\lambda_1+\lambda_n)^2}{4\lambda_1\lambda_n}(\mathbf{x}'\mathbf{V}^{-1}\mathbf{x})^{-1}. \tag{10.140b}$$

Marshall & Olkin (1990) gave a matrix version of (10.140b) as

$$(\mathbf{U}'\mathbf{V}^{-1}\mathbf{U})^{-1} \leq_{\mathsf{L}} \mathbf{U}'\mathbf{V}\mathbf{U} \leq_{\mathsf{L}} \frac{(\lambda_1 + \lambda_n)^2}{4\lambda_1\lambda_n}(\mathbf{U}'\mathbf{V}^{-1}\mathbf{U})^{-1}, \tag{10.141}$$

where $\mathbf{U}'\mathbf{U} = \mathbf{I}_p$; see also Magness & McGuire (1962). If we substitute $\mathbf{U} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1/2}$ in (10.141), where $\mathbf{X} \in \mathbb{R}^{n\times p}$ has full column rank, then (10.141) becomes

$$(\mathbf{X}'\mathbf{X})^{-1/2}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1/2} \leq_{\mathsf{L}} \frac{(\lambda_1 + \lambda_n)^2}{4\lambda_1\lambda_n}(\mathbf{X}'\mathbf{X})^{1/2}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})^{1/2}, \tag{10.142}$$

or equivalently

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \leq_{\mathsf{L}} \frac{(\lambda_1 + \lambda_n)^2}{4\lambda_1\lambda_n}\,(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}, \tag{10.143a}$$

$$\operatorname{cov}(\tilde{\boldsymbol{\beta}}) \leq_{\mathsf{L}} \operatorname{cov}(\hat{\boldsymbol{\beta}}) \leq_{\mathsf{L}} \frac{(\lambda_1 + \lambda_n)^2}{4\lambda_1\lambda_n}\,\operatorname{cov}(\tilde{\boldsymbol{\beta}})\,. \tag{10.143b}$$

Further generalizations of the matrix inequalities of the type (10.141) appear in Baksalary & Puntanen (1991), Liu (2000a,b), Liu & Neudecker (1995, 1997), Pečarić, Puntanen & Styan (1996), and Drury, Liu, Lu, Puntanen et al. (2002).

One frequently used measure for the efficiency of the OLSE($\boldsymbol{\beta}$) in the general situation is the "Watson efficiency" (Watson 1951, §3.3, Watson 1955, p. 330) which is defined as the ratio of the determinants of the corresponding covariance matrices:

$$\begin{aligned}\phi = \operatorname{eff}(\hat{\boldsymbol{\beta}}) = \frac{|\operatorname{cov}(\tilde{\boldsymbol{\beta}})|}{|\operatorname{cov}(\hat{\boldsymbol{\beta}})|} &= \frac{|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}|^{-1}}{|(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}|} \\ &= \frac{|\mathbf{X}'\mathbf{X}|^2}{|\mathbf{X}'\mathbf{V}\mathbf{X}| \cdot |\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}|}\,.\end{aligned} \tag{10.144}$$

Clearly we have again

$$0 < \operatorname{eff}(\hat{\boldsymbol{\beta}}) \leq 1\,, \tag{10.145}$$

and the upper bound is obtained if and only if OLSE equals BLUE.

We recall that Bloomfield & Watson (1975) and Knott (1975) proved the following inequality which we shortly call the BWK inequality:

$$\phi \geq \frac{4\lambda_1\lambda_n}{(\lambda_1 + \lambda_n)^2} \cdot \frac{4\lambda_2\lambda_{n-1}}{(\lambda_2 + \lambda_{n-1})^2} \cdots \frac{4\lambda_p\lambda_{n-p+1}}{(\lambda_p + \lambda_{n-p+1})^2} = \tau_1^2\,\tau_2^2 \cdots \tau_p^2\,, \tag{10.146}$$

i.e.,

$$\min_{\mathbf{X}} \phi = \prod_{i=1}^{p} \frac{4\lambda_i\lambda_{n-i+1}}{(\lambda_i + \lambda_{n-i+1})^2} = \prod_{i=1}^{p} \tau_i^2, \tag{10.147}$$

where $\lambda_i = \operatorname{ch}_i(\mathbf{V})$, and $\tau_i = i$th antieigenvalue of $\mathbf{V}$. Assuming that $p \leq n/2$, the minimum of ϕ is attained when the model matrix $\mathbf{X}$ is chosen as

$$\mathbf{X}_{\text{bad}} = (\mathbf{t}_1 \pm \mathbf{t}_n : \mathbf{t}_2 \pm \mathbf{t}_{n-1} : \dots : \mathbf{t}_p \pm \mathbf{t}_{n-p+1}) := \mathbf{TL}\,, \tag{10.148}$$

where, when $p = 3$,

$$\mathbf{L} = (\mathbf{i}_1 \pm \mathbf{i}_n : \mathbf{i}_2 \pm \mathbf{i}_{n-1} : \mathbf{i}_3 \pm \mathbf{i}_{n-2}) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & \pm 1 \\ 0 & \pm 1 & 0 \\ \pm 1 & 0 & 0 \end{pmatrix}. \tag{10.149}$$

With this choice (and for $p = 3$) we get

$$\mathbf{X}'\mathbf{X} = 2\mathbf{I}_3\,, \tag{10.150a}$$

$$\mathbf{X}'\mathbf{V}\mathbf{X} = \mathbf{L}'\boldsymbol{\Lambda}\mathbf{L} = \begin{pmatrix} \lambda_1 + \lambda_n & 0 & 0 \\ 0 & \lambda_2 + \lambda_{n-1} & 0 \\ 0 & 0 & \lambda_3 + \lambda_{n-2} \end{pmatrix}, \tag{10.150b}$$

$$\begin{aligned} \mathbf{X}'\mathbf{V}^{-1}\mathbf{X} = \mathbf{L}'\boldsymbol{\Lambda}^{-1}\mathbf{L} &= \begin{pmatrix} \frac{1}{\lambda_1} + \frac{1}{\lambda_n} & 0 & 0 \\ 0 & \frac{1}{\lambda_2} + \frac{1}{\lambda_{n-1}} & 0 \\ 0 & 0 & \frac{1}{\lambda_3} + \frac{1}{\lambda_{n-2}} \end{pmatrix} \\ &= \begin{pmatrix} \frac{\lambda_1+\lambda_n}{\lambda_1\lambda_n} & 0 & 0 \\ 0 & \frac{\lambda_2+\lambda_{n-1}}{\lambda_2\lambda_{n-1}} & 0 \\ 0 & 0 & \frac{\lambda_3+\lambda_{n-2}}{\lambda_3\lambda_{n-2}} \end{pmatrix}. \end{aligned} \tag{10.150c}$$

If $p > n/2$, then only the first $m := n - p$ columns of $\mathbf{X}_{\text{bad}}$ can be constructed as above while the rest of the columns can be chosen from the set of "unused" eigenvectors $\mathbf{t}_{m+1}, \dots, \mathbf{t}_{n-m}$.

The inequality (10.146) was originally conjectured in 1955 by James Durbin (see Watson 1955, p. 331), but first established (for $p > 1$) only twenty years later by Bloomfield & Watson (1975) and Knott (1975). For further proofs (and related considerations), see Khatri & Rao (1981, 1982).

It is obvious that the Watson efficiency ϕ remains invariant if $\mathbf{X}$ is replaced with $\mathbf{XA}$, where $\mathbf{A}$ is an arbitrary $p \times p$ nonsingular matrix. Hence ϕ is a function of the column space $\mathscr{C}(\mathbf{X})$. We can choose, e.g., an orthonormal basis for $\mathscr{C}(\mathbf{X})$ and then the BWK inequality becomes

$$\min_{\mathbf{X}} \phi = \min_{\mathbf{X}'\mathbf{X}=\mathbf{I}_p} \frac{1}{|\mathbf{X}'\mathbf{V}\mathbf{X}| \cdot |\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}|} = \prod_{i=1}^{\min(p,n-p)} \frac{4\lambda_i\lambda_{n-i+1}}{(\lambda_i + \lambda_{n-i+1})^2} \,. \tag{10.151}$$

Another measure of efficiency of OLS, introduced by Bloomfield & Watson (1975), is based on the Euclidean size of the commutator $\mathbf{HV} - \mathbf{VH}$:

$$\begin{aligned}\psi &= \tfrac{1}{2}\|\mathbf{HV} - \mathbf{VH}\|_F^2 = \tfrac{1}{2}\operatorname{tr}(\mathbf{HV} - \mathbf{VH})(\mathbf{HV} - \mathbf{VH})' \\ &= \|\mathbf{HVM}\|_F^2 = \operatorname{tr}(\mathbf{HVMV}) = \operatorname{tr}(\mathbf{HV}^2) - \operatorname{tr}(\mathbf{HV})^2. \end{aligned} \tag{10.152}$$

Clearly $\psi_{12} = 0$ whenever OLSE equals BLUE. We will call ψ_{12} the “Bloomfield–Watson efficiency”. We immediately observe that in order to use (10.152), we do not need to impose rank assumptions on $\mathbf{X}$ and $\mathbf{V}$.

Assuming that $\mathbf{V}$ is positive definite, Bloomfield & Watson (1975) proved that the commutator criterion satisfies the inequality

$$\psi \leq \tfrac{1}{4}\sum_{i=1}^{p}(\lambda_i - \lambda_{n-i-1})^2, \tag{10.153}$$

and that the equality is attained in the same situation as the minimum of ϕ. It is easy to see that (10.153) holds also for a singular $\mathbf{V}$; replace $\mathbf{V}$ with a positive definite matrix $\mathbf{V} + a^2\mathbf{I}$, and the result follows.

Rao (1985a), studied the trace of the difference between the covariance matrices of the OLSE and BLUE of $\mathbf{X}\boldsymbol{\beta}$:

$$\eta = \operatorname{tr}[\operatorname{cov}(\mathbf{X}\hat{\boldsymbol{\beta}}) - \operatorname{cov}(\mathbf{X}\tilde{\boldsymbol{\beta}})] = \operatorname{tr}[\mathbf{HVH} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'] \,. \tag{10.154}$$

Just as ϕ and ψ, also η is a function of the column space $\mathscr{C}(\mathbf{X})$, and hence we can choose an orthonormal basis for $\mathscr{C}(\mathbf{X})$, in which case η becomes

$$\eta = \operatorname{tr}[\mathbf{X}'\mathbf{V}\mathbf{X} - (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}] \,. \tag{10.155}$$

Rao (1985a) showed that

$$\eta \leq \sum_{i=1}^{p}\left(\sqrt{\lambda_i} - \sqrt{\lambda_{n-i+1}}\right)^2 . \tag{10.156}$$

Styan (1983) considered (10.154) when $p = 1$ and obtained a version of (10.156). If we consider the general representations (10.154) of $\mathbf{X}\hat{\boldsymbol{\beta}}$ and $\mathbf{X}\tilde{\boldsymbol{\beta}}$, then we get

$$\operatorname{cov}(\mathbf{X}\hat{\boldsymbol{\beta}}) - \operatorname{cov}(\mathbf{X}\tilde{\boldsymbol{\beta}}) = \mathbf{HVM}(\mathbf{MVM})^{-}\mathbf{MVH} \,, \tag{10.157}$$

whose trace’s upper bound was considered by Liski, Puntanen & Wang (1992).

The efficiency problem raises a great amount of interesting matrix problems and has received a lot of attention in statistical literature; see Drury, Liu, Lu, Puntanen et al. (2002), and Chu, Isotalo, Puntanen & Styan (2007), and the references therein. For further measures and applications of the relative efficiency of the OLSE, see, for example, Baltagi (1989), Liu (2000a,b), Liu & King (2002), Krämer (1980a), Luati & Proietti (2010), Wang & Shao (1992), and Yang & Wang (2009). For the efficiency when $\mathbf{V}$ has autocorrelation structure, see Chipman (1979), Krämer (1980a, 1982, 1984, 1986), and Krämer & Donninger (1987). For the equality of OLSE and BLUE under specific model matrices, see Baksalary (1988) and Herzberg & Aleong (1985).

Photograph 10.5 Geoffrey S. Watson (Montréal, 1995).

For the efficiency of the subvector $\hat{\boldsymbol{\beta}}_2$, see Section 15.5 (p. 331). Efficiency of the OLSE was also in a central role in the Ph.D. theses of Chu (2004) and Isotalo (2007).

10.8 Efficiency and Canonical Correlations

In this section we consider an interesting relation between the Watson efficiency and specific canonical correlations.

When $\mathbf{X}$ has full column rank, then, in view of (5.107) (p. 137), the general expression for the covariance matrix of the BLUE of $\boldsymbol{\beta}$ can be written as

$$\begin{aligned} \operatorname{cov}(\tilde{\boldsymbol{\beta}}) &= (\mathbf{X}'\mathbf{X})^{-1}[\mathbf{X}'\mathbf{V}\mathbf{X} - \mathbf{X}'\mathbf{V}\mathbf{Z}(\mathbf{Z}'\mathbf{V}\mathbf{Z})^{-}\mathbf{Z}'\mathbf{V}\mathbf{X}](\mathbf{X}'\mathbf{X})^{-1} \\ &= \operatorname{cov}(\hat{\boldsymbol{\beta}}) - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{Z}(\mathbf{Z}'\mathbf{V}\mathbf{Z})^{-}\mathbf{Z}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \operatorname{cov}(\hat{\boldsymbol{\beta}}) - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^{-}\mathbf{M}\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &:= \operatorname{cov}(\hat{\boldsymbol{\beta}}) - \mathbf{D}\,, \end{aligned} \tag{10.158}$$

where $\mathbf{Z}$ is a matrix with the property $\mathscr{C}(\mathbf{Z}) = \mathscr{C}(\mathbf{M})$, i.e., $\mathbf{Z} \in \{\mathbf{X}^{\perp}\}$, and

$$\operatorname{cov}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{X}^{+}\mathbf{V}(\mathbf{X}^{+})', \tag{10.159a}$$

$$\operatorname{cov}(\hat{\boldsymbol{\beta}}) - \operatorname{cov}(\tilde{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{Z}(\mathbf{Z}'\mathbf{V}\mathbf{Z})^{-}\mathbf{Z}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}. \tag{10.159b}$$

It is interesting to note that in (10.158) the covariance matrix $\mathbf{V}$ need not be positive definite. If $\mathbf{V}$ is positive definite, then, of course,

$$\operatorname{cov}(\tilde{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} = \operatorname{cov}(\hat{\boldsymbol{\beta}}) - \mathbf{D}\,, \tag{10.160}$$

while under the weakly singular linear model, i.e., when $\mathscr{C}(\mathbf{X}) \subset \mathscr{C}(\mathbf{V})$,

$$\operatorname{cov}(\tilde{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{V}^{+}\mathbf{X})^{-1} = \operatorname{cov}(\hat{\boldsymbol{\beta}}) - \mathbf{D}\,. \tag{10.161}$$

Now we can express the Watson efficiency $\phi = \operatorname{eff}(\hat{\boldsymbol{\beta}})$ as

$$\begin{aligned}\phi = \frac{|\operatorname{cov}(\tilde{\boldsymbol{\beta}})|}{|\operatorname{cov}(\hat{\boldsymbol{\beta}})|} &= \frac{|\mathbf{X}'\mathbf{V}\mathbf{X} - \mathbf{X}'\mathbf{V}\mathbf{Z}(\mathbf{Z}'\mathbf{V}\mathbf{Z})^{-}\mathbf{Z}'\mathbf{V}\mathbf{X}| \cdot |\mathbf{X}'\mathbf{X}|^{-2}}{|\mathbf{X}'\mathbf{V}\mathbf{X}| \cdot |\mathbf{X}'\mathbf{X}|^{-2}} \\ &= \frac{|\mathbf{X}'\mathbf{V}\mathbf{X} - \mathbf{X}'\mathbf{V}\mathbf{Z}(\mathbf{Z}'\mathbf{V}\mathbf{Z})^{-}\mathbf{Z}'\mathbf{V}\mathbf{X}|}{|\mathbf{X}'\mathbf{V}\mathbf{X}|} \\ &= |\mathbf{I}_p - (\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{Z}(\mathbf{Z}'\mathbf{V}\mathbf{Z})^{-}\mathbf{Z}'\mathbf{V}\mathbf{X}|\,. \end{aligned} \tag{10.162}$$

For (10.162) to be properly defined, we clearly must have

$$|\mathbf{X}'\mathbf{V}\mathbf{X}| \neq 0\,, \quad \text{i.e.,} \quad \operatorname{rank}(\mathbf{X}'\mathbf{V}\mathbf{X}) = p\,, \tag{10.163}$$

that is, see Proposition 6.1 (p. 147),

$$\operatorname{rank}(\mathbf{X}'\mathbf{V}) = \operatorname{rank}(\mathbf{X}) - \dim \mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{V})^{\perp} = p\,, \tag{10.164}$$

which shows that for the Watson efficiency to be properly defined it is necessary that

$$\mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{V})^{\perp} = \{\mathbf{0}\}\,. \tag{10.165}$$

Moreover, because (see Proposition 5.6, page 137)

$$\operatorname{rank}[\operatorname{cov}(\tilde{\boldsymbol{\beta}})] = \dim \mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{V}) \leq \operatorname{rank}(\mathbf{X})\,, \tag{10.166}$$

we observe that $|\operatorname{cov}(\tilde{\boldsymbol{\beta}})| = 0$, i.e., the Watson efficiency becomes 0 if the model is not a weakly singular linear model—even though $\mathbf{X}$ has full column rank. Hence, in order to have a properly defined nonzero Watson efficiency, we have to assume that

$$\mathscr{C}(\mathbf{X}) \subset \mathscr{C}(\mathbf{V})\,, \tag{10.167}$$

which trivially holds if $\mathbf{V}$ is positive definite.

Let us next consider the canonical correlations between the random vectors $\mathbf{X}'\mathbf{y}$ and $\mathbf{Z}'\mathbf{y}$, and similarly between $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$. Then we have

$$\operatorname{cov}\begin{pmatrix}\mathbf{X}'\mathbf{y}\\ \mathbf{Z}'\mathbf{y}\end{pmatrix} = \begin{pmatrix}\mathbf{X}'\mathbf{V}\mathbf{X} & \mathbf{X}'\mathbf{V}\mathbf{Z}\\ \mathbf{Z}'\mathbf{V}\mathbf{X} & \mathbf{Z}'\mathbf{V}\mathbf{Z}\end{pmatrix} := \begin{pmatrix}\mathbf{K}'\mathbf{K} & \mathbf{K}'\mathbf{L}\\ \mathbf{L}'\mathbf{K} & \mathbf{L}'\mathbf{L}\end{pmatrix} := \boldsymbol{\Sigma}\,, \tag{10.168}$$

where we have denoted

$$\mathbf{K} = \mathbf{V}^{1/2}\mathbf{X}\,, \quad \mathbf{L} = \mathbf{V}^{1/2}\mathbf{Z}\,. \tag{10.169}$$

Notice that with this notation we have

$$\mathbf{X}'\mathbf{V}\mathbf{Z}(\mathbf{Z}'\mathbf{V}\mathbf{Z})^-\mathbf{Z}'\mathbf{V}\mathbf{X} = \mathbf{K}'\mathbf{P}_\mathbf{L}\mathbf{K}\,, \tag{10.170a}$$

$$\operatorname{cov}(\hat{\boldsymbol\beta}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}'\mathbf{K}(\mathbf{X}'\mathbf{X})^{-1}, \tag{10.170b}$$

$$\operatorname{cov}(\tilde{\boldsymbol\beta}) = (\mathbf{X}'\mathbf{X})^{-1}[\mathbf{K}'(\mathbf{I}_n - \mathbf{P}_\mathbf{L})\mathbf{K}](\mathbf{X}'\mathbf{X})^{-1}. \tag{10.170c}$$

Under a weakly singular linear model we have $\tilde{\boldsymbol\beta} = (\mathbf{X}'\mathbf{V}^+\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^+\mathbf{y}$ and hence

$$\begin{aligned}\operatorname{cov}(\hat{\boldsymbol\beta}, \tilde{\boldsymbol\beta}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \cdot \mathbf{V} \cdot \mathbf{V}^+\mathbf{X}(\mathbf{X}'\mathbf{V}^+\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \cdot \mathbf{P}_\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{V}^+\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{V}^+\mathbf{X})^{-1} = \operatorname{cov}(\tilde{\boldsymbol\beta})\,, \end{aligned}\tag{10.171}$$

and so, just as in (10.122) (p. 235), we have

$$\begin{aligned}\operatorname{cov}\begin{pmatrix}\hat{\boldsymbol\beta}\\ \tilde{\boldsymbol\beta}\end{pmatrix} &= \begin{pmatrix}\operatorname{cov}(\hat{\boldsymbol\beta}) & \operatorname{cov}(\tilde{\boldsymbol\beta})\\ \operatorname{cov}(\tilde{\boldsymbol\beta}) & \operatorname{cov}(\tilde{\boldsymbol\beta})\end{pmatrix} \\ &= \begin{pmatrix}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} & (\mathbf{X}'\mathbf{V}^+\mathbf{X})^{-1}\\ (\mathbf{X}'\mathbf{V}^+\mathbf{X})^{-1} & (\mathbf{X}'\mathbf{V}^+\mathbf{X})^{-1}\end{pmatrix} := \boldsymbol\Gamma\,. \end{aligned}\tag{10.172}$$

Let and $\kappa_1 \geq \kappa_2 \geq \cdots \geq \kappa_p \geq 0$ and $\theta_1 \geq \theta_2 \geq \cdots \geq \theta_p > 0$ denote the (ordered) canonical correlations between $\mathbf{X}'\mathbf{y}$ and $\mathbf{Z}'\mathbf{y}$, and $\hat{\boldsymbol\beta}$ and $\tilde{\boldsymbol\beta}$, respectively, i.e.,

$$\kappa_i = \operatorname{cc}_i(\mathbf{X}'\mathbf{y}, \mathbf{Z}'\mathbf{y})\,, \qquad \theta_i = \operatorname{cc}_i(\hat{\boldsymbol\beta}, \tilde{\boldsymbol\beta})\,, \quad i = 1, \dots, p\,. \tag{10.173}$$

We assume that $p \leq n/2$ in which case the number of the canonical correlations, i.e., the number of pairs of canonical variables based on $\mathbf{X}'\mathbf{y}$ and $\mathbf{Z}'\mathbf{y}$, is p.

In view of Proposition 18.11 (p. 381), the number of the nonzero κ_i^2's is

$$m = \operatorname{rank}(\mathbf{X}'\mathbf{V}\mathbf{Z}) = \operatorname{rank}(\mathbf{K}'\mathbf{L}) = \operatorname{rank}(\boldsymbol\Sigma_{12})\,, \tag{10.174}$$

and they can be expressed as the nonzero eigenvalues as follows:

$$\begin{aligned}\{\kappa_1^2, \dots, \kappa_m^2\} &= \operatorname{cc}_+^2(\mathbf{X}'\mathbf{y}, \mathbf{Z}'\mathbf{y}) \\ &= \operatorname{nzch}[(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{Z}(\mathbf{Z}'\mathbf{V}\mathbf{Z})^-\mathbf{Z}'\mathbf{V}\mathbf{X}] \\ &= \operatorname{nzch}(\mathbf{P}_\mathbf{K}\mathbf{P}_\mathbf{L})\,. \end{aligned}\tag{10.175}$$

[Notice that $\operatorname{cc}_+^2(\mathbf{u}, \mathbf{v})$ is a short notation for the set of the squared nonzero canonical correlations between the random vectors $\mathbf{u}$ and $\mathbf{v}$.] Similarly, there are

$$p = \operatorname{rank}(\boldsymbol\Gamma_{12}) \tag{10.176}$$

nonzero θ_i^2's, and they can be expressed as the (necessarily nonzero) eigenvalues of the matrix

$$\begin{aligned}\mathbf{N}_1 &:= [\operatorname{cov}(\hat{\boldsymbol{\beta}})]^{-1}\operatorname{cov}(\tilde{\boldsymbol{\beta}})\\ &= (\mathbf{X}'\mathbf{X})(\mathbf{K}'\mathbf{K})^{-1}(\mathbf{X}'\mathbf{X})\cdot(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}'(\mathbf{I}_n-\mathbf{P}_{\mathbf{L}})\mathbf{K}(\mathbf{X}'\mathbf{X})^{-1}\\ &= (\mathbf{X}'\mathbf{X})(\mathbf{K}'\mathbf{K})^{-1}\mathbf{K}'(\mathbf{I}_n-\mathbf{P}_{\mathbf{L}})\mathbf{K}(\mathbf{X}'\mathbf{X})^{-1}\\ &= (\mathbf{X}'\mathbf{X})[\mathbf{I}_p-(\mathbf{K}'\mathbf{K})^{-1}\mathbf{K}'\mathbf{P}_{\mathbf{L}}\mathbf{K}](\mathbf{X}'\mathbf{X})^{-1},\end{aligned}\tag{10.177}$$

which are the same as those of the matrix

$$\mathbf{N}_2 := \mathbf{I}_p-(\mathbf{K}'\mathbf{K})^{-1}\mathbf{K}'\mathbf{P}_{\mathbf{L}}\mathbf{K}\,.\tag{10.178}$$

Hence the p (all nonzero) θ_i^2's are

$$\begin{aligned}\{\theta_1^2,\dots,\theta_p^2\} &= \operatorname{cc}^2(\hat{\boldsymbol{\beta}},\tilde{\boldsymbol{\beta}})\\ &= \operatorname{ch}[\mathbf{I}_p-(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{Z}(\mathbf{Z}'\mathbf{V}\mathbf{Z})^{-}\mathbf{Z}'\mathbf{V}\mathbf{X}]\,.\end{aligned}\tag{10.179}$$

Thus, in light of (10.175) and (10.179), we can conclude the following interesting connection:

$$\theta_i^2 = 1-\kappa_{p-i+1}^2\,,\quad i=1,\dots,p\,,\tag{10.180a}$$

$$\operatorname{cc}_i^2(\hat{\boldsymbol{\beta}},\,\tilde{\boldsymbol{\beta}}) = 1-\operatorname{cc}_i^2(\mathbf{X}'\mathbf{y},\mathbf{Z}'\mathbf{y})\,,\quad i=1,\dots,p\,,\tag{10.180b}$$

Notice also that the θ_i^2's are the roots of the equation

$$|\operatorname{cov}(\tilde{\boldsymbol{\beta}})-\theta^2\operatorname{cov}(\hat{\boldsymbol{\beta}})| = 0\,.\tag{10.181}$$

Now under a weakly singular model and $\mathbf{X}$ having full column rank, the Watson efficiency can be written as

$$\begin{aligned}\operatorname{eff}(\hat{\boldsymbol{\beta}}) = \phi &= \frac{|\mathbf{X}'\mathbf{X}|^2}{|\mathbf{X}'\mathbf{V}\mathbf{X}|\cdot|\mathbf{X}'\mathbf{V}^{+}\mathbf{X}|}\\ &= \frac{|\mathbf{X}'\mathbf{V}\mathbf{X}-\mathbf{X}'\mathbf{V}\mathbf{Z}(\mathbf{Z}'\mathbf{V}\mathbf{Z})^{-}\mathbf{Z}'\mathbf{V}\mathbf{X}|}{|\mathbf{X}'\mathbf{V}\mathbf{X}|}\\ &= |\mathbf{I}_p-(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{Z}(\mathbf{Z}'\mathbf{V}\mathbf{Z})^{-}\mathbf{Z}'\mathbf{V}\mathbf{X}|\\ &= \prod_{i=1}^{p}\theta_i^2 = \prod_{i=1}^{p}(1-\kappa_i^2)\,.\end{aligned}\tag{10.182}$$

The efficiency formula (10.182) in terms of κ_i's and θ_i's was first introduced ($\mathbf{V}$ being positive definite) by Bartmann & Bloomfield (1981).

In (10.182) there are no worries about the zero θ_i's nor unit (1's) κ_i's once we are dealing with a weakly singular model. As a matter of fact, the number of unit canonical correlations between $\mathbf{X}'\mathbf{y}$ and $\mathbf{Z}'\mathbf{y}$, say u, is the number of unit eigenvalues of the product $\mathbf{P}_{\mathbf{K}}\mathbf{P}_{\mathbf{L}}$, which in view of (5.83) (p. 134) and Proposition 5.5 (p. 135), can be expressed as

$$\begin{aligned} u &= \#\{\mathrm{ch}(\mathbf{P_K P_L}) = 1\} = \dim \mathscr{C}(\mathbf{K}) \cap \mathscr{C}(\mathbf{L}) \\ &= \dim \mathscr{C}(\mathbf{V}^{1/2}\mathbf{X}) \cap \mathscr{C}(\mathbf{V}^{1/2}\mathbf{Z}) = \mathrm{rank}(\mathbf{X}'\mathbf{P_V Z}) . \end{aligned} \tag{10.183}$$

Trivially $u = 0$ if $\mathscr{C}(\mathbf{X}) \subset \mathscr{C}(\mathbf{V})$.

It is noteworthy that for the nonzero κ_i^2's we have

$$\begin{aligned} \{\kappa_1^2, \dots, \kappa_m^2\} &= \mathrm{cc}_+^2(\mathbf{X}'\mathbf{y}, \mathbf{Z}'\mathbf{y}) = \mathrm{nzch}(\mathbf{P_K P_L}) \\ &= \mathrm{nzch}(\mathbf{P}_{\mathbf{V}^{1/2}\mathbf{X}} \mathbf{P}_{\mathbf{V}^{1/2}\mathbf{Z}}) = \mathrm{nzch}(\mathbf{P}_{\mathbf{V}^{1/2}\mathbf{H}} \mathbf{P}_{\mathbf{V}^{1/2}\mathbf{M}}) , \end{aligned} \tag{10.184}$$

and hence

$$\{\kappa_1^2, \dots, \kappa_m^2\} = \mathrm{cc}_+^2(\mathbf{X}'\mathbf{y}, \mathbf{Z}'\mathbf{y}) = \mathrm{cc}_+^2(\mathbf{Hy}, \mathbf{My}) . \tag{10.185}$$

Similarly one can show (left to the reader) that

$$\{\theta_1^2, \dots, \theta_p^2\} = \mathrm{cc}^2(\hat{\boldsymbol{\beta}}, \tilde{\boldsymbol{\beta}}) = \mathrm{cc}^2(\mathbf{X}\hat{\boldsymbol{\beta}}, \mathbf{X}\tilde{\boldsymbol{\beta}}) . \tag{10.186}$$

On account of (10.184), the Watson efficiency can be written also as

$$\phi = |\mathbf{I}_n - \mathbf{P}_{\mathbf{V}^{1/2}\mathbf{X}} \mathbf{P}_{\mathbf{V}^{1/2}\mathbf{Z}}| = |\mathbf{I}_n - \mathbf{P}_{\mathbf{V}^{1/2}\mathbf{H}} \mathbf{P}_{\mathbf{V}^{1/2}\mathbf{M}}| . \tag{10.187}$$

In this section we have considered the weakly singular model which is required so that the Watson efficiency would be properly defined. However, even if the Watson efficiency collapses, so to say, it might still make sense to utilize the canonical correlations θ_i or equivalently κ_i while characterizing the relative goodness of OLS. Some considerations are done in Section 18.10 (p. 382).

10.9 Best Linear Unbiased Predictor

Consider the general linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} , \quad \text{i.e.,} \quad \mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{V}\} . \tag{10.188}$$

Let $\mathbf{y}_f$ denote an $m \times 1$ unobservable random vector containing *new observations.* The new observations are assumed to follow the linear model

$$\mathbf{y}_f = \mathbf{X}_f\boldsymbol{\beta} + \boldsymbol{\varepsilon}_f , \tag{10.189}$$

where $\mathbf{X}_f$ is a known $m \times p$ model matrix associated with new observations, $\boldsymbol{\beta}$ is the same vector of unknown parameters as in (10.188), and $\boldsymbol{\varepsilon}_f$ is an $m \times 1$ random error vector associated with new observations. It is essential to understand the difference between the random vectors $\mathbf{y}$ and $\mathbf{y}_f$: $\mathbf{y}$ is observable but $\mathbf{y}_f$ is *not*; $\mathbf{y}_f$ is a future value of a random vector rising from the model (10.189). Our object is to predict the random vector $\mathbf{y}_f$ on the basis of $\mathbf{y}$.

The expectations are

$$\mathrm{E}\begin{pmatrix}\mathbf{y}\\ \mathbf{y}_f\end{pmatrix} = \begin{pmatrix}\mathbf{X}\boldsymbol{\beta}\\ \mathbf{X}_f\boldsymbol{\beta}\end{pmatrix} = \begin{pmatrix}\mathbf{X}\\ \mathbf{X}_f\end{pmatrix}\boldsymbol{\beta}, \tag{10.190}$$

and the joint covariance matrix of $\mathbf{y}$ and $\mathbf{y}_f$ is

$$\operatorname{cov}\begin{pmatrix}\mathbf{y}\\ \mathbf{y}_f\end{pmatrix} = \sigma^2\begin{pmatrix}\mathbf{V} = \mathbf{V}_{11} & \mathbf{V}_{12}\\ \mathbf{V}_{21} & \mathbf{V}_{22}\end{pmatrix} := \sigma^2\boldsymbol{\Sigma}\,. \tag{10.191}$$

We can denote shortly

$$\mathscr{M}_f = \left\{\begin{pmatrix}\mathbf{y}\\ \mathbf{y}_f\end{pmatrix}, \begin{pmatrix}\mathbf{X}\boldsymbol{\beta}\\ \mathbf{X}_f\boldsymbol{\beta}\end{pmatrix}, \sigma^2\begin{pmatrix}\mathbf{V} & \mathbf{V}_{12}\\ \mathbf{V}_{21} & \mathbf{V}_{22}\end{pmatrix}\right\}. \tag{10.192}$$

Notice that we naturally could have denoted the upper-left-block of $\boldsymbol{\Sigma}$ as $\mathbf{V}_{11}$ but we have deliberately left the subscripts out so that the reader could more easily see that $\mathscr{M}_f$ is simply the standard linear model $\{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{V}\}$ extended with a new future vector of observations $\mathbf{y}_f$.

A linear predictor $\mathbf{Ay}$ is said to be unbiased for $\mathbf{y}_f$ if

$$\mathrm{E}(\mathbf{Ay}) = \mathrm{E}(\mathbf{y}_f) = \mathbf{X}_f\boldsymbol{\beta} \quad \text{for all } \boldsymbol{\beta} \in \mathbb{R}^p\,, \tag{10.193}$$

i.e., the expected prediction error is $\mathbf{0}$:

$$\mathrm{E}(\mathbf{y}_f - \mathbf{Ay}) = \mathbf{0}\,. \tag{10.194}$$

We may then say that the random vector $\mathbf{y}_f$ is *unbiasedly predictable.* It is easy to confirm that (10.193) is equivalent to $\mathbf{AX} = \mathbf{X}_f$ i.e., $\mathbf{X}_f' = \mathbf{X}'\mathbf{A}$, which means that $\mathbf{y}_f$ is unbiasedly predictable if and only if

$$\mathscr{C}(\mathbf{X}_f') \subset \mathscr{C}(\mathbf{X}')\,, \quad \text{i.e.,} \quad \mathbf{X}_f\boldsymbol{\beta} \text{ is estimable.} \tag{10.195}$$

Now an unbiased linear predictor $\mathbf{Ay}$ is the *best linear unbiased predictor,* BLUP, a term introduced by Goldberger (1962), if the Löwner ordering

$$\operatorname{cov}(\mathbf{Ay} - \mathbf{y}_f) \leq_{\mathsf{L}} \operatorname{cov}(\mathbf{By} - \mathbf{y}_f) \tag{10.196}$$

holds for all $\mathbf{B}$ such that $\mathbf{By}$ is an unbiased linear predictor for $\mathbf{y}_f$.

Notice that if the parameter vector $\boldsymbol{\beta}$ were known, then $\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} := \boldsymbol{\mu}_{\mathbf{y}}$ and $\mathrm{E}(\mathbf{y}_f) = \mathbf{X}_f\boldsymbol{\beta} := \boldsymbol{\mu}_{\mathbf{y}_f}$ and thus in view of (9.45) (p. 196), the BLP of $\mathbf{y}_f$ on the basis of $\mathbf{y}$ would be

$$\begin{aligned}\mathrm{BLP}(\mathbf{y}_f;\mathbf{y}) &= \boldsymbol{\mu}_{\mathbf{y}_f} - \mathbf{V}_{21}\mathbf{V}^-\boldsymbol{\mu}_{\mathbf{y}} + \mathbf{V}_{21}\mathbf{V}^-\mathbf{y}\\ &= \boldsymbol{\mu}_{\mathbf{y}_f} + \mathbf{V}_{21}\mathbf{V}^-(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}})\end{aligned}$$

$$= \mathbf{X}_f \boldsymbol{\beta} + \mathbf{V}_{21}\mathbf{V}^{-}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \,. \tag{10.197}$$

The following proposition characterizes the BLUP; see, e.g., Christensen (2002, p. 283) and Isotalo & Puntanen (2006a, p. 1015).

Proposition 10.6 (Fundamental BLUP **equation).** *Consider the linear model* $\mathscr{M}_f$*, where* $\mathbf{X}_f\boldsymbol{\beta}$ *is a given estimable parametric function. Then the linear estimator* $\mathbf{Ay}$ *is the best linear unbiased predictor* (BLUP) *for* $\mathbf{y}_f$ *if and only if* $\mathbf{A}$ *satisfies the equation*

$$\mathbf{A}(\mathbf{X} : \mathbf{V}\mathbf{X}^{\perp}) = (\mathbf{X}_f : \mathbf{V}_{21}\mathbf{X}^{\perp}) \,. \tag{10.198}$$

Proof. The proof is parallel to that of Theorem 10 (p. 216)—but in spite of that, we'll go through it.

Let $\mathbf{A}$ be any matrix satisfying $\mathbf{A}(\mathbf{X} : \mathbf{VM}) = (\mathbf{X}_f : \mathbf{V}_{21}\mathbf{M})$, where $\mathbf{M} = \mathbf{I} - \mathbf{H}$, and let $\mathbf{By}$ be any other linear unbiased predictor for $\mathbf{y}_f$, i.e., $\mathbf{BX} = \mathbf{X}_f$. Hence $(\mathbf{B} - \mathbf{A})\mathbf{X} = \mathbf{0}$ and so $\mathbf{B}' - \mathbf{A}' = \mathbf{ML}$, for some matrix $\mathbf{L}$. Moreover, the assumption $\mathbf{AVM} = \mathbf{V}_{21}\mathbf{M}$ implies that

$$\begin{aligned}
\operatorname{cov}[\mathbf{Ay} - \mathbf{y}_f, (\mathbf{B} - \mathbf{A})\mathbf{y}] &= \mathbf{AV}(\mathbf{B} - \mathbf{A})' - \mathbf{V}_{21}(\mathbf{B} - \mathbf{A})' \\
&= (\mathbf{AV} - \mathbf{V}_{21})(\mathbf{B} - \mathbf{A})' \\
&= (\mathbf{AV} - \mathbf{V}_{21})\mathbf{ML} = \mathbf{0} \,,
\end{aligned} \tag{10.199}$$

and hence we have the Löwner ordering

$$\begin{aligned}
\operatorname{cov}(\mathbf{By} - \mathbf{y}_f) &= \operatorname{cov}[(\mathbf{By} - \mathbf{Ay}) + (\mathbf{Ay} - \mathbf{y}_f)] \\
&= \operatorname{cov}(\mathbf{By} - \mathbf{Ay}) + \operatorname{cov}(\mathbf{Ay} - \mathbf{y}_f) \\
&\geq_{\mathsf{L}} \operatorname{cov}(\mathbf{Ay} - \mathbf{y}_f) \,.
\end{aligned} \tag{10.200}$$

Conversely, let $\mathbf{Ay}$ be the BLUP for $\mathbf{y}_f$. Then the inequality (10.200) must also hold when $\mathbf{B}$ satisfies the equation $\mathbf{B}(\mathbf{X} : \mathbf{VM}) = (\mathbf{X}_f : \mathbf{V}_{21}\mathbf{M})$. Because $\mathbf{Ay}$ is an unbiased predictor, we have $\mathbf{A}' - \mathbf{B}' = \mathbf{MK}$, for some matrix $\mathbf{K}$, and just as in (10.199),

$$\operatorname{cov}[\mathbf{By} - \mathbf{y}_f, (\mathbf{A} - \mathbf{B})\mathbf{y}] = (\mathbf{BV} - \mathbf{V}_{21})\mathbf{MK} = \mathbf{0} \,. \tag{10.201}$$

Hence we have the Löwner ordering

$$\begin{aligned}
\operatorname{cov}(\mathbf{Ay} - \mathbf{y}_f) &= \operatorname{cov}[(\mathbf{Ay} - \mathbf{By}) + (\mathbf{By} - \mathbf{y}_f)] \\
&= \operatorname{cov}(\mathbf{Ay} - \mathbf{By}) + \operatorname{cov}(\mathbf{By} - \mathbf{y}_f) \\
&\geq_{\mathsf{L}} \operatorname{cov}(\mathbf{By} - \mathbf{y}_f) \,.
\end{aligned} \tag{10.202}$$

Because $\mathbf{Ay}$ is the BLUP for $\mathbf{y}_f$ we must have

$$\operatorname{cov}(\mathbf{By} - \mathbf{y}_f) = \operatorname{cov}(\mathbf{Ay} - \mathbf{y}_f) \tag{10.203}$$

which holds if and only if

$$\operatorname{cov}(\mathbf{Ay}-\mathbf{By}) = (\mathbf{A}-\mathbf{B})\mathbf{V}(\mathbf{A}-\mathbf{B})' = \mathbf{0}\,, \tag{10.204}$$

that is, if and only if $(\mathbf{A}-\mathbf{B})\mathbf{V} = \mathbf{0}$. Hence $\mathbf{AVM} = \mathbf{BVM} = \mathbf{V}_{21}\mathbf{M}$, which concludes the proof. □

It is easy to conclude that the BLUP can be characterized via Pandora's Box technique as follows.

Proposition 10.7 (Pandora's Box and the BLUP). *Consider the linear model $\mathscr{M}_f$, where $\mathbf{X}_f\boldsymbol{\beta}$ is a given estimable parametric function. Then the linear estimator $\mathbf{Ay}$ is the* BLUP *for $\mathbf{y}_f$ if and only if there exists a matrix $\mathbf{L}$ such that $\mathbf{A}$ satisfies the equation*

$$\begin{pmatrix} \mathbf{V} & \mathbf{X} \\ \mathbf{X}' & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{A}' \\ \mathbf{L} \end{pmatrix} = \begin{pmatrix} \mathbf{V}_{12} \\ \mathbf{X}_f' \end{pmatrix}. \tag{10.205}$$

According to Theorem 11 (p. 267), the general solution to (10.198) (whose consistency confirmation is left as an exercise) can be written, for example, as

$$\mathbf{A}_0 = (\mathbf{X}_f : \mathbf{V}_{21}\mathbf{M})(\mathbf{X} : \mathbf{VM})^+ + \mathbf{F}(\mathbf{I}_n - \mathbf{P}_{(\mathbf{X}:\mathbf{VM})})\,, \tag{10.206}$$

where the matrix $\mathbf{F}$ is free to vary. Even though the multiplier $\mathbf{A}$ may not be unique, the observed value $\mathbf{Ay}$ of the BLUP is unique with probability 1. We can get, for example, the following matrices $\mathbf{A}_i$ such that $\mathbf{A}_i\mathbf{y}$ equals the $\mathrm{BLUP}(\mathbf{y}_f)$:

$$\mathbf{A}_1 = \mathbf{X}_f\mathbf{B} + \mathbf{V}_{21}\mathbf{W}^-(\mathbf{I}_n - \mathbf{XB})\,, \tag{10.207a}$$

$$\mathbf{A}_2 = \mathbf{X}_f\mathbf{B} + \mathbf{V}_{21}\mathbf{V}^-(\mathbf{I}_n - \mathbf{XB})\,, \tag{10.207b}$$

$$\mathbf{A}_3 = \mathbf{X}_f\mathbf{B} + \mathbf{V}_{21}\mathbf{M}(\mathbf{MVM})^-\mathbf{M}\,, \tag{10.207c}$$

$$\mathbf{A}_4 = \mathbf{X}_f(\mathbf{X}'\mathbf{X})^-\mathbf{X}' + [\mathbf{V}_{21} - \mathbf{X}_f(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{V}]\mathbf{M}(\mathbf{MVM})^-\mathbf{M}\,, \tag{10.207d}$$

where $\mathbf{W} = \mathbf{V} + \mathbf{XUX}'$, with $\mathbf{U}$ satisfying $\mathscr{C}(\mathbf{W}) = \mathscr{C}(\mathbf{X} : \mathbf{V})$, and

$$\mathbf{B} = (\mathbf{XW}^-\mathbf{X})^-\mathbf{X}'\mathbf{W}^-\,. \tag{10.208}$$

Proofs of (10.207) are left to the reader.

In view of (10.207), the $\mathrm{BLUP}(\mathbf{y}_f)$ can be written as

$$\begin{aligned} \mathrm{BLUP}(\mathbf{y}_f) &= \mathbf{X}_f\tilde{\boldsymbol{\beta}} + \mathbf{V}_{21}\mathbf{W}^-(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) \\ &= \mathbf{X}_f\tilde{\boldsymbol{\beta}} + \mathbf{V}_{21}\mathbf{W}^-\tilde{\boldsymbol{\varepsilon}} \\ &= \mathbf{X}_f\tilde{\boldsymbol{\beta}} + \mathbf{V}_{21}\mathbf{V}^-\tilde{\boldsymbol{\varepsilon}} \\ &= \mathbf{X}_f\tilde{\boldsymbol{\beta}} + \mathbf{V}_{21}\mathbf{M}(\mathbf{MVM})^-\mathbf{My} \\ &= \mathrm{BLUE}(\mathbf{X}_f\boldsymbol{\beta}) + \mathbf{V}_{21}\dot{\mathbf{M}}\mathbf{y} \end{aligned}$$

$$\begin{aligned} &= \mathbf{X}_f\hat{\boldsymbol{\beta}} + (\mathbf{V}_{21} - \mathbf{X}_f\mathbf{X}^+\mathbf{V})\dot{\mathbf{M}}\mathbf{y} \\ &= \mathrm{OLSE}(\mathbf{X}_f\boldsymbol{\beta}) + (\mathbf{V}_{21} - \mathbf{X}_f\mathbf{X}^+\mathbf{V})\dot{\mathbf{M}}\mathbf{y}\,, \qquad (10.209) \end{aligned}$$

where $\tilde{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}$ is the vector of the BLUE's residual:

$$\mathbf{X}\tilde{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{VM}(\mathbf{MVM})^-\mathbf{My} = \mathbf{y} - \mathbf{V}\dot{\mathbf{M}}\mathbf{y}\,, \qquad (10.210a)$$

$$\tilde{\boldsymbol{\varepsilon}} = \mathbf{V}\dot{\mathbf{M}}\mathbf{y} = \mathbf{W}\dot{\mathbf{M}}\mathbf{y}\,, \quad \dot{\mathbf{M}} = \mathbf{M}(\mathbf{MVM})^-\mathbf{M}\,. \qquad (10.210b)$$

It may be worth emphasizing that if $\mathbf{A}$ and $\mathbf{B}$ are arbitrary matrices satisfying

$$\begin{pmatrix} \mathbf{A} \\ \mathbf{B} \end{pmatrix} (\mathbf{X} : \mathbf{VM}) = \begin{pmatrix} \mathbf{X}_f & \mathbf{V}_{21}\mathbf{M} \\ \mathbf{X}_f & \mathbf{0} \end{pmatrix}, \qquad (10.211)$$

then (10.209) means, e.g., that

$$\mathbf{Ay} = \mathbf{By} + \mathbf{V}_{21}\dot{\mathbf{M}}\mathbf{y} \quad \text{for all } \mathbf{y} \in \mathscr{C}(\mathbf{X} : \mathbf{V}) = \mathscr{C}(\mathbf{X} : \mathbf{VM})\,. \qquad (10.212)$$

Consider next an arbitrary representation $\mathbf{Gy}$ for the BLUE of $\mathbf{X}_f\boldsymbol{\beta}$, so that $\mathbf{G}$ satisfies $\mathbf{G}(\mathbf{X} : \mathbf{VM}) = (\mathbf{X} : \mathbf{0})$. Then $\mathbf{Gy}$ is also the BLUP for $\mathbf{y}_f$ if $\mathbf{G}$ satisfies $\mathbf{G}(\mathbf{X} : \mathbf{VM}) = (\mathbf{X} : \mathbf{V}_{21}\mathbf{M})$. Hence we can conclude that

$$\begin{aligned} \mathrm{BLUP}(\mathbf{y}_f) = \mathrm{BLUE}(\mathbf{X}_f\boldsymbol{\beta}) &\iff \mathbf{V}_{21}\mathbf{M} = \mathbf{0} \\ &\iff \mathscr{C}(\mathbf{V}_{12}) \subset \mathscr{C}(\mathbf{X})\,. \qquad (10.213) \end{aligned}$$

The above problem was considered by Elian (2000), and Puntanen, Styan & Werner (2000a). For related considerations, see also Baksalary & Kala (1981b, Th. 1), Watson (1972, §3)], and Liu (2009).

The following proposition characterizes the equality of $\mathrm{BLUP}(\mathbf{y}_f)$ and $\mathrm{OLSE}(\mathbf{X}_f\boldsymbol{\beta})$, the proof is left as an exercise (see the above references).

Proposition 10.8. *Consider the linear model $\mathscr{M}_f$. Then the following statements are equivalent:*

(a) $\mathrm{BLUP}(\mathbf{y}_f) = \mathrm{OLSE}(\mathbf{X}_f\boldsymbol{\beta}) = \mathbf{X}_f\hat{\boldsymbol{\beta}}$ *for a fixed* $\mathbf{X}_f = \mathbf{LX}$,

(b) $\mathscr{C}[\mathbf{V}_{12} - \mathbf{V}(\mathbf{X}')^+\mathbf{X}_f'] \subset \mathscr{C}(\mathbf{X})$,

(c) $\mathscr{C}(\mathbf{V}_{12} - \mathbf{VHL}') \subset \mathscr{C}(\mathbf{X})$.

Moreover, the following statements are equivalent:

(d) $\mathrm{BLUP}(\mathbf{y}_f) = \mathrm{OLSE}(\mathbf{X}_f\boldsymbol{\beta}) = \mathbf{X}_f\hat{\boldsymbol{\beta}}$ *for all* $\mathbf{X}_f$ *of the form* $\mathbf{X}_f = \mathbf{LX}$,

(e) $\mathscr{C}(\mathbf{VX}) \subset \mathscr{C}(\mathbf{X})$ *and* $\mathscr{C}(\mathbf{V}_{12}) \subset \mathscr{C}(\mathbf{X})$,

(f) $\mathrm{OLSE}(\mathbf{X}\boldsymbol{\beta}) = \mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta})$ *and* $\mathrm{BLUP}(\mathbf{y}_f) = \mathrm{BLUE}(\mathbf{X}_f\boldsymbol{\beta})$.

10.10 Examples on the BLUPs

In Section 9.6 (p. 208) we considered the best linear prediction, BLP, in the autoregressive case, where the covariance matrix is of the type

$$\begin{aligned}\mathbf{\Sigma} = \sigma^2\{\varrho^{|i-j|}\} &= \sigma^2 \begin{pmatrix} 1 & \varrho & \varrho^2 & \dots & \varrho^n \\ \varrho & 1 & \varrho & \dots & \varrho^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ \varrho^n & \varrho^{n-1} & \varrho^{n-2} & \dots & 1 \end{pmatrix} \\ &= \sigma^2 \begin{pmatrix} \mathbf{V} & \mathbf{v}_{12} \\ \mathbf{v}_{12}' & 1 \end{pmatrix} = \operatorname{cov}\begin{pmatrix} \mathbf{y} \\ y_{n+1} \end{pmatrix}, \end{aligned} \tag{10.214}$$

where $|\varrho| < 1$, and the cross-covariance matrix (vector)

$$\mathbf{v}_{12} = \sigma^2(\varrho^n, \varrho^{n-1}, \dots, \varrho^2, \varrho)' \in \mathbb{R}^n . \tag{10.215}$$

Putting, for simplicity, $\mathrm{E}(\mathbf{y}) = \mathbf{0}$, we obtained, cf. page 209,

$$\mathrm{BLP}(y_{n+1}; \mathbf{y}) = \varrho y_n . \tag{10.216}$$

Next we consider the BLUP for the new single observation $y_f = y_{n+1}$ when $\begin{pmatrix} \mathbf{y} \\ y_{n+1} \end{pmatrix}$ follows the linear model where the covariance matrix is as in (10.214).

As noted on page 209, the column $\mathbf{v}_{12}$ can be obtained by multiplying the last column of $\mathbf{V}$ by ϱ:

$$\mathbf{v}_{12} = \varrho \mathbf{V} \mathbf{i}_n , \tag{10.217}$$

where $\mathbf{i}_n$ is the last column of $\mathbf{I}_n$. Therefore

$$\mathbf{V}^{-1}\mathbf{v}_{12} = \varrho \mathbf{V}^{-1}\mathbf{V}\mathbf{i}_n = \varrho \mathbf{i}_n , \tag{10.218}$$

and the BLUP for the new single observation $y_f = y_{n+1}$ is

$$\mathrm{BLUP}(y_{n+1}) = \mathbf{x}_*'\tilde{\boldsymbol{\beta}} + \mathbf{v}_{12}'\mathbf{V}^{-1}\tilde{\boldsymbol{\varepsilon}} = \mathbf{x}_*'\tilde{\boldsymbol{\beta}} + \varrho\mathbf{i}_n'\tilde{\boldsymbol{\varepsilon}} = \mathbf{x}_*'\tilde{\boldsymbol{\beta}} + \varrho\tilde{\varepsilon}_n , \tag{10.219}$$

where $\mathbf{x}_*'$ is the row vector comprising the x_i-values of the new observation, and $\tilde{\boldsymbol{\varepsilon}}$ is the vector of the BLUE's residual. Notice how "simple" it is to make predictions in this situation.

Consider then the situation where the covariance matrix of $\begin{pmatrix} \mathbf{y} \\ y_f \end{pmatrix}$ has an intraclass correlation structure:

$$\begin{aligned}\mathbf{\Sigma} &= (1-\varrho)\mathbf{I}_{n+m} + \varrho\mathbf{1}_{n+m}\mathbf{1}_{n+m}' \\ &= \begin{pmatrix} (1-\varrho)\mathbf{I}_n + \varrho\mathbf{1}_n\mathbf{1}_n' & \varrho\mathbf{1}_n\mathbf{1}_m' \\ \varrho\mathbf{1}_m\mathbf{1}_n' & (1-\varrho)\mathbf{I}_m + \varrho\mathbf{1}_m\mathbf{1}_m' \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{V} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{pmatrix}, \end{aligned} \tag{10.220}$$

and $\mathbf{1}_n \in \mathscr{C}(\mathbf{X})$, $\mathbf{1}_m \in \mathscr{C}(\mathbf{X}_f)$. Then $\mathscr{C}(\mathbf{V}_{12}) \subset \mathscr{C}(\mathbf{X})$ and so in view of (10.213) (p. 249), we have

$$\mathrm{BLUP}(\mathbf{y}_f) = \mathrm{BLUE}(\mathbf{X}_f\boldsymbol{\beta}) \,. \tag{10.221}$$

Because $\mathrm{BLUE}(\mathbf{X}_f\boldsymbol{\beta}) = \mathrm{OLSE}(\mathbf{X}_f\boldsymbol{\beta})$ under the intraclass correlation structure when $\mathbf{1} \in \mathscr{C}(\mathbf{X})$, we have

$$\mathrm{BLUP}(\mathbf{y}_f) = \mathrm{OLSE}(\mathbf{X}_f\boldsymbol{\beta}) = \mathbf{X}_f(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y} = \mathbf{X}_f\mathbf{X}^{+}\mathbf{y} \,. \tag{10.222}$$

Isotalo & Puntanen (2006a, §4) considered the linear prediction under the following time series model:

$$y_t = \beta t + \varepsilon_t \,, \tag{10.223}$$

where t denotes the discrete time variable, β is an unknown parameter, and u_t is the standardized random walk process, cf. Davidson & MacKinnon (2004, p. 606), with a form

$$\varepsilon_t = \varepsilon_{t-1} + u_i \,, \quad \varepsilon_0 = 0 \,, \quad \mathrm{E}\begin{pmatrix} \mathbf{u} \\ u_{n+1} \end{pmatrix} = \mathbf{0} \,, \quad \mathrm{cov}\begin{pmatrix} \mathbf{u} \\ u_{n+1} \end{pmatrix} = \mathbf{I}_{n+1} \,, \tag{10.224}$$

where $\mathbf{u} = (u_1, u_2, \dots, u_n)'$. Then

$$\varepsilon_j = u_1 + u_2 + \dots + u_j \,, \quad j = 1, 2, \dots, n+1 \,, \tag{10.225}$$

and the random vector $\left(\begin{smallmatrix} \boldsymbol{\varepsilon} \\ \varepsilon_{n+1} \end{smallmatrix}\right) = (\varepsilon_1, \dots, \varepsilon_n, \varepsilon_{n+1})'$ can be written as

$$\begin{pmatrix} \boldsymbol{\varepsilon} \\ \varepsilon_{n+1} \end{pmatrix} = \mathbf{K}\begin{pmatrix} \mathbf{u} \\ u_{n+1} \end{pmatrix} , \tag{10.226}$$

where

$$\mathbf{K} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 1 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \dots & 1 \end{pmatrix} := \begin{pmatrix} \mathbf{L} & \mathbf{0}_n \\ \mathbf{1}_n' & 1 \end{pmatrix} \in \mathbb{R}^{(n+1)\times(n+1)}. \tag{10.227}$$

Let us assume that y_t is observable at times of $t = 1, 2, \dots, n$. Moreover, we are interested in predicting the outcome $y_{n+1} = y_f$ at time of $t = n+1$ based on observable random variables $y_1, y_2, \dots, y_n$. Now we can write this model as

$$\mathscr{M}_f = \left\{ \begin{pmatrix} \mathbf{y} \\ y_{n+1} \end{pmatrix} , \begin{pmatrix} \mathbf{x}\beta \\ x_f\beta \end{pmatrix} , \begin{pmatrix} \mathbf{V} & \mathbf{v}_{12} \\ \mathbf{v}_{12}' & v_{22} \end{pmatrix} \right\} , \tag{10.228}$$

where $(\mathbf{y}', y_{n+1}) = (y_1, y_2, \dots, y_n, y_{n+1})$, $(\mathbf{x}', x_f) = (1, 2, \dots, n, n+1)$, and

$$\mathbf{V} = \operatorname{cov}(\mathbf{y}) = \operatorname{cov}(\boldsymbol{\varepsilon}) = \operatorname{cov}(\mathbf{Lu}) = \mathbf{LL}'$$
$$= \begin{pmatrix} 1 & 1 & 1 & \dots & 1 & 1 \\ 1 & 2 & 2 & \dots & 2 & 2 \\ 1 & 2 & 3 & \dots & 3 & 3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 2 & 3 & \dots & n-1 & n-1 \\ 1 & 2 & 3 & \dots & n-1 & n \end{pmatrix}, \tag{10.229a}$$

and

$$\mathbf{v}_{12} = (1, 2, 3, \dots, n-1, n)', \quad v_{22} = n + 1\,. \tag{10.229b}$$

According to Proposition 10.6 (p. 247), $\mathbf{g}'\mathbf{y}$ is the BLUP for y_{n+1} under (10.228) if and only if $\mathbf{g}$ satisfies the equation

$$\mathbf{g}'(\mathbf{x} : \mathbf{V}\mathbf{x}^{\perp}) = (n + 1 : \mathbf{v}_{12}'\mathbf{x}^{\perp})\,. \tag{10.230}$$

We see immediately that choosing $\mathbf{g} = \frac{n+1}{n}\,\mathbf{i}_n$, where $\mathbf{i}_n$ is the last column of $\mathbf{I}_n$, gives the BLUP for y_{n+1}:

$$\operatorname{BLUP}(y_{n+1}) = \frac{n+1}{n} y_n\,. \tag{10.231}$$

Using Proposition 10.13 (p. 259) Isotalo & Puntanen (2006a, §3) showed that y_n is actually a linear prediction sufficient statistic for y_{n+1}.

We may use the above considerations to show that $\mathbf{A}^{-1} = \mathbf{B}$, where

Photograph 10.6 Jarkko Isotalo, ice hockey defenceman, Tappara (Tampere, 1994).

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 1 & \dots & 1 & 1 \\ 1 & 2 & 2 & \dots & 2 & 2 \\ 1 & 2 & 3 & \dots & 3 & 3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 2 & 3 & \dots & n-1 & n-1 \\ 1 & 2 & 3 & \dots & n-1 & n \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 2 & -1 & 0 & \dots & 0 & 0 \\ -1 & 2 & -1 & \dots & 0 & 0 \\ 0 & -1 & 2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 2 & -1 \\ 0 & 0 & 0 & \dots & -1 & 1 \end{pmatrix}. \tag{10.232}$$

In addition, we show that $\det(\mathbf{A}) = \det(\mathbf{B}) = 1$.

According to (10.226) (p. 251), we have $\boldsymbol{\varepsilon} = \mathbf{Lu}$, where

$$\mathbf{L} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 1 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \dots & 1 \end{pmatrix} \in \mathbb{R}^{n \times n}, \tag{10.233}$$

and $\mathbf{A} = \mathbf{L}\mathbf{L}' = \operatorname{cov}(\boldsymbol{\varepsilon})$. The determinant is $\det(\mathbf{L}) = 1$ and so $\mathbf{L}$ is nonsingular and also $\det(\mathbf{A}) = 1$. Now $\varepsilon_j - \varepsilon_{j-1} = u_j$ for $j = 2, \dots, n$, $\varepsilon_1 = u_1$, and so

$$\mathbf{u} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ -1 & 1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & 0 & \dots & -1 & 1 \end{pmatrix} \boldsymbol{\varepsilon} = \mathbf{L}^{-1}\boldsymbol{\varepsilon}\,, \tag{10.234}$$

because $\boldsymbol{\varepsilon} = \mathbf{L}\mathbf{u}$. Hence

$$\mathbf{A}^{-1} = (\mathbf{L}\mathbf{L}')^{-1} = (\mathbf{L}^{-1})'\mathbf{L}^{-1} = \begin{pmatrix} 2 & -1 & 0 & \dots & 0 & 0 \\ -1 & 2 & -1 & \dots & 0 & 0 \\ 0 & -1 & 2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 2 & -1 \\ 0 & 0 & 0 & \dots & -1 & 1 \end{pmatrix}, \tag{10.235}$$

which is $\mathbf{B}$ as claimed, and $\det(\mathbf{B}) = 1/\det(\mathbf{A}) = 1$.

For the relation (10.232), see also Watson (1996a) and Zhang (2009, Exercise 2.20).

10.11 Mixed Linear Models

A mixed linear model can be presented as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}\,, \tag{10.236}$$

where $\mathbf{X} \in \mathbb{R}^{n\times p}$ and $\mathbf{Z} \in \mathbb{R}^{n\times q}$ are known matrices, $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of unknown fixed effects, $\boldsymbol{\gamma}$ is an unobservable vector (q elements) of *random effects* with

$$\operatorname{E}(\boldsymbol{\gamma}) = \mathbf{0}_q\,, \quad \operatorname{cov}(\boldsymbol{\gamma}) = \mathbf{D}_{q\times q}\,, \quad \operatorname{cov}(\boldsymbol{\gamma}, \boldsymbol{\varepsilon}) = \mathbf{0}_{q\times p}\,, \tag{10.237}$$

and $\operatorname{E}(\boldsymbol{\varepsilon}) = \mathbf{0}_n$, $\operatorname{cov}(\boldsymbol{\varepsilon}) = \mathbf{R}_{n\times n}$. We may denote this setup briefly as

$$\mathscr{M}_{\text{mix}} = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma},\, \mathbf{D},\, \mathbf{R}\}\,. \tag{10.238}$$

Writing

$$\boldsymbol{\xi} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}\,, \quad \operatorname{cov}(\boldsymbol{\xi}) = \mathbf{Z}\mathbf{D}\mathbf{Z}' + \mathbf{R} := \boldsymbol{\Sigma}\,, \tag{10.239}$$

we can re-express (10.236) as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\xi}\,, \quad \operatorname{E}(\boldsymbol{\xi}) = \mathbf{0}\,, \quad \operatorname{cov}(\boldsymbol{\xi}) = \operatorname{cov}(\mathbf{y}) = \boldsymbol{\Sigma}\,, \tag{10.240}$$

i.e., as regards the estimation of the fixed effects under the mixed model, we can consider the following "standard fixed" model

$$\mathscr{M}_{\text{fix}} = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \mathbf{Z}\mathbf{D}\mathbf{Z}' + \mathbf{R}\} = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \boldsymbol{\Sigma}\}\,. \tag{10.241}$$

Clearly for the BLUE($\mathbf{X}\boldsymbol{\beta}$) we have the following characterization:

Proposition 10.9. *Consider the mixed model* $\mathscr{M}_{\text{mix}} = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\boldsymbol{\gamma},\, \mathbf{D},\, \mathbf{R}\}$. *Then the linear estimator* $\mathbf{B}\mathbf{y}$ *is the* BLUE *for* $\mathbf{X}\boldsymbol{\beta}$ *if and only if*

$$\mathbf{B}(\mathbf{X} : \boldsymbol{\Sigma}\mathbf{X}^{\perp}) = (\mathbf{X} : \mathbf{0})\,, \tag{10.242}$$

where $\boldsymbol{\Sigma} = \mathbf{Z}\mathbf{D}\mathbf{Z}' + \mathbf{R}$.

Assuming that the covariance matrix $\boldsymbol{\Sigma} = \mathbf{Z}\mathbf{D}\mathbf{Z}' + \mathbf{R}$ is positive definite and $\mathbf{X}$ has full column rank, we of course have

$$\text{BLUE}(\boldsymbol{\beta} \mid \mathscr{M}_{\text{mix}}) = \tilde{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{y}\,. \tag{10.243}$$

How to find the BLUP of $\boldsymbol{\gamma}$? One approach is to express the mixed model as two equations:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}\,, \tag{10.244a}$$

$$\boldsymbol{\gamma} = \mathbf{0}\cdot\boldsymbol{\beta} + \boldsymbol{\varepsilon}_f\,, \tag{10.244b}$$

where $\text{cov}(\boldsymbol{\varepsilon}_f) = \text{cov}(\boldsymbol{\gamma}) = \mathbf{D}$ and $\text{cov}(\boldsymbol{\gamma},\boldsymbol{\varepsilon}) = \text{cov}(\boldsymbol{\varepsilon}_f,\boldsymbol{\varepsilon}) = \mathbf{0}$. Since $\text{cov}(\mathbf{y},\boldsymbol{\gamma}) = \mathbf{Z}\mathbf{D}$, we have

$$\text{cov}\begin{pmatrix}\mathbf{y}\\ \boldsymbol{\gamma}\end{pmatrix} = \begin{pmatrix}\mathbf{Z}\mathbf{D}\mathbf{Z}'+\mathbf{R} & \mathbf{Z}\mathbf{D}\\ \mathbf{D}\mathbf{Z}' & \mathbf{D}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\Sigma} & \mathbf{Z}\mathbf{D}\\ \mathbf{D}\mathbf{Z}' & \mathbf{D}\end{pmatrix}. \tag{10.245}$$

Hence, corresponding to (10.192) (p. 246), the mixed model can be expressed as a version of the model with "new observations"; the new observations being now in $\boldsymbol{\gamma}$:

$$\mathscr{M}_{\text{mixed}} = \left\{\begin{pmatrix}\mathbf{y}\\ \boldsymbol{\gamma}\end{pmatrix},\, \begin{pmatrix}\mathbf{X}\\ \mathbf{0}\end{pmatrix}\boldsymbol{\beta},\, \begin{pmatrix}\mathbf{Z}\mathbf{D}\mathbf{Z}'+\mathbf{R} & \mathbf{Z}\mathbf{D}\\ \mathbf{D}\mathbf{Z}' & \mathbf{D}\end{pmatrix}\right\}. \tag{10.246}$$

Now Proposition 10.6 (p. 247) gives the following:

Proposition 10.10. *Consider the mixed linear model* $\mathscr{M}_{\text{mix}} = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma},\, \mathbf{D},\, \mathbf{R}\}$. *Then the linear estimator* $\mathbf{A}\mathbf{y}$ *is the* BLUP *for* $\boldsymbol{\gamma}$ *if and only if*

$$\mathbf{A}(\mathbf{X} : \boldsymbol{\Sigma}\mathbf{X}^{\perp}) = (\mathbf{0} : \mathbf{D}\mathbf{Z}'\mathbf{X}^{\perp})\,, \tag{10.247}$$

where $\boldsymbol{\Sigma} = \mathbf{Z}\mathbf{D}\mathbf{Z}' + \mathbf{R}$. *In terms of Pandora's Box,* $\mathbf{A}\mathbf{y} = \text{BLUP}(\boldsymbol{\gamma})$ *if and only if there exists a matrix* $\mathbf{L}$ *such that* $\mathbf{A}$ *satisfies the equation*

$$\begin{pmatrix} \mathbf{\Sigma} & \mathbf{X} \\ \mathbf{X}' & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{A}' \\ \mathbf{L} \end{pmatrix} = \begin{pmatrix} \mathbf{ZD} \\ \mathbf{0} \end{pmatrix}. \tag{10.248}$$

In view of (10.209) (p. 249), the BLUP(γ) can be written as

$$\begin{aligned} \mathrm{BLUP}(\boldsymbol{\gamma}) &= \tilde{\boldsymbol{\gamma}} = \mathbf{DZ}'\mathbf{W}^-(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) \\ &= \mathbf{DZ}'\mathbf{W}^-\tilde{\boldsymbol{\varepsilon}} = \mathbf{DZ}'\mathbf{\Sigma}^-\tilde{\boldsymbol{\varepsilon}} \\ &= \mathbf{DZ}'\mathbf{M}(\mathbf{M\Sigma M})^-\mathbf{My} = \mathbf{DZ}'\dot{\mathbf{M}}\mathbf{y}, \end{aligned} \tag{10.249}$$

where $\tilde{\boldsymbol{\varepsilon}}$ is the vector of the BLUE's residual and

$$\mathbf{W} = \mathbf{\Sigma} + \mathbf{XUX}' = \mathbf{ZDZ}' + \mathbf{R} + \mathbf{XUX}', \quad \mathscr{C}(\mathbf{W}) = \mathscr{C}(\mathbf{X} : \mathbf{\Sigma}), \tag{10.250}$$

and the matrix $\dot{\mathbf{M}}$ is defined as $\dot{\mathbf{M}} = \mathbf{M}(\mathbf{M\Sigma M})^-\mathbf{M}$.

As Henderson & Searle (1981a) point out, a difficulty with (10.243) is, in many applications, that the $n \times n$ matrix $\mathbf{\Sigma} = \mathbf{ZDZ}' + \mathbf{R}$ is often large and nondiagonal, so that inverting it is quite impractical. An alternative set of equations for solving $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\gamma}}$ is, as suggested by Henderson (1950, 1963),

$$\begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{pmatrix}. \tag{10.251}$$

If $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\gamma}}$ are solutions to (10.251), then $\mathbf{X}\tilde{\boldsymbol{\beta}}$ appears to be a BLUE of $\mathbf{X}\boldsymbol{\beta}$ and $\tilde{\boldsymbol{\gamma}}$ is a BLUP of $\boldsymbol{\gamma}$; see, e.g., Henderson, Kempthorne, Searle & von Krosigk (1959), Christensen (2002, p. 285), Searle (1997, p. 302), Searle, Casella & McCulloch (1992, §7.6), and Robinson (1991, p. 15). The equation (10.251) is called *Henderson's mixed model equation*. The proof is based on the equality

$$(\mathbf{ZDZ}' + \mathbf{R})^{-1} = \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1})^{-1}\mathbf{Z}'\mathbf{R}^{-1}, \tag{10.252}$$

which comes from the Duncan inversion formula (13.71) (p. 301).

Henderson (1950, 1963) minimized the following quadratic form with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ (keeping $\boldsymbol{\gamma}$ as a non-random vector):

$$f(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \begin{pmatrix} \mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\gamma} \\ \boldsymbol{\gamma} \end{pmatrix}' \begin{pmatrix} \mathbf{R} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\gamma} \\ \boldsymbol{\gamma} \end{pmatrix}, \tag{10.253}$$

where $\mathbf{R}$ and $\mathbf{D}$ are assumed to be positive definite. Denoting

$$\mathbf{y}_\# = \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix}, \quad \mathbf{X}_* = \begin{pmatrix} \mathbf{X} & \mathbf{Z} \\ \mathbf{0} & \mathbf{I}_q \end{pmatrix}, \quad \mathbf{V}_* = \begin{pmatrix} \mathbf{R} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{pmatrix}, \quad \boldsymbol{\pi} = \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{pmatrix}, \tag{10.254}$$

yields the following:

$$f(\boldsymbol{\beta}, \boldsymbol{\gamma}) = (\mathbf{y}_\# - \mathbf{X}_*\boldsymbol{\pi})'\mathbf{V}_*^{-1}(\mathbf{y}_\# - \mathbf{X}_*\boldsymbol{\pi}). \tag{10.255}$$

The normal equation resulting from minimizing $f(\boldsymbol{\beta},\boldsymbol{\gamma})$ is now precisely (10.251) and the minimizing values of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are (when $\mathbf{X}$ has full column rank)

$$(\mathbf{X}_*'\mathbf{V}_*^{-1}\mathbf{X}_*)^{-1}\mathbf{X}_*'\mathbf{V}_*^{-1}\mathbf{y}_\# = \begin{pmatrix}\tilde{\boldsymbol{\beta}}\\ \tilde{\boldsymbol{\gamma}}\end{pmatrix}. \tag{10.256}$$

Notice that

$$\mathbf{X}_*\tilde{\boldsymbol{\pi}} = \begin{pmatrix}\mathbf{X}\tilde{\boldsymbol{\beta}} + \mathbf{Z}\tilde{\boldsymbol{\gamma}}\\ \tilde{\boldsymbol{\gamma}}\end{pmatrix}. \tag{10.257}$$

We will return to Henderson's approach in Section 11.3 (p. 272).

Photograph 10.7 Stephen J. Haslett (Auckland, 2005).

For further references to mixed models, see, e.g., Christensen (2002, §12.3), Demidenko (2004), Harville (1990a), Hayes & Haslett (1999), Jiang (1997, 2007), McCulloch, Searle & Neuhaus (2008), Searle (1994, 1996, 1997), Rao (1987), Robinson (1991), and Haslett & Puntanen (2010b,d). For a discussion whether the effects should be considered fixed or random, see Gelman (2005, §6). For a proposal "To a non-Bayesian, all things are BLUPs", see Speed (1991, p. 44).

We complete this section with a result due to Isotalo, Möls & Puntanen (2006). They considered

$$\begin{aligned}\mathscr{A} &= \{\mathbf{y},\, \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2,\, \mathbf{R}\}\,,\\ \mathscr{B} &= \{\mathbf{y},\, \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\gamma}_2,\, \mathbf{D},\, \mathbf{R}\}\,,\end{aligned} \tag{10.258}$$

where $\mathscr{A}$ is a partitioned fixed model and $\mathscr{B}$ is a mixed model:

$$\mathscr{A}\colon \mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}\,, \quad \mathscr{B}\colon \mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\gamma}_2 + \boldsymbol{\varepsilon}\,, \tag{10.259}$$

where

$$\mathrm{E}(\boldsymbol{\gamma}_2) = \mathbf{0}\,, \quad \mathrm{cov}(\boldsymbol{\gamma}_2) = \mathbf{D}\,, \quad \mathrm{cov}(\boldsymbol{\gamma}_2, \boldsymbol{\varepsilon}) = \mathbf{0}\,, \tag{10.260}$$

and $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$, $\mathrm{cov}(\boldsymbol{\varepsilon}) = \mathbf{R}$, and $\mathrm{cov}(\mathbf{y} \mid \mathscr{B}) = \mathbf{X}_2\mathbf{D}\mathbf{X}_2' + \mathbf{R} := \boldsymbol{\Sigma}$. The proof of the following result is left as an exercise.

Proposition 10.11. *Consider the models $\mathscr{A}$ and $\mathscr{B}$ defined above. Then every representation of the* BLUE *for $\mathbf{M}_2\mathbf{X}_1\boldsymbol{\beta}_1$ under the model $\mathscr{A}$ is also the* BLUE *for $\mathbf{M}_2\mathbf{X}_1\boldsymbol{\beta}_1$ under the model $\mathscr{B}$ if and only if*

$$\mathscr{C}(\boldsymbol{\Sigma}\mathbf{X}_1^{\perp}) \subset \mathscr{C}(\mathbf{X}_2 : \mathbf{R}\mathbf{X}^{\perp})\,. \tag{10.261}$$

10.12 Linear Sufficiency

Photograph 10.8 Radosław Kala (Poznań, 2003).

The concept of linear sufficiency was introduced by Barnard (1963), Baksalary & Kala (1981a), and Drygas (1983)—who was the first to use the term linear sufficiency—while investigating those linear statistics $\mathbf{Fy}$, which are "sufficient" for estimation of $\mathbf{X}\boldsymbol{\beta}$ in the model $\mathscr{M}$. Formally, a linear statistic $\mathbf{Fy}$ is defined to be *linearly sufficient* for $\mathbf{X}\boldsymbol{\beta}$ under the model $\mathscr{M} = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{V}\}$ if there exists a matrix $\mathbf{A}$ such that $\mathbf{AFy}$ is the BLUE of $\mathbf{X}\boldsymbol{\beta}$. Baksalary & Kala (1981a, p. 913) illustrate the situation in the following "concrete" way (in our notation):

> If the vector $\mathbf{y}$ subject to the model $\{\mathbf{y},\, \mathbf{X}\beta,\, \mathbf{I}\}$ were transformed into the $\mathbf{w} = \mathbf{X}'\mathbf{y}$, then the BLUE of $\mathbf{X}\beta$, $\mathbf{X}\tilde{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{y}$, would be obtainable as a linear function of $\mathbf{w}$, namely as $\mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{w}$. If, however, the same transformation were adopted under the model $\{\mathbf{y},\, \mathbf{X}\beta,\, \mathbf{V}\}$ ($\mathbf{V}$ positive define but different from $\mathbf{I}_n$) then the BLUE of $\mathbf{X}\beta$, having now the form
>
> $$\mathbf{X}\tilde{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}\,,$$
>
> would no longer be obtainable as a linear function of $\mathbf{w} = \mathbf{X}'\mathbf{y}$ unless $\mathscr{C}(\mathbf{V}^{-1}\mathbf{X}) \subset \mathscr{C}(\mathbf{X})$. This exception might in fact be expected as the inclusion is a necessary and sufficient condition for the OLSE and BLUE to be identical.

Baksalary & Kala (1981a, p. 913) and Drygas (1983) showed that a linear statistic $\mathbf{Fy}$ is linearly sufficient for $\mathbf{X}\boldsymbol{\beta}$ under the model $\mathscr{M}$ if and only if the column space inclusion $\mathscr{C}(\mathbf{X}) \subset \mathscr{C}(\mathbf{WF}')$ holds; here $\mathbf{W} = \mathbf{V} + \mathbf{XLL}'\mathbf{X}'$ with $\mathbf{L}$ being an arbitrary matrix such that $\mathscr{C}(\mathbf{W}) = \mathscr{C}(\mathbf{X} : \mathbf{V})$. The hard-working reader may prove the following proposition.

Proposition 10.12. *Let* $\mathbf{W} = \mathbf{V} + \mathbf{XLL}'\mathbf{X}'$ *be an arbitrary matrix satisfying* $\mathscr{C}(\mathbf{W}) = \mathscr{C}(\mathbf{X} : \mathbf{V})$*. Then* $\mathbf{Fy}$ *is linearly sufficient for* $\mathbf{X}\boldsymbol{\beta}$ *under* $\mathscr{M} = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{V}\}$ *if and only if any of the following equivalent statements holds:*

(a) $\mathscr{C}(\mathbf{X}) \subset \mathscr{C}(\mathbf{WF}')$,

(b) $\mathscr{N}(\mathbf{F}) \cap \mathscr{C}(\mathbf{X} : \mathbf{V}) \subset \mathscr{C}(\mathbf{VX}^{\perp})$,

(c) $\operatorname{rank}(\mathbf{X} : \mathbf{VF}') = \operatorname{rank}(\mathbf{WF}')$,

(d) $\mathscr{C}(\mathbf{X}'\mathbf{F}') = \mathscr{C}(\mathbf{X}')$ *and* $\mathscr{C}(\mathbf{FX}) \cap \mathscr{C}(\mathbf{FVX}^{\perp}) = \{\mathbf{0}\}$,

(e) *the best linear predictor of* $\mathbf{y}$ *based on* $\mathbf{Fy}$, $\operatorname{BLP}(\mathbf{y}; \mathbf{Fy})$, *is almost surely equal to a linear function of* $\mathbf{Fy}$ *which does not depend on* $\boldsymbol{\beta}$.

Moreover, let $\mathbf{Fy}$ *be linearly sufficient for* $\mathbf{X}\boldsymbol{\beta}$ *under* $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{V}\}$*. Then each* BLUE *of* $\mathbf{X}\boldsymbol{\beta}$ *under the transformed model* $\{\mathbf{Fy}, \mathbf{FX}\boldsymbol{\beta}, \sigma^2\mathbf{FVF}'\}$ *is the* BLUE *of* $\mathbf{X}\boldsymbol{\beta}$ *under the original model* $\mathscr{M}$ *and vice versa.*

In addition to linear sufficiency, Drygas (1983) also considered related concepts of *linear minimal sufficiency* and *linear completeness*. A linearly sufficient statistic $\mathbf{Fy}$ is called linearly minimal sufficient for $\mathbf{X}\boldsymbol{\beta}$ under the model $\mathscr{M}$, if for any other linearly sufficient statistic $\mathbf{Sy}$, there exists a matrix $\mathbf{A}$ such that $\mathbf{Fy} = \mathbf{ASy}$ almost surely. Drygas (1983) showed that $\mathbf{Fy}$ is linearly minimal sufficient for $\mathbf{X}\boldsymbol{\beta}$ if and only if the equality

$$\mathscr{C}(\mathbf{X}) = \mathscr{C}(\mathbf{WF}') \tag{10.262}$$

holds. Moreover, Drygas (1983) called $\mathbf{Fy}$ linearly complete if for every linear transformation of it, $\mathbf{LFy}$, such that $\mathrm{E}(\mathbf{LFy}) = \mathbf{0}$, it follows that $\mathbf{LFy} = \mathbf{0}$ almost surely. According to Drygas (1983), $\mathbf{Fy}$ is linearly complete if and only if

$$\mathscr{C}(\mathbf{FV}) \subset \mathscr{C}(\mathbf{FX})\,. \tag{10.263}$$

It was then shown by Drygas (1983) that $\mathbf{Fy}$ is linearly minimal sufficient for $\mathbf{X}\boldsymbol{\beta}$ if and only if it is simultaneously linearly sufficient and linearly complete for $\mathbf{X}\boldsymbol{\beta}$. For related references, see Müller (1987), Müller, Rao & Sinha (1984), Liu (2002), and Sengupta & Jammalamadaka (2003, Ch. 11).

Baksalary & Kala (1986) extended the notions of linear sufficiency and linear minimal sufficiency to concern estimation of the given estimable parametric function $\mathbf{K}'\boldsymbol{\beta}$. They proved that $\mathbf{Fy}$ is linearly sufficient for $\mathbf{K}'\boldsymbol{\beta}$ under the model $\mathscr{M}$ if and only if

$$\mathscr{N}(\mathbf{FX} : \mathbf{FVX}^{\perp}) \subset \mathscr{N}(\mathbf{K}' : \mathbf{0})\,, \tag{10.264}$$

which can be shown to be equivalent to

$$\mathscr{C}[\mathbf{X}(\mathbf{X}'\mathbf{W}^{-}\mathbf{X})^{-}\mathbf{K}] \subset \mathscr{C}(\mathbf{WF}')\,. \tag{10.265}$$

Moreover, $\mathbf{Fy}$ is linearly minimal sufficient for $\mathbf{K}'\boldsymbol{\beta}$ if and only if

$$\mathscr{N}(\mathbf{FX} : \mathbf{FVX}^{\perp}) = \mathscr{N}(\mathbf{K}' : \mathbf{0})\,. \tag{10.266}$$

Isotalo & Puntanen (2006a) considered the model

$$\mathscr{M}_f = \left\{ \begin{pmatrix} \mathbf{y} \\ \mathbf{y}_f \end{pmatrix}, \begin{pmatrix} \mathbf{X}\boldsymbol{\beta} \\ \mathbf{X}_f\boldsymbol{\beta} \end{pmatrix}, \sigma^2 \begin{pmatrix} \mathbf{V} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{pmatrix} \right\}, \tag{10.267}$$

where $\mathbf{y}_f$ is the new future vector of observations, and investigated those linear statistics $\mathbf{Fy}$, which preserve enough information for obtaining the BLUP of $\mathbf{y}_f$ as a linear function of them. They call such statistics *linearly prediction sufficient* for $\mathbf{y}_f$. More formally, under the model $\mathscr{M}_f$, a linear statistic $\mathbf{Ty}$ is called linearly prediction sufficient for $\mathbf{y}_f$, if there exists a matrix $\mathbf{A}$ such that

$\mathbf{AFy}$ is the BLUP for $\mathbf{y}_f$. Moreover, $\mathbf{Ty}$ is called linearly minimal prediction sufficient for $\mathbf{y}_f$ if for any other linearly prediction sufficient $\mathbf{Ay}$, there exists a matrix $\mathbf{A}$ such that $\mathbf{Ty} = \mathbf{ASy}$ almost surely. Isotalo & Puntanen (2006a) proved the following.

Proposition 10.13. *Under model* $\mathscr{M}_f$, *a linear statistic* $\mathbf{Fy}$ *is linearly prediction sufficient for* $\mathbf{y}_f$ *if and only if*

$$\mathscr{N}(\mathbf{FX} : \mathbf{FVX}^{\perp}) \subset \mathscr{N}(\mathbf{X}_f : \mathbf{V}_{21}\mathbf{X}^{\perp})\,. \tag{10.268}$$

Moreover, $\mathbf{Fy}$ *is linearly minimal prediction sufficient if and only if the equality holds in* (10.268).

Various properties of linear sufficiency, particularly in a partitioned linear model, were considered by Tian & Puntanen (2009). They also introduced the concept of OLSE-sufficiency: $\mathbf{Fy}$ is linearly OLSE-sufficient for $\mathbf{K}'\boldsymbol{\beta}$ if there exist a matrix $\mathbf{A}$ such that $\mathrm{OLSE}(\mathbf{K}'\boldsymbol{\beta}) = \mathbf{AFy}$.

10.13 Admissibility

Consider the linear model $\mathscr{M} = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{V}\}$ and let $\mathbf{K}'\boldsymbol{\beta}$ be an estimable parametric function, $\mathbf{K} \in \mathbb{R}^{p\times q}$. Denote the set of all linear (homogeneous) estimators of $\mathbf{K}'\boldsymbol{\beta}$ as $\mathrm{LE}_q(\mathbf{y}) = \{\, \mathbf{Fy} : \mathbf{F} \in \mathbb{R}^{q\times n}\,\}$. The mean squared error matrix of $\mathbf{Fy}$ with respect to $\mathbf{K}'\boldsymbol{\beta}$ is defined as

$$\begin{aligned}\mathrm{MSEM}(\mathbf{Fy}; \mathbf{K}'\boldsymbol{\beta}) &= \mathrm{E}(\mathbf{Fy} - \mathbf{K}'\boldsymbol{\beta})(\mathbf{Fy} - \mathbf{K}'\boldsymbol{\beta})' \\ &= \mathrm{E}[\mathbf{F}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{F}'] + (\mathbf{FX} - \mathbf{K}')\boldsymbol{\beta}[(\mathbf{FX} - \mathbf{K}')\boldsymbol{\beta}]' \\ &= \mathrm{cov}(\mathbf{Fy}) + (\mathbf{FX} - \mathbf{K}')\boldsymbol{\beta}[(\mathbf{FX} - \mathbf{K}')\boldsymbol{\beta}]', \end{aligned} \tag{10.269}$$

and the quadratic risk of $\mathbf{Fy}$ under $\mathscr{M}$ is

$$\begin{aligned}\mathrm{risk}(\mathbf{Fy}; \mathbf{K}'\boldsymbol{\beta}) &= \mathrm{tr}[\mathrm{MSEM}(\mathbf{Fy}; \mathbf{K}'\boldsymbol{\beta})] \\ &= \sigma^2\,\mathrm{tr}(\mathbf{FVF}') + \|(\mathbf{FX} - \mathbf{K}')\boldsymbol{\beta}\|^2 \\ &= \mathrm{tr}[\mathrm{cov}(\mathbf{Fy})] + \|\mathrm{bias}\|^2. \end{aligned} \tag{10.270}$$

A linear estimator $\mathbf{Ay}$ is said to be *admissible* for $\mathbf{K}'\boldsymbol{\beta}$ among $\mathrm{LE}_q(\mathbf{y})$ under $\mathscr{M}$ if there does not exist $\mathbf{Fy} \in \mathrm{LE}_q(\mathbf{y})$ such that the inequality

$$\mathrm{risk}(\mathbf{Fy}; \mathbf{K}'\boldsymbol{\beta}) \le \mathrm{risk}(\mathbf{Ay}; \mathbf{K}'\boldsymbol{\beta}) \tag{10.271}$$

holds for every $(\boldsymbol{\beta}, \sigma^2) \in \mathbb{R}^p \times (0, \infty)$ and is strict for at least one point $(\boldsymbol{\beta}, \sigma^2)$. The set of admissible estimators of $\mathbf{K}'\boldsymbol{\beta}$ is denoted as $\mathrm{AD}(\mathbf{K}'\boldsymbol{\beta})$.

A general characterization of the class $\mathrm{AD}(\mathbf{K}'\boldsymbol{\beta})$ for an estimable vector $\mathbf{K}'\boldsymbol{\beta}$ and for nonsingular $\mathbf{V}$ is given in Rao (1976, Th. 6.6) and it was extended to a general Gauss–Markov model by Baksalary & Markiewicz (1988, p. 351). The result is presented in the following lemma.

Proposition 10.14. *Consider the model* $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{V}\}$ *and let* $\mathbf{K}'\boldsymbol{\beta}$ *be an estimable parametric function, i.e.,* $\mathbf{K}' = \mathbf{LX}$ *for some* $\mathbf{L} \in \mathbb{R}^{k\times n}$. *Then* $\mathbf{Fy}$ *is admissible for* $\mathbf{K}'\boldsymbol{\beta}$ *if and only if the following three conditions hold:*

$$\mathscr{C}(\mathbf{VF}') \subset \mathscr{C}(\mathbf{X})\,, \tag{10.272a}$$

$$\mathbf{FVL}' - \mathbf{FVF}' \geq_{\mathsf{L}} \mathbf{0}\,, \tag{10.272b}$$

$$\mathscr{C}[(\mathbf{F}-\mathbf{L})\mathbf{X}] = \mathscr{C}[(\mathbf{F}-\mathbf{L})\mathbf{S}]\,, \tag{10.272c}$$

where $\mathbf{S}$ *is a matrix satisfying* $\mathscr{C}(\mathbf{S}) = \mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{V})$.


Photograph 10.9 Augustyn Markiewicz (Tomar, 2006).

It is obvious that under a weakly singular linear model, i.e., when $\mathscr{C}(\mathbf{X}) \subset \mathscr{C}(\mathbf{V})$, the condition (10.272c) is trivially satisfied and hence the first two conditions guarantee that $\mathbf{Fy}$ is admissible for $\mathbf{K}'\boldsymbol{\beta}$. For further references on the linear admissibility, see, for example, Baksalary & Markiewicz (1989, 1990), Baksalary, Markiewicz & Rao (1995), Mathew, Rao & Sinha (1984), and Drygas & Zmyślony (1988).

We may complete this section by mentioning the paper by Markiewicz & Puntanen (2009), where the admissibility and linear sufficiency are studied in a partitioned linear model.

10.14 Exercises

10.1. Complete the proof of Proposition 10.1 (p. 218).

10.2. Solve the problems presented in Section 10.6 (p. 234).

10.3. Confirm that (10.152) (p. 240) holds:

$$\psi = \tfrac{1}{2}\|\mathbf{HV}-\mathbf{VH}\|_F^2 = \|\mathbf{HVM}\|_F^2 = \operatorname{tr}(\mathbf{HVMV}) = \operatorname{tr}(\mathbf{HV}^2) - \operatorname{tr}(\mathbf{HV})^2.$$

10.4 (Continued . . .). Let the columns of $\mathbf{X}_*$ and $\mathbf{Z}$ comprise the orthonormal bases for $\mathscr{C}(\mathbf{X})$ and $\mathscr{C}(\mathbf{X})^\perp$, respectively. Show that

$$\psi = \|\mathbf{X}_*'\mathbf{VZ}\|_F^2 = \operatorname{tr}(\mathbf{X}_*'\mathbf{V}^2\mathbf{X}_*) - \operatorname{tr}(\mathbf{X}_*'\mathbf{VX}_*)^2.$$

10.5. Consider the models $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$ and $\underline{\mathscr{M}} = \{\mathbf{y}, \underline{\mathbf{X}}\boldsymbol{\beta}, \mathbf{V}\}$, where $\underline{\mathbf{X}} = \mathbf{XA}$ and $\mathbf{A}$ is nonsingular. Show that

$$\operatorname{cov}[\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta} \mid \mathscr{M})] = \operatorname{cov}[\mathrm{BLUE}(\underline{\mathbf{X}}\boldsymbol{\beta} \mid \underline{\mathscr{M}})]\,.$$

10.6. Consider the linear model $\{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$, where $\mathbf{V}$ is positive definite and the columns of $\mathbf{X}_b$ form a basis for $\mathscr{C}(\mathbf{X})$. Show that then

$$\begin{aligned}\operatorname{cov}[\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta})] &= \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}' = \mathbf{X}_b(\mathbf{X}_b'\mathbf{V}^{-1}\mathbf{X}_b)^{-1}\mathbf{X}_b' \\ &= \mathbf{H}(\mathbf{H}\mathbf{V}^{-1}\mathbf{H})^{-}\mathbf{H} = (\mathbf{H}\mathbf{V}^{-1}\mathbf{H})^{+}.\end{aligned}$$

10.7. Show that under a weakly singular model, i.e., $\mathscr{C}(\mathbf{X}) \subset \mathscr{C}(\mathbf{V})$, we have

$$\begin{aligned}\operatorname{cov}[\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta})] &= \mathbf{X}(\mathbf{X}'\mathbf{V}^{-}\mathbf{X})^{-}\mathbf{X}' = \mathbf{X}_b(\mathbf{X}_b'\mathbf{V}^{-}\mathbf{X}_b)^{-1}\mathbf{X}_b' \\ &= \mathbf{H}(\mathbf{H}\mathbf{V}^{-}\mathbf{H})^{-}\mathbf{H} = (\mathbf{H}\mathbf{V}^{-}\mathbf{H})^{+},\end{aligned}$$

where the generalized inverses can be any generalized inverses; in the last representation, however, we must have the Moore–Penrose inverse.

10.8. Consider models $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$ and $\mathscr{M}^{*} = \{\mathbf{V}^{-1/2}\mathbf{y}, \mathbf{V}^{-1/2}\mathbf{X}\boldsymbol{\beta}, \mathbf{I}\}$ where $\mathbf{V} \in \mathrm{PD}_n$. Show that $\mathbf{V}^{-1/2}\mathbf{y}$ is linearly sufficient for $\mathbf{X}\boldsymbol{\beta}$ under $\mathscr{M}$ and thereby the BLUE for $\mathbf{X}\boldsymbol{\beta}$ under $\mathscr{M}^{*}$ equals the BLUE under $\mathscr{M}$. See also Exercise 0.16 (p. 50).

10.9. Consider the models $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$ and $\mathscr{M}_W = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{W}\}$, where $\mathbf{W} = \mathbf{V} + \mathbf{XLL}'\mathbf{X}'$, and $\mathscr{C}(\mathbf{W}) = \mathscr{C}(\mathbf{X} : \mathbf{V})$. Show the following:

(a) $\operatorname{cov}(\mathbf{X}\hat{\boldsymbol{\beta}} \mid \mathscr{M}_W) = \mathbf{HWH}$,
(b) $\operatorname{cov}(\mathbf{X}\tilde{\boldsymbol{\beta}} \mid \mathscr{M}_W) = \mathbf{X}(\mathbf{X}'\mathbf{W}^{-}\mathbf{X})^{-}\mathbf{X}' = (\mathbf{H}\mathbf{W}^{-}\mathbf{H})^{+}$,
(c) $\operatorname{cov}(\mathbf{X}\hat{\boldsymbol{\beta}} \mid \mathscr{M}_W) - \operatorname{cov}(\mathbf{X}\tilde{\boldsymbol{\beta}} \mid \mathscr{M}_W) = \operatorname{cov}(\mathbf{X}\hat{\boldsymbol{\beta}} \mid \mathscr{M}) - \operatorname{cov}(\mathbf{X}\tilde{\boldsymbol{\beta}} \mid \mathscr{M})$,
(d) $\{\,\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta} \mid \mathscr{M}_W)\,\} = \{\,\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta} \mid \mathscr{M})\,\}$.

Liski, Puntanen & Wang (1992, p. 126).

10.10. Let $\mathbf{B}$ be an $n \times n$ matrix and let $\mathbf{A}$ be an $n \times p$ matrix of rank a. Show that then, for any $n \times a$ matrix $\mathbf{A}_*$ such that $\mathscr{C}(\mathbf{A}_*) = \mathscr{C}(\mathbf{A})$ and $\mathbf{A}_*'\mathbf{A}_* = \mathbf{I}_a$, the equality

$$\mathbf{A}_*(\mathbf{A}_*'\mathbf{B}\mathbf{A}_*)^{+}\mathbf{A}_*' = \mathbf{P}_{\mathbf{A}}(\mathbf{P}_{\mathbf{A}}\mathbf{B}\mathbf{P}_{\mathbf{A}})^{+}\mathbf{P}_{\mathbf{A}} = (\mathbf{P}_{\mathbf{A}}\mathbf{B}\mathbf{P}_{\mathbf{A}})^{+}$$

is always true, whereas the equality

$$\mathbf{A}_*(\mathbf{A}_*'\mathbf{B}\mathbf{A}_*)^{+}\mathbf{A}_*' = \mathbf{A}(\mathbf{A}'\mathbf{B}\mathbf{A})^{+}\mathbf{A}'$$

holds if and only if $\mathscr{C}(\mathbf{A}'\mathbf{A}\mathbf{A}'\mathbf{B}\mathbf{A}) = \mathscr{C}(\mathbf{A}'\mathbf{B}\mathbf{A})$.

Baksalary, Puntanen & Styan (1990b, Cor. 2).

10.11. In an example of autocorrelation in a coin-tossing problem, Jensen & McDonald (1976) consider the following situation:

$y_1 = 1$, if coin results head in the first toss, 0 otherwise,
$y_2 =$ number of heads observed in the next 2 tosses,
$y_3 =$ number of heads observed in the next 3 tosses,
...
$y_n =$ number of heads observed in the last n tosses.

Total number of tosses is therefore $n(n+1)/2$, all tosses being independent. Let π denote the probability that the toss of a coin results in a head. Denoting $\mathbf{y} = (y_1, \dots, y_n)'$, we can write a linear model

$$\mathbf{y} = \pi\mathbf{x} + \boldsymbol{\varepsilon}, \quad \operatorname{cov}(\mathbf{y}) = \sigma^2\mathbf{V}.$$

Construct $\mathbf{x}$, $\mathbf{V}$, express σ^2 as a function of π, and find $\operatorname{OLSE}(\pi) = \hat{\pi}$, and $\operatorname{BLUE}(\pi) = \tilde{\pi}$, and show that $\frac{\operatorname{var}(\tilde{\pi})}{\operatorname{var}(\hat{\pi})} \to \frac{8}{9}$ when $n \to \infty$. Moreover, confirm that when y_i's are defined so that

$$y_i = \text{number of heads observed in the first } i \text{ tosses,}$$

so that the total number of tosses is n, then the resulting $\mathbf{V}$ is of the form $\mathbf{A}$ in (10.232) (p. 252). Amaze yourself that in both situations (regarding the definition of y_i) we get

$$\tilde{\pi} = \frac{\text{total number of heads}}{\text{total number of tosses}}.$$

10.12 (Continued ...). Show that the correlation matrix $\mathbf{R}$, say, associated with $\mathbf{A}$, is

$$\mathbf{R} = \begin{pmatrix} 1 & \sqrt{\frac{1}{2}} & \dots & \sqrt{\frac{1}{n}} \\ \sqrt{\frac{1}{2}} & 1 & \dots & \sqrt{\frac{2}{n}} \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{\frac{1}{n}} & \sqrt{\frac{2}{n}} & \dots & 1 \end{pmatrix},$$

and $\det(\mathbf{R}) = \det\big(\mathbf{A}_\delta^{-1/2}\mathbf{A}\mathbf{A}_\delta^{-1/2}\big) = \frac{\det(\mathbf{A})}{\det(\mathbf{A}_\delta)} = \frac{1}{n!}$, where $\mathbf{A}_\delta = \operatorname{diag}(\mathbf{A})$.

10.13. Consider matrices

$$\mathbf{A} = \begin{pmatrix} a_1 & a_1 & a_1 & \dots & a_1 \\ a_1 & a_2 & a_2 & \dots & a_2 \\ a_1 & a_2 & a_3 & \dots & a_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_1 & a_2 & a_3 & \dots & a_n \end{pmatrix}, \quad \mathbf{F} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ -1 & 1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & 0 & \dots & -1 & 1 \end{pmatrix},$$

where a_i are real numbers, $i = 1, \dots, n$. Show that $\mathbf{F}\mathbf{A}\mathbf{F}' = \mathbf{D}$, where $\mathbf{D} = \operatorname{diag}(a_1, a_2 - a_1, a_3 - a_2, \dots, a_n - a_{n-1})$. Hence obviously

$$\det(\mathbf{A}) = a_1(a_2 - a_1)(a_3 - a_2)\cdots(a_n - a_{n-1})\,,$$

and $\mathbf{A}$ is positive definite if and only if $0 < a_1 < a_2 < \cdots < a_n$. Confirm that the inverse of $\mathbf{A}$ is

$$\begin{aligned}\mathbf{A}^{-1} &= \mathbf{F}'\mathbf{D}^{-1}\mathbf{F} \\ &= \begin{pmatrix} \alpha_1+\alpha_2 & -\alpha_2 & 0 & 0 & \dots & 0 & 0 \\ -\alpha_2 & \alpha_2+\alpha_3 & -\alpha_3 & 0 & \dots & 0 & 0 \\ 0 & -\alpha_3 & \alpha_3+\alpha_4 & -\alpha_4 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \alpha_{n-1}-\alpha_n & -\alpha_n \\ 0 & 0 & 0 & 0 & \dots & -\alpha_n & \alpha_n \end{pmatrix},\end{aligned}$$

where $\alpha_i = \frac{1}{a_i - a_{i-1}}$, $i = 1, \dots, n$; $a_0 = 0$.

Chu, Puntanen & Styan (2011), Neudecker, Trenkler & Liu (2009), Moyé (2006, Appendix B).

10.14. Prove Proposition 10.11 (p. 256).

10.15. Consider the model $\{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \mathbf{V}\}$. Show that the condition $\mathbf{X}\mathbf{X}'\mathbf{V} = \mathbf{V}\mathbf{X}\mathbf{X}'$ is sufficient for $\mathrm{OLSE}(\mathbf{X}\boldsymbol{\beta}) = \mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta})$ but not necessary.

Trenkler & Trenkler (2008), Puntanen (2010b).

10.16 (Canner's data). Consider the model $\{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \mathbf{V}\}$, where $\mathbf{X} = \left(\begin{smallmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 1 \end{smallmatrix}\right)$, and

$$\mathbf{V} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & c & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & \varrho & \varrho \\ \varrho & 1 & \varrho \\ \varrho & \varrho & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & c & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & c\varrho & \varrho \\ c\varrho & c^2 & c\varrho \\ \varrho & c\varrho & 1 \end{pmatrix}.$$

(a) Prove that $\mathbf{V} >_{\mathsf{L}} \mathbf{0} \iff c \neq 0$ and $-1/2 < \varrho < 1$.

(b) Choose c and ϱ so that $0 < c < \varrho < 1$. Show that $\hat{\beta}_1$ and $\tilde{\beta}_1$ are identical so that the OLSE- and BLUE-lines are parallel.

(c) Is it possible that all data points are below or above the BLUE-line $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x$? (Yes!)

(d) Draw the OLSE- and BLUE-lines when $\varrho = 0.8$, $c = 0.6$ and $\mathbf{y} = (4, 8, 19)'$.

Canner (1969), Chu, Isotalo, Puntanen & Styan (2007, p. 3348).

10.17. Prove (10.95) (p. 231):

$$\mathscr{C}[(\boldsymbol{\Sigma} \otimes \mathbf{I}_n)(\mathbf{I}_d \otimes \mathbf{X})] = \mathscr{C}(\boldsymbol{\Sigma} \otimes \mathbf{X}) \subset \mathscr{C}(\mathbf{I}_d \otimes \mathbf{X})\,.$$

Hint: Show that $(\boldsymbol{\Sigma} \otimes \mathbf{I}_n)(\mathbf{I}_d \otimes \mathbf{X}) = (\mathbf{I}_d \otimes \mathbf{X})(\boldsymbol{\Sigma} \otimes \mathbf{I}_p)$ or confirm that $\mathscr{C}(\mathbf{A}\mathbf{U} \otimes \mathbf{B}\mathbf{V}) \subset \mathscr{C}(\mathbf{A} \otimes \mathbf{B})$ for conformable matrices involved. See also Exercise 0.23 (p. 52).

10.18 (Pandora's Box). Consider the model $\{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{V}\}$ and denote

$$\boldsymbol{\Gamma} = \begin{pmatrix} \mathbf{V} & \mathbf{X} \\ \mathbf{X}' & \mathbf{0} \end{pmatrix}, \quad \text{and} \quad \mathbf{C} = \begin{pmatrix} \mathbf{C}_1 & \mathbf{C}_2 \\ \mathbf{C}_3 & -\mathbf{C}_4 \end{pmatrix} = \begin{pmatrix} \mathbf{V} & \mathbf{X} \\ \mathbf{X}' & \mathbf{0} \end{pmatrix}^{-} \in \{\boldsymbol{\Gamma}^-\}\,.$$

Confirm the following:

(a) $\mathscr{C}\begin{pmatrix} \mathbf{V} \\ \mathbf{X}' \end{pmatrix} \cap \mathscr{C}\begin{pmatrix} \mathbf{X} \\ \mathbf{0} \end{pmatrix} = \{\mathbf{0}\}$, (b) $\operatorname{rk}\begin{pmatrix} \mathbf{V} & \mathbf{X} \\ \mathbf{X}' & \mathbf{0} \end{pmatrix} = \operatorname{rk}(\mathbf{V} : \mathbf{X}) + \operatorname{rk}(\mathbf{X})$,

(i) $\mathbf{X}\mathbf{C}_2'\mathbf{X} = \mathbf{X}$, $\mathbf{X}\mathbf{C}_3\mathbf{X} = \mathbf{X}$,

(ii) $\mathbf{X}\mathbf{C}_4\mathbf{X}' = \mathbf{X}\mathbf{C}_4'\mathbf{X}' = \mathbf{V}\mathbf{C}_3'\mathbf{X}' = \mathbf{X}\mathbf{C}_3\mathbf{V} = \mathbf{V}\mathbf{C}_2\mathbf{X}' = \mathbf{X}\mathbf{C}_2'\mathbf{V}$,

(iii) $\mathbf{X}'\mathbf{C}_1\mathbf{X}$, $\mathbf{X}'\mathbf{C}_1\mathbf{V}$ and $\mathbf{V}\mathbf{C}_1\mathbf{X}$ are all zero matrices,

(iv) $\mathbf{V}\mathbf{C}_1\mathbf{V}\mathbf{C}_1\mathbf{V} = \mathbf{V}\mathbf{C}_1\mathbf{V} = \mathbf{V}\mathbf{C}_1'\mathbf{V}\mathbf{C}_1\mathbf{V} = \mathbf{V}\mathbf{C}_1'\mathbf{V}$,

(v) $\operatorname{tr}(\mathbf{V}\mathbf{C}_1) = \operatorname{rk}(\mathbf{V} : \mathbf{X}) - \operatorname{rk}(\mathbf{X}) = \operatorname{tr}(\mathbf{V}\mathbf{C}_1')$,

(vi) $\mathbf{V}\mathbf{C}_1\mathbf{V}$ and $\mathbf{X}\mathbf{C}_4\mathbf{X}'$ are invariant for any choice of $\mathbf{C}_1$ and $\mathbf{C}_4$,

(vii) $\mathbf{X}\tilde{\boldsymbol{\beta}} = \mathbf{X}\mathbf{C}_2'\mathbf{y}$,

(viii) $\operatorname{cov}(\mathbf{X}\tilde{\boldsymbol{\beta}}) = \mathbf{X}\mathbf{C}_4\mathbf{X}'$,

(ix) $\tilde{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}} = \mathbf{V}\mathbf{C}_1\mathbf{y}$,

(x) the BLUE of an estimable function $\mathbf{k}'\boldsymbol{\beta}$ is $\mathbf{k}'\tilde{\boldsymbol{\beta}} = \mathbf{k}'\mathbf{C}_2'\mathbf{y} = \mathbf{k}'\mathbf{C}_3\mathbf{y}$,

(xi) $\operatorname{var}(\mathbf{k}'\tilde{\boldsymbol{\beta}}) = \sigma^2\mathbf{k}'\mathbf{C}_4\mathbf{k}$,

(xii) $\tilde{\sigma}^2 = \mathbf{y}'\mathbf{C}_1\mathbf{y}/f$ is an unbiased estimator of σ^2; $f = \operatorname{rk}(\mathbf{V} : \mathbf{X}) - \operatorname{rk}(\mathbf{X}) = \operatorname{rk}(\mathbf{V}\mathbf{M})$.

Rao (1971c, 1972), Rao (1973a, pp. 298–300), Hall & Meyer (1975), Mitra (1982, 1989), Pringle & Rayner (1970, 1971), Rao & Yanai (1985b).

10.19. Consider the linear model $\mathscr{A}$ which is partitioned row-wise so that

$$\mathscr{A} = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \mathbf{V}\} = \left\{ \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix}, \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}, \begin{pmatrix} \mathbf{V}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_2 \end{pmatrix} \right\},$$

where $\mathbf{X}_i \in \mathbb{R}^{n_i \times p}$, $\mathbf{V}_i \in \mathrm{NND}_{n_i}$, $i = 1, 2$, and $n_1 + n_2 = n$. The above partition leads to two submodels $\mathscr{A}_i = \{\mathbf{y}_i,\, \mathbf{X}_i\boldsymbol{\beta},\, \mathbf{V}_i\}, i = 1, 2$, and special choices of $\mathbf{X}_i$ and $\mathbf{V}_i$ correspond to some well-known statistical estimators. Confirm the following statements:

(a) Under the choice $\mathbf{V}_1 = \mathbf{I}_{n_1}$, $\mathbf{V}_2 = \mathbf{0}$ and $\mathbf{X}_2 = \mathbf{R}$, the BLUE of $\boldsymbol{\beta}$ is the restricted OLSE

$$\hat{\boldsymbol{\beta}}_{\mathrm{r}} = \hat{\boldsymbol{\beta}}_1 - (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}}_1 - \mathbf{r})\,,$$

where $\hat{\boldsymbol{\beta}}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}_1$ with $\mathbf{X}_1$ of full column rank, $\mathbf{R}$ of full row rank and the restriction being $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$.

(b) If $\mathbf{V}_1 = \mathbf{I}_{n_1}$ and $\mathbf{X}_2 = \mathbf{R}$, then the BLUE of $\boldsymbol{\beta}$ is the ordinary Theil–Goldberger mixed estimator

$$\hat{\boldsymbol{\beta}}_{\mathrm{m}} = (\mathbf{X}_1'\mathbf{X}_1 + \mathbf{R}'\mathbf{V}_2^{-1}\mathbf{R})^{-1}(\mathbf{X}_1'\mathbf{y}_1 + \mathbf{R}'\mathbf{V}_2^{-1}\mathbf{y}_2)\,,$$

where $\mathbf{X}_1$ has full column rank and $\mathbf{V}_2$ is positive definite.

Theil & Goldberger (1961), Theil (1968), Diderrich (1985).

10.20 (Continued ...). Show that under the model $\mathscr{A}$ the following statements are equivalent:

(a) $\mathrm{OLSE}(\mathbf{X}\boldsymbol{\beta}) = \mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta})$,

(b) (i) $\mathscr{C}(\mathbf{V}_1\mathbf{X}_1) \subset \mathscr{C}(\mathbf{X}_1)$, $\mathscr{C}(\mathbf{V}_2\mathbf{X}_2) \subset \mathscr{C}(\mathbf{X}_2)$, and
  (ii) $\mathbf{L}\mathbf{X}_1^{-}\mathbf{V}_1\mathbf{X}_1 = \mathbf{L}\mathbf{X}_2^{-}\mathbf{V}_2\mathbf{X}_2$ for some $\mathbf{X}_1^{-}$ and $\mathbf{X}_2^{-}$, and $\mathbf{L}$ is any matrix satisfying $\mathscr{C}(\mathbf{L}') = \mathscr{C}(\mathbf{X}_1') \cap \mathscr{C}(\mathbf{X}_2')$.

Groß, Puntanen & Trenkler (1996).

10.21. Consider the model $\{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$.

(a) Given $\mathbf{X} \in \mathbb{R}^{n\times p}$ and $\mathbf{V} \in \mathrm{NND}_n$, what is the set $\mathcal{A}$ of matrices $\mathbf{A} \in \mathbb{R}^{k\times n}$ satisfying $\mathbf{A}\cdot\mathrm{OLSE}(\mathbf{X}\boldsymbol{\beta}) = \mathbf{A}\cdot\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta})$?

(b) Given $\mathbf{X} \in \mathbb{R}^{n\times p}$ and $\mathbf{A} \in \mathbb{R}^{k\times n}$, what is the set $\mathcal{V}$ of nonnegative definite matrices $\mathbf{V}$ satisfying $\mathbf{A}\cdot\mathrm{OLSE}(\mathbf{X}\boldsymbol{\beta}) = \mathbf{A}\cdot\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta})$?

(c) Given $\mathbf{X} \in \mathbb{R}^{n\times p}$, $\mathbf{A} \in \mathbb{R}^{k\times n}$, and $\mathbf{V} \in \mathrm{NND}_n$, what is the set $\mathcal{Y}$ of vectors $\mathbf{y}$ satisfying $\mathbf{A}\cdot\mathrm{OLSE}(\mathbf{X}\boldsymbol{\beta}) = \mathbf{A}\cdot\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta})$?

Krämer (1980b), Baksalary (1984), Jaeger & Krämer (1998),
Groß & Trenkler (1997), Groß, Trenkler & Werner (2001).

10.22. Consider the model $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$ and let $\mathbf{K}'\boldsymbol{\beta}$ be an estimable parametric function with $\mathbf{K}$ having full column rank. Show that then the following statements are equivalent:

(a) $\mathbf{Fy}$ is unbiased for $\mathbf{K}'\boldsymbol{\beta}$ and $\mathbf{Fy}$ is linearly sufficient for $\mathbf{K}'\boldsymbol{\beta}$,

(b) $\mathbf{Fy}$ is the BLUE for $\mathbf{K}'\boldsymbol{\beta}$.

Baksalary & Kala (1986, p. 334).

10.23. Consider the model $\mathscr{A} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$, where

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}, \quad \mathbf{V} = \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{pmatrix}, \quad \mathbf{X}_i \in \mathbb{R}^{n_i\times p}.$$

Suppose that $\mathbf{X}_1$ has full column rank and $\mathbf{G}\mathbf{y}_1$ is the BLUE for $\boldsymbol{\beta}$ under the submodel $\mathscr{A}_1 = \{\mathbf{y}_1, \mathbf{X}_1, \mathbf{V}_{11}\}$, i.e., $\mathbf{G}(\mathbf{X}_1 : \mathbf{V}_{11}\mathbf{M}_1) = (\mathbf{X}_1 : \mathbf{0})$.

(a) Show that $\begin{pmatrix} \mathbf{M}_1 & -\mathbf{G}'\mathbf{X}_2' \\ \mathbf{0} & \mathbf{I}_{n_2} \end{pmatrix} \in \left\{ \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}^{\perp} \right\}$.

(b) Show that $\mathbf{y}_2$ is superfluous with respect to estimating $\boldsymbol{\beta}$ in the model $\mathscr{A}$, i.e., $\mathbf{G}\mathbf{y}_1 = (\mathbf{G} : \mathbf{0})\mathbf{y} = \mathbf{Fy}$ is linearly sufficient for $\boldsymbol{\beta}$, if and only if

$$\mathbf{G}\mathbf{V}_{11}\mathbf{M}_1 = \mathbf{0} \quad \text{and} \quad \mathbf{G}\mathbf{V}_{12} = \mathbf{G}\mathbf{V}_{11}\mathbf{G}'\mathbf{X}_2'\,,$$

which reduces to $\mathbf{X}_2\mathbf{G}\mathbf{V}_{11} = \mathbf{V}_{21}\mathbf{G}'\mathbf{X}_1'$, and in particular, if $\mathbf{V}$ is positive definite, to $\mathbf{X}_2 = \mathbf{V}_{21}\mathbf{V}_{11}^{-1}\mathbf{X}_1$.

Baksalary & Kala (1986, p. 337), Gourieroux & Monfort (1980, p. 1093).

10.24. Let $\mathbf{X}_1\boldsymbol{\beta}_1$ be estimable under $\{\mathbf{y},\, \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2,\, \mathbf{V}\}$ and denote $\dot{\mathbf{M}}_2 = \mathbf{M}_2(\mathbf{M}_2\mathbf{W}_1\mathbf{M}_2)^-\mathbf{M}_2$, where $\mathbf{W}_1 = \mathbf{V} + \mathbf{X}_1\mathbf{X}_1'$. Show that $\mathbf{T}\mathbf{y}$ is linearly minimal sufficient for $\mathbf{X}_1\boldsymbol{\beta}_1$ if and only if any of the following statements holds:

(a) $\mathscr{N}(\mathbf{T}\mathbf{X}_1 : \mathbf{T}\mathbf{X}_2 : \mathbf{T}\mathbf{V}\mathbf{X}^{\perp}) = \mathscr{N}(\mathbf{X}_1 : \mathbf{0} : \mathbf{0})$,

(b) $\mathscr{C}(\mathbf{W}_1\dot{\mathbf{M}}_2\mathbf{X}_1) = \mathscr{C}(\mathbf{W}_1\mathbf{T}')$,

(c) $\mathscr{N}(\mathbf{T}) \cap \mathscr{C}(\mathbf{W}_1) = \mathscr{N}(\mathbf{X}_1'\dot{\mathbf{M}}_2) \cap \mathscr{C}(\mathbf{W}_1)$,

(d) $\mathbf{T}\mathbf{y}$ is linearly sufficient for $\mathbf{X}_1\boldsymbol{\beta}_1$ and there exists $\mathbf{L}$ such that $\mathbf{T}\mathbf{y} = \mathbf{L}\mathbf{X}_1'\dot{\mathbf{M}}_2\mathbf{y}$ almost surely.

Confirm that $\mathbf{X}_1'\dot{\mathbf{M}}_2\mathbf{y}$ is linearly minimal sufficient for $\mathbf{X}_1\boldsymbol{\beta}_1$.

Isotalo & Puntanen (2006b, Th. 2).

10.25. Prove that equation (10.198) (p. 247),

$$\mathbf{A}(\mathbf{X} : \mathbf{V}\mathbf{X}^{\perp}) = (\mathbf{X}_f : \mathbf{V}_{21}\mathbf{X}^{\perp}),$$

where $\mathscr{C}(\mathbf{X}_f') \subset \mathscr{C}(\mathbf{X}')$, has a solution to $\mathbf{A}$.

10.26. Consider the models $\mathscr{M}_i = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \mathbf{V}_i\}, i = 1, 2$, and let $\mathbf{W}_i$ be defined so that $\mathbf{W}_i = \mathbf{V}_i + \mathbf{X}\mathbf{U}_i\mathbf{U}_i'\mathbf{X}'$, where $\mathscr{C}(\mathbf{W}_i) = \mathscr{C}(\mathbf{X} : \mathbf{V}_i), i = 1, 2$. Suppose that $\mathbf{W}_1$ has the eigenvalue decomposition $\mathbf{W}_1 = \mathbf{Z}\boldsymbol{\Lambda}\mathbf{Z}'$, where the columns of $\mathbf{Z} \in \mathbb{R}^{n\times w_1}$ are orthonormal eigenvectors of $\mathbf{W}_1$ with respect to nonzero eigenvalues $\lambda_1 \geq \cdots \geq \lambda_{w_1} > 0$ of $\mathbf{W}_1$, $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \dots, \lambda_{w_1})$, and $w_1 = \operatorname{rank}(\mathbf{W}_1)$. Denote $\mathbf{Q} = \boldsymbol{\Lambda}^{-1/2}\mathbf{Z}' \in \mathbb{R}^{w_1\times n}$. Premultiplying the model $\mathscr{M}_2$ by $\mathbf{Q}$ gives $\mathscr{M}_2^* = \{\mathbf{Q}\mathbf{y}, \mathbf{Q}\mathbf{X}\boldsymbol{\beta}, \mathbf{Q}\mathbf{V}_2\mathbf{Q}'\}$. Show that $\operatorname{OLSE}(\mathbf{X}\boldsymbol{\beta} \mid \mathscr{M}_2^*) = \operatorname{BLUE}(\mathbf{X}\boldsymbol{\beta} \mid \mathscr{M}_2^*)$ if and only if

$$\mathbf{X}'\mathbf{W}_1^+\mathbf{V}_2\mathbf{P}_{\mathbf{W}_1}\mathbf{M} = \mathbf{0}.$$

Suppose that $\mathbf{Q}\mathbf{y}$ is linearly sufficient for $\mathbf{X}\boldsymbol{\beta}$ under $\mathscr{M}_2$. Show that then the above condition becomes $\mathbf{X}'\mathbf{W}_1^+\mathbf{V}_2\mathbf{M} = \mathbf{0}$.

Hauke, Markiewicz & Puntanen (2011).

Chapter 11
General Solution to $\mathbf{AYB} = \mathbf{C}$

> *It seemed the world was divided into good people and bad people. The good ones slept their nights better, while the bad ones seemed to enjoy the wakening hours much more.*
>
> WOODY ALLEN: *The Condemned*

In almost every chapter of this book we meet linear equations whose explicit solutions we wish to write up. This is what generalized inverses make nicely possible. In this chapter we represent the basic result which shows how to express the general solution for $\mathbf{Y}$ satisfying the equation $\mathbf{AYB} = \mathbf{C}$.

Theorem 11 (General solution to $\mathbf{AYB} = \mathbf{C}$). *A necessary and sufficient condition for the equation*

$$\mathbf{AYB} = \mathbf{C} \tag{11.1}$$

to have a solution (for $\mathbf{Y}$) is that, for some $\mathbf{A}^- \in \{\mathbf{A}^-\}$, $\mathbf{B}^- \in \{\mathbf{B}^-\}$,

$$\mathbf{AA}^-\mathbf{CB}^-\mathbf{B} = \mathbf{C}\,, \tag{11.2}$$

in which case the general solution is

$$\mathbf{Y} = \mathbf{A}^-\mathbf{CB}^- + \mathbf{Z} - \mathbf{A}^-\mathbf{AZBB}^-, \tag{11.3}$$

where $\mathbf{Z}$ is an arbitrary matrix, and $\mathbf{A}^-$ and $\mathbf{B}^-$ are fixed (but arbitrary) generalized inverses. In particular, if $\mathscr{C}(\mathbf{C}) \subset \mathscr{C}(\mathbf{A})$, the general solution to $\mathbf{AY} = \mathbf{C}$ is

$$\mathbf{Y} = \mathbf{A}^-\mathbf{C} + (\mathbf{I} - \mathbf{A}^-\mathbf{A})\mathbf{Z}\,. \tag{11.4}$$

Similarly, if $\mathscr{C}(\mathbf{C}') \subset \mathscr{C}(\mathbf{B}')$, the general solution to $\mathbf{YB} = \mathbf{C}$ is

$$\mathbf{Y} = \mathbf{CB}^- + \mathbf{Z}(\mathbf{I} - \mathbf{BB}^-)\,. \tag{11.5}$$

Proof. If there exists $\mathbf{Y}$ such that (11.1) is satisfied, then pre- and postmultiplying (11.1) by $\mathbf{AA}^-$ and $\mathbf{B}^-\mathbf{B}$, respectively, yields

$$\mathbf{AA}^-\mathbf{AYBB}^-\mathbf{B} = \mathbf{AA}^-\mathbf{CB}^-\mathbf{B}\,, \tag{11.6}$$

i.e.,

$$\mathbf{AYB} = \mathbf{C} = \mathbf{AA}^-\mathbf{CB}^-\mathbf{B}\,. \tag{11.7}$$

S. Puntanen et al., *Matrix Tricks for Linear Statistical Models: Our Personal Top Twenty*,
DOI 10.1007/978-3-642-10473-2_12,

Hence the consistency of (11.1) implies (11.2). On the other hand, if (11.2) holds, then trivially $\mathbf{A}^-\mathbf{C}\mathbf{B}^-$ is one solution to (11.1) and hence (11.1) is consistent.

It remains to show that (11.3) is the general solution to (11.1). First, denote

$$\mathbf{Y}_0 = \mathbf{A}^-\mathbf{C}\mathbf{B}^- + \mathbf{Z} - \mathbf{A}^-\mathbf{A}\mathbf{Z}\mathbf{B}\mathbf{B}^- . \tag{11.8}$$

Pre- and postmultiplying $\mathbf{Y}_0$ by $\mathbf{A}$ and $\mathbf{B}$, respectively, we see at once that $\mathbf{Y}_0$ is one solution to (11.1). On the other hand, let $\mathbf{Y}_*$ be an arbitrary solution to (11.1). Then $\mathbf{Y}_*$ can be written in the form (11.3) by a suitable choice of $\mathbf{Z}$; for example, $\mathbf{Z} = \mathbf{Y}_*$ is such a choice:

$$\mathbf{Y}_* = \mathbf{A}^-\mathbf{C}\mathbf{B}^- + \mathbf{Y}_* - \mathbf{A}^-\mathbf{A}\mathbf{Y}_*\mathbf{B}\mathbf{B}^- . \tag{11.9}$$

Thus we we have shown that (11.3) provides the general solution.

For the above proof, see Rao & Mitra (1971b, p. 24), and Ben-Israel & Greville (2003, p. 52). □

Notice that according to Exercise 4.10 (p. 120) the general solutions to homogeneous equations $\mathbf{AY} = \mathbf{0}$ and $\mathbf{XA} = \mathbf{0}$ are, respectively,

$$\mathbf{Y} = (\mathbf{I} - \mathbf{A}^-\mathbf{A})\mathbf{L} = (\mathbf{A}')^{\perp}\mathbf{L} , \tag{11.10a}$$

$$\mathbf{X} = \mathbf{K}(\mathbf{I} - \mathbf{A}\mathbf{A}^-) = \mathbf{K}(\mathbf{A}^{\perp})' , \tag{11.10b}$$

where $\mathbf{L}$ and $\mathbf{K}$ are free to vary, i.e.,

$$\mathbf{Y} = \mathbf{RL} , \quad \mathbf{X} = \mathbf{KS} , \quad \text{where } \mathscr{C}(\mathbf{R}) = \mathscr{N}(\mathbf{A}) , \, \mathscr{N}(\mathbf{S}) = \mathscr{C}(\mathbf{A}) , \tag{11.11}$$

and so $\mathscr{C}(\mathbf{S}') = \mathscr{C}(\mathbf{A})^{\perp} = \mathscr{N}(\mathbf{A}')$.

It is left as an exercise to confirm that

$$\operatorname{vec}(\mathbf{AYB}) = (\mathbf{B}' \otimes \mathbf{A}) \operatorname{vec}(\mathbf{Y}) , \tag{11.12}$$

and hence the equation $\mathbf{AYB} = \mathbf{C}$ can be rewritten in the usual linear equation form as

$$(\mathbf{B}' \otimes \mathbf{A}) \operatorname{vec}(\mathbf{Y}) = \operatorname{vec}(\mathbf{C}) , \tag{11.13}$$

where $\otimes$ refers to the Kronecker product and vec to the vec-operation.

Notice also that the class of all solutions to a consistent equation $\mathbf{AY} = \mathbf{C}$ is $\{\mathbf{A}^-\mathbf{C}\}$, where $\mathbf{A}^-$ varies through all generalized inverses of $\mathbf{A}$; cf. Proposition 4.3 (p. 114).

Let us take a quick look at the general expression for $\mathbf{Y}'$ where $\mathbf{Y}$ satisfies $\mathbf{AY} = \mathbf{C}$. Because the general solution to a consistent equation $\mathbf{AY} = \mathbf{C}$ is $\mathbf{Y} = \mathbf{A}^-\mathbf{C} + (\mathbf{I} - \mathbf{A}^-\mathbf{A})\mathbf{Z}$, where $\mathbf{Z}$ is free to vary, we have trivially

$$\mathbf{Y}' = \mathbf{C}'(\mathbf{A}^-)' + \mathbf{Z}'(\mathbf{I} - \mathbf{A}^-\mathbf{A})' . \tag{11.14}$$

Choosing $\mathbf{A}^-$ as $[(\mathbf{A}')^-]'$ (please confirm that this is possible) shows that one general expression for $\mathbf{Y}'$ is

$$\mathbf{Y}' = \mathbf{C}'(\mathbf{A}')^- + \mathbf{T}[\mathbf{I} - \mathbf{A}'(\mathbf{A}')^-]\,, \tag{11.15}$$

where $\mathbf{T}$ is free to vary. Of course, (11.15) could be obtained directly by applying (11.5) into the equation $\mathbf{Y}'\mathbf{A}' = \mathbf{C}'$. In particular, for a symmetric $\mathbf{A}$,

$$\mathbf{A}' = \mathbf{A} \implies \mathbf{Y}' = \mathbf{C}'\mathbf{A}^- + \mathbf{T}(\mathbf{I} - \mathbf{A}\mathbf{A}^-), \quad \text{where } \mathbf{T} \text{ is free to vary}\,. \tag{11.16}$$

One further comment is worth making (repeating, actually): the general solution $\mathbf{Y}_0$ to the consistent equation $\mathbf{YB} = \mathbf{C}$ can of course be expressed as

$$\begin{aligned} \mathbf{Y}_0 = {} & \text{a particular solution to } \mathbf{YB} = \mathbf{C} \\ & + \text{the general solution to } \mathbf{YB} = \mathbf{0}\,. \end{aligned} \tag{11.17}$$

11.1 Equality of the BLUEs of $\mathbf{X}\boldsymbol{\beta}$ under Two Models

Consider now two linear models

$$\mathscr{M}_1 = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}_1\}\,, \quad \mathscr{M}_2 = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}_2\}\,, \tag{11.18}$$

which differ only in their covariance matrices. Then $\mathbf{Gy}$ is the BLUE for $\mathbf{X}\boldsymbol{\beta}$ under $\mathscr{M}_1$ if and only if $\mathbf{G}$ satisfies the equation

$$\mathbf{G}(\mathbf{X} : \mathbf{V}_1\mathbf{M}) = (\mathbf{X} : \mathbf{0})\,. \tag{11.19}$$

Similarly, $\mathbf{Gy}$ is the BLUE for $\mathbf{X}\boldsymbol{\beta}$ under $\mathscr{M}_2$ whenever

$$\mathbf{G}(\mathbf{X} : \mathbf{V}_2\mathbf{M}) = (\mathbf{X} : \mathbf{0})\,. \tag{11.20}$$

The set of matrices $\mathbf{G}$ satisfying $\mathbf{G}(\mathbf{X} : \mathbf{V}_i\mathbf{M}) = (\mathbf{X} : \mathbf{0})$ may be denoted as $\{\mathbf{P}_{\mathbf{X}|\mathbf{V}_i\mathbf{M}}\}$. Now according to Mitra & Moore (1973, p. 139), the comparison of the BLUEs for $\mathbf{X}\boldsymbol{\beta}$ under the models $\mathscr{M}_1$ and $\mathscr{M}_2$ can be divided into three questions:

(a) Problem [MM-1]: When is a specific linear representation of the BLUE of $\mathbf{X}\boldsymbol{\beta}$ under $\mathscr{M}_1$ also a BLUE under $\mathscr{M}_2$?

(b) Problem [MM-2]: When does $\mathbf{X}\boldsymbol{\beta}$ have a common BLUE under $\mathscr{M}_1$ and $\mathscr{M}_2$?

(c) Problem [MM-3]: When is the BLUE of $\mathbf{X}\boldsymbol{\beta}$ under $\mathscr{M}_1$ irrespective of the linear representation used in its expression, also a BLUE under $\mathscr{M}_2$?

We may cite Mitra & Moore (1973):

> "When $\mathbf{V}_1$ is singular, it is conceivable that the BLUE of an estimable linear functional $\mathbf{p}'\boldsymbol{\beta}$ may have distinct linear representations which are equal with probability 1 under $\mathscr{M}_1$, but need not be so under $\mathscr{M}_2$. This shows, in such cases, it is important to recognize the existence of three separate problems as stated above. ... When $\operatorname{rank}(\mathbf{V}_1\mathbf{X}^{\perp}) = n - \operatorname{rank}(\mathbf{X})$, the BLUE of $\mathbf{p}'\boldsymbol{\beta}$ has a unique linear presentation. Here, naturally, the three problems merge into one."

The estimator $\mathbf{Gy}$ is the BLUE for $\mathbf{X}\boldsymbol{\beta}$ under the both models if and only if

$$\mathbf{G}(\mathbf{X} : \mathbf{V}_1\mathbf{M} : \mathbf{V}_2\mathbf{M}) = (\mathbf{X} : \mathbf{0} : \mathbf{0})\,. \tag{11.21}$$

In view of Proposition 5.8 (p. 139), the above equation has a solution for $\mathbf{G}$ if and only if

$$\mathscr{C}(\mathbf{V}_1\mathbf{M} : \mathbf{V}_2\mathbf{M}) \cap \mathscr{C}(\mathbf{X}) = \{\mathbf{0}\}\,. \tag{11.22}$$

Equivalent conditions are given in Exercise 11.3 (p. 280).

Let us denote the set of all representations of BLUE for $\mathbf{X}\boldsymbol{\beta}$ under $\mathscr{M}_i$ as

$$\begin{aligned}\{\,\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta} \mid \mathscr{M}_i)\,\} &= \{\,\mathbf{Gy} : \mathbf{G}(\mathbf{X} : \mathbf{V}_i\mathbf{M}) = (\mathbf{X} : \mathbf{0})\,\} \\ &= \{\,\mathbf{Gy} : \mathbf{G} \in \{\mathbf{P}_{\mathbf{X}|\mathbf{V}_i\mathbf{M}}\}\,\}\,. \end{aligned} \tag{11.23}$$

In light of Proposition 10.5 (p. 228), the general solution to (11.19) can be written as

$$\mathbf{G}_0 = \mathbf{X}(\mathbf{X}'\mathbf{W}_1^{+}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}_1^{+} + \mathbf{F}(\mathbf{I}_n - \mathbf{P}_{(\mathbf{X}:\mathbf{V}_1\mathbf{M})})\,, \tag{11.24}$$

where $\mathbf{F}$ is free to vary and

$$\mathbf{W}_1 = \mathbf{V}_1 + \mathbf{XUX}'\,, \quad \mathscr{C}(\mathbf{W}_1) = \mathscr{C}(\mathbf{X} : \mathbf{V}_1)\,. \tag{11.25}$$

If $\mathbf{G}_0\mathbf{y}$ continues to be BLUE under $\mathscr{M}_2$, then $\mathbf{G}_0$ must satisfy also (11.20):

$$\mathbf{G}_0(\mathbf{X} : \mathbf{V}_2\mathbf{M}) = (\mathbf{X} : \mathbf{0})\,. \tag{11.26}$$

We may shorten the notation as

$$\mathbf{P}_{\mathbf{X};\mathbf{W}_1^{+}} := \mathbf{X}(\mathbf{X}'\mathbf{W}_1^{+}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}_1^{+}\,, \tag{11.27}$$

and then the $\mathbf{X}$-part of (11.26) becomes

$$\mathbf{P}_{\mathbf{X};\mathbf{W}_1^{+}}\mathbf{X} + \mathbf{F}(\mathbf{I}_n - \mathbf{P}_{(\mathbf{X}:\mathbf{V}_1\mathbf{M})})\mathbf{X} = \mathbf{X}\,, \tag{11.28}$$

which is trivially true because $\mathbf{P}_{\mathbf{X};\mathbf{W}_1^{+}} \in \{\mathbf{P}_{\mathbf{X}|\mathbf{V}_1\mathbf{M}}\}$. The $\mathbf{V}_2\mathbf{M}$-part of (11.26) is

$$\mathbf{P}_{\mathbf{X};\mathbf{W}_1^{+}}\mathbf{V}_2\mathbf{M} + \mathbf{F}(\mathbf{I}_n - \mathbf{P}_{(\mathbf{X}:\mathbf{V}_1\mathbf{M})})\mathbf{V}_2\mathbf{M} = \mathbf{0}\,. \tag{11.29}$$

Because (11.29) must hold for *every* choice of $\mathbf{F}$, we necessarily must have $(\mathbf{I}_n - \mathbf{P}_{(\mathbf{X}:\mathbf{V}_1\mathbf{M})})\mathbf{V}_2\mathbf{M} = \mathbf{0}$, i.e., $\mathscr{C}(\mathbf{V}_2\mathbf{M}) \subset \mathscr{C}(\mathbf{X} : \mathbf{V}_1\mathbf{M})$; in other words,

$$\mathbf{V}_2\mathbf{M} = \mathbf{X}\mathbf{N}_1 + \mathbf{V}_1\mathbf{M}\mathbf{N}_2\,, \quad \text{for some } \mathbf{N}_1,\, \mathbf{N}_2\,. \tag{11.30}$$

Substituting (11.30) into (11.29) yields

$$\mathbf{P}_{\mathbf{X};\mathbf{W}_1^+}(\mathbf{X}\mathbf{N}_1 + \mathbf{V}_1\mathbf{M}\mathbf{N}_2) = \mathbf{0}\,. \tag{11.31}$$

Now

$$\mathbf{P}_{\mathbf{X};\mathbf{W}_1^+}\mathbf{X}\mathbf{N}_1 = \mathbf{X}\mathbf{N}_1\,, \quad \mathbf{P}_{\mathbf{X};\mathbf{W}_1^+}\mathbf{V}_1\mathbf{M}\mathbf{N}_2 = \mathbf{0}\,, \tag{11.32}$$

and thereby $\mathbf{X}\mathbf{N}_1 = \mathbf{0}$. Substituting this into (11.30) yields

$$\mathscr{C}(\mathbf{V}_2\mathbf{M}) \subset \mathscr{C}(\mathbf{V}_1\mathbf{M})\,. \tag{11.33}$$

To go the other way, assume that $\mathbf{V}_2\mathbf{M} = \mathbf{V}_1\mathbf{M}\mathbf{N}_2$ for some $\mathbf{N}_2$. Then postmultiplying

$$\mathbf{G}(\mathbf{X} : \mathbf{V}_1\mathbf{M}) = (\mathbf{X} : \mathbf{0}) \tag{11.34}$$

by $\left(\begin{smallmatrix} \mathbf{I}_p & \mathbf{0} \\ \mathbf{0} & \mathbf{N}_2 \end{smallmatrix}\right)$ yields

$$\mathbf{G}(\mathbf{X} : \mathbf{V}_2\mathbf{M}) = (\mathbf{X} : \mathbf{0})\,. \tag{11.35}$$

Hence we have proved the part "(a) $\iff$ (c)" of the following proposition. We may note that the equivalence of (a) and (c) follows also at once from the result of Exercise 2.6 (p. 89), due to Kala (1981, Lemma 2.5):

$$\{\,\mathbf{P}_{\mathbf{C}|\mathbf{D}}\,\} \subset \{\,\mathbf{P}_{\mathbf{A}|\mathbf{B}}\,\} \iff \mathscr{C}(\mathbf{A}) \subset \mathscr{C}(\mathbf{C}) \;\text{ and }\; \mathscr{C}(\mathbf{B}) \subset \mathscr{C}(\mathbf{D})\,. \tag{11.36}$$

Proposition 11.1. *Consider the linear models* $\mathscr{M}_1 = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \mathbf{V}_1\}$ *and* $\mathscr{M}_2 = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \mathbf{V}_2\}$*, and let the notation*

$$\{\,\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta} \mid \mathscr{M}_1)\,\} \subset \{\,\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta} \mid \mathscr{M}_2)\,\} \tag{11.37}$$

mean that every representation of the BLUE *for* $\mathbf{X}\boldsymbol{\beta}$ *under* $\mathscr{M}_1$ *remains the* BLUE *for* $\mathbf{X}\boldsymbol{\beta}$ *under* $\mathscr{M}_2$*. Then the following statements are equivalent:*

(a) $\{\,\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta} \mid \mathscr{M}_1)\,\} \subset \{\,\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta} \mid \mathscr{M}_2)\,\}$,

(b) $\{\,\mathrm{BLUE}(\mathbf{K}'\boldsymbol{\beta} \mid \mathscr{M}_1)\,\} \subset \{\,\mathrm{BLUE}(\mathbf{K}'\boldsymbol{\beta}) \mid \mathscr{M}_2)\,\}$ *for every estimable* $\mathbf{K}'\boldsymbol{\beta}$,

(c) $\mathscr{C}(\mathbf{V}_2\mathbf{X}^{\perp}) \subset \mathscr{C}(\mathbf{V}_1\mathbf{X}^{\perp})$,

(d) $\mathbf{V}_2 = \mathbf{V}_1 + \mathbf{X}\mathbf{N}_1\mathbf{X}' + \mathbf{V}_1\mathbf{M}\mathbf{N}_2\mathbf{M}\mathbf{V}_1$, *for some* $\mathbf{N}_1$ *and* $\mathbf{N}_2$,

(e) $\mathbf{V}_2 = \mathbf{X}\mathbf{N}_3\mathbf{X}' + \mathbf{V}_1\mathbf{M}\mathbf{N}_4\mathbf{M}\mathbf{V}_1$, *for some* $\mathbf{N}_3$ *and* $\mathbf{N}_4$.

For the proof of the above Proposition and related discussion, see, e.g., Baksalary & Mathew (1986, Th. 3), Mitra & Moore (1973, Th. 4.1–4.2), and Rao (1968, Lemma 5), Rao (1971c, Th. 5.2, Th. 5.5), Rao (1973b, p. 289), and Tian (2009b). The optimality of the BLUEs in a linear model with incorrect design matrix is considered in Mathew & Bhimasankaram (1983c).

Notice that obviously the following statements are equivalent:

$$\{\,\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta} \mid \mathscr{M}_1)\,\} = \{\,\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta} \mid \mathscr{M}_2)\,\}\,, \tag{11.38a}$$

$$\mathscr{C}(\mathbf{V}_2\mathbf{X}^{\perp}) = \mathscr{C}(\mathbf{V}_1\mathbf{X}^{\perp})\,. \tag{11.38b}$$

For the situation of positive definite $\mathbf{V}_1$ and $\mathbf{V}_2$, see Exercise 2.3 (p. 88).

For the equality of the BLUEs of the subvector $\mathbf{X}_1\boldsymbol{\beta}_1$ under two linear models, see Section 15.6 (p. 333).

11.2 Model $\mathscr{M}_I$ vs Model $\mathscr{M}_V$

One may wonder what happens if we replace in (11.38) the model $\mathscr{M}_1$ with the model $\mathscr{M}_I = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}\}$ and $\mathscr{M}_2$ with the model $\mathscr{M}_V = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{V}\}$, where $\mathbf{X}$ has full column rank. Then we can ask the following question:

> When does every BLUE of $\boldsymbol{\beta}$ under $\mathscr{M}_I$ continue to be the BLUE of $\boldsymbol{\beta}$ also under $\mathscr{M}_V$ and vice versa? (11.39)

The answer follows from the equivalence of the following two statements:

$$\{\,\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta} \mid \mathscr{M}_I)\,\} = \{\,\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta} \mid \mathscr{M}_V)\,\}\,, \tag{11.40a}$$

$$\mathscr{C}(\mathbf{VM}) = \mathscr{C}(\mathbf{M})\,. \tag{11.40b}$$

However, the condition $\mathscr{C}(\mathbf{VM}) = \mathscr{C}(\mathbf{M})$ is *not* exactly the same as the classic condition for OLSE($\boldsymbol{\beta}$) to be equal to the BLUE of $\boldsymbol{\beta}$ which is $\mathscr{C}(\mathbf{VM}) \subset \mathscr{C}(\mathbf{M})$.

Markiewicz, Puntanen & Styan (2010) discuss the above problem concluding that the classic condition $\mathscr{C}(\mathbf{VM}) \subset \mathscr{C}(\mathbf{M})$ is the answer to the following question: What is a necessary and sufficient condition that OLSE($\boldsymbol{\beta}$), i.e., $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, is the BLUE for $\boldsymbol{\beta}$ under the model $\mathscr{M}_V = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{V}\}$?

Moreover, the “somewhat unexpected” condition $\mathscr{C}(\mathbf{M}) \subset \mathscr{C}(\mathbf{VM})$ is the answer to the following question: What is a necessary and sufficient condition that every BLUE for $\boldsymbol{\beta}$ under $\mathscr{M}_V = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{V}\}$ is also the BLUE for $\boldsymbol{\beta}$ under $\mathscr{M}_I = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}\}$?

The latter question is rather unusual, and it is the former one that most statisticians have in mind when they ask for the conditions for the equality $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}$ to hold.

11.3 Stochastic Restrictions and the BLUPs

In Section 10.11 (p. 253) we considered a mixed linear model

$$\mathscr{L} = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\gamma,\, \mathbf{D},\, \mathbf{R}\}\,, \tag{11.41}$$

where $\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \varepsilon$, $\mathbf{X} \in \mathbb{R}^{n\times p}$, $\mathbf{Z} \in \mathbb{R}^{n\times q}$,

$$\mathrm{E}(\gamma) = \mathbf{0}\,, \quad \mathrm{cov}(\gamma) = \mathbf{D}\,, \quad \mathrm{cov}(\gamma, \varepsilon) = \mathbf{0}\,, \tag{11.42}$$

and $\mathrm{E}(\varepsilon) = \mathbf{0}$, $\mathrm{cov}(\varepsilon) = \mathbf{R}$, $\mathrm{cov}(\mathbf{y}) = \mathbf{ZDZ}' + \mathbf{R} := \mathbf{\Sigma}$.

On page 255 we mentioned that Henderson (1950, 1963) obtained the BLUE for $\boldsymbol{\beta}$ and BLUP for $\boldsymbol{\gamma}$ by minimizing the following quadratic form with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ (keeping $\boldsymbol{\gamma}$ as a non-random vector):

$$f(\boldsymbol{\beta}, \gamma) = (\mathbf{y}_\# - \mathbf{X}_*\boldsymbol{\pi})'\mathbf{V}_*^{-1}(\mathbf{y}_\# - \mathbf{X}_*\boldsymbol{\pi})\,, \tag{11.43}$$

where

$$\mathbf{y}_\# = \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix}, \quad \mathbf{X}_* = \begin{pmatrix} \mathbf{X} & \mathbf{Z} \\ \mathbf{0} & \mathbf{I}_q \end{pmatrix}, \quad \mathbf{V}_* = \begin{pmatrix} \mathbf{R} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{pmatrix}, \quad \boldsymbol{\pi} = \begin{pmatrix} \beta \\ \gamma \end{pmatrix}. \tag{11.44}$$

Photograph 11.1 David A. Harville (Fort Lauderdale, 1998).

The purpose of this section is to give a generalized form for Henderson's result, without any rank assumptions, along the lines of Haslett & Puntanen (2010b). Our considerations are based on the use of stochastic restrictions; see, e.g., Rao, Toutenburg, Shalabh & Heumann (2008, §5.10). Using different approach, the results of Henderson have been generalized, for example, by Harville (1976, 1979).

We recall that $\mathbf{Ay}$ is the BLUP for $\boldsymbol{\gamma}$ under the mixed model $\mathscr{L}$ if and only if

$$\mathbf{A}(\mathbf{X} : \mathbf{\Sigma M}) = (\mathbf{0} : \mathbf{DZ}'\mathbf{M})\,, \tag{11.45}$$

and $\mathbf{Gy}$ is the BLUE for $\mathbf{X}\beta$ under $\mathscr{L}$ if and only if

$$\mathbf{G}(\mathbf{X} : \mathbf{\Sigma M}) = (\mathbf{X} : \mathbf{0})\,, \tag{11.46}$$

where $\mathbf{M} = \mathbf{I} - \mathbf{H}$.

Let us consider the *fixed effects* partitioned model

$$\mathscr{F} : \mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \varepsilon\,, \quad \mathrm{cov}(\mathbf{y}) = \mathrm{cov}(\varepsilon) = \mathbf{R}\,, \tag{11.47}$$

where both $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are fixed (but of course unknown) coefficients, and supplement $\mathscr{F}$ with the *stochastic restrictions*

$$\mathbf{y}_0 = \boldsymbol{\gamma} + \boldsymbol{\varepsilon}_0\,, \quad \mathrm{cov}(\boldsymbol{\varepsilon}_0) = \mathbf{D}\,. \tag{11.48}$$

This supplement can be expressed as the partitioned model:

$$\begin{aligned}\mathscr{F}_* &= \{\mathbf{y}_*,\, \mathbf{X}_*\boldsymbol{\pi},\, \mathbf{V}_*\}\\ &= \left\{\begin{pmatrix}\mathbf{y}\\ \mathbf{y}_0\end{pmatrix},\, \begin{pmatrix}\mathbf{X} & \mathbf{Z}\\ \mathbf{0} & \mathbf{I}_q\end{pmatrix}\begin{pmatrix}\boldsymbol{\beta}\\ \boldsymbol{\gamma}\end{pmatrix},\, \begin{pmatrix}\mathbf{R} & \mathbf{0}\\ \mathbf{0} & \mathbf{D}\end{pmatrix}\right\}. \end{aligned} \tag{11.49}$$

Here $\mathscr{F}_*$ is just an ordinary partitioned model with fixed effects. It is a partitioned linear model $\mathscr{F} = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma},\, \mathbf{R}\}$, supplemented with stochastic restrictions on $\boldsymbol{\gamma}$.

Above we have

$$\mathbf{y}_* = \begin{pmatrix}\mathbf{y}\\ \mathbf{y}_0\end{pmatrix}, \quad \mathbf{X}_* = \begin{pmatrix}\mathbf{X} & \mathbf{Z}\\ \mathbf{0} & \mathbf{I}_q\end{pmatrix} = (\mathbf{X}_{*1} : \mathbf{X}_{*2})\,. \tag{11.50}$$

Notice that $\mathscr{C}(\mathbf{X}_{*1}) \cap \mathscr{C}(\mathbf{X}_{*2}) = \{\mathbf{0}\}$, which means that

$$\mathbf{X}_{*1}\boldsymbol{\beta} = \begin{pmatrix}\mathbf{X}\boldsymbol{\beta}\\ \mathbf{0}\end{pmatrix} \quad \text{and} \quad \mathbf{X}_{*2}\boldsymbol{\gamma} = \begin{pmatrix}\mathbf{Z}\boldsymbol{\gamma}\\ \boldsymbol{\gamma}\end{pmatrix} \tag{11.51}$$

are estimable under $\mathscr{F}_*$, and in particular, $\boldsymbol{\gamma}$ itself is estimable under $\mathscr{F}_*$. We need later the matrix $\mathbf{X}_*^{\perp}$ for which one choice is (please confirm)

$$\mathbf{X}_*^{\perp} = \begin{pmatrix}\mathbf{I}_n\\ -\mathbf{Z}'\end{pmatrix}\mathbf{M}\,, \tag{11.52}$$

and so we have

$$\mathbf{V}_*\mathbf{X}_*^{\perp} = \begin{pmatrix}\mathbf{R} & \mathbf{0}\\ \mathbf{0} & \mathbf{D}\end{pmatrix}\begin{pmatrix}\mathbf{I}_n\\ -\mathbf{Z}'\end{pmatrix}\mathbf{M} = \begin{pmatrix}\mathbf{RM}\\ -\mathbf{DZ}'\mathbf{M}\end{pmatrix}. \tag{11.53}$$

We can now prove the following result.

Proposition 11.2. *With the above notation, the following statements are equivalent:*

(a) *The estimator* $\mathbf{By}_*$ *is the* BLUE *for* $\mathbf{X}_*\boldsymbol{\pi}$ *under the augmented model* $\mathscr{F}_*$, *i.e.,*

$$\begin{aligned}\mathbf{By}_* = \begin{pmatrix}\mathbf{B}_{11} & \mathbf{B}_{12}\\ \mathbf{B}_{21} & \mathbf{B}_{22}\end{pmatrix}\mathbf{y}_* &= \begin{pmatrix}\mathbf{B}_{11}\\ \mathbf{B}_{21}\end{pmatrix}\mathbf{y} + \begin{pmatrix}\mathbf{B}_{12}\\ \mathbf{B}_{22}\end{pmatrix}\mathbf{y}_0\\ &= \mathrm{BLUE}(\mathbf{X}_*\boldsymbol{\pi} \mid \mathscr{F}_*)\,,\end{aligned} \tag{11.54}$$

where $\mathbf{B}_{12} = \mathbf{Z} - \mathbf{B}_{11}\mathbf{Z}$ *and* $\mathbf{B}_{22} = \mathbf{I}_q - \mathbf{B}_{21}\mathbf{Z}$.

(b) *The predictor* $\mathbf{B}_{21}\mathbf{y}$ *is the* BLUP *for* $\boldsymbol{\gamma}$ *and* $(\mathbf{B}_{11} - \mathbf{ZB}_{21})\mathbf{y}$ *is the* BLUE *of* $\mathbf{X}\boldsymbol{\beta}$ *under the mixed model* $\mathscr{L}$; *in other words,*

$$\mathbf{KB}\begin{pmatrix}\mathbf{y}\\ \mathbf{0}\end{pmatrix} = \begin{pmatrix}\mathbf{B}_{11} - \mathbf{ZB}_{21}\\ \mathbf{B}_{21}\end{pmatrix}\mathbf{y} = \begin{pmatrix}\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta} \mid \mathscr{L})\\ \mathrm{BLUP}(\boldsymbol{\gamma} \mid \mathscr{L})\end{pmatrix}, \tag{11.55}$$

where $\mathbf{K} = \left(\begin{smallmatrix}\mathbf{I}_n & -\mathbf{Z}\\ \mathbf{0} & \mathbf{I}_q\end{smallmatrix}\right)$, *or equivalently*

$$\mathbf{B}\begin{pmatrix}\mathbf{y}\\ \mathbf{0}\end{pmatrix} = \begin{pmatrix}\mathbf{B}_{11}\\ \mathbf{B}_{21}\end{pmatrix}\mathbf{y} = \begin{pmatrix}\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta} \mid \mathscr{L}) + \mathrm{BLUP}(\mathbf{Z}\boldsymbol{\gamma} \mid \mathscr{L})\\ \mathrm{BLUP}(\boldsymbol{\gamma} \mid \mathscr{L})\end{pmatrix}. \tag{11.56}$$

Proof. The estimator $\mathbf{B}\mathbf{y}_*$ is the BLUE for $\mathbf{X}_*\boldsymbol{\pi}$ under the model $\mathscr{F}_*$ if and only if $\mathbf{B}$ satisfies the equation

$$\mathbf{B}(\mathbf{X}_* : \mathbf{V}_*\mathbf{X}_*^{\perp}) = (\mathbf{X}_* : \mathbf{0}), \tag{11.57}$$

i.e.,

$$\begin{pmatrix}\mathbf{B}_{11} & \mathbf{B}_{12}\\ \mathbf{B}_{21} & \mathbf{B}_{22}\end{pmatrix}\begin{pmatrix}\mathbf{X} & \mathbf{Z} & \mathbf{R}\mathbf{M}\\ \mathbf{0} & \mathbf{I}_q & -\mathbf{D}\mathbf{Z}'\mathbf{M}\end{pmatrix} = \begin{pmatrix}\mathbf{X} & \mathbf{Z} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_q & \mathbf{0}\end{pmatrix}. \tag{11.58}$$

From (11.58) we see immediately that

$$\mathbf{B}_{12} = \mathbf{Z} - \mathbf{B}_{11}\mathbf{Z}, \quad \mathbf{B}_{22} = \mathbf{I}_q - \mathbf{B}_{21}\mathbf{Z}, \tag{11.59}$$

which yields an important conclusion: once the blocks $\mathbf{B}_{11}$ and $\mathbf{B}_{21}$ are given, the other two blocks of $\mathbf{B}$ become fixed as in (11.59). In view of (11.59), (11.58) can be equivalently expressed as

$$\begin{pmatrix}\mathbf{B}_{11} & \mathbf{Z} - \mathbf{B}_{11}\mathbf{Z}\\ \mathbf{B}_{21} & \mathbf{I}_q - \mathbf{B}_{21}\mathbf{Z}\end{pmatrix}\begin{pmatrix}\mathbf{X} & \mathbf{R}\mathbf{M}\\ \mathbf{0} & -\mathbf{D}\mathbf{Z}'\mathbf{M}\end{pmatrix} = \begin{pmatrix}\mathbf{X} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}. \tag{11.60}$$

Moreover, premultiplying (11.60) by the nonsingular matrix $\mathbf{K} = \left(\begin{smallmatrix}\mathbf{I}_n & -\mathbf{Z}\\ \mathbf{0} & \mathbf{I}_q\end{smallmatrix}\right)$, yields

$$\begin{pmatrix}\mathbf{B}_{11} - \mathbf{Z}\mathbf{B}_{21} & -(\mathbf{B}_{11} - \mathbf{Z}\mathbf{B}_{21})\mathbf{Z}\\ \mathbf{B}_{21} & \mathbf{I}_q - \mathbf{B}_{21}\mathbf{Z}\end{pmatrix}\begin{pmatrix}\mathbf{X} & \mathbf{R}\mathbf{M}\\ \mathbf{0} & -\mathbf{D}\mathbf{Z}'\mathbf{M}\end{pmatrix} = \begin{pmatrix}\mathbf{X} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}, \tag{11.61}$$

which can be equivalently written as

$$\begin{pmatrix}\mathbf{B}_{11} - \mathbf{Z}\mathbf{B}_{21}\\ \mathbf{B}_{21}\end{pmatrix}(\mathbf{X} : \boldsymbol{\Sigma}\mathbf{M}) = \begin{pmatrix}\mathbf{X} & \mathbf{0}\\ \mathbf{0} & \mathbf{D}\mathbf{Z}'\mathbf{M}\end{pmatrix}. \tag{11.62}$$

Now, in light of (11.45) and (11.46) (p. 273), equation (11.62) is a necessary and sufficient condition for

$$\begin{pmatrix}\mathbf{B}_{11} - \mathbf{Z}\mathbf{B}_{21}\\ \mathbf{B}_{21}\end{pmatrix}\mathbf{y} = \begin{pmatrix}\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta} \mid \mathscr{L})\\ \mathrm{BLUP}(\boldsymbol{\gamma} \mid \mathscr{L})\end{pmatrix}. \tag{11.63}$$

□

The following proposition shows that actually *all* representations of the BLUEs and BLUPs in the mixed model can be generated through the matrix $\mathbf{B}$; see Haslett & Puntanen (2010b).

Proposition 11.3. *Assume that* $\mathbf{B} = \left(\begin{smallmatrix}\mathbf{B}_{11} & \mathbf{B}_{12}\\ \mathbf{B}_{21} & \mathbf{B}_{22}\end{smallmatrix}\right)$ *satisfies the equation*

$$\mathbf{B}(\mathbf{X}_* : \mathbf{V}_*\mathbf{X}_*^{\perp}) = (\mathbf{X}_* : \mathbf{0})\,. \tag{11.64}$$

Then all representations of the BLUE*s for* $\mathbf{X}\boldsymbol{\beta}$ *and the* BLUP*s for* $\boldsymbol{\gamma}$ *in the mixed model* $\mathscr{L}$ *can be generated through*

$$\begin{pmatrix}\mathbf{B}_{11} - \mathbf{Z}\mathbf{B}_{21}\\ \mathbf{B}_{21}\end{pmatrix}\mathbf{y} \tag{11.65}$$

by varying $\mathbf{B}_{11}$ *and* $\mathbf{B}_{21}$.

Proof. Proposition 11.2 tells that for the given $\mathbf{B}$, the upper part of

$$\begin{pmatrix}\mathbf{B}_{11} - \mathbf{Z}\mathbf{B}_{21}\\ \mathbf{B}_{21}\end{pmatrix}\mathbf{y} \tag{11.66}$$

is *one* representation for $\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta} \mid \mathscr{L})$ and the lower part is *one* representation for $\mathrm{BLUP}(\boldsymbol{\gamma} \mid \mathscr{L})$. Our claim is that if $\mathbf{Ay}$ is an arbitrary given representation for $\mathrm{BLUP}(\boldsymbol{\gamma} \mid \mathscr{L})$ and $\mathbf{Gy}$ is an arbitrary given representation for $\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta} \mid \mathscr{L})$, then there exists $\mathbf{B}_{11}$ satisfying (11.64) such that the matrix $\mathbf{G}$ can be represented in the form

$$\mathbf{G} = \mathbf{B}_{11} - \mathbf{Z}\mathbf{A}\,. \tag{11.67}$$

According to Theorem 11 (p. 267), the general representations for $\mathbf{G}$, $\mathbf{A}$, and $\mathbf{B}_{11}$ are the following:

$$\mathbf{G} = (\mathbf{X} : \mathbf{0})(\mathbf{X} : \boldsymbol{\Sigma}\mathbf{M})^{+} + \mathbf{L}_1(\mathbf{I}_n - \mathbf{P}_{(\mathbf{X}\,:\,\boldsymbol{\Sigma})})\,, \tag{11.68a}$$

$$\mathbf{A} = (\mathbf{0} : \mathbf{D}\mathbf{Z}'\mathbf{M})(\mathbf{X} : \boldsymbol{\Sigma}\mathbf{M})^{+} + \mathbf{L}_2(\mathbf{I}_n - \mathbf{P}_{(\mathbf{X}\,:\,\boldsymbol{\Sigma})})\,, \tag{11.68b}$$

$$\mathbf{B}_{11} = (\mathbf{X} : \mathbf{Z}\mathbf{D}\mathbf{Z}'\mathbf{M})(\mathbf{X} : \boldsymbol{\Sigma}\mathbf{M})^{+} + \mathbf{L}_3(\mathbf{I}_n - \mathbf{P}_{(\mathbf{X}\,:\,\boldsymbol{\Sigma})})\,, \tag{11.68c}$$

where $\mathbf{L}_1$, $\mathbf{L}_2$, and $\mathbf{L}_3$ are free to vary. In our particular situation, $\mathbf{L}_1$ and $\mathbf{L}_2$ are given. Choosing $\mathbf{L}_3 = \mathbf{L}_1 + \mathbf{Z}\mathbf{L}_2$ yields immediately (11.67) and thus the proof is completed. □

11.4 Two Mixed Models

Consider two mixed models

$$\mathscr{L}_i = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma},\, \mathbf{D}_i,\, \mathbf{R}_i\}\,, \quad i = 1, 2\,, \tag{11.69}$$

and their augmented versions

$$\begin{aligned}\mathscr{F}_{*i} &= \{\mathbf{y}_*,\, \mathbf{X}_*\boldsymbol{\pi},\, \mathbf{V}_{*i}\} \\ &= \left\{ \begin{pmatrix}\mathbf{y}\\ \mathbf{y}_0\end{pmatrix}, \begin{pmatrix}\mathbf{X} & \mathbf{Z}\\ \mathbf{0} & \mathbf{I}_q\end{pmatrix}\begin{pmatrix}\boldsymbol{\beta}\\ \boldsymbol{\gamma}\end{pmatrix}, \begin{pmatrix}\mathbf{R}_i & \mathbf{0}\\ \mathbf{0} & \mathbf{D}_i\end{pmatrix} \right\}, \quad i = 1, 2\,. \end{aligned} \tag{11.70}$$

Then, every representation of the BLUE for $\mathbf{X}_*\boldsymbol{\pi}$ under $\mathscr{F}_{*1}$ remains the BLUE for $\mathbf{X}_*\boldsymbol{\pi}$ under $\mathscr{F}_{*2}$ if and only if the following implication holds:

$$\mathbf{B}(\mathbf{X}_* : \mathbf{V}_{*1}\mathbf{X}_*^{\perp}) = (\mathbf{X}_* : \mathbf{0}) \implies \mathbf{B}(\mathbf{X}_* : \mathbf{V}_{*2}\mathbf{X}_*^{\perp}) = (\mathbf{X}_* : \mathbf{0})\,. \tag{11.71}$$

We may denote this relation shortly

$$\{\,\mathrm{BLUE}(\mathbf{X}_*\boldsymbol{\pi} \mid \mathscr{F}_{*1})\,\} \subset \{\,\mathrm{BLUE}(\mathbf{X}_*\boldsymbol{\pi} \mid \mathscr{F}_{*2})\,\}\,. \tag{11.72}$$

In view of Proposition 11.2 (p. 274), (11.71) is equivalent to

$$\begin{pmatrix}\mathbf{B}_{11} - \mathbf{Z}\mathbf{B}_{21}\\ \mathbf{B}_{21}\end{pmatrix}(\mathbf{X} : \boldsymbol{\Sigma}_1\mathbf{M}) = \begin{pmatrix}\mathbf{X} & \mathbf{0}\\ \mathbf{0} & \mathbf{D}_1\mathbf{Z}'\mathbf{M}\end{pmatrix} \tag{11.73a}$$

$$\implies$$

$$\begin{pmatrix}\mathbf{B}_{11} - \mathbf{Z}\mathbf{B}_{21}\\ \mathbf{B}_{21}\end{pmatrix}(\mathbf{X} : \boldsymbol{\Sigma}_2\mathbf{M}) = \begin{pmatrix}\mathbf{X} & \mathbf{0}\\ \mathbf{0} & \mathbf{D}_2\mathbf{Z}'\mathbf{M}\end{pmatrix}. \tag{11.73b}$$

On account of Proposition 11.3 (p. 276), the above implication is equivalent to that both

$$\{\,\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta} \mid \mathscr{L}_1)\,\} \subset \{\,\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta} \mid \mathscr{L}_2)\,\} \tag{11.74a}$$

and

$$\{\,\mathrm{BLUP}(\boldsymbol{\gamma} \mid \mathscr{L}_1)\,\} \subset \{\,\mathrm{BLUP}(\boldsymbol{\gamma} \mid \mathscr{L}_2)\,\} \tag{11.74b}$$

hold. On the other hand, according to Proposition 11.1 (p. 271), every representation of the BLUE for $\mathbf{X}_*\boldsymbol{\pi}$ under $\mathscr{F}_{*1}$ remains the BLUE for $\mathbf{X}_*\boldsymbol{\pi}$ under $\mathscr{F}_{*2}$ if and only if

$$\mathscr{C}(\mathbf{V}_{*2}\mathbf{X}_*^{\perp}) \subset \mathscr{C}(\mathbf{V}_{*1}\mathbf{X}_*^{\perp})\,, \tag{11.75}$$

i.e.,

$$\mathscr{C}\begin{pmatrix}\mathbf{R}_2\mathbf{M}\\ \mathbf{D}_2\mathbf{Z}'\mathbf{M}\end{pmatrix} \subset \mathscr{C}\begin{pmatrix}\mathbf{R}_1\mathbf{M}\\ \mathbf{D}_1\mathbf{Z}'\mathbf{M}\end{pmatrix}. \tag{11.76}$$

We can collect our observations in the following proposition; see Haslett & Puntanen (2010b).

Proposition 11.4. *Using the notation introduced above, the following statements are equivalent:*

(a) $\{\,\mathrm{BLUE}(\mathbf{X}_*\boldsymbol{\pi} \mid \mathscr{F}_{*1})\,\} \subset \{\,\mathrm{BLUE}(\mathbf{X}_*\boldsymbol{\pi} \mid \mathscr{F}_{*2})\,\}$,

(b) $\{\,\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta} \mid \mathscr{L}_1)\,\} \subset \{\,\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta} \mid \mathscr{L}_2)\,\}$ *and* $\{\,\mathrm{BLUP}(\boldsymbol{\gamma} \mid \mathscr{L}_1)\,\} \subset \{\,\mathrm{BLUP}(\boldsymbol{\gamma} \mid \mathscr{L}_2)\,\}$,

(c) $\mathscr{C}\begin{pmatrix}\mathbf{\Sigma}_2\mathbf{M}\\ \mathbf{D}_2\mathbf{Z}'\mathbf{M}\end{pmatrix} \subset \mathscr{C}\begin{pmatrix}\mathbf{\Sigma}_1\mathbf{M}\\ \mathbf{D}_1\mathbf{Z}'\mathbf{M}\end{pmatrix}$,

(d) $\mathscr{C}\begin{pmatrix}\mathbf{R}_2\mathbf{M}\\ \mathbf{D}_2\mathbf{Z}'\mathbf{M}\end{pmatrix} \subset \mathscr{C}\begin{pmatrix}\mathbf{R}_1\mathbf{M}\\ \mathbf{D}_1\mathbf{Z}'\mathbf{M}\end{pmatrix}$.

Haslett & Puntanen (2010d) studied also the equality of the BLUPs for $\boldsymbol{\gamma}$ under two different mixed models—omitting the equality between the BLUEs for $\mathbf{X}\boldsymbol{\beta}$. So the question is: when does every representation of the BLUP for $\boldsymbol{\gamma}$ under the mixed model $\mathscr{L}_1$ continue to be the BLUP of $\boldsymbol{\gamma}$ under the other mixed model $\mathscr{L}_2$, i.e.,

$$\{\,\mathrm{BLUP}(\boldsymbol{\gamma} \mid \mathscr{L}_1)\,\} \subset \{\,\mathrm{BLUP}(\boldsymbol{\gamma} \mid \mathscr{L}_2)\,\}\,. \tag{11.77}$$

Rong & Liu (2010) considered the consequences of misspecification of the covariance matrix in a mixed model using different approach. We may mention that Möls (2004, Th. 5.3) gave a sufficient condition for (11.77), assuming $\mathbf{R}_i$ and $\mathbf{D}_i$ to be positive definite.

Below is the result of Haslett & Puntanen (2010d).

Proposition 11.5. *Using the notation introduced above, the following statements are equivalent:*

(a) $\{\,\mathrm{BLUP}(\boldsymbol{\gamma} \mid \mathscr{L}_1)\,\} \subset \{\,\mathrm{BLUP}(\boldsymbol{\gamma} \mid \mathscr{L}_2)\,\}$,

(b) $\mathscr{C}\begin{pmatrix}\mathbf{\Sigma}_2\mathbf{M}\\ \mathbf{D}_2\mathbf{Z}'\mathbf{M}\end{pmatrix} \subset \mathscr{C}\begin{pmatrix}\mathbf{X} & \mathbf{\Sigma}_1\mathbf{M}\\ \mathbf{0} & \mathbf{D}_1\mathbf{Z}'\mathbf{M}\end{pmatrix}$,

(c) $\mathscr{C}\begin{pmatrix}\mathbf{R}_2\mathbf{M}\\ \mathbf{D}_2\mathbf{Z}'\mathbf{M}\end{pmatrix} \subset \mathscr{C}\begin{pmatrix}\mathbf{X} & \mathbf{R}_1\mathbf{M}\\ \mathbf{0} & \mathbf{D}_1\mathbf{Z}'\mathbf{M}\end{pmatrix}$,

(d) $\mathscr{C}\begin{pmatrix}\mathbf{M}\mathbf{\Sigma}_2\mathbf{M}\\ \mathbf{D}_2\mathbf{Z}'\mathbf{M}\end{pmatrix} \subset \mathscr{C}\begin{pmatrix}\mathbf{M}\mathbf{\Sigma}_1\mathbf{M}\\ \mathbf{D}_1\mathbf{Z}'\mathbf{M}\end{pmatrix}$,

(e) $\mathscr{C}\begin{pmatrix}\mathbf{M}\mathbf{R}_2\mathbf{M}\\ \mathbf{D}_2\mathbf{Z}'\mathbf{M}\end{pmatrix} \subset \mathscr{C}\begin{pmatrix}\mathbf{M}\mathbf{R}_1\mathbf{M}\\ \mathbf{D}_1\mathbf{Z}'\mathbf{M}\end{pmatrix}$.

11.5 Two Models with New Unobserved Future Observations

Here we take a look at the two linear models (with new unobserved future observations) $\mathscr{M}_f$ and $\underline{\mathscr{M}}_f$, where $\mathbf{X}_f\boldsymbol{\beta}$ is a given estimable parametric function:

$$\mathscr{M}_f = \left\{ \begin{pmatrix}\mathbf{y}\\ \mathbf{y}_f\end{pmatrix}, \begin{pmatrix}\mathbf{X}\\ \mathbf{X}_f\end{pmatrix}\boldsymbol{\beta}, \begin{pmatrix}\mathbf{V}_{11} & \mathbf{V}_{12}\\ \mathbf{V}_{21} & \mathbf{V}_{22}\end{pmatrix} \right\}, \tag{11.78a}$$

$$\underline{\mathscr{M}}_f = \left\{ \begin{pmatrix}\mathbf{y}\\ \mathbf{y}_f\end{pmatrix}, \begin{pmatrix}\underline{\mathbf{X}}\\ \underline{\mathbf{X}}_f\end{pmatrix}\boldsymbol{\beta}, \begin{pmatrix}\underline{\mathbf{V}}_{11} & \underline{\mathbf{V}}_{12}\\ \underline{\mathbf{V}}_{21} & \underline{\mathbf{V}}_{22}\end{pmatrix} \right\}. \tag{11.78b}$$

Thus the models may have different model matrices and covariance matrices.

One may now ask for a condition that every representation of the BLUP for $\mathbf{y}_f$ under the model $\mathscr{M}_f$ is also a BLUP for $\mathbf{y}_f$ under the model $\underline{\mathscr{M}}_f$. The answer is given in the following proposition; see Haslett & Puntanen (2010d,c).

Proposition 11.6. *Consider the models* $\mathscr{M}_f$ *and* $\underline{\mathscr{M}}_f$ *(with new unobserved future observations), where* $\mathscr{C}(\mathbf{X}_f') \subset \mathscr{C}(\mathbf{X}')$ *and* $\mathscr{C}(\underline{\mathbf{X}}_f') \subset \mathscr{C}(\underline{\mathbf{X}}')$*. Then every representation of the* BLUP *for* $\mathbf{y}_f$ *under the model* $\mathscr{M}_f$ *is also a* BLUP *for* $\mathbf{y}_f$ *under the model* $\underline{\mathscr{M}}_f$ *if and only if*

$$\mathscr{C}\begin{pmatrix} \underline{\mathbf{X}} & \underline{\mathbf{V}}_{11}\underline{\mathbf{M}} \\ \underline{\mathbf{X}}_f & \underline{\mathbf{V}}_{21}\underline{\mathbf{M}} \end{pmatrix} \subset \mathscr{C}\begin{pmatrix} \mathbf{X} & \mathbf{V}_{11}\mathbf{M} \\ \mathbf{X}_f & \mathbf{V}_{21}\mathbf{M} \end{pmatrix}. \tag{11.79}$$

Proof. For the proof it is convenient to observe that (11.79) holds if and only if

$$\mathscr{C}\begin{pmatrix} \underline{\mathbf{V}}_{11}\underline{\mathbf{M}} \\ \underline{\mathbf{V}}_{21}\underline{\mathbf{M}} \end{pmatrix} \subset \mathscr{C}\begin{pmatrix} \mathbf{X} & \mathbf{V}_{11}\mathbf{M} \\ \mathbf{X}_f & \mathbf{V}_{21}\mathbf{M} \end{pmatrix}, \tag{11.80a}$$

and

$$\mathscr{C}\begin{pmatrix} \underline{\mathbf{X}} \\ \underline{\mathbf{X}}_f \end{pmatrix} \subset \mathscr{C}\begin{pmatrix} \mathbf{X} & \mathbf{V}_{11}\mathbf{M} \\ \mathbf{X}_f & \mathbf{V}_{21}\mathbf{M} \end{pmatrix}. \tag{11.80b}$$

Let us first assume that every representation of the BLUP of $\mathbf{y}_f$ under $\mathscr{M}_f$ continues to be BLUP under $\underline{\mathscr{M}}_f$. Let $\mathbf{G}_0$ be a general solution to

$$\mathbf{G}(\mathbf{X} : \mathbf{V}_{11}\mathbf{M}) = (\mathbf{X}_f : \mathbf{V}_{21}\mathbf{M}) \tag{11.81}$$

i.e.,

$$\begin{aligned} \mathbf{G}_0 = \mathbf{X}_f(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}'\mathbf{W}^- + \mathbf{V}_{21}\mathbf{M}(\mathbf{M}\mathbf{V}_{11}\mathbf{M})^-\mathbf{M} \\ + \mathbf{F}(\mathbf{I}_n - \mathbf{P}_{(\mathbf{X}\,:\,\mathbf{V}_{11}\mathbf{M})}), \end{aligned} \tag{11.82}$$

where $\mathbf{F}$ is free to vary and $\mathbf{W}$ (and the related $\mathbf{U}$) are any matrices such that

$$\mathbf{W} = \mathbf{V}_{11} + \mathbf{X}\mathbf{U}\mathbf{X}', \quad \mathscr{C}(\mathbf{W}) = \mathscr{C}(\mathbf{X} : \mathbf{V}_{11}). \tag{11.83}$$

Then $\mathbf{G}_0$ has to satisfy also the other fundamental BLUP equation:

$$\mathbf{G}_0(\underline{\mathbf{X}} : \underline{\mathbf{V}}_{11}\underline{\mathbf{M}}) = (\underline{\mathbf{X}}_f : \underline{\mathbf{V}}_{21}\underline{\mathbf{M}}), \tag{11.84}$$

where $\underline{\mathbf{M}} = \mathbf{I}_n - \mathbf{P}_{\underline{\mathbf{X}}}$. The $\mathbf{X}$-part of the condition (11.84) is

$$\begin{aligned} \mathbf{X}_f(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}'\mathbf{W}^-\underline{\mathbf{X}} \\ + \mathbf{V}_{21}\mathbf{M}(\mathbf{M}\mathbf{V}_{11}\mathbf{M})^-\mathbf{M}\underline{\mathbf{X}} \\ + \mathbf{F}(\mathbf{I}_n - \mathbf{P}_{(\mathbf{X}\,:\,\mathbf{V}_{11}\mathbf{M})})\underline{\mathbf{X}} \\ = \underline{\mathbf{X}}_f\,. \end{aligned} \tag{11.85}$$

Because (11.85) must hold for all matrices $\mathbf{F}$, we necessarily have

$$\mathbf{F}(\mathbf{I}_n - \mathbf{P}_{(\mathbf{X}\,:\,\mathbf{V}_{11}\mathbf{M})})\underline{\mathbf{X}} = \mathbf{0} \quad \text{for all } \mathbf{F}\,, \tag{11.86}$$

which further implies that $\mathscr{C}(\underline{\mathbf{X}}) \subset \mathscr{C}(\mathbf{X} : \mathbf{V}_{11}\mathbf{M})$, and hence

$$\underline{\mathbf{X}} = \mathbf{X}\mathbf{K}_1 + \mathbf{V}_{11}\mathbf{M}\mathbf{K}_2 \quad \text{for some } \mathbf{K}_1 \text{ and } \mathbf{K}_2\,. \tag{11.87}$$

The rest of the proof is left as an exercise; see Haslett & Puntanen (2010c). □

Of course, if the models $\mathscr{M}_f$ and $\underline{\mathscr{M}}_f$ we have the same model matrix part $\binom{\mathbf{X}}{\mathbf{X}_f}$, then every representation of the BLUP for $\mathbf{y}_f$ under the model $\mathscr{M}_f$ is also a BLUP for $\mathbf{y}_f$ under the model $\underline{\mathscr{M}}_f$ if and only if

$$\mathscr{C}\begin{pmatrix} \underline{\mathbf{V}}_{11}\mathbf{M} \\ \underline{\mathbf{V}}_{21}\mathbf{M} \end{pmatrix} \subset \mathscr{C}\begin{pmatrix} \mathbf{X} & \mathbf{V}_{11}\mathbf{M} \\ \mathbf{X}_f & \mathbf{V}_{21}\mathbf{M} \end{pmatrix}. \tag{11.88}$$

Notice that (11.88) implies Proposition 11.5 (p. 278).

11.6 Exercises

11.1. Prove that $\mathbf{AYB} = \mathbf{C}$ can be rewritten as $(\mathbf{B}' \otimes \mathbf{A})\operatorname{vec}(\mathbf{Y}) = \operatorname{vec}(\mathbf{C})$. See also Exercise 0.23 (p. 52).

Abadir & Magnus (2005, p. 282), Harville (1997, p. 341).

11.2. Introduce the general solution to $\mathbf{AYB} = \mathbf{C}$ using Exercise 11.1 and the fact that the general solution to a consistent equation $\mathbf{Uz} = \mathbf{t}$ is $\mathbf{z}_0 = \mathbf{U}^-\mathbf{t} + (\mathbf{I} - \mathbf{U}^-\mathbf{U})\mathbf{w}$, where $\mathbf{w}$ is free to vary.

Rao & Mitra (1971b, pp. 24–25).

11.3. Consider linear models $\mathscr{M}_1 = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}_1\}$ and $\mathscr{M}_2 = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}_2\}$. Prove the equivalence of the following statements:

(a) $\mathbf{G}(\mathbf{X} : \mathbf{V}_1\mathbf{M} : \mathbf{V}_2\mathbf{M}) = (\mathbf{X} : \mathbf{0} : \mathbf{0})$ has a solution for $\mathbf{G}$,

(b) $\mathscr{C}(\mathbf{V}_1\mathbf{M} : \mathbf{V}_2\mathbf{M}) \cap \mathscr{C}(\mathbf{X}) = \{\mathbf{0}\}$,

(c) $\mathscr{C}\begin{pmatrix} \mathbf{MV}_1 \\ \mathbf{MV}_2 \end{pmatrix} \subset \mathscr{C}\begin{pmatrix} \mathbf{MV}_1\mathbf{M} \\ \mathbf{MV}_2\mathbf{M} \end{pmatrix}$.

Mitra & Moore (1973, Th. 3.2).

11.4. Show that $\mathscr{C}(\mathbf{X}_{*1}) \cap \mathscr{C}(\mathbf{X}_{*2}) = \{\mathbf{0}\}$ in (11.50) (p. 274).

11.5. Prove that $\begin{pmatrix} \mathbf{I}_n \\ -\mathbf{Z}' \end{pmatrix}\mathbf{M} \in \left\{\begin{pmatrix} \mathbf{X} & \mathbf{Z} \\ \mathbf{0} & \mathbf{I}_q \end{pmatrix}^{\perp}\right\}$.

11.6. Confirm the equivalence between (b)–(e) in Proposition 11.4 (p. 277).

11.7. Complete the proof of Proposition 11.6 (p. 279).

11.8. Let $\mathbf{X}$ be an $n \times p$ matrix satisfying $\mathbf{1}_n \in \mathscr{C}(\mathbf{X})$ and suppose that $\mathbf{Z} \in \mathbb{R}^{n \times q}$ is partitioned as $\mathbf{Z} = (\mathbf{Z}_1 : \mathbf{Z}_2)$, $\mathbf{Z}_i \in \mathbb{R}^{n \times q_i}$, $i = 1, 2$. Let each row of $\mathbf{Z}_1$ have one element 1 and all others 0 thus implying $\mathbf{Z}_1 \mathbf{1}_{q_1} = \mathbf{1}_n$. Moreover, let $\mathbf{C} = \mathbf{I}_{q_1} - \mathbf{1}_{q_1}\mathbf{1}'_{q_1}/q_1$ denote the $q_1 \times q_1$ centering matrix, and denote

$$\mathbf{D}_1 = \begin{pmatrix} \mathbf{I}_{q_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{G} \end{pmatrix}, \qquad \mathbf{D}_2 = \begin{pmatrix} \mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{G} \end{pmatrix},$$

where $\mathbf{G}$ is an arbitrary nonnegative definite $q_2 \times q_2$ matrix. Consider now two mixed models:

$$\mathscr{L}_1 = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma},\, \mathbf{D}_1,\, \mathbf{R}\}, \qquad \mathscr{L}_2 = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma},\, \mathbf{D}_2,\, \mathbf{R}\},$$

where the only difference appears in $\mathbf{D}_1$ and $\mathbf{D}_2$. Confirm that in this situation $\mathbf{D}_1\mathbf{Z}'\mathbf{M} = \mathbf{D}_2\mathbf{Z}'\mathbf{M}$ and so the condition of Proposition 11.4 (p. 277) is satisfied, and thus the BLUEs and BLUPs are the same under $\mathscr{L}_1$ and $\mathscr{L}_2$.

Möls (2004, p. 55).

11.9. Consider the models $\mathscr{M}_1 = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \mathbf{V}_1\}$ and $\mathscr{M}_2 = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \mathbf{V}_2\}$, where $\operatorname{rank}(\mathbf{X}) = r$. Denote $\mathbf{P}_{\mathbf{X};\mathbf{W}_1^+} = \mathbf{X}(\mathbf{X}'\mathbf{W}_1^+\mathbf{X})^-\mathbf{X}'\mathbf{W}_1^+$, where $\mathbf{W}_1 = \mathbf{V}_1 + \mathbf{X}\mathbf{U}\mathbf{U}'\mathbf{X}'$, and $\mathscr{C}(\mathbf{W}_1) = \mathscr{C}(\mathbf{X} : \mathbf{V}_1)$ and so $\mathbf{P}_{\mathbf{X};\mathbf{W}_1^+}\mathbf{y}$ is the BLUE for $\mathbf{X}\boldsymbol{\beta}$ under $\mathscr{M}_1$. Show that $\mathbf{P}_{\mathbf{X};\mathbf{W}_1^+}\mathbf{y}$ is the BLUE for $\mathbf{X}\boldsymbol{\beta}$ also under $\mathscr{M}_2$ if and only if any of the following equivalent conditions holds:

(a) $\mathbf{X}'\mathbf{W}_1^+\mathbf{V}_2\mathbf{M} = \mathbf{0}$,
(b) $\mathscr{C}(\mathbf{V}_2\mathbf{W}_1^+\mathbf{X}) \subset \mathscr{C}(\mathbf{X})$,
(c) $\mathscr{C}(\mathbf{V}_2\mathbf{M}) \subset \mathscr{N}(\mathbf{X}'\mathbf{W}_1^+) = \mathscr{C}(\mathbf{W}_1^+\mathbf{X})^\perp = \mathscr{C}(\mathbf{G})$; $\mathbf{G} \in \{(\mathbf{W}_1^+\mathbf{X})^\perp\}$,
(d) $\mathscr{C}(\mathbf{W}_1^+\mathbf{X})$ is spanned by a set of r proper eigenvectors of $\mathbf{V}_2$ with respect to $\mathbf{W}_1$,
(e) $\mathscr{C}(\mathbf{X})$ is spanned by a set of r eigenvectors of $\mathbf{V}_2\mathbf{W}_1^+$,
(f) $\mathbf{P}_{\mathbf{X};\mathbf{W}_1^+}\mathbf{V}_2$ is symmetric,
(g) $\mathbf{V}_2 \in \{\mathbf{V}_2 \in \mathrm{NND}_n : \mathbf{V}_2 = \mathbf{X}\mathbf{N}_1\mathbf{X}' + \mathbf{G}\mathbf{N}_2\mathbf{G}'$ for some $\mathbf{N}_1$ and $\mathbf{N}_2\}$.

Hint: Notice that according to Exercise 5.6 (p. 143),

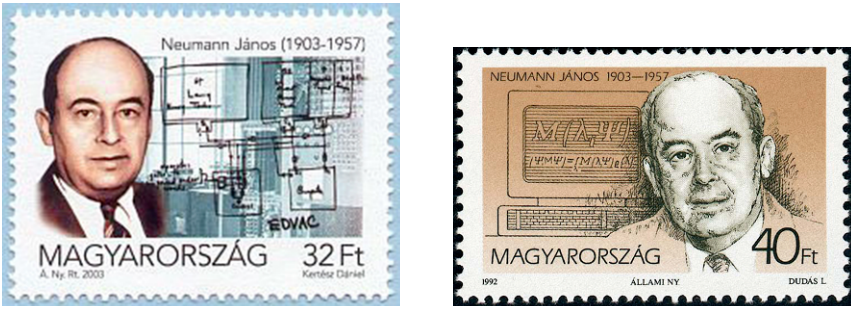
$$\mathscr{C}(\mathbf{G}) \oplus \mathscr{C}(\mathbf{X}) = \mathbb{R}^n, \quad \mathscr{C}(\mathbf{G})^\perp \oplus \mathscr{C}(\mathbf{X})^\perp = \mathbb{R}^n,$$

where $\mathbf{G} \in \{(\mathbf{W}_1^+\mathbf{X})^\perp\}$ and hence every $\mathbf{V}_2 \in \mathrm{NND}_n$ can be written as $\mathbf{V}_2 = \mathbf{F}\mathbf{F}'$ for some $\mathbf{A}_1$ and $\mathbf{A}_2$ such that $\mathbf{F} = \mathbf{X}\mathbf{A}_1 + \mathbf{G}\mathbf{A}_2$, and so

$$\mathbf{V}_2 = \mathbf{X}\mathbf{A}_1\mathbf{A}_1'\mathbf{X}' + \mathbf{G}\mathbf{A}_2\mathbf{A}_2'\mathbf{G}' + \mathbf{X}\mathbf{A}_1\mathbf{A}_2\mathbf{G}' + \mathbf{G}\mathbf{A}_2\mathbf{A}_1'\mathbf{X}'.$$

For positive definite $\mathbf{V}_1$ and $\mathbf{V}_2$, see Exercise 2.3 (p. 88).

Mitra & Moore (1973, Th. 2.2), Rao & Mitra (1971b, p. 157).



Philatelic Item 11.1 John von Neumann (1903–1957) was a Hungarian–American mathematician who made major contributions to a vast range of fields, including set theory, functional analysis, quantum mechanics, ergodic theory, continuous geometry, economics and game theory, computer science, numerical analysis, hydrodynamics (of explosions), and statistics. For the von Neumann trace inequality see Exercise 19.5 (p. 410. The stamps were issued by Hungary, in 2003 (left panel, *Scott* 3824) and in 1992 (right panel, *Scott* 3354).

Philatelic Item 11.2 Jules Henri Poincaré (1854–1912) was one of France's greatest mathematicians and theoretical physicists and in 1886 he was appointed to the Chair of Mathematical Physics and Probability at the Sorbonne. For the Poincaré separation theorem for eigenvalues, see Proposition 19.2 (p. 398). The stamp on the left was issued by France in 1952 (*Scott* B270) and the stamp on the right, which was issued by Portugal in 2000 (*Scott* 2345b), also depicts (in the centre) Kurt Gödel (1906–1978) and (on the right) Andrei Nikolaevich Kolmogorov (1903–1987).

Chapter 12
Invariance with Respect to the Choice of Generalized Inverse

> *Apart from scoring two goals, Finland did nothing spectacular.*
>
> TIMES OF INDIA: *Comment after Finland had beaten India in football 2–0 in Madras, 25 January 1993.*

Because we are heavily using generalized inverses, it is very important to know something about the invariance of the matrix expressions with respect to the choice of generalized inverses appearing in the expression. Our main attention in this chapter is focused on the matrix product $\mathbf{AB^-C}$, but one can make corresponding questions concerning, for example, the invariance of the column space $\mathscr{C}(\mathbf{AB^-C})$, the rank of $\mathbf{AB^-C}$, and the eigenvalues of $\mathbf{AB^-C}$.

Theorem 12 (Invariance wrt the choice of generalized inverse). *Let* $\mathbf{A} \neq \mathbf{0}$*, and* $\mathbf{C} \neq \mathbf{0}$*. Then the following two statements are equivalent:*

(a) $\mathbf{AB^-C}$ *is invariant with respect to the choice of* $\mathbf{B}^-$*,*

(b) $\mathscr{C}(\mathbf{C}) \subset \mathscr{C}(\mathbf{B})$ *and* $\mathscr{C}(\mathbf{A}') \subset \mathscr{C}(\mathbf{B}')$*.*

In particular, the product $\mathbf{BB^-C}$ *is invariant with respect to the choice of* $\mathbf{B}^-$ *if and only if* $\mathscr{C}(\mathbf{C}) \subset \mathscr{C}(\mathbf{B})$*; in other words,*

$$\mathbf{BB^-C} = \mathbf{C} \ \text{ for all } \mathbf{B}^- \iff \mathscr{C}(\mathbf{C}) \subset \mathscr{C}(\mathbf{B})\,. \tag{12.1}$$

Proof. Suppose that (b) holds. Then there exist matrices $\mathbf{U}$ and $\mathbf{Z}$ such that $\mathbf{C} = \mathbf{BU}$, $\mathbf{A}' = \mathbf{B}'\mathbf{Z}$ and hence

$$\mathbf{AB^-C} = \mathbf{Z}'\mathbf{BB^-BU} = \mathbf{Z}'\mathbf{BU}\,, \tag{12.2}$$

which does not depend on the choice of $\mathbf{B}^-$. As pointed out by Bapat (2000, p. 35), the matrices $\mathbf{U}$ and $\mathbf{Z}$ are not unique, but if $\mathbf{U}_1$ and $\mathbf{Z}_1$ are some other choices, then $\mathbf{AB^-C} = \mathbf{Z}'\mathbf{BU} = \mathbf{Z}_1'\mathbf{BU}_1$. To prove that (a) implies (b), we first write, in view of Proposition 4.1 (p. 111), the general representation of $\mathbf{B}^-$ ($= \mathbf{G}$) as follows:

$$\mathbf{G} = \mathbf{B}^+ + \mathbf{V}(\mathbf{I} - \mathbf{P_B}) + (\mathbf{I} - \mathbf{P_{B'}})\mathbf{W}\,, \tag{12.3}$$

where $\mathbf{V}$ and $\mathbf{W}$ are arbitrary matrices. This means that varying $\mathbf{V}$ and $\mathbf{W}$ in (12.3) we can generate all generalized inverses of $\mathbf{B}$, or, in other words,

S. Puntanen et al., *Matrix Tricks for Linear Statistical Models: Our Personal Top Twenty*,
DOI 10.1007/978-3-642-10473-2_13, © Springer-Verlag Berlin Heidelberg 2011

$$\{\mathbf{B}^-\} = \{\, \mathbf{G} : \mathbf{G} = \mathbf{B}^+ + \mathbf{V}(\mathbf{I}-\mathbf{P}_{\mathbf{B}}) + (\mathbf{I}-\mathbf{P}_{\mathbf{B}'})\mathbf{W} \text{ for some } \mathbf{V},\, \mathbf{W} \,\}\,. \tag{12.4}$$

Premultiplying (12.3) by $\mathbf{A}$ and postmultiplying (12.3) by $\mathbf{C}$, yields

$$\mathbf{AGC} = \mathbf{AB}^+\mathbf{C} + \mathbf{AV}(\mathbf{I}-\mathbf{P}_{\mathbf{B}})\mathbf{C} + \mathbf{A}(\mathbf{I}-\mathbf{P}_{\mathbf{B}'})\mathbf{WC}\,. \tag{12.5}$$

The invariance assumption means that the equality $\mathbf{AGC} = \mathbf{AB}^+\mathbf{C}$ must always hold, that is,

$$\mathbf{AV}(\mathbf{I}-\mathbf{P}_{\mathbf{B}})\mathbf{C} + \mathbf{A}(\mathbf{I}-\mathbf{P}_{\mathbf{B}'})\mathbf{WC} = \mathbf{0} \quad \text{for all } \mathbf{V},\, \mathbf{W}\,. \tag{12.6}$$

Choosing $\mathbf{W} = \mathbf{B}'$, we obtain

$$\mathbf{AV}(\mathbf{I}-\mathbf{P}_{\mathbf{B}})\mathbf{C} = \mathbf{0} \quad \text{for all } \mathbf{V}\,. \tag{12.7}$$

We may now recall the following result:

$$\text{Let } \mathbf{y} \neq \mathbf{0} \text{ be a given vector. Then } \mathbf{y}'\mathbf{Qx} = 0 \text{ for all } \mathbf{Q} \implies \mathbf{x} = \mathbf{0}\,. \tag{12.8}$$

By the assumption, at least one row of $\mathbf{A}$, say $\mathbf{a}'$, has property $\mathbf{a}' \neq \mathbf{0}'$. Using this row as $\mathbf{y}'$, and using $\mathbf{V}$ as $\mathbf{Q}$, then, according to (12.8), every column of $(\mathbf{I}-\mathbf{P}_{\mathbf{B}})\mathbf{C}$ equals $\mathbf{0}$, i.e.,

$$(\mathbf{I}-\mathbf{P}_{\mathbf{B}})\mathbf{C} = \mathbf{0}\,, \tag{12.9}$$

which means that $\mathscr{C}(\mathbf{C}) \subset \mathscr{C}(\mathbf{B})$. The necessity of $\mathscr{C}(\mathbf{A}') \subset \mathscr{C}(\mathbf{B}')$ can be shown in the same way. □

The above proof follows that of Rao & Mitra (1971b, Lemma 2.2.4, and the Complement 14, p. 43), and Rao, Mitra & Bhimasankaram (1972, Lemma 1).

12.1 Invariance of $\mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'$

To take an extremely simple example of the use of Theorem 12, we may consider the matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'. \tag{12.10}$$

We surely have met $\mathbf{H}$ already repeatedly, but there is no harm to notice that condition $\mathscr{C}(\mathbf{X}') \subset \mathscr{C}(\mathbf{X}'\mathbf{X}) = \mathscr{C}(\mathbf{X}')$ guarantees the uniqueness of $\mathbf{H}$.

12.2 Estimability and Invariance

A parametric function $\mathbf{K}'\boldsymbol{\beta}$ is said to be estimable if it has a linear unbiased estimator, i.e., there exists a matrix $\mathbf{A}$ such that

$$\mathrm{E}(\mathbf{A}\mathbf{y}) = \mathbf{A}\mathbf{X}\boldsymbol{\beta} = \mathbf{K}'\boldsymbol{\beta} \quad \text{for all } \boldsymbol{\beta} \in \mathbb{R}^p\,; \tag{12.11}$$

and hence

$$\mathbf{K}'\boldsymbol{\beta} \text{ is estimable} \iff \exists\, \mathbf{A}\colon \mathbf{K} = \mathbf{X}'\mathbf{A}' \iff \mathscr{C}(\mathbf{K}) \subset \mathscr{C}(\mathbf{X}')\,. \tag{12.12}$$

The OLSE for $\mathbf{K}'\boldsymbol{\beta}$ it is defined as $\mathbf{K}'\hat{\boldsymbol{\beta}}$ where $\hat{\boldsymbol{\beta}}$ is an arbitrary solution to the normal equation

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}\,. \tag{12.13}$$

The general solution to the normal equation (which is always consistent) is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y} + [\mathbf{I}_p - (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X}]\mathbf{z}\,, \tag{12.14}$$

where $\mathbf{z} \in \mathbb{R}^p$ is free to vary and $(\mathbf{X}'\mathbf{X})^{-}$ is an arbitrary (but fixed) generalized inverse of $\mathbf{X}'\mathbf{X}$. We can express (12.14) also as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y} + (\mathbf{I}_p - \mathbf{P}_{\mathbf{X}'})\mathbf{z}_1\,, \tag{12.15}$$

or

$$\hat{\boldsymbol{\beta}} = \mathbf{X}^{+}\mathbf{y} + (\mathbf{I}_p - \mathbf{P}_{\mathbf{X}'})\mathbf{z}_2\,, \tag{12.16}$$

where $\mathbf{z}_1$ and $\mathbf{z}_2$ are free to vary. On the other hand, all solutions to the normal equation can be generated via

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}\,, \tag{12.17}$$

by letting $(\mathbf{X}'\mathbf{X})^{-}$ vary through all generalized inverses.

Hence the expression for OLSE of $\mathbf{K}'\boldsymbol{\beta}$ is

$$\begin{aligned}\mathrm{OLSE}(\mathbf{K}'\boldsymbol{\beta}) &= \mathbf{K}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y} + \mathbf{K}'(\mathbf{I}_p - \mathbf{P}_{\mathbf{X}'})\mathbf{z}_1 \\ &= \mathbf{K}'\mathbf{X}^{+}\mathbf{y} + \mathbf{K}'(\mathbf{I}_p - \mathbf{P}_{\mathbf{X}'})\mathbf{z}_2\,, \end{aligned} \tag{12.18}$$

where $\mathbf{z}_1$ and $\mathbf{z}_2$ are free to vary. An alternative way to express the set of OLSEs for $\mathbf{K}'\boldsymbol{\beta}$ is of course simply

$$\mathrm{OLSE}(\mathbf{K}'\boldsymbol{\beta}) = \mathbf{K}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}\,, \tag{12.19}$$

where $(\mathbf{X}'\mathbf{X})^{-}$ varies through all generalized inverses of $\mathbf{X}'\mathbf{X}$. Therefore, it is easy to confirm the following result:

$$\mathbf{K}'\boldsymbol{\beta} \text{ is estimable} \tag{12.20a}$$

$$\iff$$

$$\mathbf{K}'\hat{\boldsymbol{\beta}} = \mathbf{K}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y} \text{ is invariant for all } (\mathbf{X}'\mathbf{X})^{-} \tag{12.20b}$$

$$\iff$$

$$\mathscr{C}(\mathbf{K}) \subset \mathscr{C}(\mathbf{X}')\,. \tag{12.20c}$$

12.3 Properties of $\mathbf{X}'\mathbf{W}^-\mathbf{X}$

Proposition 12.1. *Let* $\mathbf{V}$ *be an* $n\times n$ *symmetric nonnegative definite matrix, let* $\mathbf{X}$ *be an* $n \times p$ *matrix, and define* $\mathbf{W}$ *as*

$$\mathbf{W} = \mathbf{V} + \mathbf{X}\mathbf{U}\mathbf{X}', \tag{12.21}$$

where $\mathbf{U}$ *is a* $p \times p$ *matrix. Then the following statements are equivalent:*

(a) $\mathscr{C}(\mathbf{X}) \subset \mathscr{C}(\mathbf{W})$,

(b) $\mathscr{C}(\mathbf{X} : \mathbf{V}) = \mathscr{C}(\mathbf{W})$,

(c) $\operatorname{rank}(\mathbf{X} : \mathbf{V}) = \operatorname{rank}(\mathbf{W})$,

(d) $\mathbf{X}'\mathbf{W}^-\mathbf{X}$ *is invariant for any choice of* $\mathbf{W}^-$,

(e) $\mathscr{C}(\mathbf{X}'\mathbf{W}^-\mathbf{X})$ *is invariant for any choice of* $\mathbf{W}^-$,

(f) $\mathscr{C}(\mathbf{X}'\mathbf{W}^-\mathbf{X}) = \mathscr{C}(\mathbf{X}')$ *for any choice of* $\mathbf{W}^-$,

(g) $\operatorname{rank}(\mathbf{X}'\mathbf{W}^-\mathbf{X}) = \operatorname{rank}(\mathbf{X})$ *irrespective of the choice of* $\mathbf{W}^-$,

(h) $\operatorname{rank}(\mathbf{X}'\mathbf{W}^-\mathbf{X})$ *is invariant with respect to the choice of* $\mathbf{W}^-$,

(i) $\mathbf{X}(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}'\mathbf{W}^-\mathbf{X} = \mathbf{X}$ *for any choices of the generalized inverses involved.*

Moreover, each of these statements is equivalent to

(a') $\mathscr{C}(\mathbf{X}) \subset \mathscr{C}(\mathbf{W}')$,

and hence to the statements (b')–(i') *obtained from* (b)–(i), *by setting* $\mathbf{W}'$ *in place of* $\mathbf{W}$.

Proof. We will prove only some of the implications between the above statements. Further references: Baksalary & Puntanen (1989); Baksalary, Puntanen & Styan (1990b, Th. 2); Baksalary & Mathew (1990, Th. 2), Harville (1997, p. 468).

We begin by noting that the inclusion

$$\mathscr{C}(\mathbf{W}) = \mathscr{C}(\mathbf{V} + \mathbf{X}\mathbf{U}\mathbf{X}') = \mathscr{C}\left[(\mathbf{V} : \mathbf{X})\begin{pmatrix}\mathbf{I}_n \\ \mathbf{U}\mathbf{X}'\end{pmatrix}\right] \subset \mathscr{C}(\mathbf{X} : \mathbf{V}) \tag{12.22}$$

trivially holds, and similarly we have

$$\mathscr{C}(\mathbf{W}') = \mathscr{C}(\mathbf{V} + \mathbf{X}\mathbf{U}'\mathbf{X}') \subset \mathscr{C}(\mathbf{X} : \mathbf{V})\,. \tag{12.23}$$

Moreover, we recall the decomposition

$$\mathscr{C}(\mathbf{X} : \mathbf{V}) = \mathscr{C}(\mathbf{X} : \mathbf{V}\mathbf{M}) = \mathscr{C}(\mathbf{X}) \oplus \mathscr{C}(\mathbf{V}\mathbf{M})\,, \tag{12.24}$$

and the obvious inclusions

$$\mathscr{C}(\mathbf{VM}) = \mathscr{C}[(\mathbf{V} + \mathbf{XUX}')\mathbf{M}] = \mathscr{C}(\mathbf{WM}) \subset \mathscr{C}(\mathbf{W}), \tag{12.25a}$$

$$\mathscr{C}(\mathbf{VM}) = \mathscr{C}[(\mathbf{V} + \mathbf{XU}'\mathbf{X}')\mathbf{M}] = \mathscr{C}(\mathbf{W}'\mathbf{M}) \subset \mathscr{C}(\mathbf{W}'). \tag{12.25b}$$

(a) $\iff$ (b): Inclusion (12.22) means that (b) is equivalent to

$$\mathscr{C}(\mathbf{X} : \mathbf{V}) = \mathscr{C}(\mathbf{X} : \mathbf{VM}) \subset \mathscr{C}(\mathbf{W}). \tag{12.26}$$

In view of (12.25a), (12.26) holds if and only if $\mathscr{C}(\mathbf{X}) \subset \mathscr{C}(\mathbf{W})$, and hence it seen that (a) $\iff$ (b).

(b) $\iff$ (c): This follows immediately from $\mathscr{C}(\mathbf{W}) \subset \mathscr{C}(\mathbf{X} : \mathbf{V})$.

(c) $\iff$ (c'): In view of (12.23), we see that (c'), which is

$$\operatorname{rank}(\mathbf{X} : \mathbf{V}) = \operatorname{rank}(\mathbf{W}'), \tag{12.27}$$

is clearly equivalent to $\operatorname{rank}(\mathbf{X} : \mathbf{V}) = \operatorname{rank}(\mathbf{W})$, which is (c).

(a) $\iff$ (a'). Since (a) $\iff$ (c'), i.e., $\operatorname{rank}(\mathbf{X} : \mathbf{V}) = \operatorname{rank}(\mathbf{W}')$, the inclusions (12.23) and (12.25b) mean that (c') holds if and only if $\mathscr{C}(\mathbf{X}) \subset \mathscr{C}(\mathbf{W}')$. This shows that (a) $\iff$ (a').

(a) $\iff$ (d): The product $\mathbf{X}'\mathbf{W}^-\mathbf{X}$ is invariant for any choice of $\mathbf{W}^-$ if and only if both (a) and (a') hold. But because (a) $\iff$ (a'), we have indeed (a) $\iff$ (d).

(f) $\iff$ (g) $\iff$ (i): This is easy to confirm. Notice that applying the rank cancellation rule and (g) into the equation

$$\mathbf{X}'\mathbf{W}^-\mathbf{X}(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}'\mathbf{W}^-\mathbf{X} = \mathbf{X}'\mathbf{W}^-\mathbf{X}, \tag{12.28}$$

yields immediately the claim (i):

$$\mathbf{X}(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}'\mathbf{W}^-\mathbf{X} = \mathbf{X}. \tag{12.29}$$

(a') $\implies$ (g): Since $\mathbf{VM} = \mathbf{WM}$, we have

$$\mathbf{X}'\mathbf{W}^-\mathbf{VM} = \mathbf{X}'\mathbf{W}^-\mathbf{WM} = \mathbf{X}'\mathbf{M} = \mathbf{0}, \tag{12.30}$$

where we have used (a'), i.e., $\mathscr{C}(\mathbf{X}) \subset \mathscr{C}(\mathbf{W}')$. Moreover,

$$\begin{aligned}\operatorname{rank}(\mathbf{X}'\mathbf{W}^-\mathbf{X}) &= \operatorname{rank}[\mathbf{X}'\mathbf{W}^-(\mathbf{X} : \mathbf{VM})] \\ &= \operatorname{rank}(\mathbf{X}'\mathbf{W}^-\mathbf{W}) = \operatorname{rank}(\mathbf{X}).\end{aligned} \tag{12.31}$$

The second equality in (12.31) follows from (b) because

$$\mathscr{C}(\mathbf{X} : \mathbf{VM}) = \mathscr{C}(\mathbf{W}) \implies \operatorname{rank}[\mathbf{A}(\mathbf{X} : \mathbf{VM})] = \operatorname{rank}(\mathbf{AW}) \tag{12.32}$$

for any matrix $\mathbf{A}$, and the third equality in (12.31) comes from (a').

Notice that (g) implies trivially (e) as well as (h).

We do not show that (f) implies (a) nor that (h) implies (a). For these proofs, see Baksalary & Puntanen (1989); Baksalary, Puntanen & Styan

(1990b, Th. 2); Baksalary & Mathew (1990, Th. 2). For related considerations, see also Baksalary & Puntanen (1989, §3), and Harville (1997, pp. 468–472). □

We can now also easily check that the matrix

$$\mathbf{G} = \mathbf{X}(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}'\mathbf{W}^- \tag{12.33}$$

satisfies the BLUE equations. First, part (i) of Proposition 12.1 guarantees that $\mathbf{GX} = \mathbf{X}$. Furthermore, because $\mathbf{VM} = \mathbf{WM}$, we have, in view of $\mathbf{X}'\mathbf{W}^-\mathbf{W} = \mathbf{X}'$ and (12.30),

$$\begin{aligned}\mathbf{X}(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}'\mathbf{W}^-\mathbf{VM} &= \mathbf{X}(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}'\mathbf{W}^-\mathbf{WM} \\ &= \mathbf{X}(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}'\mathbf{M} = \mathbf{0}\,. \end{aligned} \tag{12.34}$$

Therefore, one representation for the BLUE of $\mathbf{X}\boldsymbol{\beta}$ is

$$\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta}) = \mathbf{X}[\mathbf{X}'(\mathbf{V}+\mathbf{XUX}')^-\mathbf{X}]^-\mathbf{X}'(\mathbf{V}+\mathbf{XUX}')^-\mathbf{y}\,, \tag{12.35}$$

where $\mathbf{U}$ is defined so that it satisfies any of the equivalent conditions (a)–(i) of the above Proposition 12.1.

12.4 Exercises

12.1. Show that the generalized normal equation $\mathbf{X}'\mathbf{W}^-\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{W}^-\mathbf{y}$ is consistent, i.e., it has a solution for $\boldsymbol{\beta}$ when $\mathbf{W}$ is defined as

$$\mathbf{W} = \mathbf{V} + \mathbf{XUX}', \quad \mathscr{C}(\mathbf{W}) = \mathscr{C}(\mathbf{X} : \mathbf{V})\,,$$

and $\mathbf{y} \in \mathscr{C}(\mathbf{X} : \mathbf{V})$.

12.2. Prove the following: Let $\mathbf{y} \neq \mathbf{0}$ be a given vector. Then $\mathbf{y}'\mathbf{Qx} = 0$ for all $\mathbf{Q} \implies \mathbf{x} = \mathbf{0}$.

12.3. Complete the proof of Proposition 12.1 (p. 286).

12.4. Consider the linear model $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$. Show that the following statements are equivalent:

(a) $\mathbf{M}(\mathbf{MVM})^-\mathbf{M} \in \{\mathbf{V}^-\}$ for any choice of $(\mathbf{MVM})^-$,
(b) $\mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{V}) = \{\mathbf{0}\}$,
(c) $\mathrm{cov}(\mathbf{X}\tilde{\boldsymbol{\beta}}) = \mathbf{0}$.

Furthermore, confirm the following statements:

(d) $\mathbf{M}(\mathbf{MVM})^-\mathbf{M} \in \{(\mathbf{MVM})^-\}$ for any choice of $(\mathbf{MVM})^-$,

(e) $\mathbf{M}(\mathbf{MVM})^{-}\mathbf{M} = (\mathbf{MVM})^{+}$ for any choice of $(\mathbf{MVM})^{-}$ if and only if $\mathscr{C}(\mathbf{X} : \mathbf{V}) = \mathbb{R}^n$.

Puntanen & Scott (1996, Th. 2.2).

12.5. Consider again the linear model $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$, and let $\mathbf{F}$ be a given $n \times n$ matrix. Show that the equation

$$\mathbf{X}'\mathbf{F}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{F}\mathbf{y}$$

has a solution for $\boldsymbol{\beta}$ for all $\mathbf{y} \in \mathscr{C}(\mathbf{X} : \mathbf{V})$ if and only if $\mathscr{C}(\mathbf{X}'\mathbf{F}\mathbf{V}) \subset \mathscr{C}(\mathbf{X}'\mathbf{F}\mathbf{X})$.

12.6 (Continued ...). Denote $\mathbf{P}_{\mathbf{X};\mathbf{F}} = \mathbf{X}(\mathbf{X}'\mathbf{F}\mathbf{X})^{-}\mathbf{X}'\mathbf{F}$, with $\mathbf{F}$ satisfying $\mathscr{C}(\mathbf{X}'\mathbf{F}\mathbf{V}) \subset \mathscr{C}(\mathbf{X}'\mathbf{F}\mathbf{X}) \neq \{\mathbf{0}\}$. Show the equivalence of the following statements:

(a) $\mathbf{P}_{\mathbf{X};\mathbf{F}}\mathbf{y}$ is invariant for any choice of $(\mathbf{X}'\mathbf{F}\mathbf{X})^{-}$ for all $\mathbf{y} \in \mathscr{C}(\mathbf{X} : \mathbf{V})$,
(b) $\mathbf{P}_{\mathbf{X};\mathbf{F}}\mathbf{y}$ is unbiased for $\mathbf{X}\boldsymbol{\beta}$,
(c) $\operatorname{rank}(\mathbf{X}'\mathbf{F}\mathbf{X}) = \operatorname{rank}(\mathbf{X})$.

Baksalary & Puntanen (1989, Lemma 1).

12.7. The nonnull $n \times m$ matrices $\mathbf{A}$ and $\mathbf{B}$ are said to be *parallel summable* if the matrix $\mathbf{A}(\mathbf{A} + \mathbf{B})^{-}\mathbf{B}$ is invariant with respect to the choice of $(\mathbf{A} + \mathbf{B})^{-}$. Show that the rank of $\mathbf{A}(\mathbf{A} + \mathbf{B})^{-}\mathbf{B}$ is invariant with respect to the choice of $(\mathbf{A} + \mathbf{B})^{-}$ if and only if $\mathbf{A}$ and $\mathbf{B}$ are parallel summable.

Rao & Mitra (1971b, pp. 188–189), Puntanen, Šemrl & Styan (1996).

Philatelic Item 12.1 Tadeusz Banachiewicz (1882–1954) was a Polish astronomer, mathematician and geodesist. For the Banachiewicz inversion formula for partitioned matrices, see Section 13.1 (p. 294). The stamp (left panel) was issued by Poland in 1983 (*Scott* 2565). For more about Banachiewicz and his inversion formula see Grala, Markiewicz & Styan (2000). The French mathematician Baron Augustin-Louis Cauchy (1789–1857) started the project of formulating and proving the theorems of infinitesimal calculus in a rigorous manner. For the Cauchy–Schwarz inequality see Chapter 20 (p. 415). The stamp (right panel) was issued by France in 1989 (*Scott* 2176).

Chapter 13
Block-Diagonalization and the Schur Complement

It's a great place for trees.
Alongside the trees you will find other trees,
which blend into a forest running for a thousand miles.
If you've never seen a forest running,
then Finland is the place for you.

Monty Python's Spamalot

In this chapter we present a block-diagonalization result for a symmetric nonnegative definite matrix. We may emphasize that the block-diagonalization result, sometimes called the Aitken block-diagonalization formula, due to Aitken (1939, Ch. 3, §29), is mathematically indeed quite simple just as it is. However, it is exceptionally handy and powerful tool for various situations arising in linear models and multivariate analysis, see, e.g., the derivation of the conditional multinormal distribution in Anderson (2003, §2.5); cf. also (9.21)–(9.22) (p. 193). We also consider the Schur complements whose usefulness in linear models and related areas can hardly be overestimated.

Theorem 13 (Block-diagonalization of an nnd matrix). *Let* $\mathbf{A}$ *be a symmetric nonnegative definite matrix partitioned as*

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}, \tag{13.1}$$

where $\mathbf{A}_{11}$ *is a square matrix. Then the following decomposition holds:*

$$\begin{aligned} \mathbf{A} &= \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{A}_{21}\mathbf{A}_{11}^{=} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-}\mathbf{A}_{12} \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{A}_{11}^{\sim}\mathbf{A}_{12} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \\ &:= \mathbf{BCD}, \end{aligned} \tag{13.2}$$

where $\mathbf{A}_{11}^{=}$*,* $\mathbf{A}_{11}^{-}$*, and* $\mathbf{A}_{11}^{\sim}$ *are arbitrary generalized inverses of* $\mathbf{A}_{11}$*, and*

$$\begin{aligned} \mathbf{A}_{22\cdot 1} = \mathbf{A}/\mathbf{A}_{11} &= \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-}\mathbf{A}_{12} \\ &= \textit{the Schur complement of } \mathbf{A}_{11} \textit{ in } \mathbf{A}\,. \end{aligned} \tag{13.3}$$

Moreover,

$$\begin{aligned} \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{A}_{21}\mathbf{A}_{11}^{=} & \mathbf{I} \end{pmatrix} \cdot \mathbf{A} \cdot \begin{pmatrix} \mathbf{I} & -\mathbf{A}_{11}^{\sim}\mathbf{A}_{12} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} &= \begin{pmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-}\mathbf{A}_{12} \end{pmatrix} \\ &= \mathbf{B}^{-1}\mathbf{A}\mathbf{D}^{-1}. \end{aligned} \tag{13.4}$$

S. Puntanen et al., *Matrix Tricks for Linear Statistical Models: Our Personal Top Twenty*,
DOI 10.1007/978-3-642-10473-2_14,

Proof. Straightforward calculation proves the result. In the proof we use the column space inclusion, see Section 1.3 (p. 62):

$$\mathscr{C}(\mathbf{A}_{12}) \subset \mathscr{C}(\mathbf{A}_{11})\,, \tag{13.5}$$

and Theorem 12 (p. 283); notice that also $\mathscr{C}(\mathbf{A}_{21}) \subset \mathscr{C}(\mathbf{A}_{22})$ holds. First,

$$\begin{aligned}\mathbf{BC} = \begin{pmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{A}_{21}\mathbf{A}_{11}^{=} & \mathbf{I}\end{pmatrix}\begin{pmatrix}\mathbf{A}_{11} & \mathbf{0}\\ \mathbf{0} & \mathbf{A}_{22\cdot 1}\end{pmatrix} &= \begin{pmatrix}\mathbf{A}_{11} & \mathbf{0}\\ \mathbf{A}_{21}\mathbf{A}_{11}^{=}\mathbf{A}_{11} & \mathbf{A}_{22\cdot 1}\end{pmatrix}\\ &= \begin{pmatrix}\mathbf{A}_{11} & \mathbf{0}\\ \mathbf{A}_{21} & \mathbf{A}_{22\cdot 1}\end{pmatrix}, \end{aligned} \tag{13.6}$$

where we have utilized the fact

$$\mathbf{A}_{21}\mathbf{A}_{11}^{=}\mathbf{A}_{11} = \mathbf{A}_{21}\,, \tag{13.7}$$

which follows from $\mathscr{C}(\mathbf{A}_{21}') = \mathscr{C}(\mathbf{A}_{12}) \subset \mathscr{C}(\mathbf{A}_{11}) = \mathscr{C}(\mathbf{A}_{11}')$. Moreover, on account of $\mathbf{A}_{11}\mathbf{A}_{11}^{\sim}\mathbf{A}_{12} = \mathbf{A}_{12}$, we have

$$\begin{aligned}\mathbf{BCD} &= \begin{pmatrix}\mathbf{A}_{11} & \mathbf{0}\\ \mathbf{A}_{21} & \mathbf{A}_{22\cdot 1}\end{pmatrix}\begin{pmatrix}\mathbf{I} & \mathbf{A}_{11}^{\sim}\mathbf{A}_{12}\\ \mathbf{0} & \mathbf{I}\end{pmatrix}\\ &= \begin{pmatrix}\mathbf{A}_{11} & \mathbf{A}_{11}\mathbf{A}_{11}^{\sim}\mathbf{A}_{12}\\ \mathbf{A}_{21} & \mathbf{A}_{21}\mathbf{A}_{11}^{\sim}\mathbf{A}_{12} + \mathbf{A}_{22\cdot 1}\end{pmatrix}\\ &= \begin{pmatrix}\mathbf{A}_{11} & \mathbf{A}_{12}\\ \mathbf{A}_{21} & \mathbf{A}_{21}\mathbf{A}_{11}^{\sim}\mathbf{A}_{12} + \mathbf{A}_{22\cdot 1}\end{pmatrix}. \end{aligned} \tag{13.8}$$

Finally

$$\mathbf{A}_{21}\mathbf{A}_{11}^{\sim}\mathbf{A}_{12} + \mathbf{A}_{22\cdot 1} = \mathbf{A}_{21}\mathbf{A}_{11}^{\sim}\mathbf{A}_{12} + \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-}\mathbf{A}_{12} = \mathbf{A}_{22}\,, \tag{13.9}$$

because

$$\mathbf{A}_{21}\mathbf{A}_{11}^{\sim}\mathbf{A}_{12} = \mathbf{A}_{21}\mathbf{A}_{11}^{-}\mathbf{A}_{12} \quad \text{for all } \mathbf{A}_{11}^{\sim},\, \mathbf{A}_{11}^{-} \in \{\mathbf{A}_{11}^{-}\}\,. \tag{13.10}$$

The property (13.4) follows from

$$\begin{pmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{U} & \mathbf{I}\end{pmatrix}^{-1} = \begin{pmatrix}\mathbf{I} & \mathbf{0}\\ -\mathbf{U} & \mathbf{I}\end{pmatrix}, \tag{13.11}$$

and thus the proof is completed. □

It is obvious that we can also write

$$\begin{aligned}\mathbf{A} &= \begin{pmatrix}\mathbf{I} & \mathbf{A}_{12}\mathbf{A}_{22}^{=}\\ \mathbf{0} & \mathbf{I}\end{pmatrix}\begin{pmatrix}\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-}\mathbf{A}_{21} & \mathbf{0}\\ \mathbf{0} & \mathbf{A}_{22}\end{pmatrix}\begin{pmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{A}_{22}^{\sim}\mathbf{A}_{21} & \mathbf{I}\end{pmatrix}\\ &:= \mathbf{EFG}\,, \end{aligned} \tag{13.12}$$

where $\mathbf{A}_{22}^{=}$, $\mathbf{A}_{22}^{-}$, and $\mathbf{A}_{22}^{\sim}$ are arbitrary generalized inverses of $\mathbf{A}_{22}$.

Moreover, consider $\mathbf{A}$ which is partitioned as (13.1) but we only know that

$$\mathscr{C}(\mathbf{A}_{12}) \subset \mathscr{C}(\mathbf{A}_{11}) \quad \text{and} \quad \mathscr{C}(\mathbf{A}'_{21}) \subset \mathscr{C}(\mathbf{A}'_{11}), \tag{13.13}$$

so that $\mathbf{A}$ is not necessarily symmetric nor nonnegative definite and $\mathbf{A}_{11}$ need not be a square matrix. Even then, proceeding as in the proof of Theorem 13, we see that the decomposition

$$\begin{aligned}\mathbf{A} &= \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{A}_{21}\mathbf{A}_{11}^{=} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-}\mathbf{A}_{12} \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{A}_{11}^{\sim}\mathbf{A}_{12} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{A}_{21}\mathbf{A}_{11}^{=} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22\cdot 1} \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{A}_{11}^{\sim}\mathbf{A}_{12} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}\end{aligned} \tag{13.14}$$

is valid. Similarly, if

$$\mathscr{C}(\mathbf{A}_{21}) \subset \mathscr{C}(\mathbf{A}_{22}) \quad \text{and} \quad \mathscr{C}(\mathbf{A}'_{12}) \subset \mathscr{C}(\mathbf{A}'_{22}), \tag{13.15}$$

then we have

$$\begin{aligned}\mathbf{A} &= \begin{pmatrix} \mathbf{I} & \mathbf{A}_{12}\mathbf{A}_{22}^{=} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-}\mathbf{A}_{21} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{A}_{22}^{\sim}\mathbf{A}_{21} & \mathbf{I} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{I} & \mathbf{A}_{12}\mathbf{A}_{22}^{=} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{A}_{11\cdot 2} & \\ \mathbf{0} & \mathbf{A}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{A}_{22}^{\sim}\mathbf{A}_{21} & \mathbf{I} \end{pmatrix}.\end{aligned} \tag{13.16}$$

The matrix $\mathbf{A}_{22\cdot 1}$ is the *Schur complement* of $\mathbf{A}_{11}$ in $\mathbf{A}$, denoted also as

$$\mathbf{A}_{22\cdot 1} = \mathbf{A}/\mathbf{A}_{11}\,. \tag{13.17}$$

For a comprehensive review in the literature, see, e.g., Ouellette (1981), and Styan (1985). Other surveys of the Schur complement and its applications include Carlson (1986), Cottle (1974), Henderson & Searle (1981a), and the monograph edited by Zhang (2005b) including chapters by Puntanen & Styan (2005a,b). The term "Schur complement" was introduced by Haynsworth (1968a,b) following work by Schur[1] (1917), who showed that (for $\mathbf{A}_{11}$ being a nonsingular square matrix)

$$|\mathbf{A}| = |\mathbf{A}_{11}||\mathbf{A}/\mathbf{A}_{11}| = |\mathbf{A}_{11}||\mathbf{A}_{22\cdot 1}| = |\mathbf{A}_{11}||\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}|\,, \tag{13.18}$$

---

[1] Issai Schur (1875–1941) published under "I. Schur" and under "J. Schur". He was a superb lecturer and his lectures were meticulously prepared and were exceedingly popular. Walter Ledermann (1983) remembers attending Schur's algebra course which was held in a lecture theatre filled with about 400 students: "Sometimes, when I had to be content with a seat at the back of the lecture theatre, I used a pair of opera glasses to get a glimpse of the speaker."

and so the determinant is multiplicative on the Schur complement. In view of (13.18), Haynsworth (1968a,b) introduced notation (13.17). Because for the block-diagonal matrix we have

$$\begin{pmatrix} \mathbf{F}_{f\times f} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_{g\times g} \end{pmatrix} = \begin{pmatrix} \mathbf{F} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_g \end{pmatrix} \begin{pmatrix} \mathbf{I}_f & \mathbf{0} \\ \mathbf{0} & \mathbf{G} \end{pmatrix} \implies \begin{vmatrix} \mathbf{F} & \mathbf{0} \\ \mathbf{0} & \mathbf{G} \end{vmatrix} = |\mathbf{F}||\mathbf{G}|\,, \tag{13.19}$$

we see from (13.14) (p. 293) that

$$|\mathbf{A}| = |\mathbf{A}_{11}||\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-}\mathbf{A}_{12}|\,, \tag{13.20}$$

holds when $\mathscr{C}(\mathbf{A}_{12}) \subset \mathscr{C}(\mathbf{A}_{11})$ and $\mathscr{C}(\mathbf{A}_{21}') \subset \mathscr{C}(\mathbf{A}_{11}')$; these column space inclusions trivially hold, for example, if $\mathbf{A}$ is nonnegative definite or if $\mathbf{A}_{11}$ is invertible.

13.1 "Inverting" a Partitioned nnd Matrix—Schur Complement

An immediate consequence of Theorem 13 is the standard formula for the inverse of a positive definite matrix. Since it is of the same form as the corresponding Moore–Penrose inverse, we state here the following result whose proof is left to the reader as an exercise; see Marsaglia & Styan (1974b).

Proposition 13.1. *Let* $\mathbf{A}$ *be a symmetric nonnegative definite matrix. Then there exists a matrix* $\mathbf{L}$ *such that* $\mathbf{A} = \mathbf{L}'\mathbf{L}$, *where* $\mathbf{L} = (\mathbf{L}_1 : \mathbf{L}_2)$ *and hence*

$$\mathbf{A} = \mathbf{L}'\mathbf{L} = \begin{pmatrix} \mathbf{L}_1'\mathbf{L}_1 & \mathbf{L}_1'\mathbf{L}_2 \\ \mathbf{L}_2'\mathbf{L}_1 & \mathbf{L}_2'\mathbf{L}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}. \tag{13.21}$$

If any of the following three statements hold, then all three hold:

(a) $\operatorname{rank}(\mathbf{A}) = \operatorname{rank}(\mathbf{A}_{11}) + \operatorname{rank}(\mathbf{A}_{22})$, *i.e.,* $\mathscr{C}(\mathbf{L}_1) \cap \mathscr{C}(\mathbf{L}_2) = \{\mathbf{0}\}$,

(b) $\mathbf{A}^{+} = \begin{pmatrix} \mathbf{A}_{11\cdot 2}^{+} & -\mathbf{A}_{11\cdot 2}^{+}\mathbf{A}_{12}\mathbf{A}_{22}^{+} \\ -\mathbf{A}_{22}^{+}\mathbf{A}_{21}\mathbf{A}_{11\cdot 2}^{+} & \mathbf{A}_{22}^{+} + \mathbf{A}_{22}^{+}\mathbf{A}_{21}\mathbf{A}_{11\cdot 2}^{+}\mathbf{A}_{12}\mathbf{A}_{22}^{+} \end{pmatrix}$,

(c) $\mathbf{A}^{+} = \begin{pmatrix} \mathbf{A}_{11}^{+} + \mathbf{A}_{11}^{+}\mathbf{A}_{12}\mathbf{A}_{22\cdot 1}^{+}\mathbf{A}_{21}\mathbf{A}_{11}^{+} & -\mathbf{A}_{11}^{+}\mathbf{A}_{12}\mathbf{A}_{22\cdot 1}^{+} \\ -\mathbf{A}_{22\cdot 1}^{+}\mathbf{A}_{21}\mathbf{A}_{11}^{+} & \mathbf{A}_{22\cdot 1}^{+} \end{pmatrix}$,

where

$$\mathbf{A}_{22\cdot 1} = \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-}\mathbf{A}_{12} = \mathbf{L}_2'(\mathbf{I} - \mathbf{P}_{\mathbf{L}_1})\mathbf{L}_2\,, \tag{13.22a}$$

$$\mathbf{A}_{11\cdot 2} = \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-}\mathbf{A}_{21} = \mathbf{L}_1'(\mathbf{I} - \mathbf{P}_{\mathbf{L}_2})\mathbf{L}_1\,. \tag{13.22b}$$

We may draw the reader's attention to the fact that the condition $\mathscr{C}(\mathbf{L}_1) \cap \mathscr{C}(\mathbf{L}_2) = \{\mathbf{0}\}$ must necessarily hold if we want to find the Moore–Penrose inverse of a partitioned nonnegative definite matrix using the formulas (b) and (c). However, the following matrix $\mathbf{A}^{\#}$ is *always* (why?) a generalized inverse of $\mathbf{A}$:

$$\mathbf{A}^{\#} = \begin{pmatrix} \mathbf{A}_{11\cdot 2}^{\sim} & -\mathbf{A}_{11\cdot 2}^{\sim}\mathbf{A}_{12}\mathbf{A}_{22}^{\sim} \\ -\mathbf{A}_{22}^{\sim}\mathbf{A}_{21}\mathbf{A}_{11\cdot 2}^{\sim} & \mathbf{A}_{22}^{\sim} + \mathbf{A}_{22}^{\sim}\mathbf{A}_{21}\mathbf{A}_{11\cdot 2}^{\sim}\mathbf{A}_{12}\mathbf{A}_{22}^{\sim} \end{pmatrix}, \tag{13.23}$$

where $\mathbf{A}_{11\cdot 2} = \mathbf{L}_1'\mathbf{Q}_2\mathbf{L}_1$, with $\mathbf{B}^{\sim}$ denoting a generalized inverse of $\mathbf{B}$ and $\mathbf{Q}_2 = \mathbf{I} - \mathbf{P}_{\mathbf{L}_2} = \mathbf{I} - \mathbf{L}_2(\mathbf{L}_2'\mathbf{L}_2)^{-}\mathbf{L}_2'$. In particular, the matrix $\mathbf{A}^{\#}$ is a symmetric reflexive generalized inverse of $\mathbf{A}$ for any choices of symmetric reflexive generalized inverses $(\mathbf{L}_2'\mathbf{L}_2)^{\sim}$ and $(\mathbf{L}_1'\mathbf{Q}_2\mathbf{L}_1)^{\sim}$. We say that $\mathbf{A}^{\#}$, defined in (13.23), is in Banachiewicz–Schur form. For further references on the Banachiewicz–Schur form, see, e.g., Banachiewicz (1937a,b), Baksalary, Puntanen & Yanai (1992), Baksalary & Styan (2002), Grala, Markiewicz & Styan (2000), and Tian & Takane (2009a).

13.2 Inverting $\mathbf{X}'\mathbf{X}$

As an example, let us consider a linear model $\{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}\}$, where the $n \times (k+1)$ model matrix $\mathbf{X}$, having full column rank, is partitioned as

$$\mathbf{X} = (\mathbf{1} : \mathbf{x}_1 : \ldots : \mathbf{x}_k) = (\mathbf{1} : \mathbf{X}_0)\,, \qquad \mathbf{X} = (\mathbf{X}_1 : \mathbf{X}_2)\,. \tag{13.24}$$

Here $\mathbf{X}_0$ is an $n \times k$ matrix, $\mathbf{X}_1$ is $n \times p_1$ and $\mathbf{X}_1$ is $n \times p_2$; $p_1 + p_2 = p = k+1$. Now we have

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_2 & \mathbf{X}_2'\mathbf{X}_2 \end{pmatrix}, \qquad (\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} \cdot & \cdot \\ \cdot & (\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1} \end{pmatrix}, \tag{13.25}$$

where $\mathbf{M}_1 = \mathbf{I}_n - \mathbf{P}_{\mathbf{X}_1}$. If

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}'\mathbf{X}_0 \\ \mathbf{X}_0'\mathbf{1} & \mathbf{X}_0'\mathbf{X}_0 \end{pmatrix}, \tag{13.26}$$

then, with $\mathbf{C}$ denoting the centering matrix,

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} \cdot & \cdot \\ \cdot & (\mathbf{X}_0'\mathbf{C}\mathbf{X}_0)^{-1} \end{pmatrix} = \begin{pmatrix} \cdot & \cdot \\ \cdot & \mathbf{T}_{\mathbf{xx}}^{-1} \end{pmatrix}. \tag{13.27}$$

Denoting $\mathbf{T}_{\mathbf{xx}}^{-1} = \{t^{ij}\}$, we obtain the last diagonal element of $\mathbf{T}_{\mathbf{xx}}^{-1}$, i.e., the last diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$:

$$t^{kk} = (\mathbf{x}_k'\mathbf{M}_1\mathbf{x}_k)^{-1} = \frac{1}{\mathbf{x}_k'(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_1})\mathbf{x}_k} = \frac{1}{\|(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_1})\mathbf{x}_k\|^2}\,, \tag{13.28}$$

where $\mathbf{P}_{\mathbf{X}_1}$ is the orthogonal projector onto $\mathscr{C}(\mathbf{1} : \mathbf{x}_1 : \ldots : \mathbf{x}_{k-1}) = \mathscr{C}(\mathbf{X}_1)$. Therefore,

$$t^{kk} \text{ is huge } \iff \mathbf{x}_k \text{ belongs almost to } \mathscr{C}(\mathbf{X}_1)\,. \tag{13.29}$$

Under the model $\{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}\}$, the variance of $\hat{\beta}_k$ is $\operatorname{var}(\hat{\beta}_k) = \sigma^2 t^{kk}$, and hence it is not good news for the estimation of β_k if the last column of $\mathbf{X}$ is almost a linear combination of the other columns; cf. (8.52) (p. 162),

Denoting

$$\mathbf{T} = (\mathbf{X}_0 : \mathbf{y})'\mathbf{C}(\mathbf{X}_0 : \mathbf{y}) = \begin{pmatrix} \mathbf{X}_0'\mathbf{C}\mathbf{X}_0 & \mathbf{X}_0'\mathbf{C}\mathbf{y} \\ \mathbf{y}'\mathbf{C}\mathbf{X}_0 & \mathbf{y}'\mathbf{C}\mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{T}_{\mathbf{xx}} & \mathbf{t}_{\mathbf{x}y} \\ \mathbf{t}_{\mathbf{x}y}' & t_{yy} \end{pmatrix}, \tag{13.30}$$

we get

$$\mathbf{T}^{-1} := \begin{pmatrix} \cdot & \cdot \\ \cdot & t^{yy} \end{pmatrix} = \begin{pmatrix} \cdot & \cdot \\ \cdot & (t_{yy} - \mathbf{t}_{\mathbf{x}y}'\mathbf{T}_{\mathbf{xx}}^{-1}\mathbf{t}_{\mathbf{x}y})^{-1} \end{pmatrix}, \tag{13.31}$$

where

$$t^{yy} = \frac{1}{t_{yy\cdot\mathbf{x}}} = \frac{1}{t_{yy} - \mathbf{t}_{\mathbf{x}y}'\mathbf{T}_{\mathbf{xx}}^{-1}\mathbf{t}_{\mathbf{x}y}} = \frac{1}{\mathrm{SSE}}\,. \tag{13.32}$$

If $\mathbf{R}$ is a sample correlation matrix of variables $x_1, \ldots, x_k$, y,

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_{\mathbf{xx}} & \mathbf{r}_{\mathbf{x}y} \\ \mathbf{r}_{\mathbf{x}y}' & 1 \end{pmatrix}, \tag{13.33}$$

then the last diagonal element of $\mathbf{R}^{-1}$ is

$$r^{yy} = \frac{1}{1 - \mathbf{r}_{\mathbf{x}y}'\mathbf{R}_{\mathbf{xx}}^{-1}\mathbf{r}_{\mathbf{x}y}} = \frac{1}{1 - R^2} = \frac{1}{1 - R_{y\cdot\mathbf{x}}^2}\,, \tag{13.34}$$

where R^2 is the multiple correlation squared when y is explained by the variables $x_1, \ldots, x_k$ (and constant). Moreover, we of course have

$$R^2 = 1 - \frac{1}{r^{yy}}\,. \tag{13.35}$$

As stated on page 172, the ith diagonal element of $\mathbf{R}_{\mathbf{xx}}^{-1}$ can be expressed as

$$r^{ii} = \frac{1}{1 - R_{i\cdot 1\ldots i-1,i+1,\ldots k}^2} = \mathrm{VIF}_i = i\text{th variance inflation factor}\,, \tag{13.36}$$

while

$$R_{i\cdot 1\ldots i-1,i+1,\ldots k}^2 = 1 - \frac{1}{r^{ii}}\,. \tag{13.37}$$

13.3 Determinants, Ranks and Schur Complements

Let $\mathbf{A}$ be a symmetric nonnegative definite matrix partitioned as in (13.1). Then from

$$\begin{aligned}\mathbf{A} &= \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{A}_{21}\mathbf{A}_{11}^{-} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-}\mathbf{A}_{12} \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{A}_{11}^{-}\mathbf{A}_{12} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{I} & \mathbf{A}_{12}\mathbf{A}_{22}^{-} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-}\mathbf{A}_{21} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{A}_{22}^{-}\mathbf{A}_{21} & \mathbf{I} \end{pmatrix} \qquad (13.38)\end{aligned}$$

we observe immediately the formulas for the determinant:

$$\begin{aligned}|\mathbf{A}| &= |\mathbf{A}_{11}||\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-}\mathbf{A}_{12}| & (13.39\text{a}) \\ &= |\mathbf{A}_{22}||\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-}\mathbf{A}_{21}|\,. & (13.39\text{b})\end{aligned}$$

Furthermore, from (13.39) we get the inequality

$$|\mathbf{A}| \leq |\mathbf{A}_{11}| \cdot |\mathbf{A}_{22}|\,. \qquad (13.40)$$

In particular, if the random vector $\mathbf{z} = \binom{\mathbf{x}}{y}$, where $\mathbf{x}$ is a p-dimensional random vector, has the covariance matrix

$$\operatorname{cov}(\mathbf{z}) = \operatorname{cov}\begin{pmatrix} \mathbf{x} \\ y \end{pmatrix} = \begin{pmatrix} \mathbf{\Sigma}_{\mathbf{xx}} & \boldsymbol{\sigma}_{\mathbf{x}y} \\ \boldsymbol{\sigma}'_{\mathbf{x}y} & \sigma_y^2 \end{pmatrix} = \mathbf{\Sigma}\,, \qquad (13.41)$$

then we get immediately the formulas for the determinant:

$$\begin{aligned}|\mathbf{\Sigma}| &= |\mathbf{\Sigma}_{\mathbf{xx}}|(\sigma_y^2 - \boldsymbol{\sigma}'_{\mathbf{x}y}\mathbf{\Sigma}_{\mathbf{xx}}^{-}\boldsymbol{\sigma}_{\mathbf{x}y}) \\ &= |\mathbf{\Sigma}_{\mathbf{xx}}|\sigma_{y\cdot\mathbf{x}}^2 \\ &= |\mathbf{\Sigma}_{\mathbf{xx}}|\sigma_y^2\left(1 - \frac{\boldsymbol{\sigma}'_{\mathbf{x}y}\mathbf{\Sigma}_{\mathbf{xx}}^{-}\boldsymbol{\sigma}_{\mathbf{x}y}}{\sigma_y^2}\right) \\ &= |\mathbf{\Sigma}_{\mathbf{xx}}|\sigma_y^2(1 - \varrho_{y\cdot\mathbf{x}}^2)\,, \qquad (13.42)\end{aligned}$$

where we assume that $\sigma_y^2 > 0$. If $\sigma_y^2 = 0$, then of course $\boldsymbol{\sigma}_{\mathbf{x}y} = \mathbf{0}$ and $|\mathbf{\Sigma}| = 0$. Similarly, denoting $\operatorname{var}(x_p) = \sigma_p^2$, and $\mathbf{x}_{(p-1)} = (x_1, \dots, x_{p-1})'$, we get

$$\begin{aligned}|\mathbf{\Sigma}_{\mathbf{xx}}| &= |\operatorname{cov}(\mathbf{x}_{(p-1)})|\big[\sigma_p^2 - [\operatorname{cov}(\mathbf{x}_{(p-1)}, x_p)]'[\operatorname{cov}(\mathbf{x}_{(p-1)})]^{-}\operatorname{cov}(\mathbf{x}_{(p-1)}, x_p)\big] \\ &= |\operatorname{cov}(\mathbf{x}_{(p-1)})|\sigma_{p\cdot\mathbf{x}_{(p-1)}}^2 \\ &= |\operatorname{cov}(\mathbf{x}_{(p-1)})|\sigma_p^2(1 - \varrho_{p\cdot\mathbf{x}_{(p-1)}}^2)\,. \qquad (13.43)\end{aligned}$$

Hence we can conclude that the following holds; see also page 208:

$$|\mathbf{\Sigma}| = |\operatorname{cov}(\mathbf{x}_{(p-1)})|\sigma_p^2\sigma_y^2(1 - \varrho_{p\cdot\mathbf{x}_{(p-1)}}^2)(1 - \varrho_{y\cdot\mathbf{x}}^2)\,, \qquad (13.44)$$

and

$$\begin{aligned}|\mathbf{\Sigma}| &= \sigma_1^2\sigma_2^2\cdots\sigma_p^2\sigma_y^2(1-\varrho_{2\cdot 1}^2)(1-\varrho_{3\cdot 12}^2)\cdots(1-\varrho_{p\cdot 12\ldots p-1}^2)(1-\varrho_{y\cdot\mathbf{x}}^2)\\ &= \sigma_1^2\sigma_{2\cdot 1}^2\cdots\sigma_{p\cdot 1\ldots p-1}^2\sigma_{y\cdot 1\ldots p}^2\,,\end{aligned} \tag{13.45}$$

where we can define $\varrho_{i\cdot 12\ldots i-1}^2 = 1$ if $\sigma_i^2 = 0$; see (9.20) (p. 193). It is obvious that

$$|\mathbf{\Sigma}| \leq \sigma_1^2\sigma_2^2\cdots\sigma_p^2\sigma_y^2\,, \tag{13.46}$$

where the equality holds if and only if $\mathbf{\Sigma}$ is a diagonal matrix or contains a zero column (row). The result (13.46) is often called the Hadamard's determinantal inequality; it can be written also as

$$|\mathbf{A}|^2 \leq (\mathbf{a}_1'\mathbf{a}_1)^2(\mathbf{a}_2'\mathbf{a}_2)^2\cdots(\mathbf{a}_p'\mathbf{a}_p)^2, \tag{13.47}$$

where $\mathbf{A} = (\mathbf{a}_1 : \mathbf{a}_2 : \ldots : \mathbf{a}_p) \in \mathbb{R}^{p\times p}$.

If $\boldsymbol{\rho}$ is the correlation matrix based on the covariance matrix $\mathbf{\Sigma}$, then $|\boldsymbol{\rho}| = |\mathbf{\Sigma}|/|\mathrm{diag}(\mathbf{\Sigma})|$, and so

$$|\boldsymbol{\rho}| = (1-\varrho_{2\cdot 1}^2)(1-\varrho_{3\cdot 12}^2)\cdots(1-\varrho_{p\cdot 12\ldots p-1}^2)(1-\varrho_{y\cdot\mathbf{x}}^2) \leq 1\,, \tag{13.48}$$

Moreover, if $\mathbf{X} = (\mathbf{1} : \mathbf{X}_0) \in \mathbb{R}^{n\times(k+1)}$ and

$$\mathbf{T} = (\mathbf{X}_0 : \mathbf{y})'\mathbf{C}(\mathbf{X}_0 : \mathbf{y}) = \begin{pmatrix}\mathbf{X}_0'\mathbf{C}\mathbf{X}_0 & \mathbf{X}_0'\mathbf{C}\mathbf{y}\\ \mathbf{y}'\mathbf{C}\mathbf{X}_0 & \mathbf{y}'\mathbf{C}\mathbf{y}\end{pmatrix} = \begin{pmatrix}\mathbf{T}_{\mathbf{xx}} & \mathbf{t}_{\mathbf{x}y}\\ \mathbf{t}_{\mathbf{x}y}' & t_{yy}\end{pmatrix}, \tag{13.49}$$

$$\mathbf{A} = (\mathbf{X} : \mathbf{y})'(\mathbf{X} : \mathbf{y}) = \begin{pmatrix}\mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{y}\\ \mathbf{y}'\mathbf{X} & \mathbf{y}'\mathbf{y}\end{pmatrix}, \tag{13.50}$$

then

$$|\mathbf{T}| = |\mathbf{T}_{\mathbf{xx}}|(t_{yy} - \mathbf{t}_{\mathbf{x}y}'\mathbf{T}_{\mathbf{xx}}^-\mathbf{t}_{\mathbf{x}y}) = |\mathbf{T}_{\mathbf{xx}}|\cdot\mathrm{SSE}\,, \tag{13.51a}$$

$$|\mathbf{A}| = |\mathbf{X}'\mathbf{X}|(\mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{H}\mathbf{y}) = |\mathbf{X}'\mathbf{X}|\cdot\mathrm{SSE}\,. \tag{13.51b}$$

Let us then consider the eigenvalues of the matrix products $\mathbf{UV}$ and $\mathbf{VU}$ (for conformable $\mathbf{U}$ and $\mathbf{V}$). This can be done conveniently using the Schur complement. First, it is easy to confirm that

$$\det\begin{pmatrix}\mathbf{I}_u & \mathbf{U}_{u\times v}\\ \mathbf{V}_{v\times u} & \mathbf{I}_v\end{pmatrix} = |\mathbf{I}_u|\cdot|\mathbf{I}_v - \mathbf{VU}| = |\mathbf{I}_v|\cdot|\mathbf{I}_u - \mathbf{UV}|\,, \tag{13.52}$$

and hence

$$|\mathbf{I}_v - \mathbf{VU}| = |\mathbf{I}_u - \mathbf{UV}|\,. \tag{13.53}$$

Suppose that $\lambda \neq 0$. Then

$$\begin{aligned}|\lambda\mathbf{I}_v - \mathbf{VU}| &= |\lambda(\mathbf{I}_v - \mathbf{VU}\lambda^{-1})| = \lambda^v|\mathbf{I}_v - \mathbf{VU}\lambda^{-1}|\\ &= \lambda^v|\mathbf{I}_u - \mathbf{UV}\lambda^{-1}| = \lambda^v|\lambda^{-1}(\lambda\mathbf{I}_u - \mathbf{UV})|\end{aligned}$$

$$= \lambda^{v-u} |\lambda \mathbf{I}_u - \mathbf{UV}| \,, \tag{13.54}$$

which implies the following very important result:

Proposition 13.2. *The matrices* $\mathbf{UV}$ *and* $\mathbf{VU}$ *have the same set of nonzero eigenvalues, including multiplicities, i.e.,*

$$\operatorname{nzch}(\mathbf{UV}) = \operatorname{nzch}(\mathbf{VU}) \,. \tag{13.55}$$

An alternative proof for Proposition 13.2 can be obtained by noting first that

$$\begin{pmatrix} \mathbf{I}_u & \mathbf{U} \\ \mathbf{0} & \mathbf{I}_v \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{UV} & \mathbf{0} \\ \mathbf{V} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{I}_u & \mathbf{U} \\ \mathbf{0} & \mathbf{I}_v \end{pmatrix} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{V} & \mathbf{VU} \end{pmatrix} . \tag{13.56}$$

Thus the matrices

$$\mathbf{A} = \begin{pmatrix} \mathbf{UV} & \mathbf{0} \\ \mathbf{V} & \mathbf{0} \end{pmatrix}, \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{V} & \mathbf{VU} \end{pmatrix} \tag{13.57}$$

are similar and hence they have the same eigenvalues; see (18.22) (p. 360). Because $\mathbf{A}$ and $\mathbf{B}$ are block-triangular matrices, their eigenvalues are the eigenvalues of the diagonal blocks $\mathbf{UV}$, $\mathbf{0}_{u\times u}$, and $\mathbf{VU}$, $\mathbf{0}_{v\times v}$, respectively. From this (13.55) follows.

In this context we may briefly consider the concept of the *inertia*, denoted as $\operatorname{In}(\mathbf{A})$. The inertia of a symmetric $n \times n$ matrix $\mathbf{A}$ is defined as the ordered triple

$$\operatorname{In}(\mathbf{A}) = \{i_+(\mathbf{A}),\, i_-(\mathbf{A}),\, i_0(\mathbf{A})\} \,, \tag{13.58}$$

where $i_+(\mathbf{A})$, $i_-(\mathbf{A})$, and $i_0(\mathbf{A})$ are the number of positive, negative, and zero eigenvalues of $\mathbf{A}$, respectively, all counting multiplicities. Then

$$\operatorname{rank}(\mathbf{A}) = i_+(\mathbf{A}) + i_-(\mathbf{A}) \,, \qquad n = i_+(\mathbf{A}) + i_-(\mathbf{A}) + i_0(\mathbf{A}) \,. \tag{13.59}$$

From the block-diagonalization theorem we then conclude that the inertia of a (symmetric) partitioned nonnegative definite $\mathbf{A}$ has the property

$$\operatorname{In}(\mathbf{A}) = \operatorname{In}(\mathbf{A}_{11}) + \operatorname{In}(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-}\mathbf{A}_{12}) \tag{13.60a}$$

$$= \operatorname{In}(\mathbf{A}_{22}) + \operatorname{In}(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-}\mathbf{A}_{21}) \,. \tag{13.60b}$$

Consequently, we can conclude that

$$\operatorname{rank}(\mathbf{A}) = \operatorname{rank}(\mathbf{A}_{11}) + \operatorname{rank}(\mathbf{A}_{22\cdot 1}) \,, \tag{13.61}$$

as shown by Guttman (1946), and so the rank is *additive* on the Schur complement. Actually, see (13.14) (p. 293), (13.61) holds if

$$\mathscr{C}(\mathbf{A}_{12}) \subset \mathscr{C}(\mathbf{A}_{11}) \quad \text{and} \quad \mathscr{C}(\mathbf{A}_{21}') \subset \mathscr{C}(\mathbf{A}_{11}') \,, \tag{13.62}$$

which conditions of course are satisfied e.g. if $\mathbf{A}$ is (symmetric) nonnegative definite or if $\mathbf{A}_{11}^{-1}$ exists. Similarly $\mathscr{C}(\mathbf{A}_{21}) \subset \mathscr{C}(\mathbf{A}_{22})$ and $\mathscr{C}(\mathbf{A}'_{12}) \subset \mathscr{C}(\mathbf{A}'_{22})$ imply

$$\operatorname{rank}(\mathbf{A}) = \operatorname{rank}(\mathbf{A}_{22}) + \operatorname{rank}(\mathbf{A}_{11\cdot 2}) . \tag{13.63}$$

For further details on the rank additivity on the Schur complement, see Marsaglia & Styan (1974a, Th. 19, Cor. 19.1).

For interesting remarks on the determinants, and in particular, their historical development, see Tee (2003).

13.4 Inverting a Sum of Matrices

Consider now a nonsingular (not necessarily positive definite) matrix $\mathbf{A}_{n\times n}$ partitioned as

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} , \tag{13.64}$$

where $\mathbf{A}_{11}$ is a square matrix. Then, according to (a) and (b) of Section 13.1 (p. 294), we can express the inverse $\mathbf{A}^{-1}$ in the following two ways:

$$\begin{aligned} \mathbf{A}^{-1} &= \begin{pmatrix} \mathbf{A}_{11\cdot 2}^{-1} & -\mathbf{A}_{11\cdot 2}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}_{11\cdot 2}^{-1} & \mathbf{A}_{22}^{-1} + \mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}_{11\cdot 2}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{A}_{11}^{-1} + \mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}_{22\cdot 1}^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}_{22\cdot 1}^{-1} \\ -\mathbf{A}_{22\cdot 1}^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & \mathbf{A}_{22\cdot 1}^{-1} \end{pmatrix} \\ &:= \begin{pmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{A}^{22} \end{pmatrix} , \end{aligned} \tag{13.65}$$

where

$$\mathbf{A}_{22\cdot 1} = \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12} = \mathbf{A}/\mathbf{A}_{11} , \tag{13.66a}$$

$$\mathbf{A}_{11\cdot 2} = \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} = \mathbf{A}/\mathbf{A}_{22} . \tag{13.66b}$$

From (13.65), we immediately observe the following matrix equality:

$$\mathbf{A}_{11\cdot 2}^{-1} = \mathbf{A}_{11}^{-1} + \mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}_{22\cdot 1}^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} , \tag{13.67}$$

that is,

$$\begin{aligned} &(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1} \\ &\qquad = \mathbf{A}_{11}^{-1} + \mathbf{A}_{11}^{-1}\mathbf{A}_{12}(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} . \end{aligned} \tag{13.68}$$

If $\mathbf{A}$ is symmetric and denoted as

$$\mathbf{A} = \begin{pmatrix} \mathbf{E} & \mathbf{F} \\ \mathbf{F}' & \mathbf{G} \end{pmatrix}, \tag{13.69}$$

then (13.68) becomes

$$(\mathbf{E} - \mathbf{F}\mathbf{G}^{-1}\mathbf{F}')^{-1} = \mathbf{E}^{-1} + \mathbf{E}^{-1}\mathbf{F}(\mathbf{G} - \mathbf{F}'\mathbf{E}^{-1}\mathbf{F})^{-1}\mathbf{F}'\mathbf{E}^{-1}. \tag{13.70}$$

Denoting $\mathbf{B} = -\mathbf{G}^{-1}$, we can express (13.70) as

$$\begin{aligned}(\mathbf{E} + \mathbf{F}\mathbf{B}\mathbf{F}')^{-1} &= \mathbf{E}^{-1} - \mathbf{E}^{-1}\mathbf{F}(\mathbf{B}^{-1} + \mathbf{F}'\mathbf{E}^{-1}\mathbf{F})^{-1}\mathbf{F}'\mathbf{E}^{-1} \\ &= \mathbf{E}^{-1} - \mathbf{E}^{-1}\mathbf{F}\mathbf{B}(\mathbf{B} + \mathbf{B}\mathbf{F}'\mathbf{E}^{-1}\mathbf{F}\mathbf{B})^{-1}\mathbf{B}\mathbf{F}'\mathbf{E}^{-1}. \end{aligned} \tag{13.71}$$

As illustrated by Henderson & Searle (1981a), and Ouellette (1981), formula (13.71) has many applications in statistics. Formula (13.71) is considered to be due to Duncan (1944) and Woodbury (1950), and is often referred to as Duncan formula, or Woodbury formula. The case

$$(\mathbf{E} + \alpha\mathbf{f}\mathbf{f}')^{-1} = \mathbf{E}^{-1} - \frac{\alpha\mathbf{E}^{-1}\mathbf{f}\mathbf{f}'\mathbf{E}^{-1}}{1 + \alpha\mathbf{f}'\mathbf{E}^{-1}\mathbf{f}}, \quad \text{where } \alpha \in \mathbb{R}, \tag{13.72}$$

is often attributed to Sherman & Morrison (1949, 1950). For a survey on updating the inverse of a matrix, see Hager (1989). The motivation for diagonal increments to ease matrix inversion in least squares problems is discussed in Piegorsch & Casella (1989).

Partitioning the model matrix row-wise as

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_{(1)} \\ \mathbf{x}'_{(n)} \end{pmatrix}, \tag{13.73}$$

we get one application of (13.72):

$$\begin{aligned}(\mathbf{X}'_{(1)}\mathbf{X}_{(1)})^{-1} &= (\mathbf{X}'\mathbf{X} - \mathbf{x}_{(n)}\mathbf{x}'_{(n)})^{-1} = [\mathbf{X}'(\mathbf{I}_n - \mathbf{i}_n\mathbf{i}'_n)\mathbf{X}]^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1} + \frac{1}{1 - h_{nn}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_{(n)}\mathbf{x}'_{(n)}(\mathbf{X}'\mathbf{X})^{-1}, \end{aligned} \tag{13.74}$$

where h_{nn} is the last diagonal element of the hat matrix $\mathbf{H} = \mathbf{P}_{\mathbf{X}}$.

13.5 Generating Observations with a Given Covariance Matrix

Let $\mathbf{w} = \binom{w_1}{w_2}$ be a two-dimensional random vector with covariance matrix

$$\operatorname{cov}(\mathbf{w}) = \operatorname{cov}\begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} 1 & \varrho \\ \varrho & 1 \end{pmatrix}, \quad \varrho^2 < 1\,. \tag{13.75}$$

Then the block-diagonalization theorem gives us decompositions

$$\begin{pmatrix} 1 & 0 \\ -\varrho & 1 \end{pmatrix} \begin{pmatrix} 1 & \varrho \\ \varrho & 1 \end{pmatrix} \begin{pmatrix} 1 & -\varrho \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1-\varrho^2 \end{pmatrix}, \tag{13.76a}$$

$$\begin{pmatrix} 1 & 0 \\ \varrho & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1-\varrho^2 \end{pmatrix} \begin{pmatrix} 1 & \varrho \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & \varrho \\ \varrho & 1 \end{pmatrix}, \tag{13.76b}$$

i.e.,

$$\begin{pmatrix} 1 & 0 \\ \varrho & \sqrt{1-\varrho^2} \end{pmatrix} \begin{pmatrix} 1 & \varrho \\ 0 & \sqrt{1-\varrho^2} \end{pmatrix} := \mathbf{A}\mathbf{A}' = \begin{pmatrix} 1 & \varrho \\ \varrho & 1 \end{pmatrix}. \tag{13.77}$$

This implies that if

$$\operatorname{cov}(\mathbf{t}) = \operatorname{cov}\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{13.78}$$

then

$$\operatorname{cov}(\mathbf{A}\mathbf{t}) = \operatorname{cov}\begin{pmatrix} 1 & 0 \\ \varrho & \sqrt{1-\varrho^2} \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} 1 & \varrho \\ \varrho & 1 \end{pmatrix}. \tag{13.79}$$

Choosing $\mathbf{B} = \left(\begin{smallmatrix} \sigma_x & 0 \\ 0 & \sigma_y \end{smallmatrix}\right)$, we get a random vector $\mathbf{z} = \left(\begin{smallmatrix} x \\ y \end{smallmatrix}\right) = \mathbf{BAt}$ whose covariance matrix is

$$\operatorname{cov}(\mathbf{z}) = \operatorname{cov}\left[\begin{pmatrix} \sigma_x & 0 \\ \varrho\sigma_y & \sqrt{1-\varrho^2}\sigma_y \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix}\right] = \begin{pmatrix} \sigma_x^2 & \varrho\sigma_x\sigma_y \\ \varrho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix} = \mathbf{\Sigma}. \tag{13.80}$$

Transforming $\mathbf{z} = \mathbf{BAt}$, that is,

$$x = \sigma_x \cdot u\,, \tag{13.81a}$$

$$y = \sigma_y[\varrho \cdot u + \sqrt{1-\varrho^2} \cdot v]\,, \tag{13.81b}$$

is a very handy way to generate 2-dimensional observations having a given covariance matrix $\mathbf{\Sigma}$. We can first generate n observations from the random vector $\mathbf{t}$ whose covariance matrix is $\mathbf{I}_2$, put these n observations as the rows of the matrix $\mathbf{T} \in \mathbb{R}^{n\times 2}$, and then postmultiply $\mathbf{T}$ by the matrix $\mathbf{L}'$, where

$$\mathbf{L} = \mathbf{BA} = \begin{pmatrix} \sigma_x & 0 \\ \sigma_y\varrho & \sigma_y\sqrt{1-\varrho^2} \end{pmatrix}. \tag{13.82}$$

An alternative way to do this would, of course, be to postmultiply $\mathbf{T}$ by $\mathbf{\Sigma}^{1/2}$.

If we generate observations of $\mathbf{z}$ from $\mathrm{N}_2(\mathbf{0}, \mathbf{\Sigma})$ according to (13.81), then we can see a couple of interesting things at once. First, let's keep x fixed at level x_* in which case u is fixed as $u_* = x_*/\sigma_x$, and hence the conditional variance of y is

$$\operatorname{var}(y \mid x = x_*) = \operatorname{var}(\sigma_y\sqrt{1-\varrho^2} \cdot v) = \sigma_y^2(1-\varrho^2)\,. \tag{13.83}$$

Similarly we get

$$\begin{aligned}\mathrm{E}(y \mid x = x_*) &= \mathrm{E}\big(y \mid u = \frac{x_*}{\sigma_x}\big) \\ &= \sigma_y \varrho \cdot \frac{x_*}{\sigma_x} + \sigma_y \sqrt{1-\varrho^2} \cdot \mathrm{E}(v) = \frac{\sigma_{xy}}{\sigma_x^2} x_* \,. \end{aligned} \tag{13.84}$$

It is noteworthy that (13.83) and (13.84) are, of course, precisely the conditional expectation and variance of y under multinormality $\mathrm{N}_2(\mathbf{0}, \mathbf{\Sigma})$. The point here is that (13.81) offers an interesting way to introduce these representations.

13.6 Exercises

13.1. Prove Proposition 13.1 (p. 294).

13.2. Suppose that $\mathbf{A} = \big(\begin{smallmatrix}\mathbf{B} & \mathbf{C}\\ \mathbf{D} & \mathbf{F}\end{smallmatrix}\big)$, where $\mathbf{B}, \mathbf{C}, \mathbf{D}, \mathbf{F} \in \mathbb{R}^{n\times n}$ and $\mathbf{B}$ and $\mathbf{D}$ commute: $\mathbf{BD} = \mathbf{DB}$. Confirm that for invertible $\mathbf{B}$ the formula (13.18) (p. 293) implies that $\det(\mathbf{A}) = \det(\mathbf{BF} - \mathbf{DC})$. Actually this holds for even if $\mathbf{B}$ is not invertible; the proof can be done using a continuity argument.

Abadir & Magnus (2005, p. 116),
Puntanen & Styan (2005a, p. 4), Schur (1917, pp. 215–216).

13.3. Prove that for $\mathbf{B} \in \mathbb{R}^{b\times b}$ and $\mathbf{D} \in \mathbb{R}^{d\times d}$: $\begin{vmatrix}\mathbf{B} & \mathbf{C}\\ \mathbf{0} & \mathbf{D}\end{vmatrix} = |\mathbf{B}||\mathbf{D}|$.

Hint: For invertible $\mathbf{B}$, use (13.18) (p. 293). If $\operatorname{rank}(\mathbf{B}) < b$, then confirm that $\begin{vmatrix}\mathbf{B} & \mathbf{C}\\ \mathbf{0} & \mathbf{D}\end{vmatrix} = 0 = |\mathbf{B}|$.

13.4. Let $\mathbf{P} \in \mathbb{R}^{n\times n}$ orthogonal projector (symmetric idempotent) which is partitioned as

$$\mathbf{P} = \begin{pmatrix}\mathbf{P}_{11} & \mathbf{P}_{12}\\ \mathbf{P}_{21} & \mathbf{P}_{22}\end{pmatrix},$$

where $\mathbf{P}_{11}$ is a square matrix. Show that then the Schur complement $\mathbf{P}_{22\cdot 1} = \mathbf{P}_{22} - \mathbf{P}_{21}\mathbf{P}_{11}^{-}\mathbf{P}_{12}$ is also an orthogonal projector.

Baksalary, Baksalary & Szulc (2004).

13.5 (Wedderburn–Guttman theorem). Consider $\mathbf{A} \in \mathbb{R}^{n\times p}$, $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{y} \in \mathbb{R}^p$ and assume that $\alpha := \mathbf{x}'\mathbf{A}\mathbf{y} \neq 0$, and denote

$$\mathbf{B} = \begin{pmatrix}\mathbf{A} & \mathbf{A}\mathbf{y}\\ \mathbf{x}'\mathbf{A} & \mathbf{x}'\mathbf{A}\mathbf{y}\end{pmatrix}. \tag{13.85}$$

Using the rank-additivity on the Schur complement, see (13.61) (p. 299), prove the Wedderburn–Guttman theorem:

$$\operatorname{rank}(\mathbf{A} - \alpha^{-1}\mathbf{A}\mathbf{y}\mathbf{x}'\mathbf{A}) = \operatorname{rank}(\mathbf{A}) - 1 . \tag{13.86}$$

Citing Chu, Funderlic & Golub (1995, p. 512): “This Weddenburn rank-one reduction formula is easy to prove, yet the idea is so powerful that perhaps all matrix factorizations can be derived from it.”

Wedderburn (1934), Guttman (1944, 1946, 1952, 1957),
Hubert, Meulman & Heiser (2000), Takane & Yanai (2005, 2007).

13.6. Let the covariance matrix of the random vector $\mathbf{y} = \left(\begin{smallmatrix} y_1 \\ y_2 \end{smallmatrix}\right)$ be $\mathbf{I}_2$. Find such random variables $z_1 = ay_1 + by_2$ and $z_2 = cy_1 + dy_2$ that

$$\operatorname{cov}\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \operatorname{cov}(\mathbf{z}) = \operatorname{cov}(\mathbf{A}\mathbf{y}) = \begin{pmatrix} 1 & 2 \\ 2 & 9 \end{pmatrix} = \mathbf{\Sigma} .$$

Confirm that (13.81) (p. 302) gives a solution $\mathbf{A}_1 = \left(\begin{smallmatrix} 1 & 0 \\ 2 & \sqrt{5} \end{smallmatrix}\right)$, and that further solutions are obtained by choosing $\mathbf{A}_2 = \mathbf{T}\mathbf{\Lambda}^{1/2}$, and $\mathbf{A}_3 = \mathbf{T}\mathbf{\Lambda}^{1/2}\mathbf{T}' = \mathbf{\Sigma}^{1/2}$, where $\mathbf{\Sigma} = \mathbf{T}\mathbf{\Lambda}\mathbf{T}'$ is the eigenvalue decomposition of $\mathbf{\Sigma}$.

13.7. Let $\mathbf{A} \in \mathrm{NND}_n$ and $0 \neq \alpha \in \mathbb{R}$. Show that

$$(\mathbf{I}_n + \alpha\mathbf{A})^{-1} = \mathbf{A}^{+}(\alpha\mathbf{I}_n + \mathbf{A}^{+})^{-1} + \mathbf{I}_n - \mathbf{P}_{\mathbf{A}} .$$

Chapter 14
Nonnegative Definiteness of a Partitioned Matrix

> *. . . And don't believe anyone who says*
> *you can't get tomatoes here. Of course*
> *you can. And bags to put them in.*
> *It all adds up to a must-see experience.*
>
> Monty Python's Spamalot

The nonnegative definiteness of a symmetric partitioned matrix can be characterized in an interesting way in terms of submatrices. Our experience is that this characterization is a very handy tool in various problems related to the Löwner partial ordering.

Theorem 14 (Nonnegative definiteness of a partitioned matrix). *Let* $\mathbf{A}$ *be a symmetric matrix partitioned as*

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}, \tag{14.1}$$

where $\mathbf{A}_{11}$ *is a square matrix. Then the following three statements are equivalent:*

(a) $\mathbf{A} \geq_{\mathsf{L}} \mathbf{0}$,

(b$_1$) $\mathbf{A}_{11} \geq_{\mathsf{L}} \mathbf{0}$, (b$_2$) $\mathscr{C}(\mathbf{A}_{12}) \subset \mathscr{C}(\mathbf{A}_{11})$, (b$_3$) $\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-}\mathbf{A}_{12} \geq_{\mathsf{L}} \mathbf{0}$,

(c$_1$) $\mathbf{A}_{22} \geq_{\mathsf{L}} \mathbf{0}$, (c$_2$) $\mathscr{C}(\mathbf{A}_{21}) \subset \mathscr{C}(\mathbf{A}_{22})$, (c$_3$) $\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-}\mathbf{A}_{21} \geq_{\mathsf{L}} \mathbf{0}$.

Proof. Let $\mathbf{A} \in \mathbb{R}^{p\times p}$, $\mathbf{A}_{11} \in \mathbb{R}^{p_1\times p_1}$ and $\mathbf{A}_{22} \in \mathbb{R}^{p_2\times p_2}$, and assume first that (a) holds. Then there exists a matrix $\mathbf{L}_{q\times p}$ (for some q) such that $\mathbf{A} = \mathbf{L}'\mathbf{L}$, where $\mathbf{L} = (\mathbf{L}_1 : \mathbf{L}_2)$, $\mathbf{L}_i \in \mathbb{R}^{q\times p_i}$, and hence

$$\mathbf{A} = \mathbf{L}'\mathbf{L} = \begin{pmatrix} \mathbf{L}_1'\mathbf{L}_1 & \mathbf{L}_1'\mathbf{L}_2 \\ \mathbf{L}_2'\mathbf{L}_1 & \mathbf{L}_2'\mathbf{L}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}. \tag{14.2}$$

From (14.2) we observe immediately that (b$_1$) and (b$_2$) hold. Writing

$$\mathbf{A}_{22\cdot 1} = \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-}\mathbf{A}_{12} = \mathbf{L}_2'\mathbf{L}_2 - \mathbf{L}_2'\mathbf{L}_1(\mathbf{L}_1'\mathbf{L}_1)^{-}\mathbf{L}_1'\mathbf{L}_2 \tag{14.3}$$

we get

$$\mathbf{A}_{22\cdot 1} = \mathbf{L}_2'(\mathbf{I}_q - \mathbf{P}_{\mathbf{L}_1})\mathbf{L}_2 \geq_{\mathsf{L}} \mathbf{0}, \tag{14.4}$$

and hence (a) implies (b). Similarly we can prove that (a) implies (c). To go the other direction, note first that the column space inclusion (b$_2$) means

S. Puntanen et al., *Matrix Tricks for Linear Statistical Models: Our Personal Top Twenty*,
DOI 10.1007/978-3-642-10473-2_15,

that the Schur complement $\mathbf{A}_{22\cdot1}$ is invariant for any choice of a generalized inverse $\mathbf{A}_{11}^{-}$. Moreover, (b_2) implies that the following decomposition holds:

$$\begin{aligned}\mathbf{A} &= \begin{pmatrix}\mathbf{I}_{p_1} & \mathbf{0}\\ \mathbf{A}_{21}\mathbf{A}_{11}^{+} & \mathbf{I}_{p_2}\end{pmatrix}\cdot\begin{pmatrix}\mathbf{A}_{11} & \mathbf{0}\\ \mathbf{0} & \mathbf{A}_{22}-\mathbf{A}_{21}\mathbf{A}_{11}^{-}\mathbf{A}_{12}\end{pmatrix}\cdot\begin{pmatrix}\mathbf{I}_{p_1} & \mathbf{A}_{11}^{+}\mathbf{A}_{12}\\ \mathbf{0} & \mathbf{I}_{p_2}\end{pmatrix}\\ &:= \mathbf{F}'\mathbf{G}\mathbf{F}\,. \end{aligned} \tag{14.5}$$

The verification of (14.5) is straightforward in the spirit of the block-diagonalization Theorem 13 (p. 291). The matrix $\mathbf{F}$ in (14.5) is invertible and thus $\mathbf{A}$ is nonnegative definite if and only if $\mathbf{G}$ is nonnegative definite, and so

$$\mathbf{A}_{11} \geq_{\mathsf{L}} \mathbf{0} \text{ and } \mathbf{A}_{22\cdot1} \geq_{\mathsf{L}} \mathbf{0} \implies \begin{pmatrix}\mathbf{A}_{11} & \mathbf{0}\\ \mathbf{0} & \mathbf{A}_{22\cdot1}\end{pmatrix} \geq_{\mathsf{L}} \mathbf{0} \implies \mathbf{A} \geq_{\mathsf{L}} \mathbf{0}\,. \tag{14.6}$$

In the same way we can prove that (c) implies (a). □

Dey, Hande & Tiku (1994) give a statistical proof for Theorem 14. While we believe that Albert (1969) and (1972, Th. 9.1.6) was the first to establish these nonnegative definiteness conditions we are well aware of Stigler's Law of Eponymy, Stigler (1999, Ch. 14, p. 277), which "in its simplest form" states that "no scientific discovery is named after its original discoverer".

Consider a special case when we have the symmetric matrix

$$\mathbf{A} = \begin{pmatrix}\mathbf{B} & \mathbf{b}\\ \mathbf{b}' & \alpha\end{pmatrix}, \quad \text{where } \alpha > 0\,. \tag{14.7}$$

Then, on account of Theorem 14, the following statements are equivalent:

$$\mathbf{A} \geq_{\mathsf{L}} \mathbf{0}\,, \tag{14.8a}$$

$$\mathbf{B} - \frac{\mathbf{b}\mathbf{b}'}{\alpha} \geq_{\mathsf{L}} \mathbf{0}\,, \tag{14.8b}$$

$$\mathbf{B} \geq_{\mathsf{L}} \mathbf{0}\,, \qquad \mathbf{b} \in \mathscr{C}(\mathbf{B})\,, \qquad \alpha - \mathbf{b}'\mathbf{B}^{-}\mathbf{b} \geq 0\,. \tag{14.8c}$$

For related considerations, see also Farebrother (1976), Baksalary & Kala (1983b), Bekker & Neudecker (1989), and Baksalary, Schipp & Trenkler (1992, Th. 1).

When $\mathbf{A}_{11}$ and $\mathbf{A}_{22}$ are positive definite, it follows at once that $\mathscr{C}(\mathbf{A}_{12}) \subset \mathscr{C}(\mathbf{A}_{11})$ and $\mathscr{C}(\mathbf{A}_{21}) \subset \mathscr{C}(\mathbf{A}_{22})$. Hence the following three statements are equivalent:

(a) $\mathbf{A} >_{\mathsf{L}} \mathbf{0}$,

(b_1) $\mathbf{A}_{11} >_{\mathsf{L}} \mathbf{0}$, ($\mathrm{b}_3$) $\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12} >_{\mathsf{L}} \mathbf{0}$,

(c_1) $\mathbf{A}_{22} >_{\mathsf{L}} \mathbf{0}$, ($\mathrm{c}_3$) $\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} >_{\mathsf{L}} \mathbf{0}$.

In particular, putting $\mathbf{A}_{12} = \mathbf{I}_{p_1}$ and denoting $\mathbf{A}_{11} = \mathbf{U}$, $\mathbf{A}_{22}^{-1} = \mathbf{V}$, we get

$$\mathbf{U} \geq_{\mathsf{L}} \mathbf{V} \iff \mathbf{V}^{-1} \geq_{\mathsf{L}} \mathbf{U}^{-1} \quad \text{and} \quad \mathbf{U} >_{\mathsf{L}} \mathbf{V} \iff \mathbf{V}^{-1} >_{\mathsf{L}} \mathbf{U}^{-1}. \tag{14.9}$$

14.1 When Is a “Correlation Matrix” a Correlation Matrix?

Consider a symmetric matrix

$$\mathbf{R} = \begin{pmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{pmatrix}, \quad \text{where all } r_{ij}^2 \leq 1 \,. \tag{14.10}$$

When is this matrix $\mathbf{R}$ a proper correlation matrix, that is, such a correlation matrix which could be obtained from some real data? It is very crucial to realize that this is not the case for arbitrary chosen “correlations” r_{ij}. The key point here is that $\mathbf{R}$ must be nonnegative definite.

We define a *correlation matrix* $\mathbf{R}$ be a square $p \times p$ symmetric nonnegative definite matrix with all diagonal elements equal to 1. This definition is natural in the sense that given such a matrix $\mathbf{R}$, we can always construct an $n \times p$ data matrix whose correlation matrix is $\mathbf{R}$. Consider p variables $u_1, \dots, u_p$ whose observed centered values are the columns of $\mathbf{U} = (\mathbf{u}_1 : \dots : \mathbf{u}_p)$, and assume that each variable has a nonzero variance, i.e., $\mathbf{u}_i \neq \mathbf{0}$ for each i. Let each column of $\mathbf{U}$ have unit length. Now since the correlation coefficient r_{ij} is the cosine between the centered vectors $\mathbf{u}_i$ and $\mathbf{u}_j$, the correlation matrix $\mathbf{R}$ is simply $\mathbf{U}'\mathbf{U}$, and thereby nonnegative definite. Note that $\mathbf{U}'\mathbf{U}$ is not necessarily a correlation matrix of the u_i-variables if $\mathbf{U}$ is not centered (even though the columns of $\mathbf{U}$ have unit length). It is also clear that orthogonality and uncorrelatedness are equivalent concepts when dealing with centered data.

Interpreted as cosines, the off-diagonal elements r_{ij} of such a correlation matrix then satisfy the inequality $r_{ij}^2 \leq 1$ for all $i \neq j$; $i, j = 1, 2, \dots, p$. To go the other way, suppose that the square $p \times p$ symmetric matrix $\mathbf{R}$ has all its diagonal elements equal to 1 and all its off-diagonal elements $r_{ij}^2 \leq 1$; $i \neq j$ $(i, j = 1, 2, \dots, p)$. Then when is $\mathbf{R}$ a correlation matrix? That is, when is it nonnegative definite?

The special case when $p = 3$ is of particular interest, see, e.g., Hauke & Pomianowska (1987), Baksalary (1990), Yanai (2003), Yanai & Takane (2007), Arav, Hall & Li (2008), and in particular, Puntanen & Styan (2005b, §6.2.1), which we closely follow in this section. Olkin (1981) considers the above results (i)–(vii) in the case where three sets of variables are available.

Photograph 14.1 Haruo Yanai (New Delhi, 1992).

Let us consider the 3×3 symmetric matrix (14.10). Using Theorem 14 it is now easy to prove the following result.

Proposition 14.1. *The following statements concerning* (14.10) *are equivalent:*

(i) $\mathbf{R}$ *is a correlation matrix,*

(ii) $\det(\mathbf{R}) = 1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{13}r_{23} \geq 0$*,*

(iii) $(r_{12} - r_{13}r_{23})^2 \leq (1 - r_{13}^2)(1 - r_{23}^2)$*, or equivalently,*

$$r_{13}r_{23} - \sqrt{(1 - r_{13}^2)(1 - r_{23}^2)} \leq r_{12} \leq r_{13}r_{23} + \sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}\,,$$

(iv) $(r_{23} - r_{12}r_{31})^2 \leq (1 - r_{12}^2)(1 - r_{31}^2)$*,*

(v) $(r_{13} - r_{12}r_{32})^2 \leq (1 - r_{12}^2)(1 - r_{32}^2)$*,*

(vi) (a) $\mathbf{r}_{\mathbf{x}y} \in \mathscr{C}(\mathbf{R}_{\mathbf{xx}})$ *and* (b) $r_{yy\cdot\mathbf{x}} = 1 - \mathbf{r}'_{\mathbf{x}y}\mathbf{R}^-_{\mathbf{xx}}\mathbf{r}_{\mathbf{x}y} \geq 0$*, where, for convenience,*

$$\mathbf{R} = \begin{pmatrix} 1 & r_{12} & r_{1y} \\ r_{21} & 1 & r_{2y} \\ r_{y1} & r_{y2} & 1 \end{pmatrix} = \begin{pmatrix} \mathbf{R}_{\mathbf{xx}} & \mathbf{r}_{\mathbf{x}y} \\ \mathbf{r}'_{\mathbf{x}y} & 1 \end{pmatrix} = \begin{pmatrix} \mathbf{R}_{\mathbf{xx}} & \mathbf{r}_{\mathbf{x}y} \\ \mathbf{r}'_{\mathbf{x}y} & r_{yy} \end{pmatrix}, \tag{14.11}$$

(vii) $\mathbf{R}_{\mathbf{xx}\cdot y} = \mathbf{R}_{\mathbf{xx}} - \mathbf{r}_{\mathbf{x}y}\mathbf{r}'_{\mathbf{x}y} \geq_{\mathsf{L}} \mathbf{0}$*, where*

$$\mathbf{R}_{\mathbf{xx}} - \mathbf{r}_{\mathbf{x}y}\mathbf{r}'_{\mathbf{x}y} = \begin{pmatrix} 1 - r_{1y}^2 & r_{12} - r_{1y}r_{2y} \\ r_{12} - r_{1y}r_{2y} & 1 - r_{2y}^2 \end{pmatrix}. \tag{14.12}$$

Proof. The proof relies on the fact that (i) is equivalent to the nonnegative definiteness of $\mathbf{R}$. Using Albert's result, and since $r_{yy} = 1$ is certainly nonnegative and clearly $\mathscr{C}(\mathbf{r}'_{\mathbf{x}y}) \subset \mathbb{R}$, (i) holds if and only if

$$\mathbf{R}_{\mathbf{xx}\cdot y} = \mathbf{R}_{\mathbf{xx}} - \mathbf{r}_{\mathbf{x}y}\mathbf{r}'_{\mathbf{x}y} \geq_{\mathsf{L}} \mathbf{0}\,, \tag{14.13}$$

and thus (vii) is proved. Similarly, $\mathbf{R} \geq_{\mathsf{L}} \mathbf{0}$ if and only if

$$\mathbf{r}_{\mathbf{x}y} \in \mathscr{C}(\mathbf{R}_{\mathbf{xx}})\,, \quad r_{yy\cdot\mathbf{x}} = 1 - \mathbf{r}'_{\mathbf{x}y}\mathbf{R}^-_{\mathbf{xx}}\mathbf{r}_{\mathbf{x}y} \geq 0\,, \quad \mathbf{R}_{\mathbf{xx}} \geq_{\mathsf{L}} \mathbf{0}\,. \tag{14.14}$$

In (14.14) the last condition is always true since $r_{12}^2 \leq 1$, and hence (vi) is obtained. Conditions (ii)–(v) follow from (vii) at once. □

Notice that one way to consider the nonnegative definiteness of $\mathbf{R}$ in (14.10), would, of course, be to calculate all principal minors of $\mathbf{R}$, i.e., the determinants of all submatrices of $\mathbf{R}$ that have been obtained by deleting the same rows and columns from $\mathbf{R}$. Then $\mathbf{R}$ is nonnegative definite if and only if all principal minors are nonnegative. Similarly, $\mathbf{R}$ is positive definite if and only if all leading principal minors are positive. The seven principal minors of $\mathbf{R}$ are

$$\det(\mathbf{R})\,,\ \det\begin{pmatrix}1 & r_{12}\\ r_{12} & 1\end{pmatrix},\ \det\begin{pmatrix}1 & r_{13}\\ r_{13} & 1\end{pmatrix},\ \det\begin{pmatrix}1 & r_{23}\\ r_{23} & 1\end{pmatrix},\ 1\,,\ 1\,,\ 1\,. \tag{14.15}$$

Trivially the last six principal minors are nonnegative (assuming all $r_{ij}^2 \leq 1$) and hence $\mathbf{R}$ is nonnegative definite if and only if $\det(\mathbf{R}) \geq 0$.

The quantity

$$r_{12\cdot y}^2 = \frac{(r_{12} - r_{1y}r_{2y})^2}{(1-r_{1y}^2)(1-r_{2y}^2)}\,, \tag{14.16}$$

is well defined only when both $r_{1y}^2 \neq 1$ and $r_{2y}^2 \neq 1$ and then (when $\mathbf{R}$ is a correlation matrix) it is the formula for the squared partial correlation coefficient between variables (say) x_1 and x_2 when y is held constant.

When $\mathbf{R}$ is a correlation matrix, the quadratic form $\mathbf{r}'_{\mathbf{x}y}\mathbf{R}^-_{\mathbf{xx}}\mathbf{r}_{\mathbf{x}y}$ represents the squared multiple correlation coefficient when y is regressed on the variables x_1 and x_2:

$$R_{y\cdot 12}^2 = \mathbf{r}'_{\mathbf{x}y}\mathbf{R}^-_{\mathbf{xx}}\mathbf{r}_{\mathbf{x}y}\,. \tag{14.17}$$

The matrix $\mathbf{R}_{\mathbf{xx}}$ is singular if and only if $r_{12}^2 = 1$. If $r_{12} = 1$, then (a) of (vi) forces $r_{1y} = r_{2y}$ and so $\mathbf{r}_{\mathbf{x}y} = r_{1y}\mathbf{1}_2$. Choosing $\mathbf{R}^-_{\mathbf{xx}} = \mathbf{R}^+_{\mathbf{xx}} = \frac{1}{4}\left(\begin{smallmatrix}1 & 1\\ 1 & 1\end{smallmatrix}\right)$, we get $\mathbf{r}'_{\mathbf{x}y}\mathbf{R}^-_{\mathbf{xx}}\mathbf{r}_{\mathbf{x}y} = r_{1y}^2$. Similarly, if $r_{12} = -1$, then (a) of (vi) implies that $r_{1y} = -r_{2y}$, and again $\mathbf{r}'_{\mathbf{x}y}\mathbf{R}^-_{\mathbf{xx}}\mathbf{r}_{\mathbf{x}y} = r_{1y}^2$. If $r_{12}^2 \neq 1$, then we have

$$R_{y\cdot 12}^2 = \frac{r_{1y}^2 + r_{2y}^2 - 2r_{12}r_{1y}r_{2y}}{1 - r_{12}^2}\,. \tag{14.18}$$

Of course, if $r_{12} = 0$ then $R_{y\cdot 12}^2 = r_{1y}^2 + r_{2y}^2$, but it is interesting to observe that both

$$R_{y\cdot 12}^2 < r_{1y}^2 + r_{2y}^2 \tag{14.19}$$

and

$$R_{y\cdot 12}^2 > r_{1y}^2 + r_{2y}^2 \tag{14.20}$$

can occur. As Shieh (2001) points out, "...the inequality [(14.20)] may seem surprising and counterintuitive at first ... but it occurs more often than one may think." A geometric illustration of (14.20) is in Figure 8.5 (p. 184). For related literature, the reader is referred, e.g., to Hamilton (1987), Bertrand & Holder (1988), Cuadras (1993, 1995), Freund (1988), and Mitra (1988).

Let us now take two simple examples. First, consider the intraclass correlation structure:

$$\mathbf{A} = \begin{pmatrix}1 & r & r\\ r & 1 & r\\ r & r & 1\end{pmatrix} = (1-r)\mathbf{I}_3 + r\mathbf{1}_3\mathbf{1}_3'\,. \tag{14.21}$$

The determinant $\det(\mathbf{A}) = (1-r)^2(1+2r)$, and hence $\mathbf{A}$ in (14.21) is indeed a correlation matrix if and only if

$$-\frac{1}{2} \leq r \leq 1\,. \tag{14.22}$$

If $r = -0.5$, then the corresponding centered variable vectors $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{y}$ lie in the same plane so that the angle of each pair is 120°. In general, the $p \times p$ intraclass correlation matrix

$$\mathbf{R} = \begin{pmatrix} 1 & r & \dots & r \\ r & 1 & \dots & r \\ \vdots & \vdots & \ddots & \vdots \\ r & r & \dots & 1 \end{pmatrix} = \begin{pmatrix} \mathbf{R}_{\mathbf{xx}} & \mathbf{r}_{\mathbf{x}y} \\ \mathbf{r}'_{\mathbf{x}y} & 1 \end{pmatrix} = \begin{pmatrix} \mathbf{R}_{\mathbf{xx}} & r\mathbf{1}_{p-1} \\ r\mathbf{1}'_{p-1} & 1 \end{pmatrix} \tag{14.23}$$

must satisfy

$$-\frac{1}{p-1} \leq r \leq 1\,, \tag{14.24}$$

which follows from the fact, see Proposition 18.2 (p. 366), that the eigenvalues of $\mathbf{R}$ are

$$\operatorname{ch}(\mathbf{R}) = \begin{cases} 1 + (p-1)r & \text{with multiplicity } 1\,, \\ 1 - r & \text{with multiplicity } p-1\,. \end{cases} \tag{14.25}$$

Notice that $\mathbf{R}_{\mathbf{xx}}\mathbf{1}_{p-1} = [1 + (p-2)r]\mathbf{1}_{p-1}$ and thereby, if $r \neq -\frac{1}{p-2}$,

$$\mathbf{1}_{p-1} = [1 + (p-2)r]^{-1}\mathbf{R}_{\mathbf{xx}}\mathbf{1}_{p-1}\,. \tag{14.26}$$

Now for a nonnegative definite $\mathbf{R}$, (14.22) guarantees that $r \neq -\frac{1}{p-2}$ and, see (18.73) (p. 367),

$$\begin{aligned} \mathbf{r}'_{\mathbf{x}y}\mathbf{R}^{-}_{\mathbf{xx}}\mathbf{r}_{\mathbf{x}y} &= \frac{r^2}{[1 + (p-2)r]^2}\,\mathbf{1}'_{p-1}\mathbf{R}_{\mathbf{xx}}\mathbf{R}^{-}_{\mathbf{xx}}\mathbf{R}_{\mathbf{xx}}\mathbf{1}_{p-1} \\ &= \frac{r^2}{[1 + (p-2)r]^2}\,\mathbf{1}'_{p-1}\mathbf{R}_{\mathbf{xx}}\mathbf{1}_{p-1} = \frac{(p-1)r^2}{1 + (p-2)r}\,. \end{aligned} \tag{14.27}$$

If $\mathbf{R}$ is a correlation matrix, then $R^2_{y\cdot\mathbf{x}} = \mathbf{r}'_{\mathbf{x}y}\mathbf{R}^{-}_{\mathbf{xx}}\mathbf{r}_{\mathbf{x}y}$, with $R^2_{y\cdot\mathbf{x}}$ referring to the squared multiple correlation when y is explained by the x_i-variables (plus the constant). It is left as an exercise to confirm that (14.24) is equivalent to

$$\mathbf{r}'_{\mathbf{x}y}\mathbf{R}^{-}_{\mathbf{xx}}\mathbf{r}_{\mathbf{x}y} = \frac{(p-1)r^2}{1 + (p-2)r} \leq 1\,. \tag{14.28}$$

14.2 $\mathbf{A} \leq_{\mathsf{L}} \mathbf{B} \iff \mathscr{C}(\mathbf{A}) \subset \mathscr{C}(\mathbf{B})$ and $\mathbf{ch_1}(\mathbf{AB}^+) \leq 1$

The result mentioned in the header above does not fit 100 per cent in this section—we are not using Albert's result—but in any event, it gives a handy characterization for the Löwner ordering $\mathbf{A} \leq_{\mathsf{L}} \mathbf{B}$.

We begin with the following result.

Proposition 14.2. *Let* $\mathbf{A}_{n\times n}$ *and* $\mathbf{B}_{n\times n}$ *be nonnegative definite symmetric matrices. Then*

$$\mathbf{A} \leq_{\mathsf{L}} \mathbf{B} \implies \mathscr{C}(\mathbf{A}) \subset \mathscr{C}(\mathbf{B})\,. \tag{14.29}$$

Proof. One way to prove (14.29) is to write $\mathbf{A} = \mathbf{UU}'$, $\mathbf{B} = \mathbf{VV}'$, and $\mathbf{B} - \mathbf{A} = \mathbf{NN}'$, and then use the fact that $\mathbf{B} = \mathbf{UU}' + \mathbf{NN}'$, which implies

$$\mathscr{C}(\mathbf{B}) = \mathscr{C}(\mathbf{U} : \mathbf{N}) = \mathscr{C}(\mathbf{U}) + \mathscr{C}(\mathbf{N}) = \mathscr{C}(\mathbf{A}) + \mathscr{C}(\mathbf{N})\,. \tag{14.30}$$

An alternative way is to note that

$$\mathbf{B} \geq_{\mathsf{L}} \mathbf{A} \geq_{\mathsf{L}} \mathbf{0} \iff \mathbf{x}'\mathbf{Bx} \geq \mathbf{x}'\mathbf{Ax} \geq 0 \quad \text{for all } \mathbf{x} \in \mathbb{R}^n. \tag{14.31}$$

Choose now $\mathbf{x} \in \mathscr{N}(\mathbf{B}) = \mathscr{C}(\mathbf{B})^\perp$ and note that in view of (14.31), we must have

$$0 \geq \mathbf{x}'\mathbf{Ax} \geq 0\,, \tag{14.32}$$

and hence $\mathbf{x} \in \mathscr{N}(\mathbf{A}) = \mathscr{C}(\mathbf{A})^\perp$, i.e., $\mathscr{N}(\mathbf{B}) \subset \mathscr{N}(\mathbf{A})$. Obviously, cf. part (c) of Theorem 1 (p. 57),

$$\mathscr{N}(\mathbf{B}) \subset \mathscr{N}(\mathbf{A}) \Leftrightarrow [\mathscr{N}(\mathbf{A})]^\perp \subset [\mathscr{N}(\mathbf{B})]^\perp \Leftrightarrow \mathscr{C}(\mathbf{A}) \subset \mathscr{C}(\mathbf{B})\,. \tag{14.33}$$

□

Proposition 14.3. *Let* $\mathbf{A}_{n\times n}$ *and* $\mathbf{B}_{n\times n}$ *be nonnegative definite symmetric matrices. Then the following two statements are equivalent:*

(a) $\mathbf{A} \leq_{\mathsf{L}} \mathbf{B}$,

(b) $\mathscr{C}(\mathbf{A}) \subset \mathscr{C}(\mathbf{B})$ *and* $\mathrm{ch}_1(\mathbf{AB}^+) \leq 1$.

Proof. Let $\mathbf{B}$ have an eigenvalue decomposition $\mathbf{B} = \mathbf{T}_1\mathbf{\Lambda}_1\mathbf{T}_1'$, where the columns of $\mathbf{T}_1 \in \mathbb{R}^{n\times b}$ are the orthonormal eigenvectors corresponding to nonzero eigenvalues of $\mathbf{B}$, and $\mathbf{\Lambda}_1$ is the diagonal matrix of these eigenvalues and $\mathrm{rank}(\mathbf{B}) = b$. Denoting $\mathbf{L} = \mathbf{T}_1\mathbf{\Lambda}_1^{1/2}$ and $\mathbf{K} = \mathbf{T}_1\mathbf{\Lambda}_1^{-1/2}$, we have

$$\mathbf{LL}' = \mathbf{B}, \quad \mathbf{KK}' = \mathbf{B}^+, \quad \mathbf{LK}' = \mathbf{T}_1\mathbf{T}_1' = \mathbf{P}_\mathbf{B}\,. \tag{14.34}$$

Assume first that $\mathbf{A} \leq_{\mathsf{L}} \mathbf{B}$, i.e.,

$$\mathbf{T}_1\mathbf{\Lambda}_1\mathbf{T}_1' - \mathbf{A} \geq_{\mathsf{L}} \mathbf{0}\,. \tag{14.35}$$

Pre- and postmultiplying (14.35) by $\mathbf{K}'$ and $\mathbf{K}$, respectively, yields

$$\mathbf{I}_b - \mathbf{K}'\mathbf{A}\mathbf{K} \geq_{\mathsf{L}} \mathbf{0}\,, \tag{14.36}$$

which holds if and only if

$$1 \geq \operatorname{ch}_1(\mathbf{K}'\mathbf{A}\mathbf{K}) = \operatorname{ch}_1(\mathbf{K}\mathbf{K}'\mathbf{A}) = \operatorname{ch}_1(\mathbf{B}^{+}\mathbf{A})\,, \tag{14.37}$$

and thus we have confirmed that (a) implies (b).

Suppose then that (b) holds. Then $\operatorname{ch}_1(\mathbf{B}^{+}\mathbf{A}) \leq 1$ means that necessarily (14.36) holds. Pre- and postmultiplying (14.36) by $\mathbf{L}$ and $\mathbf{L}'$, respectively, gives

$$\mathbf{L}\mathbf{L}' - \mathbf{L}\mathbf{K}'\mathbf{A}\mathbf{K}\mathbf{L}' = \mathbf{B} - \mathbf{P}_{\mathbf{B}}\mathbf{A}\mathbf{P}_{\mathbf{B}} \geq_{\mathsf{L}} \mathbf{0}\,. \tag{14.38}$$

Assumption $\mathscr{C}(\mathbf{A}) \subset \mathscr{C}(\mathbf{B})$ implies that $\mathbf{P}_{\mathbf{B}}\mathbf{A}\mathbf{P}_{\mathbf{B}} = \mathbf{A}$, and hence (14.38) means that (a) holds. Thus the proof is completed.

Note that the nonzero eigenvalues of $\mathbf{A}\mathbf{B}^{-}$ are invariant with respect to the choice of $\mathbf{B}^{-}$. Namely, writing $\mathbf{A} = \mathbf{F}\mathbf{F}'$, we have

$$\operatorname{nzch}(\mathbf{A}\mathbf{B}^{-}) = \operatorname{nzch}(\mathbf{F}\mathbf{F}'\mathbf{B}^{-}) = \operatorname{nzch}(\mathbf{F}'\mathbf{B}^{-}\mathbf{F})\,, \tag{14.39}$$

where the matrix $\mathbf{F}'\mathbf{B}^{-}\mathbf{F}$ is invariant for any choice of $\mathbf{B}^{-}$ because $\mathscr{C}(\mathbf{F}) \subset \mathscr{C}(\mathbf{B})$. □

Photograph 14.2 Friedrich Pukelsheim (Smolenice Castle, Slovakia, 2009).

The proof above follows that of Liski & Puntanen (1989); see also Stępniak (1985) and Baksalary, Liski & Trenkler (1989). Note that the equivalence of (a) and (b) gives immediately the result (14.9) (p. 306), i.e.,

$$\mathbf{A} \leq_{\mathsf{L}} \mathbf{B} \iff \mathbf{B}^{-1} \leq_{\mathsf{L}} \mathbf{A}^{-1}, \tag{14.40}$$

where $\mathbf{A}$ and $\mathbf{B}$ are positive definite. For a quick proof of (14.40), see also Andrews & Phillips (1986), and Pukelsheim (1993, p. 13).

14.3 BLUE's Covariance Matrix as a Shorted Matrix

Following Isotalo, Puntanen & Styan (2008a, §4), let us consider a simple linear model $\{\mathbf{y},\, \mathbf{1}\beta,\, \mathbf{V}\}$, where $\mathbf{V}$ is positive definite. Let our task be to find a nonnegative definite matrix $\mathbf{S}$ which belongs to the set

$$\mathcal{U} = \{\, \mathbf{U} : \mathbf{0} \leq_{\mathsf{L}} \mathbf{U} \leq_{\mathsf{L}} \mathbf{V},\ \mathscr{C}(\mathbf{U}) \subset \mathscr{C}(\mathbf{1}) \,\}\,, \tag{14.41}$$

and which is maximal in the Löwner sense; that is, a nonnegative definite matrix which is "as close to $\mathbf{V}$ as possible" in the Löwner partial ordering, but whose column space is in that of $\mathbf{1}$. This matrix $\mathbf{S}$ is called the *shorted matrix* of $\mathbf{V}$ with respect to $\mathbf{1}$, and denoted as $\operatorname{Sh}(\mathbf{V} \mid \mathbf{1})$.

Because $\mathbf{S}$ is nonnegative definite, we must have $\mathbf{S} = \mathbf{L}\mathbf{L}'$ for some $\mathbf{L}$ of full column rank. Further, the condition $\mathscr{C}(\mathbf{S}) \subset \mathscr{C}(\mathbf{1})$ implies that $\mathbf{L} = \alpha\mathbf{1}$ for some nonzero scalar α and hence $\mathbf{S} = \alpha^2\mathbf{1}\mathbf{1}'$. Our objective is to find a scalar α so that $\alpha^2\mathbf{1}\mathbf{1}'$ is maximal in the Löwner sense, which means that α^2 must be maximal. The choice of α^2 must be made under the condition

$$\alpha^2\mathbf{1}\mathbf{1}' \leq_{\mathsf{L}} \mathbf{V}. \tag{14.42}$$

We show that the maximal value for α^2 is

$$\alpha^2 = (\mathbf{1}'\mathbf{V}^{-1}\mathbf{1})^{-1}. \tag{14.43}$$

To confirm that (14.43) gives the maximal α^2, consider the matrix

$$\mathbf{A} = \begin{pmatrix} \mathbf{V} & \mathbf{1} \\ \mathbf{1}' & \alpha^{-2} \end{pmatrix}. \tag{14.44}$$

Then, according to (14.7) and (14.8) (p. 306), the equivalent condition sets (b) and (c) can in this situation be written as

$$(\mathrm{b}_1)\ \mathbf{V} \geq_{\mathsf{L}} \mathbf{0}\,, \qquad (\mathrm{b}_2)\ \mathbf{1} \in \mathscr{C}(\mathbf{V})\,, \qquad (\mathrm{b}_3)\ \alpha^{-2} - \mathbf{1}'\mathbf{V}^{-1}\mathbf{1} \geq 0\,, \tag{14.45a}$$

$$(\mathrm{c}_1)\ \alpha^{-2} \geq 0\,, \qquad (\mathrm{c}_2)\ \alpha \neq 0\,, \qquad (\mathrm{c}_3)\ \mathbf{V} - \alpha^2\mathbf{1}\mathbf{1}' \geq_{\mathsf{L}} \mathbf{0}\,. \tag{14.45b}$$

Clearly the first two conditions vanish when $\mathbf{V}$ is positive definite and hence we have

$$\alpha^{-2} - \mathbf{1}'\mathbf{V}^{-1}\mathbf{1} \geq 0 \iff \mathbf{V} - \alpha^2\mathbf{1}\mathbf{1}' \geq_{\mathsf{L}} \mathbf{0}\,. \tag{14.46}$$

Therefore the maximal choice is as in (14.43) and the shorted matrix $\mathbf{S}$ is

$$\mathrm{Sh}(\mathbf{V} \mid \mathbf{1}) = (\mathbf{1}'\mathbf{V}^{-1}\mathbf{1})^{-1}\mathbf{1}\mathbf{1}', \tag{14.47}$$

which is precisely the covariance matrix of $\mathrm{BLUE}(\mathbf{1}\beta)$ under $\{\mathbf{y},\, \mathbf{1}\beta,\, \mathbf{V}\}$. This result can be also generalized as shown below.

Consider now a general case of $\mathcal{U}$:

$$\mathcal{U} = \{\,\mathbf{U} : \mathbf{0} \leq_{\mathsf{L}} \mathbf{U} \leq_{\mathsf{L}} \mathbf{V},\ \mathscr{C}(\mathbf{U}) \subset \mathscr{C}(\mathbf{X})\,\}\,. \tag{14.48}$$

The maximal element $\mathbf{U}$ in $\mathcal{U}$ is the shorted matrix of $\mathbf{V}$ with respect to $\mathbf{X}$, and denoted as $\mathrm{Sh}(\mathbf{V} \mid \mathbf{X})$. The concept of shorted matrix (or operator) was first introduced by Krein (1947), and later rediscovered by W. N. Anderson (1971), who introduced the term "shorted operator". Mitra & Puri (1979, 1982) were apparently the first to consider statistical applications of the shorted matrix and the shorted operator.

As shown by W. N. Anderson (1971), and by W. N. Anderson and Trapp (1975), the set $\mathcal{U}$ in (14.48) indeed has a maximal element and it, the shorted matrix, is unique.

Consider now the general linear model $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$. Then, as Mitra & Puntanen (1991) proved,

$$\operatorname{cov}[\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta})] = \mathrm{Sh}(\mathbf{V} \mid \mathbf{X})\,. \tag{14.49}$$

Let us go through the proof which is rather easy, while the result (14.49) itself is somewhat unexpected. To prove (14.49), let $\mathbf{Gy} = \mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta})$ and so

$$\operatorname{cov}(\mathbf{Gy}) = \mathbf{GVG}' \leq_{\mathsf{L}} \mathbf{V} = \operatorname{cov}(\mathbf{y})\,, \tag{14.50}$$

because $\mathbf{y}$ is an unbiased estimator of $\mathbf{X}\boldsymbol{\beta}$. Let $\mathbf{U}$ be an arbitrary member of $\mathcal{U}$, which implies that $\mathbf{U} = \mathbf{HAA}'\mathbf{H}$ for some matrix $\mathbf{A}$ and

$$\mathbf{U} = \mathbf{HAA}'\mathbf{H} \leq_{\mathsf{L}} \mathbf{V}\,, \quad \text{where } \mathbf{H} = \mathbf{P}_{\mathbf{X}}\,. \tag{14.51}$$

Premultiplying (14.51) by $\mathbf{G}$ and postmultiplying it by $\mathbf{G}'$ yields

$$\mathbf{GUG}' = \mathbf{GHAA}'\mathbf{HG}' = \mathbf{HAA}'\mathbf{H} = \mathbf{U} \leq_{\mathsf{L}} \mathbf{GVG}', \tag{14.52}$$

where we have used the fact that $\mathbf{GH} = \mathbf{H}$. Now (14.52) confirms that $\mathbf{GVG}'$ is the maximal element in the class $\mathcal{U}$, i.e., (14.49) holds.

For a review of shorted matrices and their applications in statistics, see Mitra, Puntanen & Styan (1995) and Mitra, Bhimasankaram & Malik (2010).

14.4 Exercises

14.1. Consider $\mathbf{B} = \begin{pmatrix} 1 & a & r \\ a & 1 & r \\ r & r & 1 \end{pmatrix}$, where a is a given real number, $a^2 \leq 1$. What are the possible values for r such that $\mathbf{B}$ is a correlation matrix? Answer: $r^2 \leq \frac{1+a}{2}$.

14.2. Consider

$$\mathbf{R} = \begin{pmatrix} 1 & 0 & r & r \\ 0 & 1 & r & -r \\ r & r & 1 & 0 \\ r & -r & 0 & 1 \end{pmatrix}.$$

What are the possible values for r such that $\mathbf{R}$ is a correlation matrix? Find the squared multiple correlation coefficients $R^2_{4\cdot123}$, $R^2_{3\cdot124}$, ...

Yanai & Takane (2007, p. 352).

14.3. Consider $(p+q)$-dimensional random vectors $\binom{\mathbf{x}}{\mathbf{y}}$ and $\binom{\mathbf{u}}{\mathbf{v}}$, where $\mathbf{x}$ and $\mathbf{u}$ are p-dimensional. Let

$$\operatorname{cov}\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{xx}} & \boldsymbol{\Sigma}_{\mathbf{xy}} \\ \boldsymbol{\Sigma}_{\mathbf{yx}} & \boldsymbol{\Sigma}_{\mathbf{yy}} \end{pmatrix} = \boldsymbol{\Sigma}\,, \quad \operatorname{cov}\begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{uu}} & \boldsymbol{\Sigma}_{\mathbf{uv}} \\ \boldsymbol{\Sigma}_{\mathbf{vu}} & \boldsymbol{\Sigma}_{\mathbf{vv}} \end{pmatrix}.$$

Prove: $\operatorname{rank}(\boldsymbol{\Sigma}) = \operatorname{rank}(\boldsymbol{\Sigma}_{\mathbf{xx}}) \iff \varrho_{y_i\cdot\mathbf{x}} = 1$, $i = 1, \dots, q$, where $\varrho^2_{y_i\cdot\mathbf{x}} = \boldsymbol{\sigma}'_{\mathbf{x}y_i}\boldsymbol{\Sigma}^-_{\mathbf{xx}}\boldsymbol{\sigma}_{\mathbf{x}y_i}/\sigma^2_{y_i}$ = squared population multiple correlation coefficient. If $\sigma_{y_i} = 0$, then $\varrho_{y_i\cdot\mathbf{x}}$ may be put 1; see (9.20) (p. 193).

Mitra (1973a).

14.4 (Continued ...). Suppose that $\operatorname{cov}\big(\binom{\mathbf{x}}{\mathbf{y}}, \binom{\mathbf{u}}{\mathbf{v}}\big) = \mathbf{0}$, and denote $\binom{\mathbf{z}}{\mathbf{w}} = \binom{\mathbf{x}}{\mathbf{y}} + \binom{\mathbf{u}}{\mathbf{v}}$. Prove: $\varrho_{w_i\cdot\mathbf{z}} = 1 \implies \varrho_{y_i\cdot\mathbf{x}} = 1$ and $\varrho_{v_i\cdot\mathbf{u}} = 1$, $i = 1, \dots, q$.

Mitra (1973a).

14.5. Let $\mathbf{0} \leq_{\mathsf{L}} \mathbf{A} \leq_{\mathsf{L}} \mathbf{B}$ and and let $\mathbf{A}^{\sim}$ and $\mathbf{B}^{\sim}$ be symmetric reflexive generalized inverses of $\mathbf{A}$ and $\mathbf{B}$, respectively. Show that then

$$\mathbf{A}^{\sim} \geq_{\mathsf{L}} \mathbf{B}^{\sim} \iff \mathscr{C}(\mathbf{A}^{\sim}) = \mathscr{C}(\mathbf{B}^{\sim}).$$

Liski & Puntanen (1989).

14.6. Let $\mathbf{0} \leq_{\mathsf{L}} \mathbf{A} \leq_{\mathsf{L}} \mathbf{B}$. Using the Exercise 14.5 prove the result due to Milliken & Akdeniz (1977):

$$\mathbf{A}^{+} \geq_{\mathsf{L}} \mathbf{B}^{+} \iff \operatorname{rank}(\mathbf{A}) = \operatorname{rank}(\mathbf{B}).$$

14.7. Show that the following five statements are equivalent when considering the linear model $\{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$ with $\mathbf{G}$ a generalized inverse of $\mathbf{X}$.

(a) $\mathbf{XGVG'X'} \leq_{\mathsf{L}} \mathbf{V}$,
(b) $\mathbf{G'}$ is a minimum-$\mathbf{V}$-seminorm generalized inverse of $\mathbf{X'}$,
(c) $\mathbf{XGy}$ is the BLUE for $\mathbf{X}\boldsymbol{\beta}$,
(d) $\mathbf{XGVG'X'} \leq_{\text{rs}} \mathbf{V}$, i.e., $\operatorname{rk}(\mathbf{V}-\mathbf{XGVG'X'}) = \operatorname{rk}(\mathbf{V})-\operatorname{rk}(\mathbf{XGVG'X'})$,
(e) $\mathbf{XGVG'X'} = \operatorname{Sh}(\mathbf{V} \mid \mathbf{X})$.

In (d) the symbol $\leq_{\text{rs}}$ denotes the rank-subtractivity partial ordering or the minus ordering introduced by Hartwig (1980).

Mitra, Puntanen & Styan (1995),
Chipman (1968, Lemma 1.2; 1976, Lemma 1.2).

14.8. Show that the shorted matrix $\mathbf{S} = \operatorname{Sh}(\mathbf{V} \mid \mathbf{X})$ has the following properties:

(a) $\mathscr{C}(\mathbf{S}) = \mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{V})$,
(b) $\mathscr{C}(\mathbf{V}) = \mathscr{C}(\mathbf{S}) \oplus \mathscr{C}(\mathbf{V} - \mathbf{S})$,
(c) $\operatorname{rank}(\mathbf{V}) = \operatorname{rank}(\mathbf{S}) + \operatorname{rank}(\mathbf{V} - \mathbf{S})$,
(d) $\mathbf{SV}^{+}(\mathbf{V} - \mathbf{S}) = \mathbf{0}$,
(e) $\mathbf{V}^{-} \in \{\mathbf{S}^{-}\}$ for some (and hence for all) $\mathbf{V}^{-} \in \{\mathbf{V}^{-}\}$, i.e., $\{\mathbf{V}^{-}\} \subset \{\mathbf{S}^{-}\}$.

14.9. The minus (or rank-subtractivity) partial ordering for the two $n \times m$ matrices $\mathbf{A}$ and $\mathbf{B}$ is defined as

$$\mathbf{A} \leq_{\mathsf{rs}} \mathbf{B} \iff \operatorname{rank}(\mathbf{B} - \mathbf{A}) = \operatorname{rank}(\mathbf{B}) - \operatorname{rank}(\mathbf{A}).$$

Confirm that $\mathbf{A} \leq_{\mathsf{rs}} \mathbf{B}$ holds if and only if any of the following conditions holds:

(a) $\mathbf{A}^-\mathbf{A} = \mathbf{A}^-\mathbf{B}$ and $\mathbf{A}\mathbf{A}^- = \mathbf{B}\mathbf{A}^-$ for some $\mathbf{A}^- \in \{\mathbf{A}^-\}$,
(b) $\mathbf{A}^-\mathbf{A} = \mathbf{A}^-\mathbf{B}$ and $\mathbf{A}\mathbf{A}^\sim = \mathbf{B}\mathbf{A}^\sim$ for some $\mathbf{A}^-$, $\mathbf{A}^\sim \in \{\mathbf{A}^-\}$,
(c) $\{\mathbf{B}^-\} \subset \{\mathbf{A}^-\}$,
(d) $\mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B} - \mathbf{A}) = \{\mathbf{0}\}$ and $\mathscr{C}(\mathbf{A}') \cap \mathscr{C}(\mathbf{B}' - \mathbf{A}') = \{\mathbf{0}\}$,
(e) $\mathscr{C}(\mathbf{A}) \subset \mathscr{C}(\mathbf{B})$, $\mathscr{C}(\mathbf{A}') \subset \mathscr{C}(\mathbf{B}')$ & $\mathbf{A}\mathbf{B}^-\mathbf{A} = \mathbf{A}$ for some (and hence for all) $\mathbf{B}^-$.

Hartwig (1980), Hartwig & Styan (1986),
Baksalary, Pukelsheim & Styan (1989), Mitra (1972, 1986).

14.10. Let $\mathbf{A} \leq_{\mathsf{L}} \mathbf{B}$, where $\mathbf{A} \geq_{\mathsf{L}} \mathbf{0}$ and $\mathbf{B} >_{\mathsf{L}} \mathbf{0}$. Prove that then $\det(\mathbf{A}) = \det(\mathbf{B}) \implies \mathbf{A} = \mathbf{B}$. What about if we if we only assume that $\mathbf{B} \geq_{\mathsf{L}} \mathbf{0}$?

14.11. Let $\mathbf{V} \in \mathrm{NND}_n$ and $\mathbf{G} \in \{\mathbf{V}^-\}$. Find necessary and sufficient conditions for

(a) $\mathbf{G} \in \mathrm{NND}_n$,
(b) $\mathbf{G} \in \mathrm{NND}_n$ and $\mathbf{G}$ is a reflexive generalized inverse of $\mathbf{V}$.

Hint: Use the general representation of$\mathbf{V}^- = \mathbf{T}\begin{pmatrix} \mathbf{\Lambda}_1^{-1} & \mathbf{L} \\ \mathbf{M} & \mathbf{N} \end{pmatrix}\mathbf{T}'$, and Albert's conditions for a partitioned matrix being nonnegative definite. The answer to (b) is then

$$\mathbf{G} = \mathbf{T}\begin{pmatrix} \mathbf{\Lambda}_1^{-1} & \mathbf{L} \\ \mathbf{L}' & \mathbf{L}'\mathbf{\Lambda}_1\mathbf{L} \end{pmatrix}\mathbf{T}', \quad \text{where } \mathbf{L} \text{ is an arbitrary matrix.}$$

14.12. Let $\mathbf{A}$ and $\mathbf{B}$ be nonnegative definite $n \times n$ matrices. Show the following:

(a) $\mathbf{A} \leq_{\mathsf{L}} \mathbf{B} \implies \mathbf{A}^{1/2} \leq_{\mathsf{L}} \mathbf{B}^{1/2}$.
(b) In general, $\mathbf{A} \leq_{\mathsf{L}} \mathbf{B}$ does not imply $\mathbf{A}^2 \leq_{\mathsf{L}} \mathbf{B}^2$.
(c) $\mathbf{A} \leq_{\mathsf{L}} \mathbf{B}$ & $\mathbf{A}\mathbf{B} = \mathbf{B}\mathbf{A} \implies \mathbf{A}^2 \leq_{\mathsf{L}} \mathbf{B}^2$.

Baksalary & Pukelsheim (1991), Zhang (1999, pp. 169–170),
Abadir & Magnus (2005, p. 332), Baksalary, Hauke, Liu & Liu (2004).

14.13. Show that for conformable matrices $\mathbf{A}$ and $\mathbf{B}$ we have

$$\mathbf{A}\mathbf{B}' + \mathbf{B}\mathbf{A}' \geq_{\mathsf{L}} \mathbf{0} \iff \mathrm{ch}_1\left[(\mathbf{A} - \mathbf{B})'(\mathbf{A}\mathbf{A}' + \mathbf{B}\mathbf{B}')^-(\mathbf{A} - \mathbf{B})\right] \leq 1\,.$$

Hint: Write $\mathbf{A}\mathbf{B}' + \mathbf{B}\mathbf{A}' = (\mathbf{A}\mathbf{A}' + \mathbf{B}\mathbf{B}') - (\mathbf{A} - \mathbf{B})(\mathbf{A} - \mathbf{B})'$ and apply Proposition 14.3 (p. 311).

14.14. Suppose that $\mathbf{V} \in \mathrm{NND}_n$, $\mathbf{A} \in \mathbb{R}^{n\times n}$, and $\mathbf{V}\mathbf{A}^2 = \mathbf{V}\mathbf{A}$. Show that then $\mathbf{A}'\mathbf{V}\mathbf{A} \leq_{\mathsf{L}} \mathbf{V} \iff \mathbf{A}'\mathbf{V} = \mathbf{V}\mathbf{A}$.

Baksalary, Kala & Kłaczyński (1983).

Chapter 15

The Matrix $\dot{\mathbf{M}}$

> *A child of five would understand this.*
> *Send someone to fetch a child of five.*
>
> Groucho Marx

It is well known that if $\mathbf{V}$ is a symmetric positive definite $n \times n$ matrix, and $(\mathbf{X} : \mathbf{Z})$ is a partitioned orthogonal $n \times n$ matrix, then

$$(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} = \mathbf{X}'\mathbf{V}\mathbf{X} - \mathbf{X}'\mathbf{V}\mathbf{Z}(\mathbf{Z}'\mathbf{V}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{V}\mathbf{X}\,. \tag{15.1}$$

In this chapter we illustrate the usefulness the above formula, and in particular its version

$$\mathbf{Z}(\mathbf{Z}'\mathbf{V}\mathbf{Z})^{-1}\mathbf{Z}' = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}, \tag{15.2}$$

and present several related formulas, as well as some generalized versions. We also include several statistical applications. One way to confirm (15.1) is to observe that

$$[(\mathbf{X} : \mathbf{Z})'\mathbf{V}(\mathbf{X} : \mathbf{Z})]^{-1} = (\mathbf{X} : \mathbf{Z})'\mathbf{V}^{-1}(\mathbf{X} : \mathbf{Z})\,, \tag{15.3a}$$

$$\begin{pmatrix} \mathbf{X}'\mathbf{V}\mathbf{X} & \mathbf{X}'\mathbf{V}\mathbf{Z} \\ \mathbf{Z}'\mathbf{V}\mathbf{X} & \mathbf{Z}'\mathbf{V}\mathbf{Z} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{X}'\mathbf{V}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{V}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{V}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z} \end{pmatrix}, \tag{15.3b}$$

and then note, using the Schur complement, that the upper-left block of $[(\mathbf{X} : \mathbf{Z})'\mathbf{V}(\mathbf{X} : \mathbf{Z})]^{-1}$ is the inverse of the right-hand side of (15.1).

The center of attention in this chapter is the matrix

$$\dot{\mathbf{M}} = \mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^{-}\mathbf{M}\,, \tag{15.4}$$

where $\mathbf{M} = \mathbf{I}_n - \mathbf{H} = \mathbf{I}_n - \mathbf{P}_{\mathbf{X}}$. In particular, if $\mathbf{V}$ is positive definite and $\mathbf{V}^{1/2}$ is its positive definite symmetric square root, and $\mathbf{Z}$ is a matrix with the property $\mathscr{C}(\mathbf{Z}) = \mathscr{C}(\mathbf{M})$, then we obviously have

$$\begin{aligned} \dot{\mathbf{M}} &= \mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^{-}\mathbf{M} \\ &= \mathbf{V}^{-1/2}\mathbf{V}^{1/2}\mathbf{M}(\mathbf{M}\mathbf{V}^{1/2}\mathbf{V}^{1/2}\mathbf{M})\mathbf{M}\mathbf{V}^{1/2}\mathbf{V}^{-1/2} \\ &= \mathbf{V}^{-1/2}\mathbf{P}_{\mathbf{V}^{1/2}\mathbf{M}}\mathbf{V}^{-1/2} \end{aligned}$$

S. Puntanen et al., *Matrix Tricks for Linear Statistical Models: Our Personal Top Twenty*,
DOI 10.1007/978-3-642-10473-2_16,

$$= \mathbf{V}^{-1/2}\mathbf{P}_{\mathbf{V}^{1/2}\mathbf{Z}}\mathbf{V}^{-1/2} = \mathbf{Z}(\mathbf{Z}'\mathbf{V}\mathbf{Z})^{-}\mathbf{Z}', \tag{15.5}$$

which is clearly unique. In general, the matrix $\dot{\mathbf{M}}$ is not necessarily unique with respect to the choice of $(\mathbf{MVM})^{-}$. In view of Theorem 12 (p. 283), $\dot{\mathbf{M}}$ is unique if and only if $\mathscr{C}(\mathbf{M}) \subset \mathscr{C}(\mathbf{MV})$, which is easily seen to be equivalent to $\mathbb{R}^n = \mathscr{C}(\mathbf{X} : \mathbf{V})$; see also Exercise 12.4 (p. 288). Even though $\dot{\mathbf{M}}$ is not necessarily unique, the matrix product

$$\mathbf{P}_{\mathbf{V}}\dot{\mathbf{M}}\mathbf{P}_{\mathbf{V}} = \mathbf{P}_{\mathbf{V}}\mathbf{M}(\mathbf{MVM})^{-}\mathbf{M}\mathbf{P}_{\mathbf{V}} := \ddot{\mathbf{M}} \tag{15.6}$$

is, however, invariant for any choice of $(\mathbf{MVM})^{-}$, i.e.,

$$\ddot{\mathbf{M}} = \mathbf{P}_{\mathbf{V}}\dot{\mathbf{M}}\mathbf{P}_{\mathbf{V}} = \mathbf{P}_{\mathbf{V}}\mathbf{M}(\mathbf{MVM})^{-}\mathbf{M}\mathbf{P}_{\mathbf{V}} = \mathbf{P}_{\mathbf{V}}\mathbf{M}(\mathbf{MVM})^{+}\mathbf{M}\mathbf{P}_{\mathbf{V}}\,. \tag{15.7}$$

The matrices $\dot{\mathbf{M}}$ and $\ddot{\mathbf{M}}$ appear to be very handy in many ways related to linear model $\mathscr{M} = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \mathbf{V}\}$; we meet them there repeatedly, sometimes even unexpectedly. In this chapter, we collect together various properties of $\dot{\mathbf{M}}$ and $\ddot{\mathbf{M}}$ and show several examples illustrating their usefulness in the context of linear models. Our presentation follows much that of Isotalo, Puntanen & Styan (2008d). We will by no means cover all applications of $\dot{\mathbf{M}}$; for example, we skip over the so-called restricted maximum likelihood (REML) method, introduced by Patterson & Thompson (1971); see also LaMotte (2007).

Theorem 15 (Matrix $\dot{\mathbf{M}}$). *Consider the linear model* $\mathscr{M} = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \mathbf{V}\}$, *and let the matrices* $\mathbf{M}$, $\dot{\mathbf{M}}$, *and* $\ddot{\mathbf{M}}$ *be defined as*

$$\mathbf{M} = \mathbf{I} - \mathbf{P}_{\mathbf{X}}\,, \qquad \dot{\mathbf{M}} = \mathbf{M}(\mathbf{MVM})^{-}\mathbf{M}\,, \qquad \ddot{\mathbf{M}} = \mathbf{P}_{\mathbf{V}}\dot{\mathbf{M}}\mathbf{P}_{\mathbf{V}}\,. \tag{15.8}$$

Assume that the condition

$$\mathbf{H}\mathbf{P}_{\mathbf{V}}\mathbf{M} = \mathbf{0} \tag{15.9}$$

holds. Then

(a) $\ddot{\mathbf{M}} = \mathbf{P}_{\mathbf{V}}\mathbf{M}(\mathbf{MVM})^{-}\mathbf{M}\mathbf{P}_{\mathbf{V}} = \mathbf{V}^{+} - \mathbf{V}^{+}\mathbf{X}(\mathbf{X}'\mathbf{V}^{+}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{+}$,

(b) $\ddot{\mathbf{M}} = \mathbf{M}\mathbf{V}^{+}\mathbf{M} - \mathbf{M}\mathbf{V}^{+}\mathbf{X}(\mathbf{X}'\mathbf{V}^{+}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{+}\mathbf{M}$,

(c) $\ddot{\mathbf{M}} = \mathbf{M}\ddot{\mathbf{M}} = \ddot{\mathbf{M}}\mathbf{M} = \mathbf{M}\ddot{\mathbf{M}}\mathbf{M}$,

(d) $\mathbf{M}(\mathbf{MVM})^{+}\mathbf{M} = (\mathbf{MVM})^{+}\mathbf{M} = \mathbf{M}(\mathbf{MVM})^{+} = (\mathbf{MVM})^{+}$,

(e) $\ddot{\mathbf{M}}\mathbf{V}\ddot{\mathbf{M}} = \ddot{\mathbf{M}}$, *i.e.*, $\mathbf{V} \in \{(\ddot{\mathbf{M}})^{-}\}$,

(f) $\operatorname{rk}(\ddot{\mathbf{M}}) = \operatorname{rk}(\mathbf{VM}) = \operatorname{rk}(\mathbf{V}) - \dim \mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{V}) = \operatorname{rk}(\mathbf{X} : \mathbf{V}) - \operatorname{rk}(\mathbf{X})$,

(g) *If* $\mathbf{Z}$ *is a matrix with property* $\mathscr{C}(\mathbf{Z}) = \mathscr{C}(\mathbf{M})$, *then*

$$\ddot{\mathbf{M}} = \mathbf{P}_{\mathbf{V}}\mathbf{Z}(\mathbf{Z}'\mathbf{V}\mathbf{Z})^{-}\mathbf{Z}'\mathbf{P}_{\mathbf{V}}\,, \qquad \mathbf{V}\dot{\mathbf{M}}\mathbf{V} = \mathbf{V}\mathbf{Z}(\mathbf{Z}'\mathbf{V}\mathbf{Z})^{-}\mathbf{Z}'\mathbf{V}\,. \tag{15.10}$$

(h) *Let* $(\mathbf{X} : \mathbf{Z})$ *be orthogonal. Then always* (*even if* $\mathbf{H}\mathbf{P}_{\mathbf{V}}\mathbf{M} = \mathbf{0}$ *does not hold*)

$$[(\mathbf{X} : \mathbf{Z})'\mathbf{V}(\mathbf{X} : \mathbf{Z})]^{+} = (\mathbf{X} : \mathbf{Z})'\mathbf{V}^{+}(\mathbf{X} : \mathbf{Z})\,. \tag{15.11}$$

Moreover, if in addition we have $\mathbf{H}\mathbf{P}_{\mathbf{V}}\mathbf{M} = \mathbf{0}$, *i.e.,* $\mathscr{C}(\mathbf{V}\mathbf{H}) \cap \mathscr{C}(\mathbf{V}\mathbf{M}) = \{\mathbf{0}\}$, *then*

$$\begin{aligned}
&[(\mathbf{X} : \mathbf{Z})'\mathbf{V}(\mathbf{X} : \mathbf{Z})]^{+} \\
&= \begin{pmatrix} [\mathbf{X}'\mathbf{V}\mathbf{X} - \mathbf{X}'\mathbf{V}\mathbf{Z}(\mathbf{Z}'\mathbf{V}\mathbf{Z})^{-}\mathbf{Z}'\mathbf{V}\mathbf{X}]^{+} & -\mathbf{X}'\mathbf{V}^{+}\mathbf{X}\mathbf{X}'\mathbf{V}\mathbf{Z}(\mathbf{Z}'\mathbf{V}\mathbf{Z})^{+} \\ -(\mathbf{Z}'\mathbf{V}\mathbf{Z})^{+}\mathbf{Z}'\mathbf{V}\mathbf{X}\mathbf{X}'\mathbf{V}^{+}\mathbf{X} & [\mathbf{Z}'\mathbf{V}\mathbf{Z} - \mathbf{Z}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{V}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}\mathbf{Z}]^{+} \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{X}'\mathbf{V}^{+}\mathbf{X} & \mathbf{X}'\mathbf{V}^{+}\mathbf{Z} \\ \mathbf{Z}'\mathbf{V}^{+}\mathbf{X} & \mathbf{Z}'\mathbf{V}^{+}\mathbf{Z} \end{pmatrix}. \qquad (15.12)
\end{aligned}$$

(i) *If* $\mathbf{V}$ *is positive definite and* $\mathscr{C}(\mathbf{Z}) = \mathscr{C}(\mathbf{M})$, *then*

$$\begin{aligned}
\dot{\mathbf{M}} = \ddot{\mathbf{M}} = \mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^{-}\mathbf{M} &= (\mathbf{M}\mathbf{V}\mathbf{M})^{+} = \mathbf{Z}(\mathbf{Z}'\mathbf{V}\mathbf{Z})^{-}\mathbf{Z}' \\
&= \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1} \\
&= \mathbf{V}^{-1}(\mathbf{I} - \mathbf{P}_{\mathbf{X};\mathbf{V}^{-1}}), \qquad (15.13) \\
\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}' &= \mathbf{V} - \mathbf{V}\mathbf{Z}(\mathbf{Z}'\mathbf{V}\mathbf{Z})^{-}\mathbf{Z}'\mathbf{V} \\
&= \mathbf{V} - \mathbf{V}\dot{\mathbf{M}}\mathbf{V}, \qquad (15.14) \\
\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1} &= \mathbf{I} - \mathbf{V}\mathbf{Z}(\mathbf{Z}'\mathbf{V}\mathbf{Z})^{-}\mathbf{Z}' \\
&= \mathbf{I} - \mathbf{V}\dot{\mathbf{M}} = \mathbf{I} - \mathbf{P}'_{\mathbf{Z};\mathbf{V}} \\
&= \mathbf{I} - \mathbf{P}_{\mathbf{V}\mathbf{Z};\mathbf{V}^{-1}}. \qquad (15.15)
\end{aligned}$$

(j) *If* $\mathbf{V}$ *is positive definite and* $(\mathbf{X} : \mathbf{Z})$ *is orthogonal, then*

$$(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} = \mathbf{X}'\mathbf{V}\mathbf{X} - \mathbf{X}'\mathbf{V}\mathbf{Z}(\mathbf{Z}'\mathbf{V}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{V}\mathbf{X}. \qquad (15.16)$$

Proof. Proceeding as Puntanen & Scott (1996, p. 323), we denote

$$\mathbf{A} = (\mathbf{V}^{+})^{1/2}\mathbf{X}, \quad \mathbf{B} = \mathbf{V}^{1/2}\mathbf{M}, \qquad (15.17)$$

where $\mathbf{V}^{1/2} = \mathbf{T}_1\boldsymbol{\Lambda}_1^{1/2}\mathbf{T}_1'$, $(\mathbf{V}^{+})^{1/2} = \mathbf{T}_1\boldsymbol{\Lambda}_1^{-1/2}\mathbf{T}_1'$; here $\boldsymbol{\Lambda}_1$ is a diagonal matrix of the nonzero (necessarily positive) eigenvalues of $\mathbf{V}$, and $\mathbf{T}_1$ is a matrix comprising the corresponding orthonormal eigenvectors. Now $\mathbf{V}^{1/2}(\mathbf{V}^{+})^{1/2} = \mathbf{T}_1\mathbf{T}_1' = \mathbf{P}_{\mathbf{V}}$, and hence, in view of Theorem 7 (p. 151),

$$\mathbf{P}_{\mathbf{A}} + \mathbf{P}_{\mathbf{B}} = \mathbf{P}_{(\mathbf{A} : \mathbf{B})} \qquad (15.18)$$

whenever the equation

$$\mathbf{A}'\mathbf{B} = \mathbf{X}'\mathbf{P}_{\mathbf{V}}\mathbf{M} = \mathbf{0} \qquad (15.19)$$

holds. Let us next confirm that if (15.19) holds, then

$$\mathscr{C}(\mathbf{A} : \mathbf{B}) = \mathscr{C}(\mathbf{V}). \qquad (15.20)$$

It is obvious that $\mathscr{C}(\mathbf{A} : \mathbf{B}) \subset \mathscr{C}(\mathbf{V})$, and that $\operatorname{rank}(\mathbf{A}) = \operatorname{rank}(\mathbf{V}\mathbf{X})$, $\operatorname{rank}(\mathbf{B}) = \operatorname{rank}(\mathbf{V}\mathbf{M})$. Moreover,

$$\operatorname{rank}[\mathbf{V}(\mathbf{X} : \mathbf{M})] = \operatorname{rank}[\mathbf{P}_{\mathbf{V}}(\mathbf{X} : \mathbf{M})] = \operatorname{rank}(\mathbf{V})\,, \tag{15.21}$$

from which, in view of (15.19), it follows that

$$\operatorname{rank}(\mathbf{P}_{\mathbf{V}}\mathbf{X}) + \operatorname{rank}(\mathbf{P}_{\mathbf{V}}\mathbf{M}) = \operatorname{rank}(\mathbf{V})\,, \tag{15.22}$$

and thereby $\operatorname{rank}(\mathbf{A}) + \operatorname{rank}(\mathbf{B}) = \operatorname{rank}(\mathbf{V})$, and so (15.20) indeed holds. Hence the equation (15.18) can be written as

$$\mathbf{P}_{\mathbf{A}} + \mathbf{P}_{\mathbf{B}} = \mathbf{P}_{\mathbf{V}}\,. \tag{15.23}$$

Substituting the explicit formulas of $\mathbf{P}_{\mathbf{A}}$ and $\mathbf{P}_{\mathbf{B}}$ into (15.23), we have

$$(\mathbf{V}^{+})^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{V}^{+}\mathbf{X})^{-}\mathbf{X}'(\mathbf{V}^{+})^{1/2} + \mathbf{V}^{1/2}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^{-}\mathbf{M}\mathbf{V}^{1/2} = \mathbf{P}_{\mathbf{V}}\,. \tag{15.24}$$

Pre- and postmultiplying (15.24) by $(\mathbf{V}^{+})^{1/2}$ gives

$$\mathbf{V}^{+}\mathbf{X}(\mathbf{X}'\mathbf{V}^{+}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{+} + \mathbf{P}_{\mathbf{V}}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^{-}\mathbf{M}\mathbf{P}_{\mathbf{V}} = \mathbf{V}^{+}, \tag{15.25}$$

i.e., we have obtained our claim (a):

$$\mathbf{P}_{\mathbf{V}}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^{-}\mathbf{M}\mathbf{P}_{\mathbf{V}} = \mathbf{V}^{+} - \mathbf{V}^{+}\mathbf{X}(\mathbf{X}'\mathbf{V}^{+}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{+}. \tag{15.26}$$

The claim (b) follows from the following (see Proposition 15.1, page 321):

$$\mathbf{H}\mathbf{P}_{\mathbf{V}}\mathbf{M} = \mathbf{0} \iff \mathbf{P}_{\mathbf{V}}\mathbf{M} = \mathbf{M}\mathbf{P}_{\mathbf{V}} \iff \mathbf{P}_{\mathbf{V}}\mathbf{M} = \mathbf{M}\mathbf{P}_{\mathbf{V}}\mathbf{M}\,. \tag{15.27}$$

Property (c) follows from (a) at once. To prove (d) we can use the identity

$$\mathbf{A}^{+} = (\mathbf{A}'\mathbf{A})^{+}\mathbf{A}' = \mathbf{A}'(\mathbf{A}\mathbf{A}')^{+}. \tag{15.28}$$

Now (15.28) gives, for example,

$$\begin{aligned} \mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^{+}\mathbf{M} &= \mathbf{M}[(\mathbf{M}\mathbf{V}\mathbf{M})' \cdot \mathbf{M}\mathbf{V}\mathbf{M}]^{+}(\mathbf{M}\mathbf{V}\mathbf{M})' \cdot \mathbf{M} \\ &= \mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^{+}. \end{aligned} \tag{15.29}$$

Property (e) is obvious, and (f) holds because

$$\begin{aligned} \operatorname{rank}[\mathbf{P}_{\mathbf{V}}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^{+}\mathbf{M}\mathbf{P}_{\mathbf{V}}] &= \operatorname{rank}[\mathbf{P}_{\mathbf{V}}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^{+}] \\ &= \operatorname{rank}[\mathbf{P}_{\mathbf{V}}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})] \\ &= \operatorname{rank}(\mathbf{P}_{\mathbf{V}}\mathbf{M}\mathbf{V}\mathbf{M}) \\ &= \operatorname{rank}(\mathbf{V}\mathbf{M}\mathbf{V}\mathbf{M}) = \operatorname{rank}(\mathbf{V}\mathbf{M})\,. \end{aligned} \tag{15.30}$$

Result (g) comes from the following:

$$\begin{aligned} \mathbf{P}_{\mathbf{V}}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^{-}\mathbf{M}\mathbf{P}_{\mathbf{V}} &= (\mathbf{V}^{+})^{1/2}\mathbf{P}_{\mathbf{V}^{1/2}\mathbf{M}}(\mathbf{V}^{+})^{1/2} \\ &= (\mathbf{V}^{+})^{1/2}\mathbf{P}_{\mathbf{V}^{1/2}\mathbf{Z}}(\mathbf{V}^{+})^{1/2} \end{aligned}$$

$$= \mathbf{P_V Z(Z'VZ)^- Z'P_V}\,. \tag{15.31}$$

The proof of (h) comes immediately from Proposition 13.1, (p. 294), which states that if a symmetric nonnegative definite matrix $\mathbf{A}$ is partitioned as $\mathbf{A} = \big(\begin{smallmatrix}\mathbf{A}_{11} & \mathbf{A}_{12}\\ \mathbf{A}_{21} & \mathbf{A}_{22}\end{smallmatrix}\big)$, then

$$\operatorname{rank}(\mathbf{A}) = \operatorname{rank}(\mathbf{A}_{11}) + \operatorname{rank}(\mathbf{A}_{22}) \tag{15.32a}$$

$$\iff$$

$$\mathbf{A}^+ = \begin{pmatrix} \mathbf{A}_{11\cdot 2}^+ & -\mathbf{A}_{11\cdot 2}^+\mathbf{A}_{12}\mathbf{A}_{22}^+ \\ -\mathbf{A}_{22}^+\mathbf{A}_{21}\mathbf{A}_{11\cdot 2}^+ & \mathbf{A}_{22}^+ + \mathbf{A}_{22}^+\mathbf{A}_{21}\mathbf{A}_{11\cdot 2}^+\mathbf{A}_{12}\mathbf{A}_{22}^+ \end{pmatrix}, \tag{15.32b}$$

where $\mathbf{A}_{11\cdot 2} = \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^-\mathbf{A}_{21}$ is the Schur complement of $\mathbf{A}_{22}$ in $\mathbf{A}$.

The results (i) and (j) are easy to obtain. □

It is helpful to give various equivalent characterizations for the important condition $\mathbf{HP_V M} = \mathbf{0}$.

Proposition 15.1. *Consider the linear model* $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$. *Then the following statements are equivalent:*

(a) $\mathbf{HP_V M} = \mathbf{0}$,

(b) $\mathbf{P_V M} = \mathbf{MP_V}$,

(c) $\mathscr{C}(\mathbf{P_V H}) \subset \mathscr{C}(\mathbf{H})$,

(d) $\mathscr{C}(\mathbf{VH}) \cap \mathscr{C}(\mathbf{VM}) = \{\mathbf{0}\}$,

(e) $\mathscr{C}(\mathbf{V}^{1/2}\mathbf{H}) \cap \mathscr{C}(\mathbf{V}^{1/2}\mathbf{M}) = \{\mathbf{0}\}$,

(f) $\mathscr{C}(\mathbf{X}) = \mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{V}) \boxplus \mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{V})^{\perp}$,

(g) $u = \dim \mathscr{C}(\mathbf{VH}) \cap \mathscr{C}(\mathbf{VM}) = \operatorname{rank}(\mathbf{HP_V M}) = 0$, *where* u *is the number of unit canonical correlations between* $\mathbf{Hy}$ *and* $\mathbf{My}$.

The equivalence of (a)–(c) follows from Proposition 10.1 (p. 218) by replacing $\mathbf{V}$ with $\mathbf{P_V}$. For other parts, see Proposition 5.5 (p. 135).

We now present a generalized version of Theorem 15 as a proposition. The proof when $\mathbf{W}$ is (symmetric) nonnegative definite follows along the lines of Theorem 15 with obvious changes. The general case, however, is a bit more complicated.

Proposition 15.2 (Matrix $\dot{\mathbf{M}}$: general case). *Consider the linear model* $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$. *Let* $\mathbf{U}$ *be any* $p \times p$ *matrix such that the matrix* $\mathbf{W} = \mathbf{V} + \mathbf{XUX'}$ *satisfies the condition* $\mathscr{C}(\mathbf{W}) = \mathscr{C}(\mathbf{X} : \mathbf{V})$. *Then*

$$\mathbf{P_W M(MVM)^- MP_W} = \mathbf{(MVM)}^+ \tag{15.33a}$$

$$= \mathbf{W}^+ - \mathbf{W}^+\mathbf{X}(\mathbf{X'W^-X})^-\mathbf{X'W}^+, \tag{15.33b}$$

that is,

$$\mathbf{P_W}\dot{\mathbf{M}}\mathbf{P_W} := \ddot{\mathbf{M}}_{\mathrm{W}} = \mathbf{W}^+ - \mathbf{W}^+\mathbf{X}(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}'\mathbf{W}^+. \tag{15.34}$$

Moreover, the matrix $\ddot{\mathbf{M}}_{\mathrm{W}}$ *has the corresponding properties as* $\ddot{\mathbf{M}}$ *in Theorem* 15.

Proof. Equality (15.33a) follows by observing that $\mathbf{P_W} = \mathbf{H} + \mathbf{P_{MVM}}$. We note that if $\mathbf{W}$ is a symmetric nonnegative definite matrix, then substituting $\mathbf{A} = (\mathbf{W}^+)^{1/2}\mathbf{X}$ and $\mathbf{B} = \mathbf{W}^{1/2}\mathbf{M}$ into (15.17), and proceeding as in the proof of Theorem 15, we can introduce the decomposition (15.33b).

Consider now the general case. Let us denote

$$\mathbf{G}_1 = \mathbf{I}_n - \mathbf{VM}(\mathbf{MVM})^-\mathbf{M}\,, \quad \mathbf{G}_2 = \mathbf{X}(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}'\mathbf{W}^-. \tag{15.35}$$

Now both $\mathbf{G}_1$ and $\mathbf{G}_2$ satisfy the fundamental BLUE equation, i.e.,

$$(\mathbf{G}_1 - \mathbf{G}_2)(\mathbf{X} : \mathbf{VM}) = \mathbf{0}\,, \tag{15.36}$$

which is equivalent to

$$\mathscr{C}(\mathbf{G}_1' - \mathbf{G}_2') \subset \mathscr{C}(\mathbf{X} : \mathbf{VM})^\perp = \mathscr{C}(\mathbf{W})^\perp, \tag{15.37}$$

i.e., there exists a matrix $\mathbf{F}'$ such that

$$\mathbf{G}_1' - \mathbf{G}_2' = (\mathbf{I}_n - \mathbf{P_W})\mathbf{F}'. \tag{15.38}$$

Postmultiplying the equation

$$\mathbf{X}(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}'\mathbf{W}^- = \mathbf{I}_n - \mathbf{VM}(\mathbf{MVM})^-\mathbf{M} + \mathbf{F}(\mathbf{I} - \mathbf{P_W}) \tag{15.39}$$

by $\mathbf{W}$ yields

$$\begin{aligned}\mathbf{X}(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}' &= \mathbf{W} - \mathbf{VM}(\mathbf{MVM})^-\mathbf{MV} \\ &= \mathbf{W} - \mathbf{WM}(\mathbf{MVM})^-\mathbf{MW}\,.\end{aligned} \tag{15.40}$$

Pre- and postmultiplying (15.40) by $\mathbf{W}^+$ yields

$$\mathbf{W}^+\mathbf{X}(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}'\mathbf{W}^+ = \mathbf{W}^+ - \mathbf{P_{W'}}\mathbf{M}(\mathbf{MVM})^-\mathbf{MP_W}\,, \tag{15.41}$$

In view of Proposition 12.1 (p. 286), we have

$$\mathscr{C}(\mathbf{X} : \mathbf{V}) = \mathscr{C}(\mathbf{W}) \iff \mathscr{C}(\mathbf{X} : \mathbf{V}) = \mathscr{C}(\mathbf{W}') \tag{15.42}$$

and hence in this situation $\mathbf{P_{W'}} = \mathbf{P_W}$, and thereby (15.41) implies our claim (15.33b). □

We complete this section by mentioning some references which are related to Theorem 15 and Proposition 15.2. Let us first recall that when $\mathbf{X}$ has full column rank and $\mathbf{V}$ is positive definite, then the OLSE and BLUE of

$\boldsymbol{\beta}$ under $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$ are, respectively, $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, and $\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$, while the corresponding covariance matrices are

$$\operatorname{cov}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}, \quad \operatorname{cov}(\tilde{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}. \tag{15.43}$$

Hence we have the Löwner ordering

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \geq_{\mathsf{L}} (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}, \tag{15.44}$$

i.e., the matrix

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} := \mathbf{D} \tag{15.45}$$

is nonnegative definite. What is interesting here in (15.45), is that the explicit alternative expression for the nonnegative definite matrix $\mathbf{D}$ is available. Namely, according to part (i) of Theorem 15 (p. 318),

$$\begin{aligned} \operatorname{cov}(\tilde{\boldsymbol{\beta}}) &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}[\mathbf{X}'\mathbf{V}\mathbf{X} - \mathbf{X}'\mathbf{V}\mathbf{Z}(\mathbf{Z}'\mathbf{V}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{V}\mathbf{X}](\mathbf{X}'\mathbf{X})^{-1} \\ &= \mathbf{X}^{+}[\mathbf{V} - \mathbf{V}\mathbf{Z}(\mathbf{Z}'\mathbf{V}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{V}](\mathbf{X}^{+})' \\ &= \operatorname{cov}(\hat{\boldsymbol{\beta}}) - \mathbf{X}^{+}\mathbf{V}\mathbf{Z}(\mathbf{Z}'\mathbf{V}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{V}(\mathbf{X}^{+})', \end{aligned} \tag{15.46}$$

and hence

$$\operatorname{cov}(\mathbf{X}\tilde{\boldsymbol{\beta}}) = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}' = \mathbf{H}\mathbf{V}\mathbf{H} - \mathbf{H}\mathbf{V}\mathbf{Z}(\mathbf{Z}'\mathbf{V}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{V}\mathbf{H}. \tag{15.47}$$

Among the first places where (15.46) occurs are probably the papers by Khatri (1966, Lemma 1) and Rao (1967, Lemmas 2a, 2b and 2c); see also Rao (1973a, Problem 33, p. 77).

Srivastava & Khatri (1979, Cor. 1.9) present part (a) of Theorem 15 when $\mathbf{V}$ is positive definite. Khatri (1982, Lemma 6) generalized the considerations towards our Proposition 15.2 (using a different proof).

Other related references include de Hoog, Speed & Williams (1990), Koch (1969), Khatri (1990, Th. 1), Puntanen (1997), Bhimasankaram & Saha Ray (1997, Lemma 2.1), Satorra & Neudecker (2003, Lemma 1), and Searle, Casella & McCulloch (1992, pp. 451–452).

The most general formulation related to our discussion, is given, according to our understanding, in Baksalary, Puntanen & Styan (1990b, Th. 3):

Proposition 15.3. *Consider the model* $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$ *and let* $\mathbf{U}$ *be such that* $\mathbf{W} = \mathbf{V} + \mathbf{X}\mathbf{U}\mathbf{X}'$ *satisfies* $\mathscr{C}(\mathbf{W}) = \mathscr{C}(\mathbf{X} : \mathbf{V})$. *Then the equality*

$$\mathbf{W} = \mathbf{V}\mathbf{B}(\mathbf{B}'\mathbf{V}\mathbf{B})^{-}\mathbf{B}'\mathbf{V} + \mathbf{X}(\mathbf{X}'\mathbf{W}^{-}\mathbf{X})^{-}\mathbf{X}' \tag{15.48}$$

holds for an $n \times p$ *matrix* $\mathbf{B}$ *if and only if*

$$\mathscr{C}(\mathbf{V}\mathbf{W}^{-}\mathbf{X}) \subset \mathscr{C}(\mathbf{B})^{\perp} \text{ and } \mathscr{C}(\mathbf{V}\mathbf{M}) \subset \mathscr{C}(\mathbf{V}\mathbf{B}), \tag{15.49}$$

or, equivalently,

$$\mathscr{C}(\mathbf{V}\mathbf{W}^{-}\mathbf{X}) = \mathscr{C}(\mathbf{B})^{\perp} \cap \mathscr{C}(\mathbf{V}), \tag{15.50}$$

the subspace $\mathscr{C}(\mathbf{V}\mathbf{W}^{-}\mathbf{X})$ *being independent of the choice of* $\mathbf{W}^{-}$.

15.1 Representations for the BLUE

Consider first a weakly singular model $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$, where

$$\mathscr{C}(\mathbf{X}) \subset \mathscr{C}(\mathbf{V}). \tag{15.51}$$

Now (15.51) immediately implies $\mathbf{H}\mathbf{P}_{\mathbf{V}}\mathbf{M} = \mathbf{0}$ and hence

$$\mathbf{P}_{\mathbf{V}}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^{-}\mathbf{M}\mathbf{P}_{\mathbf{V}} = \mathbf{V}^{+} - \mathbf{V}^{+}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{+}. \tag{15.52}$$

Premultiplying (15.52) by $\mathbf{V}$ and using (15.51) yields

$$\begin{aligned}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{+} &= \mathbf{P}_{\mathbf{V}} - \mathbf{V}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^{-}\mathbf{M}\mathbf{P}_{\mathbf{V}} \\ &= \mathbf{P}_{\mathbf{V}} - \mathbf{V}\dot{\mathbf{M}}\mathbf{P}_{\mathbf{V}}.\end{aligned} \tag{15.53}$$

It is important to note that (15.53) is a “plain” matrix equality that holds if $\mathscr{C}(\mathbf{X}) \subset \mathscr{C}(\mathbf{V})$ holds. Now the consistency of the model $\mathscr{M}$ implies that (with probability 1) $\mathbf{y} \in \mathscr{C}(\mathbf{X} : \mathbf{V})$, which now becomes $\mathbf{y} \in \mathscr{C}(\mathbf{V})$. Postmultiplying (15.53) by $\mathbf{y}$ yields the following:

Proposition 15.4. *Under a weakly singular linear model* $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$, *for every* $\mathbf{y} \in \mathscr{C}(\mathbf{V})$, *we have,*

$$\mathbf{X}(\mathbf{X}'\mathbf{V}^{-}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-}\mathbf{y} = \mathbf{y} - \mathbf{V}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^{-}\mathbf{M}\mathbf{y} = \mathbf{y} - \mathbf{V}\dot{\mathbf{M}}\mathbf{y}. \tag{15.54}$$

It is worth emphasizing that (15.54) is a *statistical* equation; it does not hold for every $\mathbf{y}$, but we know that the realized value of $\mathbf{y}$ belongs $\mathscr{C}(\mathbf{V})$, and that is enough when dealing with a weakly singular linear model.

In Proposition 15.4 we have two representations for the BLUE of $\mathbf{X}\boldsymbol{\beta}$; note that even though $\mathbf{A}\mathbf{y}$ and $\mathbf{B}\mathbf{y}$ were the BLUEs, the multipliers $\mathbf{A}$ and $\mathbf{B}$ need not be equal. The left-hand representation in (15.54) is sometimes called the Zyskind–Martin estimator, cf. Zyskind & Martin (1969), while the right-hand formula is a general representation in the sense that it is valid without any further assumptions. Note that the Zyskind–Martin estimator is invariant with respect to the choice of the generalized inverses involved;see also Proposition 6.1 (p. 147).

The matrices

$$\mathbf{G}_1 = \mathbf{I} - \mathbf{V}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^{-}\mathbf{M} = \mathbf{I} - \mathbf{V}\dot{\mathbf{M}}, \tag{15.55a}$$

$$\mathbf{G}_2 = \mathbf{X}(\mathbf{X}'\mathbf{W}^{-}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}^{-}, \tag{15.55b}$$

satisfy

$$\mathbf{G}_i(\mathbf{X} : \mathbf{VM}) = (\mathbf{X} : \mathbf{0}) , \quad i = 1, 2 , \tag{15.56}$$

and hence $\mathbf{G}_1\mathbf{y}$ and $\mathbf{G}_2\mathbf{y}$ are general expressions for BLUE($\mathbf{X}\boldsymbol{\beta}$) in the sense that they are valid without any rank conditions. Note that

$$\mathbf{y} - \mathbf{G}_1\mathbf{y} = \mathbf{V}\dot{\mathbf{M}}\mathbf{y} = \tilde{\boldsymbol{\varepsilon}} \tag{15.57}$$

is a general representation for the "raw" residual of the BLUE($\mathbf{X}\boldsymbol{\beta}$).

Premultiplying (15.56) by $\mathbf{H}$, we see immediately that $\mathbf{HG}_1 := \mathbf{G}_3$ provides one further representation for the BLUE of $\mathbf{X}\boldsymbol{\beta}$:

$$\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta}) = \mathbf{Hy} - \mathbf{HVM}(\mathbf{MVM})^{-}\mathbf{My} = \mathbf{G}_3\mathbf{y} , \tag{15.58}$$

and so the difference between the OLSE($\mathbf{X}\boldsymbol{\beta}$) and BLUE($\mathbf{X}\boldsymbol{\beta}$) is

$$\mathrm{OLSE}(\mathbf{X}\boldsymbol{\beta}) - \mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta}) = \mathbf{HVM}(\mathbf{MVM})^{-}\mathbf{My} = \mathbf{HV}\dot{\mathbf{M}}\mathbf{y} . \tag{15.59}$$

The expressions $\mathbf{G}_1$, $\mathbf{G}_2$ and $\mathbf{G}_3$ appear, e.g., in Albert (1973, p. 182), Drygas (1970, p. 78), and Rao (1973b, p. 285). In passing we may mention that utilizing (15.59), Baksalary & Kala (1980) provided an upper bound for the Euclidean norm $\|\mathrm{OLSE}(\mathbf{X}\boldsymbol{\beta}) - \mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta})\|$ in terms of the singular values of $\mathbf{HVM}$, eigenvalues of $\mathbf{MVM}$ and the residual vector $\mathbf{My}$; see also Haberman (1975), Baksalary & Kala (1978a), and Trenkler (1994, p. 261).

Premultiplying the equation

$$\mathbf{P}_{\mathbf{W}}\mathbf{M}(\mathbf{MVM})^{-}\mathbf{MP}_{\mathbf{W}} = \mathbf{W}^{+} - \mathbf{W}^{+}\mathbf{X}(\mathbf{X}'\mathbf{W}^{-}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}^{+} \tag{15.60}$$

by $\mathbf{W}$ yields

$$\mathbf{X}(\mathbf{X}'\mathbf{W}^{-}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}^{+} = \mathbf{P}_{\mathbf{W}} - \mathbf{VM}(\mathbf{MVM})^{-}\mathbf{MP}_{\mathbf{W}} , \tag{15.61}$$

and hence for every $\mathbf{y} \in \mathscr{C}(\mathbf{X} : \mathbf{V}) = \mathscr{C}(\mathbf{W})$, we have

$$\mathbf{X}(\mathbf{X}'\mathbf{W}^{-}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}^{-}\mathbf{y} = \mathbf{y} - \mathbf{V}\dot{\mathbf{M}}\mathbf{y} . \tag{15.62}$$

Equality (15.61) is a "plain" mathematical matrix equality (which always holds), while (15.62) is a "statistical" equality which necessarily holds (only) if $\mathbf{y} \in \mathscr{C}(\mathbf{X} : \mathbf{V})$. Note also that premultiplying (15.61) with $\mathbf{H}$ yields another plain matrix decomposition:

$$\mathbf{X}(\mathbf{X}'\mathbf{W}^{-}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}^{+} = \mathbf{H} - \mathbf{HVM}(\mathbf{MVM})^{-}\mathbf{MP}_{\mathbf{W}} . \tag{15.63}$$

15.2 The BLUE's Covariance Matrix

Under the linear model $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$,

$$\begin{aligned}\operatorname{cov}[\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta})] &= \mathbf{HVH} - \mathbf{HVM}(\mathbf{MVM})^-\mathbf{MVH} \\ &= \mathbf{V} - \mathbf{VM}(\mathbf{MVM})^-\mathbf{MV} = \mathbf{V} - \mathbf{V}\dot{\mathbf{M}}\mathbf{V}\,, \end{aligned} \tag{15.64}$$

and thereby, see Exercise 0.15 (p. 50),

$$\operatorname{cov}(\mathbf{Hy} - \mathbf{Gy}) = \operatorname{cov}(\mathbf{Hy}) - \operatorname{cov}(\mathbf{Gy}) = \mathbf{HVM}(\mathbf{MVM})^-\mathbf{MVH}\,, \tag{15.65}$$

and

$$\begin{aligned}\operatorname{cov}(\mathbf{y} - \mathbf{Gy}) &= \operatorname{cov}(\tilde{\boldsymbol{\varepsilon}}) = \operatorname{cov}(\mathbf{y}) - \operatorname{cov}(\mathbf{Gy}) \\ &= \mathbf{VM}(\mathbf{MVM})^-\mathbf{MV} = \mathbf{V}\dot{\mathbf{M}}\mathbf{V}\,, \end{aligned} \tag{15.66}$$

where $\mathbf{Gy} = \mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta})$ and $\tilde{\boldsymbol{\varepsilon}}$ is the BLUE's residual

$$\tilde{\boldsymbol{\varepsilon}} = \mathbf{VM}(\mathbf{MVM})^-\mathbf{My} = \mathbf{V}\dot{\mathbf{M}}\mathbf{y}\,. \tag{15.67}$$

In particular, if $\mathbf{HP_V M} = \mathbf{0}$, then according to Theorem 15 (p. 318),

$$\mathbf{P_V M}(\mathbf{MVM})^-\mathbf{MP_V} = \mathbf{V}^+ - \mathbf{V}^+\mathbf{X}(\mathbf{X}'\mathbf{V}^+\mathbf{X})^-\mathbf{X}'\mathbf{V}^+. \tag{15.68}$$

Pre- and postmultiplying (15.68) by $\mathbf{V}$ implies

$$\mathbf{P_V X}(\mathbf{X}'\mathbf{V}^+\mathbf{X})^-\mathbf{X}'\mathbf{P_V} = \mathbf{V} - \mathbf{VM}(\mathbf{MVM})^-\mathbf{MV}\,. \tag{15.69}$$

It is straightforward to confirm that (15.69) correspondingly implies that $\mathbf{HP_V M} = \mathbf{0}$. Actually it is possible to prove the following result, see, e.g., Puntanen & Scott (1996, Th. 2.6)

Proposition 15.5. *Consider the linear model $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$, and let $\mathbf{X}\tilde{\boldsymbol{\beta}}$ be the* BLUE *of $\mathbf{X}\boldsymbol{\beta}$. Then the following statements are equivalent:*

(a) $\mathbf{HP_V M} = \mathbf{0}$,

(b) *there are no unit canonical correlations between $\mathbf{Hy}$ and $\mathbf{My}$,*

(c) $\operatorname{cov}(\mathbf{X}\tilde{\boldsymbol{\beta}}) = \mathbf{X}_*(\mathbf{X}_*'\mathbf{V}^+\mathbf{X}_*)^+\mathbf{X}_*'$,

(d) $\operatorname{cov}(\mathbf{X}\tilde{\boldsymbol{\beta}}) = \mathbf{H}(\mathbf{HV}^+\mathbf{H})^+\mathbf{H} = (\mathbf{HV}^+\mathbf{H})^+$,

(e) $\operatorname{cov}(\mathbf{X}\tilde{\boldsymbol{\beta}}) = \mathbf{P_V H}(\mathbf{HV}^+\mathbf{H})^-\mathbf{HP_V}$,

(f) $\operatorname{cov}(\mathbf{X}\tilde{\boldsymbol{\beta}}) = \mathbf{P_V X}_*(\mathbf{X}_*'\mathbf{V}^+\mathbf{X}_*)^-\mathbf{X}_*'\mathbf{P_V}$,

(g) $\operatorname{cov}(\mathbf{X}\tilde{\boldsymbol{\beta}}) = \mathbf{P_V X}(\mathbf{X}'\mathbf{V}^+\mathbf{X})^-\mathbf{X}'\mathbf{P_V}$,

where $\mathbf{X}_*$ is a matrix whose columns form an orthonormal basis for $\mathscr{C}(\mathbf{X})$.

We may also note that the decomposition (15.33b) (p. 321) can be written equivalently as in (15.40) (p. 322):

$$\mathbf{X}(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}' = \mathbf{W} - \mathbf{VM}(\mathbf{MVM})^-\mathbf{MV}, \tag{15.70}$$

where $\mathbf{W} = \mathbf{V} + \mathbf{XUX}'$, and hence the BLUE's covariance matrix has the representation

$$\mathbf{X}(\mathbf{X}'\mathbf{W}^{-}\mathbf{X})^{-}\mathbf{X}' - \mathbf{XUX}' = \mathbf{V} - \mathbf{V}\dot{\mathbf{M}}\mathbf{V} = \operatorname{cov}(\mathbf{X}\tilde{\boldsymbol{\beta}})\,. \tag{15.71}$$

Choosing $\mathbf{U}$ nonnegative definite, i.e., $\mathbf{U} = \mathbf{LL}'$, (15.71) implies the Löwner ordering

$$\operatorname{cov}(\mathbf{X}\tilde{\boldsymbol{\beta}}) \leq_{\mathsf{L}} \mathbf{X}[\mathbf{X}'(\mathbf{V} + \mathbf{XUX}')^{-}\mathbf{X}]^{-}\mathbf{X}' \tag{15.72}$$

for all nonnegative definite $\mathbf{U}$ such that $\mathscr{C}(\mathbf{V} + \mathbf{XUX}') = \mathscr{C}(\mathbf{X} : \mathbf{V})$.

15.3 Partitioned Linear Model

Consider the full rank partitioned linear model

$$\mathscr{M}_{12} = \{\mathbf{y},\, \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2,\, \mathbf{V}\}\,, \tag{15.73}$$

where $\mathbf{X}_1$ and $\mathbf{X}_2$ are $n \times p_1$ and $n \times p_2$ matrices, respectively. Denoting

$$\begin{aligned}\dot{\mathbf{M}}_1 &= \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}_1(\mathbf{X}_1'\mathbf{V}^{-1}\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{V}^{-1} \\ &= \mathbf{V}^{-1}(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_1;\mathbf{V}^{-1}}) = \mathbf{M}_1(\mathbf{M}_1\mathbf{VM}_1)^{-}\mathbf{M}_1\,, \end{aligned} \tag{15.74}$$

we obtain the well-known result:

Proposition 15.6. *Consider the linear model* $\mathscr{M}_{12} = \{\mathbf{y},\, \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2,\, \mathbf{V}\}$, *where* $\mathbf{X}$ *has full column rank and* $\mathbf{V}$ *is positive definite. Then*

$$\operatorname{BLUE}(\boldsymbol{\beta}_2) = (\mathbf{X}_2'\dot{\mathbf{M}}_1\mathbf{X}_2)^{-1}\mathbf{X}_2'\dot{\mathbf{M}}_1\mathbf{y} = \tilde{\boldsymbol{\beta}}_2(\mathscr{M}_{12})\,, \tag{15.75}$$

where $\dot{\mathbf{M}}_1$ *is defined as in* (15.74).

Proof. Observe first that in this situation $\operatorname{rank}(\mathbf{X}_2'\dot{\mathbf{M}}_1\mathbf{X}_2) = p_2$:

$$\begin{aligned}\operatorname{rank}(\mathbf{X}_2'\dot{\mathbf{M}}_1\mathbf{X}_2) &= \operatorname{rank}[\mathbf{X}_2'\mathbf{M}_1(\mathbf{M}_1\mathbf{VM}_1)^{+}\mathbf{M}_1\mathbf{X}_2] \\ &= \operatorname{rank}[\mathbf{X}_2'\mathbf{M}_1(\mathbf{M}_1\mathbf{VM}_1)] \\ &= \operatorname{rank}(\mathbf{X}_2'\mathbf{M}_1\mathbf{V}) = \operatorname{rank}(\mathbf{X}_2'\mathbf{M}_1) = \operatorname{rank}(\mathbf{X}_2)\,. \end{aligned} \tag{15.76}$$

Denoting $\mathbf{N} = (\mathbf{X}_2'\dot{\mathbf{M}}_1\mathbf{X}_2)^{-1}\mathbf{X}_2'\dot{\mathbf{M}}_1 \in \mathbb{R}^{p_2 \times n}$, we observe, on account of (10.5) (p. 217), that (15.75) holds if and only if $\mathbf{N}$ satisfies the equation

$$\mathbf{N}(\mathbf{X}_1 : \mathbf{X}_2 : \mathbf{VM}) = (\mathbf{0} : \mathbf{I}_{p_2} : \mathbf{0})\,. \tag{15.77}$$

The equation $\mathbf{N}(\mathbf{X}_1 : \mathbf{X}_2) = (\mathbf{0} : \mathbf{I}_{p_2})$ is obvious. It remains to show that $\mathbf{X}_2'\dot{\mathbf{M}}_1\mathbf{VM} = \mathbf{0}$, which follows from (15.74) at once. □

Notice that Proposition 15.6 could of course be proved by using the rules of inverting a partitioned positive definite matrix.

We recall that

$$\mathrm{OLSE}(\boldsymbol{\beta}_2) = (\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{M}_1\mathbf{y} = \hat{\boldsymbol{\beta}}_2(\mathscr{M}_{12})\,, \tag{15.78}$$

and so we have a striking correspondence between the representations of the OLSE and BLUE of $\boldsymbol{\beta}_2$. Representation (15.75) is one of the most useful results due to Theorem 15.

15.4 The Frisch–Waugh–Lovell Theorem

Let us consider the partitioned linear model $\mathscr{M}_{12} = \{\mathbf{y},\, \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2,\, \mathbf{V}\}$. Premultiplying $\mathscr{M}_{12}$ by the orthogonal projector $\mathbf{M}_1$ yields the reduced model:

$$\mathscr{M}_{12\cdot 1} = \{\mathbf{M}_1\mathbf{y},\, \mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2,\, \mathbf{M}_1\mathbf{V}\mathbf{M}_1\}\,. \tag{15.79}$$

Obviously the OLS estimators of $\mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2$ under $\mathscr{M}_{12}$ and $\mathscr{M}_{12\cdot 1}$ coincide, see Section 8.4 (p. 163):

$$\mathbf{M}_1\mathbf{X}_2\hat{\boldsymbol{\beta}}_2(\mathscr{M}_{12}) = \mathbf{M}_1\mathbf{X}_2\hat{\boldsymbol{\beta}}_2(\mathscr{M}_{12\cdot 1}) = \mathbf{M}_1\mathbf{X}_2(\mathbf{X}_2'\mathbf{M}_2\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{M}_2\mathbf{y}\,. \tag{15.80}$$

What about the BLUE of $\mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2$ in the reduced model $\mathscr{M}_{12\cdot 1}$?

Before proceeding we notice that in view of Proposition 16.1 (p. 345), the following statements are equivalent:

$$\mathbf{K}_2'\boldsymbol{\beta}_2 \text{ is estimable under } \mathscr{M}_{12}, \tag{15.81a}$$

$$\mathscr{C}(\mathbf{K}_2) \subset \mathscr{C}(\mathbf{X}_2'\mathbf{M}_1)\,, \tag{15.81b}$$

$$\mathbf{K}_2' = \mathbf{L}\mathbf{M}_1\mathbf{X}_2 \quad \text{for some } \mathbf{L}\,. \tag{15.81c}$$

Moreover, it is easy to confirm that $\mathbf{K}_2'\boldsymbol{\beta}_2$ is estimable under $\mathscr{M}_{12}$ if and only if $\mathbf{K}_2'\boldsymbol{\beta}_2$ is estimable under $\mathscr{M}_{12\cdot 1}$.

Let us denote

$$\{\,\mathrm{BLUE}(\mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2 \mid \mathscr{M}_{12})\,\} = \{\,\mathbf{A}\mathbf{y} : \mathbf{A}\mathbf{y} \text{ is BLUE for } \mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2\,\}\,. \tag{15.82}$$

Then we can formulate the generalized Frisch–Waugh–Lovell theorem as follows; see, e.g., Groß & Puntanen (2000a, Th. 4).

Proposition 15.7 (Generalized Frisch–Waugh–Lovell Theorem). *Every representation of the* BLUE *of* $\mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2$ *under* $\mathscr{M}_{12}$ *remains the* BLUE *under* $\mathscr{M}_{12\cdot 1}$ *and vice versa, i.e., the sets of the* BLUE*s coincide:*

$$\{\,\mathrm{BLUE}(\mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2 \mid \mathscr{M}_{12})\,\} = \{\,\mathrm{BLUE}(\mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2 \mid \mathscr{M}_{12\cdot 1})\,\}\,. \tag{15.83}$$

In other words: Let $\mathbf{K}_2'\boldsymbol{\beta}_2$ *be an arbitrary estimable parametric function under* $\mathscr{M}_{12}$*. Then every representation of the* BLUE *of* $\mathbf{K}_2'\boldsymbol{\beta}_2$ *under* $\mathscr{M}_{12}$ *remains the* BLUE *under* $\mathscr{M}_{12\cdot 1}$ *and vice versa.*

Proof. Assume that $\mathbf{Fy}$ is the BLUE for $\mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2$ under $\mathscr{M}_{12}$, which means that $\mathbf{F}$ is an arbitrary solution to

$$\mathbf{F}(\mathbf{X}_1 : \mathbf{X}_2 : \mathbf{VM}) = (\mathbf{0} : \mathbf{M}_1\mathbf{X}_2 : \mathbf{0})\,. \tag{15.84}$$

We first show that (15.84) is equivalent to

$$\mathbf{F} = \mathbf{AM}_1\,, \text{ where } \mathbf{A}(\mathbf{M}_1\mathbf{X}_2 : \mathbf{M}_1\mathbf{VM}_1\mathbf{Q}_{\mathbf{M}_1\mathbf{X}_2}) = (\mathbf{M}_1\mathbf{X}_2 : \mathbf{0})\,. \tag{15.85}$$

Now $\mathbf{FX}_1 = \mathbf{0}$ holds if and only if $\mathbf{F} = \mathbf{AM}_1$ for some matrix $\mathbf{A}$. Hence (15.84) holds if and only if $\mathbf{F}$ can be expressed as $\mathbf{F} = \mathbf{AM}_1$ for some $\mathbf{A}$, where $\mathbf{A}$ satisfies

$$\mathbf{AM}_1(\mathbf{X}_2 : \mathbf{VM}) = (\mathbf{M}_1\mathbf{X}_2 : \mathbf{0})\,. \tag{15.86}$$

Because, cf. (8.7) (p. 156),

$$\mathbf{M} = \mathbf{I}_n - \mathbf{P}_{(\mathbf{X}_1 : \mathbf{X}_2)} = \mathbf{I}_n - \mathbf{P}_{\mathbf{X}_1} - \mathbf{P}_{\mathbf{M}_1\mathbf{X}_2} = \mathbf{M}_1\mathbf{Q}_{\mathbf{M}_1\mathbf{X}_2}\,, \tag{15.87}$$

we can rewrite (15.86) as

$$\mathbf{AM}_1(\mathbf{X}_2 : \mathbf{VM}_1\mathbf{Q}_{\mathbf{M}_1\mathbf{X}_2}) = (\mathbf{M}_1\mathbf{X}_2 : \mathbf{0})\,, \tag{15.88}$$

i.e.,

$$\mathbf{A}(\mathbf{M}_1\mathbf{X}_2 : \mathbf{M}_1\mathbf{VM}_1\mathbf{Q}_{\mathbf{M}_1\mathbf{X}_2}) = (\mathbf{M}_1\mathbf{X}_2 : \mathbf{0})\,, \tag{15.89}$$

which confirms the equivalence of (15.84) and (15.85).

Obviously (15.89) can be written as

$$\mathbf{AM}_1(\mathbf{M}_1\mathbf{X}_2 : \mathbf{M}_1\mathbf{VM}_1\mathbf{Q}_{\mathbf{M}_1\mathbf{X}_2}) = (\mathbf{M}_1\mathbf{X}_2 : \mathbf{0})\,, \tag{15.90}$$

which means that $\mathbf{AM}_1\mathbf{y} = \mathbf{Fy}$ is the BLUE for $\mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2$ under the reduced model $\mathscr{M}_{12\cdot 1} = \{\mathbf{M}_1\mathbf{y},\, \mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2,\, \mathbf{M}_1\mathbf{VM}_1\}$ and so we have shown that

$$\{\,\mathrm{BLUE}(\mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2 \mid \mathscr{M}_{12})\,\} \subset \{\,\mathrm{BLUE}(\mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2 \mid \mathscr{M}_{12\cdot 1})\,\}\,. \tag{15.91}$$

To prove the reverse inclusion in (15.91), we suppose that $\mathbf{BM}_1\mathbf{y}$ is an arbitrary representation for the BLUE of $\mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2$ under $\mathscr{M}_{12\cdot 1}$, i.e., $\mathbf{B}$ is an arbitrary solution to

$$\mathbf{B}(\mathbf{M}_1\mathbf{X}_2 : \mathbf{M}_1\mathbf{VM}_1\mathbf{Q}_{\mathbf{M}_1\mathbf{X}_2}) = (\mathbf{M}_1\mathbf{X}_2 : \mathbf{0})\,. \tag{15.92}$$

Then, in light of the equivalence of (15.84) and (15.85), $\mathbf{BM}_1\mathbf{y}$ is also the BLUE for $\mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2$ under $\mathscr{M}_{12}$. □

Let us next consider the explicit expression for the BLUE of $\mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2$ under $\mathscr{M}_{12}$; these being the same as under $\mathscr{M}_{12\cdot 1}$. Applying part (b) of Proposition 10.5 (p. 228), one representation for the BLUE of $\mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2$ under $\mathscr{M}_{12\cdot1} = \{\mathbf{M}_1\mathbf{y},\, \mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2,\, \mathbf{M}_1\mathbf{V}\mathbf{M}_1\}$ is $\mathbf{A}\mathbf{M}_1\mathbf{y}$, where

$$\mathbf{A} = \mathbf{M}_1\mathbf{X}_2(\mathbf{X}_2'\mathbf{M}_1\mathbf{W}_*^-\mathbf{M}_1\mathbf{X}_2)^-\mathbf{X}_2'\mathbf{M}_1\mathbf{W}_*^-, \tag{15.93}$$

with $\mathbf{W}_*$ being any matrix of the form

$$\mathbf{W}_* = \mathbf{M}_1\mathbf{V}\mathbf{M}_1 + \mathbf{M}_1\mathbf{X}_2\mathbf{U}_*\mathbf{X}_2'\mathbf{M}_1 \tag{15.94}$$

satisfying the condition

$$\mathscr{C}(\mathbf{W}_*) = \mathscr{C}(\mathbf{M}_1\mathbf{X}_2 : \mathbf{M}_1\mathbf{V}\mathbf{M}_1) = \mathscr{C}(\mathbf{M}_1\mathbf{X}_2 : \mathbf{M}_1\mathbf{V})\,. \tag{15.95}$$

Let $\mathbf{W}_2$ and the corresponding $\mathbf{U}_2$ be defined as

$$\mathbf{W}_2 = \mathbf{V} + \mathbf{X}_2\mathbf{U}_2\mathbf{X}_2'\,, \quad \text{where } \mathscr{C}(\mathbf{W}_2) = \mathscr{C}(\mathbf{X}_2 : \mathbf{V})\,,\ \mathbf{U}_2 \geq_{\mathsf{L}} \mathbf{0}\,. \tag{15.96}$$

Now $\mathscr{C}(\mathbf{W}_2) = \mathscr{C}(\mathbf{X}_2 : \mathbf{V})$ implies that $\mathscr{C}(\mathbf{M}_1\mathbf{W}_2) = \mathscr{C}(\mathbf{M}_1\mathbf{X}_2 : \mathbf{M}_1\mathbf{V})$ and thereby an appropriate choice for $\mathbf{W}_*$ is $\mathbf{W}_* = \mathbf{M}_1\mathbf{W}_2\mathbf{M}_1$ where $\mathbf{W}_2$ is defined in (15.96). Hence

$$\mathrm{BLUE}(\mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2) = \mathbf{A}\mathbf{M}_1\mathbf{y} = \mathbf{M}_1\mathbf{X}_2(\mathbf{X}_2'\dot{\mathbf{M}}_{1\mathrm{W}}\mathbf{X}_2)^-\mathbf{X}_2'\dot{\mathbf{M}}_{1\mathrm{W}}\mathbf{y}\,, \tag{15.97}$$

where

$$\dot{\mathbf{M}}_{1\mathrm{W}} = \mathbf{M}_1\mathbf{W}_*^-\mathbf{M}_1 = \mathbf{M}_1(\mathbf{M}_1\mathbf{W}\mathbf{M}_1)^-\mathbf{M}_1\,. \tag{15.98}$$

Let us denote $\mathbf{L} = (\mathbf{X}_2'\dot{\mathbf{M}}_{1\mathrm{W}}\mathbf{X}_2)^-\mathbf{X}_2'\dot{\mathbf{M}}_{1\mathrm{W}}$ so that

$$\mathbf{M}_1\mathbf{X}_2\mathbf{L}(\mathbf{X}_1 : \mathbf{X}_2 : \mathbf{V}\mathbf{M}) = (\mathbf{0} : \mathbf{M}_1\mathbf{X}_2 : \mathbf{0})\,. \tag{15.99}$$

If $\mathrm{rank}(\mathbf{M}_1\mathbf{X}_2) = \mathrm{rank}(\mathbf{X}_2)$, i.e., $\mathscr{C}(\mathbf{X}_1) \cap \mathscr{C}(\mathbf{X}_2) = \{\mathbf{0}\}$, i.e., $\mathbf{X}_2\boldsymbol{\beta}_2$ is estimable, then we can cancel the left-most $\mathbf{M}_1$ on each side of (15.99) and thus conclude that $\mathbf{X}_2\mathbf{L}\mathbf{y}$ is one representation for BLUE of $\mathbf{X}_2\boldsymbol{\beta}_2$.

For a convenience, we may write the following proposition:

Proposition 15.8. *Consider the linear model* $\mathscr{M}_{12} = \{\mathbf{y},\, \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2,\, \mathbf{V}\}$, *where* $\mathbf{X}$ *has full column rank. Then*

$$\hat{\boldsymbol{\beta}}_2(\mathscr{M}_{12}) = \hat{\boldsymbol{\beta}}_2(\mathscr{M}_{12\cdot1}) \quad \textit{and} \quad \tilde{\boldsymbol{\beta}}_2(\mathscr{M}_{12}) = \tilde{\boldsymbol{\beta}}_2(\mathscr{M}_{12\cdot1})\,. \tag{15.100}$$

One gentle warning regarding (15.100) may be worth mentioning. Namely, assuming that $\mathbf{X}$ has full column rank, and writing

$$\mathbf{A}_1\mathbf{y} = \hat{\boldsymbol{\beta}}_2(\mathscr{M}_{12}) = \mathrm{OLSE}(\boldsymbol{\beta}_2 \mid \mathscr{M}_{12})\,, \tag{15.101a}$$

$$\mathbf{A}_2\mathbf{y} = \hat{\boldsymbol{\beta}}_2(\mathscr{M}_{12\cdot1}) = \mathrm{OLSE}(\boldsymbol{\beta}_2 \mid \mathscr{M}_{12\cdot1})\,, \tag{15.101b}$$

then the multipliers $\mathbf{A}_1$ and $\mathbf{A}_2$ are unique. However, writing

$$\mathbf{B}_1\mathbf{y} = \tilde{\boldsymbol{\beta}}_2(\mathscr{M}_{12}) = \mathrm{BLUE}(\boldsymbol{\beta}_2 \mid \mathscr{M}_{12})\,, \tag{15.102a}$$

$$\mathbf{B}_2\mathbf{y} = \tilde{\boldsymbol{\beta}}_2(\mathscr{M}_{12\cdot 1}) = \mathrm{BLUE}(\boldsymbol{\beta}_2 \mid \mathscr{M}_{12\cdot 1})\,, \tag{15.102b}$$

does not necessarily mean that $\mathbf{B}_1$ and $\mathbf{B}_2$ are unique. Equation $\tilde{\boldsymbol{\beta}}_2(\mathscr{M}_{12}) = \tilde{\boldsymbol{\beta}}_2(\mathscr{M}_{12\cdot 1})$ is actually a short notation for the claim that every representation of the BLUE for $\tilde{\boldsymbol{\beta}}_2$ under $\mathscr{M}_{12}$ remains the BLUE under $\mathscr{M}_{12\cdot 1}$ and vice versa.

The first equality in (15.100) is known as the Frisch–Waugh–Lovell theorem; see Frisch & Waugh (1933) and Lovell (1963, 2008). For related results, see Groß & Puntanen (2000a), Gourieroux & Monfort (1980), Bhimasankaram & Sengupta (1996, Th. 6.1), and Sengupta & Jammalamadaka (2003, §7.10). Further considerations are being made by Groß & Puntanen (2000b, 2005), Werner & Yapar (1995), Chipman (1998, §2.2), and Fiebig, Bartels & Krämer (1996).

The reduced model has been studied, for example, in Aigner & Balestra (1988, p. 970), Nurhonen & Puntanen (1992b), Puntanen (1996, 1997), Bhimasankaram & Saha Ray (1997), Bhimasankaram, Shah & Saha Ray (1998), and Zhang, Liu & Lu (2004).

Groß & Puntanen (2000a) considered also extensively the model

$$\mathscr{M}_r = \{\mathbf{y},\, \mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2,\, \mathbf{V}\}\,. \tag{15.103}$$

Under $\mathscr{M}_r$, the OLSE of $\mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2$ is obviously the same as under $\mathscr{M}_{12}$ and $\mathscr{M}_{12\cdot 1}$. Groß & Puntanen (2000a) provided, e.g., conditions under which the BLUE of $\mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2$ in $\mathscr{M}_r$ remains the BLUE under $\mathscr{M}_{12}$. For related results, see also Kala & Pordzik (2009).

15.5 The Watson Efficiency of a Subvector

Let us again consider the partitioned linear model $\mathscr{M}_{12}$, where $\mathbf{X}_2$ has full column rank and assume further that

$$\mathscr{C}(\mathbf{X}_1) \cap \mathscr{C}(\mathbf{X}_2) = \{\mathbf{0}\}\,. \tag{15.104}$$

Then according to Proposition 15.8 (p. 330),

$$\hat{\boldsymbol{\beta}}_2(\mathscr{M}_{12}) = \hat{\boldsymbol{\beta}}_2(\mathscr{M}_{12\cdot 1}) \quad \text{and} \quad \tilde{\boldsymbol{\beta}}_2(\mathscr{M}_{12}) = \tilde{\boldsymbol{\beta}}_2(\mathscr{M}_{12\cdot 1})\,, \tag{15.105}$$

where $\mathscr{M}_{12\cdot 1} = \{\mathbf{M}_1\mathbf{y},\, \mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2,\, \mathbf{M}_1\mathbf{V}\mathbf{M}_1\}$. Suppose we have the following problem:

$$\text{When is the OLSE of } \boldsymbol{\beta}_2 \text{ under } \mathscr{M}_{12} \text{ the BLUE of } \boldsymbol{\beta}_2\,, \tag{15.106}$$

or in a short notation:

$$\text{When does } \hat{\boldsymbol\beta}_2(\mathscr{M}_{12}) = \tilde{\boldsymbol\beta}_2(\mathscr{M}_{12}) \text{ hold?} \tag{15.107}$$

Because of (15.105) we can transfer these questions to concern the reduced model $\mathscr{M}_{12\cdot1}$. Moreover, applying the Proposition 10.1 (p. 218), we get immediately the following proposition.

Proposition 15.9. *Consider a partitioned linear model $\mathscr{M}_{12}$, where $\mathbf{X}_2$ has full column rank and the disjointness property $\mathscr{C}(\mathbf{X}_1) \cap \mathscr{C}(\mathbf{X}_2) = \{\mathbf{0}\}$ holds. Then the following statements are equivalent:*

(a) $\hat{\boldsymbol\beta}_2(\mathscr{M}_{12}) = \tilde{\boldsymbol\beta}_2(\mathscr{M}_{12})$,

(b) $\hat{\boldsymbol\beta}_2(\mathscr{M}_{12\cdot1}) = \tilde{\boldsymbol\beta}_2(\mathscr{M}_{12\cdot1})$,

(c) $\mathscr{C}(\mathbf{M}_1\mathbf{V}\mathbf{M}_1\mathbf{X}_2) \subset \mathscr{C}(\mathbf{M}_1\mathbf{X}_2)$,

(d) $\mathscr{C}[\mathbf{M}_1\mathbf{V}\mathbf{M}_1(\mathbf{M}_1\mathbf{X}_2)^\perp] \subset \mathscr{C}(\mathbf{M}_1\mathbf{X}_2)^\perp$,

(e) $\mathbf{P}_{\mathbf{M}_1\mathbf{X}_2}\mathbf{M}_1\mathbf{V}\mathbf{M}_1 = \mathbf{M}_1\mathbf{V}\mathbf{M}_1\mathbf{P}_{\mathbf{M}_1\mathbf{X}_2}$,

(f) $\mathbf{P}_{\mathbf{M}_1\mathbf{X}_2}\mathbf{M}_1\mathbf{V}\mathbf{M}_1\mathbf{Q}_{\mathbf{M}_1\mathbf{X}_2} = \mathbf{0}$, *where* $\mathbf{Q}_{\mathbf{M}_1\mathbf{X}_2} = \mathbf{I}_n - \mathbf{P}_{\mathbf{M}_1\mathbf{X}_2}$,

(g) $\mathbf{P}_{\mathbf{M}_1\mathbf{X}_2}\mathbf{V}\mathbf{M}_1 = \mathbf{M}_1\mathbf{V}\mathbf{P}_{\mathbf{M}_1\mathbf{X}_2}$,

(h) $\mathbf{P}_{\mathbf{M}_1\mathbf{X}_2}\mathbf{V}\mathbf{M} = \mathbf{0}$,

(i) $\mathscr{C}(\mathbf{M}_1\mathbf{X}_2)$ *has a basis comprising p_2 eigenvectors of* $\mathbf{M}_1\mathbf{V}\mathbf{M}_1$.

There is also an alternative elementary way to prove Proposition 15.9. Namely, noting that $\hat{\boldsymbol\beta}_2(\mathscr{M}_{12}) = (\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{M}_1\mathbf{y}$ and recalling from (10.5) (p. 217) that $\mathbf{A}\mathbf{y}$ is the BLUE for an estimable parametric function $\mathbf{K}'\boldsymbol\beta$ if and only if $\mathbf{A}(\mathbf{X} : \mathbf{V}\mathbf{M}) = (\mathbf{K}' : \mathbf{0})$, we see that (15.107) holds if and only if

$$(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{M}_1(\mathbf{X}_1 : \mathbf{X}_2 : \mathbf{V}\mathbf{M}) = (\mathbf{0} : \mathbf{I}_{p_2} : \mathbf{0})\,, \tag{15.108}$$

i.e., $\mathbf{M}_1\mathbf{X}_2\mathbf{V}\mathbf{M} = \mathbf{0}$, which is the condition (h) of Proposition 15.9.

Let $\mathbf{X} = (\mathbf{X}_1 : \mathbf{X}_2)$ have full column rank, and $\mathscr{C}(\mathbf{X}) \subset \mathscr{C}(\mathbf{V})$, but $\mathbf{V}$ is possibly singular. Then

$$\tilde{\boldsymbol\beta}_2(\mathscr{M}_{12}) = (\mathbf{X}_2'\dot{\mathbf{M}}_1\mathbf{X}_2)^{-1}\mathbf{X}_2'\dot{\mathbf{M}}_1\mathbf{y}\,, \tag{15.109}$$

where $\dot{\mathbf{M}}_1 = \mathbf{M}_1(\mathbf{M}_1\mathbf{V}\mathbf{M}_1)^-\mathbf{M}_1$. Moreover,

$$\operatorname{cov}(\tilde{\boldsymbol\beta}_2 \mid \mathscr{M}_{12}) = \operatorname{cov}(\tilde{\boldsymbol\beta}_2 \mid \mathscr{M}_{12\cdot1}) = (\mathbf{X}_2'\dot{\mathbf{M}}_1\mathbf{X}_2)^{-1}, \tag{15.110}$$

and

$$\begin{aligned}\operatorname{cov}(\hat{\boldsymbol\beta}_2 \mid \mathscr{M}_{12}) &= \operatorname{cov}(\hat{\boldsymbol\beta}_2 \mid \mathscr{M}_{12\cdot1})\\ &= (\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{M}_1\mathbf{V}\mathbf{M}_1\mathbf{X}_2(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}.\end{aligned} \tag{15.111}$$

The corresponding Watson efficiencies are

$$\operatorname{eff}(\hat{\boldsymbol{\beta}} \mid \mathscr{M}_{12}) = \frac{|\operatorname{cov}(\tilde{\boldsymbol{\beta}} \mid \mathscr{M}_{12})|}{|\operatorname{cov}(\hat{\boldsymbol{\beta}} \mid \mathscr{M}_{12})|} = \frac{|\mathbf{X}'\mathbf{X}|^2}{|\mathbf{X}'\mathbf{V}\mathbf{X}| \cdot |\mathbf{X}'\mathbf{V}^+\mathbf{X}|}, \tag{15.112a}$$

$$\operatorname{eff}(\hat{\boldsymbol{\beta}}_2 \mid \mathscr{M}_{12}) = \frac{|\operatorname{cov}(\tilde{\boldsymbol{\beta}}_2 \mid \mathscr{M}_{12})|}{|\operatorname{cov}(\hat{\boldsymbol{\beta}}_2 \mid \mathscr{M}_{12})|} = \frac{|\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2|^2}{|\mathbf{X}_2'\mathbf{M}_1\mathbf{V}\mathbf{M}_1\mathbf{X}_2||\mathbf{X}_2'\dot{\mathbf{M}}_1\mathbf{X}_2|}, \tag{15.112b}$$

$$\operatorname{eff}(\hat{\boldsymbol{\beta}}_1 \mid \mathscr{M}_1) = \frac{|\mathbf{X}_1'\mathbf{X}_1|^2}{|\mathbf{X}_1'\mathbf{V}\mathbf{X}_1| \cdot |\mathbf{X}_1'\mathbf{V}^+\mathbf{X}_1|}, \tag{15.112c}$$

$$\operatorname{eff}(\hat{\boldsymbol{\beta}}_1 \mid \mathscr{M}_{1\mathrm{H}}) = \frac{|\mathbf{X}_1'\mathbf{X}_1|^2}{|\mathbf{X}_1'\mathbf{V}\mathbf{X}_1| \cdot |\mathbf{X}_1'(\mathbf{HVH})^-\mathbf{X}_1|}, \tag{15.112d}$$

where

$$\mathscr{M}_{1\mathrm{H}} = \{\mathbf{Hy},\, \mathbf{X}_1\boldsymbol{\beta}_1,\, \mathbf{HVH}\}. \tag{15.112e}$$

In a series of papers, Chu, Isotalo, Puntanen & Styan (2004, 2005, 2007, 2008), considered a particular decomposition of the Watson efficiency. They observed that the total Watson efficiency $\operatorname{eff}(\hat{\boldsymbol{\beta}} \mid \mathscr{M}_{12})$ of the $\operatorname{OLSE}(\boldsymbol{\beta})$ under the partitioned weakly singular linear model $\mathscr{M}_{12}$, where $\mathbf{X}$ has full column rank, can be expressed as the product

$$\operatorname{eff}(\hat{\boldsymbol{\beta}} \mid \mathscr{M}_{12}) = \operatorname{eff}(\hat{\boldsymbol{\beta}}_1 \mid \mathscr{M}_1) \cdot \operatorname{eff}(\hat{\boldsymbol{\beta}}_2 \mid \mathscr{M}_{12}) \cdot \frac{1}{\operatorname{eff}(\hat{\boldsymbol{\beta}}_1 \mid \mathscr{M}_{1\mathrm{H}})}, \tag{15.113}$$

where

$$\operatorname{eff}(\hat{\boldsymbol{\beta}}_1 \mid \mathscr{M}_{1\mathrm{H}}) = \frac{|\mathbf{X}_1'\mathbf{X}_1|^2}{|\mathbf{X}_1'\mathbf{V}\mathbf{X}_1| \cdot |\mathbf{X}_1'(\mathbf{HVH})^-\mathbf{X}_1|}. \tag{15.114}$$

In particular, Chu, Isotalo, Puntanen & Styan (2005) proved that the following three statements are equivalent:

$$\operatorname{eff}(\hat{\boldsymbol{\beta}} \mid \mathscr{M}_{12}) = \operatorname{eff}(\hat{\boldsymbol{\beta}}_2 \mid \mathscr{M}_{12}), \tag{15.115a}$$

$$\mathscr{C}(\mathbf{X}_1) \subset \mathscr{C}(\mathbf{VX}), \tag{15.115b}$$

$$\mathbf{Hy} \text{ is linearly sufficient for } \mathbf{X}_1\boldsymbol{\beta}_1 \text{ under model } \mathscr{M}_1. \tag{15.115c}$$

Some properties of the Watson efficiency of the estimable parametric function $\mathbf{K}'\boldsymbol{\beta}$, using the Frisch–Waugh–Lovell theorem, are studied by Isotalo, Puntanen & Styan (2009).

15.6 Equality of the BLUEs of $\mathbf{X}_1\boldsymbol{\beta}_1$ under Two Models

In Section 11.1 (p. 269), we considered the equality of the BLUEs for $\mathbf{X}\boldsymbol{\beta}$ under two linear models which were different only in their covariance matrices. Following Haslett & Puntanen (2010a), we will now consider the corresponding equality for the $\mathbf{X}_1\boldsymbol{\beta}_1$. A further, different proof (for (a) $\iff$ (b) below) appears in Mathew & Bhimasankaram (1983a, Th. 2.4).

Proposition 15.10. *Consider the partitioned linear models*

$$\mathscr{M}_{12} = \{\mathbf{y},\, \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2,\, \mathbf{V}\}\,, \quad \underline{\mathscr{M}}_{12} = \{\mathbf{y},\, \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2,\, \underline{\mathbf{V}}\}\,, \tag{15.116}$$

and assume that $\mathbf{X}_1\boldsymbol{\beta}_1$ *is estimable under* $\mathscr{M}_{12}$*. Then the following statements are equivalent:*

(a) $\{\,\mathrm{BLUE}(\mathbf{X}_1\boldsymbol{\beta}_1 \mid \mathscr{M}_{12})\,\} \subset \{\,\mathrm{BLUE}(\mathbf{X}_1\boldsymbol{\beta}_1 \mid \underline{\mathscr{M}}_{12})\,\}$,

(b) $\mathscr{C}(\underline{\mathbf{V}}\mathbf{M}) \subset \mathscr{C}(\mathbf{X}_2 : \mathbf{V}\mathbf{M})$,

(c) $\mathscr{C}(\mathbf{M}_2\underline{\mathbf{V}}\mathbf{M}) \subset \mathscr{C}(\mathbf{M}_2\mathbf{V}\mathbf{M})$.

Proof. Consider first the equivalence of (b) and (c). Clearly if (b) holds, then

$$\mathscr{C}(\mathbf{M}_2\underline{\mathbf{V}}\mathbf{M}) \subset \mathscr{C}[\mathbf{M}_2(\mathbf{X}_2 : \mathbf{V}\mathbf{M})] = \mathscr{C}(\mathbf{M}_2\mathbf{V}\mathbf{M})\,, \tag{15.117}$$

i.e., (c) holds too. To go the other way, (c) implies the column space inclusion

$$\mathscr{C}(\mathbf{X}_2 : \mathbf{M}_2\underline{\mathbf{V}}\mathbf{M}) \subset \mathscr{C}(\mathbf{X}_2 : \mathbf{M}_2\mathbf{V}\mathbf{M})\,, \tag{15.118}$$

which, in view of $\mathscr{C}(\mathbf{K} : \mathbf{L}) = \mathscr{C}(\mathbf{K} : \mathbf{Q_K}\mathbf{L})$ (for any conformable $\mathbf{K}$, $\mathbf{L}$), is equivalent to

$$\mathscr{C}(\mathbf{X}_2 : \underline{\mathbf{V}}\mathbf{M}) \subset \mathscr{C}(\mathbf{X}_2 : \mathbf{V}\mathbf{M})\,. \tag{15.119}$$

Now (c) follows from (15.119).

The proof for "(a) $\iff$ (b)", or, equivalently, for "(a) $\iff$ (c)", can be constructed using the generalized Frisch–Waugh–Lovell theorem, i.e., Proposition 15.7 (p. 328). For this purpose, we consider the reduced models

$$\mathscr{M}_{12\cdot 2} = \{\mathbf{M}_2\mathbf{y},\, \mathbf{M}_2\mathbf{X}_1\boldsymbol{\beta}_1,\, \mathbf{M}_2\mathbf{V}\mathbf{M}_2\}\,, \tag{15.120a}$$

$$\underline{\mathscr{M}}_{12\cdot 2} = \{\mathbf{M}_2\mathbf{y},\, \mathbf{M}_2\mathbf{X}_1\boldsymbol{\beta}_1,\, \mathbf{M}_2\underline{\mathbf{V}}\mathbf{M}_2\}\,. \tag{15.120b}$$

Notice that all estimable parametric functions of $\boldsymbol{\beta}_1$ under $\mathscr{M}_{12}$ are of the form

$$\mathbf{L}\mathbf{M}_2\mathbf{X}_1\boldsymbol{\beta}_1 \quad \text{for some matrix } \mathbf{L}\,. \tag{15.121}$$

On the basis of the generalized Frisch–Waugh–Lovell theorem, we know that every BLUE for $\mathbf{M}_2\mathbf{X}_1\boldsymbol{\beta}_1$ under the reduced model $\mathscr{M}_{12\cdot 2}$ remains BLUE for $\mathbf{M}_2\mathbf{X}_1\boldsymbol{\beta}_1$ under the partitioned model $\mathscr{M}_{12}$ and vice versa; in short,

$$\{\,\mathrm{BLUE}(\mathbf{M}_2\mathbf{X}_1\boldsymbol{\beta}_1 \mid \underline{\mathscr{M}}_{12})\,\} = \{\,\mathrm{BLUE}(\mathbf{M}_2\mathbf{X}_1\boldsymbol{\beta}_1 \mid \underline{\mathscr{M}}_{12\cdot 2})\,\}\,, \tag{15.122a}$$

$$\{\,\mathrm{BLUE}(\mathbf{M}_2\mathbf{X}_1\boldsymbol{\beta}_1 \mid \mathscr{M}_{12})\,\} = \{\,\mathrm{BLUE}(\mathbf{M}_2\mathbf{X}_1\boldsymbol{\beta}_1 \mid \mathscr{M}_{12\cdot 2})\,\}\,. \tag{15.122b}$$

Therefore we can base the equality studies of the $\mathrm{BLUE}(\mathbf{M}_2\mathbf{X}_1\boldsymbol{\beta}_1)$ on the reduced models, and so

$$\{\,\mathrm{BLUE}(\mathbf{M}_2\mathbf{X}_1\boldsymbol{\beta}_1 \mid \underline{\mathscr{M}}_{12})\,\} \subset \{\,\mathrm{BLUE}(\mathbf{M}_2\mathbf{X}_1\boldsymbol{\beta}_1 \mid \mathscr{M}_{12})\,\} \tag{15.123a}$$

$$\iff$$

$$\{\,\mathrm{BLUE}(\mathbf{M}_2\mathbf{X}_1\boldsymbol{\beta}_1 \mid \underline{\mathscr{M}}_{12\cdot 2})\,\} \subset \{\,\mathrm{BLUE}(\mathbf{M}_2\mathbf{X}_1\boldsymbol{\beta}_1 \mid \mathscr{M}_{12\cdot 2})\,\} \tag{15.123b}$$

$$\iff$$

$$\mathscr{C}[\mathbf{M}_2\underline{\mathbf{V}}\mathbf{M}_2(\mathbf{M}_2\mathbf{X}_1)^{\perp}] \subset \mathscr{C}[\mathbf{M}_2\mathbf{V}\mathbf{M}_2(\mathbf{M}_2\mathbf{X}_1)^{\perp}]\,, \tag{15.123c}$$

where the last statement follows from Proposition 11.1 (p. 271). Now, cf. (8.7) (p. 156),

$$\mathbf{M} = \mathbf{I}_n - \mathbf{P}_{(\mathbf{X}_1 : \mathbf{X}_2)} = \mathbf{I}_n - \mathbf{P}_{\mathbf{X}_2} - \mathbf{P}_{\mathbf{M}_2\mathbf{X}_1} = \mathbf{M}_2\mathbf{Q}_{\mathbf{M}_2\mathbf{X}_1}\,, \tag{15.124}$$

and thereby (15.123c) can be expressed as (c). On the other hand, when $\mathbf{X}_1\boldsymbol{\beta}_1$ is estimable under $\mathscr{M}_{12}$, it is easy to conclude that (15.123a) is equivalent to (a) which confirms "(a) $\iff$ (c)". □

15.7 Adding a Variable: a Recursive Formula for the BLUE

In Proposition 8.4 (p. 161), we showed that if

$$\mathscr{C}(\mathbf{X}_1) \cap \mathscr{C}(\mathbf{X}_2) = \{\mathbf{0}\}\,, \quad \text{i.e.,} \quad \operatorname{rank}(\mathbf{M}_1\mathbf{X}_2) = \operatorname{rank}(\mathbf{X}_2)\,, \tag{15.125}$$

then

$$\mathbf{X}_1\hat{\boldsymbol{\beta}}_1(\mathscr{M}_{12}) = \mathbf{X}_1\hat{\boldsymbol{\beta}}_1(\mathscr{M}_1) - \mathbf{P}_{\mathbf{X}_1}\mathbf{X}_2\hat{\boldsymbol{\beta}}_2(\mathscr{M}_{12})\,, \tag{15.126}$$

and if, in addition to (15.125), the matrix $\mathbf{X}_1$ has full column rank, then

$$\hat{\boldsymbol{\beta}}_1(\mathscr{M}_{12}) = \hat{\boldsymbol{\beta}}_1(\mathscr{M}_1) - (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\hat{\boldsymbol{\beta}}_2(\mathscr{M}_{12})\,. \tag{15.127}$$

In this section we consider a generalization of these results.

Let us denote

$$\mathbf{G}\mathbf{y} = \mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta} \mid \mathscr{M}_{12})\,, \qquad \mathbf{G}_{11}\mathbf{y} = \mathrm{BLUE}(\mathbf{X}_1\boldsymbol{\beta}_1 \mid \mathscr{M}_1)\,, \tag{15.128a}$$

$$\mathbf{G}_1\mathbf{y} = \mathrm{BLUE}(\mathbf{X}_1\boldsymbol{\beta}_1 \mid \mathscr{M}_{12})\,, \qquad \mathbf{G}_2\mathbf{y} = \mathrm{BLUE}(\mathbf{X}_2\boldsymbol{\beta}_2 \mid \mathscr{M}_{12})\,. \tag{15.128b}$$

What this means is that $\mathbf{G}$, $\mathbf{G}_1$, $\mathbf{G}_2$, and $\mathbf{G}_{11}$ are any matrices satisfying the following equations:

$$\mathbf{G}(\mathbf{X}_1 : \mathbf{X}_2 : \mathbf{V}\mathbf{M}) = (\mathbf{X}_1 : \mathbf{X}_2 : \mathbf{0})\,, \tag{15.129a}$$

$$\mathbf{G}_{11}(\mathbf{X}_1 : \mathbf{V}\mathbf{M}_1) = (\mathbf{X}_1 : \mathbf{0})\,, \tag{15.129b}$$

$$\mathbf{G}_1(\mathbf{X}_1 : \mathbf{X}_2 : \mathbf{V}\mathbf{M}) = (\mathbf{X}_1 : \mathbf{0} : \mathbf{0})\,, \tag{15.129c}$$

$$\mathbf{G}_2(\mathbf{X}_1 : \mathbf{X}_2 : \mathbf{V}\mathbf{M}) = (\mathbf{0} : \mathbf{X}_2 : \mathbf{0})\,. \tag{15.129d}$$

In other words, $\mathbf{G}$, $\mathbf{G}_{11}$, $\mathbf{G}_1$, and $\mathbf{G}_2$ are arbitrary members of the sets

$$\{\mathbf{P}_{\mathbf{X}|\mathbf{VM}}\}\,,\ \{\mathbf{P}_{\mathbf{X}_1|\mathbf{VM}_1}\}\,,\ \{\mathbf{P}_{\mathbf{X}_1|(\mathbf{X}_2:\mathbf{VM})}\}\,,\ \text{and}\ \{\mathbf{P}_{\mathbf{X}_2|(\mathbf{X}_1:\mathbf{VM})}\}\,, \tag{15.130}$$

respectively.

Notice that now we consider at the same time the full model $\mathscr{M}_{12}$ and the small model $\mathscr{M}_1$. These models have different consistency conditions and to avoid any contradiction we will assume the following:

$$\mathscr{C}(\mathbf{X}_2) \subset \mathscr{C}(\mathbf{X}_1 : \mathbf{V})\,, \quad \text{i.e.,} \quad \mathscr{C}(\mathbf{X}_1 : \mathbf{X}_2 : \mathbf{V}) = \mathscr{C}(\mathbf{X}_1 : \mathbf{V})\,. \tag{15.131}$$

The following decomposition of the BLUE($\mathbf{X}_1\boldsymbol{\beta}_1$) appears to be useful.

Proposition 15.11. *Assume that* (15.125) *and* (15.131) *hold, that is,* $\mathbf{X}_1\boldsymbol{\beta}_1$ *is estimable and* $\mathscr{C}(\mathbf{X}_2) \subset \mathscr{C}(\mathbf{X}_1 : \mathbf{V})$*, and denote*

$$\mathbf{F} = \mathbf{X}_1(\mathbf{X}_1'\mathbf{W}_1^-\mathbf{X}_1)^-\mathbf{X}_1'\mathbf{W}_1^-\,, \tag{15.132}$$

where $\mathbf{W}_1 = \mathbf{V} + \mathbf{X}_1\mathbf{U}_1\mathbf{X}_1'\,,\ \mathscr{C}(\mathbf{W}_1) = \mathscr{C}(\mathbf{X}_1 : \mathbf{V})\,.$ *Then*

$$\begin{aligned}\mathbf{X}_1\tilde{\boldsymbol{\beta}}_1(\mathscr{M}_{12}) &= \mathbf{X}_1\tilde{\boldsymbol{\beta}}_1(\mathscr{M}_1) - \mathbf{X}_1(\mathbf{X}_1'\mathbf{W}_1^-\mathbf{X}_1)^-\mathbf{X}_1'\mathbf{W}_1^-\mathbf{X}_2\tilde{\boldsymbol{\beta}}_2(\mathscr{M}_{12})\\ &= \mathbf{X}_1\tilde{\boldsymbol{\beta}}_1(\mathscr{M}_1) - \mathbf{F}\mathbf{X}_2\tilde{\boldsymbol{\beta}}_2(\mathscr{M}_{12})\,, \end{aligned} \tag{15.133}$$

or in other notation, for all $\mathbf{y} \in \mathscr{C}(\mathbf{X} : \mathbf{VM})$:

$$\mathrm{BLUE}(\mathbf{X}_1\boldsymbol{\beta}_1 \mid \mathscr{M}_{12}) = \mathrm{BLUE}(\mathbf{X}_1\boldsymbol{\beta}_1 \mid \mathscr{M}_1) - \mathbf{F}\cdot \mathrm{BLUE}(\mathbf{X}_2\boldsymbol{\beta}_2 \mid \mathscr{M}_{12})\,, \tag{15.134}$$

$$\mathbf{G}_1\mathbf{y} = \mathbf{G}_{11}\mathbf{y} - \mathbf{F}\mathbf{y}\,, \tag{15.135}$$

where the $\mathbf{G}$*-matrices are defined as in* (15.129)*.*

Proof. Assumption $\mathscr{C}(\mathbf{X}_2) \subset \mathscr{C}(\mathbf{X}_1 : \mathbf{VM})$ implies that

$$\mathbf{X}_2 = \mathbf{X}_1\mathbf{L}_1 + \mathbf{VML}_2 \quad \text{for some } \mathbf{L}_1,\ \mathbf{L}_2\,. \tag{15.136}$$

Now

$$\mathbf{G}_1(\mathbf{X}_1 : \mathbf{X}_2 : \mathbf{VM}) = (\mathbf{X}_1 : \mathbf{0} : \mathbf{0})\,, \tag{15.137a}$$

$$\mathbf{G}_{11}(\mathbf{X}_1 : \mathbf{X}_2 : \mathbf{VM}) = (\mathbf{X}_1 : \mathbf{G}_{11}\mathbf{X}_2 : \mathbf{G}_{11}\mathbf{VM})\,, \tag{15.137b}$$

$$\mathbf{F}\mathbf{G}_2(\mathbf{X}_1 : \mathbf{X}_2 : \mathbf{VM}) = (\mathbf{0} : \mathbf{F}\mathbf{X}_2 : \mathbf{0})\,. \tag{15.137c}$$

Corresponding to (15.124) (p. 335), we have $\mathbf{M} = \mathbf{M}_1\mathbf{Q}_{\mathbf{M}_1\mathbf{X}_2}$, and hence

$$\mathbf{G}_{11}\mathbf{VM} = \mathbf{G}_{11}\mathbf{VM}_1\mathbf{Q}_{\mathbf{M}_1\mathbf{X}_2} = \mathbf{0}\,. \tag{15.138}$$

Noting that $\mathbf{F}$ is one member of the class $\{\mathbf{P}_{\mathbf{X}_1|\mathbf{VM}_1}\}$ and using (15.136) and $\mathbf{M} = \mathbf{M}_1\mathbf{Q}_{\mathbf{M}_1\mathbf{X}_2}$, we get

$$\mathbf{G}_{11}\mathbf{X}_2 = \mathbf{G}_{11}(\mathbf{X}_1\mathbf{L}_1 + \mathbf{VML}_2) = \mathbf{X}_1\mathbf{L}_1\,, \tag{15.139a}$$

$$\mathbf{FX}_2 = \mathbf{X}_1(\mathbf{X}_1'\mathbf{W}_1^-\mathbf{X}_1)^-\mathbf{X}_1'\mathbf{W}_1^-(\mathbf{X}_1\mathbf{L}_1 + \mathbf{VML}_2) = \mathbf{X}_1\mathbf{L}_1\,. \tag{15.139b}$$

Hence we can conclude that

$$\mathbf{G}_1(\mathbf{X}_1 : \mathbf{X}_2 : \mathbf{VM}) = \mathbf{G}_{11}(\mathbf{X}_1 : \mathbf{X}_2 : \mathbf{VM}) - \mathbf{FG}_2(\mathbf{X}_1 : \mathbf{X}_2 : \mathbf{VM})\,, \tag{15.140}$$

and thus the proof is completed. □

The above proof appears in Haslett & Puntanen (2010a). Proposition 15.11, using different formulation and proof, appears also in Werner & Yapar (1996, Th. 2.3). In the full rank model, i.e., $\mathbf{X}$ has full column rank, and $\mathbf{V}$ is positive definite, see, e.g., Haslett (1996).

Notice that under a weakly singular partitioned linear model $\mathscr{M}_{12}$, where the column space disjointness condition $\mathscr{C}(\mathbf{X}_1) \cap \mathscr{C}(\mathbf{X}_2) = \{\mathbf{0}\}$ holds, we have

$$\mathbf{X}_1\tilde{\boldsymbol{\beta}}_1(\mathscr{M}_{12}) + \mathbf{X}_2\tilde{\boldsymbol{\beta}}_1(\mathscr{M}_{12}) = \mathbf{X}_1\tilde{\boldsymbol{\beta}}_1(\mathscr{M}_1) + \mathbf{V}\dot{\mathbf{M}}_1\mathbf{X}_2\tilde{\boldsymbol{\beta}}_2(\mathscr{M}_{12})\,. \tag{15.141}$$

Further, if $\mathbf{X}_1$ has full column rank, then

$$\tilde{\boldsymbol{\beta}}_1(\mathscr{M}_{12}) = \tilde{\boldsymbol{\beta}}_1(\mathscr{M}_1) - (\mathbf{X}_1'\mathbf{V}^-\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{V}^-\mathbf{X}_2\tilde{\boldsymbol{\beta}}_2(\mathscr{M}_{12})\,. \tag{15.142}$$

Using Proposition 15.11 (p. 336), Haslett & Puntanen (2010a) proved the following result; for related results, see Nurhonen & Puntanen (1992a) and Isotalo, Puntanen & Styan (2007), who studied whether the equality of the OLSE and BLUE continues after elimination of one observation.

Proposition 15.12. *Consider the linear models* $\mathscr{M}_{12}$ *and* $\underline{\mathscr{M}}_{12}$*, and assume that* $\mathbf{X}_1\boldsymbol{\beta}_1$ *is estimable and* $\mathscr{C}(\mathbf{X}_2) \subset \mathscr{C}(\mathbf{X}_1 : \mathbf{V})$*. Moreover, assume that every representation of* $\mathrm{BLUE}(\mathbf{X}_1\boldsymbol{\beta}_1)$ *under the small model* $\mathscr{M}_1$ *continues to be* BLUE *of* $\mathbf{X}_1\boldsymbol{\beta}_1$ *under* $\underline{\mathscr{M}}_1$*, and vice versa. Then the following statements are equivalent:*

(a) $\mathbf{X}_1\tilde{\boldsymbol{\beta}}_1(\mathscr{M}_{12}) = \mathbf{X}_1\tilde{\boldsymbol{\beta}}_1(\underline{\mathscr{M}}_{12})$,

(b) $\mathbf{X}_1'\mathbf{W}_1^-\mathbf{X}_2[\tilde{\boldsymbol{\beta}}_2(\mathscr{M}_{12}) - \tilde{\boldsymbol{\beta}}_2(\underline{\mathscr{M}}_{12})] = \mathbf{0}$.

In particular, under the condition $\operatorname{rank}(\mathbf{X}_1'\mathbf{W}_1^-\mathbf{X}_2) = \operatorname{rank}(\mathbf{X}_2) = p_2$*, we have*

$$\mathbf{X}_1\tilde{\boldsymbol{\beta}}_1(\mathscr{M}_{12}) = \mathbf{X}_1\tilde{\boldsymbol{\beta}}_1(\underline{\mathscr{M}}_{12}) \iff \tilde{\boldsymbol{\beta}}_2(\mathscr{M}_{12}) = \tilde{\boldsymbol{\beta}}_2(\underline{\mathscr{M}}_{12})\,. \tag{15.143}$$

15.8 Deleting an Observation: $\operatorname{cov}(\mathbf{y}) = \sigma^2\mathbf{V}$

Consider the generalization of Section 8.12 (p. 180) so that we have the following three linear models:

$$\mathscr{M} = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{V}\}\,, \tag{15.144a}$$

$$\mathscr{M}_{(i)} = \{\mathbf{y}_{(i)},\, \mathbf{X}_{(i)}\boldsymbol{\beta},\, \sigma^2\mathbf{V}_{(i)}\}\,, \tag{15.144b}$$

$$\mathscr{M}_Z = \{\mathbf{y},\, \mathbf{Z}\gamma,\, \sigma^2\mathbf{V}\} = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta} + \mathbf{i}_i\delta,\, \sigma^2\mathbf{V}\}\,. \tag{15.144c}$$

We assume that both $\mathbf{X}$ and $\mathbf{Z}$ have full column rank and that $\mathbf{V}$ is positive definite. By $\mathscr{M}_{(i)}$ we mean such a version of $\mathscr{M}$, where the ith case is deleted; thus $\mathbf{y}_{(i)}$ has $n-1$ elements and $\mathbf{X}_{(i)}$ has $n-1$ rows. (For notational simplicity we delete the last one.) Let us denote

$$\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\mathscr{M}) = \mathrm{BLUE}(\boldsymbol{\beta}) \text{ under } \mathscr{M}\,, \tag{15.145a}$$

$$\tilde{\boldsymbol{\beta}}_{(i)} = \tilde{\boldsymbol{\beta}}(\mathscr{M}_{(i)}) = \mathrm{BLUE}(\boldsymbol{\beta}) \text{ under } \mathscr{M}_{(i)}\,, \tag{15.145b}$$

$$\tilde{\boldsymbol{\beta}}_Z = \tilde{\boldsymbol{\beta}}(\mathscr{M}_Z) = \mathrm{BLUE}(\boldsymbol{\beta}) \text{ under } \mathscr{M}_Z\,. \tag{15.145c}$$

Model $\mathscr{M}_Z$ is an extended version of $\mathscr{M}$:

$$\mathbf{Z} = (\mathbf{X} : \mathbf{i}_i)\,, \qquad \mathbf{i}_i = (0, \dots, 0, 1)', \qquad \gamma = \begin{pmatrix} \boldsymbol{\beta} \\ \delta \end{pmatrix}. \tag{15.146}$$

Using the Frisch–Waugh–Lovell theorem, it can be shown that $\tilde{\boldsymbol{\beta}}(\mathscr{M}_{(i)}) = \tilde{\boldsymbol{\beta}}(\mathscr{M}_Z)$, and moreover,

$$\mathrm{DFBETA}_i(\mathbf{V}) = \tilde{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_{(i)} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{i}_i\, \frac{\dot{\varepsilon}_i}{\dot{m}_{ii}}\,, \tag{15.147}$$

where $\dot{\boldsymbol{\varepsilon}} = \dot{\mathbf{M}}\mathbf{y}$, and

$$\begin{aligned} \dot{m}_{ii} &= \mathbf{i}_i'\dot{\mathbf{M}}\mathbf{i}_i = \mathbf{i}_i'\mathbf{M}(\mathbf{MVM})^{-}\mathbf{M}\mathbf{i}_i \\ &= \mathbf{i}_i'[\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}]\mathbf{i}_i\,. \end{aligned} \tag{15.148}$$

The term $\dot{m}_{ii}$ corresponds now to m_{ii} in the formulas of the ordinary Studentized residuals; cf. (8.192) (p. 182) and Exercise 15.4 (p. 341).

We complete this section by noting that the generalized Cook's distance is

$$\mathrm{COOK}_i^2(\mathbf{V}) = (\tilde{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_{(i)})'\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}(\tilde{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_{(i)})/(p\tilde{\sigma}^2)\,, \tag{15.149}$$

where $\tilde{\sigma}^2 = \mathrm{SSE}(\mathbf{V})/(n-p) = \mathbf{y}'\dot{\mathbf{M}}\mathbf{y}/(n-p)$; for $\mathrm{SSE}(\mathbf{V})$, see Section 15.9 below. For further related references, see Schall & Dunne (1988, p. 164), Haslett (1996), Hayes & Haslett (1999), Christensen, Pearson & Johnson (1992), Puterman (1988), and Martin (1992).

15.9 Weighted Sum of Squares of Errors

The ordinary, unweighted sum of squares of errors SSE is defined as

$$\mathrm{SSE} = \mathrm{SSE}(\mathbf{I}) = \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = \mathbf{y}'\mathbf{M}\mathbf{y}\,, \tag{15.150}$$

while the weighted SSE is (when $\mathbf{V}$ is positive definite)

$$\begin{aligned} \mathrm{SSE}(\mathbf{V}) &= \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2_{\mathbf{V}^{-1}} = \|\mathbf{y} - \mathbf{P}_{\mathbf{X};\mathbf{V}^{-1}}\mathbf{y}\|^2_{\mathbf{V}^{-1}} \\ &= \mathbf{y}'[\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}]\mathbf{y} \\ &= \mathbf{y}'\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^{-}\mathbf{M}\mathbf{y} = \mathbf{y}'\dot{\mathbf{M}}\mathbf{y}\,. \end{aligned} \tag{15.151}$$

In the general case, the weighted SSE can be defined as

$$\mathrm{SSE}(\mathbf{V}) = (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'\mathbf{W}^{-}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})\,, \tag{15.152}$$

where $\mathbf{X}\tilde{\boldsymbol{\beta}} = \mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta})$ and $\mathbf{W} = \mathbf{V} + \mathbf{X}\mathbf{U}\mathbf{X}'$, with $\mathscr{C}(\mathbf{W}) = \mathscr{C}(\mathbf{X} : \mathbf{V})$. Then, recalling that the BLUE's residual is $\tilde{\varepsilon} = \mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}} = \mathbf{V}\dot{\mathbf{M}}\mathbf{y}$, it is straightforward to confirm the following:

$$\begin{aligned} \mathrm{SSE}(\mathbf{V}) &= (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'\mathbf{W}^{-}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) \\ &= \tilde{\varepsilon}'\mathbf{W}^{-}\tilde{\varepsilon} = \mathbf{y}'\dot{\mathbf{M}}\mathbf{V}\mathbf{W}^{-}\mathbf{V}\dot{\mathbf{M}}\mathbf{y} \\ &= \mathbf{y}'\dot{\mathbf{M}}\mathbf{W}\mathbf{W}^{-}\mathbf{W}\dot{\mathbf{M}}\mathbf{y} = \mathbf{y}'\dot{\mathbf{M}}\mathbf{W}\dot{\mathbf{M}}\mathbf{y} \\ &= \mathbf{y}'\dot{\mathbf{M}}\mathbf{V}\dot{\mathbf{M}}\mathbf{y} = \mathbf{y}'\dot{\mathbf{M}}\mathbf{V}\mathbf{V}^{-}\mathbf{V}\dot{\mathbf{M}}\mathbf{y} \\ &= \tilde{\varepsilon}'\mathbf{V}^{-}\tilde{\varepsilon} = \mathbf{y}'\dot{\mathbf{M}}\mathbf{y}\,. \end{aligned} \tag{15.153}$$

Note that $\mathrm{SSE}(\mathbf{V})$ is invariant with respect to the choice of $\mathbf{W}^{-}$.

We recall, in light of Proposition 15.2 (p. 321), that the following holds:

$$\ddot{\mathbf{M}}_{\mathrm{W}} = \mathbf{P}_{\mathbf{W}}\dot{\mathbf{M}}\mathbf{P}_{\mathbf{W}} = \mathbf{W}^{+} - \mathbf{W}^{+}\mathbf{X}(\mathbf{X}'\mathbf{W}^{-}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}^{+}. \tag{15.154}$$

From (15.154) it follows that for every $\mathbf{y} \in \mathscr{C}(\mathbf{W}) = \mathscr{C}(\mathbf{X} : \mathbf{V})$,

$$\mathbf{y}'\mathbf{P}_{\mathbf{W}}\dot{\mathbf{M}}\mathbf{P}_{\mathbf{W}}\mathbf{y} = \mathbf{y}'[\mathbf{W}^{+} - \mathbf{W}^{+}\mathbf{X}(\mathbf{X}'\mathbf{W}^{-}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}^{+}]\mathbf{y}\,, \tag{15.155}$$

i.e.,

$$\mathrm{SSE}(\mathbf{V}) = \mathbf{y}'\dot{\mathbf{M}}\mathbf{y} = \mathbf{y}'[\mathbf{W}^{-} - \mathbf{W}^{-}\mathbf{X}(\mathbf{X}'\mathbf{W}^{-}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}^{-}]\mathbf{y}\,. \tag{15.156}$$

Therefore, we can conclude the following result where the matrix $\dot{\mathbf{M}}$ nicely shows its usefulness.

Proposition 15.13. *Under the linear model* $\{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$*, the weighted sum of squares of errors can be written as*

$$\mathrm{SSE}(\mathbf{V}) = \tilde{\varepsilon}'\mathbf{W}^{-}\tilde{\varepsilon} = \tilde{\varepsilon}'\mathbf{V}^{-}\tilde{\varepsilon} = \mathbf{y}'\dot{\mathbf{M}}\mathbf{y} = \mathbf{y}'\ddot{\mathbf{M}}\mathbf{y}\,. \tag{15.157}$$

The sum $\mathrm{SSE}(\mathbf{V})$ provides an unbiased estimator of σ^2:

$$\mathrm{E}(\mathbf{y}'\dot{\mathbf{M}}\mathbf{y}/f) := \mathrm{E}(\tilde{\sigma}^2) = \sigma^2, \quad f = \operatorname{rank}(\mathbf{VM})\,. \tag{15.158}$$

This is easy to see because using

$$\begin{aligned} \mathrm{E}(\mathbf{y}'\mathbf{A}\mathbf{y}) &= \mathrm{E}[\operatorname{tr}(\mathbf{y}'\mathbf{A}\mathbf{y})] = \mathrm{E}[\operatorname{tr}(\mathbf{A}\mathbf{y}\mathbf{y}')] \\ &= \operatorname{tr}[\mathbf{A}\,\mathrm{E}(\mathbf{y}\mathbf{y}')] = \operatorname{tr}[\mathbf{A}(\sigma^2\mathbf{V} + \mathbf{X}\boldsymbol{\beta}(\mathbf{X}\boldsymbol{\beta})'] \\ &= \sigma^2 \operatorname{tr}(\mathbf{AV}) + \boldsymbol{\beta}'\mathbf{X}'\mathbf{A}\mathbf{X}\boldsymbol{\beta}\,, \end{aligned} \tag{15.159}$$

where $\mathbf{A}$ is a symmetric matrix, we get

$$\begin{aligned} \mathrm{E}(\mathbf{y}'\dot{\mathbf{M}}\mathbf{y}) &= \sigma^2 \operatorname{tr}(\dot{\mathbf{M}}\mathbf{V}) = \sigma^2 \operatorname{tr}[\mathbf{M}(\mathbf{MVM})^-\mathbf{MV}] \\ &= \sigma^2 \operatorname{tr}[\mathbf{MVM}(\mathbf{MVM})^-] = \sigma^2 \operatorname{rank}[\mathbf{MVM}(\mathbf{MVM})^-] \\ &= \sigma^2 \operatorname{rank}(\mathbf{MVM})\,. \end{aligned} \tag{15.160}$$

The proof of the following proposition, originally proved (by other means) by Groß (1997a), shows another interesting application of the matrix $\dot{\mathbf{M}}$.

Proposition 15.14. *Under the linear model* $\{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$*, the following statements are equivalent:*

(a) $\mathrm{SSE}(\mathbf{I}) = \mathbf{y}'\mathbf{M}\mathbf{y} = \mathbf{y}'\dot{\mathbf{M}}\mathbf{y} = \mathrm{SSE}(\mathbf{V})$ *for all* $\mathbf{y} \in \mathscr{C}(\mathbf{X} : \mathbf{V})$*,*

(b) $(\mathbf{VM})^2 = \mathbf{VM}$*.*

Proof. Because $\mathbf{y} \in \mathscr{C}(\mathbf{X} : \mathbf{VM})$ there exist vectors $\mathbf{a}$ and $\mathbf{b}$ such that $\mathbf{y} = \mathbf{Xa} + \mathbf{VMb}$ and so (a) holds if and only if

$$\mathbf{b}'\mathbf{MV}\cdot\mathbf{M}\cdot\mathbf{VMb} = \mathbf{b}'\mathbf{MV}\cdot\mathbf{M}(\mathbf{MVM})^-\mathbf{M}\cdot\mathbf{VMb} \quad \text{for all } \mathbf{b} \in \mathbb{R}^n, \tag{15.161}$$

i.e.,

$$\mathbf{b}'\mathbf{MVMVMb} = \mathbf{b}'\mathbf{MVMb} \quad \text{for all } \mathbf{b} \in \mathbb{R}^n, \tag{15.162}$$

which is further equivalent to (cf. Harville 1997, Cor. 14.1.3)

$$\mathbf{MVMVM} = \mathbf{MVM}\,. \tag{15.163}$$

Clearly (b) implies (15.163). On the other hand, in view of $\operatorname{rank}(\mathbf{MVM}) = \operatorname{rank}(\mathbf{VM})$, the rank cancellation rule allows us to cancel the left-most $\mathbf{M}$ in (15.163), and so we obtain (b). □

For related references, see Groß (1997a, Prop. 1), Koch (1969, p. 970), Young, Odell & Hahn (2000), Young, Scariano & Hallum (2005), and Wang, Wu & Ma (2003).

15.10 Exercises

15.1. Prove Proposition 15.5 (p. 326).

15.2. Prove Proposition 15.12 (p. 337).

15.3. Prove (see page 338):

$$\tilde{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_{(i)} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{i}_i \, \frac{\dot{\varepsilon}_i}{\dot{m}_{ii}} \,, \qquad \frac{\dot{\varepsilon}_i}{\dot{m}_{ii}} = \tilde{\delta} = \mathrm{BLUE}(\delta \mid \mathscr{M}_Z) \,.$$

15.4. Consider the models defined as in (15.144) (p. 337). Show that the F-test statistics for testing $\delta = 0$ under $\mathscr{M}_Z$ is

$$t_i^2(\mathbf{V}) = \frac{\dot{\varepsilon}_i^2}{\dot{m}_{ii}\tilde{\sigma}_{(i)}^2} = \frac{\tilde{\delta}_i^2}{\widetilde{\mathrm{var}}(\tilde{\delta}_i)} \,,$$

where $\dot{\boldsymbol{\varepsilon}} = \dot{\mathbf{M}}\mathbf{y} = \mathbf{V}^{-1}\tilde{\boldsymbol{\varepsilon}}$, $\tilde{\boldsymbol{\varepsilon}}$ is the BLUE's residual, and $(n-p-1)\tilde{\sigma}_{(i)}^2 = \mathrm{SSE}_{(i)}(\mathbf{V}) = \mathrm{SSE}_Z(\mathbf{V})$. The statistics $t_i^2(\mathbf{V})$ is a generalized version of the externally Studentized residual, cf. (8.192) (p. 182).

15.5. Suppose that $\mathbf{V}$ has the intraclass correlation structure and $\mathbf{1} \in \mathscr{C}(\mathbf{X})$, where $\mathbf{X}_{n\times p}$ has full column rank. Show that then $\mathrm{DFBETA}_i(\mathbf{V}) = \mathrm{DFBETA}_i(\mathbf{I})$. What about the equality between Cook's distance COOK_i^2 and the generalized Cook's distance $\mathrm{COOK}_i^2(\mathbf{V})$?

15.6. Express the Cook's distance $\mathrm{COOK}_i^2(\mathbf{V})$ as a function of an appropriate Mahalanobis distance.

15.7. Consider a partitioned model $\mathscr{M}_{12} = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \mathbf{V}\}$, where $\mathbf{X}$ has full column rank and $\mathbf{V}$ is positive definite. Suppose that $\hat{\boldsymbol{\beta}}_1$ is fully efficient under the small model $\mathscr{M}_1 = \{\mathbf{y},\, \mathbf{X}_1\boldsymbol{\beta}_1,\, \mathbf{V}\}$. Show that

$$\mathrm{eff}(\hat{\boldsymbol{\beta}}_1 \mid \mathscr{M}_{12}) = 1 \iff \mathbf{X}_1'\mathbf{X}_2\tilde{\boldsymbol{\beta}}_2(\mathscr{M}_{12}) = \mathbf{X}_1'\mathbf{X}_2\hat{\boldsymbol{\beta}}_2(\mathscr{M}_{12}) \,.$$

Isotalo, Puntanen & Styan (2007).

15.8. Consider the models $\mathscr{M}$ and $\mathscr{M}_{(i)}$, defined in (15.144) (p. 337), where $\mathbf{X}$ and $(\mathbf{X} : \mathbf{i}_i)$ have full column ranks. Suppose that $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}$ under $\mathscr{M}$. Show that the equality $\hat{\boldsymbol{\beta}}_{(i)} = \tilde{\boldsymbol{\beta}}_{(i)}$ for all $i = 1, \dots, n$ holds if and only if $\mathbf{MVM} = c^2\mathbf{M}$ for some nonzero $c \in \mathbb{R}$.

Nurhonen & Puntanen (1992a), Isotalo, Puntanen & Styan (2007).

15.9. According to (8.192) (p. 182), the F-test statistic for testing the hypothesis $\delta = 0$ under the model $\mathscr{M}_Z(\mathbf{I}) = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta} + \mathbf{u}\delta,\, \sigma^2\mathbf{I}\}$, where $\mathbf{u} = \mathbf{i}_i$, becomes

$$t_i^2 = \frac{\mathbf{y}'\mathbf{P}_{\mathbf{Mu}}\mathbf{y}}{\frac{1}{n-p-1}\mathbf{y}'(\mathbf{M} - \mathbf{P}_{\mathbf{Mu}})\mathbf{y}} = \frac{\hat{\varepsilon}_i^2}{\frac{1}{n-p-1}\,\mathrm{SSE}_{(i)}\, m_{ii}} \,,$$

which is the squared externally Studentized residual. Under the model $\mathscr{M}_Z(\mathbf{I})$ the test statistic t_i^2 follows an F-distribution with 1 and $n-p-1$

degrees of freedom. Denote $\mathscr{M}_Z(\mathbf{V}) = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta} + \mathbf{u}\delta,\, \sigma^2\mathbf{V}\}$. Prove that under the model $\mathscr{M}_Z(\mathbf{V})$: $t_i^2 \sim \mathrm{F}(1, n-p-1) \iff \mathbf{MVM} = c^2\mathbf{M}$ for some $c \neq 0$.

Nurhonen & Puntanen (1991), Rao & Mitra (1971b, Ch. 9).

15.10. Write up the solution to Exercise 15.9 when $\mathbf{Z} = (\mathbf{1} : \mathbf{i}_n)$. Moreover, confirm that the following statements are equivalent:

(a) $\mathbf{CVC} = c^2\mathbf{C}$ for some $c \neq 0$,

(b) $\mathbf{V} = \alpha^2\mathbf{I} + \mathbf{a}\mathbf{1}' + \mathbf{1}\mathbf{a}'$, where $\mathbf{a}$ is an arbitrary vector and α is any scalar ensuring the positive definiteness of $\mathbf{V}$.

Above $\mathbf{C}$ denotes the centering matrix. Confirm also that the eigenvalues of $\mathbf{V}$ in (b) are $\alpha^2 + \mathbf{1}'\mathbf{a} \pm \sqrt{n\mathbf{a}'\mathbf{a}}$, each with multiplicity one, and α^2 with multiplicity $n-2$. These results appear useful when studying the robustness of the Grubbs's test for detecting a univariate outlier.

Baksalary & Puntanen (1990b), Baksalary, Nurhonen & Puntanen (1992), Lehman & Young (1993), Markiewicz (2001).

15.11. Consider the models $\mathscr{M}_\varrho = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{V}\}$ and $\mathscr{M}_0 = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{I}\}$, where $\mathbf{X} = (\mathbf{1} : \mathbf{x}_1 : \dots : \mathbf{x}_k)$ and $\mathbf{V}$ has the intraclass correlation structure $\mathbf{V} = (1-\varrho)\mathbf{I}_n + \varrho\mathbf{1}_n\mathbf{1}_n'$, with $-\frac{1}{n-1} < \varrho < 1$. Show that the weighted sum of squares of errors under $\mathscr{M}_\varrho$ is

$$\begin{aligned}\mathrm{SSE}(\mathbf{V}) &= \mathbf{y}'\dot{\mathbf{M}}\mathbf{y} = \mathbf{y}'\mathbf{M}(\mathbf{MVM})^-\mathbf{My}\\ &= (1-\varrho)^{-1}\mathbf{y}'\mathbf{My} = (1-\varrho)^{-1}\,\mathrm{SSE}(\mathbf{I})\,.\end{aligned}$$

15.12 (Continued ...). Denote $\mathbf{X} = (\mathbf{1} : \mathbf{x}_1 : \dots : \mathbf{x}_{k-1} : \mathbf{x}_k) = (\mathbf{X}_1 : \mathbf{x}_k)$. The F-test statistics for testing the hypothesis $H\colon \beta_k = 0$ under $\mathscr{M}_0$ and under $\mathscr{M}_\varrho$ are

$$F_0 = \frac{\mathbf{y}'(\mathbf{M}_1 - \mathbf{M})\mathbf{y}}{\mathbf{y}'\mathbf{My}/f}\,, \quad F_\varrho = \frac{\mathbf{y}'(\dot{\mathbf{M}}_1 - \dot{\mathbf{M}})\mathbf{y}}{\mathbf{y}'\dot{\mathbf{M}}\mathbf{y}/f}\,,$$

respectively, where $\mathbf{M}_1 = \mathbf{I} - \mathbf{P}_{\mathbf{X}_1}$, $\dot{\mathbf{M}}_1 = \mathbf{M}_1(\mathbf{M}_1\mathbf{V}\mathbf{M}_1)^-\mathbf{M}_1$ and $f = n - \operatorname{rank}(\mathbf{X})$. Confirm that $F_0 = F_\varrho$. For the robustness of the F-test, see Khatri (1981), Mathew & Bhimasankaram (1983a,b), Mathew (1985).

15.13. Suppose that $\mathbf{X}$ has full column rank, $\mathbf{V} \in \mathrm{PD}_n$, and $\mathbf{L} = \mathbf{X}_1'\mathbf{V}^{-1}\mathbf{X}_1$, $\mathbf{N} = (\mathbf{X}_1'\mathbf{V}^{-1}\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{V}^{-1}\mathbf{X}_2$. Show using (15.142) (p. 337):

$$\begin{aligned}\operatorname{cov}[\tilde{\boldsymbol{\beta}}_1(\mathscr{M}_{12})] &= (\mathbf{X}_2'\dot{\mathbf{M}}_1\mathbf{X}_2)^{-1} = \operatorname{cov}[\tilde{\boldsymbol{\beta}}_1(\mathscr{M}_1)] + \operatorname{cov}[\mathbf{N}\tilde{\boldsymbol{\beta}}_2(\mathscr{M}_{12})]\\ &= \mathbf{L}^{-1} + \mathbf{N}(\mathbf{X}_2'\dot{\mathbf{M}}_1\mathbf{X}_2)^{-1}\mathbf{N}'.\end{aligned}$$

Chapter 16
Disjointness of Column Spaces

> *My good friend Jacques Monod spoke often of the randomness of the cosmos. He believed everything in existence occurred by pure chance with the possible extension of his breakfast, which he felt certain was made by his housekeeper.*
>
> WOODY ALLEN: *My Speech to the Graduates*

In this chapter we collect together various equivalent characterizations for the disjointness of the column spaces $\mathscr{C}(\mathbf{A})$ and $\mathscr{C}(\mathbf{B})$, by which we mean that $\mathscr{C}(\mathbf{A})$ and $\mathscr{C}(\mathbf{B})$ have only the null vector in common; here $\mathbf{A}$ and $\mathbf{B}$ are $n \times p$ and $n \times q$ matrices. There are numerous situations in linear models and multivariate analysis when we meet the problem of disjointness.

Theorem 16 (Disjointness of column spaces). *For conformable matrices* $\mathbf{A}$ *and* $\mathbf{B}$*, the following statements are equivalent:*

(a) $\mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B}) = \{\mathbf{0}\}$,

(b) $\begin{pmatrix} \mathbf{A}' \\ \mathbf{B}' \end{pmatrix} (\mathbf{AA}' + \mathbf{BB}')^{-}(\mathbf{AA}' : \mathbf{BB}') = \begin{pmatrix} \mathbf{A}' & \mathbf{0} \\ \mathbf{0} & \mathbf{B}' \end{pmatrix}$,

(c) $\mathbf{A}'(\mathbf{AA}' + \mathbf{BB}')^{-}\mathbf{AA}' = \mathbf{A}'$,

(d) $\mathbf{A}'(\mathbf{AA}' + \mathbf{BB}')^{-}\mathbf{B} = \mathbf{0}$,

(e) $(\mathbf{AA}' + \mathbf{BB}')^{-}$ *is a generalized inverse of* $\mathbf{AA}'$,

(f) $\mathbf{A}'(\mathbf{AA}' + \mathbf{BB}')^{-}\mathbf{A} = \mathbf{P}_{\mathbf{A}'}$,

(g) $\mathscr{C}\begin{pmatrix} \mathbf{0} \\ \mathbf{B}' \end{pmatrix} \subset \mathscr{C}\begin{pmatrix} \mathbf{A}' \\ \mathbf{B}' \end{pmatrix}$,

(h) $\mathscr{N}(\mathbf{A} : \mathbf{B}) \subset \mathscr{N}(\mathbf{0} : \mathbf{B})$,

(i) $\mathbf{Y}(\mathbf{A} : \mathbf{B}) = (\mathbf{0} : \mathbf{B})$ *has a solution for* $\mathbf{Y}$,

(j) $\mathbf{P}_{\binom{\mathbf{A}'}{\mathbf{B}'}} = \begin{pmatrix} \mathbf{P}_{\mathbf{A}'} & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_{\mathbf{B}'} \end{pmatrix}$,

(k) $\operatorname{rank}(\mathbf{Q_B A}) = \operatorname{rank}(\mathbf{A})$,

(l) $\mathscr{C}(\mathbf{A}'\mathbf{Q_B}) = \mathscr{C}(\mathbf{A}')$,

(m) $\mathbf{P}_{\mathbf{A}'\mathbf{Q_B}}\mathbf{A}' = \mathbf{A}'$,

(n) $\operatorname{ch}_1(\mathbf{P_A P_B}) < 1$,

(o) $\det(\mathbf{I} - \mathbf{P_A P_B}) \neq 0$.

S. Puntanen et al., *Matrix Tricks for Linear Statistical Models: Our Personal Top Twenty*, 343

DOI 10.1007/978-3-642-10473-2_17,

Proof. Applying the rank cancellation rule (p. 145) to the equation

$$(\mathbf{A} : \mathbf{B}) \begin{pmatrix} \mathbf{A}' \\ \mathbf{B}' \end{pmatrix} (\mathbf{AA}' + \mathbf{BB}')^{-} (\mathbf{A} : \mathbf{B}) \begin{pmatrix} \mathbf{A}' \\ \mathbf{B}' \end{pmatrix} = (\mathbf{A} : \mathbf{B}) \begin{pmatrix} \mathbf{A}' \\ \mathbf{B}' \end{pmatrix}, \tag{16.1}$$

we can cancel $(\mathbf{A} : \mathbf{B})$ from the left side on each part of the equation. Rewrite (16.1) as

$$\begin{pmatrix} \mathbf{A}' \\ \mathbf{B}' \end{pmatrix} (\mathbf{AA}' + \mathbf{BB}')^{-} (\mathbf{A} : \mathbf{B}) \begin{pmatrix} \mathbf{A}' & \mathbf{0} \\ \mathbf{0} & \mathbf{B}' \end{pmatrix} \begin{pmatrix} \mathbf{I} \\ \mathbf{I} \end{pmatrix} = \begin{pmatrix} \mathbf{A}' & \mathbf{0} \\ \mathbf{0} & \mathbf{B}' \end{pmatrix} \begin{pmatrix} \mathbf{I} \\ \mathbf{I} \end{pmatrix}. \tag{16.2}$$

The disjointness condition (a) is clearly equivalent to

$$\operatorname{rank} \left[\begin{pmatrix} \mathbf{A}' & \mathbf{0} \\ \mathbf{0} & \mathbf{B}' \end{pmatrix} \begin{pmatrix} \mathbf{I} \\ \mathbf{I} \end{pmatrix} \right] = \operatorname{rank} \begin{pmatrix} \mathbf{A}' & \mathbf{0} \\ \mathbf{0} & \mathbf{B}' \end{pmatrix}, \tag{16.3}$$

and, therefore, (a) implies that $(\mathbf{I} : \mathbf{I})'$ can be cancelled from both sides of (16.2) which yields (b). Clearly (b) implies (c) and (d), and (c) & (d) together is equivalent to (b). Since $\mathscr{C}(\mathbf{A}) \subset \mathscr{C}(\mathbf{AA}' + \mathbf{BB}')$, we must always have

$$(\mathbf{AA}' + \mathbf{BB}')(\mathbf{AA}' + \mathbf{BB}')^{-}\mathbf{A} = \mathbf{A}, \tag{16.4}$$

that is, we always have

$$\mathbf{AA}'(\mathbf{AA}' + \mathbf{BB}')^{-}\mathbf{A} = \mathbf{A} - \mathbf{BB}'(\mathbf{AA}' + \mathbf{BB}')^{-}\mathbf{A}. \tag{16.5}$$

Now equation (16.5) proves the equivalence between (c) and (d). Thereby also (c) implies (b). Furthermore, (e) clearly implies (f) and if (f) is postmultiplied by $\mathbf{A}'$, we obtain (c). To prove that (c) implies (a), we take a vector $\mathbf{z} \in \mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B})$, i.e., there exist vectors $\mathbf{a}$ and $\mathbf{b}$ such that

$$\mathbf{u} = \mathbf{Aa} = \mathbf{Bb}. \tag{16.6}$$

Premultiplying (16.6) by $\mathbf{AA}'(\mathbf{AA}' + \mathbf{BB}')^{-}$ and using (c) yields (a). Statement (g) is equivalent to

$$\mathbf{P}_{\begin{pmatrix} \mathbf{A}' \\ \mathbf{B}' \end{pmatrix}} \begin{pmatrix} \mathbf{0} \\ \mathbf{B}' \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{B}' \end{pmatrix}. \tag{16.7}$$

Writing out the explicit expression of the orthogonal projector in (16.7) proves the equivalence between (g) and the preceding statements. The equivalence between (g) and (f) is obvious.

The rest of the proof is left to the reader. □

16.1 Estimability of $\mathbf{X}_2\boldsymbol{\beta}_2$

Let us consider the partitioned linear model

$$\mathscr{M}_{12} = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \mathbf{V}\} = \{\mathbf{y},\, \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2,\, \sigma^2\mathbf{V}\}\,, \tag{16.8}$$

where $\mathbf{X}_1$ is an $n \times p_1$ matrix and $\mathbf{X}_2$ is an $n \times p_2$; $p = p_1 + p_2$. We may now ask when is the parametric function $\mathbf{X}_2\boldsymbol{\beta}_2$ estimable under $\mathscr{M}_{12}$. We know that $\mathbf{X}_2\boldsymbol{\beta}_2$ is estimable if it has a linear unbiased estimator, i.e., there exists a matrix $\mathbf{A}$ such that

$$\mathrm{E}(\mathbf{Ay}) = \mathbf{AX}\boldsymbol{\beta} = \mathbf{AX}_1\boldsymbol{\beta}_1 + \mathbf{AX}_2\boldsymbol{\beta}_2 = \mathbf{X}_2\boldsymbol{\beta}_2 \quad \text{for all } \boldsymbol{\beta} \in \mathbb{R}^p, \tag{16.9}$$

i.e., $\mathbf{A}(\mathbf{X}_1 : \mathbf{X}_2) = (\mathbf{0} : \mathbf{X}_2)$. Therefore, we can conclude the following:

$$\mathbf{X}_2\boldsymbol{\beta}_2 \text{ is estimable } \iff \mathscr{C}\begin{pmatrix} \mathbf{0} \\ \mathbf{X}_2' \end{pmatrix} \subset \mathscr{C}\begin{pmatrix} \mathbf{X}_1' \\ \mathbf{X}_2' \end{pmatrix}, \tag{16.10}$$

which in view of Theorem 16 is equivalent to

$$\mathscr{C}(\mathbf{X}_1) \cap \mathscr{C}(\mathbf{X}_2) = \{\mathbf{0}\}\,. \tag{16.11}$$

Our interest may focus on estimation of estimable parametric functions $\mathbf{L}\boldsymbol{\beta}_2$. Then $\mathbf{L}\boldsymbol{\beta}_2 = (\mathbf{0} : \mathbf{L})\binom{\boldsymbol{\beta}_1}{\boldsymbol{\beta}_2}$ is estimable if and only if there exists a matrix $\mathbf{A}$ such that $\mathbf{A}(\mathbf{X}_1 : \mathbf{X}_2)\boldsymbol{\beta} = (\mathbf{0} : \mathbf{L})\boldsymbol{\beta}$ for all $\boldsymbol{\beta} \in \mathbb{R}^p$, i.e.,

$$\mathbf{A}(\mathbf{X}_1 : \mathbf{X}_2) = (\mathbf{0} : \mathbf{L}) \quad \text{for some } \mathbf{A}\,. \tag{16.12}$$

We can now prove, see Groß & Puntanen (2000a, Lemma 1), that (16.12) is equivalent to

$$\mathbf{L} = \mathbf{BM}_1\mathbf{X}_2 \quad \text{for some } \mathbf{B}\,, \tag{16.13}$$

i.e., $\mathscr{C}(\mathbf{L}') \subset \mathscr{C}(\mathbf{X}_2'\mathbf{M}_1)$. If $\mathbf{L} = \mathbf{BM}_1\mathbf{X}_2$, then $\mathbf{A} = \mathbf{BM}_1$ satisfies (16.12). Conversely, if (16.12) holds for some $\mathbf{A}$, then $\mathbf{AX}_1 = \mathbf{0}$, implying $\mathbf{A} = \mathbf{BM}_1$ for some $\mathbf{B}$. Hence $\mathbf{L} = \mathbf{BM}_1\mathbf{X}_2$.

We may collect the summary of our results in the following proposition; see also Baksalary, Puntanen & Yanai (1992, Lemma 1).

Proposition 16.1. *Let us consider the partitioned linear model* $\mathscr{M}_{12} = \{\mathbf{y},\, \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2,\, \sigma^2\mathbf{V}\}$. *Then the following statements are equivalent:*

(a) $\mathbf{L}\boldsymbol{\beta}_2$ *is estimable,*

(b) $\mathscr{C}(\mathbf{L}') \subset \mathscr{C}(\mathbf{X}_2'\mathbf{M}_1)$,

(c) $\mathbf{L} = \mathbf{BM}_1\mathbf{X}_2$ *for some* $\mathbf{B}$.

In particular, $\boldsymbol{\beta}_2$ *is estimable if and only if*

$$\operatorname{rank}(\mathbf{X}_2'\mathbf{M}_1) = \operatorname{rank}(\mathbf{X}_2) = p_2\,. \tag{16.14}$$

Moreover, denoting

$$\mathbf{P}_{\mathbf{X}_2 \cdot \mathbf{X}_1} = \mathbf{X}_2(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^-\mathbf{X}_2'\mathbf{M}_1 \,, \tag{16.15}$$

the following statements are equivalent:

(i) $\mathbf{X}_2\boldsymbol{\beta}_2$ *is estimable,*
(ii) $\mathscr{C}(\mathbf{X}_2') = \mathscr{C}(\mathbf{X}_2'\mathbf{M}_1)$,
(iii) $\operatorname{rank}(\mathbf{X}_2') = \operatorname{rank}(\mathbf{X}_2'\mathbf{M}_1)$,
(iv) $\mathscr{C}(\mathbf{X}_1) \cap \mathscr{C}(\mathbf{X}_2) = \{\mathbf{0}\}$,
(v) $\mathbf{P}_{\mathbf{X}_2 \cdot \mathbf{X}_1}\mathbf{X}_2 = \mathbf{X}_2$,
(vi) $\mathbf{P}_{\mathbf{X}_2 \cdot \mathbf{X}_1}$ *is invariant with respect to the choice of* $(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^-$,
(vii) $\mathbf{P}_{\mathbf{X}_2 \cdot \mathbf{X}_1}$ *is a projector onto* $\mathscr{C}(\mathbf{X}_2)$ *along* $\mathscr{C}(\mathbf{X}_1) \oplus \mathscr{C}(\mathbf{X})^\perp$,
(viii) $\mathbf{H} = \mathbf{P}_{\mathbf{X}_2 \cdot \mathbf{X}_1} + \mathbf{P}_{\mathbf{X}_1 \cdot \mathbf{X}_2}$.

For an extensive study of the estimability, see Alalouf & Styan (1979a,b).

16.2 When is $(\mathbf{V} + \mathbf{XUX}')^-$ a G-inverse of $\mathbf{V}$?

Consider a linear model $\{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$ and let $\mathbf{W}$ be defined as

$$\mathbf{W} = \mathbf{V} + \mathbf{XAA}'\mathbf{X} = (\mathbf{V}^{1/2} : \mathbf{XA})(\mathbf{V}^{1/2} : \mathbf{XA})', \tag{16.16}$$

where $\mathbf{A}$ satisfies the condition

$$\mathscr{C}(\mathbf{X} : \mathbf{V}) = \mathscr{C}(\mathbf{W}) \,. \tag{16.17}$$

Note that (16.17) is now equivalent to $\operatorname{rank}(\mathbf{X} : \mathbf{V}) = \operatorname{rank}(\mathbf{XA} : \mathbf{V})$. Then one representation for the BLUE of $\mathbf{X}\boldsymbol{\beta}$ is

$$\operatorname{BLUE}(\mathbf{X}\boldsymbol{\beta}) = \mathbf{X}(\mathbf{X}'\mathbf{W}^-\mathbf{X})^-\mathbf{X}'\mathbf{W}^-\mathbf{y} \,. \tag{16.18}$$

Now one may ask which choice of $\mathbf{A}$ makes $\mathbf{W}^-$ to be also a generalized inverse of $\mathbf{V}$, that is,

$$\mathbf{V}(\mathbf{V} + \mathbf{XAA}'\mathbf{X})^-\mathbf{V} = \mathbf{V} \,. \tag{16.19}$$

The answer comes immediately from Theorem 16: $\mathbf{A}$ must satisfy the disjointness condition

$$\mathscr{C}(\mathbf{V}) \cap \mathscr{C}(\mathbf{XA}) = \{\mathbf{0}\} \,. \tag{16.20}$$

As suggested by Groß (1997b), one possible choice to $\mathbf{A}$ to satisfy both (16.17) and (16.20) is for example $\mathbf{A} = \mathbf{X}^+(\mathbf{I}_n - \mathbf{P}_\mathbf{V})$. For further discussion on this topic, see also Mitra (1973b).

16.3 Usual Constraints

Proposition 16.2. *Consider the matrix* $\mathbf{X} \in \mathbb{R}^{n\times p}$, $\operatorname{rank}(\mathbf{X}) = r$, *and let* $\mathbf{L}$ *be an* $q \times p$ *matrix. Then the equation*

$$\begin{pmatrix}\mathbf{X}\\ \mathbf{L}\end{pmatrix} \boldsymbol{\beta} = \begin{pmatrix}\boldsymbol{\mu}\\ \mathbf{0}\end{pmatrix} \tag{16.21}$$

has a unique solution for $\boldsymbol{\beta}$ *for every given vector* $\boldsymbol{\mu} \in \mathscr{C}(\mathbf{X})$ *if and only if the following two conditions hold:*

(a) $\mathscr{C}(\mathbf{X}') \cap \mathscr{C}(\mathbf{L}') = \{\mathbf{0}\}$,

(b) $\operatorname{rank}(\mathbf{X}' : \mathbf{L}') = p$.

Proof. Equation (16.21) has a solution for $\boldsymbol{\beta}$ if and only if

$$\begin{pmatrix}\boldsymbol{\mu}\\ \mathbf{0}\end{pmatrix} \in \mathscr{C}\begin{pmatrix}\mathbf{X}\\ \mathbf{L}\end{pmatrix}, \quad \text{i.e.,} \quad \mathbf{P}_{\left(\begin{smallmatrix}\mathbf{X}\\ \mathbf{L}\end{smallmatrix}\right)}\begin{pmatrix}\boldsymbol{\mu}\\ \mathbf{0}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\mu}\\ \mathbf{0}\end{pmatrix}, \tag{16.22}$$

which can be written as

$$\begin{pmatrix}\mathbf{X}\\ \mathbf{L}\end{pmatrix} (\mathbf{X}'\mathbf{X} + \mathbf{L}'\mathbf{L})^{-}(\mathbf{X}' : \mathbf{L}') \begin{pmatrix}\boldsymbol{\mu}\\ \mathbf{0}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\mu}\\ \mathbf{0}\end{pmatrix}. \tag{16.23}$$

Requiring (16.23) to hold for every $\boldsymbol{\mu} \in \mathscr{C}(\mathbf{X})$ yields

$$\begin{pmatrix}\mathbf{X}\\ \mathbf{L}\end{pmatrix} (\mathbf{X}'\mathbf{X} + \mathbf{L}'\mathbf{L})^{-}(\mathbf{X}' : \mathbf{L}') \begin{pmatrix}\mathbf{X}\\ \mathbf{0}\end{pmatrix} = \begin{pmatrix}\mathbf{X}\\ \mathbf{0}\end{pmatrix}. \tag{16.24}$$

In light of Theorem 16 (p. 343), (16.24) holds if and only if $\mathscr{C}(\mathbf{X}') \cap \mathscr{C}(\mathbf{L}') = \{\mathbf{0}\}$. Moreover, equation (16.21) has a unique solution if and only if

$$\operatorname{rank}(\mathbf{X}' : \mathbf{L}') = \operatorname{rank}(\mathbf{X}) + \operatorname{rank}(\mathbf{L}) - \dim \mathscr{C}(\mathbf{X}') \cap \mathscr{C}(\mathbf{L}') = p\,. \tag{16.25}$$

Thus the proposition is proved. □

Proposition 16.2 is useful when imposing "usual" constraints on the normal equation $\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}$ in order to obtain a unique solution; see, e.g., Searle (1971, §5.7), Seber (1980, §3.4, Appendix 1), and Monahan (2008, §3.8). We note that if we want to find $\mathbf{L}$ so that there is only one solution to $\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}$ which satisfies the constraint $\mathbf{L}\boldsymbol{\beta} = \mathbf{0}$, we need to consider the equation

$$\begin{pmatrix}\mathbf{X}'\mathbf{X}\\ \mathbf{L}\end{pmatrix} \boldsymbol{\beta} = \begin{pmatrix}\mathbf{X}'\mathbf{y}\\ \mathbf{0}\end{pmatrix}, \tag{16.26}$$

which is equivalent (why?) to

$$\begin{pmatrix}\mathbf{X}\\ \mathbf{L}\end{pmatrix} \boldsymbol{\beta} = \begin{pmatrix}\mathbf{P}_{\mathbf{X}}\mathbf{y}\\ \mathbf{0}\end{pmatrix}. \tag{16.27}$$

Once the vector $\mathbf{y}$ is given, $\mathbf{P}_{\mathbf{X}}\mathbf{y}$ becomes a vector $\boldsymbol{\mu}$, say, belonging to $\mathscr{C}(\mathbf{X})$. Finding $\mathbf{L}$ so that whatever the value of $\boldsymbol{\mu} \in \mathscr{C}(\mathbf{X})$ is, then the solution $\boldsymbol{\beta}$ for (16.27) is unique, yields the task whose solution is given in Proposition 16.2. The unique solution to (16.27) is $(\mathbf{X}'\mathbf{X} + \mathbf{L}'\mathbf{L})^{-1}\mathbf{X}'\mathbf{y}$.

16.4 Exercises

16.1. Complete the proof of Theorem 16 (p. 343).

16.2. Consider the block model given in Example 0.3 (p. 31):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} = (\mathbf{1}_n : \mathbf{T} : \mathbf{B}) \begin{pmatrix} \mu \\ \boldsymbol{\tau} \\ \boldsymbol{\beta} \end{pmatrix} + \boldsymbol{\varepsilon}\,, \text{ where } \mathbf{T} \in \mathbb{R}^{n\times t},\ \mathbf{B} \in \mathbb{R}^{n\times b}.$$

A block design is said to be *connected* if all elementary contrasts of treatment effects, i.e., those of form $\tau_i - \tau_j$, are estimable. The matrix $\mathbf{T}'\mathbf{Q}_{\mathbf{B}}\mathbf{T}$ is often called a C-matrix: $\mathbf{C}_d = \mathbf{T}'\mathbf{Q}_{\mathbf{B}}\mathbf{T}$. Confirm the following:

(a) $\boldsymbol{\ell}'\boldsymbol{\tau}$ is estimable $\iff$ $\boldsymbol{\ell} \in \mathscr{C}(\mathbf{T}'\mathbf{Q}_{\mathbf{B}}) = \mathscr{C}(\mathbf{T}'\mathbf{Q}_{\mathbf{B}}\mathbf{T})$.

(b) $\boldsymbol{\ell}'\boldsymbol{\tau}$ is estimable $\implies$ $\boldsymbol{\ell}'\boldsymbol{\tau}$ is a treatment contrast.

(c) The block design is connected $\iff$ $\operatorname{rank}(\mathbf{T}'\mathbf{Q}_{\mathbf{B}}\mathbf{T}) = t - 1$ $\iff$ $\dim \mathscr{C}(\mathbf{T}) \cap \mathscr{C}(\mathbf{B}) = 1$.

Bapat (2000, pp. 99-103).

16.3. Consider the same situation as in Exercise 16.2 above but denote, for convenience, $\mathbf{X}_1 = \mathbf{T}_{n\times t}$ and $\mathbf{X}_2 = \mathbf{B}_{n\times b}$, and assume that we have the linear model $\{\mathbf{y},\ \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2,\ \sigma^2\mathbf{I}_n\}$. Confirm that

$$\operatorname{cov}\begin{pmatrix} \mathbf{X}_1\mathbf{y} \\ \mathbf{X}_2\mathbf{y} \end{pmatrix} = \sigma^2 \begin{pmatrix} \mathbf{D}_r & \mathbf{N} \\ \mathbf{N}' & \mathbf{D}_c \end{pmatrix},$$

where $\mathbf{D}_r = \operatorname{diag}(\mathbf{r})$, $\mathbf{D}_c = \operatorname{diag}(\mathbf{c})$, and $\mathbf{N} = \{n_{ij}\} \in \mathbb{R}^{t\times b}$ are defined as in Exercise 0.21 (p. 51). Show that the canonical correlations between $\mathbf{X}_1\mathbf{y}$ and $\mathbf{X}_2\mathbf{y}$ are the singular values of the $t \times b$ matrix $\mathbf{D}_t^{-1/2}\mathbf{N}\mathbf{D}_b^{-1/2} = \{n_{ij}/\sqrt{r_i c_j}\}$. Moreover, show that the block design in this situation is connected if there is precisely one unit canonical correlation between $\mathbf{X}_1\mathbf{y}$ and $\mathbf{X}_2\mathbf{y}$. (See also Exercise 19.12, page 412.)

Styan (1985, §2.2), Puntanen & Styan (2005b, §6.4).

Chapter 17
Full Rank Decomposition

> *The lion and the calf shall lie down together*
> *but the calf won't get much sleep.*
>
> Woody Allen: *The Scrolls*

This chapter shows how helpful it is to express a matrix $\mathbf{A}$ as a product $\mathbf{UV}'$ where both $\mathbf{U}$ and $\mathbf{V}$ have full column ranks.

Theorem 17 (Full rank decomposition). *Let* $\mathbf{A}$ *be an* $n \times m$ *matrix with rank* $r > 0$. *Then* $\mathbf{A}$ *can be written as a product*

$$\mathbf{A} = \mathbf{UV}', \tag{17.1}$$

where

$$\operatorname{rank}(\mathbf{U}_{n\times r}) = \operatorname{rank}(\mathbf{V}_{m\times r}) = r\,, \tag{17.2}$$

i.e., $\mathbf{U}$ *and* $\mathbf{V}$ *have full column ranks.*

Proof. Let $\mathbf{U}$ be an $n \times r$ matrix whose columns form a basis for the column space of $\mathbf{A}$. Then every vector in the column space of $\mathbf{A}$ can be expressed as a linear combination of the columns of $\mathbf{U}$. In particular, every column $\mathbf{a}_i$ of $\mathbf{A}$ can be written as

$$\mathbf{a}_i = \mathbf{U}\mathbf{v}_i\,, \quad i = 1, \dots, m\,, \tag{17.3}$$

for some $\mathbf{v}_i \in \mathbb{R}^r$ (which is unique after fixing the basis $\mathbf{U}$). Hence there exists a matrix $\mathbf{V}_{m\times r}$ such that

$$\mathbf{A} = \mathbf{UV}' = \mathbf{U}(\mathbf{v}_1 : \dots : \mathbf{v}_m)\,. \tag{17.4}$$

Because

$$r = \operatorname{rank}(\mathbf{A}) = \operatorname{rank}(\mathbf{UV}') \le \operatorname{rank}(\mathbf{V}') \le r\,, \tag{17.5}$$

we observe that $\mathbf{V}$ has full column rank, and the theorem is proved. □

We note that in the full rank decomposition (17.1) the columns of $\mathbf{U}$ form a basis for the column space $\mathscr{C}(\mathbf{A})$, and clearly the columns of $\mathbf{V}$ form a basis for $\mathscr{C}(\mathbf{A}')$:

$$\mathscr{C}(\mathbf{U}) = \mathscr{C}(\mathbf{A})\,, \quad \mathscr{C}(\mathbf{V}) = \mathscr{C}(\mathbf{A}')\,. \tag{17.6}$$

Sometimes it is convenient to choose the columns of $\mathbf{U}$ or $\mathbf{V}$ orthonormal.

S. Puntanen et al., *Matrix Tricks for Linear Statistical Models: Our Personal Top Twenty*,
DOI 10.1007/978-3-642-10473-2_18,

It is worth emphasizing that the full rank decomposition is (like also some other tricks in this book) mathematically very simple—but it can be an amazingly handy tool at appropriate situations. As references, we may mention Marsaglia & Styan (1974a, Th. 1), Bhimasankaram (1988), and Piziak & Odell (1999).

Photograph 17.1 Pochiraju Bhimasankaram (Hyderabad, 2007).

17.1 Some Properties of an Idempotent Matrix

In this section we consider three properties of an idempotent matrix $\mathbf{A}_{n\times n}$ that can be easily proved using the full rank decomposition

$$\mathbf{A} = \mathbf{U}\mathbf{V}', \tag{17.7}$$

where $\operatorname{rank}(\mathbf{U}_{n\times r}) = \operatorname{rank}(\mathbf{V}_{n\times r}) = r$. The first property is the following:

Proposition 17.1. *With the above notation,*

$$\mathbf{A} = \mathbf{A}^2 \iff \mathbf{V}'\mathbf{U} = \mathbf{I}_r\,. \tag{17.8}$$

Proof. To prove (17.8), we first assume that $\mathbf{A}$ is idempotent:

$$\mathbf{U}\mathbf{V}' = \mathbf{U}\mathbf{V}'\mathbf{U}\mathbf{V}', \tag{17.9}$$

Premultiplying (17.9) by the matrix $(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'$ yields the equality

$$\mathbf{V}' = \mathbf{V}'\mathbf{U}\mathbf{V}', \tag{17.10}$$

Now postmultiplying (17.10) by $\mathbf{V}(\mathbf{V}'\mathbf{V})^{-1}$ gives our claim $\mathbf{V}'\mathbf{U} = \mathbf{I}_r$. On the other hand, if $\mathbf{V}'\mathbf{U} = \mathbf{I}_r$, then

$$\mathbf{A}^2 = \mathbf{U}\mathbf{V}'\mathbf{U}\mathbf{V}' = \mathbf{U}\mathbf{I}_r\mathbf{V}' = \mathbf{A}\,, \tag{17.11}$$

and thus (17.8) is proved. □

As a second result, we prove the following implication:

Proposition 17.2.

$$\mathbf{A} = \mathbf{A}^2 \implies \operatorname{rank}(\mathbf{A}) = \operatorname{tr}(\mathbf{A})\,. \tag{17.12}$$

Proof. This comes at once from (17.8):

$$\operatorname{tr}(\mathbf{A}) = \operatorname{tr}(\mathbf{U}\mathbf{V}') = \operatorname{tr}(\mathbf{V}'\mathbf{U}) = \operatorname{tr}(\mathbf{I}_r) = r = \operatorname{rank}(\mathbf{A})\,. \tag{17.13}$$

□

Also the following result can be easily proved using the full rank decomposition (see Groß, Trenkler & Troschke 1997):

Proposition 17.3. *Let* $\mathbf{A} = \mathbf{A}^2$. *Then*

$$\mathbf{A} = \mathbf{A}' \iff \mathscr{C}(\mathbf{A}) = \mathscr{C}(\mathbf{A}')\,. \tag{17.14}$$

Proof. The idempotency of $\mathbf{A}$ implies that $\mathbf{A}$ has a full rank decomposition $\mathbf{A} = \mathbf{U}\mathbf{V}'$, where $\mathbf{V}'\mathbf{U} = \mathbf{I}_r$, and

$$\mathscr{C}(\mathbf{A}) = \mathscr{C}(\mathbf{U})\,, \quad \mathscr{C}(\mathbf{A}') = \mathscr{C}(\mathbf{V})\,. \tag{17.15}$$

If $\mathscr{C}(\mathbf{A}) = \mathscr{C}(\mathbf{A}')$, then the orthogonal projectors onto $\mathscr{C}(\mathbf{A}) = \mathscr{C}(\mathbf{U})$ and onto $\mathscr{C}(\mathbf{A}') = \mathscr{C}(\mathbf{V})$ must be identical, i.e.,

$$\mathbf{U}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}' = \mathbf{V}(\mathbf{V}'\mathbf{V})^{-1}\mathbf{V}'. \tag{17.16}$$

Premultiplying (17.16) by $\mathbf{U}\mathbf{V}'$ we obtain, using $\mathbf{V}'\mathbf{U} = \mathbf{I}_r$,

$$\mathbf{U}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}' = \mathbf{U}\mathbf{V}' = \mathbf{A}\,, \tag{17.17}$$

and so $\mathbf{A}$ indeed is symmetric. □

17.2 Rank Additivity

We prove the following result:

Proposition 17.4. *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be non-null* $n \times m$ *matrices, and let* $\operatorname{rank}(\mathbf{A}) = a$, $\operatorname{rank}(\mathbf{B}) = b$. *Then the following statements are equivalent:*

(a) $\operatorname{rank}(\mathbf{A} + \mathbf{B}) = \operatorname{rank}(\mathbf{A}) + \operatorname{rank}(\mathbf{B})$,

(b) $\dim \mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B}) = \dim \mathscr{C}(\mathbf{A}') \cap \mathscr{C}(\mathbf{B}') = 0$.

Proof. Assume first that (a) holds. Then, in view of

$$\begin{aligned} \operatorname{rk}(\mathbf{A}) + \operatorname{rk}(\mathbf{B}) &= \operatorname{rk}(\mathbf{A} + \mathbf{B}) = \operatorname{rk}\left[(\mathbf{A} : \mathbf{B}) \begin{pmatrix} \mathbf{I}_m \\ \mathbf{I}_m \end{pmatrix}\right] \\ &\le \operatorname{rk}(\mathbf{A} : \mathbf{B}) = \operatorname{rk}(\mathbf{A}) + \operatorname{rk}(\mathbf{B}) - \dim \mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B})\,, \qquad (17.18\text{a}) \\ \operatorname{rk}(\mathbf{A}) + \operatorname{rk}(\mathbf{B}) &= \operatorname{rk}(\mathbf{A} + \mathbf{B}) = \operatorname{rk}\left[(\mathbf{I}_n : \mathbf{I}_n) \begin{pmatrix} \mathbf{A} \\ \mathbf{B} \end{pmatrix}\right] \\ &\le \operatorname{rk}\begin{pmatrix} \mathbf{A} \\ \mathbf{B} \end{pmatrix} = \operatorname{rk}(\mathbf{A}) + \operatorname{rk}(\mathbf{B}) - \dim \mathscr{C}(\mathbf{A}') \cap \mathscr{C}(\mathbf{B}')\,, \qquad (17.18\text{b}) \end{aligned}$$

it is clear that (a) implies (b).

To go the other way round, let $\mathbf{A}$ and $\mathbf{B}$ have the full rank decompositions

$$\mathbf{A} = \mathbf{A}_1\mathbf{A}_2' , \quad \mathbf{B} = \mathbf{B}_1\mathbf{B}_2' , \tag{17.19}$$

where

$$\mathscr{C}(\mathbf{A}) = \mathscr{C}(\mathbf{A}_1) ,\ \mathbf{A}_1 \in \mathbb{R}^{n\times a}, \quad \mathscr{C}(\mathbf{B}) = \mathscr{C}(\mathbf{B}_1) ,\ \mathbf{B}_1 \in \mathbb{R}^{n\times b}, \tag{17.20a}$$

$$\mathscr{C}(\mathbf{A}') = \mathscr{C}(\mathbf{A}_2) ,\ \mathbf{A}_2 \in \mathbb{R}^{m\times a}, \quad \mathscr{C}(\mathbf{B}') = \mathscr{C}(\mathbf{B}_2) ,\ \mathbf{B}_2 \in \mathbb{R}^{m\times b}. \tag{17.20b}$$

Then

$$\mathbf{A} + \mathbf{B} = \mathbf{A}_1\mathbf{A}_2' + \mathbf{B}_1\mathbf{B}_2' = (\mathbf{A}_1 : \mathbf{B}_1) \begin{pmatrix} \mathbf{A}_2' \\ \mathbf{B}_2' \end{pmatrix} := \mathbf{U}\mathbf{V}'. \tag{17.21}$$

In view of the disjointness assumption $\mathscr{C}(\mathbf{A}_2) \cap \mathscr{C}(\mathbf{B}_2) = \{\mathbf{0}\}$, we have

$$\operatorname{rk}(\mathbf{V}_{m\times(a+b)}) = \operatorname{rk}(\mathbf{A}_2 : \mathbf{B}_2) = \operatorname{rk}(\mathbf{A}_2) + \operatorname{rk}(\mathbf{B}_2) = a + b , \tag{17.22}$$

which means that $\mathbf{V}$ has full column rank. This further implies that $\operatorname{rk}(\mathbf{U}) = \operatorname{rk}(\mathbf{U}\mathbf{V}')$ since

$$\operatorname{rk}(\mathbf{U}) \geq \operatorname{rk}(\mathbf{U}\mathbf{V}') \geq \operatorname{rk}[\mathbf{U}\mathbf{V}'\mathbf{V}(\mathbf{V}'\mathbf{V})^{-1}] = \operatorname{rk}(\mathbf{U}) . \tag{17.23}$$

Hence we have

$$\operatorname{rk}(\mathbf{A} + \mathbf{B}) = \operatorname{rk}(\mathbf{U}\mathbf{V}') = \operatorname{rk}(\mathbf{U}) = \operatorname{rk}(\mathbf{A}_1 : \mathbf{B}_1) , \tag{17.24}$$

which, in light of the disjointness assumption $\mathscr{C}(\mathbf{A}_1) \cap \mathscr{C}(\mathbf{B}_1) = \{\mathbf{0}\}$, yields

$$\operatorname{rk}(\mathbf{A} + \mathbf{B}) = \operatorname{rk}(\mathbf{A}_1 : \mathbf{B}_1) = \operatorname{rk}(\mathbf{A}_1) + \operatorname{rk}(\mathbf{B}_1) , \tag{17.25}$$

and hence our claim is proved. ☐

As references to Proposition 17.4, we may mention Marsaglia & Styan (1972), and Marsaglia & Styan (1974a, Th. 11); see also Rao & Bhimasankaram (2000, p. 132).

17.3 Cochran's Theorem: a Simple Version

In this example we consider the following simple version of the Cochran's Theorem; for extended versions, see, e.g., Marsaglia & Styan (1974a), Anderson & Styan (1982), and Bapat (2000, p. 60).

Proposition 17.5. *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be* $n \times n$ *matrices satisfying the condition*

$$\mathbf{A} + \mathbf{B} = \mathbf{I}_n . \tag{17.26}$$

Then the following statements are equivalent:

(a) $\operatorname{rank}(\mathbf{A}) + \operatorname{rank}(\mathbf{B}) = n$,

(b) $\mathbf{A}^2 = \mathbf{A}$ *and* $\mathbf{B}^2 = \mathbf{B}$,

(c) $\mathbf{AB} = \mathbf{0}$.

Proof. "(c) $\implies$ (b)". If (c) holds, then, in view of (17.26), $\mathbf{B} = \mathbf{I}_n - \mathbf{A}$, and

$$\mathbf{AB} = \mathbf{A}(\mathbf{I}_n - \mathbf{A}) = \mathbf{A} - \mathbf{A}^2 = \mathbf{0}\,, \tag{17.27}$$

and hence $\mathbf{A} = \mathbf{A}^2$ (and similarly $\mathbf{B} = \mathbf{B}^2$).

"(b) $\implies$ (a)". If (b) holds, then $\operatorname{rank}(\mathbf{A}) = \operatorname{tr}(\mathbf{A})$ and $\operatorname{rank}(\mathbf{B}) = \operatorname{tr}(\mathbf{B})$, and (17.26) implies

$$n = \operatorname{tr}(\mathbf{I}_n) = \operatorname{tr}(\mathbf{A} + \mathbf{B}) = \operatorname{tr}(\mathbf{A}) + \operatorname{tr}(\mathbf{B}) = \operatorname{rank}(\mathbf{A}) + \operatorname{rank}(\mathbf{B})\,. \tag{17.28}$$

"(a) $\implies$ (c)". We assume that $a + b = n$, when $a = \operatorname{rank}(\mathbf{A})$, and $b = \operatorname{rank}(\mathbf{B}) = n - a$. Consider the full rank decompositions of $\mathbf{A}$ and $\mathbf{B}$:

$$\mathbf{A} = \mathbf{A}_1\mathbf{A}_2'\,, \quad \mathbf{A}_1 \in \mathbb{R}^{n\times a}\,, \qquad \mathbf{B} = \mathbf{B}_1\mathbf{B}_2'\,, \quad \mathbf{B}_1 \in \mathbb{R}^{n\times(n-a)}\,, \tag{17.29}$$

which means that

$$\mathbf{A}_1\mathbf{A}_2' + \mathbf{B}_1\mathbf{B}_2' = (\mathbf{A}_1 : \mathbf{B}_1)\begin{pmatrix}\mathbf{A}_2'\\ \mathbf{B}_2'\end{pmatrix} := \mathbf{FG} = \mathbf{I}_n\,. \tag{17.30}$$

Because $(\mathbf{A}_1 : \mathbf{B}_1) = \mathbf{F}$ is an $n \times n$ matrix satisfying the equation $\mathbf{FG} = \mathbf{I}_n$, the matrix $\mathbf{G}$ is the inverse of $\mathbf{F}$, and hence satisfies the equation

$$\mathbf{GF} = \begin{pmatrix}\mathbf{A}_2'\\ \mathbf{B}_2'\end{pmatrix}(\mathbf{A}_1 : \mathbf{B}_1) = \mathbf{I}_n\,, \tag{17.31}$$

that is,

$$\begin{pmatrix}\mathbf{A}_2'\\ \mathbf{B}_2'\end{pmatrix}(\mathbf{A}_1 : \mathbf{B}_1) = \begin{pmatrix}\mathbf{A}_2'\mathbf{A}_1 & \mathbf{A}_2'\mathbf{B}_1\\ \mathbf{B}_2'\mathbf{A}_1 & \mathbf{B}_2'\mathbf{B}_1\end{pmatrix} = \begin{pmatrix}\mathbf{I}_a & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_{n-a}\end{pmatrix} = \mathbf{I}_n\,. \tag{17.32}$$

The final claim $\mathbf{AB} = \mathbf{0}$ is achieved when the equation

$$\mathbf{A}_2'\mathbf{B}_1 = \mathbf{0} \tag{17.33}$$

is premultiplied by $\mathbf{A}_1$ and postmultiplied by $\mathbf{B}_2'$. □

17.4 Proof of the RCR Using the Full Rank Decomposition

In Chapter 6 (p. 145) we have already proved the rank cancellation rule:

$$\mathbf{LAY} = \mathbf{MAY} \text{ and } \operatorname{rank}(\mathbf{AY}) = \operatorname{rank}(\mathbf{A}) \implies \mathbf{LA} = \mathbf{MA}\,, \tag{17.34}$$

but it is of interest to give a proof using the full rank decomposition (as done in Marsaglia & Styan 1974a). To do this, let $\mathbf{A}_{n\times m}$ have a full rank decomposition

$$\mathbf{A} = \mathbf{UV}', \tag{17.35}$$

where

$$\operatorname{rank}(\mathbf{U}_{n\times r}) = \operatorname{rank}(\mathbf{V}_{m\times r}) = r = \operatorname{rank}(\mathbf{A}_{n\times m})\,. \tag{17.36}$$

Assumption $\mathbf{LAY} = \mathbf{MAY}$ can be written as

$$\mathbf{L}\cdot\mathbf{UV}'\cdot\mathbf{Y} = \mathbf{M}\cdot\mathbf{UV}'\cdot\mathbf{Y}\,, \quad \text{i.e.,} \quad \mathbf{LU}(\mathbf{V}'\mathbf{Y}) = \mathbf{MU}(\mathbf{V}'\mathbf{Y})\,. \tag{17.37}$$

Therefore, if the $r\times p$ matrix $\mathbf{V}'\mathbf{Y}$ has full row rank, then we can postmultiply (17.37) by $(\mathbf{V}'\mathbf{Y})'[(\mathbf{V}'\mathbf{Y})(\mathbf{V}'\mathbf{Y})']^{-1}$ and obtain the equality

$$\mathbf{LU} = \mathbf{MU}\,. \tag{17.38}$$

Our claim would then follow by postmultiplying (17.38) by $\mathbf{V}'$. Our task is, therefore, to show that

$$\operatorname{rank}(\mathbf{V}'\mathbf{Y}) = r\,. \tag{17.39}$$

Assumption $\operatorname{rank}(\mathbf{AY}) = \operatorname{rank}(\mathbf{A})$ implies that $\operatorname{rank}(\mathbf{UV}'\mathbf{Y}) = r$. but since $\mathbf{U}$ has full column rank, we get

$$\operatorname{rank}(\mathbf{V}'\mathbf{Y}) \geq \operatorname{rank}(\mathbf{UV}'\mathbf{Y}) \geq \operatorname{rank}[(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{UV}'\mathbf{Y}] \geq \operatorname{rank}(\mathbf{V}'\mathbf{Y})\,, \tag{17.40}$$

and hence indeed (17.39) holds.

17.5 Exercises

17.1. Let $\mathbf{A}$ have a full rank decomposition $\mathbf{A} = \mathbf{UV}'$. Prove the claim (4.13) (p. 107): $\mathbf{A}^{+} = \mathbf{V}(\mathbf{V}'\mathbf{V})^{-1}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'$.

17.2. Proposition 17.5 (p. 352) can be generalized as follows: Let $\mathbf{A}_1, \dots, \mathbf{A}_m$ be $n\times n$ matrices such that $\mathbf{I}_n = \mathbf{A}_1 + \cdots + \mathbf{A}_m$. Then the following three conditions are equivalent:

(a) $n = \operatorname{rank}(\mathbf{A}_1) + \cdots + \operatorname{rank}(\mathbf{A}_m)$,

(b) $\mathbf{A}_i^2 = \mathbf{A}_i$ for $i = 1, \dots, m$,
(c) $\mathbf{A}_i\mathbf{A}_j = \mathbf{0}$ for all $i \neq j$.

Confirm the following: Let $\mathbf{z} \sim \mathrm{N}_n(\boldsymbol{\mu}, \mathbf{I}_n)$ and let $\mathbf{z}'\mathbf{z} = \mathbf{z}'\mathbf{A}_1\mathbf{z} + \cdots + \mathbf{z}'\mathbf{A}_m\mathbf{z}$. Then any of the three above conditions is a necessary and sufficient condition for $\mathbf{z}'\mathbf{A}_i\mathbf{z}$ to be independently distributed as $\chi^2[\mathrm{rank}(\mathbf{A}_i), \cdot]$. For the χ^2-distribution, see page 18.

17.3. Let

$$\mathbf{W} = \mathbf{I}_n + \mathbf{X} - 2\mathbf{J}\,,$$

where $\mathbf{X}$ is an $n \times n$ symmetric involutory doubly-stochastic matrix, i.e., $\mathbf{X}^2 = \mathbf{I}_n$ and $\mathbf{X}\mathbf{1}_n = \mathbf{1}_n$, and $\mathbf{J} = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'$.

(a) Show that $\mathbf{W}$ is *scalar-potent,* i.e., $\mathbf{W}^2 = c\mathbf{W}$ for some scalar c.
(b) Find the scalar c and confirm that $\mathbf{W}$ is nonnegative definite.
(c) Find the $\mathrm{rank}(\mathbf{W})$ as a function of the trace $\mathrm{tr}(\mathbf{X})$ and hence show that $\mathrm{tr}(\mathbf{X})$ is even if and only if n is even.
(d) Show that the $\mathrm{rank}(\mathbf{W}) = 1$ if and only if $n + \mathrm{tr}(\mathbf{X}) = 4$. [When $\mathrm{rank}(\mathbf{W}) = 1$ then the matrix $\mathbf{I}_n - \mathbf{W}$ is a Householder transformation, see Exercise 18.23 (p. 390).]

For an application of the results in this exercise to magic squares see Chu, Drury, Styan & Trenkler (2010).

17.4. (a) Let $\mathbf{A}$ and $\mathbf{B}$ be symmetric $n \times n$ idempotent matrices and let $\mathbf{Z}$ be $n \times n$ nonnegative definite, such that $\mathbf{A} + \mathbf{Z} = \mathbf{B}$. Show that $\mathbf{Z}$ is idempotent and $\mathbf{AZ} = \mathbf{0}$.
(b) Let the random variables x_1 and x_2 follow central chi-squared distributions with degrees of freedom f_1 and f_2, respectively, such that $x_1 - x_2 = x_3 \geq 0$ with probability 1. Then show that x_3 follows a chi-squared distribution and find the number of degrees of freedom. Show also that x_2 and x_3 are independently distributed.
The results in this exercise may be called the *Hogg–Craig theorem,* following results in Hogg & Craig (1958); see also Ogasawara & Takahashi (1951) and Styan (1970).

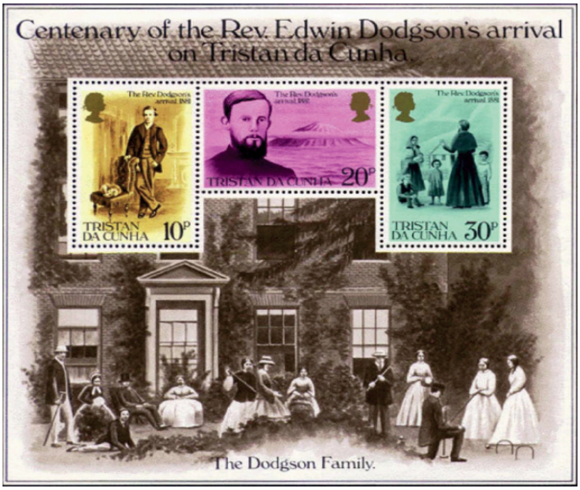
17.5 (Group inverse). Let $\mathbf{A}$ be an $n \times n$ matrix such that $\mathrm{rank}(\mathbf{A}) = \mathrm{rank}(\mathbf{A}^2)$. Then the *group inverse* $\mathbf{A}^{\#}$ is the unique matrix satisfying

$$\text{(a) } \mathbf{A}\mathbf{A}^{\#}\mathbf{A} = \mathbf{A}, \quad \text{(b) } \mathbf{A}^{\#}\mathbf{A}\mathbf{A}^{\#} = \mathbf{A}^{\#}, \quad \text{and} \quad \text{(c) } \mathbf{A}^{\#}\mathbf{A}^2 = \mathbf{A}.$$

If $\mathbf{A}$ has full-rank decomposition $\mathbf{UV}'$ show that $\mathbf{V}'\mathbf{U}$ is nonsingular and that $\mathbf{A}^{\#} = \mathbf{U}(\mathbf{V}'\mathbf{U})^{-2}\mathbf{V}'$.

For more about the group inverse see Ben-Israel & Greville (2003, pp. 156–161).

Philatelic Item 17.1 As observed by Tee (2003), “Determinants were applied in 1683 by the Japanese mathematician Takakazu Seki Kôwa (1642–1708) in the construction of the resolvent of a system of polynomial equations [but] were independently invented in 1693 by Gottfried Wilhelm von Leibniz (1646–1716)”; see also Farebrother, Styan & Tee (2003). Leibniz was a German mathematician and philosopher who invented infinitesimal calculus independently of Isaac Newton (1643–1727). The stamp (left panel) for Seki was issued by Japan in 1992 (*Scott* 2147) and the stamp (right panel) for Leibniz by St. Vincent in 1991 (*Scott* 1557) as “Head librarian for the electors of Hannover (& co-inventor of calculus)”.



Philatelic Item 17.2 Charles Lutwidge Dodgson (1832–1898), better known by the pen name Lewis Carroll, was an English author, mathematician, Anglican clergyman and photographer. As a mathematician, Dodgson was the author of *Condensation of Determinants* (1866) and *Elementary Treatise on Determinants* (1867). His most famous writings, however, are *Alice's Adventures in Wonderland* and its sequel *Through the Looking-Glass*. The sheetlet was issued by Tristan da Cunha in 1981 (*Scott* 287a) and shows the Dodgson family outside the Croft Rectory (in Darlington, Yorkshire), c. 1860; Charles is shown seated on the ground at the left of the group. Edwin Heron Dodgson (1846–1918), Charles's youngest brother (shown kneeling third from the right), apparently saved the population of Tristan da Cunha from starvation. For more see Farebrother, Jensen & Styan (2000).

Chapter 18
Eigenvalue Decomposition

> *The late night slowed him down just enough to give a lecture the whole class could follow.*
>
> IMS Bulletin[a]: *About Walter T. Federer, after he had gone downhill skiing all weekend and had not gotten back until after midnight.*
>
> ---
>
> [a] McCulloch, Hedayat & Wells (2008, p. 13)

There is no way to survive in the middle of statistical considerations without being pretty well aware of the main properties of the eigenvalues and eigenvectors. This chapter provides a summary of some central results.

Theorem 18 (Eigenvalue decomposition). *Let* $\mathbf{A}$ *be an* $n \times n$ *symmetric matrix. Then* $\mathbf{A}$ *can be written as a product*

$$\mathbf{A} = \mathbf{T\Lambda T}' = \lambda_1 \mathbf{t}_1 \mathbf{t}_1' + \cdots + \lambda_n \mathbf{t}_n \mathbf{t}_n' \,, \tag{EVD}$$

and thereby

$$(\mathbf{At}_1 : \mathbf{At}_2 : \ldots : \mathbf{At}_n) = (\lambda_1 \mathbf{t}_1 : \lambda_2 \mathbf{t}_2 : \ldots : \lambda_n \mathbf{t}_n) \,, \quad \mathbf{AT} = \mathbf{\Lambda T} \,, \tag{18.1}$$

where $\mathbf{T}_{n\times n}$ *is orthogonal,* $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$*, and* $\lambda_1 \geq \cdots \geq \lambda_n$ *are the ordered eigenvalues of* $\mathbf{A}$*; we denote* $\mathrm{ch}_i(\mathbf{A}) = \lambda_i$*. The columns* $\mathbf{t}_i$ *of* $\mathbf{T}$ *are the orthonormal eigenvectors of* $\mathbf{A}$*.*

Consider the distinct eigenvalues of $\mathbf{A}$*,* $\lambda_{\{1\}} > \cdots > \lambda_{\{s\}}$*, and let* $\mathbf{T}_{\{i\}}$ *be an* $n \times m_i$ *matrix consisting of the orthonormal eigenvectors corresponding to* $\lambda_{\{i\}}$*;* m_i *is the multiplicity of* $\lambda_{\{i\}}$*. Then*

$$\mathbf{A} = \mathbf{T\Lambda T}' = \lambda_{\{1\}} \mathbf{T}_{\{1\}} \mathbf{T}_{\{1\}}' + \cdots + \lambda_{\{s\}} \mathbf{T}_{\{s\}} \mathbf{T}_{\{s\}}' \,. \tag{18.2}$$

With this ordering, $\mathbf{\Lambda}$ *is unique and* $\mathbf{T}$ *is unique up to postmultiplying by a blockdiagonal matrix* $\mathbf{U} = \mathrm{blockdiag}(\mathbf{U}_1, \ldots, \mathbf{U}_s)$*, where* $\mathbf{U}_i$ *is an orthogonal* $m_i \times m_i$ *matrix. If all the eigenvalues are distinct, then* $\mathbf{U}$ *is a diagonal matrix with diagonal elements equal to* ± 1*.*

In particular, for a nonnegative definite $n \times n$ *matrix* $\mathbf{A}$ *with rank* $r > 0$ *we have*

$$\begin{aligned} \mathbf{A} = \mathbf{T\Lambda T}' &= (\mathbf{T}_1 : \mathbf{T}_0) \begin{pmatrix} \mathbf{\Lambda}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{T}_1' \\ \mathbf{T}_0' \end{pmatrix} = \mathbf{T}_1 \mathbf{\Lambda}_1 \mathbf{T}_1' \\ &= \lambda_1 \mathbf{t}_1 \mathbf{t}_1' + \cdots + \lambda_r \mathbf{t}_r \mathbf{t}_r' \,, \end{aligned} \tag{18.3}$$

S. Puntanen et al., *Matrix Tricks for Linear Statistical Models: Our Personal Top Twenty*,
DOI 10.1007/978-3-642-10473-2_19, © Springer-Verlag Berlin Heidelberg 2011

where $\lambda_1 \geq \cdots \geq \lambda_r > 0$, $\mathbf{\Lambda}_1 = \mathrm{diag}(\lambda_1, \dots, \lambda_r)$, *and*

$$\mathbf{T}_1 = (\mathbf{t}_1 : \ldots : \mathbf{t}_r)\,, \quad \mathbf{T}_0 = (\mathbf{t}_{r+1} : \ldots : \mathbf{t}_n)\,. \tag{18.4}$$

The above eigenvalue decomposition, also called the spectral decomposition, is in heavy use in our book. In this chapter we go through some central properties of eigenvalues and eigenvectors—not proving everything. There is an enormous amount of literature on this area and we try to pick some results which are particularly useful for a statistician. For further details, including thorough proofs (which we skip over) of Theorem 18, the reader is referred, e.g., to Meyer (2000, Ch. 7) Harville (1997, Ch. 21), Horn & Johnson (1990, Ch. 1), Searle (1982, Ch. 11), Seber (2008, Ch. 6), and Stewart (2001, Ch. 1).

Before refreshing the reader's memory about the eigenvalues, we may mention two obvious but important consequences of (18.3):

$$\mathrm{tr}(\mathbf{A}) = \mathrm{tr}(\mathbf{T\Lambda T}') = \mathrm{tr}(\mathbf{T}'\mathbf{T\Lambda}) = \mathrm{tr}(\mathbf{\Lambda})\,, \tag{18.5a}$$

$$|\mathbf{A}| = |\mathbf{T\Lambda T}'| = |\mathbf{T}||\mathbf{\Lambda}||\mathbf{T}'| = |\mathbf{T}'||\mathbf{T}||\mathbf{\Lambda}| = |\mathbf{T}'\mathbf{T}||\mathbf{\Lambda}| = |\mathbf{\Lambda}|\,, \tag{18.5b}$$

where $|\cdot|$ denotes the determinant; recall that instead of $\det(\cdot)$ we may occasionally use notation $|\cdot|$ (to save space). It is worth mentioning that the properties (18.5) hold even if $\mathbf{A}$ is not symmetric.

We recall that the scalar λ (complex or real) is an eigenvalue of $\mathbf{A}_{n\times n}$ if

$$\mathbf{At} = \lambda\mathbf{t} \quad \text{for some nonzero vector } \mathbf{t}\,, \tag{18.6}$$

in which case $\mathbf{t}$ is an eigenvector of $\mathbf{A}$ corresponding to λ and $(\lambda, \mathbf{t})$ is an eigenpair for $\mathbf{A}$. In particular, let λ_i and λ_j be two distinct eigenvalues of a symmetric $\mathbf{A}$ with $\mathbf{t}_i$ and $\mathbf{t}_j$ being the corresponding eigenvectors. Then premultiplying $\mathbf{At}_i = \lambda_i\mathbf{t}_i$ by $\mathbf{t}_j'$ shows that necessarily $\mathbf{t}_i'\mathbf{t}_j = 0$. If $\lambda_{\{i\}}$ is a multiple root of a symmetric matrix, then, as stated in Theorem 18, it is possible to construct a set of m_i orthonormal eigenvectors corresponding to $\lambda_{\{i\}}$.

Writing (18.6) as

$$(\mathbf{A} - \lambda\mathbf{I}_n)\mathbf{t} = \mathbf{0}, \quad \mathbf{t} \neq \mathbf{0}\,, \tag{18.7}$$

we see that λ is an eigenvalue of $\mathbf{A}$ if and only if $\mathbf{A} - \lambda\mathbf{I}_n$ is a singular matrix, i.e., λ is a root for the characteristic equation

$$p_{\mathbf{A}}(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I}_n) = 0\,. \tag{18.8}$$

The function $p_{\mathbf{A}}(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I}_n)$ is called the characteristic polynomial of $\mathbf{A}$. Moreover, the set of all eigenvectors associated with λ is

$$\{\,\mathbf{t} \neq \mathbf{0} : (\mathbf{A} - \lambda\mathbf{I}_n)\mathbf{t} = \mathbf{0}\,\} = \{\,\mathbf{t} \neq \mathbf{0} : \mathbf{t} \in \mathscr{N}(\mathbf{A} - \lambda\mathbf{I}_n)\,\}\,. \tag{18.9}$$

The eigenvalues of $\mathbf{A}$ are the n roots of the characteristic equation (18.8). That these roots indeed exist, is not not a trivial matter, as Harville (1997,

p. 533) points out. The set of all n eigenvalues is called the spectrum of $\mathbf{A}$ and we denote it as

$$\operatorname{ch}(\mathbf{A}) = \{\lambda_1, \lambda_2, \dots, \lambda_n\}\,. \tag{18.10}$$

Some authors define the spectrum as a set of *distinct* eigenvalues, but in our notation, for example,

$$\operatorname{ch}(\mathbf{I}_2) = \{1, 1\}\,. \tag{18.11}$$

If $\mathbf{A}$ is symmetric, then all eigenvalues appear to be real and then we can order them so that

$$\operatorname{ch}_i(\mathbf{A}) = \lambda_i = i\text{th largest eigenvalue of } \mathbf{A}\,. \tag{18.12}$$

Notice, for example, that the characteristic equation $\mathbf{A} = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$ is $\lambda^2 + 1 = 0$, and hence the eigenvalues of $\mathbf{A}$ are not real numbers.

The characteristic polynomial can be uniquely decomposed as

$$p_{\mathbf{A}}(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I}_n) = (\lambda - \lambda_{\{1\}})^{m_1}(\lambda - \lambda_{\{2\}})^{m_2} \cdots (\lambda - \lambda_{\{s\}})^{m_s}, \tag{18.13}$$

where the numbers $\lambda_{\{i\}}$ are distinct and

$$m_1 + m_2 + \dots + m_s = n\,. \tag{18.14}$$

The *algebraic multiplicity* of the eigenvalue $\lambda_{\{i\}}$ is m_i. We may denote

$$\operatorname{alg\,mult}_{\mathbf{A}}(\lambda_{\{i\}}) = m_i\,. \tag{18.15}$$

We call

$$\mathscr{N}(\mathbf{A} - \lambda\mathbf{I}_n) = \text{the eigenspace of } \mathbf{A} \text{ corresponding to } \lambda\,. \tag{18.16}$$

The *geometric multiplicity* of λ is the dimension of the eigenspace of $\mathbf{A}$ corresponding to λ:

$$\operatorname{geo\,mult}_{\mathbf{A}}(\lambda) = \dim \mathscr{N}(\mathbf{A} - \lambda\mathbf{I}_n) = n - \operatorname{rank}(\mathbf{A} - \lambda\mathbf{I}_n)\,. \tag{18.17}$$

In other words, the geometric multiplicity of λ is the maximal number of linearly independent eigenvectors corresponding to the eigenvalue λ. It can be shown that

$$\operatorname{geo\,mult}_{\mathbf{A}}(\lambda) \le \operatorname{alg\,mult}_{\mathbf{A}}(\lambda) \quad \text{for each } \lambda \in \operatorname{ch}(\mathbf{A})\,, \tag{18.18}$$

where the equality holds e.g. when $\mathbf{A}$ is symmetric. Actually, the following statements concerning the matrix $\mathbf{A}_{n\times n}$ are equivalent:

$$\mathbf{A} \text{ is diagonalizable,} \tag{18.19a}$$

$$\operatorname{geo\,mult}_{\mathbf{A}}(\lambda) = \operatorname{geo\,alg}_{\mathbf{A}}(\lambda) \quad \text{for all } \lambda \in \operatorname{ch}(\mathbf{A})\,, \tag{18.19b}$$

$$\mathbf{A} \text{ has } n \text{ linearly independent eigenvectors.} \tag{18.19c}$$

The matrix $\mathbf{A}_{n\times n}$ is said to be diagonalizable whenever there exists a nonsingular matrix $\mathbf{F}_{n\times n}$ such that

$$\mathbf{F}^{-1}\mathbf{A}\mathbf{F} = \mathbf{D} \quad \text{for some diagonal matrix } \mathbf{D}_{n\times n}\,, \tag{18.20}$$

i.e., $\mathbf{A}$ is similar to a diagonal matrix. In particular, any symmetric $\mathbf{A}$ is diagonalizable. The matrices $\mathbf{A}_{n\times n}$ and $\mathbf{B}_{n\times n}$ are said to be similar if there exists a matrix $\mathbf{L}_{n\times n}$ such that

$$\mathbf{L}^{-1}\mathbf{A}\mathbf{L} = \mathbf{B}\,. \tag{18.21}$$

It is easy to confirm that

$$\mathbf{L}^{-1}\mathbf{A}\mathbf{L} = \mathbf{B} \implies \operatorname{ch}(\mathbf{A}) = \operatorname{ch}(\mathbf{B})\,. \tag{18.22}$$

We complete this short introductory review by mentioning a few well-known features of the eigenvalues. One very frequently used property of eigenvalues is that for any matrices $\mathbf{A}_{n\times p}$ and $\mathbf{B}_{p\times n}$, the products $\mathbf{AB}$ and $\mathbf{BA}$ have the same nonzero eigenvalues, see Proposition 13.2 (p. 299):

$$\operatorname{nzch}(\mathbf{AB}) = \operatorname{nzch}(\mathbf{BA})\,. \tag{18.23}$$

Recall also that the nonnegative definite square root of the nonnegative definite matrix $\mathbf{A} = \mathbf{T}\boldsymbol{\Lambda}\mathbf{T}'$ can be defined as

$$\mathbf{A}^{1/2} = \mathbf{T}\boldsymbol{\Lambda}^{1/2}\mathbf{T}' = \mathbf{T}_1\boldsymbol{\Lambda}_1^{1/2}\mathbf{T}_1'\,. \tag{18.24}$$

The matrix $\mathbf{T}\boldsymbol{\Lambda}^{1/2}\mathbf{T}'$ is the only symmetric nonnegative definite matrix whose squared value equals $\mathbf{A}$; for the proof, see Bapat (2000, p. 21) and Harville (1997, §21.9). Moreover,

$$(\mathbf{A}^{+})^{1/2} = \mathbf{T}(\boldsymbol{\Lambda}^{+})^{1/2}\mathbf{T}' = \mathbf{T}_1\boldsymbol{\Lambda}_1^{-1/2}\mathbf{T}_1'\,, \tag{18.25}$$

and hence

$$\mathbf{A}^{1/2}(\mathbf{A}^{+})^{1/2} = \mathbf{T}_1\mathbf{T}_1' = \mathbf{P}_{\mathbf{A}}\,. \tag{18.26}$$

It is also obvious that

$$\mathbf{A}^{k} = \mathbf{T}\boldsymbol{\Lambda}^{k}\mathbf{T}'\,, \quad \mathbf{A}^{0} = \mathbf{T}\boldsymbol{\Lambda}^{0}\mathbf{T}' = \mathbf{T}_1\mathbf{T}_1' = \mathbf{P}_{\mathbf{A}}\,. \tag{18.27}$$

As was mentioned earlier, there are far more useful results about the eigenvalues but we better proceed forward and ...

18.1 The Maximal Value of $\mathbf{x}'\mathbf{A}\mathbf{x}$

Let a symmetric matrix $\mathbf{A}_{n\times n}$ have the eigenvalue decomposition

$$\mathbf{A} = \mathbf{T}\boldsymbol{\Lambda}\mathbf{T}' = \mathbf{T}_1\boldsymbol{\Lambda}_1\mathbf{T}_1' = \lambda_1\mathbf{t}_1\mathbf{t}_1' + \cdots + \lambda_r\mathbf{t}_n\mathbf{t}_n' \,, \tag{18.28}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ are the ordered eigenvalues of $\mathbf{A}$. We show now that

$$\max_{\mathbf{x}'\mathbf{x}=1} \mathbf{x}'\mathbf{A}\mathbf{x} = \lambda_1 \,, \tag{18.29}$$

and that the maximum is attained when $\mathbf{x} = \mathbf{t}_1$. Our claim comes at once from the following:

$$\begin{aligned} \mathbf{x}'\mathbf{A}\mathbf{x} &= \mathbf{x}'\mathbf{T}\boldsymbol{\Lambda}\mathbf{T}'\mathbf{x} = \mathbf{y}'\boldsymbol{\Lambda}\mathbf{y} \\ &= \lambda_1 y_1^2 + \lambda_2 y_2^2 + \cdots + \lambda_n y_n^2 \\ &\leq \lambda_1(y_1^2 + y_2^2 + \cdots + y_n^2) = \lambda_1\mathbf{y}'\mathbf{y} = \lambda_1 \,, \end{aligned} \tag{18.30}$$

where we have denoted $\mathbf{y} = \mathbf{T}\mathbf{x}$ and thereby $\mathbf{y}'\mathbf{y} = 1$. The maximum is clearly attained when $\mathbf{x}$ is chosen as $\mathbf{t}_1$. In the same way we can prove that

$$\min_{\mathbf{x}'\mathbf{x}=1} \mathbf{x}'\mathbf{A}\mathbf{x} = \lambda_n \,. \tag{18.31}$$

Note that we can of course write the following:

$$\lambda_1 = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}'\mathbf{A}\mathbf{x}}{\mathbf{x}'\mathbf{x}} \,, \qquad \lambda_n = \min_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}'\mathbf{A}\mathbf{x}}{\mathbf{x}'\mathbf{x}} \,. \tag{18.32}$$

The expression $\mathbf{x}'\mathbf{A}\mathbf{x}/\mathbf{x}'\mathbf{x}$ is known as a Rayleigh ratio (quotient) or Rayleigh–Ritz ratio according to two physicists, third Baron Rayleigh (John William Strutt) and Walter Ritz.

Let us denote

$$\mathbf{T}_{(k)} = (\mathbf{t}_1 : \ldots : \mathbf{t}_k)\text{:} \quad \text{the first } k \text{ columns of } \mathbf{T} \,, \tag{18.33a}$$

$$\mathbf{T}_{[k]} = (\mathbf{t}_{n-k+1} : \ldots : \mathbf{t}_n)\text{:} \quad \text{the last } k \text{ columns of } \mathbf{T} \,. \tag{18.33b}$$

The second largest eigenvalue λ_2 can be obtained as follows:

$$\operatorname{ch}_2(\mathbf{A}) = \lambda_2 = \max_{\mathbf{x}\in\mathcal{L}} \mathbf{x}'\mathbf{A}\mathbf{x} \,, \tag{18.34}$$

where the maximum is taken over the set

$$\begin{aligned} \mathcal{L} &= \{\, \mathbf{x} \in \mathbb{R}^n : \mathbf{x}'\mathbf{x} = 1,\ \mathbf{x}'\mathbf{t}_1 = 0 \,\} \\ &= \{\, \mathbf{x} \in \mathbb{R}^n : \mathbf{x}'\mathbf{x} = 1,\ \mathbf{x} \in \mathscr{C}(\mathbf{t}_2 : \ldots : \mathbf{t}_n) \,\} \\ &= \{\, \mathbf{x} \in \mathbb{R}^n : \mathbf{x}'\mathbf{x} = 1,\ \mathbf{x} \in \mathscr{C}(\mathbf{T}_{[n-1]}) \,\} \,. \end{aligned} \tag{18.35}$$

Above we have used the obvious fact that

$$\mathbb{R}^n = \mathscr{C}(\mathbf{t}_1) \boxplus \mathscr{C}(\mathbf{T}_{[n-1]}) \,, \tag{18.36}$$

and thereby

$$\mathbf{x}'\mathbf{t}_1 = 0 \iff \mathbf{x} \in \mathscr{C}(\mathbf{T}_{[n-1]})\,. \tag{18.37}$$

Obviously λ_2 can be characterized also as

$$\mathrm{ch}_2(\mathbf{A}) = \lambda_2 = \min_{\mathbf{x}\neq\mathbf{0},\, \mathbf{x}\in\mathscr{C}(\mathbf{T}_{(2)})} \mathbf{x}'\mathbf{A}\mathbf{x} = \min_{\mathbf{x}\neq\mathbf{0},\, \mathbf{T}'_{[n-2]}\mathbf{x}=\mathbf{0}} \mathbf{x}'\mathbf{A}\mathbf{x}\,. \tag{18.38}$$

The ith largest eigenvalue can be defined in the corresponding way. We may present the results as a proposition.

Proposition 18.1. *Let $\mathbf{A}_{n\times n}$ be symmetric. Then*

(a) $\mathrm{ch}_1(\mathbf{A}) = \lambda_1 = \max_{\mathbf{x}\neq\mathbf{0}} \dfrac{\mathbf{x}'\mathbf{A}\mathbf{x}}{\mathbf{x}'\mathbf{x}} = \max_{\mathbf{x}'\mathbf{x}=1} \mathbf{x}'\mathbf{A}\mathbf{x},$

(b) $\mathrm{ch}_n(\mathbf{A}) = \lambda_n = \min_{\mathbf{x}\neq\mathbf{0}} \dfrac{\mathbf{x}'\mathbf{A}\mathbf{x}}{\mathbf{x}'\mathbf{x}} = \min_{\mathbf{x}'\mathbf{x}=1} \mathbf{x}'\mathbf{A}\mathbf{x},$

(c) $\mathrm{ch}_n(\mathbf{A}) = \lambda_n \leq \dfrac{\mathbf{x}'\mathbf{A}\mathbf{x}}{\mathbf{x}'\mathbf{x}} \leq \lambda_1 = \mathrm{ch}_1(\mathbf{A}),$

(d) $\mathrm{ch}_2(\mathbf{A}) = \lambda_2 = \max_{\mathbf{x}'\mathbf{x}=1,\, \mathbf{t}_1'\mathbf{x}=0} \mathbf{x}'\mathbf{A}\mathbf{x},$

(e) $\mathrm{ch}_{k+1}(\mathbf{A}) = \lambda_{k+1} = \max_{\mathbf{x}'\mathbf{x}=1,\, \mathbf{T}'_{(k)}\mathbf{x}=\mathbf{0}} \mathbf{x}'\mathbf{A}\mathbf{x} = \lambda_{k+1},\ k = 1,\dots,n-1.$

18.2 Principal Components

Let a p-dimensional random vector $\mathbf{x}$ have $\mathrm{E}(\mathbf{x}) = \boldsymbol{\mu}$ and $\mathrm{cov}(\mathbf{x}) = \boldsymbol{\Sigma}$ and let the eigenvalue decomposition of $\boldsymbol{\Sigma}$ be

$$\boldsymbol{\Sigma} = \mathbf{T}\boldsymbol{\Lambda}\mathbf{T}', \quad \boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1,\dots,\lambda_p)\,, \quad \mathbf{T}'\mathbf{T} = \mathbf{I}_p\,, \tag{18.39}$$

where $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ are the ordered eigenvalues of $\boldsymbol{\Sigma}$. We might be interested in knowing which linear combination $\mathbf{b}'\mathbf{x}$ has the largest variance under the condition $\mathbf{b}'\mathbf{b} = 1$. Now, the Proposition 18.1 gives the answer at once: the random variable $w_i = \mathbf{t}_1'\mathbf{x}$, which is the first element of the random vector $\mathbf{w} = \mathbf{T}'\mathbf{x}$:

$$\mathbf{w} = \mathbf{T}'\mathbf{x} = \begin{pmatrix} \mathbf{t}_1'\mathbf{x} \\ \vdots \\ \mathbf{t}_p'\mathbf{x} \end{pmatrix}. \tag{18.40}$$

The centered random variable

$$y_1 = w_1 - \mathrm{E}(w_1) = \mathbf{t}_1'(\mathbf{x} - \boldsymbol{\mu}) \tag{18.41}$$

is the first (population) principal component of $\mathbf{x}$. Thus we have

$$\max_{\mathbf{b}'\mathbf{b}=1} \mathrm{var}(\mathbf{b}'\mathbf{x}) = \mathrm{var}(\mathbf{t}_1'\mathbf{x}) = \mathrm{var}[\mathbf{t}_1'(\mathbf{x} - \boldsymbol{\mu})] = \mathrm{var}(y_1) = \lambda_1\,. \tag{18.42}$$

The second principal component of $\mathbf{x}$ is the random variable $\mathbf{b}'(\mathbf{x}-\boldsymbol{\mu})$ where $\mathbf{b}$ is a solution to the task

$$\max \operatorname{var}(\mathbf{b}'\mathbf{x}) \quad \text{subject to } \operatorname{cov}(y_1, \mathbf{b}'\mathbf{x}) = 0,\ \mathbf{b}'\mathbf{b} = 1\,. \tag{18.43}$$

Because (assuming $\lambda_1 > 0$)

$$\operatorname{cov}(y_1, \mathbf{b}'\mathbf{x}) = \mathbf{t}_1'\boldsymbol{\Sigma}\mathbf{b} = \lambda_1\mathbf{t}_1'\mathbf{b} = 0 \iff \mathbf{t}_1'\mathbf{b} = 0\,, \tag{18.44}$$

the task (18.43) becomes

$$\max \operatorname{var}(\mathbf{b}'\mathbf{x}) \quad \text{subject to } \mathbf{t}_1'\mathbf{b} = 0,\ \mathbf{b}'\mathbf{b} = 1\,, \tag{18.45}$$

for which the solution is $\mathbf{b} = \mathbf{t}_2$ and so the random variable $y_2 = \mathbf{t}_2'(\mathbf{x}-\boldsymbol{\mu})$ is the second principal component of $\mathbf{x}$. In general,

$$\max_{\mathbf{b}'\mathbf{b}=1,\mathbf{T}_{(i-1)}'\mathbf{b}=\mathbf{0}} \operatorname{var}(\mathbf{b}'\mathbf{x}) = \operatorname{var}(\mathbf{t}_i'\mathbf{x}) = \operatorname{var}(y_i) = \lambda_i\,, \tag{18.46}$$

i.e., $\mathbf{t}_i'\mathbf{x}$ has maximum variance of all normalized linear combinations uncorrelated with the elements of $\mathbf{T}_{(i-1)}'\mathbf{x}$.

All in all, the principal components y_i are centered and uncorrelated random variables such that

$$\mathbf{y} = \mathbf{T}'(\mathbf{x}-\boldsymbol{\mu}) = \mathbf{w} - \boldsymbol{\mu}_{\mathbf{w}} = \begin{pmatrix} w_1 - \mu_{w_1} \\ \vdots \\ w_p - \mu_{w_p} \end{pmatrix}, \tag{18.47a}$$

$$\mathrm{E}(\mathbf{y}) = \mathbf{0}\,, \quad \operatorname{cov}(\mathbf{y}) = \boldsymbol{\Lambda}\,, \tag{18.47b}$$

$$\operatorname{tr}\operatorname{cov}(\mathbf{y}) = \operatorname{tr}(\boldsymbol{\Sigma}) = \lambda_1 + \lambda_2 + \dots + \lambda_p\,. \tag{18.47c}$$

We can scale the random vector $\mathbf{y}$ so that ($\boldsymbol{\Sigma}$ being positive definite)

$$\mathbf{z} = \boldsymbol{\Lambda}^{-1/2}\mathbf{y} = \boldsymbol{\Lambda}^{-1/2}\mathbf{T}'(\mathbf{x}-\boldsymbol{\mu})\,, \tag{18.48}$$

and thereby

$$\mathrm{E}(\mathbf{z}) = \mathbf{0}\,, \quad \operatorname{cov}(\mathbf{z}) = \mathbf{I}_p\,. \tag{18.49}$$

. Now

$$\begin{aligned}
\mathbf{z}'\mathbf{z} &= \mathbf{y}'\boldsymbol{\Lambda}^{-1}\mathbf{y} = (\mathbf{x}-\boldsymbol{\mu})'\mathbf{T}\boldsymbol{\Lambda}^{-1}\mathbf{T}'(\mathbf{x}-\boldsymbol{\mu}) = (\mathbf{w}-\boldsymbol{\mu}_{\mathbf{w}})'\boldsymbol{\Lambda}^{-1}(\mathbf{w}-\boldsymbol{\mu}_{\mathbf{w}}) \\
&= \frac{(w_1-\mu_{w_1})^2}{\lambda_1} + \frac{(w_2-\mu_{w_2})^2}{\lambda_2} + \dots + \frac{(w_p-\mu_{w_p})^2}{\lambda_p} \\
&= \frac{y_1^2}{\lambda_1} + \frac{y_2^2}{\lambda_2} + \dots + \frac{y_p^2}{\lambda_p} = z_1^2 + z_2^2 + \dots + z_p^2 \\
&= (\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \operatorname{MHLN}^2(\mathbf{x},\, \boldsymbol{\mu},\, \boldsymbol{\Sigma})\,.
\end{aligned} \tag{18.50}$$

From above we see that the squared Mahalanobis distance $\mathrm{MHLN}^2(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the sum of squares of the centered and normed (variance 1) uncorrelated random variables

$$\frac{w_i - \mathrm{E}(w_i)}{\sqrt{\mathrm{var}(w_i)}} = \frac{y_i}{\sqrt{\mathrm{var}(y_i)}} = z_i \,. \tag{18.51}$$

In particular, if $\mathbf{x} \sim \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\mathbf{z} \sim \mathrm{N}_p(\mathbf{0}, \mathbf{I}_p)$ and

$$(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = \mathbf{z}'\mathbf{z} \sim \chi^2(p) \,. \tag{18.52}$$

If $\mathrm{rank}(\boldsymbol{\Sigma}) = r < p$, then we can define

$$\mathbf{z} = (\boldsymbol{\Lambda}^+)^{1/2}\mathbf{y} \,, \quad \boldsymbol{\Lambda}^+ = \begin{pmatrix} \boldsymbol{\Lambda}_1^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \quad \boldsymbol{\Lambda}_1 = \mathrm{diag}(\lambda_1, \dots, \lambda_r) \,, \tag{18.53}$$

and so

$$\begin{aligned} \mathbf{z}'\mathbf{z} &= \mathbf{y}'(\boldsymbol{\Lambda}^+)^{1/2}\mathbf{y} \\ &= \frac{(w_1 - \mu_{w_1})^2}{\lambda_1} + \frac{(w_2 - \mu_{w_2})^2}{\lambda_2} + \dots + \frac{(w_r - \mu_{w_r})^2}{\lambda_r} \\ &= \frac{y_1^2}{\lambda_1} + \frac{y_2^2}{\lambda_2} + \dots + \frac{y_r^2}{\lambda_r} = z_1^2 + z_2^2 + \dots + z_r^2 \\ &= (\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^+(\mathbf{x} - \boldsymbol{\mu}) = \mathrm{MHLN}^2(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \,. \end{aligned} \tag{18.54}$$

Notice that actually (18.54) is invariant with respect to the choice of $\boldsymbol{\Sigma}^-$; see page 65.

We will return to the principal components in connection with the singular value decomposition, see Section 19.4 (p. 402). For a predictive approach to principal component analysis, see Section 9.4 (p. 203).

18.3 Eigenvalues of $\boldsymbol{\Sigma}_{2\times 2}$

Let's look at a simple example where we have a covariance matrix

$$\mathrm{cov}(\mathbf{z}) = \mathrm{cov}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{yx} & \sigma_y^2 \end{pmatrix} = \boldsymbol{\Sigma} \,. \tag{18.55}$$

Then the characteristic polynomial for $\boldsymbol{\Sigma}$ is

$$p_{\boldsymbol{\Sigma}}(\lambda) = \lambda^2 - (\sigma_x^2 + \sigma_y^2)\lambda + (\sigma_x^2\sigma_y^2 - \sigma_{xy}^2) = \lambda^2 - \mathrm{tr}(\boldsymbol{\Sigma})\lambda + \det(\boldsymbol{\Sigma}) \,, \tag{18.56}$$

and the eigenvalues are

$$\lambda_1, \lambda_2 = \frac{1}{2}\left[\sigma_x^2 + \sigma_y^2 \pm \sqrt{(\sigma_x^2 - \sigma_y^2)^2 + 4\sigma_{xy}^2}\,\right]. \tag{18.57}$$

In light of (18.5) (p. 358), we of course have

$$\operatorname{tr}(\mathbf{\Sigma}) = \sigma_x^2 + \sigma_y^2 = \lambda_1 + \lambda_2\,, \tag{18.58a}$$

$$\det(\mathbf{\Sigma}) = \sigma_x^2\sigma_y^2\left(1 - \frac{\sigma_{xy}^2}{\sigma_x^2\sigma_y^2}\right) = \sigma_x^2\sigma_y^2(1 - \varrho_{xy}^2) = \lambda_1\lambda_2\,. \tag{18.58b}$$

The eigenvector $\mathbf{t}_1 = \binom{t_{11}}{t_{21}}$ with respect to λ_1 is a nonzero solution to the equation $(\mathbf{\Sigma} - \lambda_1\mathbf{I}_2)\mathbf{t}_1 = \mathbf{0}$, i.e.,

$$(\sigma_x^2 - \lambda_1)t_{11} + \sigma_{xy}t_{21} = 0\,, \tag{18.59a}$$

$$\sigma_{xy}t_{11} + (\sigma_y^2 - \lambda_1)t_{21} = 0\,. \tag{18.59b}$$

Suppose that $\sigma_x^2 > \sigma_y^2$. Then it is easy to conclude that t_{21} can be put 0 if and only if $\sigma_{xy} = 0$ (please confirm!), in which case t_{11} can be any nonzero real number. Notice that it is the ratio t_{21}/t_{11} which specifies the eigenvector; see also Exercise 18.14 (p. 387).

In particular, if $\sigma_x^2 = \sigma_y^2 = \sigma^2$ and $\sigma_{xy} = \sigma^2\varrho_{xy} > 0$, we have

$$\lambda_1 = \sigma^2 + \sigma_{xy} = \sigma^2(1 + \varrho_{xy})\,, \quad \lambda_2 = \sigma^2 - \sigma_{xy} = \sigma^2(1 - \varrho_{xy})\,, \tag{18.60}$$

while the standardized eigenvectors are

$$\mathbf{t}_1 = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\1\end{pmatrix} = \begin{pmatrix}\cos\theta\\ \sin\theta\end{pmatrix}, \quad \mathbf{t}_2 = \frac{1}{\sqrt{2}}\begin{pmatrix}-1\\1\end{pmatrix} = \begin{pmatrix}-\sin\theta\\ \cos\theta\end{pmatrix}, \tag{18.61}$$

where $\theta = 45°$. If $\sigma_{xy} < 0$, then $\lambda_1 = \sigma^2 - \sigma_{xy}$ and $\mathbf{t}_1 = \frac{1}{\sqrt{2}}(-1, 1)'$. If $\sigma_{xy} = 0$, then $\lambda_1 = \lambda_2 = \sigma^2$ and $\mathbf{t}_1$ can be chosen as any nonzero vector in $\mathbb{R}^2$.

Let $\mathbf{z} \sim \mathrm{N}_2(\boldsymbol{\mu}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is positive definite. Then $\mathbf{z}$ has the density

$$n(\mathbf{z}; \boldsymbol{\mu}, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{1/2}|\mathbf{\Sigma}|^{1/2}}\mathrm{e}^{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu})}. \tag{18.62}$$

It is obvious that the contours of constant density for $\mathrm{N}_2(\boldsymbol{\mu}, \mathbf{\Sigma})$ are ellipses defined by $\mathbf{z}$ such that

$$\mathcal{A} = \{\, \mathbf{z} \in \mathbb{R}^2 : (\mathbf{z} - \boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{z} - \boldsymbol{\mu}) = c^2 \,\}\,. \tag{18.63}$$

How are the axes of this ellipse located? The major axis is the longest diameter (line through $\boldsymbol{\mu}$) of the ellipse, that is, we want to find a point $\mathbf{z}_1$ of the ellipse which has the maximal Euclidean distance from the mean point $\boldsymbol{\mu}$:

$$\max\|\mathbf{z} - \boldsymbol{\mu}\|^2 \quad \text{subject to } \mathbf{z} \in \mathcal{A}\,. \tag{18.64}$$

Denoting $\mathbf{u} = \mathbf{z} - \boldsymbol{\mu}$, the above task becomes

$$\max \mathbf{u}'\mathbf{u} \quad \text{subject to } \mathbf{u}'\boldsymbol{\Sigma}^{-1}\mathbf{u} = c^2. \tag{18.65}$$

It is not complicated to confirm that the solution to (18.65) is

$$\mathbf{u}_1 = \mathbf{z}_1 - \boldsymbol{\mu} = \pm c\sqrt{\lambda_1}\mathbf{t}_1\,, \quad \text{and } \mathbf{u}_1'\mathbf{u}_1 = c^2\lambda_1. \tag{18.66}$$

Correspondingly, the minor axis is the shortest diameter of the ellipse $\mathcal{A}$.

One way to prove that (18.66) is the solution to (18.65), is to use the method of Lagrangian multipliers, by determining the stationary points by considering

$$\psi = \mathbf{u}'\mathbf{u} - \lambda\mathbf{u}'\boldsymbol{\Sigma}^{-1}\mathbf{u}\,, \tag{18.67}$$

where λ is a Lagrangian multiplier. Differentiating ψ with respect to the components of $\mathbf{u}$ yields the result; see Anderson (2003, p. 466). The other way is to note that (18.65) can be expressed as

$$\max \mathbf{v}'\boldsymbol{\Sigma}\mathbf{v} \quad \text{subject to } \mathbf{v}'\mathbf{v} = c^2, \tag{18.68}$$

where $\mathbf{v} = \boldsymbol{\Sigma}^{-1/2}\mathbf{u}$ and hence $\mathbf{u} = \boldsymbol{\Sigma}^{1/2}\mathbf{v}$. The details of the proof are left as an exercise.

The ellipse $\mathcal{A}$ is centered at $\boldsymbol{\mu}$ and the axes are determined by $\mathbf{u}_i = \pm c\sqrt{\lambda_i}\mathbf{t}_i$, where $\lambda_i = \operatorname{ch}_i(\boldsymbol{\Sigma})$ and $\mathbf{t}_i$ is the corresponding eigenvector. If $\sigma_x^2 = \sigma_y^2 = \sigma^2$ and $\sigma_{xy} = \sigma^2\varrho_{xy} > 0$, then the first eigenvector $\mathbf{t}_1$ lies along the 45° line through the point $\boldsymbol{\mu} = (\mu_x, \mu_y)'$. The lengths (or actually half-lengths) of the axes are then

$$c\sqrt{\sigma^2 + \sigma_{xy}} = c\sigma\sqrt{1 + \varrho_{xy}} \quad \text{and} \quad c\sqrt{\sigma^2 - \sigma_{xy}} = c\sigma\sqrt{1 - \varrho_{xy}}\,. \tag{18.69}$$

18.4 Eigenvalues of the Intraclass Correlation Matrix

If $\mathbf{A}_{p\times p} = (a-b)\mathbf{I}_p + b\mathbf{1}_p\mathbf{1}_p'$ for some $a, b \in \mathbb{R}$, then the matrix $\mathbf{A}$ is called a completely symmetric matrix. If it is a covariance matrix then it is said to have an intraclass correlation structure. The following proposition collects together some useful properties of completely symmetric matrices. The proof is left to the energetic readers; see also Exercise 18.17 (p. 388).

Proposition 18.2. *Consider the matrices*

$$\mathbf{A}_{p\times p} = (a-b)\mathbf{I}_p + b\mathbf{1}_p\mathbf{1}_p'\,, \quad \boldsymbol{\Sigma}_{p\times p} = (1-\varrho)\mathbf{I}_p + \varrho\mathbf{1}_p\mathbf{1}_p'\,. \tag{18.70}$$

Then

(a) $\det(\mathbf{A}) = (a-b)^{n-1}[a + (p-1)b]$,

(b) *the eigenvalues of* $\mathbf{A}$ *are* $a + (p-1)b$ (*with multiplicity* 1), *and* $a - b$ (*with multiplicity* $n-1$),

(c) $\mathbf{A}$ *is nonsingular if and only if* $a \neq b$ *and* $a \neq -(p-1)b$*, in which case* $\mathbf{A}^{-1} = \frac{1}{a-b}\Big(\mathbf{I}_p - \frac{b}{a+(p-1)b}\mathbf{1}_p\mathbf{1}_p'\Big)$,

(d) $\operatorname{ch}(\mathbf{\Sigma}) = \begin{cases} 1+(p-1)\varrho & \textit{with multiplicity } 1, \\ 1-\varrho & \textit{with multiplicity } p-1, \end{cases}$

(e) $\mathbf{\Sigma}$ *is nonnegative definite* $\iff -\frac{1}{p-1} \leq \varrho \leq 1$,

(f) $\mathbf{t}_1 = \alpha\mathbf{1}_p =$ *eigenvector w.r.t.* $\lambda_1 = 1+(p-1)\varrho$, $0 \neq \alpha \in \mathbb{R}$,
$\mathbf{t}_2, \dots, \mathbf{t}_p$ *are orthonormal eigenvectors w.r.t.* $\lambda_i = 1-\varrho$, $i = 2, \dots, p$,
$\mathbf{t}_2, \dots, \mathbf{t}_p$ *form an orthonormal basis for* $\mathscr{C}(\mathbf{1}_p)^\perp$,

(g) $\det(\mathbf{\Sigma}) = (1-\varrho)^{p-1}[1+(p-1)\varrho]$,

(h) $\mathbf{\Sigma}\mathbf{1}_p = [1+(p-1)\varrho]\mathbf{1}_p := \lambda_1\mathbf{1}_p$; *if* $\varrho \neq -\frac{1}{p-1}$, *then* $\mathbf{1}_p = \lambda_1^{-1}\mathbf{\Sigma}\mathbf{1}_p$ *in which case*

$$\mathbf{1}_p'\mathbf{\Sigma}^-\mathbf{1}_p = \lambda_1^{-2}\mathbf{1}_p'\mathbf{\Sigma}\mathbf{\Sigma}^-\mathbf{\Sigma}\mathbf{1}_p = \lambda_1^{-2}\mathbf{1}_p'\mathbf{\Sigma}\mathbf{1}_p = \frac{p}{1+(p-1)\varrho}\,, \tag{18.71}$$

(i) $\mathbf{\Sigma}^{-1} = \frac{1}{1-\varrho}\Big(\mathbf{I}_p - \frac{\varrho}{1+(p-1)\varrho}\mathbf{1}_p\mathbf{1}_p'\Big)$, *for* $\varrho \neq 1$, $\varrho \neq -\frac{1}{p-1}$.

(j) *Let* $\mathbf{z} = \binom{\mathbf{x}}{y}$ *be a* $(p+1)$*-dimensional random vector with intraclass covariance structure*

$$\operatorname{cov}\begin{pmatrix}\mathbf{x}\\ y\end{pmatrix} = \sigma^2\begin{pmatrix}\mathbf{\Sigma}_{\mathbf{xx}} & \boldsymbol{\sigma}_{\mathbf{x}y}\\ \boldsymbol{\sigma}_{\mathbf{x}y}' & 1\end{pmatrix} = \sigma^2[(1-\varrho)\mathbf{I}_{p+1} + \varrho\mathbf{1}_{p+1}\mathbf{1}_{p+1}']\,, \tag{18.72}$$

where (necessarily) $-\frac{1}{p} \leq \varrho \leq 1$. *Then*

$$\varrho_{y\cdot\mathbf{x}}^2 = \boldsymbol{\sigma}_{\mathbf{x}y}'\mathbf{\Sigma}_{\mathbf{xx}}^-\boldsymbol{\sigma}_{\mathbf{x}y} = \frac{p\varrho^2}{1+(p-1)\varrho}\,. \tag{18.73}$$

(k) *Let* $\mathbf{x}$ *and* $\mathbf{y}$ *be* p*-dimensional random vectors with intraclass covariance structure*

$$\operatorname{cov}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix} = \begin{pmatrix}\mathbf{\Sigma} & \varrho\mathbf{1}_p\mathbf{1}_p'\\ \varrho\mathbf{1}_p\mathbf{1}_p' & \mathbf{\Sigma}\end{pmatrix}, \tag{18.74}$$

where $\mathbf{\Sigma}_{p\times p} = (1-\varrho)\mathbf{I}_p + \varrho\mathbf{1}_p\mathbf{1}_p'$. *Then*

$$\operatorname{cor}(\mathbf{1}_p'\mathbf{x}, \mathbf{1}_p'\mathbf{y}) = \frac{p\varrho}{1+(p-1)\varrho}\,. \tag{18.75}$$

18.5 Proper Eigenvalues; B Positive Definite

In this section, as well as in Section 18.7, we consider the *proper eigenvalues* and *proper eigenvectors* following Rao & Mitra (1971b, §6.3); see also Mitra &

Rao (1968), as well as de Leeuw (1982), McDonald, Torii & Nishisato (1979) Mitra & Moore (1973, Appendix), Scott & Styan (1985), SenGupta (1991), and Isotalo, Puntanen & Styan (2008a, §2).

Let $\mathbf{A}$ and $\mathbf{B}$ be two symmetric $n \times n$ matrices of which $\mathbf{B}$ is nonnegative definite. Let λ be a scalar and $\mathbf{w}$ a vector such that

$$\mathbf{A}\mathbf{w} = \lambda\mathbf{B}\mathbf{w}, \quad \mathbf{B}\mathbf{w} \neq \mathbf{0}\,. \tag{18.76}$$

Then we call λ a proper eigenvalue and $\mathbf{w}$ a proper eigenvector of $\mathbf{A}$ with respect to $\mathbf{B}$, or shortly, λ is a proper eigenvalue of the pair $(\mathbf{A}, \mathbf{B})$.

Before allowing $\mathbf{B}$ to be singular (but still nonnegative definite), let us consider the situation when $\mathbf{B}$ is assumed to be positive definite. In such a case, (18.76) becomes

$$\mathbf{A}\mathbf{w} = \lambda\mathbf{B}\mathbf{w}\,, \quad \mathbf{w} \neq \mathbf{0}\,. \tag{18.77}$$

Premultiplying (18.77) by $\mathbf{B}^{-1}$ yields the usual eigenvalue equation

$$\mathbf{B}^{-1}\mathbf{A}\mathbf{w} = \lambda\mathbf{w}\,, \quad \mathbf{w} \neq \mathbf{0}\,. \tag{18.78}$$

Now the matrix $\mathbf{B}^{-1}\mathbf{A}$ is not necessarily symmetric but that is no problem because

$$\operatorname{nzch}(\mathbf{B}^{-1}\mathbf{A}) = \operatorname{nzch}(\mathbf{B}^{-1/2}\mathbf{B}^{-1/2}\mathbf{A}) = \operatorname{nzch}(\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2})\,, \tag{18.79}$$

where the matrix $\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}$ is symmetric and thereby the eigenvalues of $\mathbf{B}^{-1}\mathbf{A}$ are all real. Notice that premultiplying (18.77) by $\mathbf{B}^{-1/2}$ yields

$$\mathbf{B}^{-1/2}\mathbf{A}\mathbf{w} = \lambda\mathbf{B}^{1/2}\mathbf{w}\,, \tag{18.80}$$

i.e., $\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2} \cdot \mathbf{B}^{1/2}\mathbf{w} = \lambda\mathbf{B}^{1/2}\mathbf{w}$, which shows that

$$(\lambda, \mathbf{w}) \text{ is an eigenpair for } (\mathbf{A}, \mathbf{B}) \tag{18.81a}$$

$$\Longleftrightarrow$$

$$(\lambda, \mathbf{B}^{1/2}\mathbf{w}) \text{ is an eigenpair for } \mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2} \tag{18.81b}$$

$$\Longleftrightarrow$$

$$(\lambda, \mathbf{w}) \text{ is an eigenpair for } \mathbf{B}^{-1}\mathbf{A}\,. \tag{18.81c}$$

Rewriting (18.77) as

$$(\mathbf{A} - \lambda\mathbf{B})\mathbf{w} = \mathbf{0}\,, \tag{18.82}$$

we observe that nontrivial solutions $\mathbf{w}$ for (18.77) exist if and only if

$$\det(\mathbf{A} - \lambda\mathbf{B}) = 0\,. \tag{18.83}$$

The expression $\mathbf{A} - \lambda\mathbf{B}$, with indeterminate λ, is called a matrix pencil or simply a pencil.

It is left as an exercise to confirm that if $\lambda \neq \mu$ then

$$(\lambda, \mathbf{w}) \text{ and } (\mu, \mathbf{v}) \text{ are eigenpairs for } (\mathbf{A}, \mathbf{B}) \implies \mathbf{w}'\mathbf{B}\mathbf{v} = 0\,. \tag{18.84}$$

It is straightforward to obtain the following very important result:

$$\begin{aligned}\max_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}'\mathbf{A}\mathbf{x}}{\mathbf{x}'\mathbf{B}\mathbf{x}} &= \max_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}'\mathbf{B}^{1/2}\cdot\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}\cdot\mathbf{B}^{1/2}\mathbf{x}}{\mathbf{x}'\mathbf{B}^{1/2}\cdot\mathbf{B}^{1/2}\mathbf{x}} \\ &= \max_{\mathbf{z}\in\mathscr{C}(\mathbf{B}^{1/2}),\,\mathbf{z}\neq\mathbf{0}} \frac{\mathbf{z}'\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}\mathbf{z}}{\mathbf{z}'\mathbf{z}} \\ &= \max_{\mathbf{z}\neq\mathbf{0}} \frac{\mathbf{z}'\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}\mathbf{z}}{\mathbf{z}'\mathbf{z}} \\ &= \operatorname{ch}_1(\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}) = \operatorname{ch}_1(\mathbf{B}^{-1}\mathbf{A})\,. \end{aligned} \tag{18.85}$$

We cannot resist the temptation to write up some useful (in particular for a statistician) properties of the eigenvalues of the pair $(\mathbf{A}, \mathbf{B})$ when $\mathbf{B}$ is positive definite.

Proposition 18.3. *Let* $\mathbf{A}_{n\times n}$ *be symmetric and let* $\mathbf{B}_{n\times n}$ *be positive definite. Then*

(a) $\max_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}'\mathbf{A}\mathbf{x}}{\mathbf{x}'\mathbf{B}\mathbf{x}} = \operatorname{ch}_1(\mathbf{A}\mathbf{B}^{-1}) := \lambda_1$, *i.e.,* $\lambda_1 =$ *the largest root of*

$$\det(\mathbf{A} - \lambda\mathbf{B}) = 0\,. \tag{18.86}$$

(b) *The vectors* $\mathbf{w}_1, \dots, \mathbf{w}_n$ *satisfy*

$$\max_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}'\mathbf{A}\mathbf{x}}{\mathbf{x}'\mathbf{B}\mathbf{x}} = \frac{\mathbf{w}_1'\mathbf{A}\mathbf{w}_1}{\mathbf{w}_1'\mathbf{B}\mathbf{w}_1} = \operatorname{ch}_1(\mathbf{A}\mathbf{B}^{-1})\,, \tag{18.87a}$$

$$\max_{\substack{\mathbf{x}'\mathbf{A}\mathbf{w}_j=0,\,\mathbf{x}\neq\mathbf{0}\\ j=1,\dots,i-1}} \frac{\mathbf{x}'\mathbf{A}\mathbf{x}}{\mathbf{x}'\mathbf{B}\mathbf{x}} = \frac{\mathbf{w}_i'\mathbf{A}\mathbf{w}_i}{\mathbf{w}_i'\mathbf{B}\mathbf{w}_i} = \operatorname{ch}_i(\mathbf{A}\mathbf{B}^{-1})\,, \quad i > 1\,, \tag{18.87b}$$

if and only if $\mathbf{w}_i$ *is an eigenvector of* $\mathbf{A}\mathbf{B}^{-1}$ *corresponding to the eigenvalue* $\operatorname{ch}_i(\mathbf{A}\mathbf{B}^{-1}) = \lambda_i$, *i.e.,* λ_i *is the* i*th largest root of* (18.86).

(c) $\max_{\mathbf{x}\neq\mathbf{0}} \frac{(\mathbf{a}'\mathbf{x})^2}{\mathbf{x}'\mathbf{B}\mathbf{x}} = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}'\cdot\mathbf{a}\mathbf{a}'\cdot\mathbf{x}}{\mathbf{x}'\mathbf{B}\mathbf{x}} = \operatorname{ch}_1(\mathbf{a}\mathbf{a}'\mathbf{B}^{-1}) = \mathbf{a}'\mathbf{B}^{-1}\mathbf{a}$.

Notice that in part (c) of Proposition 18.3, we utilize the handy fact

$$\operatorname{nzch}(\mathbf{a}\mathbf{a}'\mathbf{B}^{-1}) = \operatorname{nzch}(\mathbf{a}'\mathbf{B}^{-1}\mathbf{a}) = \mathbf{a}'\mathbf{B}^{-1}\mathbf{a}\,. \tag{18.88}$$

It is important to observe that (c) can be obtained also using the following version of the Cauchy–Schwarz inequality, see Section 20.1 (p. 416):

$$(\mathbf{a}'\mathbf{x})^2 \leq \mathbf{x}'\mathbf{B}\mathbf{x}\cdot\mathbf{a}'\mathbf{B}^{-1}\mathbf{a} \quad \text{for } \mathbf{B} >_{\mathsf{L}} \mathbf{0}, \text{ for all } \mathbf{a}, \mathbf{x} \in \mathbb{R}^n. \tag{18.89}$$

In light of the above inequality we get, for $\mathbf{x} \neq \mathbf{0}$,

$$\frac{(\mathbf{a}'\mathbf{x})^2}{\mathbf{x}'\mathbf{B}\mathbf{x}} \leq \frac{\mathbf{x}'\mathbf{B}\mathbf{x} \cdot \mathbf{a}'\mathbf{B}^{-1}\mathbf{a}}{\mathbf{x}'\mathbf{B}\mathbf{x}} = \mathbf{a}'\mathbf{B}^{-1}\mathbf{a}\,. \tag{18.90}$$

The equality in (18.89) occurs if and only if

$$\mathbf{x} = \alpha\mathbf{B}^{-1}\mathbf{a} \quad \text{for some } \alpha \in \mathbb{R}\,. \tag{18.91}$$

What happens if the matrix $\mathbf{B}$ in part (d) of Proposition 18.3 is nonnegative definite and thus possibly singular? Again, according to Section 20.1 (p. 416), we have

$$(\mathbf{a}'\mathbf{x})^2 \leq \mathbf{x}'\mathbf{B}\mathbf{x} \cdot \mathbf{a}'\mathbf{B}^{+}\mathbf{a} \quad \text{for } \mathbf{B} \geq_{\mathsf{L}} \mathbf{0}, \text{ for all } \mathbf{a} \in \mathscr{C}(\mathbf{B})\ \mathbf{x} \in \mathbb{R}^n, \tag{18.92}$$

which yields, for $\mathbf{B}\mathbf{x} \neq \mathbf{0}$,

$$\frac{(\mathbf{a}'\mathbf{x})^2}{\mathbf{x}'\mathbf{B}\mathbf{x}} \leq \frac{\mathbf{x}'\mathbf{B}\mathbf{x} \cdot \mathbf{a}'\mathbf{B}^{+}\mathbf{a}}{\mathbf{x}'\mathbf{B}\mathbf{x}} = \mathbf{a}'\mathbf{B}^{+}\mathbf{a}\,. \tag{18.93}$$

Above we can replace $\mathbf{B}^{+}$ with any $\mathbf{B}^{-}$ because $\mathbf{a}'\mathbf{B}^{-}\mathbf{a}$ is invariant with respect to the choice of $\mathbf{B}^{-}$. To ease the referencing, we'll put this observation as a separate proposition.

Proposition 18.4. *Let* $\mathbf{B}_{n\times n}$ *be nonnegative definite and* $\mathbf{a} \in \mathscr{C}(\mathbf{B})$. *Then*

$$\max_{\mathbf{B}\mathbf{x}\neq\mathbf{0}} \frac{(\mathbf{a}'\mathbf{x})^2}{\mathbf{x}'\mathbf{B}\mathbf{x}} = \mathbf{a}'\mathbf{B}^{-}\mathbf{a}\,, \tag{18.94}$$

where the equality is obtained if and only if $\mathbf{B}\mathbf{x} = \alpha\mathbf{a}$ *for some* $\alpha \in \mathbb{R}$, *i.e.,*

$$\mathbf{x} = \alpha\mathbf{B}^{-}\mathbf{a} \quad \text{for some } \alpha \in \mathbb{R}\,. \tag{18.95}$$

The following result gives the crucial condition for simultaneous diagonalization by orthogonal matrix (see, e.g., Rao 1973a, p. 41, Searle 1982, p. 312).

Proposition 18.5. *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be symmetric* $n \times n$ *matrices. Then there exist an orthogonal* $\mathbf{Q}_{n\times n}$ *such that both* $\mathbf{Q}'\mathbf{A}\mathbf{Q}$ *and* $\mathbf{Q}'\mathbf{B}\mathbf{Q}$ *are diagonal if and only if* $\mathbf{A}\mathbf{B} = \mathbf{B}\mathbf{A}$.

Proposition 18.6. *Consider the symmetric matrices* $\mathbf{A}_{n\times n}$ *and* $\mathbf{B}_{n\times n}$.

(a) *If* $\mathbf{B}$ *positive definite, then there exists a nonsingular matrix* $\mathbf{Q}$ *such that*

$$\mathbf{Q}'\mathbf{A}\mathbf{Q} = \mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \dots, \lambda_n)\,, \quad \mathbf{Q}'\mathbf{B}\mathbf{Q} = \mathbf{I}_n\,, \tag{18.96}$$

where $\mathbf{\Lambda}$ *is the diagonal matrix of the solutions for* λ *to* $\det(\mathbf{A} - \lambda\mathbf{B}) = 0$, *i.e., the eigenvalues of* $\mathbf{B}^{-1}\mathbf{A}$. *The columns of* $\mathbf{Q}$ *are the eigenvectors of* $\mathbf{B}^{-1}\mathbf{A}$.

(b) *Let* $\mathbf{B}$ *be nonnegative definite with* $\operatorname{rank}(\mathbf{B}) = b$. *Then there exists a matrix* $\mathbf{L}_{n\times b}$ *such that*

$$\mathbf{L}'\mathbf{A}\mathbf{L} = \mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \dots, \lambda_b)\,, \quad \mathbf{L}'\mathbf{B}\mathbf{L} = \mathbf{I}_b\,. \tag{18.97}$$

(c) *Let* $\mathbf{B}$ *be nonnegative definite,* $\operatorname{rank}(\mathbf{B}) = b$, *and assume that*

$$\operatorname{rank}(\mathbf{N}'\mathbf{A}\mathbf{N}) = \operatorname{rank}(\mathbf{N}'\mathbf{A})\,, \text{ where } \mathbf{N} = \mathbf{B}^{\perp}. \tag{18.98}$$

Then there exists a nonsingular matrix $\mathbf{Q}_{n\times n}$ *such that*

$$\mathbf{Q}'\mathbf{A}\mathbf{Q} = \begin{pmatrix} \mathbf{\Lambda}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{\Lambda}_2 \end{pmatrix}, \quad \mathbf{Q}'\mathbf{B}\mathbf{Q} = \begin{pmatrix} \mathbf{I}_b & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \tag{18.99}$$

where $\mathbf{\Lambda}_1 \in \mathbb{R}^{b\times b}$ *and* $\mathbf{\Lambda}_2 =\in \mathbb{R}^{(n-b)\times(n-b)}$ *are diagonal matrices.*

Proof. Consider the claim (b), and let $\mathbf{B}$ have the eigenvalue decomposition

$$\mathbf{B} = (\mathbf{T}_1 : \mathbf{T}_0) \begin{pmatrix} \mathbf{D}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{T}_1' \\ \mathbf{T}_0' \end{pmatrix} = \mathbf{T}_1\mathbf{D}_1\mathbf{T}_1'\,, \tag{18.100}$$

where $\mathbf{D}_1 = \operatorname{diag}(d_1, \dots, d_b)$ with d_i being the nonzero (necessarily positive) eigenvalues of $\mathbf{B}$. Pre- and postmultiplying (18.100) by $\mathbf{K}' = \mathbf{D}_1^{-1/2}\mathbf{T}_1'$ and $\mathbf{K} = \mathbf{T}_1\mathbf{D}_1^{-1/2}$, respectively, yields $\mathbf{K}'\mathbf{B}\mathbf{K} = \mathbf{I}_b$. Now $\mathbf{K}'\mathbf{A}\mathbf{K}$ is a symmetric $b \times b$ matrix and we may denote its eigenvalue decomposition as $\mathbf{K}'\mathbf{A}\mathbf{K} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}'$, and thereby

$$\mathbf{U}'\mathbf{K}'\mathbf{A}\mathbf{K}\mathbf{U} = \mathbf{\Lambda}\,. \tag{18.101}$$

It is easy to confirm that the matrix $\mathbf{K}\mathbf{U} = \mathbf{T}_1\mathbf{D}_1^{-1/2}\mathbf{U}$ is precisely the matrix $\mathbf{L}$ satisfying both $\mathbf{L}'\mathbf{A}\mathbf{L} = \mathbf{\Lambda}$ and $\mathbf{L}'\mathbf{B}\mathbf{L} = \mathbf{I}_b$.

The proof of (c) appears in Rao & Mitra (1971b, Th. 6.2.2) and in Mitra & Rao (1968, Th. 3.2). The key in their proof is to utilize the matrix

$$\mathbf{F} = \mathbf{I}_n - \mathbf{N}(\mathbf{N}'\mathbf{A}\mathbf{N})^{-}\mathbf{N}'\mathbf{A}\,, \tag{18.102}$$

which, in light of (18.98), has the property $\mathbf{N}'\mathbf{A}\mathbf{F} = \mathbf{0}$. Defining

$$\mathbf{S} = (\mathbf{F}\mathbf{L} : \mathbf{N})\,, \tag{18.103}$$

where $\mathbf{L}_{n\times b}$ satisfies (18.97), and $\mathbf{N}_{n\times(n-b)}$ has full column rank, it is seen that

$$\mathbf{S}'\mathbf{A}\mathbf{S} = \begin{pmatrix} \mathbf{L}'\mathbf{F}'\mathbf{A}\mathbf{F}\mathbf{L} & \mathbf{0} \\ \mathbf{0} & \mathbf{N}'\mathbf{A}\mathbf{N} \end{pmatrix} := \begin{pmatrix} \mathbf{E} & \mathbf{0} \\ \mathbf{0} & \mathbf{G} \end{pmatrix}, \quad \mathbf{S}'\mathbf{B}\mathbf{S} = \begin{pmatrix} \mathbf{I}_b & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}. \tag{18.104}$$

The next step in the proof is to consider the eigenvalue decompositions $\mathbf{E} = \mathbf{U}\mathbf{\Lambda}_1\mathbf{U}'$ and $\mathbf{G} = \mathbf{V}\mathbf{\Lambda}_1\mathbf{V}'$, and choose

$$\mathbf{Q} = \mathbf{S} \begin{pmatrix} \mathbf{U} & \mathbf{0} \\ \mathbf{0} & \mathbf{V} \end{pmatrix}. \tag{18.105}$$

Then the above matrix **Q** satisfies (18.99). It is left as an exercise to confirm that **Q** is nonsingular. □

18.6 Statistical Distance

Consider the $n \times 2$ data matrix $\mathbf{U}$:

$$\mathbf{U} = (\mathbf{x} : \mathbf{y}) = \begin{pmatrix} x_1 & y_1 \\ \vdots & \vdots \\ x_n & y_n \end{pmatrix} = \begin{pmatrix} \mathbf{u}'_{(1)} \\ \vdots \\ \mathbf{u}'_{(n)} \end{pmatrix}, \tag{18.106}$$

so that the vectors $\mathbf{u}_{(i)}$ are observations from a two-dimensional variable $\mathbf{u} = \binom{x}{y}$. The Euclidean distance (squared) of the ith observation $\mathbf{u}_{(i)}$ from the mean $\bar{\mathbf{u}}$ is of course

$$\|\mathbf{u}_{(i)} - \bar{\mathbf{u}}\|^2 = (\mathbf{u}_{(i)} - \bar{\mathbf{u}})'(\mathbf{u}_{(i)} - \bar{\mathbf{u}}) = (x_i - \bar{x})^2 + (y_i - \bar{y})^2 \,. \tag{18.107}$$

While describing how far $\mathbf{u}_{(i)}$ is from the mean $\bar{\mathbf{u}}$, the Euclidean distance is surely doing mathematically good job—as stated on page 24. But given the data matrix $\mathbf{U}$, one may wonder if there is a more informative way, particularly in statistical sense, to measure the distance between $\mathbf{u}_{(i)}$ and $\bar{\mathbf{u}}$. In univariate case it is natural to calculate

$$D_i = \frac{|x_i - \bar{x}|}{\sqrt{\operatorname{var}_{\mathrm{s}}(x)}} \,, \tag{18.108}$$

because then the distance is measured in units of standard deviation and then we might well know that D_i is large if it exceeds 2 or 3. Moreover, D_i does not depend on the unit of measurement like the Euclidean distance. We can now consider a generalization of (18.108) into a multivariate situation in the spirit of Flury (1997, §5.2); see also Flury & Riedwyl (1986). For that purpose, let us define a new variable

$$z = \mathbf{a}'\mathbf{u} = a_1 x + a_2 y \,, \tag{18.109}$$

so that the n values of this new variable are represented by the variable vector

$$\mathbf{z} = \mathbf{U}\mathbf{a} = \begin{pmatrix} \mathbf{a}'\mathbf{u}_{(1)} \\ \vdots \\ \mathbf{a}'\mathbf{u}_{(n)} \end{pmatrix}. \tag{18.110}$$

Then $z_i = \mathbf{a}'\mathbf{u}_{(i)}$ and $\bar{z} = \mathbf{a}'\bar{\mathbf{u}}$, and so the distance D_i for $\mathbf{z}$ is

$$D_i(\mathbf{a}) = \frac{|z_i - \bar{z}|}{\sqrt{\mathrm{var_s}(z)}} = \frac{|\mathbf{a}'(\mathbf{u}_{(i)} - \bar{\mathbf{u}})|}{\sqrt{\mathbf{a}'\mathbf{S}\mathbf{a}}}\,, \tag{18.111}$$

where $\mathrm{cov_d}(\mathbf{U}) = \mathbf{S}$. Let us then find a vector $\mathbf{a}_*$ which maximizes the univariate distance $D_i(\mathbf{a})$.

In view of Proposition 18.4 (p. 370),

$$\max_{\mathbf{S}\mathbf{a}\neq\mathbf{0}} D_i^2(\mathbf{a}) = (\mathbf{u}_{(i)} - \bar{\mathbf{u}})'\mathbf{S}^-(\mathbf{u}_{(i)} - \bar{\mathbf{u}})\,, \tag{18.112}$$

which is nothing but the squared (sample) Mahalanobis distance

$$\mathrm{MHLN}^2(\mathbf{u}_{(i)}, \bar{\mathbf{u}}, \mathbf{S}) = \max_{\mathbf{S}\mathbf{a}\neq\mathbf{0}} D_i^2(\mathbf{a})\,. \tag{18.113}$$

The maximum is attained for any vector $\mathbf{a}_*$ proportional to $\mathbf{S}^-(\mathbf{u}_{(i)} - \bar{\mathbf{u}})$. According to (1.59) (p. 66), the sample Mahalanobis distance is invariant with respect to the choice of $\mathbf{S}^-$.

Flury (1997, §5.2) calls the squared Mahalanobis distance as the squared standard distance while some authors call it the squared statistical distance.

Following Flury, let's repeat one more time the principle used above: We look at all possible linear combinations $\mathbf{U}\mathbf{a}$ and find the one or the ones for which $\mathbf{a}'\mathbf{u}_{(i)}$ and $\mathbf{a}'\bar{\mathbf{u}}$ are as distant from each other as possible—in terms of the standard deviation of $\mathbf{U}\mathbf{a}$. The resulting maximal squared distance is the squared sample Mahalanobis distance $\mathrm{MHLN}^2(\mathbf{u}_{(i)}, \bar{\mathbf{u}}, \mathbf{S})$.

Let $\mathbf{S} = \mathbf{T}\boldsymbol{\Lambda}\mathbf{T}'$ be the eigenvalue decomposition of the sample covariance matrix $\mathbf{S}$ based on the $n \times 2$ data matrix $\mathbf{U}$ defined in (18.106). Let $\widetilde{\mathbf{U}}$ be the centered data matrix, and let $\mathbf{w}$ be a new two-dimensional variable defined as

$$\mathbf{w} = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \mathbf{T}'(\mathbf{u} - \bar{\mathbf{u}})\,. \tag{18.114}$$

Then the n values of the variable $\mathbf{w}$ can be written as

$$\mathbf{W}' = (\mathbf{w}_{(1)} : \ldots : \mathbf{w}_{(n)}) = \mathbf{T}'(\mathbf{u}_{(1)} - \bar{\mathbf{u}} : \ldots : \mathbf{u}_{(n)} - \bar{\mathbf{u}}) = \begin{pmatrix} \mathbf{w}_1' \\ \mathbf{w}_2' \end{pmatrix}, \tag{18.115}$$

so that the data matrix of the new variables w_1 and w_2 is

$$\begin{aligned} \mathbf{W} &= (\mathbf{w}_1 : \mathbf{w}_2) = \widetilde{\mathbf{U}}\mathbf{T} \\ &= \begin{pmatrix} (\mathbf{u}_{(1)} - \bar{\mathbf{u}})'\mathbf{T} \\ \vdots \\ (\mathbf{u}_{(n)} - \bar{\mathbf{u}})'\mathbf{T} \end{pmatrix} = \begin{pmatrix} \mathbf{w}_{(1)}' \\ \vdots \\ \mathbf{w}_{(n)}' \end{pmatrix} = \begin{pmatrix} w_{11} & w_{12} \\ \vdots & \vdots \\ w_{n1} & w_{n2} \end{pmatrix}. \end{aligned} \tag{18.116}$$

Then $\bar{\mathbf{w}} = \left(\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right)$ and $\mathrm{cov_d}(\mathbf{W}) = \boldsymbol{\Lambda} = \left(\begin{smallmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{smallmatrix}\right)$, and, corresponding to (18.50), for a positive definite $\mathbf{S}$, we have

$$\begin{aligned}
\mathrm{MHLN}^2(\mathbf{u}_{(i)}, \bar{\mathbf{u}}, \mathbf{S}) &= (\mathbf{u}_{(i)} - \bar{\mathbf{u}})'\mathbf{S}^{-1}(\mathbf{u}_{(i)} - \bar{\mathbf{u}}) \\
&= (\mathbf{u}_{(i)} - \bar{\mathbf{u}})'\mathbf{T}\boldsymbol{\Lambda}^{-1}\mathbf{T}'(\mathbf{u}_{(i)} - \bar{\mathbf{u}}) \\
&= \mathbf{w}'_{(i)}\boldsymbol{\Lambda}^{-1}\mathbf{w}_{(i)} \\
&= \frac{w_{i1}^2}{\mathrm{var}_\mathrm{s}(w_1)} + \frac{w_{i2}^2}{\mathrm{var}_\mathrm{s}(w_2)} \\
&= \mathrm{MHLN}^2(\mathbf{w}_{(i)}, \mathbf{0}, \boldsymbol{\Lambda}) \\
&= \mathrm{MHLN}^2\big(\mathbf{w}_{(i)}, \bar{\mathbf{w}}, \mathrm{cov}_\mathrm{d}(\mathbf{W})\big)\,. \qquad (18.117)
\end{aligned}$$

Thus we see that the squared sample Mahalanobis distance is the sum of squares of centered and normed (variance 1) uncorrelated variables; see also page 364.

Given two samples, represented by the data matrices $\mathbf{U}_1 \in \mathbb{R}^{n_1 \times p}$ and $\mathbf{U}_2 \in \mathbb{R}^{n_2 \times p}$, it is natural to measure the statistical distance of the corresponding sample means $\bar{\mathbf{u}}_1$ and $\bar{\mathbf{u}}_2$ by

$$D_{\bar{u}}(\mathbf{a}) = \frac{|\mathbf{a}'(\bar{\mathbf{u}}_1 - \bar{\mathbf{u}}_2)|}{\sqrt{\mathbf{a}'\mathbf{S}_\#\mathbf{a}}}\,. \qquad (18.118)$$

Above $\mathbf{S}_\#$ is an appropriately calculated unbiased pooled sample covariance matrix:

$$\mathbf{S}_\# = \frac{1}{n_1 + n_2 - 2}(\mathbf{U}_1'\mathbf{C}_{n_1}\mathbf{U}_1 + \mathbf{U}_2'\mathbf{C}_{n_2}\mathbf{U}_2)\,, \qquad (18.119)$$

where $\mathbf{C}_{n_1}$ and $\mathbf{C}_{n_2}$ are centering matrices. The maximum for $D^2_{\bar{u}}(\mathbf{a})$ is

$$\mathrm{MHLN}^2(\bar{\mathbf{u}}_1, \bar{\mathbf{u}}_2, \mathbf{S}_\#) = (\bar{\mathbf{u}}_1 - \bar{\mathbf{u}}_2)'\mathbf{S}_\#^-(\bar{\mathbf{u}}_1 - \bar{\mathbf{u}}_2)\,; \qquad (18.120)$$

see also (10.114) (p. 233). The maximizing vector $\mathbf{a}_*$ is now any vector proportional to

$$\mathbf{S}_\#^-(\bar{\mathbf{u}}_1 - \bar{\mathbf{u}}_2)\,, \qquad (18.121)$$

which is called a *linear discriminant function* for the given samples. For the linear discriminant function in the case of possibly singular covariance matrix, see, e.g., Ben-Israel & Levin (2006, p. 80).

18.7 Proper Eigenvalues; B Nonnegative Definite

Now it's high time to return to the equation

$$\mathbf{A}\mathbf{w} = \lambda\mathbf{B}\mathbf{w}, \quad \mathbf{B}\mathbf{w} \neq \mathbf{0}\,, \qquad (18.122)$$

where $\mathbf{A}_{n\times n}$ is symmetric and $\mathbf{B}_{n\times n}$ is nonnegative definite and thus possibly singular. As stated earlier, we call λ a proper eigenvalue and $\mathbf{w}$ a proper eigenvector of the pair $(\mathbf{A}, \mathbf{B})$. In such a situation there may exist a vector

$\mathbf{w} \neq \mathbf{0}$ such that $\mathbf{Aw} = \mathbf{Bw} = \mathbf{0}$, in which case

$$(\mathbf{A} - \lambda\mathbf{B})\mathbf{w} = \mathbf{0} \tag{18.123}$$

is satisfied with arbitrary λ. We call such a vector $\mathbf{w}$ an improper eigenvector of $\mathbf{A}$ with respect to $\mathbf{B}$. The space of improper eigenvectors is precisely

$$\mathscr{N}(\mathbf{A}) \cap \mathscr{N}(\mathbf{B}) = \mathscr{C}(\mathbf{A} : \mathbf{B})^{\perp}, \tag{18.124}$$

which has dimension

$$f = \dim \mathscr{N}(\mathbf{A}) \cap \mathscr{N}(\mathbf{B}) = n - \operatorname{rank}(\mathbf{A} : \mathbf{B}) . \tag{18.125}$$

We call f the number of improper eigenvalues of $\mathbf{A}$ with respect to $\mathbf{B}$.

Notice that if (18.123) holds for some nonzero $\mathbf{w}$, then necessarily

$$\det(\mathbf{A} - \lambda\mathbf{B}) = 0 \tag{18.126}$$

is satisfied for any $\lambda \in \mathbb{R}$. Then the pencil $\mathbf{A} - \lambda\mathbf{B}$ (with indeterminate λ) is said to be singular; otherwise the pencil is regular. Clearly if $f = \dim \mathscr{N}(\mathbf{A}) \cap \mathscr{N}(\mathbf{B}) > 0$, then the pencil is singular. In particular, if $\mathbf{A}$ and $\mathbf{B}$ are nonnegative definite, then according to part (c) of Proposition 18.6 (p. 370),

$$|\mathbf{A} - \lambda\mathbf{B}| = 0 \iff |\mathbf{Q}'(\mathbf{A} - \lambda\mathbf{B})\mathbf{Q}| = 0 \iff |\mathbf{\Lambda}_1 - \lambda\mathbf{I}_b||\mathbf{\Lambda}_2| = 0 . \tag{18.127}$$

From above we immediately see that if at least one element of $\mathbf{\Lambda}_2$ equals 0, then the determinant $|\mathbf{A} - \lambda\mathbf{B}|$ vanishes for every $\lambda \in \mathbb{R}$ and so the pencil $\mathbf{A} - \lambda\mathbf{B}$ is singular.

For a slightly different approach to the concept of the proper eigenvalue (for singular pencils), using the term generalized eigenvalue, see Gantmacher (1959, Ch. 12), Golub & Van Loan (1996, §7.7), Stewart & Sun (1990, Ch. 6), and Watkins (2007).

For a positive definite $\mathbf{B}$, we have, as seen earlier on page 368,

$$\operatorname{ch}(\mathbf{A}, \mathbf{B}) = \operatorname{ch}(\mathbf{B}^{-1}\mathbf{A}, \mathbf{I}_n) = \operatorname{ch}(\mathbf{B}^{-1}\mathbf{A}) = \operatorname{ch}(\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}) , \tag{18.128}$$

and

$$(\lambda, \mathbf{w}) \text{ is an eigenpair for } (\mathbf{A}, \mathbf{B}) \iff (\lambda, \mathbf{w}) \text{ is an eigenpair for } \mathbf{B}^{-1}\mathbf{A} . \tag{18.129}$$

If $\mathbf{B}$ is singular, we might wonder whether, for instance, the following might be true:

$$\operatorname{ch}(\mathbf{A}, \mathbf{B}) = \operatorname{ch}(\mathbf{B}^{+}\mathbf{A}) \, ? \tag{18.130}$$

We will soon see that (18.130) does not always hold, but, if $\mathscr{C}(\mathbf{A}) \subset \mathscr{C}(\mathbf{B})$, then indeed (18.130) holds for nonzero eigenvalues:

$$\mathscr{C}(\mathbf{A}) \subset \mathscr{C}(\mathbf{B}) \implies \operatorname{nzch}(\mathbf{A}, \mathbf{B}) = \operatorname{nzch}(\mathbf{B}^{+}\mathbf{A})\,. \tag{18.131}$$

Before returning (18.131), we will make a quick visit to some results of Rao & Mitra (1971b, §6.3); see also Mitra & Rao (1968). The following result is a consequence of part (c) of Proposition 18.6 (p. 370).

Proposition 18.7. (Rao & Mitra, 1971b, Th. 6.3.1) *Let* $\mathbf{A}_{n\times n}$ *and* $\mathbf{B}_{n\times n}$ *both be symmetric matrices of which* $\mathbf{B}$ *is nonnegative definite,* $\operatorname{rank}(\mathbf{B}) = b$*, and assume that*

$$\operatorname{rank}(\mathbf{N}'\mathbf{A}\mathbf{N}) = \operatorname{rank}(\mathbf{N}'\mathbf{A})\,, \ \textit{where } \mathbf{N} = \mathbf{B}^{\perp}. \tag{18.132}$$

Then there are precisely b *proper eigenvalues of* $\mathbf{A}$ *with respect to* $\mathbf{B}$*,*

$$\operatorname{ch}(\mathbf{A}, \mathbf{B}) = \{\lambda_1, \dots, \lambda_b\}\,, \tag{18.133}$$

some of which may be repeated or null. Also $\mathbf{w}_1, \dots, \mathbf{w}_b$*, the corresponding eigenvectors, can be so chosen that if* $\mathbf{W}$ *is an* $n \times b$ *matrix with* $\mathbf{w}_i$ *its* i*th column, then*

$$\mathbf{W}'\mathbf{B}\mathbf{W} = \mathbf{I}_b\,, \quad \mathbf{W}'\mathbf{A}\mathbf{W} = \mathbf{\Lambda}_1\,, \quad \mathscr{C}(\mathbf{A}\mathbf{W}) \subset \mathscr{C}(\mathbf{B})\,, \tag{18.134}$$

where $\mathbf{\Lambda}_1$ *is a diagonal matrix of* $\lambda_1, \dots, \lambda_b$*.*

We note that condition (18.132) holds, for example, if $\mathbf{A}$ is nonnegative definite, or if $\mathscr{C}(\mathbf{A}) \subset \mathscr{C}(\mathbf{B})$.

Proposition 18.8. (Rao & Mitra, 1971b, Th. 6.3.2) *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be as in Proposition* 18.7*, and assume that* (18.132) *holds. Then the nonzero proper eigenvalues of* $\mathbf{A}$ *with respect to* $\mathbf{B}$ *are the same as the nonzero eigenvalues of*

$$[\mathbf{A} - \mathbf{A}\mathbf{N}(\mathbf{N}'\mathbf{A}\mathbf{N})^{-}\mathbf{N}'\mathbf{A}]\mathbf{B}^{-} \tag{18.135}$$

and vice versa for any generalized inverses involved; i.e.,

$$\operatorname{nzch}(\mathbf{A}, \mathbf{B}) = \operatorname{nzch}\big([\mathbf{A} - \mathbf{A}\mathbf{N}(\mathbf{N}'\mathbf{A}\mathbf{N})^{-}\mathbf{N}'\mathbf{A}]\mathbf{B}^{-}\big)\,. \tag{18.136}$$

In particular,

$$\mathscr{C}(\mathbf{A}) \subset \mathscr{C}(\mathbf{B}) \implies \operatorname{nzch}(\mathbf{A}, \mathbf{B}) = \operatorname{nzch}(\mathbf{A}\mathbf{B}^{-})\,. \tag{18.137}$$

Is the set $\operatorname{nzch}(\mathbf{A}\mathbf{B}^{-})$ invariant with respect to the choice of the generalized inverse $\mathbf{B}^{-}$? This is indeed the case for (symmetric) nonnegative definite $\mathbf{A}$ and $\mathbf{B}$, satisfying $\mathscr{C}(\mathbf{A}) \subset \mathscr{C}(\mathbf{B})$. To confirm this, write $\mathbf{A} = \mathbf{C}\mathbf{C}'$ and

$$\operatorname{nzch}(\mathbf{A}\mathbf{B}^{-}) = \operatorname{nzch}(\mathbf{C}\mathbf{C}'\mathbf{B}^{-}) = \operatorname{nzch}(\mathbf{C}'\mathbf{B}^{-}\mathbf{C})\,. \tag{18.138}$$

Now, in view of Theorem 12 (p. 283), the matrix product $\mathbf{C}'\mathbf{B}^{-}\mathbf{C}$ is invariant with respect to the choice of the generalized inverse $\mathbf{B}^{-}$ if and only if $\mathscr{C}(\mathbf{C}) \subset \mathscr{C}(\mathbf{B})$, which in this case of course holds.

The following proposition generalizes the part (b) of Proposition 18.3 (p. 369).

Proposition 18.9. *Suppose* $\mathbf{A}_{n\times n}$ *is symmetric and* $\mathbf{B}_{n\times n}$ *is a symmetric nonnegative definite matrix satisfying*

$$\mathscr{C}(\mathbf{A}) \subset \mathscr{C}(\mathbf{B}) . \tag{18.139}$$

Then

$$\max_{\mathbf{Bx}\neq\mathbf{0}} \frac{\mathbf{x}'\mathbf{Ax}}{\mathbf{x}'\mathbf{Bx}} = \frac{\mathbf{w}_1'\mathbf{Aw}_1}{\mathbf{w}_1'\mathbf{Bw}_1} = \lambda_1 = \mathrm{ch}_1(\mathbf{B}^{+}\mathbf{A}) = \mathrm{ch}_1(\mathbf{A},\mathbf{B}) , \tag{18.140}$$

where λ_1 *is the largest proper eigenvalue of* $\mathbf{A}$ *with respect to* $\mathbf{B}$ *and* $\mathbf{w}_1$ *the corresponding proper eigenvector satisfying*

$$\mathbf{Aw}_1 = \lambda_1\mathbf{Bw}_1 , \quad \mathbf{Bw}_1 \neq \mathbf{0} . \tag{18.141}$$

Proof. Because $\mathbf{B}^{+1/2}\mathbf{B}^{1/2} = \mathbf{P}_{\mathbf{B}}$ and $\mathbf{P}_{\mathbf{B}}\mathbf{A} = \mathbf{A}$, we obtain, proceeding as in (18.85),

$$\begin{aligned}
\max_{\mathbf{Bx}\neq\mathbf{0}} \frac{\mathbf{x}'\mathbf{Ax}}{\mathbf{x}'\mathbf{Bx}} &= \max_{\mathbf{Bx}\neq\mathbf{0}} \frac{\mathbf{x}'\mathbf{B}^{1/2}\cdot\mathbf{B}^{+1/2}\mathbf{AB}^{+1/2}\cdot\mathbf{B}^{1/2}\mathbf{x}}{\mathbf{x}'\mathbf{B}^{1/2}\cdot\mathbf{B}^{1/2}\mathbf{x}} \\
&= \max_{\mathbf{z}\in\mathscr{C}(\mathbf{B}^{1/2}),\,\mathbf{z}\neq\mathbf{0}} \frac{\mathbf{z}'\mathbf{B}^{+1/2}\mathbf{AB}^{+1/2}\mathbf{z}}{\mathbf{z}'\mathbf{z}} \\
&\leq \mathrm{ch}_1(\mathbf{B}^{+1/2}\mathbf{AB}^{+1/2}) \\
&= \mathrm{ch}_1(\mathbf{B}^{+}\mathbf{A}) = \mathrm{ch}_1(\mathbf{B}^{-}\mathbf{A}) \quad \text{for all } \mathbf{z} = \mathbf{B}^{1/2}\mathbf{x} .
\end{aligned} \tag{18.142}$$

Now, if $\mathbf{B}$ is singular, the vector $\mathbf{z} \in \mathscr{C}(\mathbf{B}^{1/2}) = \mathscr{C}(\mathbf{B})$ is not free to vary through the whole $\mathbb{R}^n$ and hence we have "$\leq$" above. However, if $\lambda_1 = \mathrm{ch}_1(\mathbf{B}^{+}\mathbf{A})$ and $\mathbf{w}_1$ is the corresponding eigenvector of $\mathbf{B}^{+}\mathbf{A}$, then

$$\mathbf{B}^{+}\mathbf{Aw}_1 = \lambda_1\mathbf{w}_1 . \tag{18.143}$$

Premultiplying (18.143) by $\mathbf{B}$ yields, in light of (18.139), $\mathbf{Aw}_1 = \lambda_1\mathbf{Bw}_1$, and thereby $\mathbf{w}_1'\mathbf{Aw}_1 = \lambda_1\mathbf{w}_1'\mathbf{Bw}_1$. Hence, in case of $\mathscr{C}(\mathbf{A}) \subset \mathscr{C}(\mathbf{B})$, we have

$$\max_{\mathbf{Bx}\neq\mathbf{0}} \frac{\mathbf{x}'\mathbf{Ax}}{\mathbf{x}'\mathbf{Bx}} = \frac{\mathbf{w}_1'\mathbf{Aw}_1}{\mathbf{w}_1'\mathbf{Bw}_1} = \lambda_1 = \mathrm{ch}_1(\mathbf{B}^{+}\mathbf{A}) = \mathrm{ch}_1(\mathbf{A},\mathbf{B}) . \tag{18.144}$$

On the other hand, if $\mathscr{C}(\mathbf{A}) \subset \mathscr{C}(\mathbf{B})$ does not hold, then, according to de Leeuw (1982, p. 91) the ratio $\mathbf{x}'\mathbf{Ax}/\mathbf{x}'\mathbf{Bx}$ can be made as big as one wants. See also Ben-Israel & Levin (2006, p. 81). □

Assume now that $\mathscr{C}(\mathbf{A}) \subset \mathscr{C}(\mathbf{B})$ and that λ is a nonzero solution to

$$\mathbf{Aw} = \lambda\mathbf{Bw} , \quad \text{for some } \mathbf{Bw} \neq \mathbf{0} , \tag{18.145}$$

that is, $(\lambda, \mathbf{w})$ is a proper eigenpair of $(\mathbf{A}, \mathbf{B})$. Premultiplying (18.145) by $\mathbf{B}^+$ yields

$$\mathbf{B}^+\mathbf{A}\mathbf{w} = \lambda\mathbf{B}^+\mathbf{B}\mathbf{w} = \lambda\mathbf{P}_{\mathbf{B}}\mathbf{w}\,, \tag{18.146}$$

and hence, in light of $\mathbf{A}\mathbf{P}_{\mathbf{B}} = \mathbf{A}$, we have

$$\mathbf{B}^+\mathbf{A}\mathbf{P}_{\mathbf{B}}\mathbf{w} = \lambda\mathbf{P}_{\mathbf{B}}\mathbf{w}\,. \tag{18.147}$$

Because in (18.145) the scalar λ is nonzero, also the vector $\mathbf{A}\mathbf{w} = \mathbf{A}\mathbf{P}_{\mathbf{B}}\mathbf{w}$ is a nonzero vector which further implies that $\mathbf{P}_{\mathbf{B}}\mathbf{w} \neq \mathbf{0}$. Hence the equation (18.147) shows that $(\lambda, \mathbf{P}_{\mathbf{B}}\mathbf{w})$ is an eigenpair for $\mathbf{B}^+\mathbf{A}$, and so we have confirmed that for nonzero λ

$$\begin{aligned} &(\lambda, \mathbf{w}) \text{ is a proper eigenpair of } (\mathbf{A}, \mathbf{B}) \\ &\qquad \implies (\lambda, \mathbf{P}_{\mathbf{B}}\mathbf{w}) \text{ is a proper eigenpair of } \mathbf{B}^+\mathbf{A}\,. \end{aligned} \tag{18.148}$$

It is not difficult to show that the implication in (18.148) holds also in the reverse direction, and thus it is a generalization of (18.81) (p. 368).

18.8 Eigenvalues of the BLUE's Covariance Matrix

Let's take a look at the following somewhat peculiar result:

Proposition 18.10. *Consider the linear model* $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$. *The nonzero proper eigenvalues of* $\mathbf{V}$ *with respect to* $\mathbf{H}$ *are the same as the nonzero eigenvalues of the covariance matrix of* $\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta})$.

Proof. The proof follows at once from Proposition 18.8. To do this, consider the equation

$$\mathbf{V}\mathbf{w} = \lambda\mathbf{H}\mathbf{w}, \quad \mathbf{H}\mathbf{w} \neq \mathbf{0}. \tag{18.149}$$

Applying Proposition 18.8 we immediately observe that the nonzero proper eigenvalues of $\mathbf{V}$ with respect to $\mathbf{H}$ are the nonzero eigenvalues of

$$[\mathbf{V} - \mathbf{V}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^-\mathbf{M}\mathbf{V}]\mathbf{H}^- \tag{18.150}$$

which are exactly the same as the nonzero eigenvalues of

$$\mathbf{H}[\mathbf{V} - \mathbf{V}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^-\mathbf{M}\mathbf{V}]\mathbf{H} \tag{18.151}$$

which is the covariance matrix of the $\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta})$.

An alternative proof, without using Proposition 18.8, is done by Isotalo, Puntanen & Styan (2008a, Th. 2.3). □

18.9 Canonical Correlations and Proper Eigenvalues

The maximal value of the ratio

$$\frac{(\boldsymbol{\alpha}'\mathbf{A}'\mathbf{B}\boldsymbol{\beta})^2}{\boldsymbol{\alpha}'\mathbf{A}'\mathbf{A}\boldsymbol{\alpha}\cdot\boldsymbol{\beta}'\mathbf{B}'\mathbf{B}\boldsymbol{\beta}}\,, \tag{18.152}$$

where $\mathbf{A}_{n\times a}$ and $\mathbf{B}_{n\times b}$ are given matrices, is being considered in various places of this book; see Section 2.1 (p. 77) and Section 5.9 (p. 133).

According to Proposition 2.4 (p. 78), we have

$$\begin{aligned}\max_{\substack{\mathbf{A}\boldsymbol{\alpha}\neq\mathbf{0},\\ \mathbf{B}\boldsymbol{\beta}\neq\mathbf{0}}}\frac{(\boldsymbol{\alpha}'\mathbf{A}'\mathbf{B}\boldsymbol{\beta})^2}{\boldsymbol{\alpha}'\mathbf{A}'\mathbf{A}\boldsymbol{\alpha}\cdot\boldsymbol{\beta}'\mathbf{B}'\mathbf{B}\boldsymbol{\beta}} &= \frac{(\boldsymbol{\alpha}_1'\mathbf{A}'\mathbf{B}\boldsymbol{\beta}_1)^2}{\boldsymbol{\alpha}_1'\mathbf{A}'\mathbf{A}\boldsymbol{\alpha}_1\cdot\boldsymbol{\beta}_1'\mathbf{B}'\mathbf{B}\boldsymbol{\beta}_2}\\ &= \frac{\boldsymbol{\alpha}_1'\mathbf{A}'\mathbf{P}_{\mathbf{B}}\mathbf{A}\boldsymbol{\alpha}_1}{\boldsymbol{\alpha}_1'\mathbf{A}'\mathbf{A}\boldsymbol{\alpha}_1} = \mathrm{ch}_1(\mathbf{P}_{\mathbf{A}}\mathbf{P}_{\mathbf{B}}) = \lambda_1^2\,,\end{aligned} \tag{18.153}$$

where λ_1^2 is the largest proper eigenvalue of $\mathbf{A}'\mathbf{P}_{\mathbf{B}}\mathbf{A}$ with respect to $\mathbf{A}'\mathbf{A}$ and $\boldsymbol{\alpha}_1$ is the corresponding proper eigenvector satisfying

$$\mathbf{A}'\mathbf{P}_{\mathbf{B}}\mathbf{A}\boldsymbol{\alpha}_1 = \lambda_1^2\mathbf{A}'\mathbf{A}\boldsymbol{\alpha}_1\,, \quad \mathbf{A}\boldsymbol{\alpha}_1 \neq \mathbf{0}\,. \tag{18.154}$$

The vector $\boldsymbol{\beta}_1$ is the solution to

$$\mathbf{B}'\mathbf{P}_{\mathbf{A}}\mathbf{B}\boldsymbol{\beta}_1 = \lambda_1^2\mathbf{B}'\mathbf{B}\boldsymbol{\beta}_1\,, \quad \mathbf{B}\boldsymbol{\beta}_1 \neq \mathbf{0}\,. \tag{18.155}$$

Consider now an n-dimensional random vector $\mathbf{u}$ whose covariance matrix is $\mathrm{cov}(\mathbf{u}) = \mathbf{I}_n$. Then $\mathbf{x} = \mathbf{A}'\mathbf{u}$ and $\mathbf{y} = \mathbf{B}'\mathbf{u}$ are a- and b-dimensional random vectors whose covariance matrix is

$$\mathrm{cov}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix} = \mathrm{cov}\begin{pmatrix}\mathbf{A}'\mathbf{u}\\ \mathbf{B}'\mathbf{u}\end{pmatrix} = \begin{pmatrix}\mathbf{A}'\mathbf{A} & \mathbf{A}'\mathbf{B}\\ \mathbf{B}'\mathbf{A} & \mathbf{B}'\mathbf{B}\end{pmatrix} := \begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{pmatrix} = \boldsymbol{\Sigma}\,. \tag{18.156}$$

Let ϱ_i denote the ith largest canonical correlation between the random vectors $\mathbf{A}'\mathbf{u}$ and $\mathbf{B}'\mathbf{u}$. Then ϱ_1 is defined as the maximum value of the correlation between arbitrary linear combinations $\boldsymbol{\alpha}'\mathbf{A}'\mathbf{u}$ and $\boldsymbol{\beta}'\mathbf{B}'\mathbf{u}$. Let $\boldsymbol{\alpha}_1$ and $\boldsymbol{\beta}_1$ be the corresponding maximizing values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Then $\boldsymbol{\alpha}_1'\mathbf{A}'\mathbf{u}$ and $\boldsymbol{\beta}_1'\mathbf{B}'\mathbf{u}$ are called the first canonical variables. In view of (18.153),

$$\varrho_1^2 = \max_{\substack{\mathbf{A}\boldsymbol{\alpha}\neq\mathbf{0},\\ \mathbf{B}\boldsymbol{\beta}\neq\mathbf{0}}}\frac{(\boldsymbol{\alpha}'\mathbf{A}'\mathbf{B}\boldsymbol{\beta})^2}{\boldsymbol{\alpha}'\mathbf{A}'\mathbf{A}\boldsymbol{\alpha}\cdot\boldsymbol{\beta}'\mathbf{B}'\mathbf{B}\boldsymbol{\beta}} = \frac{\boldsymbol{\alpha}_1'\mathbf{A}'\mathbf{P}_{\mathbf{B}}\mathbf{A}\boldsymbol{\alpha}_1}{\boldsymbol{\alpha}_1'\mathbf{A}'\mathbf{A}\boldsymbol{\alpha}_1} = \mathrm{ch}_1(\mathbf{P}_{\mathbf{A}}\mathbf{P}_{\mathbf{B}})\,. \tag{18.157}$$

In other words, ϱ_1^2 is the largest proper eigenvalue of $\mathbf{A}'\mathbf{P}_{\mathbf{B}}\mathbf{A}$ with respect to $\mathbf{A}'\mathbf{A}$ and $\boldsymbol{\alpha}_1$ is the corresponding proper eigenvector satisfying

$$\mathbf{A}'\mathbf{P}_{\mathbf{B}}\mathbf{A}\boldsymbol{\alpha}_1 = \varrho_1^2\mathbf{A}'\mathbf{A}\boldsymbol{\alpha}_1\,, \quad \mathbf{A}\boldsymbol{\alpha}_1 \neq \mathbf{0}\,. \tag{18.158}$$

Moreover, $\mathbf{A}\boldsymbol{\alpha}_1$ is the eigenvector of $\mathbf{P}_\mathbf{A}\mathbf{P}_\mathbf{B}$ with respect to the eigenvalue ϱ_1^2:

$$\mathbf{P}_\mathbf{A}\mathbf{P}_\mathbf{B}\mathbf{A}\boldsymbol{\alpha}_1 = \varrho_1^2\mathbf{A}\boldsymbol{\alpha}_1\,, \tag{18.159}$$

and similarly $\mathbf{B}\boldsymbol{\beta}_1$ is the eigenvector of $\mathbf{P}_\mathbf{B}\mathbf{P}_\mathbf{A}$ with respect to the eigenvalue ϱ_1^2:

$$\mathbf{P}_\mathbf{B}\mathbf{P}_\mathbf{A}\mathbf{B}\boldsymbol{\beta}_1 = \varrho_1^2\mathbf{B}\boldsymbol{\beta}_1\,. \tag{18.160}$$

We also have

$$\begin{aligned}\varrho_1^2 &= \mathrm{ch}_1[\mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{A}'\mathbf{B}(\mathbf{B}'\mathbf{B})^-\mathbf{B}'] \\ &= \mathrm{ch}_1[\mathbf{B}'\mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{A}'\mathbf{B}(\mathbf{B}'\mathbf{B})^-] \\ &= \mathrm{ch}_1(\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^-)\,. \end{aligned} \tag{18.161}$$

In particular, if $\mathbf{A}$ and $\mathbf{B}$ have full column ranks,

$$(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{P}_\mathbf{B}\mathbf{A}\boldsymbol{\alpha}_1 = \varrho_1^2\boldsymbol{\alpha}_1\,, \tag{18.162a}$$

$$(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\mathbf{P}_\mathbf{A}\mathbf{B}\boldsymbol{\beta}_1 = \varrho_1^2\boldsymbol{\beta}_1\,, \tag{18.162b}$$

or in other notation,

$$\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\alpha}_1 = \varrho_1^2\boldsymbol{\alpha}_1\,, \tag{18.162c}$$

$$\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\beta}_1 = \varrho_1^2\boldsymbol{\beta}_1\,. \tag{18.162d}$$

The second largest canonical correlation ϱ_2 is defined as the maximum correlation between $\boldsymbol{\alpha}'\mathbf{A}'\mathbf{u}$ and $\boldsymbol{\beta}'\mathbf{B}'\mathbf{u}$, but now $\mathbf{A}\boldsymbol{\alpha}$ and $\mathbf{B}\boldsymbol{\beta}$ are not arbitrary vectors of $\mathscr{C}(\mathbf{A})$ and $\mathscr{C}(\mathbf{B})$; they are subject to constrains that $\boldsymbol{\alpha}'\mathbf{A}'\mathbf{u}$ is uncorrelated with $\boldsymbol{\alpha}_1'\mathbf{A}'\mathbf{u}$, and $\boldsymbol{\beta}'\mathbf{B}'\mathbf{u}$ is uncorrelated with $\boldsymbol{\beta}_1'\mathbf{B}'\mathbf{u}$ so that

$$\mathrm{cov}(\boldsymbol{\alpha}'\mathbf{A}'\mathbf{u}, \boldsymbol{\alpha}_1'\mathbf{A}'\mathbf{u}) = \boldsymbol{\alpha}'\mathbf{A}'\mathbf{A}\boldsymbol{\alpha}_1 = 0\,, \tag{18.163a}$$

$$\mathrm{cov}(\boldsymbol{\beta}'\mathbf{B}'\mathbf{u}, \boldsymbol{\beta}_1'\mathbf{B}'\mathbf{u}) = \boldsymbol{\beta}'\mathbf{B}'\mathbf{B}\boldsymbol{\beta}_1 = 0\,. \tag{18.163b}$$

The third (and higher order) canonical correlations are defined in the corresponding way. We may assume that

$$\mathrm{rank}(\mathbf{A}) \leq \mathrm{rank}(\mathbf{B})\,, \tag{18.164}$$

in which case the number of the canonical correlations, i.e., the number of pairs of canonical variables $\boldsymbol{\alpha}_i'\mathbf{A}'\mathbf{u}$ and $\boldsymbol{\beta}_i'\mathbf{B}'\mathbf{u}$, is r, where $r = \mathrm{rank}(\mathbf{A})$.

Let $\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_r$ satisfy

$$\mathbf{A}'\mathbf{P}_\mathbf{B}\mathbf{A}\boldsymbol{\alpha}_i = \varrho_i^2\mathbf{A}'\mathbf{A}\boldsymbol{\alpha}_i\,, \quad \mathbf{A}\boldsymbol{\alpha}_i \neq \mathbf{0}\,, \quad i = 1, \dots, r\,, \tag{18.165}$$

or, equivalently,

$$\mathbf{P}_\mathbf{A}\mathbf{P}_\mathbf{B}\mathbf{A}\boldsymbol{\alpha}_i = \varrho_i^2\mathbf{A}\boldsymbol{\alpha}_i\,, \quad \mathbf{A}\boldsymbol{\alpha}_i \neq \mathbf{0}\,, \quad i = 1, \dots, r\,. \tag{18.166}$$

It is easy to confirm that if $\varrho_i^2 \neq \varrho_j^2$, then $\boldsymbol{\alpha}_i'\mathbf{A}'\mathbf{A}\boldsymbol{\alpha}_j = 0$, i.e., $\boldsymbol{\alpha}_i'\mathbf{A}'\mathbf{u}$ and $\boldsymbol{\alpha}_j'\mathbf{A}'\mathbf{u}$ are uncorrelated. Even in the case of multiple roots ϱ_i^2's, it is possible to choose $\boldsymbol{\alpha}_i$ so that $\boldsymbol{\alpha}_i'\mathbf{A}'\mathbf{A}\boldsymbol{\alpha}_j = 0$ for all $i \neq j$. One simple way to conclude this is to use Proposition 18.7 (p. 376).

The following proposition gives a summary about the properties of canonical correlations based on the covariance matrix

$$\operatorname{cov}\begin{pmatrix}\mathbf{A}'\mathbf{u}\\ \mathbf{B}'\mathbf{u}\end{pmatrix} = \begin{pmatrix}\mathbf{A}'\mathbf{A} & \mathbf{A}'\mathbf{B}\\ \mathbf{B}'\mathbf{A} & \mathbf{B}'\mathbf{B}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{pmatrix} = \boldsymbol{\Sigma}\,, \tag{18.167}$$

where $r = \operatorname{rank}(\mathbf{A}) \leq \operatorname{rank}(\mathbf{B})$. We may denote the ith largest canonical correlation as $\operatorname{cc}_i(\mathbf{A}'\mathbf{u}, \mathbf{B}'\mathbf{u}) = \varrho_i$, and

$$\operatorname{cc}(\mathbf{A}'\mathbf{u}, \mathbf{B}'\mathbf{u}) = \{\varrho_1, \dots, \varrho_r\} = \text{the set of all cc's}, \tag{18.168a}$$

$$\operatorname{cc}_+(\mathbf{A}'\mathbf{u}, \mathbf{B}'\mathbf{u}) = \{\varrho_1, \dots, \varrho_m\} = \text{the set of nonzero cc's}, \tag{18.168b}$$

where $m = \operatorname{rank}(\mathbf{A}'\mathbf{B})$.

Proposition 18.11. *With the above notation, the following statements hold:*

(a) *there are* $r = \operatorname{rank}(\mathbf{A})$ *pairs of canonical variables* $\boldsymbol{\alpha}_i'\mathbf{A}'\mathbf{u}, \boldsymbol{\beta}_i'\mathbf{B}'\mathbf{u}$*, and* r *corresponding canonical correlations* $\varrho_1 \geq \varrho_2 \geq \dots \geq \varrho_r$*,*

(b) *the vectors* $\boldsymbol{\alpha}_i$ *are the proper eigenvectors of* $\mathbf{A}'\mathbf{P}_\mathbf{B}\mathbf{A}$ *with respect to* $\mathbf{A}'\mathbf{A}$*, and* ϱ_i^2*'s are the corresponding proper eigenvalues,*

(c) *the* ϱ_i^2*'s are the* r *largest eigenvalues of* $\mathbf{P}_\mathbf{A}\mathbf{P}_\mathbf{B}$*,*

(d) *the nonzero* ϱ_i^2*'s are the nonzero eigenvalues of* $\mathbf{P}_\mathbf{A}\mathbf{P}_\mathbf{B}$*, or in a short notation,*

$$\operatorname{cc}_+^2(\mathbf{A}'\mathbf{u}, \mathbf{B}'\mathbf{u}) = \operatorname{nzch}(\mathbf{P}_\mathbf{A}\mathbf{P}_\mathbf{B})\,, \tag{18.169}$$

(e) *the vectors* $\boldsymbol{\beta}_i$ *are the proper eigenvectors of* $\mathbf{B}'\mathbf{P}_\mathbf{A}\mathbf{B}$ *with respect to* $\mathbf{B}'\mathbf{B}$*,*

(f) *the number of nonzero* ϱ_i*'s is*

$$m = \operatorname{rank}(\mathbf{A}'\mathbf{B}) = \operatorname{rank}(\mathbf{A}) - \dim \mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B})^\perp, \tag{18.170}$$

(g) *the number of unit* ϱ_i*'s is*

$$u = \dim \mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B})\,, \tag{18.171}$$

(h) *the number of zero* ϱ_i*'s is*

$$s = \operatorname{rank}(\mathbf{A}) - \operatorname{rank}(\mathbf{A}'\mathbf{B}) = \dim \mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\mathbf{B})^\perp. \tag{18.172}$$

18.10 More about Canonical Correlations and the Watson Efficiency

In Section 10.8 (p. 241) we considered the canonical correlations between $\mathbf{X}'\mathbf{y}$ and $\mathbf{Z}'\mathbf{y}$, and between $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$. Those consideration were made assuming that $\mathbf{X}$ has full column rank and the model is a weakly singular linear model. We denoted

$$\kappa_i = \mathrm{cc}_i(\mathbf{X}'\mathbf{y}, \mathbf{Z}'\mathbf{y}) , \quad \theta_i = \mathrm{cc}_i(\hat{\boldsymbol{\beta}}, \tilde{\boldsymbol{\beta}}) , \quad i = 1, \dots, p , \tag{18.173}$$

and observed that

$$\theta_i^2 = 1 - \kappa_{p-i+1}^2 , \quad i = 1, \dots, p , \tag{18.174a}$$

$$\mathrm{cc}_i^2(\hat{\boldsymbol{\beta}}, \tilde{\boldsymbol{\beta}}) = 1 - \mathrm{cc}_i^2(\mathbf{X}'\mathbf{y}, \mathbf{Z}'\mathbf{y}) , \quad i = 1, \dots, p . \tag{18.174b}$$

If $\mathbf{X}$ has not full column rank, then $\boldsymbol{\beta}$ is not estimable and naturally it does not have a BLUE. In such a situation we mean by $\tilde{\boldsymbol{\beta}}$ such an estimator for which $\mathbf{X}\tilde{\boldsymbol{\beta}}$ is the BLUE for $\mathbf{X}\boldsymbol{\beta}$, and correspondingly $\hat{\boldsymbol{\beta}}$ has the property $\mathbf{H}\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}}$.

Let us now relax from all rank assumptions and consider the linear model $\mathscr{M} = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \mathbf{V}\}$ and denote

$$\mathbf{G} = \mathbf{H} - \mathbf{H}\mathbf{V}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^{-}\mathbf{M} . \tag{18.175}$$

Then $\mathbf{G}\mathbf{y} = \mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta})$, and its covariance matrix is

$$\begin{aligned} \mathrm{cov}[\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta})] &= \mathbf{G}\mathbf{V}\mathbf{G}' \\ &= \mathbf{H}\mathbf{V}\mathbf{H} - \mathbf{H}\mathbf{V}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^{-}\mathbf{M}\mathbf{V}\mathbf{H} \\ &= \mathbf{H}\mathbf{V}^{1/2}(\mathbf{I} - \mathbf{P}_{\mathbf{V}^{1/2}\mathbf{M}})\mathbf{V}^{1/2}\mathbf{H} \\ &= \mathbf{K}'(\mathbf{I} - \mathbf{P}_{\mathbf{L}})\mathbf{K} , \end{aligned} \tag{18.176}$$

where $\mathbf{K} = \mathbf{V}^{1/2}\mathbf{H}$ and $\mathbf{L} = \mathbf{V}^{1/2}\mathbf{M}$. Note that in (10.169) (p. 242) we had $\mathbf{K} = \mathbf{V}^{1/2}\mathbf{X}$ and $\mathbf{L} = \mathbf{V}^{1/2}\mathbf{Z}$. Moreover,

$$\mathrm{cov}\begin{pmatrix} \mathbf{H}\mathbf{y} \\ \mathbf{M}\mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{H}\mathbf{V}\mathbf{H} & \mathbf{H}\mathbf{V}\mathbf{M} \\ \mathbf{M}\mathbf{V}\mathbf{H} & \mathbf{M}\mathbf{V}\mathbf{M} \end{pmatrix} = \begin{pmatrix} \mathbf{K}'\mathbf{K} & \mathbf{K}'\mathbf{L} \\ \mathbf{L}'\mathbf{K} & \mathbf{L}'\mathbf{L} \end{pmatrix} , \tag{18.177}$$

and in view of $\mathrm{cov}(\mathbf{G}\mathbf{y}, \mathbf{H}\mathbf{y}) = \mathbf{G}\mathbf{V}\mathbf{H} = \mathbf{G}\mathbf{V}\mathbf{G}'$, we have

$$\mathrm{cov}\begin{pmatrix} \mathbf{H}\mathbf{y} \\ \mathbf{G}\mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{H}\mathbf{V}\mathbf{H} & \mathbf{G}\mathbf{V}\mathbf{G}' \\ \mathbf{G}\mathbf{V}\mathbf{G}' & \mathbf{G}\mathbf{V}\mathbf{G}' \end{pmatrix} = \begin{pmatrix} \mathbf{K}'\mathbf{K} & \mathbf{K}'\mathbf{Q}_{\mathbf{L}}\mathbf{K} \\ \mathbf{K}'\mathbf{Q}_{\mathbf{L}}\mathbf{K} & \mathbf{K}'\mathbf{Q}_{\mathbf{L}}\mathbf{K} \end{pmatrix} , \tag{18.178}$$

where $\mathbf{Q}_{\mathbf{L}} = \mathbf{I} - \mathbf{P}_{\mathbf{L}}$. Denote

$$\mathbf{T}_1 = (\mathbf{H}\mathbf{V}\mathbf{H})^{-}\mathbf{H}\mathbf{V}\mathbf{M}(\mathbf{M}\mathbf{V}\mathbf{M})^{-}\mathbf{M}\mathbf{V}\mathbf{H} , \tag{18.179a}$$

$$\mathbf{T}_2 = (\mathbf{HVH})^-\mathbf{GVG}\,. \tag{18.179b}$$

Then the nonzero canonical correlations between the random vectors $\mathbf{Hy}$ and $\mathbf{My}$ are the nonzero eigenvalues of the matrix $\mathbf{T}_1$, i.e.,

$$\begin{aligned}\mathrm{cc}_+^2(\mathbf{Hy},\mathbf{My}) &= \{\kappa_1^2,\dots,\kappa_m^2\} = \mathrm{nzch}(\mathbf{T}_1)\\ &= \mathrm{nzch}[(\mathbf{HVH})^-\mathbf{HVM}(\mathbf{MVM})^-\mathbf{MVH}]\\ &= \mathrm{nzch}(\mathbf{P_K P_L})\,, \end{aligned}\tag{18.180}$$

where $m = \mathrm{rank}(\mathbf{HVM})$. We may assume that $\mathrm{rank}(\mathbf{VH}) \leq \mathrm{rank}(\mathbf{VM})$, in which case the number of the canonical correlations is h, where $h = \mathrm{rank}(\mathbf{VH})$. The nonzero canonical correlations between $\mathbf{Hy}$ and $\mathbf{Gy}$ are the nonzero eigenvalues of the matrix $\mathbf{T}_2$:

$$\begin{aligned}\mathrm{cc}_+^2(\mathbf{Hy},\mathbf{Gy}) &= \{\theta_1^2,\dots,\theta_g^2\} = \mathrm{nzch}(\mathbf{T}_2)\\ &= \mathrm{nzch}[(\mathbf{HVH})^-\mathbf{GVG}']\\ &= \mathrm{nzch}[(\mathbf{K}'\mathbf{K})^-\mathbf{K}'(\mathbf{I}-\mathbf{P_L})\mathbf{K}]\\ &= \mathrm{nzch}[\mathbf{P_K}(\mathbf{I}-\mathbf{P_L})]\,, \end{aligned}\tag{18.181}$$

where $g = \mathrm{rank}(\mathbf{GVG}') = \dim \mathscr{C}(\mathbf{X}) \cap \mathscr{C}(\mathbf{V})$.

Now the squared canonical correlations κ_i^2's are the proper eigenvalues of $\mathbf{K}'\mathbf{P_L K}$ with respect to $\mathbf{K}'\mathbf{K}$, i.e.,

$$\mathbf{HVM}(\mathbf{MVM})^-\mathbf{MVHw} = \kappa^2\mathbf{HVHw}\,, \quad \mathbf{HVHw} \neq \mathbf{0}\,, \tag{18.182}$$

which, in light of $\mathbf{HVM}(\mathbf{MVM})^-\mathbf{MVH} = \mathbf{HVH} - \mathbf{GVG}'$, can be written as

$$\mathbf{GVG}'\mathbf{w} = (1-\kappa^2)\mathbf{HVHw}\,, \quad \mathbf{HVHw} \neq \mathbf{0}\,. \tag{18.183}$$

Similarly the squared canonical correlations θ_i^2's are the proper eigenvalues of $\mathbf{K}'(\mathbf{I}-\mathbf{P_L})\mathbf{K} = \mathbf{GVG}'$ with respect to $\mathbf{K}'\mathbf{K} = \mathbf{HVH}$:

$$\mathbf{GVG}'\mathbf{w} = \theta^2\mathbf{HVHw}\,, \quad \mathbf{HVHw} \neq \mathbf{0}\,. \tag{18.184}$$

Hence

$$\mathbf{GVG}'\mathbf{w} = \theta^2\mathbf{HVHw} = (1-\kappa^2)\mathbf{HVHw}\,, \quad \mathbf{HVHw} \neq \mathbf{0}\,, \tag{18.185}$$

and we can conclude the following.

Proposition 18.12. *Under the linear model* $\{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$ *we have the following relation:*

$$\mathrm{cc}_i^2(\mathbf{Hy},\mathbf{My}) = 1 - \mathrm{cc}_{h-i+1}^2(\mathbf{Hy},\mathbf{Gy})\,, \quad i = 1,\dots,h\,, \tag{18.186}$$

where $h = \mathrm{rank}(\mathbf{VH})$.

We can use the following notation:

$$\kappa_1^2 = \cdots = \kappa_u^2 = 1\,, \tag{18.187a}$$

$$1 > \kappa_{u+1}^2 \geq \cdots \geq \kappa_{u+t}^2 = \kappa_m^2 > 0\,, \tag{18.187b}$$

$$\kappa_{u+t+1}^2 = \kappa_{m+1}^2 = \cdots = \kappa_{m+s}^2 = \kappa_h^2 = 0\,, \tag{18.187c}$$

and of course

$$\kappa_i^2 = 1 - \theta_{h-i+1}^2\,, \quad i = 1, \dots, h = \operatorname{rank}(\mathbf{VH})\,. \tag{18.188}$$

The value of t above is the number of unit eigenvalues of the product $\mathbf{P_K P_L}$, see (5.83) (p. 134):

$$u = \dim \mathscr{C}(\mathbf{K}) \cap \mathscr{C}(\mathbf{L}) = \dim \mathscr{C}(\mathbf{VX}) \cap \mathscr{C}(\mathbf{VZ}) = \operatorname{rank}(\mathbf{HP_V M})\,. \tag{18.189}$$

For the values of m, s and t, and for further related considerations, see Isotalo, Puntanen & Styan (2008a).

What about the Watson efficiency when $\mathbf{X}$ does not have full column rank, i.e., $\operatorname{rank}(\mathbf{X}) = r < p$? In such a situation $\det(\mathbf{X'VX}) = 0$ and the Watson efficiency as the ratio of the determinants of the covariance matrices is not properly defined. However, the nonzero squared canonical correlations $\kappa_1^2, \dots, \kappa_m^2$ between $\mathbf{X'y}$ and $\mathbf{Z'y}$ are the nonzero eigenvalues of the matrix

$$\mathbf{X'VZ(Z'VZ)^- Z'VX(X'VX)^-}, \tag{18.190}$$

and under a weakly singular linear model all κ_i's are less than 1. Hence

$$0 \neq |\mathbf{I}_p - \mathbf{X'VZ(Z'VZ)^- Z'VX(X'VX)^-}| = \prod_{i=1}^{m} (1 - \kappa_i^2) := \phi^*, \tag{18.191}$$

where $m = \operatorname{rank}(\mathbf{HVM})$. Now, as Chu, Isotalo, Puntanen & Styan (2005, p. 84) point out, a natural question arises: what is the (estimable) parametric function $\mathbf{L}'\boldsymbol{\beta}$, say, whose efficiency ϕ^* actually is? The answer is $\mathbf{F}'\boldsymbol{\beta}$, where $\mathbf{F}$ comes from the full rank decomposition of $\mathbf{X}$:

$$\mathbf{X} = \mathbf{X}_* \mathbf{F}'\,, \quad \mathbf{X}_* \in \mathbb{R}^{p\times r},\ \mathbf{F} \in \mathbb{R}^{p \times r},\ \operatorname{rank}(\mathbf{X}) = \operatorname{rank}(\mathbf{F}) = r\,. \tag{18.192}$$

Here the columns of $\mathbf{X}_*$ and $\mathbf{F}$ span the column spaces $\mathscr{C}(\mathbf{X})$ and $\mathscr{C}(\mathbf{X}')$, respectively. In this situation,

$$\mathbf{F} = \mathbf{X}'\mathbf{X}_*(\mathbf{X}_*'\mathbf{X}_*)^{-1} := \mathbf{X}'\mathbf{X}_*\mathbf{U}\,. \tag{18.193}$$

Because

$$\operatorname{cc}(\mathbf{X'y}, \mathbf{Z'y}) = \operatorname{cc}(\mathbf{X}_*'\mathbf{y}, \mathbf{Z'y}) = \{\kappa_i\}\,, \tag{18.194}$$

we see that

$$\phi^* = |\mathbf{I}_r - \mathbf{X}_*'\mathbf{VZ(Z'VZ)^- Z'VX}_*(\mathbf{X}_*'\mathbf{VX}_*)^{-1}|$$

$$
\begin{aligned}
&= \frac{|\mathbf{X}_*'\mathbf{V}\mathbf{X}_* - \mathbf{X}_*'\mathbf{V}\mathbf{Z}(\mathbf{Z}'\mathbf{V}\mathbf{Z})^-\mathbf{Z}'\mathbf{V}\mathbf{X}_*|}{|\mathbf{X}_*'\mathbf{V}\mathbf{X}_*|} \\
&= \frac{|\mathbf{X}_*'\mathbf{X}_*|^2}{|\mathbf{X}_*'\mathbf{V}\mathbf{X}_*| \cdot |\mathbf{X}_*'\mathbf{V}^+\mathbf{X}_*|}\,. \qquad (18.195)
\end{aligned}
$$

On the other hand, under a weakly singular model, we have,

$$
\begin{aligned}
\operatorname{cov}(\mathbf{X}\tilde{\boldsymbol\beta}) &= \mathbf{X}(\mathbf{X}'\mathbf{V}^+\mathbf{X})^-\mathbf{X}' \\
&= \mathbf{V}^{1/2}\mathbf{V}^{+1/2}\mathbf{X}(\mathbf{X}'\mathbf{V}^{+1/2}\mathbf{V}^{+1/2}\mathbf{X})^-\mathbf{V}^{+1/2}\mathbf{V}^{1/2}\mathbf{X}' \\
&= \mathbf{V}^{1/2}\mathbf{P}_{\mathbf{V}^{+1/2}\mathbf{X}}\mathbf{V}^{1/2} \\
&= \mathbf{V}^{1/2}\mathbf{P}_{\mathbf{V}^{+1/2}\mathbf{X}_*}\mathbf{V}^{1/2} = \mathbf{X}_*(\mathbf{X}_*'\mathbf{V}^+\mathbf{X}_*)^{-1}\mathbf{X}_*'\,; \qquad (18.196)
\end{aligned}
$$

see Exercise 10.6 (p. 261). Because $\mathbf{F}'\boldsymbol\beta = \mathbf{U}\mathbf{X}_*'\mathbf{X}\boldsymbol\beta$ where $\mathbf{U} = (\mathbf{X}_*'\mathbf{X}_*)^{-1}$, it is straightforward to observe that

$$
\begin{aligned}
\operatorname{cov}(\mathbf{F}'\tilde{\boldsymbol\beta}) &= \operatorname{cov}(\mathbf{U}\mathbf{X}_*'\mathbf{X}\tilde{\boldsymbol\beta}) \\
&= \mathbf{U}\mathbf{X}_*'\mathbf{X}_*(\mathbf{X}_*'\mathbf{V}^+\mathbf{X}_*)^{-1}\mathbf{X}_*'\mathbf{X}_*\mathbf{U} = (\mathbf{X}_*'\mathbf{V}^+\mathbf{X}_*)^{-1}, \qquad (18.197)
\end{aligned}
$$

and

$$
\operatorname{cov}(\mathbf{F}'\hat{\boldsymbol\beta}) = \mathbf{U}\mathbf{X}_*'\mathbf{V}\mathbf{X}_*\mathbf{U} = (\mathbf{X}_*'\mathbf{X}_*)^{-1}\mathbf{X}_*'\mathbf{V}\mathbf{X}_*(\mathbf{X}_*'\mathbf{X}_*)^{-1}. \qquad (18.198)
$$

Recall that $\tilde{\boldsymbol\beta}$ and $\hat{\boldsymbol\beta}$ refer to vectors satisfying $\mathrm{BLUE}(\mathbf{X}\boldsymbol\beta) = \mathbf{X}\tilde{\boldsymbol\beta}$ and $\mathrm{OLSE}(\mathbf{X}\boldsymbol\beta) = \mathbf{X}\hat{\boldsymbol\beta}$. For a clarity, we may summarize our finding in the following.

Proposition 18.13. *Consider a weakly singular model $\mathscr{M}$, and let $\mathbf{X}$ have a full rank decomposition $\mathbf{X} = \mathbf{X}_*\mathbf{F}'$. Then*

$$
\begin{aligned}
\phi^* &= |\mathbf{I}_p - \mathbf{X}'\mathbf{V}\mathbf{Z}(\mathbf{Z}'\mathbf{V}\mathbf{Z})^-\mathbf{Z}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{V}\mathbf{X})^-| \\
&= |\mathbf{I}_r - \mathbf{X}_*'\mathbf{V}\mathbf{Z}(\mathbf{Z}'\mathbf{V}\mathbf{Z})^-\mathbf{Z}'\mathbf{V}\mathbf{X}_*(\mathbf{X}_*'\mathbf{V}\mathbf{X}_*)^{-1}| \\
&= \operatorname{eff}(\mathbf{F}'\hat{\boldsymbol\beta}) = \frac{|\operatorname{cov}(\mathbf{F}'\tilde{\boldsymbol\beta})|}{|\operatorname{cov}(\mathbf{F}'\hat{\boldsymbol\beta})|} = \prod_{i=1}^{m}(1-\kappa_i^2)\,, \qquad (18.199)
\end{aligned}
$$

where $m = \operatorname{rank}(\mathbf{X}'\mathbf{V}\mathbf{Z})$, and $\kappa_1, \dots, \kappa_m$ are the (nonzero) canonical correlations between the OLS *fitted values and the residuals.*

It is worth noting that the ϕ^* is properly defined if there are no unit canonical correlations between $\mathbf{Hy}$ and $\mathbf{My}$, i.e., $u = 0$. For this the weak singularity of the model is sufficient but not necessary. Hence we could extend Proposition 18.13 by assuming only that $u = 0$. However, in this situation, (18.196) does not necessarily hold. We survive from the problems by assuming that the columns of $\mathbf{X}_*$ are orthonormal—the details are left to the reader.

For further references regarding canonical correlations in the case of singular $\mathbf{V}$, the reader is referred to Khatri (1976, 1978, 1989, 1990), Seshadri & Styan (1980), Rao (1981), Latour, Puntanen & Styan (1987), Wang & Chow (1987), SenGupta (1991), Baksalary, Puntanen & Yanai (1992), Yanai & Puntanen (1993), and Drury, Liu, Lu, Puntanen et al. (2002).

18.11 Exercises

18.1. Let $\mathbf{A}_{2\times 2}$ be symmetric. Show that

$$\mathbf{A} \geq_{\mathsf{L}} \mathbf{0} \iff \operatorname{tr}(\mathbf{A}) \geq 0 \text{ and } \det(\mathbf{A}) \geq 0\,.$$

18.2. Consider symmetric $\mathbf{A} = \left(\begin{smallmatrix} a & b \\ b & c \end{smallmatrix}\right)$ with eigenvalues $\lambda_1 \geq \lambda_2$. Show that $2|b| \leq \lambda_1 - \lambda_2$, and that if $\mathbf{A}$ is positive definite, then

$$\frac{|b|}{\sqrt{ac}} \leq \frac{\lambda_1 - \lambda_2}{\lambda_1 + \lambda_2}\,, \quad \frac{|b|}{a} \leq \frac{\lambda_1 - \lambda_2}{2\sqrt{\lambda_1 \lambda_2}}\,, \quad \frac{|b|}{\sqrt{c}} \leq \sqrt{\lambda_1} - \sqrt{\lambda_2}\,.$$

Zhang (2005a, p. 86).

18.3. Denote $\mathbf{u} = \mathbf{z} - \boldsymbol{\mu}$ as in (18.65) (p. 366). Show, using (a) (18.67) (p. 366), (b) (18.68), that the solution to

$$\max \|\mathbf{z} - \boldsymbol{\mu}\|^2 \quad \text{subject to } (\mathbf{z} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{z} - \boldsymbol{\mu}) = c^2$$

is $\mathbf{u}_1 = \mathbf{z}_1 - \boldsymbol{\mu} = \pm c\sqrt{\lambda_1}\,\mathbf{t}_1$ and $\mathbf{u}_1'\mathbf{u}_1 = c^2\lambda_1$.

18.4. Let $\mathbf{A}$ and $\mathbf{B}$ be symmetric nonnegative definite $n \times n$ matrices. Prove the following:

(a) $\operatorname{tr}(\mathbf{AB}) \leq \operatorname{ch}_1(\mathbf{A}) \cdot \operatorname{tr}(\mathbf{B})$,

(b) $\mathbf{A} \geq \mathbf{B} \implies \operatorname{ch}_i(\mathbf{A}) \geq \operatorname{ch}_i(\mathbf{B}), \quad i = 1, \dots, n.$

Show that in (b) we must have at least one strict inequality if $\mathbf{A} \neq \mathbf{B}$.

Bapat (2000, p. 83), Seber (1984, pp. 528–529).

18.5. Let $\mathbf{P_A}$ and $\mathbf{P_B}$ be orthogonal projectors of order $n \times n$. Confirm the following:

(a) $-1 \leq \operatorname{ch}_i(\mathbf{P_A} - \mathbf{P_B}) \leq 1,\ i = 1, \dots, n$,

(b) $0 \leq \operatorname{ch}_i(\mathbf{P_A}\mathbf{P_B}) \leq 1,\ i = 1, \dots, n$,

(c) $\operatorname{trace}(\mathbf{P_A}\mathbf{P_B}) \leq \operatorname{rank}(\mathbf{P_A}\mathbf{P_B})$.

18.6. Suppose that $\mathscr{C}(\mathbf{A}) \subset \mathscr{C}(\mathbf{C})$ and $\mathscr{C}(\mathbf{B}) \subset \mathscr{C}(\mathbf{D})$ for matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$, each having n rows. Show that $\operatorname{ch}_i(\mathbf{P_A}\mathbf{P_B}) \leq \operatorname{ch}_i(\mathbf{P_C}\mathbf{P_D})$, $i = 1, \dots, n$.

18.7. Confirm (18.84) (p. 369): If $\lambda \neq \mu$ and $\mathbf{A}$ and $\mathbf{B}$ are symmetric, then

$$(\lambda, \mathbf{w}) \text{ and } (\mu, \mathbf{v}) \text{ are eigenpairs for } (\mathbf{A}, \mathbf{B}) \implies \mathbf{w}'\mathbf{B}\mathbf{v} = 0\,.$$

18.8 (Simultaneous diagonalization by an orthogonal matrix). Prove Proposition 18.5 (p. 370): Let $\mathbf{A}$ and $\mathbf{B}$ be symmetric $n \times n$ matrices. Then there exists an orthogonal $\mathbf{Q}$ such that both $\mathbf{Q}'\mathbf{A}\mathbf{Q}$ and $\mathbf{Q}'\mathbf{B}\mathbf{Q}$ are diagonal if and only if $\mathbf{AB} = \mathbf{BA}$.

Rao (1973a, p. 41), Searle (1982, p. 312).

18.9. Prove that the matrix $\mathbf{Q}$ in (18.105) (p. 371) is nonsingular.

18.10. Let $\mathbf{u}$ denote a random vector with covariance matrix $\operatorname{cov}(\mathbf{u}) = \mathbf{I}_n$ and let $\mathbf{K} \in \mathbb{R}^{n\times p}$ and $\mathbf{L} \in \mathbb{R}^{n\times q}$. Confirm the following result: The canonical correlations between the random vectors $\mathbf{K}'\mathbf{Q}_{\mathbf{L}}\mathbf{u}$ and $\mathbf{L}'\mathbf{Q}_{\mathbf{K}}\mathbf{u}$ are all less than 1, and are precisely those canonical correlations between the vectors $\mathbf{K}'\mathbf{u}$ and $\mathbf{L}'\mathbf{u}$ that are not equal to 1.

Styan (1985, Th. 2.5), Jewell & Bloomfield (1983),
Puntanen (1987, p. 38).

18.11 (Continued ...). Apply the previous exercise to the situation when $\mathbf{K}$ and $\mathbf{L}$ be defined as $\mathbf{K} = \mathbf{V}^{1/2}\mathbf{H}$ and $\mathbf{L} = \mathbf{V}^{1/2}\mathbf{M}$.

18.12 (Continued ...). Let $\mathbf{F}$ be a matrix with the property $\mathscr{C}(\mathbf{F}) = \mathscr{C}(\mathbf{K}) \cap \mathscr{C}(\mathbf{L})$. Show that the nonzero eigenvalues of $\mathbf{P}_{\mathbf{K}}\mathbf{P}_{\mathbf{L}} - \mathbf{P}_{\mathbf{F}}$ are all less than 1, and are precisely the t canonical correlations between the vectors $\mathbf{K}'\mathbf{u}$ and $\mathbf{L}'\mathbf{u}$ that are not equal to 1.
Hint: According to Exercise 8.19 (p. 190): $\mathbf{P}_{\mathbf{K}}\mathbf{P}_{\mathbf{L}} - \mathbf{P}_{\mathbf{F}} = \mathbf{P}_{\mathbf{Q_F K}}\mathbf{P}_{\mathbf{Q_F L}}$.

18.13. Consider the random vector $\mathbf{z} = \binom{\mathbf{x}}{\mathbf{y}}$, with $\operatorname{cov}(\mathbf{z}) = \mathbf{\Sigma}$ where $\mathbf{\Sigma}$ is positive definite and denote $\mathbf{z}^* = \binom{\mathbf{x}^*}{\mathbf{y}^*}$, with $\operatorname{cov}(\mathbf{z}^*) = \mathbf{\Sigma}^{-1}$. Confirm that the canonical correlations between $\mathbf{x}$ and $\mathbf{y}$ are precisely the same as those between $\mathbf{x}^*$ and $\mathbf{y}^*$. What about, in the singular case, replacing $\mathbf{\Sigma}^{-1}$ with an appropriate generalized inverse?

Jewell & Bloomfield (1983), Latour, Puntanen & Styan (1987),
Baksalary, Puntanen & Yanai (1992).

18.14. As in (18.59) (p. 365), let $\mathbf{t}_1 = \binom{t_{11}}{t_{21}}$ be the eigenvector of $\mathbf{\Sigma}$ with respect to λ_1. Suppose that $\sigma_{xy} \neq 0$. Show that the ratio of the second and first element of the eigenvector $\mathbf{t}_1$ is

$$\frac{t_{21}}{t_{11}} = \frac{\lambda_1 - \sigma_x^2}{\sigma_{xy}} = \frac{\lambda_1 - \sigma_x^2}{\sigma_x\sigma_y\varrho_{xy}} = \frac{\sigma_{xy}}{\sigma_x^2 - \lambda_2}\,.$$

18.15. Show, using (10.135) (p. 237) that the first antieigenvalue of the 2×2 covariance matrix $\mathbf{\Sigma}$ is

$$\tau_1 = \cos(\mathbf{\Sigma}) = \frac{2\sqrt{\lambda_1\lambda_2}}{\lambda_1 + \lambda_2} = \frac{2\sigma_x\sigma_y\sqrt{1 - \varrho_{xy}^2}}{\sigma_x^2 + \sigma_y^2}\,.$$

Find also the corresponding antieigenvectors.

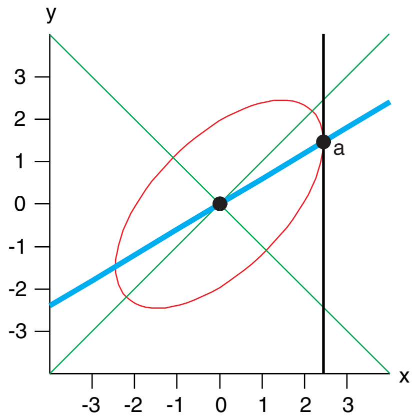


Figure 18.1 (See Exercise 18.16.) A 95% confidence region for the observations from $N_2(\mathbf{0}, \boldsymbol{\Sigma})$.

18.16. The points inside the ellipse

$$\mathcal{A} = \left\{ \mathbf{z} \in \mathbb{R}^2 \colon (\mathbf{z} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{z} - \boldsymbol{\mu}) = \chi^2_{\alpha,\,2} \right\}$$

form a $100(1-\alpha)\%$ confidence region for the observations from $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$; see Figure 18.1. Assume that we have 10 000 observations from $\mathbf{z}$ and that $\boldsymbol{\mu} = \mathbf{0}$, $\boldsymbol{\Sigma} = \left(\begin{smallmatrix} 1 & \varrho \\ \varrho & 1 \end{smallmatrix}\right)$.

(a) What is your guess for the regression line when y is explained by x (and the constant).

(b) Find the vector $\mathbf{a} = \left(\begin{smallmatrix} a_1 \\ a_2 \end{smallmatrix}\right)$ in Figure 18.1 when $\alpha = 0.05$.
$[a_1^2 = \sigma^2 \chi^2_{0.05,2} = \chi^2_{0.05,2}]$

(c) What is the cosine of the angle between the regression line and the major axis.

(d) Why is $a_1 > 1.96$?

18.17 (Eigenvalues of a completely symmetric matrix). Consider a completely symmetric matrix $\mathbf{A}_{p \times p} = (a-b)\mathbf{I}_p + b\mathbf{1}_p\mathbf{1}_p'$, where $a, b \in \mathbb{R}$; see Proposition 18.2, page 366. Show that

$$\mathbf{L}^{-1}\mathbf{A}\mathbf{L} = \begin{pmatrix} a-b & 0 & 0 & \dots & 0 & b \\ 0 & a-b & 0 & \dots & 0 & b \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \dots & a-b & b \\ 0 & 0 & 0 & \dots & 0 & a+(p-1)b \end{pmatrix} := \mathbf{F}\,,$$

where $\mathbf{L}$ carries out elementary column operations:

$$\mathbf{L} = \begin{pmatrix} \mathbf{I}_{p-1} & \mathbf{0}_{p-1} \\ -\mathbf{1}'_{p-1} & 1 \end{pmatrix}, \quad \mathbf{L}^{-1} = \begin{pmatrix} \mathbf{I}_{p-1} & \mathbf{0}_{p-1} \\ \mathbf{1}'_{p-1} & 1 \end{pmatrix},$$

and hence $\det(\mathbf{A}) = (a-b)^{p-1}[a+(p-1)b]$. Confirm that

$$\det(\mathbf{A} - \lambda \mathbf{I}_p) = 0 \iff \det(\mathbf{I}_p - \lambda \mathbf{F}) = 0 .$$

18.18. Prove the rest (in addition to the above Exercise 18.17) of Proposition 18.2 (p. 366) concerning the eigenvalues of the intraclass correlation structure.

18.19 (Canonical correlations in the intraclass correlation structure). As in (18.74) (p. 367), consider

$$\operatorname{cov}\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{\Sigma} & \varrho \mathbf{1}_p \mathbf{1}'_p \\ \varrho \mathbf{1}_p \mathbf{1}'_p & \mathbf{\Sigma} \end{pmatrix},$$

where $\mathbf{\Sigma}_{p \times p} = (1-\varrho)\mathbf{I}_p + \varrho \mathbf{1}_p \mathbf{1}'_p$. Show that there is only one nonzero canonical correlation between $\mathbf{x}$ and $\mathbf{y}$ and that it is

$$\mathrm{cc}_1(\mathbf{x}, \mathbf{y}) = \frac{p\varrho}{1+(p-1)\varrho} .$$

18.20 (Courant–Fischer minimax theorem). Let $\mathbf{A}_{n \times n}$ be symmetric with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$ and corresponding orthonormal eigenvectors $\mathbf{t}_1, \dots, \mathbf{t}_n$. Furthermore, let k be a given integer with $2 \leq k \leq n$ and let $\mathbf{B}$ denote an $n \times (k-1)$ matrix. Show that then

$$\text{(i)} \ \min_{\mathbf{B}} \max_{\mathbf{B}'\mathbf{x}=\mathbf{0}} \frac{\mathbf{x}'\mathbf{A}\mathbf{x}}{\mathbf{x}'\mathbf{x}} = \lambda_k \quad \text{and} \quad \text{(ii)} \ \max_{\mathbf{B}} \min_{\mathbf{B}'\mathbf{x}=\mathbf{0}} \frac{\mathbf{x}'\mathbf{A}\mathbf{x}}{\mathbf{x}'\mathbf{x}} = \lambda_{n-k+1} .$$

The result (i) is obtained when $\mathbf{B} = \mathbf{T}_{(k-1)} = (\mathbf{t}_1 : \cdots : \mathbf{t}_{k-1})$ and $\mathbf{x} = \mathbf{t}_k$, where $\mathbf{t}_i$ are the orthonormal eigenvectors corresponding to λ_i.

Abadir & Magnus (2005, p. 346), Bapat (2000, Th. 2.4), Seber (1984, pp. 525–526).

18.21. The *spectral radius* of $\mathbf{A}_{n \times n}$ is defined as the maximum of the absolute values of the eigenvalues of $\mathbf{A}$; often denoted in literature as $\rho(\mathbf{A})$. Suppose that each element of $\mathbf{A}$ is nonnegative and denote $\mathbf{c} = \mathbf{A}'\mathbf{1}_n$, $\mathbf{r} = \mathbf{A}\mathbf{1}_n$. Then it can be shown, see e.g. Rao & Rao (1998, p. 471), that

$$\min_i c_i \leq \rho(\mathbf{A}) \leq \max_i c_i , \quad \text{and} \quad \min_i r_i \leq \rho(\mathbf{A}) \leq \max_i r_i . \tag{18.200}$$

If each element of $\mathbf{B}$ is nonnegative and $\mathbf{B}\mathbf{1}_n = \mathbf{1}_n$, then $\mathbf{B}$ is said to be *(row) stochastic* and if also $\mathbf{B}'\mathbf{1}_n = \mathbf{1}_n$, then $\mathbf{B}$ is doubly stochastic. Use (18.200) to prove that for a row stochastic matrix we have $\rho(\mathbf{B}) = 1$.

18.22. Consider the $n \times n$ real symmetric matrix $\mathbf{A}_{pq}$, where all the elements in the first p and last p rows and columns are equal to $a \neq 0$ and all the elements in the $q \times q$ "centre" are equal to b, with $q = n - 2p$. For example, with $n = 5$, $p = 1$ and $q = 3$,

$$\mathbf{A}_{13} = \begin{pmatrix} a & a & a & a & a \\ a & b & b & b & a \\ a & b & b & b & a \\ a & b & b & b & a \\ a & a & a & a & a \end{pmatrix}.$$

Find, with $p > 1$ and $q > 1$: (1) $\operatorname{rank}(\mathbf{A}_{pq})$, (2) the eigenvalues of $\mathbf{A}_{pq}$, (3) a simple necessary and sufficient condition for $\mathbf{A}_{pq}$ to be nonnegative definite, (4) $\mathbf{A}_{pq}^{+}$, and (5) show that the orthogonal projector $\mathbf{A}_{pq}\mathbf{A}_{pq}^{+}$ does not depend on a or b.

Chu, Puntanen & Styan (2010).

18.23. Consider the $n \times n$ matrix $\mathbf{R} = \mathbf{I}_n - 2\mathbf{h}\mathbf{h}'$, where the $n \times 1$ vector $\mathbf{h}$ satisfies $\mathbf{h}'\mathbf{h} = 1$. Show that

(a) $\mathbf{R}$ is orthogonal, i.e., $\mathbf{R}\mathbf{R}' = \mathbf{I}_n$.
(b) $\mathbf{R}$ is symmetric and hence $\mathbf{R}^2 = \mathbf{I}_n$.
(c) $\mathbf{R}$ is tripotent, i.e., $\mathbf{R}^3 = \mathbf{R}$.
(d) Find the eigenvalues of $\mathbf{R}$.
(e) Show that the determinant $\det \mathbf{R} = -1$.
(f) Show that $\mathbf{h}$ is an eigenvector of $\mathbf{R}$ and identify the associated eigenvalue.

The matrix $\mathbf{R}$ here is known as a *Householder transformation* and is used in computing least-squares estimates, see, e.g., Golub & Styan (1973), Golub & Van Loan (1996, p. 209). A matrix $\mathbf{A}$ such that $\mathbf{A}^2 = \mathbf{I}_n$ is said to be *involutory*. For some useful rank equalities for involutory matrices see, e.g., Tian & Styan (2001). All involutory matrices are *tripotent:* for connections between tripotent matrices and Cochran's theorem, see, e.g., Anderson & Styan (1982).

18.24. Denote $\hat{\boldsymbol{\alpha}} = \mathbf{R}_{\mathbf{xx}}^{-1}\mathbf{r}_{\mathbf{x}y}$, where $\mathbf{R}_{\mathbf{xx}}$ is the sample correlation matrix of the variables $x_1, \dots, x_k$ and $\mathbf{r}_{\mathbf{x}y}$ is the vector of correlations between x_i and y. Given the correlation matrix $\mathbf{R}_{\mathbf{xx}}$, find the conditions for the cosine between the vectors $\hat{\boldsymbol{\alpha}}$ and $\mathbf{r}_{\mathbf{x}y}$ to be (a) ±1, (b) minimum.

Waller & Jones (2010).

Chapter 19
Singular Value Decomposition

> *The Mayor of Finland: A character who really shouldn't be in this musical.*
>
> Monty Python's Spamalot

While the eigenvalue decomposition $\mathbf{A} = \mathbf{T\Lambda T}'$, say, concerns only symmetric matrices, the singular value decomposition (SVD) $\mathbf{A} = \mathbf{U\Delta V}'$, say, concerns any $n \times m$ matrix. In this chapter we illustrate the usefulness of the SVD, particularly from the statistical point of view. Surprisingly many statistical methods have connections to the SVD.

Theorem 19 (Singular value decomposition). *Let* $\mathbf{A}$ *be an* $n \times m$ $(m \le n)$ *matrix with rank* $r > 0$. *Then* $\mathbf{A}$ *can be written as a product*

$$\begin{aligned}\mathbf{A} &= (\mathbf{U}_1 : \mathbf{U}_0)\begin{pmatrix}\mathbf{\Delta}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{V}_1'\\ \mathbf{V}_0'\end{pmatrix}\\ &= \mathbf{U\Delta V}' = \mathbf{U}_1\mathbf{\Delta}_1\mathbf{V}_1' = \mathbf{U}_*\mathbf{\Delta}_*\mathbf{V}'\\ &= \delta_1\mathbf{u}_1\mathbf{v}_1' + \dots + \delta_r\mathbf{u}_r\mathbf{v}_r' ,\end{aligned} \tag{19.1}$$

where

$$\begin{aligned}\mathbf{\Delta}_1 &= \operatorname{diag}(\delta_1, \dots, \delta_r) , \quad \delta_1 \ge \dots \ge \delta_r > 0 , \quad \mathbf{\Delta} \in \mathbb{R}^{n\times m},\\ \mathbf{\Delta} &= \begin{pmatrix}\mathbf{\Delta}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix} = \begin{pmatrix}\mathbf{\Delta}_*\\ \mathbf{0}\end{pmatrix} \in \mathbb{R}^{n\times m}, \quad \mathbf{\Delta}_1 \in \mathbb{R}^{r\times r}, \quad \mathbf{\Delta}_* \in \mathbb{R}^{m\times m},\\ \delta_{r+1} &= \delta_{r+2} = \dots = \delta_m = 0 ,\\ \delta_i &= \operatorname{sg}_i(\mathbf{A}) = \sqrt{\operatorname{ch}_i(\mathbf{A}'\mathbf{A})}\\ &= i\textit{th singular value of } \mathbf{A} , \quad i = 1, \dots, m ,\\ \mathbf{\Delta}_* &= \operatorname{diag}(\delta_1, \dots, \delta_r, \delta_{r+1}, \dots, \delta_m) = \textit{the first } m \textit{ rows of } \mathbf{\Delta} ,\\ \mathbf{U}_{n\times n} &= (\mathbf{U}_1 : \mathbf{U}_0) , \quad \mathbf{U}_1 \in \mathbb{R}^{n\times r}, \quad \mathbf{U}'\mathbf{U} = \mathbf{U}\mathbf{U}' = \mathbf{I}_n ,\\ \mathbf{V}_{m\times m} &= (\mathbf{V}_1 : \mathbf{V}_0) , \quad \mathbf{V}_1 \in \mathbb{R}^{m\times r}, \quad \mathbf{V}'\mathbf{V} = \mathbf{V}\mathbf{V}' = \mathbf{I}_m ,\\ \mathbf{U}_* &= (\mathbf{u}_1 : \ldots : \mathbf{u}_m) = \textit{the first } m \textit{ columns of } \mathbf{U} , \quad \mathbf{U}_* \in \mathbb{R}^{n\times m},\\ \mathbf{V}'\mathbf{A}'\mathbf{A}\mathbf{V} &= \mathbf{\Delta}'\mathbf{\Delta} = \mathbf{\Delta}_*^2 = \begin{pmatrix}\mathbf{\Delta}_1^2 & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix} \in \mathbb{R}^{m\times m},\end{aligned}$$

S. Puntanen et al., *Matrix Tricks for Linear Statistical Models: Our Personal Top Twenty*,
DOI 10.1007/978-3-642-10473-2_20, © Springer-Verlag Berlin Heidelberg 2011

$$\mathbf{U}'\mathbf{A}\mathbf{A}'\mathbf{U} = \mathbf{\Delta}\mathbf{\Delta}' = \mathbf{\Delta}_{\#}^2 = \begin{pmatrix} \mathbf{\Delta}_1^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \in \mathbb{R}^{n\times n},$$

$$\mathbf{u}_i = i\textit{th left singular vector of } \mathbf{A};\ i\textit{th eigenvector of } \mathbf{A}\mathbf{A}'.$$

$$\mathbf{v}_i = i\textit{th right singular vector of } \mathbf{A};\ i\textit{th eigenvector of } \mathbf{A}'\mathbf{A}\,.$$

Above we have, for simplicity, assumed that $m \leq n$. A more general approach would be to denote $p = \min(n, m)$ and express $\mathbf{A}$ as

$$\mathbf{A} = \mathbf{U}\begin{pmatrix} \mathbf{\Delta}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{V}' = \mathbf{U}\begin{pmatrix} \mathbf{\Delta}_\star & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{V}', \tag{19.2}$$

where $\mathbf{\Delta}_\star = \operatorname{diag}(\delta_1, \dots, \delta_r, \delta_{r+1}, \dots, \delta_p)$, and so

$$\mathbf{A}_{n\times m} \text{ has } r \text{ nonzero and } p - r \text{ zero singular values.} \tag{19.3}$$

However, in what follows, we assume that $m \leq n$. Notation $\operatorname{sg}(\mathbf{A})$ will stand for set of the singular values of $\mathbf{A}_{n\times m}$ ($m \leq n$):

$$\operatorname{sg}(\mathbf{A}) = \{\operatorname{sg}_1(\mathbf{A}), \dots, \operatorname{sg}_m(\mathbf{A})\} \tag{19.4}$$

while

$$\operatorname{nzsg}(\mathbf{A}) = \{\operatorname{sg}_1(\mathbf{A}), \dots, \operatorname{sg}_r(\mathbf{A})\} \tag{19.5}$$

denotes the set of the nonzero singular values.

Mustonen (1995, pp. 15–16) has a useful interpretation for the multivariate observations through the singular value decomposition. He considers the multinormal observations from $\mathrm{N}_n(\mathbf{0}, \mathbf{\Sigma})$ and emphasizes that they can be thought to be generated from $(0, 1)$-normal independent variables in three phases. This comes from the equality $\mathbf{y} = \mathbf{A}\mathbf{x} + \boldsymbol{\mu}$, where $\mathbf{x} \sim \mathrm{N}_n(\mathbf{0}, \mathbf{I}_n)$ and $\operatorname{cov}(\mathbf{y}) = \mathbf{A}\mathbf{A}' := \mathbf{\Sigma}$. Taking an SVD $\mathbf{A} = \mathbf{U}\mathbf{\Delta}\mathbf{V}'$, we see that multiplication $\mathbf{V}'\mathbf{x}$ keeps the variables uncorrelated and hence $\mathbf{y}$ could be defined as $\mathbf{y} = \mathbf{U}\mathbf{\Delta}\mathbf{z} + \boldsymbol{\mu}$, where $\mathbf{z} \sim \mathrm{N}_n(\mathbf{0}, \mathbf{I}_n)$. Now

(a) multiplication by $\mathbf{\Delta}$ means shrinking or expanding,

(b) multiplication by $\mathbf{U}$ means rotation, and

(c) adding $\boldsymbol{\mu}$ means moving the center of distribution out of the origin.

The importance of the singular value decomposition is reflected nicely, for example, in the Preface (to the first edition) of Golub & Van Loan (1996, pp. xiii–xiv):

> "Indeed, perhaps the most recurring theme in the book is the practical and theoretical value of this [SVD] decomposition. Its algorithmic and mathematical properties have a key role to play in nearly every chapter. In many respects *Matrix Computions* is an embellishment of the manuscript 'Everything You Wanted to Know About the Singular Value Decomposition (But Were Afraid to Ask)' authored by our colleague Alan Cline."

Horn & Olkin (1996, p. 470) comment on the importance of the singular value decomposition as follows:

> "The following singular value decomposition is arguably the most basic and useful factorization known for all real or complex matrices."

Interesting comments on the singular value decomposition appear also in Moler (2006).

Photograph 19.1 Gene H. Golub (Uppsala, 2006).

The proof below follows that of Searle (1982, p. 316); for other proofs, see, e.g., Stewart (1998, p. 62), Golub & Van Loan (1996, §2.5.3), and Horn & Olkin (1996, Th. 2.1). The history of the singular value decomposition has been documented in Horn & Johnson (1990, §3.0), and Stewart (1993); see also Horn & Olkin (1996, §2). For the uses of the SVD of the model matrix $\mathbf{X}$ in a regression analysis, see Mandel (1982) and Nelder (1985), and as a tool for solving estimability problems, see, e.g., Eubank & Webster (1985).

Proof of the SVD. Consider the eigenvalue decomposition of the symmetric nonnegative definite $m \times m$ matrix $\mathbf{A}'\mathbf{A}$:

$$\mathbf{A}'\mathbf{A} = \mathbf{V}\mathbf{\Delta}_*^2\mathbf{V}' = (\mathbf{V}_1 : \mathbf{V}_0)\mathbf{\Delta}_*^2(\mathbf{V}_1 : \mathbf{V}_0)', \tag{19.6}$$

where $\mathbf{\Delta}_*^2$ is an $m \times m$ diagonal matrix whose diagonal elements are the eigenvalues of $\mathbf{A}'\mathbf{A}$:

$$\mathbf{\Delta}_*^2 = \begin{pmatrix} \mathbf{\Delta}_1^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \quad \mathbf{\Delta}_1^2 \in \mathbb{R}^{r\times r}, \quad r = \operatorname{rank}(\mathbf{A}), \tag{19.7a}$$

$$\mathbf{\Delta}_1^2 = \operatorname{diag}(\delta_1^2, \dots, \delta_r^2), \quad \delta_1^2 \geq \dots \geq \delta_r^2 > 0 = \delta_{r+1}^2 = \dots = \delta_m^2, \tag{19.7b}$$

$$\delta_i^2 = \operatorname{ch}_i(\mathbf{A}'\mathbf{A}), \quad i = 1, \dots, m, \tag{19.7c}$$

$$\mathbf{V} = (\mathbf{V}_1 : \mathbf{V}_0), \quad \mathbf{V}_1 = (\mathbf{v}_1 : \dots : \mathbf{v}_r) \in \mathbb{R}^{m\times r}. \tag{19.7d}$$

Note that the matrices $\mathbf{A}'\mathbf{A}$ and $\mathbf{A}\mathbf{A}'$ have the same nonzero eigenvalues:

$$\{\delta_1^2, \dots, \delta_r^2\} = \operatorname{nzch}(\mathbf{A}'\mathbf{A}) = \operatorname{nzch}(\mathbf{A}\mathbf{A}'). \tag{19.8}$$

The columns of the matrix $\mathbf{V}_1 \in \mathbb{R}^{m\times r}$ are the orthonormal eigenvectors corresponding to the nonzero eigenvalues δ_i^2 of $\mathbf{A}'\mathbf{A}$, and the columns of $\mathbf{V}_0 \in \mathbb{R}^{m\times(m-r)}$ are the orthonormal eigenvectors corresponding to the zero eigenvalues of $\mathbf{A}'\mathbf{A}$. Furthermore we have

$$\mathbf{V}\mathbf{V}' = \mathbf{V}'\mathbf{V} = \mathbf{I}_m. \tag{19.9}$$

Now (19.6) can be written as

$$\mathbf{A}'\mathbf{A} = \mathbf{V}\boldsymbol{\Delta}_*^2\mathbf{V}' = (\mathbf{V}_1 : \mathbf{V}_0)\begin{pmatrix}\boldsymbol{\Delta}_1^2 & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{V}_1'\\ \mathbf{V}_0'\end{pmatrix} = \mathbf{V}_1\boldsymbol{\Delta}_1^2\mathbf{V}_1', \tag{19.10}$$

which implies that

$$\boldsymbol{\Delta}_1^2 = \mathbf{V}_1'\mathbf{A}'\mathbf{A}\mathbf{V}_1 . \tag{19.11}$$

Define now the $n \times r$ matrix $\mathbf{U}_1$ as follows:

$$\mathbf{U}_1 = \mathbf{A}\mathbf{V}_1\boldsymbol{\Delta}_1^{-1} \in \mathbb{R}^{n\times r}. \tag{19.12}$$

Then the columns of $\mathbf{U}_1$ are orthonormal since in view of (19.11), we have

$$\mathbf{U}_1'\mathbf{U}_1 = \boldsymbol{\Delta}_1^{-1}\mathbf{V}_1'\mathbf{A}'\mathbf{A}\mathbf{V}_1\boldsymbol{\Delta}_1^{-1} = \boldsymbol{\Delta}_1^{-1}\boldsymbol{\Delta}_1^2\boldsymbol{\Delta}_1^{-1} = \mathbf{I}_r . \tag{19.13}$$

We can now form an $n \times (n-r)$ matrix $\mathbf{U}_0$ so that

$$\mathbf{U} = (\mathbf{U}_1 : \mathbf{U}_0) \in \mathbb{R}^{n\times n} \tag{19.14}$$

is orthogonal. When (19.12) is postmultiplied by $\boldsymbol{\Delta}_1\mathbf{V}_1'$, we obtain

$$\mathbf{U}_1\boldsymbol{\Delta}_1\mathbf{V}_1' = \mathbf{A}\mathbf{V}_1\mathbf{V}_1' = \mathbf{A}, \tag{19.15}$$

where we have used the fact $\mathbf{V}_1\mathbf{V}_1' = \mathbf{P}_{\mathbf{A}'}$. Notice also that

$$\mathbf{A}\mathbf{V}_0 = \mathbf{0}. \tag{19.16}$$

Hence we finally have the singular value decomposition

$$\mathbf{A} = (\mathbf{U}_1 : \mathbf{U}_0)\begin{pmatrix}\boldsymbol{\Delta}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{V}_1'\\ \mathbf{V}_0'\end{pmatrix} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}'. \tag{19.17}$$

□

Note that because the first m rows of $\boldsymbol{\Delta}$ are the rows of $\boldsymbol{\Delta}_*$, i.e.,

$$\boldsymbol{\Delta}_{n\times m} = \begin{pmatrix}\boldsymbol{\Delta}_{*(m\times m)}\\ \mathbf{0}_{(n-m)\times m}\end{pmatrix}, \tag{19.18}$$

the singular value decomposition can also be written as

$$\mathbf{A} = \mathbf{U}_*\boldsymbol{\Delta}_*\mathbf{V}', \tag{19.19}$$

where $\mathbf{U}_* = (\mathbf{u}_1 : \dots : \mathbf{u}_m)$, i.e., $\mathbf{U}_*$ comprises the first m columns of $\mathbf{U}$. The decomposition (19.19) is often called a *thin* SVD while $\mathbf{A} = \mathbf{U}_1\boldsymbol{\Delta}_1\mathbf{V}_1'$ can be called a *reduced* SVD.

As stated above, the columns of

$$\mathbf{U} = (\underbrace{\mathbf{u}_1 : \dots : \mathbf{u}_r}_{\mathbf{U}_1} : \underbrace{\mathbf{u}_{r+1} : \dots : \mathbf{u}_m : \dots : \mathbf{u}_n}_{\mathbf{U}_0}), \tag{19.20a}$$

$$\mathbf{V} = (\underbrace{\mathbf{v}_1 : \ldots : \mathbf{v}_r}_{\mathbf{V}_1} : \underbrace{\mathbf{v}_{r+1} : \ldots : \mathbf{v}_m}_{\mathbf{V}_0}) \tag{19.20b}$$

are the *left* and *right singular vectors* of $\mathbf{A}$, respectively. Moreover, we have

$$\mathbf{A}'\mathbf{A} = \mathbf{V}\boldsymbol{\Delta}_*^2\mathbf{V}', \quad \mathbf{A}\mathbf{A}' = \mathbf{U}\boldsymbol{\Delta}_\#^2\mathbf{U}', \tag{19.21}$$

where

$$\boldsymbol{\Delta}_*^2 = \begin{pmatrix} \boldsymbol{\Delta}_1^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \in \mathbb{R}^{m\times m}, \quad \boldsymbol{\Delta}_\#^2 = \begin{pmatrix} \boldsymbol{\Delta}_1^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \in \mathbb{R}^{n\times n}, \tag{19.22}$$

and so (as in the proof of the SVD) the columns $\mathbf{v}_i$ are eigenvectors of $\mathbf{A}'\mathbf{A}$ and $\mathbf{u}_i$ are eigenvectors of $\mathbf{A}\mathbf{A}'$. We can conclude (please confirm) that the singular vectors satisfy

$$\mathscr{C}(\mathbf{V}_1) = \mathscr{C}(\mathbf{A}) \quad \mathscr{C}(\mathbf{U}_1) = \mathscr{C}(\mathbf{A}')\,, \tag{19.23a}$$

$$\mathbf{A}\mathbf{v}_i = \delta_i\mathbf{u}_i\,, \quad \mathbf{A}'\mathbf{u}_i = \delta_i\mathbf{v}_i\,, \quad i = 1,\ldots,m\,, \tag{19.23b}$$

$$\mathbf{A}'\mathbf{u}_i = \mathbf{0}\,, \quad i = m+1,\ldots,n\,, \tag{19.23c}$$

$$\mathbf{u}_i'\mathbf{A}\mathbf{v}_i = \delta_i\,,\ i = 1,\ldots,m\,, \quad \mathbf{u}_i'\mathbf{A}\mathbf{v}_j = 0 \text{ for } i \neq j\,. \tag{19.23d}$$

As stated by Stewart (1998, pp. 65–66; 2001, p. 205), the SVD is "essentially unique" and hence we usually speak about *the* singular value decomposition; the repeated singular values are a source for nonuniqueness. The singular values themselves are unique. Citing Stewart (2001, p. 205): "The right singular vectors corresponding to a multiple singular value are not unique, but the space they span is. Once the right singular vectors have been specified, the left singular vectors corresponding to nonzero singular values are uniquely determined by the relation $\mathbf{A}\mathbf{v}_i = \delta_i\mathbf{u}_i$." Moreover, Stewart points out that when $n > m$, the matrix $\mathbf{A}\mathbf{A}'$ has $n - m$ additional zero eigenvalues; Stewart (2001, p. 205) calls them "honorary singular values".

19.1 Maximal Value of $\mathbf{x}'\mathbf{A}\mathbf{y}$

Let $\mathbf{A}$ be an $n \times m$ ($m \leq n$) matrix with rank $r > 0$, and let $\mathbf{A}$ have the singular value decomposition

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}' = \mathbf{U}_1\boldsymbol{\Delta}_1\mathbf{V}_1' = \delta_1\mathbf{u}_1\mathbf{v}_1' + \cdots + \delta_r\mathbf{u}_r\mathbf{v}_r'\,. \tag{19.24}$$

We show now that

$$\max_{\mathbf{x}'\mathbf{x}=\mathbf{y}'\mathbf{y}=1} \mathbf{x}'\mathbf{A}\mathbf{y} = \delta_1\,, \tag{19.25}$$

and that the maximum is attained when

$$\mathbf{x} = \mathbf{u}_1\,, \qquad \mathbf{y} = \mathbf{v}_1\,. \tag{19.26}$$

Let $\mathbf{x}$ and $\mathbf{y}$ be arbitrary vectors of $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively, satisfying $\mathbf{x}'\mathbf{x} = \mathbf{y}'\mathbf{y} = 1$. Then applying the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned}(\mathbf{x}'\mathbf{A}\mathbf{y})^2 &= (\mathbf{x}'\mathbf{U}\boldsymbol{\Delta}\mathbf{V}'\mathbf{y})^2 = (\mathbf{x}'\mathbf{U}\cdot\boldsymbol{\Delta}\mathbf{V}'\mathbf{y})^2 \\ &\leq \mathbf{x}'\mathbf{U}\mathbf{U}'\mathbf{x}\cdot\mathbf{y}'\mathbf{V}\boldsymbol{\Delta}'\boldsymbol{\Delta}\mathbf{V}'\mathbf{y} = \mathbf{y}'\mathbf{V}\boldsymbol{\Delta}_*^2\mathbf{V}'\mathbf{y} := \mathbf{t}'\boldsymbol{\Delta}_*^2\mathbf{t} \\ &= \delta_1^2t_1^2 + \delta_2^2t_2^2 + \cdots + \delta_r^2t_r^2 \\ &\leq \delta_1^2(t_1^2 + t_2^2 + \cdots + t_r^2) \\ &\leq \delta_1^2(t_1^2 + t_2^2 + \cdots + t_r^2 + \cdots + t_m^2) \\ &= \delta_1^2\mathbf{y}'\mathbf{V}\mathbf{V}'\mathbf{y} \\ &= \delta_1^2\,,\end{aligned} \tag{19.27}$$

where we have denoted $\mathbf{t} = \mathbf{V}'\mathbf{y} \in \mathbb{R}^m$. The maximum is clearly attained when $\mathbf{x}$ and $\mathbf{y}$ are chosen as in (19.26).

Note that we can of course write the following:

$$\operatorname{sg}_1(\mathbf{A}) = \delta_1 = \max_{\mathbf{x}\neq\mathbf{0},\,\mathbf{y}\neq\mathbf{0}} \frac{\mathbf{x}'\mathbf{A}\mathbf{y}}{\sqrt{\mathbf{x}'\mathbf{x}\cdot\mathbf{y}'\mathbf{y}}}\,. \tag{19.28}$$

Furthermore, the second largest singular value δ_2 can be obtained as follows:

$$\operatorname{sg}_2(\mathbf{A}) = \delta_2 = \max \mathbf{x}'\mathbf{A}\mathbf{y}\,, \tag{19.29}$$

where the maximum is taken over the set

$$\{\,\mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^m : \mathbf{x}'\mathbf{x} = \mathbf{y}'\mathbf{y} = 1,\ \mathbf{x}'\mathbf{u}_1 = \mathbf{y}'\mathbf{v}_1 = 0\,\}\,. \tag{19.30}$$

The ith largest singular value can be defined in the corresponding way so that

$$\max_{\substack{\mathbf{x}\neq\mathbf{0},\,\mathbf{y}\neq\mathbf{0}\\ \mathbf{U}_{(k)}\mathbf{x}=\mathbf{0},\mathbf{V}_{(k)}\mathbf{y}=\mathbf{0}}} \frac{\mathbf{x}'\mathbf{A}\mathbf{y}}{\sqrt{\mathbf{x}'\mathbf{x}\cdot\mathbf{y}'\mathbf{y}}} = \delta_{k+1}\,, \quad k = 1,\dots,r-1\,, \tag{19.31}$$

where $\mathbf{U}_{(k)} = (\mathbf{u}_1 : \ldots : \mathbf{u}_k)$ and $\mathbf{V}_{(k)} = (\mathbf{v}_1 : \ldots : \mathbf{v}_k)$; the maximum occurs when $\mathbf{x} = \mathbf{u}_{k+1}$ and $\mathbf{y} = \mathbf{v}_{k+1}$.

19.2 When is $\mathbf{A}'\mathbf{A} = \mathbf{B}'\mathbf{B}$?

In linear models, and in particular, in multivariate analysis, we may meet the following problem: Let $\mathbf{A}$ and $\mathbf{B}$ be given $n \times m$ matrices. What is then the necessary and sufficient condition for

$$\mathbf{A}'\mathbf{A} = \mathbf{B}'\mathbf{B}\,. \tag{19.32}$$

The answer is:

$$\mathbf{A} = \mathbf{QB} \quad \text{for some orthogonal } \mathbf{Q}\,. \tag{19.33}$$

The sufficiency of (19.33) is obvious. To prove that (19.32) implies (19.33), we note that, in view of (19.32), there exists an orthogonal matrix $\mathbf{V}$ such that

$$\mathbf{A}'\mathbf{A} = \mathbf{B}'\mathbf{B} = \mathbf{V}\begin{pmatrix} \boldsymbol{\Delta}_1^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{V}' \in \mathbb{R}^{m\times m}, \tag{19.34}$$

where $\boldsymbol{\Delta}_1$ is an $r \times r$ diagonal matrix; $\operatorname{rank}(\mathbf{A}) = \operatorname{rank}(\mathbf{B}) = r$. From the proof of the singular value decomposition, we recall that it is possible to construct orthogonal $n \times n$ matrices $\mathbf{F}$ and $\mathbf{G}$ such that

$$\mathbf{A} = \mathbf{F}\begin{pmatrix} \boldsymbol{\Delta}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{V}', \qquad \mathbf{B} = \mathbf{G}\begin{pmatrix} \boldsymbol{\Delta}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{V}'. \tag{19.35}$$

Choosing now

$$\mathbf{A} = \mathbf{FG}' \cdot \mathbf{G}\begin{pmatrix} \boldsymbol{\Delta}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{V}' := \mathbf{QB}\,, \tag{19.36}$$

gives the required result since $\mathbf{QQ}' = \mathbf{FG}'\mathbf{GF}' = \mathbf{I}_n$.

As references concerning (19.32), we may mention Vinograde (1950), Horn & Olkin (1996, Th. 3.1), Neudecker & van de Velden (2000, 2001), and Young & Young (2004).

19.3 Approximation by a Matrix of Lower Rank

We begin with the following result.

Proposition 19.1. *Let* $\mathbf{A}_{n\times n}$ *be a given symmetric matrix of rank* $r > 0$*, whose eigenvalue decomposition is*

$$\mathbf{A} = \mathbf{T}\boldsymbol{\Lambda}\mathbf{T}', \quad \boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \dots, \lambda_n)\,, \tag{19.37}$$

and $\mathbf{T} = (\mathbf{t}_1 : \dots : \mathbf{t}_n)$, $\mathbf{T}'\mathbf{T} = \mathbf{I}_n$. *Let* k *be a given integer,* $k \le n$. *Then*

$$\max_{\mathbf{G}'\mathbf{G}=\mathbf{I}_k} \operatorname{tr}(\mathbf{G}'\mathbf{AG}) = \lambda_1 + \dots + \lambda_k\,, \tag{19.38}$$

where the upper bound is obtained when

$$\mathbf{G} = (\mathbf{t}_1 : \dots : \mathbf{t}_k) := \mathbf{T}_{(k)}\,. \tag{19.39}$$

Proof. To prove (19.38), we first write

$$\operatorname{tr}(\mathbf{G}'\mathbf{AG}) = \operatorname{tr}(\mathbf{G}'\mathbf{T}\boldsymbol{\Lambda}\mathbf{T}'\mathbf{G}) := \operatorname{tr}(\mathbf{Y}'\boldsymbol{\Lambda}\mathbf{Y}) = \operatorname{tr}(\mathbf{YY}'\boldsymbol{\Lambda})\,, \tag{19.40}$$

where $\mathbf{Y} = \mathbf{T}'\mathbf{G}$. Then $\mathbf{Y}'\mathbf{Y} = \mathbf{G}'\mathbf{T}\mathbf{T}'\mathbf{G} = \mathbf{I}_k$, and the matrix

$$\mathbf{Y}\mathbf{Y}' := \mathbf{P} \in \mathbb{R}^{n\times n} \tag{19.41}$$

is an orthogonal projector, with diagonal elements satisfying

$$0 \leq p_{ii} \leq 1, \quad i = 1,\dots,n\,, \quad p_{11} + \cdots + p_{nn} = k = \operatorname{rank}(\mathbf{P})\,. \tag{19.42}$$

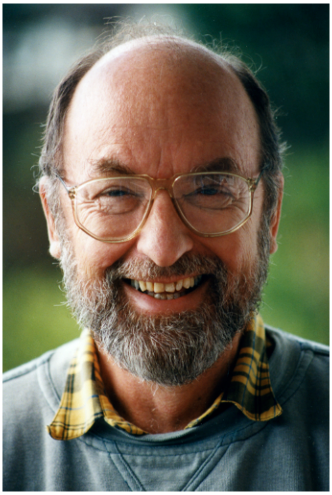
Hence we have

$$\operatorname{tr}(\mathbf{G}'\mathbf{A}\mathbf{G}) = \operatorname{tr}(\mathbf{P}\boldsymbol{\Lambda}) = p_{11}\lambda_1 + \cdots + p_{kk}\lambda_k + \cdots + p_{nn}\lambda_n\,, \tag{19.43}$$

and in view of (19.42),

$$\operatorname{tr}(\mathbf{G}'\mathbf{A}\mathbf{G}) \leq \lambda_1 + \cdots + \lambda_k\,. \tag{19.44}$$

The upper bound in (19.44) is clearly obtained when (19.39) holds. □



Photograph 19.2 Alastair J. Scott (Auckland, 1994).

For a reference of result (19.38), see, e.g., Rao (1973a, p. 63), and Harville (1997, Th. 21.12.5). We may note that (19.38) could be also easily proved using the *Poincaré separation theorem*; see, e.g., Rao (1973a, p. 64), Rao (1980, Th. 2.1), Scott & Styan (1985, p. 213), and Abadir & Magnus (2005, p. 347), and Mäkeläinen (1970a, Th. 4.1, Cor. 4.2.2):

Proposition 19.2. *Let $\mathbf{A}_{n\times n}$ be a symmetric matrix, and let $\mathbf{G}_{n\times k}$ be such that $\mathbf{G}'\mathbf{G} = \mathbf{I}_k$, $k \leq n$. Then, for $i = 1,\dots,k$,*

$$\operatorname{ch}_{n-k+i}(\mathbf{A}) \leq \operatorname{ch}_i(\mathbf{G}'\mathbf{A}\mathbf{G}) \leq \operatorname{ch}_i(\mathbf{A})\,. \tag{19.45}$$

Equality holds on the right of (19.45) simultaneously for all $i = 1,\dots,k$ if and only if there exists $\mathbf{U}_{n\times k}$ such that

$$\mathbf{U}'\mathbf{U} = \mathbf{I}_k\,, \quad \mathbf{A}\mathbf{U} = \mathbf{U}\boldsymbol{\Lambda}_{(k)}\,, \quad \textit{and} \quad \mathscr{C}(\mathbf{U}) = \mathscr{C}(\mathbf{G})\,, \tag{19.46}$$

where $\boldsymbol{\Lambda}_{(k)} \in \mathbb{R}^{k\times k}$ is a diagonal matrix containing the k largest eigenvalues of $\mathbf{A}$. Equality holds on the left of (19.45) if and only if there exists $\mathbf{L}_{n\times k}$ such that $\mathbf{L}'\mathbf{L} = \mathbf{I}_k$, $\mathbf{A}\mathbf{L} = \mathbf{L}\boldsymbol{\Lambda}_{[k]}$ and $\mathscr{C}(\mathbf{L}) = \mathscr{C}(\mathbf{G})$, where $\boldsymbol{\Lambda}_{[k]} \in \mathbb{R}^{k\times k}$ is a diagonal matrix containing the k smallest eigenvalues of $\mathbf{A}$. Equivalently we can state that the upper bound in (19.45) is attained if and only if $\mathbf{G} = \mathbf{U}\mathbf{N}$, where $\mathbf{N}_{k\times k}$ is orthogonal and the columns of $\mathbf{U}$ are any set of orthonormal eigenvectors of $\mathbf{A}$ corresponding to the k largest eigenvalues.

In particular, the upper bound in (19.45) is attained when

$$\mathbf{G} = (\mathbf{t}_1 : \ldots : \mathbf{t}_k) = \mathbf{T}_{(k)}\,, \tag{19.47}$$

and the lower bound when

$$\mathbf{G} = (\mathbf{t}_{n-k+1} : \ldots : \mathbf{t}_n) := \mathbf{T}_{[k]}\,, \tag{19.48}$$

where the columns of $\mathbf{T}$ *are the orthonormal eigenvectors of* $\mathbf{A}$.

The inequalities in (19.45) are being named after Henri Poincaré (1890, pp. 259–260); see also Philatelic Item 11.2, (p. 282). For a related version of (19.45) see Exercise 19.9 (p. 411).

In Propositions 19.3–19.6 we consider the Frobenius norm of the matrix $\mathbf{L}$: $\|\mathbf{L}\|_F^2 = \operatorname{tr}(\mathbf{LL}')$.

Proposition 19.3. *Let* $\mathbf{A}$ *be a given* $n \times m$ *matrix of rank* r*, and let* $\mathbf{B}$ *be an* $n \times m$ *matrix of rank* k $(< r)$*. Then*

$$\begin{aligned}(\mathbf{A}-\mathbf{B})(\mathbf{A}-\mathbf{B})' &\geq_{\mathsf{L}} (\mathbf{A}-\mathbf{AP}_{\mathbf{B}'})(\mathbf{A}-\mathbf{AP}_{\mathbf{B}'})' \\ &= \mathbf{A}(\mathbf{I}_m - \mathbf{P}_{\mathbf{B}'})\mathbf{A}',\end{aligned} \tag{19.49}$$

and hence

$$\|\mathbf{A}-\mathbf{B}\|_F^2 \geq \|\mathbf{A}-\mathbf{AP}_{\mathbf{B}'}\|_F^2\,. \tag{19.50}$$

Moreover, if $\mathbf{B}$ *has the full rank decomposition*

$$\mathbf{B} = \mathbf{FG}', \quad \textit{where} \quad \mathbf{G}'\mathbf{G} = \mathbf{I}_k\,, \tag{19.51}$$

then $\mathbf{P}_{\mathbf{B}'} = \mathbf{P}_{\mathbf{G}}$*, and*

$$\begin{aligned}(\mathbf{A}-\mathbf{B})(\mathbf{A}-\mathbf{B})' &\geq_{\mathsf{L}} (\mathbf{A}-\mathbf{FG}')(\mathbf{A}-\mathbf{FG}')' \\ &= \mathbf{A}(\mathbf{I}_m - \mathbf{GG}')\mathbf{A}'.\end{aligned} \tag{19.52}$$

Proof. The proof is a simple consequence from Proposition 2.5 (p. 81), but let's go through it anyways. We observe that

$$\mathbf{A}-\mathbf{B} = (\mathbf{A}-\mathbf{AP}_{\mathbf{B}'}) + (\mathbf{AP}_{\mathbf{B}'}-\mathbf{B}) := \mathbf{S}+\mathbf{R}\,, \tag{19.53}$$

where

$$\begin{aligned}\mathbf{SR}' &= (\mathbf{A}-\mathbf{AP}_{\mathbf{B}'})(\mathbf{P}_{\mathbf{B}'}\mathbf{A}'-\mathbf{B}') \\ &= \mathbf{AP}_{\mathbf{B}'}\mathbf{A}' - \mathbf{AB}' - \mathbf{AP}_{\mathbf{B}'}\mathbf{A}' + \mathbf{AB}' = \mathbf{0}\,.\end{aligned} \tag{19.54}$$

Hence

$$\begin{aligned}(\mathbf{A}-\mathbf{B})(\mathbf{A}-\mathbf{B})' &= \mathbf{SS}' + \mathbf{RR}' \\ &\geq_{\mathsf{L}} \mathbf{SS}' = \mathbf{A}(\mathbf{I}_m - \mathbf{P}_{\mathbf{B}'})\mathbf{A}',\end{aligned} \tag{19.55}$$

which proves our claim (19.50). Note that (19.52) is simply another way to express (19.49). □

Now we are (finally) ready to state and prove the following fundamental result due to Eckart & Young (1936); see also Householder & Young (1938). This result is often called the *Eckart–Young theorem.*

Proposition 19.4 (Eckart–Young Theorem). *Let* $\mathbf{A}$ *be a given* $n \times m$ *matrix of rank* r*, with the singular value decomposition*

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}' = \mathbf{U}_1\boldsymbol{\Delta}_1\mathbf{V}_1' = \delta_1\mathbf{u}_1\mathbf{v}_1' + \cdots + \delta_r\mathbf{u}_r\mathbf{v}_r' \,. \tag{19.56}$$

Let $\mathbf{B}$ *be an* $n \times m$ *matrix of rank* k $(< r)$*. Then*

$$\min_{\mathbf{B}} \|\mathbf{A} - \mathbf{B}\|_F^2 = \delta_{k+1}^2 + \cdots + \delta_r^2 \,, \tag{19.57}$$

and the minimum is attained taking

$$\mathbf{B} = \dot{\mathbf{B}}_k = \delta_1\mathbf{u}_1\mathbf{v}_1' + \cdots + \delta_k\mathbf{u}_k\mathbf{v}_k' \,. \tag{19.58}$$

Proof. The proof of (19.57) goes now very smoothly using Propositions 19.1 and 19.3. Letting the columns of $\mathbf{G}$ be an orthonormal basis of $\mathscr{C}(\mathbf{B}')$, we observe, in view of (19.52), that

$$\begin{aligned} \|\mathbf{A} - \mathbf{B}\|_F^2 \geq \|\mathbf{A} - \mathbf{A}\mathbf{G}\mathbf{G}'\|_F^2 &= \operatorname{tr}[\mathbf{A}(\mathbf{I}_m - \mathbf{G}\mathbf{G}')\mathbf{A}'] \\ &= \operatorname{tr}(\mathbf{A}\mathbf{A}') - \operatorname{tr}(\mathbf{G}'\mathbf{A}'\mathbf{A}\mathbf{G}) \,. \end{aligned} \tag{19.59}$$

Now, on the basis of (19.38) (p. 397),

$$\operatorname{tr}(\mathbf{G}'\mathbf{A}'\mathbf{A}\mathbf{G}) \leq \operatorname{ch}_1(\mathbf{A}'\mathbf{A}) + \cdots + \operatorname{ch}_k(\mathbf{A}'\mathbf{A}) = \delta_1^2 + \cdots + \delta_k^2 \tag{19.60}$$

for all $\mathbf{G}$ such that $\mathbf{G}'\mathbf{G} = \mathbf{I}_k$. Hence (19.59) implies

$$\|\mathbf{A} - \mathbf{B}\|_F^2 \geq \delta_{k+1}^2 + \cdots + \delta_r^2 \tag{19.61}$$

for all $\mathbf{B}_{n\times m}$ such that $\operatorname{rank}(\mathbf{B}) = k$. Clearly the equality is attained when (19.58) holds, and thus our claim (19.57) is proved. □

It is worth noting that the Löwner ordering (19.49) implies also

$$\operatorname{ch}_i(\mathbf{A} - \mathbf{B})(\mathbf{A} - \mathbf{B})' \geq \operatorname{ch}_i[\mathbf{A}(\mathbf{I}_m - \mathbf{P}_{\mathbf{B}'})\mathbf{A}'] \geq \operatorname{ch}_{i+k}(\mathbf{A}\mathbf{A}') \,, \tag{19.62}$$

for $i + k \leq r$, i.e.,

$$\operatorname{sg}_i(\mathbf{A} - \mathbf{B}) \geq \operatorname{sg}_{i+k}(\mathbf{A}) \,, \tag{19.63}$$

where the last inequality in (19.62) follows from Proposition 19.2 (p. 398); details are left as an exercise, see Exercise 19.8 (p. 411).

Stewart (1993, p. 563) points out that actually Eckart & Young (1936) rediscovered the result which earlier had appeared in a more general form in the paper by Schmidt (1907); see also Hubert, Meulman & Heiser (2000, p. 71). Instead of the "Eckart–Young theorem" Stewart (1993) calls Proposition 19.4 the "Schmidt's approximation theorem" and so do Ben-Israel & Greville (2003, p. 213), who also use the term *best rank-k approximation of* **A**. For further discussion on Proposition 19.4, which we still (slightly incorrectly) continue to call the Eckart–Young theorem, see Mirsky (1960), Chipman (1997), Harville (1997, pp. 556–559), and Stewart & Sun (1990, 208–210).

For a general reference to matrix approximation, see Rao (1979, 1980), and for the related comments on the use of the singular value decomposition in psychometrics, see Takane (2004), and Yanai & Takane (2007). Hubert, Meulman & Heiser (2000) review some connections between numerical linear algebra and applied statistics/psychometrics and their respective concerns with matrix factorization and the subsequent rank reduction of a matrix.

For completeness we write the Eckart–Young theorem also in the situation when **A** is symmetric.

Proposition 19.5. *Let* $\mathbf{A}_{n\times n}$ *be a symmetric matrix of rank* r*, with eigenvalue decomposition*

$$\mathbf{A} = \mathbf{T}\boldsymbol{\Delta}\mathbf{T}' = \lambda_1\mathbf{t}_1\mathbf{t}_1' + \cdots + \lambda_r\mathbf{t}_r\mathbf{t}_r' \,. \tag{19.64}$$

Then

$$\min_{\operatorname{rank}(\mathbf{B})=k} \|\mathbf{A}-\mathbf{B}\|_F^2 = \min_{\operatorname{rank}(\mathbf{B})=k} \operatorname{tr}(\mathbf{A}-\mathbf{B})(\mathbf{A}-\mathbf{B})' = \lambda_{k+1}^2 + \cdots + \lambda_r^2 \,, \tag{19.65}$$

and the minimum is attained when

$$\mathbf{B} = \mathbf{T}_{(k)}\boldsymbol{\Lambda}_{(k)}\mathbf{T}_{(k)}' = \lambda_1\mathbf{t}_1\mathbf{t}_1' + \cdots + \lambda_k\mathbf{t}_k\mathbf{t}_k' \,. \tag{19.66}$$

We complete this example with the following result useful in factor analysis (Schönemann 1966, Green 1952).

Proposition 19.6. *Let* **A** *and* **B** *be given* $n \times p$ *matrices and let* $\mathscr{O}_{p\times p}$ *be the set of orthogonal* $p \times p$ *matrices. Then*

$$\min_{\mathbf{Z}\in\mathscr{O}_{p\times p}} \|\mathbf{A} - \mathbf{B}\mathbf{Z}\|_F^2 = \|\mathbf{A} - \mathbf{B}\mathbf{Z}_*\|_F^2, \tag{19.67}$$

where $\mathbf{Z}_* = \mathbf{U}\mathbf{V}'$ *is obtained from the singular value decomposition* $\mathbf{B}'\mathbf{A} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}'$.

Proof. Following Vehkalahti, Puntanen & Tarkkonen (2007, Lemma 1), we first write

$$\begin{aligned}\|\mathbf{A}-\mathbf{BZ}\|_F^2 &= \operatorname{tr}[(\mathbf{A}-\mathbf{BZ})(\mathbf{A}-\mathbf{BZ})'] \\ &= \operatorname{tr}(\mathbf{AA}') + \operatorname{tr}(\mathbf{BB}') - 2\operatorname{tr}[\mathbf{Z}(\mathbf{B}'\mathbf{A})'] \,, \end{aligned} \tag{19.68}$$

and observe that minimizing (19.68) is equivalent to maximizing $\operatorname{tr}[\mathbf{Z}(\mathbf{B}'\mathbf{A})']$. Using the singular value decomposition $\mathbf{B}'\mathbf{A} = \mathbf{U\Delta V}'$, where $\mathbf{U},\mathbf{V} \in \mathscr{O}_{p\times p}$, we obtain

$$\operatorname{tr}[\mathbf{Z}(\mathbf{B}'\mathbf{A})'] = \operatorname{tr}[\mathbf{Z}(\mathbf{U\Delta V}')'] = \operatorname{tr}(\mathbf{U}'\mathbf{ZV\Delta}) = \operatorname{tr}(\mathbf{Q\Delta}) \,, \tag{19.69}$$

where $\mathbf{Q} = \mathbf{U}'\mathbf{ZV} \in \mathscr{O}_{p\times p}$, being the product of orthogonal matrices. Because the elements of $\mathbf{Q}$ can not exceed 1, we have

$$\operatorname{tr}(\mathbf{Q\Delta}) = \sum_{j=1}^{p} q_{jj}\delta_j \leq \sum_{j=1}^{p} \delta_j = \operatorname{tr}(\mathbf{\Delta}) \,, \tag{19.70}$$

and the maximum of (19.69) and hence the minimum of (19.68) is clearly attained when $\mathbf{Q} = \mathbf{I}_p$. Our claim (19.67) follows by selecting $\mathbf{Z} = \mathbf{Z}_* = \mathbf{UV}'$, as then $\mathbf{Q} = \mathbf{U}'\mathbf{UV}'\mathbf{V} = \mathbf{I}_p$. □

19.4 Principal Component Analysis

In Section 18.2 (p. 362) we considered the (population) principal components and defined the random variable $\mathbf{b}_1'(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}})$ to be the first principal component if the maximum of $\operatorname{var}(\mathbf{b}'\mathbf{x})$ under the condition $\mathbf{b}'\mathbf{b} = 1$ is attained for $\mathbf{b} = \mathbf{b}_1$. The second principal component is defined as $\mathbf{b}_2'(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}})$ where $\mathbf{b}_2$ maximizes $\operatorname{var}(\mathbf{b}'\mathbf{x})$ under the conditions $\mathbf{b}'\mathbf{b} = 1$ and $\operatorname{cor}(\mathbf{b}'\mathbf{x}, \mathbf{b}_1'\mathbf{x}) = 0$.

In Section 9.4 (p. 203), we were interested in predicting the p-dimensional random vector $\mathbf{x}$ on the basis of some linear combinations of $\mathbf{x}$, that is, we want to find

$$\operatorname{BLP}(\mathbf{x};\mathbf{A}'\mathbf{x}) = \text{best linear predictor of } \mathbf{x} \text{ on the basis of } \mathbf{A}'\mathbf{x} \,, \tag{19.71}$$

where $\mathbf{A}' \in \mathbb{R}^{k\times p}$. How to obtain the $\mathbf{A} \in \mathbb{R}^{p\times k}$? The criterion to construct $\mathbf{A}$ can obviously be based on the covariance matrix of the prediction error $\mathbf{x} - \operatorname{BLP}(\mathbf{x};\mathbf{A}'\mathbf{x}) = \mathbf{e}$:

$$\operatorname{cov}(\mathbf{e}) = \mathbf{\Sigma} - \mathbf{\Sigma A}(\mathbf{A}'\mathbf{\Sigma A})^{-1}\mathbf{A}'\mathbf{\Sigma} \,. \tag{19.72}$$

In Proposition 9.3 (p. 205) we concluded that the minimum of the (squared) Frobenius norm $\|\operatorname{cov}(\mathbf{e})\|_F^2 = \operatorname{tr}\left[\operatorname{cov}(\mathbf{e})\right]^2$ is attained when $\mathbf{A}$ chosen as $\mathbf{T}_{(k)}$, where the columns of $\mathbf{T}_{(k)}$ comprise the first k eigenvectors of $\operatorname{cov}(\mathbf{x}) = \mathbf{\Sigma}$. Moreover, minimizing the trace of $\operatorname{cov}(\mathbf{e})$ yields the same result.

The above two different approaches both yield the principal components. In this section we present two more approaches: one based on the observed data points and fitting there a subspace appropriately, while in the other approach we deal with the variable space. These are sample analogies to the predictive approach in population.

Let $\widetilde{\mathbf{X}}$ be an $n \times p$ centered data matrix containing n measurements of p variables. Let us denote

$$\widetilde{\mathbf{X}} = (\tilde{\mathbf{x}}_1 : \tilde{\mathbf{x}}_2 : \ldots : \tilde{\mathbf{x}}_p) = \begin{pmatrix} \tilde{\mathbf{x}}'_{(1)} \\ \tilde{\mathbf{x}}'_{(2)} \\ \vdots \\ \tilde{\mathbf{x}}'_{(n)} \end{pmatrix}. \tag{19.73}$$

Here the columns (corresponding to variables) $\tilde{\mathbf{x}}_i$ are vectors in the variable space $\mathbb{R}^n$, and (the transposes) of the rows (corresponding to observations) $\tilde{\mathbf{x}}'_{(j)}$ are vectors in the observation space $\mathbb{R}^m$.

Consider now the scatter plot of the n centered observations (or points) in the observation space $\mathbb{R}^p$. Let $\mathbf{G}_{m\times k}$, where $k < p$, be a matrix with orthonormal columns. Then $\mathscr{C}(\mathbf{G})$ is an k-dimensional subspace of the observation space $\mathbb{R}^p$, and the orthogonal projector onto that subspace is

$$\mathbf{G}\mathbf{G}' = \mathbf{P}_{\mathbf{G}}\,. \tag{19.74}$$

How should $\mathbf{G}$ be chosen if we wish that

$$\mathscr{C}(\mathbf{G}) \text{ goes "nicely" through the scatter plot?} \tag{19.75}$$

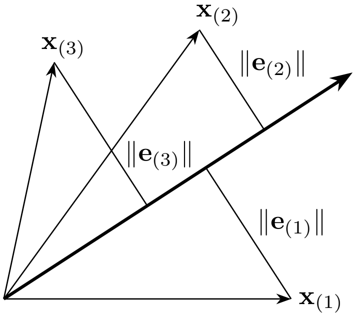


Figure 19.1 Minimizing orthogonal distances.

One natural criterion for "niceness" is based on the orthogonal distances of the data points (observations) $\tilde{\mathbf{x}}_{(i)}$ from the subspace $\mathscr{C}(\mathbf{G})$. We may denote

$$\mathbf{e}_{(i)} = (\mathbf{I}_m - \mathbf{P}_{\mathbf{G}})\tilde{\mathbf{x}}_{(i)}\,, \quad i = 1, \ldots, p\,, \tag{19.76}$$

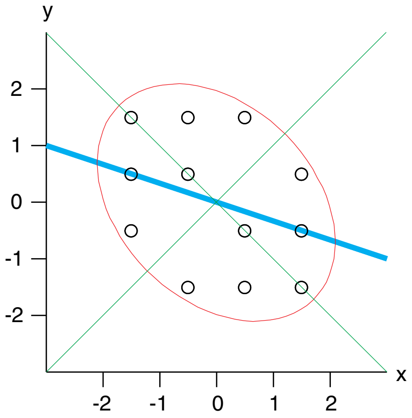


Figure 19.2 The scatter plot of the 12 observations of the centered data matrix appeared in Figures 0.1b (p. 2) and 0.8b (p. 56). Denote the centered data matrix as $\widetilde{\mathbf{X}}$.

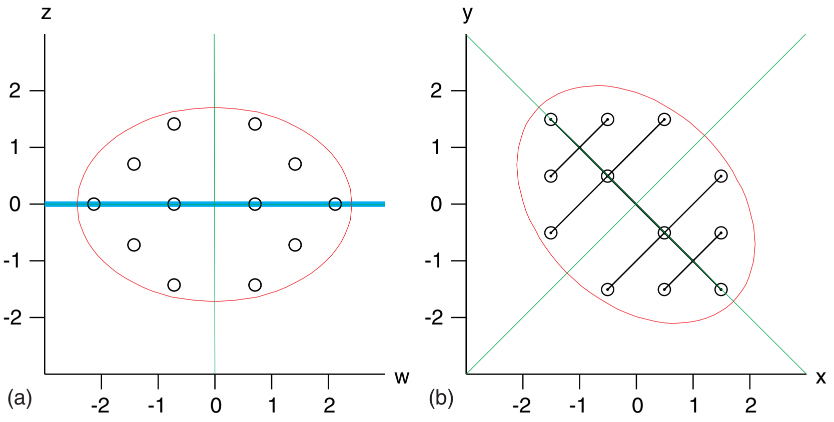


Figure 19.3 (Continuation to Figure 19.2.) (a) The centered data matrix $\widetilde{\mathbf{X}}$ is postmultiplied by the rotation matrix $\mathbf{T}$: $\widetilde{\mathbf{X}}\mathbf{T} = (\mathbf{w} : \mathbf{z})$. What is $\mathbf{T} = (\mathbf{t}_1 : \mathbf{t}_2)$? (b) The orthogonal distances of the data points from the line $\mathscr{C}(\mathbf{t}_1)$ are marked up.

and

$$\mathbf{E}' = (\mathbf{I}_m - \mathbf{P}_{\mathbf{G}})\widetilde{\mathbf{X}}'. \tag{19.77}$$

The columns of $\mathbf{E}'$ are the differences (residuals) between the data points and their projections onto $\mathscr{C}(\mathbf{G})$. Minimizing the sum of squared norms of these differences means that we have to minimize

$$\|\mathbf{E}\|_F^2 = \|\widetilde{\mathbf{X}} - \widetilde{\mathbf{X}}\mathbf{P}_{\mathbf{G}}\|_F^2, \tag{19.78}$$

that is, we have to minimize

$$\operatorname{tr}(\mathbf{E}'\mathbf{E}) = \operatorname{tr}(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}) - \operatorname{tr}(\mathbf{P}_{\mathbf{G}}\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}) , \tag{19.79}$$

which is equivalent to maximizing

$$\operatorname{tr}(\mathbf{G}'\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}\mathbf{G}) \quad \text{subject to} \quad \mathbf{G}'\mathbf{G} = \mathbf{I}_k . \tag{19.80}$$

On the basis of (19.38), the matrix $\mathbf{G}$ minimizing (19.78) comprises the first k orthonormal eigenvectors of $\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}$, which we may denote as

$$\dot{\mathbf{G}} = \mathbf{T}_{(k)} = (\mathbf{t}_1 : \ldots : \mathbf{t}_k) , \tag{19.81}$$

and

$$\begin{aligned} \min_{\mathbf{G}'\mathbf{G}=\mathbf{I}_k} \|\mathbf{E}\|_F^2 &= \|\widetilde{\mathbf{X}} - \widetilde{\mathbf{X}}\dot{\mathbf{G}}\dot{\mathbf{G}}'\|_F^2 \\ &= \|\widetilde{\mathbf{X}} - \widetilde{\mathbf{X}}\mathbf{T}_{(k)}\mathbf{T}_{(k)}'\|_F^2 \\ &= \operatorname{ch}_{k+1}(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}) + \cdots + \operatorname{ch}_p(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}) \\ &= \operatorname{sg}_{k+1}^2(\widetilde{\mathbf{X}}) + \cdots + \operatorname{sg}_p^2(\widetilde{\mathbf{X}}) . \end{aligned} \tag{19.82}$$

Note that the eigenvectors $\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}$ are the same as those of the covariance matrix, $\mathbf{S}$, say, of the observed data:

$$\mathbf{S} = \frac{1}{n-1}\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}} . \tag{19.83}$$

The new projected observations are the columns of the matrix

$$\mathbf{T}_{(k)}\mathbf{T}_{(k)}'\widetilde{\mathbf{X}}' = \mathbf{P}_{\mathbf{T}_{(k)}}\widetilde{\mathbf{X}}'. \tag{19.84}$$

In particular, if $k = 1$, the new projected observations are the columns of

$$\mathbf{t}_1\mathbf{t}_1'\widetilde{\mathbf{X}}' = \mathbf{t}_1(\mathbf{t}_1'\widetilde{\mathbf{X}}') = \mathbf{t}_1(\mathbf{t}_1'\tilde{\mathbf{x}}_{(1)}, \ldots, \mathbf{t}_1'\tilde{\mathbf{x}}_{(n)}) := \mathbf{t}_1\mathbf{s}_1' , \tag{19.85}$$

where the vector

$$\mathbf{s}_1 = \widetilde{\mathbf{X}}\mathbf{t}_1 = \begin{pmatrix} \mathbf{t}_1'\tilde{\mathbf{x}}_{(1)} \\ \vdots \\ \mathbf{t}_n'\tilde{\mathbf{x}}_{(n)} \end{pmatrix} \in \mathbb{R}^n \tag{19.86}$$

comprises the values (scores) of the first (sample) principal component; the ith individual has the score $\mathbf{t}_i'\tilde{\mathbf{x}}_{(i)}$ on the first principal component. The scores are the (signed) lengths of the new projected observations.

Let us change our approach a little bit and do some considerations in the variable space $\mathbb{R}^n$. Consider a given centered data matrix $\widetilde{\mathbf{X}} = (\tilde{\mathbf{x}}_1 : \tilde{\mathbf{x}}_2 : \ldots : \tilde{\mathbf{x}}_p)$ and let $\widetilde{\mathbf{X}}\mathbf{a}$ be a vector in the variable space $\mathbb{R}^n$, and suppose that $\mathbf{a}'\mathbf{a} = 1$. How to choose the vector $\mathbf{a} \in \mathbb{R}^p$ so that the vector $\widetilde{\mathbf{X}}\mathbf{a}$ goes “nicely” through the bunch of variable vectors $\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_p$? Now the orthogonal projection of the variable vector $\tilde{\mathbf{x}}_i$ onto the line $\mathscr{C}(\widetilde{\mathbf{X}}\mathbf{a})$ is $\mathbf{p}_i = \mathbf{P}_{\widetilde{\mathbf{X}}\mathbf{a}}\tilde{\mathbf{x}}_i$, and the corresponding difference (a kind of residual) is

$$\mathbf{f}_i = (\mathbf{I}_n - \mathbf{P}_{\widetilde{\mathbf{X}}\mathbf{a}})\tilde{\mathbf{x}}_i = \bigg(\mathbf{I}_n - \frac{\widetilde{\mathbf{X}}\mathbf{a}\mathbf{a}'\widetilde{\mathbf{X}}'}{\mathbf{a}'\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}\mathbf{a}}\bigg)\tilde{\mathbf{x}}_i\,. \tag{19.87}$$

The matrix of all $\mathbf{f}_i$-vectors is

$$\mathbf{F} = (\mathbf{f}_1 : \mathbf{f}_2 : \dots : \mathbf{f}_p) = (\mathbf{I}_n - \mathbf{P}_{\widetilde{\mathbf{X}}\mathbf{a}})\widetilde{\mathbf{X}}\,, \tag{19.88}$$

and it seems natural to choose $\mathbf{a}$ so that the matrix $\mathbf{F}$ is small in some sense. Taking the Euclidean matrix norm, we get

$$\begin{aligned} f(\mathbf{a}) := \|(\mathbf{I}_n - \mathbf{P}_{\widetilde{\mathbf{X}}\mathbf{a}})\widetilde{\mathbf{X}}\|_F^2 &= \operatorname{tr}(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}) - \operatorname{tr}(\mathbf{P}_{\widetilde{\mathbf{X}}\mathbf{a}}\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}') \\ &= \operatorname{tr}(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}) - \frac{\mathbf{a}'(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^2\mathbf{a}}{\mathbf{a}'\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}\mathbf{a}}\,. \end{aligned} \tag{19.89}$$

Now the minimum of $f(\mathbf{a})$ is attained when $\mathbf{a}$ is chosen as the eigenvector corresponding to the largest eigenvalue of $\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}} = (n-1)\mathbf{S}$. Notice that $\mathbf{S}$ and $\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}$ have the same eigenvectors. Hence we conclude that this approach, putting the vector $\widetilde{\mathbf{X}}\mathbf{a}$ into the variable space in an appropriate way, yields the same results as the corresponding task in the observation space on page 403. Corresponding considerations can be done also in the more general case by looking for a matrix $\mathbf{A} \in \mathbb{R}^{p\times k}$ with orthonormal columns such that the residual matrix $\mathbf{F}$ in (19.88) is

$$\mathbf{F} = (\mathbf{f}_1 : \mathbf{f}_2 : \dots : \mathbf{f}_p) = (\mathbf{I}_n - \mathbf{P}_{\widetilde{\mathbf{X}}\mathbf{A}})\widetilde{\mathbf{X}}\,. \tag{19.90}$$

Let us take a look at the connection between the SVD and the principal component analysis. Let $\widetilde{\mathbf{X}}$ be a centered $n \times p$ data matrix whose SVD is

$$\widetilde{\mathbf{X}} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}' := \mathbf{U}\boldsymbol{\Delta}\mathbf{T}', \tag{19.91}$$

where we have replaced the "traditional" notation $\mathbf{V}$ with $\mathbf{T}$. Then

$$\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}} = \mathbf{V}\boldsymbol{\Delta}'\boldsymbol{\Delta}\mathbf{V}' := \mathbf{T}\boldsymbol{\Lambda}\mathbf{T}', \tag{19.92}$$

and so $\mathbf{T}$ is the matrix of the orthonormal eigenvectors of $\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}$ and the diagonal matrix $\boldsymbol{\Lambda}$ comprises its eigenvalues (squared singular values of $\widetilde{\mathbf{X}}$). If we consider a best rank-1 approximation of the centered data matrix $\widetilde{\mathbf{X}}$, then the Eckart–Young theorem (p. 400) tells us that

$$\min_{\mathbf{B},\operatorname{rank}(\mathbf{B})=1} \|\widetilde{\mathbf{X}} - \mathbf{B}\|_F^2 = \delta_2^2 + \dots + \delta_p^2\,, \tag{19.93}$$

and that the minimum is attained by taking

$$\mathbf{B} = \dot{\mathbf{B}}_1 = \delta_1 \mathbf{u}_1 \mathbf{t}_1' \,, \tag{19.94}$$

where $\mathbf{t}_1$ is the first eigenvector of $\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}$ (or equivalently, the first eigenvector of covariance matrix $\mathbf{S}$), i.e., the first right singular vector of $\widetilde{\mathbf{X}}$, satisfying [see (19.23), p. 395]

$$\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}\mathbf{t}_1 = \delta_1^2 \mathbf{t}_1 \,, \quad \widetilde{\mathbf{X}}\mathbf{t}_1 = \delta_1 \mathbf{u}_1 \,. \tag{19.95}$$

Hence we have

$$\dot{\mathbf{B}}_1 = \delta_1 \mathbf{u}_1 \mathbf{t}_1' = \widetilde{\mathbf{X}}\mathbf{t}_1\mathbf{t}_1' = \mathbf{s}_1\mathbf{t}_1' \,, \quad \text{where } \mathbf{s}_1 = \widetilde{\mathbf{X}}\mathbf{t}_1 \,, \tag{19.96}$$

which is precisely the result obtained above by minimizing the sum of squared distances. To repeat one more time, the best rank-1 approximation of the centered data matrix $\widetilde{\mathbf{X}}$ is the matrix $\dot{\mathbf{B}}_1 = \mathbf{s}_1\mathbf{t}_1'$ whose rows are multiples of $\mathbf{t}_1'$; the columns of $\dot{\mathbf{B}}_1' = \mathbf{t}_1\mathbf{s}_1'$ are projections of the data points onto the one-dimensional space (line) $\mathscr{C}(\mathbf{t}_1)$.

It is interesting to observe that the matrix approximation by a matrix of a lower rank indeed yields the principal component analysis. It is also noteworthy, cf. Mustonen (1995, p. 65), that, as in (19.86) with $k = 1$, the values (scores) of the principal components can be obtained directly from the SVD of $\widetilde{\mathbf{X}}$; the values of the jth principal component are in the jth column of

$$\mathbf{U}\boldsymbol{\Delta} = \widetilde{\mathbf{X}}\mathbf{T} \,. \tag{19.97}$$

The columns $\underline{\tilde{\mathbf{x}}}_{(i)}$, say, of matrix $\mathbf{T}'\widetilde{\mathbf{X}}'$ represent the new rotated observations (data points); the cosine between the original observations $\tilde{\mathbf{x}}_{(i)}$ and $\tilde{\mathbf{x}}_{(j)}$ is the same as the cosine between the new observations $\underline{\tilde{\mathbf{x}}}_{(i)}$ and $\underline{\tilde{\mathbf{x}}}_{(j)}$ since $\mathbf{T}$ is orthogonal. Note also that the variables corresponding to the columns of $\widetilde{\mathbf{X}}\mathbf{T}$, i.e., the principal components, are uncorrelated because $\mathbf{T}'\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}\mathbf{T} = \boldsymbol{\Delta}_*^2$ is a diagonal matrix.

19.5 Generalized Inverses through the SVD

Different types of generalized inverses can be conveniently characterized through the singular value decomposition. In the following, we will simply list (without proofs) some results. The following characterizations appear, e.g., in Rao (1973a, Complement 28, pp. 76–77), and Ben-Israel & Greville (2003, Ex. 14, p. 208).

Let $\mathbf{A}_{n\times m}$ have a singular value decomposition

$$\mathbf{A} = \mathbf{U}\begin{pmatrix} \boldsymbol{\Delta}_{1(r\times r)} & \mathbf{0}_{r\times(m-r)} \\ \mathbf{0}_{(n-r)\times r} & \mathbf{0}_{(n-r)\times(m-r)} \end{pmatrix}\mathbf{V}'. \tag{19.98}$$

Then the following statements hold:

$$\mathbf{G} \in \{\mathbf{A}^-\} \iff \mathbf{G} = \mathbf{V}\begin{pmatrix} \mathbf{\Delta}_1^{-1} & \mathbf{K} \\ \mathbf{L} & \mathbf{N} \end{pmatrix}\mathbf{U}', \tag{19.99}$$

$$\mathbf{G} \in \{\mathbf{A}_{12}^-\} \iff \mathbf{G} = \mathbf{V}\begin{pmatrix} \mathbf{\Delta}_1^{-1} & \mathbf{K} \\ \mathbf{L} & \mathbf{L}\mathbf{\Delta}_1\mathbf{K} \end{pmatrix}\mathbf{U}', \tag{19.100}$$

$$\mathbf{G} \in \{\mathbf{A}_{13}^-\} = \{\mathbf{A}_{\ell}^-\} \iff \mathbf{G} = \mathbf{V}\begin{pmatrix} \mathbf{\Delta}_1^{-1} & \mathbf{0} \\ \mathbf{L} & \mathbf{N} \end{pmatrix}\mathbf{U}', \tag{19.101}$$

$$\mathbf{G} \in \{\mathbf{A}_{14}^-\} = \{\mathbf{A}_{m}^-\} \iff \mathbf{G} = \mathbf{V}\begin{pmatrix} \mathbf{\Delta}_1^{-1} & \mathbf{K} \\ \mathbf{0} & \mathbf{N} \end{pmatrix}\mathbf{U}', \tag{19.102}$$

$$\mathbf{G} = \mathbf{A}^+ \iff \mathbf{G} = \mathbf{V}\begin{pmatrix} \mathbf{\Delta}_1^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{U}', \tag{19.103}$$

where $\mathbf{K}$, $\mathbf{L}$, and $\mathbf{N}$ are arbitrary matrices.

For the proof of (19.99), see page 110 of Section 4.1.

19.6 Canonical Correlations and the SVD

Canonical correlations are considered in various parts of this book; see, e.g., Section 2.1 (p. 77), Section 5.9 (p. 133), and Section 18.9 (p. 379).

As in Section 18.9, consider now random vector $\mathbf{u}$ whose covariance matrix is $\operatorname{cov}(\mathbf{u}) = \mathbf{I}_n$. Then

$$\operatorname{cov}\begin{pmatrix} \mathbf{A}'\mathbf{u} \\ \mathbf{B}'\mathbf{u} \end{pmatrix} = \begin{pmatrix} \mathbf{A}'\mathbf{A} & \mathbf{A}'\mathbf{B} \\ \mathbf{B}'\mathbf{A} & \mathbf{B}'\mathbf{B} \end{pmatrix} := \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix} = \mathbf{\Sigma}\,. \tag{19.104}$$

Let ϱ_i denote the ith largest canonical correlation between the random vectors $\mathbf{A}'\mathbf{u}$ and $\mathbf{B}'\mathbf{u}$. Then ϱ_1 is defined as the maximum value of the correlation between arbitrary linear combinations $\boldsymbol{\alpha}'\mathbf{A}'\mathbf{u}$ and $\boldsymbol{\beta}'\mathbf{B}'\mathbf{u}$. Let $\boldsymbol{\alpha}_1$ and $\boldsymbol{\beta}_1$ be the corresponding maximizing values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Then $\boldsymbol{\alpha}_1'\mathbf{A}'\mathbf{u}$ and $\boldsymbol{\beta}_1'\mathbf{B}'\mathbf{u}$ are called the first canonical variables. In view of (18.157) (p. 379),

$$\varrho_1^2 = \max_{\substack{\mathbf{A}\boldsymbol{\alpha}\neq\mathbf{0},\\ \mathbf{B}\boldsymbol{\beta}\neq\mathbf{0}}} \frac{(\boldsymbol{\alpha}'\mathbf{A}'\mathbf{B}\boldsymbol{\beta})^2}{\boldsymbol{\alpha}'\mathbf{A}'\mathbf{A}\boldsymbol{\alpha}\cdot\boldsymbol{\beta}'\mathbf{B}'\mathbf{B}\boldsymbol{\beta}} = \frac{\boldsymbol{\alpha}_1'\mathbf{A}'\mathbf{P}_{\mathbf{B}}\mathbf{A}\boldsymbol{\alpha}_1}{\boldsymbol{\alpha}_1'\mathbf{A}'\mathbf{A}\boldsymbol{\alpha}_1} = \operatorname{ch}_1(\mathbf{P}_{\mathbf{A}}\mathbf{P}_{\mathbf{B}})\,. \tag{19.105}$$

Hence we immediately see that

$$\varrho_1 = \operatorname{sg}_1\big((\mathbf{A}'\mathbf{A})^{+1/2}\mathbf{A}'\mathbf{B}(\mathbf{B}'\mathbf{B})^{+1/2}\big) = \operatorname{sg}_1\big(\mathbf{\Sigma}_{11}^{+1/2}\mathbf{\Sigma}_{21}\mathbf{\Sigma}_{22}^{+1/2}\big)\,, \tag{19.106}$$

where we have denoted shortly $(\mathbf{U}^+)^{1/2} = \mathbf{U}^{+1/2}$. (Actually we can use any generalized inverse $\mathbf{U}^-$ which is nonnegative definite.)

Suppose, for simplicity, that $\mathbf{A}$ and $\mathbf{B}$ have full column ranks. Then, in view of (19.28) (p. 396),

$$\begin{aligned}
\varrho_1^2 &= \max_{\boldsymbol{\alpha}\neq\mathbf{0},\,\boldsymbol{\beta}\neq\mathbf{0}} \frac{(\boldsymbol{\alpha}'\mathbf{A}'\mathbf{B}\boldsymbol{\beta})^2}{\boldsymbol{\alpha}'\mathbf{A}'\mathbf{A}\boldsymbol{\alpha}\cdot\boldsymbol{\beta}'\mathbf{B}'\mathbf{B}\boldsymbol{\beta}} \\
&= \max_{\boldsymbol{\alpha}\neq\mathbf{0},\,\boldsymbol{\beta}\neq\mathbf{0}} \frac{\big[\boldsymbol{\alpha}'(\mathbf{A}'\mathbf{A})^{1/2}(\mathbf{A}'\mathbf{A})^{-1/2}\mathbf{A}'\mathbf{B}(\mathbf{B}'\mathbf{B})^{-1/2}(\mathbf{B}'\mathbf{B})^{1/2}\boldsymbol{\beta}\big]^2}{\boldsymbol{\alpha}'(\mathbf{A}'\mathbf{A})^{1/2}(\mathbf{A}'\mathbf{A})^{1/2}\boldsymbol{\alpha}\cdot\boldsymbol{\beta}'(\mathbf{B}'\mathbf{B})^{1/2}(\mathbf{B}'\mathbf{B})^{1/2}\boldsymbol{\beta}} \\
&= \max_{\mathbf{a}\neq\mathbf{0},\,\mathbf{b}\neq\mathbf{0}} \frac{\big[\mathbf{a}'(\mathbf{A}'\mathbf{A})^{-1/2}\mathbf{A}'\mathbf{B}(\mathbf{B}'\mathbf{B})^{-1/2}\mathbf{b}'\big]^2}{\mathbf{a}'\mathbf{a}\cdot\mathbf{b}'\mathbf{b}} \\
&= \max_{\mathbf{a}'\mathbf{a}=\mathbf{b}'\mathbf{b}=1} (\mathbf{a}'\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}\mathbf{b}')^2 = (\mathbf{a}_1'\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}\mathbf{b}_1')^2 \\
&= \mathrm{sg}_1^2(\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}) = \mathrm{ch}_1(\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})\,, \qquad (19.107)
\end{aligned}$$

where $\mathbf{a}_1 = \boldsymbol{\Sigma}_{11}^{1/2}\boldsymbol{\alpha}_1$ and $\mathbf{b}_1 = \boldsymbol{\Sigma}_{22}^{1/2}\boldsymbol{\beta}_1$ are the first left and right singular vectors of $\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}$, respectively.

Next we show one geometric approach to end up to canonical correlations; see also Section 2.1 (p. 77).

Let $\mathbf{X}_{n\times a}$ and $\mathbf{Y}_{n\times b}$ be centered data matrices (having full column ranks) of variables $x_1, \dots, x_a$ and $y_1, \dots, y_b$, respectively. The problem is to find vectors $\mathbf{f} = \mathbf{X}\mathbf{r} \in \mathscr{C}(\mathbf{X})$ and $\mathbf{g} = \mathbf{Y}\mathbf{s} \in \mathscr{C}(\mathbf{Y})$ such that their cosine, i.e., the sample correlation $\mathrm{cor_d}(\mathbf{X}\mathbf{r}, \mathbf{Y}\mathbf{s})$ is maximal. Keeping first $\mathbf{f} = \mathbf{X}\mathbf{r}$ fixed, the vector $\mathbf{Y}\dot{\mathbf{s}}$ closest to $\mathbf{f}$ is

$$\mathbf{Y}\dot{\mathbf{s}} = \mathbf{P}_{\mathbf{Y}}\mathbf{X}\mathbf{r} = \mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{X}\mathbf{r}\,, \qquad (19.108)$$

and hence

$$\cos^2(\mathbf{X}\mathbf{r}, \mathbf{Y}\dot{\mathbf{s}}) = \cos^2(\mathbf{X}\mathbf{r}, \mathbf{P}_{\mathbf{Y}}\mathbf{X}\mathbf{r}) = \frac{\mathbf{r}'\mathbf{X}'\mathbf{P}_{\mathbf{Y}}\mathbf{X}\mathbf{r}}{\mathbf{r}'\mathbf{X}'\mathbf{X}\mathbf{r}} := \gamma^2. \qquad (19.109)$$

The maximal value of γ^2 is

$$\mathrm{ch}_1\big[\mathbf{X}'\mathbf{P}_{\mathbf{Y}}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\big] = \mathrm{ch}_1(\mathbf{P}_{\mathbf{X}}\mathbf{P}_{\mathbf{Y}}) := \gamma_1^2\,, \qquad (19.110)$$

which is the largest sample canonical correlation (squared) between the sets of variables $\{x_1, \dots, x_a\}$ and $\{y_1, \dots, y_b\}$.

19.7 Exercises

19.1. Find the rotation matrix $\mathbf{T}$ in Figure 19.3 (p. 404).

19.2. Confirm that the singular value decomposition of the row vector $(1, 1)$ is

$$(1,1) = 1 \cdot \begin{pmatrix} \sqrt{2} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}.$$

Notice that $(1,1)$ has no zero singular values—unless we take into account the honorary singular values; see page 395.

19.3. Let $\mathbf{A} \in \mathbb{R}^{n\times a}$ with $\operatorname{rank}(\mathbf{A}) = r$ and denote

$$\mathbf{B} = \begin{pmatrix} \mathbf{0} & \mathbf{A} \\ \mathbf{A}' & \mathbf{0} \end{pmatrix} \in \mathbb{R}^{(n+a)\times(n+a)}, \quad \operatorname{sg}_i(\mathbf{A}) = \delta_i\,,\ i = 1,\dots,r\,.$$

Show that the nonzero eigenvalues of $\mathbf{B}$ are $\delta_1,\dots,\delta_r,-\delta_1,\dots,-\delta_r$.

19.4. Consider $\mathbf{A} \in \mathbb{R}^{n\times m}$ with $\operatorname{rank}(\mathbf{A}) = r$, and suppose that we construct $\mathbf{U}_1 \in \mathbb{R}^{n\times r}$ and $\mathbf{V}_1 \in \mathbb{R}^{m\times r}$ with orthonormal columns so that $\mathbf{AA}'\mathbf{U}_1 = \mathbf{U}_1\mathbf{\Delta}_1^2$, and $\mathbf{A}'\mathbf{AV}_1 = \mathbf{V}_1\mathbf{\Delta}_1^2$, where $\mathbf{\Delta}_1^2$ is a diagonal matrix comprising the r nonzero eigenvalues of $\mathbf{A}'\mathbf{A}$. Thus the columns of $\mathbf{U}_1$ and $\mathbf{V}_1$ are orthonormal eigenvectors of $\mathbf{AA}'$ and $\mathbf{A}'\mathbf{A}$, respectively. Why does not this necessarily yield the singular value decomposition $\mathbf{A} = \mathbf{U}_1\mathbf{\Delta}_1\mathbf{V}_1'$?

Abadir & Magnus (2005, p. 226).

19.5. Consider symmetric $n \times n$ matrices $\mathbf{A}$ and $\mathbf{B}$. Then

$$\sum_{i=1}^{n} \operatorname{ch}_i(\mathbf{A}) \operatorname{ch}_{n-i-1}(\mathbf{B}) \le \operatorname{tr}(\mathbf{AB}) \le \sum_{i=1}^{n} \operatorname{ch}_i(\mathbf{A}) \operatorname{ch}_i(\mathbf{B})\,. \tag{19.111}$$

Suppose that $\mathbf{A} \in \mathbb{R}^{n\times p}$, $\mathbf{B} \in \mathbb{R}^{p\times n}$. Show, using Exercise 19.3, that the above result, due to von Neumann (1937), implies the following:

$$-\sum_{i=1}^{k} \operatorname{sg}_i(\mathbf{A}) \operatorname{sg}_i(\mathbf{B}) \le \operatorname{tr}(\mathbf{AB}) \le \sum_{i=1}^{k} \operatorname{sg}_i(\mathbf{A}) \operatorname{sg}_i(\mathbf{B})\,, \tag{19.112}$$

where $k = \min\big(\operatorname{rank}(\mathbf{A}), \operatorname{rank}(\mathbf{B})\big)$.

Rao & Rao (1998, pp. 386–387).

19.6. Consider $\mathbf{A},\mathbf{B} \in \mathbb{R}^{n\times m}$, where $\operatorname{rank}(\mathbf{A}) = r > \operatorname{rank}(\mathbf{B}) = k$, and $\alpha_1 \ge \dots \ge \alpha_r$ and $\beta_1 \ge \dots \ge \beta_k$ are the singular values of $\mathbf{A}$ and $\mathbf{B}$, respectively. Confirm, using (19.112), that the following holds:

$$\begin{aligned} \|\mathbf{A}-\mathbf{B}\|_F^2 &= \operatorname{tr}(\mathbf{A}'\mathbf{A}) + \operatorname{tr}(\mathbf{B}'\mathbf{B}) - 2\operatorname{tr}(\mathbf{A}'\mathbf{B}) \\ &\ge \sum_{i=1}^{r} \alpha_i^2 + \sum_{j=1}^{k} \beta_j^2 - 2\sum_{j=1}^{k} \alpha_j\beta_j \\ &= \sum_{j=1}^{k} (\alpha_j - \beta_j)^2 + \sum_{i=k+1}^{r} \alpha_i^2 \ge \sum_{i=k+1}^{r} \alpha_i^2\,, \end{aligned} \tag{19.113}$$

thus offering an alternative proof for the Eckart–Young theorem (p. 400).

19.7. Suppose that $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{A}-\mathbf{B}$ are symmetric nonnegative definite $n \times n$ matrices. Prove that then

$$\operatorname{ch}_i(\mathbf{A}-\mathbf{B}) \geq \operatorname{ch}_{i+k}(\mathbf{A}), \quad i = 1, \dots, n-k,$$

and with equality for all i if and only if

$$\mathbf{B} = \sum_{i=1}^{k} \operatorname{ch}_i(\mathbf{A})\mathbf{t}_i\mathbf{t}_i' = \mathbf{T}_{(k)}\mathbf{\Lambda}_{(k)}\mathbf{T}_{(k)}',$$

where $\mathbf{\Lambda}_{(k)}$ comprises the first k eigenvalues of $\mathbf{A}$ and $\mathbf{T}_{(k)} = (\mathbf{t}_1 : \dots : \mathbf{t}_k)$ is the matrix of the corresponding orthonormal eigenvectors of $\mathbf{A}$.

Okamoto & Kanazawa (1968), Okamoto (1969),
Fujikoshi, Ulyanov & Shimizu (2010, p. 505).

19.8. Prove (19.62) (p. 400): For $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times m}$, where $\operatorname{rank}(\mathbf{A}) = r$ and $\operatorname{rank}(\mathbf{B}) = k$, we have

$$\operatorname{sg}_i(\mathbf{A}-\mathbf{B}) \geq \operatorname{sg}_{i+k}(\mathbf{A}), \quad i+k \leq r, \tag{19.114}$$

and $\operatorname{sg}_i(\mathbf{A}-\mathbf{B}) \geq 0$ for $i+k > r$. The equality is attained in (19.114) if and only if $k \leq r$ and $\mathbf{B} = \sum_{i=1}^{k} \operatorname{sg}_i(\mathbf{A})\mathbf{t}_i\mathbf{u}_i'$, where $\mathbf{A} = \sum_{i=1}^{c} \operatorname{sg}_i(\mathbf{A})\mathbf{t}_i\mathbf{u}_i'$ is the SVD of $\mathbf{A}$.

Rao (1980, pp. 8–9), Rao & Rao (1998, p. 382), Stewart (1993, §6).

19.9. Let $\mathbf{A}$ and $\mathbf{B}$ be nonnegative definite $n \times n$ matrices so that $\mathscr{C}(\mathbf{A}) \subset \mathscr{C}(\mathbf{B})$, $b = \operatorname{rk}(\mathbf{B})$, and let $\mathbf{X}_{n\times p}$ be a matrix with $r = \operatorname{rk}(\mathbf{BX})$. Then

$$\operatorname{ch}_{b-r+i}(\mathbf{B}^-\mathbf{A}) \leq \operatorname{ch}_i[(\mathbf{X}'\mathbf{BX})^-\mathbf{X}'\mathbf{AX}] \leq \operatorname{ch}_i(\mathbf{B}^-\mathbf{A}), \quad i = 1, \dots, r.$$

Show that the upper bound above is attained simultaneously for all $i = 1, \dots, r$ if and only if there exists $\mathbf{Q}_1 \in \mathbb{R}^{n\times r}$ such that

$$\mathbf{Q}_1'\mathbf{BQ}_1 = \mathbf{I}_r, \quad \mathbf{AQ}_1 = \mathbf{BQ}_1\mathbf{D}_1, \quad \text{and} \quad \mathscr{C}(\mathbf{BQ}_1) = \mathscr{C}(\mathbf{BX}),$$

where $\mathbf{D}_1 \in \mathbb{R}^{r\times r}$ is a diagonal matrix comprising the r largest proper eigenvalues of $\mathbf{A}$ with respect to $\mathbf{B}$ and the columns of $\mathbf{Q}_1$ are the corresponding proper eigenvectors. Confirm that the upper bound above is attained by choosing $\mathbf{X}$ so that the columns of $\mathbf{Q}_1$ span $\mathscr{C}(\mathbf{X})$.

Scott & Styan (1985, Th. 2).

19.10. Let f be a function defined in the set NND_p. Then it can be shown that a necessary and sufficient condition for f to be

(a) strictly increasing, i.e., $f(\mathbf{A}) > f(\mathbf{B})$ if $\mathbf{A} \geq_{\mathsf{L}} \mathbf{B}$ and $\mathbf{A} \neq \mathbf{B}$, and

(b) invariant under orthogonal transformation, i.e., $f(\mathbf{W}'\mathbf{AW}) = f(\mathbf{A})$ for any orthogonal $\mathbf{W}$, is that $f(\mathbf{A}) = g[\mathrm{ch}_1(\mathbf{A}), \dots, \mathrm{ch}_p(\mathbf{A})]$ for some g that is strictly increasing in each argument. This means that minimizing $f(\mathbf{A})$ with respect to $\mathbf{A}$ is equivalent to simultaneously minimizing the eigenvalues of $\mathbf{A}$.

Confirm that $\mathrm{tr}(\mathbf{A})$, the Frobenius norm $\|\mathbf{A}\|_F$, and $\det(\mathbf{A})$ satisfy the above definition. Show that if f satisfies (a) and (b), and $\mathbf{y}$ is a given p-dimensional random vector, then

$$f[\mathrm{cov}(\mathbf{y} - \mathbf{A}\mathbf{x})] \quad \text{is minimized for any } \mathbf{x} \text{ and any } \mathbf{A}_{p\times k}$$

if and only if $\mathbf{A}\mathbf{x} = \mathbf{T}_{(k)}\mathbf{T}'_{(k)}\mathbf{y}$, where $\mathbf{T}_{(k)}$ comprises the first k orthonormal eigenvectors of $\mathrm{cov}(\mathbf{y})$.

Okamoto & Kanazawa (1968), Seber (1984, pp. 177–178),
Fujikoshi, Ulyanov & Shimizu (2010, §10.2).

19.11. The variables (columns) $\mathbf{x}_1, \dots, \mathbf{x}_p$ are exactly collinear if one of the $\mathbf{x}_i$ is an exact linear combination of the others. This is exact collinearity, i.e., linear dependency, and by term collinearity or near dependency we mean inexact collinear relations. Condition number of $\mathbf{X}$ is defined as:

$$\mathrm{cond}(\mathbf{X}) = \kappa(\mathbf{X}) = \frac{\mathrm{ch}^{1/2}_{\max}(\mathbf{X}'\mathbf{X})}{\mathrm{ch}^{1/2}_{\min}(\mathbf{X}'\mathbf{X})} = \frac{\mathrm{sg}_{\max}(\mathbf{X})}{\mathrm{sg}_{\min}(\mathbf{X})}\,.$$

Show that the following result holds:

$$\kappa(\mathbf{X}_1 : \mathbf{X}_2) \geq \kappa(\mathbf{X}_1 : \mathbf{M}_1\mathbf{X}_2) \geq \kappa(\mathbf{M}_1\mathbf{X}_2)\,.$$

Belsley (1991, Th. 6.1).

19.12 (Contingency table, $r \times c$). Consider the similar situation as in Exercise 0.7 (p. 47), but suppose the variables x and y have values $A_1, \dots, A_r$ and $B_1, \dots, B_c$, respectively, and assume that we have n observations from these variables. Let us define new variables in the following way:

$$\begin{aligned}
&x_1 = 1 \text{ if } x \text{ has value } A_1, \text{ and } x_1 = 0 \text{ otherwise,}\\
&\dots\\
&x_r = 1 \text{ if } x \text{ has value } A_r, \text{ and } x_r = 0 \text{ otherwise,}
\end{aligned}$$

and let $y_1, \dots, y_c$ be defined in the corresponding way with respect to the values $B_1, \dots, B_c$. Denote the observed $n \times (r + c)$ data matrix as

$$\mathbf{U} = (\mathbf{x}_1 : \dots : \mathbf{x}_r : \mathbf{y}_1 : \dots : \mathbf{y}_c) = (\mathbf{X} : \mathbf{Y})\,.$$

Let our task be to calculate canonical correlations ϱ_i, say, between the x-variables and the y-variables, that is, we are interested in linear combinations $\mathbf{X}\boldsymbol{\alpha}$ and $\mathbf{Y}\boldsymbol{\beta}$ which have maximal correlation. (Actually it might be better to denote $\hat{\varrho}_i$ because these are a sample statistics.) Suppose that $r \leq c$, in which case we have r pairs $(\mathbf{X}\boldsymbol{\alpha}_i, \mathbf{Y}\boldsymbol{\beta}_i)$ yielding the c canonical correlations ϱ_i. Show, using Exercise 0.8 (p. 48):

$$\begin{aligned}\varrho_i &= \operatorname{sg}_i\big[\big((\mathbf{X}'\mathbf{C}\mathbf{X})^-\big)^{1/2}\mathbf{X}'\mathbf{C}\mathbf{Y}\big((\mathbf{Y}'\mathbf{C}\mathbf{Y})^-\big)^{1/2}\big] \\ &= \operatorname{sg}_i\big[\mathbf{D}_{\mathbf{r}}^{-1/2}(\mathbf{F}-\mathbf{E})\mathbf{D}_{\mathbf{c}}^{-1/2}\big] = \operatorname{sg}_i(\mathbf{G}_*)\,,\end{aligned}$$

where $\mathbf{C}$ is the centering matrix, $(\mathbf{X}'\mathbf{C}\mathbf{X})^- \in \mathrm{NND}_r$, $(\mathbf{Y}'\mathbf{C}\mathbf{Y})^- \in \mathrm{NND}_c$, and, as in Exercise 0.7 (p. 47):

$$\begin{aligned}&\mathbf{D}_{\mathbf{c}} = \mathbf{Y}'\mathbf{Y} = \operatorname{diag}(\mathbf{c})\,, \quad \mathbf{D}_{\mathbf{r}} = \mathbf{X}'\mathbf{X} = \operatorname{diag}(\mathbf{r})\,,\\ &\mathbf{c} = \mathbf{Y}'\mathbf{1}_n \in \mathbb{R}^c, \quad \mathbf{r} = \mathbf{X}'\mathbf{1}_n \in \mathbb{R}^r,\\ &\mathbf{F} = \mathbf{X}'\mathbf{Y} \in \mathbb{R}^{r\times c}\text{: contingency table of } x \text{ vs. } y\,,\\ &\mathbf{E} = \mathbf{r}\mathbf{c}'/n \in \mathbb{R}^{r\times c}\text{: ``expected'' frequencies of the contingency table } \mathbf{F}\,,\\ &\mathbf{G}_* = \mathbf{D}_{\mathbf{r}}^{-1/2}(\mathbf{F}-\mathbf{E})\mathbf{D}_{\mathbf{c}}^{-1/2} = (\mathbf{X}'\mathbf{X})^{-1/2}\mathbf{X}'\mathbf{C}\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1/2}.\end{aligned}$$

19.13 (Continued ...). Confirm:

(a) $\varrho_i^2 = \operatorname{ch}_i(\mathbf{P}_{\mathbf{CX}}\mathbf{P}_{\mathbf{CY}}) = \operatorname{ch}_i(\mathbf{C}\mathbf{P}_{\mathbf{X}}\mathbf{C}\mathbf{P}_{\mathbf{Y}})$,

(b) $\varrho_i^2 = \operatorname{ch}_i[\mathbf{D}_{\mathbf{r}}^{-1}(\mathbf{F}-\mathbf{E})\mathbf{D}_{\mathbf{c}}^{-1}(\mathbf{F}-\mathbf{E})']$.

(c) $\chi^2 = n\operatorname{tr}(\mathbf{P}_{\mathbf{CX}}\mathbf{P}_{\mathbf{CY}}) = n(\varrho_1^2 + \cdots + \varrho_r^2)$; see Exercise 0.9 (p. 48).

(d) Let $\mathbf{G}_* = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}'$ be the SVD of $\mathbf{G}_*$. Then $\boldsymbol{\alpha}_i = \mathbf{D}_{\mathbf{r}}^{-1/2}\mathbf{u}_i$ and $\boldsymbol{\beta}_i = \mathbf{D}_{\mathbf{c}}^{-1/2}\mathbf{v}_i$ yield the ϱ_i.

(e) Check from literature how close we are to the correspondence analysis.

Greenacre (2007, pp. 201–202), Izenman (2008, §17), Rao (1995).

19.14 (Continued ...). Confirm the following:

(a) $\mathbf{P}_{\mathbf{CX}}\mathbf{P}_{\mathbf{CY}} = \mathbf{P}_{\mathbf{X}}\mathbf{P}_{\mathbf{Y}} - \mathbf{J}$,

(b) $\operatorname{nzch}(\mathbf{P}_{\mathbf{X}}\mathbf{P}_{\mathbf{Y}}) = \{1, \varrho_1^2, \varrho_2^2, \dots, \varrho_r^2\}$,

(c) $\operatorname{nzch}(\mathbf{P}_{\mathbf{CX}}\mathbf{P}_{\mathbf{CY}}) = \{\varrho_1^2, \varrho_2^2, \dots, \varrho_r^2\}$,

(d) $\operatorname{nzch}(\mathbf{P}_{\mathbf{Q_Y X}}\mathbf{P}_{\mathbf{Q_X Y}}) = \{\, \varrho_i^2 : 0 < \varrho_i^2 < 1 \,\}$; see Exercise 18.10 (p. 387).

Philatelic Item 19.1 Paul Anthony Samuelson (1914–2009) was the first American to win the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel (in 1970). The Laguerre–Samuelson inequality connects an 1880 theorem of Edmond Nicolas Laguerre (1834–1886) concerning polynomials with all real roots and a 1968 inequality of Samuelson for the maximum and minimum deviation from the mean, see Section 20.5 (p. 420). The stamp (left panel) for Samuelson was issued by Guinea-Bissau in 2005. Jan Tinbergen (1903–1994) was a Dutch economist, who was awarded the first Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel in 1969 for having developed and applied dynamic models for the analysis of economic processes. Tinbergen shared this Nobel Prize with the Norwegian economist Ragnar Anton Kittil Frisch (1895–1973). For the Frisch–Waugh–Lovell theorem see Section 15.4 (p. 328). The stamp (right panel) for Tinbergen was issued by The Netherlands in 1995 (*Scott* 894).

Philatelic Item 19.2 John William Strutt, 3rd Baron Rayleigh, OM (1842–1919) was an English physicist who, with Sir William Ramsay, KCB (1852–1916), discovered the element argon, an achievement for which he earned the Nobel Prize for Physics in 1904. For a real symmetric matrix $\mathbf{A}$ the Rayleigh quotient $\mathbf{x}'\mathbf{A}\mathbf{x}/\mathbf{x}'\mathbf{x}$ lies between the minimum and maximum eigenvalues of $\mathbf{A}$, see Section 18.1 (p. 360).

Chapter 20
The Cauchy–Schwarz Inequality

> *Yes, I've learnt from my mistakes*
> *and I think I'm now able to repeat them almost exactly.*
>
> PETER COOK

As Steele (2004, p. 1) says, there is no doubt that the Cauchy–Schwarz inequality is one of the most widely and most important inequalities in all of mathematics. This chapter gives some examples of its use in statistics; further examples appear in several places in this book. The Cauchy–Schwarz inequality is also known as the Cauchy–Bouniakowsky–Schwarz inequality and is named after Augustin-Louis Cauchy (1789–1857) (see also Philatelic Item 12.1, p. 290), Viktor Yakovlevich Bouniakowsky [Buniakovskii, Bunyakovsky] (1804–1899), and [Karl] Hermann Amandus Schwarz (1843–1921); see Cauchy (1821)[1] Bouniakowsky (1859, pp. 3–4), and Schwarz (1888, pp. 343–345), and the book by Steele (2004, Ch. 1).

Theorem 20 (The Cauchy–Schwarz inequality). *Let* $\mathbf{x}$ *and* $\mathbf{y}$ *be* $n \times 1$ *nonnull real vectors. Then*

$$(\mathbf{x}'\mathbf{y})^2 \leq \mathbf{x}'\mathbf{x} \cdot \mathbf{y}'\mathbf{y}\,, \quad \textit{for all } \mathbf{x},\, \mathbf{y} \tag{20.1}$$

is the vector version of the Cauchy–Schwarz inequality. Equality holds in (20.1) *if and only if* $\mathbf{x}$ *and* $\mathbf{y}$ *are linearly dependent, i.e.,*

$$(\mathbf{x}'\mathbf{y})^2 = \mathbf{x}'\mathbf{x} \cdot \mathbf{y}'\mathbf{y} \iff \textit{there exists } \lambda \in \mathbb{R}\colon \mathbf{x} = \lambda\mathbf{y}\,. \tag{20.2}$$

Proof. There are many ways to prove (20.2), see, e.g., Marcus & Minc (1992, p. 61). We prove it here using the presentation of the orthogonal projector onto the column space $\mathscr{C}(\mathbf{x})$. Namely, it is obvious that

$$\mathbf{y}'(\mathbf{I}_n - \mathbf{P}_{\mathbf{x}})\mathbf{y} \geq 0\,, \tag{20.3}$$

i.e.,

$$\mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y} \geq 0\,, \tag{20.4}$$

---

[1] As Steele (2004, p. 10) says: "Oddly enough, Cauchy did not use his inequality in his text, except in some illustrative exercises."

S. Puntanen et al., *Matrix Tricks for Linear Statistical Models: Our Personal Top Twenty*,
DOI 10.1007/978-3-642-10473-2_21,

$$\frac{(\mathbf{x}'\mathbf{y})^2}{\mathbf{x}'\mathbf{x}} \leq \mathbf{y}'\mathbf{y}\,, \tag{20.5}$$

from which (20.1) follows. Clearly the equality in (20.1) holds if and only if

$$(\mathbf{I}_n - \mathbf{P}_{\mathbf{x}})\mathbf{y} = \mathbf{0}\,, \tag{20.6}$$

which is equivalent to the right-hand side of (20.2). Note that clearly the equality in (20.1) holds if $\mathbf{x} = \mathbf{0}$ or $\mathbf{y} = \mathbf{0}$. □

20.1 Specific Versions of the Cauchy–Schwarz Inequality

Let $\mathbf{A}$ be an $n \times n$ nonnegative definite symmetric matrix with the eigenvalue decomposition

$$\begin{gathered} \mathbf{A} = \mathbf{T}\mathbf{\Lambda}\mathbf{T}', \\ \mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \dots, \lambda_n)\,, \quad \mathbf{T} = (\mathbf{t}_1 : \dots : \mathbf{t}_n)\,, \quad \mathbf{T}'\mathbf{T} = \mathbf{I}_n\,. \end{gathered} \tag{20.7}$$

Here $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$, and $\mathbf{t}_i$ is the eigenvector corresponding to λ_i. In particular, denoting $\mathbf{T}_1 = (\mathbf{t}_1 : \dots : \mathbf{t}_r)$ and $\mathbf{\Lambda}_1 = \operatorname{diag}(\lambda_1, \dots, \lambda_r)$, where $r = \operatorname{rank}(\mathbf{A})$, we get the symmetric nonnegative definite square root of $\mathbf{A}$: $\mathbf{A}^{1/2} = \mathbf{T}_1\mathbf{\Lambda}_1^{1/2}\mathbf{T}_1'$. Similarly we get $(\mathbf{A}^+)^{1/2} = \mathbf{T}_1\mathbf{\Lambda}_1^{-1/2}\mathbf{T}_1'$ and

$$\mathbf{A}^{1/2}(\mathbf{A}^+)^{1/2} = \mathbf{T}_1\mathbf{T}_1' = \mathbf{P}_{\mathbf{A}}\,. \tag{20.8}$$

Substituting now $\mathbf{x} = \mathbf{A}^{1/2}\mathbf{u}$ and $\mathbf{y} = (\mathbf{A}^+)^{1/2}\mathbf{v}$ into (20.1) we obtain

$$[\mathbf{u}'\mathbf{A}^{1/2}(\mathbf{A}^+)^{1/2}\mathbf{v}]^2 \leq \mathbf{u}'\mathbf{A}^{1/2}\mathbf{A}^{1/2}\mathbf{u} \cdot \mathbf{v}'(\mathbf{A}^+)^{1/2}(\mathbf{A}^+)^{1/2}\mathbf{v}\,, \tag{20.9a}$$

$$(\mathbf{u}'\mathbf{P}_{\mathbf{A}}\mathbf{v})^2 \leq \mathbf{u}'\mathbf{A}\mathbf{u} \cdot \mathbf{v}'\mathbf{A}^+\mathbf{v}\,. \tag{20.9b}$$

Equality holds in (20.9) if and only if

$$\mathbf{A}^{1/2}\mathbf{u} \text{ and } (\mathbf{A}^+)^{1/2}\mathbf{v} \text{ are linearly dependent,} \tag{20.10}$$

for which one sufficient condition is

$$\mathbf{u} \in \mathscr{N}(\mathbf{A}) \quad \text{or} \quad \mathbf{v} \in \mathscr{N}(\mathbf{A})\,. \tag{20.11}$$

Requesting that $\mathbf{u} \notin \mathscr{N}(\mathbf{A})$ and $\mathbf{v} \notin \mathscr{N}(\mathbf{A})$ the equality in (20.9) can be characterized by the condition

$$\mathbf{A}^{1/2}\mathbf{u} = \lambda\mathbf{A}^{+1/2}\mathbf{v} \quad \text{for some } \lambda \in \mathbb{R}\,, \tag{20.12}$$

which is equivalent (please confirm) to

$$\mathbf{A}\mathbf{u} = \lambda\mathbf{P}_{\mathbf{A}}\mathbf{v} \quad \text{for some } \lambda \in \mathbb{R}\,. \tag{20.13}$$

Furthermore,

$$(\mathbf{u}'\mathbf{v})^2 \leq \mathbf{u}'\mathbf{A}\mathbf{u} \cdot \mathbf{v}'\mathbf{A}^-\mathbf{v} \quad \text{for all } \mathbf{v} \in \mathscr{C}(\mathbf{A})\,. \tag{20.14}$$

Equality holds in (20.14) if and only if

$$\mathbf{A}\mathbf{u} = \lambda\mathbf{v} \quad \text{for some } \lambda \in \mathbb{R}\,. \tag{20.15}$$

Clearly, if $\mathbf{A}$ positive definite, we have

$$(\mathbf{u}'\mathbf{v})^2 \leq \mathbf{u}'\mathbf{A}\mathbf{u} \cdot \mathbf{v}'\mathbf{A}^{-1}\mathbf{v}\,, \tag{20.16}$$

and

$$(\mathbf{u}'\mathbf{u})^2 \leq \mathbf{u}'\mathbf{A}\mathbf{u} \cdot \mathbf{u}'\mathbf{A}^{-1}\mathbf{u}\,, \tag{20.17a}$$

$$1 \leq \frac{\mathbf{u}'\mathbf{A}\mathbf{u} \cdot \mathbf{u}'\mathbf{A}^{-1}\mathbf{u}}{(\mathbf{u}'\mathbf{u})^2}\,, \tag{20.17b}$$

$$\frac{(\mathbf{u}'\mathbf{u})^2}{\mathbf{u}'\mathbf{A}\mathbf{u} \cdot \mathbf{u}'\mathbf{A}^{-1}\mathbf{u}} \leq 1\,, \tag{20.17c}$$

where the equality holds (assuming $\mathbf{u} \neq \mathbf{0}$) if and only if $\mathbf{u}$ is an eigenvector of $\mathbf{A}$:

$$\mathbf{A}\mathbf{u} = \lambda\mathbf{u} \quad \text{for some } \lambda \in \mathbb{R}\,. \tag{20.18}$$

20.2 Maximizing $\operatorname{cor}(y, \mathbf{b}'\mathbf{x})$

Consider a $(p+1)$-dimensional random vector $\mathbf{z} = (\mathbf{x}', y)'$ whose covariance matrix is

$$\operatorname{cov}(\mathbf{z}) = \operatorname{cov}\begin{pmatrix} \mathbf{x} \\ y \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{xx}} & \boldsymbol{\sigma}_{\mathbf{x}y} \\ \boldsymbol{\sigma}'_{\mathbf{x}y} & \sigma_y^2 \end{pmatrix} = \boldsymbol{\Sigma}\,. \tag{20.19}$$

What is then the vector $\mathbf{b}_* \in \mathbb{R}^p$ which gives the maximal value for the correlation between the random variables y and $\mathbf{b}'\mathbf{x}$, i.e.,

$$\max_{\mathbf{b}} \operatorname{cor}^2(y, \mathbf{b}'\mathbf{x}) = \operatorname{cor}^2(y, \mathbf{b}'_*\mathbf{x})\,. \tag{20.20}$$

Using the Cauchy–Schwarz inequality it is easy to prove, see (9.18) (p. 193) that

$$\mathbf{b}_* = \boldsymbol{\Sigma}^-_{\mathbf{xx}}\boldsymbol{\sigma}_{\mathbf{x}y}\,, \tag{20.21}$$

which could, of course, be multiplied with any nonzero scalar, and

$$\max_{\mathbf{b}} \operatorname{cor}^2(y, \mathbf{b}'\mathbf{x}) = \frac{\boldsymbol{\sigma}'_{\mathbf{x}y}\boldsymbol{\Sigma}^-_{\mathbf{xx}}\boldsymbol{\sigma}_{\mathbf{x}y}}{\sigma_y^2} = \varrho^2_{y\cdot\mathbf{x}}\,, \tag{20.22}$$

which is the population multiple correlation coefficient (squared).

20.3 The Kantorovich Inequality

As pointed out in Section 10.7, p. 234), the Watson efficiency under the model $\mathscr{M} = \{\mathbf{y}, \mathbf{x}\beta, \sigma^2\mathbf{V}\}$, where $\mathbf{V}$ is positive definite, is

$$\phi = \operatorname{eff}(\hat{\beta}) = \frac{\operatorname{var}(\tilde{\beta})}{\operatorname{var}(\hat{\beta})} = \frac{(\mathbf{x}'\mathbf{V}^{-1}\mathbf{x})^{-1}}{(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{V}\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}} = \frac{(\mathbf{x}'\mathbf{x})^2}{\mathbf{x}'\mathbf{V}\mathbf{x} \cdot \mathbf{x}'\mathbf{V}^{-1}\mathbf{x}}\,, \tag{20.23}$$

and by the Cauchy–Schwarz inequality, we have $0 < \phi \leq 1$. The upper bound is obtained if and only if $\hat{\beta} = \tilde{\beta}$.

The lower bound of ϕ is obtained from the *Kantorovich inequality*; cf., e.g., Watson, Alpargu & Styan (1997):

$$\tau_1^2 := \frac{4\lambda_1\lambda_n}{(\lambda_1 + \lambda_n)^2} \leq \frac{(\mathbf{x}'\mathbf{x})^2}{\mathbf{x}'\mathbf{V}\mathbf{x} \cdot \mathbf{x}'\mathbf{V}^{-1}\mathbf{x}} = \phi\,, \tag{20.24}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$ are the ordered eigenvalues of $\mathbf{V}$. The lower bound is obtained when $\mathbf{x}$ is proportional to $\mathbf{t}_1 \pm \mathbf{t}_n$, where $\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_n$ are the orthonormal eigenvectors of $\mathbf{V}$ corresponding to the eigenvalues λ_i. The other way to write the Kantorovich inequality is, of course,

$$\frac{\mathbf{x}'\mathbf{V}\mathbf{x} \cdot \mathbf{x}'\mathbf{V}^{-1}\mathbf{x}}{(\mathbf{x}'\mathbf{x})^2} \leq \frac{(\lambda_1 + \lambda_n)^2}{4\lambda_1\lambda_n} := \mu_n = \frac{1}{\tau_1^2}\,. \tag{20.25}$$

The inequality (20.25) is usually attributed to Leonid Vitaliyevich Kantorovich for the inequality he established in Kantorovich (1948, pp. 142–144), Kantorovich (1952, pp. 106–107); see also Philatelic Item 7.1 (p. 154). As observed by Watson, Alpargu & Styan (1997), the inequality (20.25) originates from Frucht (1943) and so it may be called "Frucht–Kantorovich inequality".

The upper bound μ_n in (20.24) may be written in several different ways, e.g.,

$$\mu_n = \frac{(\lambda_1 + \lambda_n)^2}{4\lambda_1\lambda_n} = \left(\frac{(\lambda_1 + \lambda_n)/2}{\sqrt{\lambda_1\lambda_n}}\right)^2, \tag{20.26}$$

the square of the ratio of the arithmetic and geometric means of λ_1 and λ_n. We may call this upper bound μ_n the "Kantorovich ratio". Moreover, we will call $1 - (1/\mu_n)$ the "Wielandt ratio" and write

$$\nu_n^2 = 1 - \frac{1}{\mu_n}\,. \tag{20.27}$$

Other ways of expressing μ_n include

$$\mu_n = \frac{(\lambda_1+\lambda_n)^2}{4\lambda_1\lambda_n} = \frac{1}{4}\left(\sqrt{\frac{\lambda_1}{\lambda_n}}+\sqrt{\frac{\lambda_n}{\lambda_1}}\right)^2 = \frac{1}{2}(\lambda_1+\lambda_n)\cdot\frac{1}{2}\left(\frac{1}{\lambda_1}+\frac{1}{\lambda_n}\right)$$

$$= \left(\sqrt{\lambda_1\lambda_n}\bigg/\frac{2}{\frac{1}{\lambda_1}+\frac{1}{\lambda_n}}\right)^2 = \frac{1}{1-\left(\frac{\lambda_1-\lambda_n}{\lambda_1+\lambda_n}\right)^2}\,. \tag{20.28}$$

As noted in Section 10.7 (p. 237), the first antieigenvalue of $\mathbf{V}$ is

$$\tau_1 = \frac{\sqrt{\lambda_1\lambda_n}}{(\lambda_1+\lambda_n)/2} = \frac{1}{\sqrt{\mu_n}}\,. \tag{20.29}$$

20.4 The Wielandt Inequality

Let $\mathbf{V}_{n\times n}$ be a positive definite matrix and let $\mathbf{x}$ and $\mathbf{y}$ be nonnull vectors in $\mathbb{R}^n$ satisfying the condition $\mathbf{x}'\mathbf{y}=0$. Then

$$\frac{(\mathbf{x}'\mathbf{V}\mathbf{y})^2}{\mathbf{x}'\mathbf{V}\mathbf{x}\cdot\mathbf{y}'\mathbf{V}\mathbf{y}} \le \left(\frac{\psi-1}{\psi+1}\right)^2 = \left(\frac{\lambda_1-\lambda_n}{\lambda_1+\lambda_n}\right)^2 = \nu_n^2\,, \tag{20.30}$$

say, where $\lambda_1 \ge \dots \ge \lambda_n > 0$ are the eigenvalues of $\mathbf{V}$ and ψ is the condition number:

$$\psi = \lambda_1/\lambda_n = \operatorname{cond}(\mathbf{V})\,. \tag{20.31}$$

We will refer to (20.30) as the "Wielandt inequality". For the geometric meaning of the Wielandt inequality, see Gustafson (1999). The first appearance of (20.30) in a statistical context seems to be by Eaton (1976).

The matrix $\dot{\mathbf{M}}$ and the Cauchy–Schwarz inequality give an interesting proof for the Wielandt inequality, see Isotalo, Puntanen & Styan (2008d).

Proposition 20.1. *Let* $\mathbf{V}$ *be an* $n\times n$ *nonnegative definite matrix with* $\operatorname{rank}(\mathbf{V}) = v$ *and let* $\lambda_1 \ge \dots \ge \lambda_v > 0$ *be the nonzero eigenvalues of* $\mathbf{V}$*, and* $\mathbf{t}_1,\dots,\mathbf{t}_v$ *the corresponding orthonormal eigenvectors. Let* $\mathbf{x}\in\mathscr{C}(\mathbf{V})$ *and* $\mathbf{y}$ *be nonnull vectors satisfying the condition* $\mathbf{x}'\mathbf{y}=0$*. Then*

$$\frac{(\mathbf{x}'\mathbf{V}\mathbf{y})^2}{\mathbf{x}'\mathbf{V}\mathbf{x}\cdot\mathbf{y}'\mathbf{V}\mathbf{y}} \le \left(\frac{\lambda_1-\lambda_v}{\lambda_1+\lambda_v}\right)^2 = 1-\frac{4\lambda_1\lambda_v}{(\lambda_1+\lambda_v)^2} := \nu_v^2. \tag{20.32}$$

The upper bound is attained when $\mathbf{x}=\mathbf{t}_1+\mathbf{t}_v$ *and* $\mathbf{y}=\mathbf{t}_1-\mathbf{t}_v$*.*

Proof. Let us denote $\mathbf{M}=\mathbf{I}_n-\mathbf{P}_{\mathbf{x}}$. Then $\mathbf{x}'\mathbf{y}=0 \iff \mathbf{y}=\mathbf{M}\mathbf{a}$ for some $\mathbf{a}\in\mathbb{R}^n$. Hence our goal is to find $\mathbf{x}$ and $\mathbf{a}$ which maximize

$$\frac{(\mathbf{x}'\mathbf{V}\mathbf{y})^2}{\mathbf{x}'\mathbf{V}\mathbf{x}\cdot\mathbf{y}'\mathbf{V}\mathbf{y}} = \frac{(\mathbf{x}'\mathbf{V}\mathbf{M}\mathbf{a})^2}{\mathbf{x}'\mathbf{V}\mathbf{x}\cdot\mathbf{a}'\mathbf{M}\mathbf{V}\mathbf{M}\mathbf{a}} := f(\mathbf{x},\mathbf{a})\,. \tag{20.33}$$

In light of (20.14) (p. 417), we can write

$$\begin{aligned}(\mathbf{x}'\mathbf{VMa})^2 &= (\mathbf{x}'\mathbf{VM}\cdot\mathbf{a})^2 \le \mathbf{x}'\mathbf{VM}(\mathbf{MVM})^-\mathbf{MVx}\cdot\mathbf{a}'\mathbf{MVMa}\\ &= \mathbf{x}'\mathbf{V}\dot{\mathbf{M}}\mathbf{Vx}\cdot\mathbf{a}'\mathbf{MVMa}\quad\text{for all } \mathbf{x},\mathbf{a}\in\mathbb{R}^n, \end{aligned}\tag{20.34}$$

and so, using $\mathbf{x}\in\mathscr{C}(\mathbf{V})$ and part (a) of Theorem 15 (p. 318),

$$\begin{aligned} f(\mathbf{x},\mathbf{a}) &\le \frac{\mathbf{x}'\mathbf{V}\dot{\mathbf{M}}\mathbf{Vx}}{\mathbf{x}'\mathbf{Vx}} = \frac{\mathbf{x}'\mathbf{V}[\mathbf{V}^+ - \mathbf{V}^+\mathbf{x}(\mathbf{x}'\mathbf{V}^+\mathbf{x})^{-1}\mathbf{x}'\mathbf{V}^+]\mathbf{Vx}}{\mathbf{x}'\mathbf{Vx}}\\ &= 1 - \frac{(\mathbf{x}'\mathbf{x})^2}{\mathbf{x}'\mathbf{Vx}\cdot\mathbf{x}'\mathbf{V}^+\mathbf{x}} = 1 - \operatorname{eff}(\hat\beta \mid \mathscr{M})\,. \end{aligned}\tag{20.35}$$

The term $\operatorname{eff}(\hat\beta \mid \mathscr{M})$ in (20.35) is the efficiency of OLSE(β) under the simple weakly singular linear model $\mathscr{M} = \{\mathbf{y},\, \mathbf{x}\beta,\, \mathbf{V}\}$. In view of the Kantorovich inequality, we have

$$\operatorname{eff}(\hat\beta \mid \mathscr{M}) \ge \frac{4\lambda_1\lambda_v}{(\lambda_1+\lambda_v)^2}\,. \tag{20.36}$$

Substituting (20.36) into (20.35) yields (20.32). □

20.5 How Deviant Can You Be?

As Olkin (1992, p. 205) states, Thompson (1935) showed that the deviation of any particular observation from the mean is bounded by a multiple of the standard deviation. Implicit in Thompson's article is the Samuelson's inequality:

Proposition 20.2. *Let us denote*

$$\bar y = \frac{1}{n}\sum_{i=1}^n y_i\,,\quad s_y^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i-\bar y)^2 = \frac{1}{n-1}\mathbf{y}'\mathbf{C}\mathbf{y}\,, \tag{20.37}$$

where $\mathbf{y} = (y_1,\dots,y_n)' \in \mathbb{R}^n$, *and* $\mathbf{C}$ *is the centering matrix. Then*

$$(y_k-\bar y)^2 \le \frac{n-1}{n}\sum_{i=1}^n (y_i-\bar y)^2 = \frac{(n-1)^2}{n}s_y^2\,,\quad k=1,\dots,n\,. \tag{20.38}$$

We may cite Olkin (1992, p. 205) who states that although the inequality (20.38) appeared in early literature, it was not popularly known until the article by Samuelson (1968). Overviews of the development of Samuelson's inequality can be found in Olkin (1992), Jensen (1999), and Jensen & Styan (1999). (See also Philatelic Item 19.1, p. 414.)

Proof (of Samuelson's inequality). To prove (20.38), along the lines of Wolkowicz & Styan (1979), we first note that $y_k - \bar y$ can be expressed as

$$y_k - \bar{y} = \mathbf{u}'\mathbf{C}\mathbf{y}\,, \tag{20.39}$$

where $\mathbf{u} = \mathbf{i}_k = k$th column of $\mathbf{I}_n$. Hence we get

$$\begin{aligned}(y_k - \bar{y})^2 &= \mathbf{y}'\mathbf{C}\mathbf{u}\mathbf{u}'\mathbf{C}\mathbf{y} = \mathbf{u}'\mathbf{C}\mathbf{u} \cdot \mathbf{y}'\mathbf{C}\mathbf{u}(\mathbf{u}'\mathbf{C}\mathbf{u})^{-1}\mathbf{u}'\mathbf{C}\mathbf{y} \\ &= c_{kk} \cdot \mathbf{y}'\mathbf{P}_{\mathbf{C}\mathbf{u}}\mathbf{y} = \tfrac{n-1}{n} \cdot \mathbf{y}'\mathbf{P}_{\mathbf{C}\mathbf{u}}\mathbf{y}\,.\end{aligned} \tag{20.40}$$

In view of the Cauchy–Schwarz inequality (and since $\mathbf{y}'\mathbf{P}_{\mathbf{C}\mathbf{u}} = \mathbf{y}'\mathbf{C}\mathbf{P}_{\mathbf{C}\mathbf{u}}$), we get

$$(\mathbf{y}'\mathbf{P}_{\mathbf{C}\mathbf{u}}\mathbf{y})^2 = (\mathbf{y}'\mathbf{C} \cdot \mathbf{P}_{\mathbf{C}\mathbf{u}}\mathbf{y})^2 \le \mathbf{y}'\mathbf{C}\mathbf{y} \cdot \mathbf{y}'\mathbf{P}_{\mathbf{C}\mathbf{u}}\mathbf{y}\,, \tag{20.41}$$

i.e.,

$$\mathbf{y}'\mathbf{P}_{\mathbf{C}\mathbf{u}}\mathbf{y} \le \mathbf{y}'\mathbf{C}\mathbf{y} = (n-1)s_y^2\,. \tag{20.42}$$

Multiplying (20.42) with $(n-1)/n$ gives our claim (20.38). The equality is obtained if all other y_i's are equal except y_k. □

Note that if $n = 5$ million, then a single observation could not be more than 2236 standard deviations above the mean.

Trenkler & Puntanen (2005) considered the following generalization of Samuelson's inequality, where

$$\bar{\mathbf{x}} = \tfrac{1}{n}(\mathbf{x}_{(1)} + \cdots + \mathbf{x}_{(n)}) = \mathbf{X}_0'\mathbf{1}\tfrac{1}{n} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_k \end{pmatrix}, \tag{20.43a}$$

$$\mathbf{S}_{\mathbf{x}\mathbf{x}} = \tfrac{1}{n-1}\mathbf{X}_0'\mathbf{C}\mathbf{X}_0 = \tfrac{1}{n-1}\mathbf{T}_{\mathbf{x}\mathbf{x}} = \tfrac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_{(i)} - \bar{\mathbf{x}})(\mathbf{x}_{(i)} - \bar{\mathbf{x}})'. \tag{20.43b}$$

Proposition 20.3. *With the above notation,*

$$\frac{(n-1)^2}{n}\mathbf{S}_{\mathbf{x}\mathbf{x}} - (\mathbf{x}_{(j)} - \bar{\mathbf{x}})(\mathbf{x}_{(j)} - \bar{\mathbf{x}})' \ge_{\mathsf{L}} \mathbf{0}\,, \tag{20.44}$$

or equivalently,

$$(\mathbf{x}_{(j)} - \bar{\mathbf{x}})'\mathbf{S}_{\mathbf{x}\mathbf{x}}^{-1}(\mathbf{x}_{(j)} - \bar{\mathbf{x}}) \le \frac{(n-1)^2}{n}\,, \quad j = 1, \dots, n\,. \tag{20.45}$$

The equality above holds if and only if all $\mathbf{x}_{(i)}$ different from $\mathbf{x}_{(j)}$ coincide with their mean.

For the proof, see Trenkler & Puntanen (2005), and Arnold & Groeneveld (1974, Cor. 1); cf. also Proposition 8.2 (p. 157). For related references see also Farebrother (2009) and Olkin & Raveh (2009).

20.6 Two Inequalities Related to Cronbach's *alpha*

Proposition 20.4. *Let* $\mathbf{V}$ *be a symmetric nonnegative definite* $p \times p$ *matrix. Then*

$$\mathbf{1}'\mathbf{V}\mathbf{1} \geq \frac{p}{p-1}[\mathbf{1}'\mathbf{V}\mathbf{1} - \operatorname{tr}(\mathbf{V})], \tag{20.46}$$

i.e., (assuming $\mathbf{1}'\mathbf{V}\mathbf{1} \neq 0$*),*

$$1 \geq \frac{p}{p-1}\left(1 - \frac{\operatorname{tr}(\mathbf{V})}{\mathbf{1}'\mathbf{V}\mathbf{1}}\right). \tag{20.47}$$

The equality in (20.46) *is obtained if and only if* $\mathbf{V} = \delta^2\mathbf{1}\mathbf{1}'$ *for some* $\delta \in \mathbb{R}$.

Proof. Following Vehkalahti, Puntanen & Tarkkonen (2009, Th. 1), let us rewrite (20.46) as $(p-1)\mathbf{1}'\mathbf{V}\mathbf{1} \geq p\mathbf{1}'\mathbf{V}\mathbf{1} - p\operatorname{tr}(\mathbf{V})$, that is,

$$\frac{\mathbf{1}'\mathbf{V}\mathbf{1}}{\mathbf{1}'\mathbf{1}} \leq \operatorname{tr}(\mathbf{V}) = \lambda_1 + \lambda_2 + \cdots + \lambda_p\,, \tag{20.48}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ are the eigenvalues of $\mathbf{V}$; we will denote $\lambda_i = \operatorname{ch}_i(\mathbf{V})$. Since

$$\max_{\mathbf{z}\neq\mathbf{0}} \frac{\mathbf{z}'\mathbf{V}\mathbf{z}}{\mathbf{z}'\mathbf{z}} = \lambda_1 = \operatorname{ch}_1(\mathbf{V})\,,$$

the inequality (20.48) indeed holds. The equality in (20.48) means that

$$\lambda_1 \geq \frac{\mathbf{1}'\mathbf{V}\mathbf{1}}{\mathbf{1}'\mathbf{1}} = \lambda_1 + \lambda_2 + \cdots + \lambda_p\,,$$

which holds if and only if $\lambda_2 = \cdots = \lambda_p = 0$ and vector $\mathbf{1}$ is the eigenvector of $\mathbf{V}$ with respect to λ_1, i.e., $\mathbf{V}$ is of the form $\mathbf{V} = \lambda_1\mathbf{1}\mathbf{1}'$. □

An alternative proof of Proposition 20.4 appears in Vehkalahti (2000, Lemma 4.1). The right-hand term of (20.47), i.e.,

$$\frac{p}{p-1}\left(1 - \frac{\operatorname{tr}(\mathbf{V})}{\mathbf{1}'\mathbf{V}\mathbf{1}}\right) := \alpha(\mathbf{1})\,, \tag{20.49}$$

can be interpreted as Cronbach's *alpha*, see Cronbach (1951).

Below we give a related result, see Vehkalahti, Puntanen & Tarkkonen (2009, Th. 3).

Proposition 20.5. *Let* $\mathbf{V}$ *be a symmetric nonnegative definite* $p \times p$ *matrix with* $\mathbf{V}_\delta = \operatorname{diag}(\mathbf{V})$ *being positive definite. Then*

$$\max_{\mathbf{a}\neq\mathbf{0}} \frac{\mathbf{a}'\mathbf{V}\mathbf{a}}{\mathbf{a}'\mathbf{V}_\delta\mathbf{a}} = \operatorname{ch}_1(\mathbf{V}_\delta^{-1/2}\mathbf{V}\mathbf{V}_\delta^{-1/2}) = \operatorname{ch}_1(\mathbf{R}_V)\,, \tag{20.50}$$

where $\mathbf{R}_V = \mathbf{V}_\delta^{-1/2}\mathbf{V}\mathbf{V}_\delta^{-1/2}$*, i.e.,* $\mathbf{R}_V$ *can be considered as a correlation matrix. Moreover,*

$$\frac{\mathbf{a}'\mathbf{V}\mathbf{a}}{\mathbf{a}'\mathbf{V}_\delta\mathbf{a}} \leq p \quad \textit{for all } \mathbf{a} \in \mathbb{R}^p, \tag{20.51}$$

where the equality is obtained if and only if $\mathbf{V} = \delta^2\mathbf{q}\mathbf{q}'$ *for some* $\delta \in \mathbb{R}$ *and some* $\mathbf{q} = (q_1, \dots, q_p)'$*, and* $\mathbf{a}$ *is a multiple of* $\mathbf{a}_* = (1/q_1, \dots, 1/q_p)'$*.*

Proof. In view of Proposition 18.3 (p. 369),

$$\max_{\mathbf{a}\neq\mathbf{0}} \frac{\mathbf{a}'\mathbf{V}\mathbf{a}}{\mathbf{a}'\mathbf{V}_\delta\mathbf{a}} = \mathrm{ch}_1(\mathbf{V}_\delta^{-1}\mathbf{V}) = \mathrm{ch}_1(\mathbf{V}_\delta^{-1/2}\mathbf{V}\mathbf{V}_\delta^{-1/2}), \tag{20.52}$$

and thus (20.50) is confirmed. It is obvious that the largest eigenvalue μ_1 of a $p \times p$ correlation matrix $\mathbf{R}_V$ has property $\mu_1 \leq p$, and clearly $\mu_1 = p$ if and only if $\mathbf{R}_V = \mathbf{1}\mathbf{1}'$, i.e., $\mathbf{V}$ must be of the form $\mathbf{V} = \gamma^2\mathbf{q}\mathbf{q}'$ for some $\gamma \in \mathbb{R}$ and $\mathbf{q} = (q_1, \dots, q_p)' \in \mathbb{R}^p$. It is easy to conclude that if $\mathbf{V} = \gamma^2\mathbf{q}\mathbf{q}'$, then the equality in (20.51) is obtained if and only if $\mathbf{a}$ is a multiple of $\mathbf{a}_* = \mathbf{V}_\delta^{-1/2}\mathbf{1} = \frac{1}{\gamma}(1/q_1, \dots, 1/q_p)'$. □

20.7 Matrix Versions of the Cauchy–Schwarz and Kantorovich Inequality

Consider the matrices $\mathbf{A}_{n\times a}$ and $\mathbf{B}_{n\times b}$. Then the matrix $\mathbf{I}_n - \mathbf{P}_\mathbf{B}$ is an orthogonal projector and hence is nonnegative definite:

$$\mathbf{I}_n - \mathbf{B}(\mathbf{B}'\mathbf{B})^-\mathbf{B}' \geq_\mathsf{L} \mathbf{0}\,. \tag{20.53}$$

Moreover, (20.53) implies, cf. Chipman (1964, p. 1093), that

$$\mathbf{A}'\mathbf{B}(\mathbf{B}'\mathbf{B})^-\mathbf{B}'\mathbf{A} \leq_\mathsf{L} \mathbf{A}'\mathbf{A} \tag{20.54}$$

which is a matrix version of the Cauchy–Schwarz inequality. Equality in (20.54) holds (for a nonnull $\mathbf{B}$) if and only if

$$\mathscr{C}(\mathbf{A}) \subset \mathscr{C}(\mathbf{B})\,. \tag{20.55}$$

Note that from (20.54) we get the determinantal version (supposing that $p = q$) of the Cauchy–Schwarz inequality:

$$|\mathbf{A}'\mathbf{B}|^2 \leq |\mathbf{A}'\mathbf{A}| \cdot |\mathbf{B}'\mathbf{B}|\,. \tag{20.56}$$

Let then $\mathbf{V}$ be a symmetric nonnegative definite $n \times n$ matrix and $\mathbf{X}$ an $n \times p$ matrix. Substituting $\mathbf{A} = \mathbf{V}^{1/2}\mathbf{X}$ and $\mathbf{B} = (\mathbf{V}^+)^{1/2}\mathbf{X}$ into (20.54), Baksalary & Puntanen (1991, Th. 1) proved the inequality

$$\mathbf{X}'\mathbf{P}_{\mathbf{V}}\mathbf{X}(\mathbf{X}'\mathbf{V}\mathbf{X})^{-}\mathbf{X}'\mathbf{P}_{\mathbf{V}}\mathbf{X} \leq_{\mathsf{L}} \mathbf{X}'\mathbf{V}^{+}\mathbf{X}\,, \tag{20.57}$$

with equality if and only if

$$\mathscr{C}(\mathbf{V}\mathbf{X}) = \mathscr{C}(\mathbf{P}_{\mathbf{V}}\mathbf{X})\,. \tag{20.58}$$

Obviously we can express (20.57) equivalently as

$$\mathbf{X}'\mathbf{P}_{\mathbf{V}}\mathbf{X}(\mathbf{X}'\mathbf{V}^{+}\mathbf{X})^{-}\mathbf{X}'\mathbf{P}_{\mathbf{V}}\mathbf{X} \leq_{\mathsf{L}} \mathbf{X}'\mathbf{V}\mathbf{X}\,. \tag{20.59}$$

Let us then have a look at the matrix versions on the Kantorovich inequality. When $\mathbf{u}'\mathbf{u} = 1$, we may express the Kantorovich inequality as

$$\mathbf{u}'\mathbf{V}\mathbf{u} \leq \frac{(\lambda_1 + \lambda_n)^2}{4\lambda_1\lambda_n}(\mathbf{u}'\mathbf{V}^{-1}\mathbf{u})^{-1}, \tag{20.60}$$

which is more convenient for matrix extensions. Marshall & Olkin (1990) generalized the Kantorovich inequality by showing that in the Löwner partial ordering

$$\mathbf{U}'\mathbf{V}\mathbf{U} \leq_{\mathsf{L}} \frac{(\lambda_1 + \lambda_n)^2}{4\lambda_1\lambda_n}(\mathbf{U}'\mathbf{V}^{-1}\mathbf{U})^{-1}, \tag{20.61}$$

where the $n \times n$ matrix $\mathbf{V}$ is positive definite with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$, while the $n \times p$ matrix $\mathbf{U}$ satisfies $\mathbf{U}'\mathbf{U} = \mathbf{I}_p$. It is easy to see that $\mathbf{V}$ and $\mathbf{V}^{-1}$ in (20.61) may be interchanged and so we also find that

$$\mathbf{U}'\mathbf{V}^{-1}\mathbf{U} \leq_{\mathsf{L}} \frac{(\lambda_1 + \lambda_n)^2}{4\lambda_1\lambda_n}(\mathbf{U}'\mathbf{V}\mathbf{U})^{-1}. \tag{20.62}$$

Let $\mathbf{X}_{n\times p}$ has full column rank. Then, as pointed out on page 238,

$$\begin{aligned}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} &\leq_{\mathsf{L}} (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &\leq_{\mathsf{L}} \frac{(\lambda_1 + \lambda_n)^2}{4\lambda_1\lambda_n}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}.\end{aligned} \tag{20.63}$$

We may complete this section by considering the following result (see Lemma 1 of Drury, Liu, Lu, Puntanen et al. 2002):

Proposition 20.6. *Let* $\mathbf{V}$ *be an* $n \times n$ *nonnull nonnegative definite matrix and let the matrices* $\mathbf{X}$ *and* $\mathbf{Y}$ *be* $n \times p$ *and* $n \times q$*, respectively. Assume that equation*

$$\mathbf{X}'\mathbf{P}_{\mathbf{V}}\mathbf{Y} = \mathbf{0} \tag{20.64}$$

holds. Then

$$\mathbf{V} - \mathbf{P}_{\mathbf{V}}\mathbf{X}(\mathbf{X}'\mathbf{V}^{+}\mathbf{X})^{-}\mathbf{X}'\mathbf{P}_{\mathbf{V}} \geq_{\mathsf{L}} \mathbf{V}\mathbf{Y}(\mathbf{Y}'\mathbf{V}\mathbf{Y})^{-}\mathbf{Y}'\mathbf{V}\,, \tag{20.65}$$

and hence

$$\mathbf{X}'\mathbf{V}\mathbf{X} - \mathbf{X}'\mathbf{P}_{\mathbf{V}}\mathbf{X}(\mathbf{X}'\mathbf{V}^{+}\mathbf{X})^{-}\mathbf{X}'\mathbf{P}_{\mathbf{V}}\mathbf{X} \geq_{\mathsf{L}} \mathbf{X}'\mathbf{V}\mathbf{Y}(\mathbf{Y}'\mathbf{V}\mathbf{Y})^{-}\mathbf{Y}'\mathbf{V}\mathbf{X} \tag{20.66}$$

for any choices of generalized inverses $(\mathbf{X}'\mathbf{V}^{+}\mathbf{X})^{-}$ *and* $(\mathbf{Y}'\mathbf{V}\mathbf{Y})^{-}$*. Equality holds in* (20.65) *if and only if*

$$\operatorname{rank}(\mathbf{V}) = \operatorname{rank}(\mathbf{V}\mathbf{X}) + \operatorname{rank}(\mathbf{V}\mathbf{Y}), \tag{20.67}$$

and then equality holds in (20.66)*.*

Proof. To prove (20.65), we first denote $\mathbf{K} = (\mathbf{V}^{+})^{1/2}\mathbf{X}$ and $\mathbf{L} = \mathbf{V}^{1/2}\mathbf{Y}$, and so, in view of (20.64), $\mathbf{F}'\mathbf{L} = \mathbf{0}$, and

$$\mathbf{P}_{\mathbf{K}} + \mathbf{P}_{\mathbf{L}} = \mathbf{P}_{(\mathbf{K}\,:\,\mathbf{L})} := \mathbf{P}_{\mathbf{U}}\,. \tag{20.68}$$

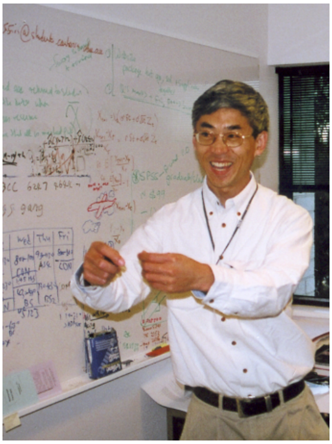
Substituting the explicit representations of the projectors into (20.68), yields

$$\begin{aligned}(\mathbf{V}^{+})^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{V}^{+}\mathbf{X})^{-}\mathbf{X}'(\mathbf{V}^{+})^{1/2} + \mathbf{V}^{1/2}\mathbf{Y}(\mathbf{Y}'\mathbf{V}\mathbf{Y})^{-}\mathbf{Y}'\mathbf{V}^{1/2}& \\ = \mathbf{P}_{\mathbf{U}} \leq_{\mathsf{L}} \mathbf{P}_{\mathbf{V}}\,,&\end{aligned} \tag{20.69}$$

where the last Löwner ordering follows from the column space inclusion $\mathscr{C}(\mathbf{U}) \subset \mathscr{C}(\mathbf{V})$. Pre- and postmultiplying (20.69) with $\mathbf{V}^{1/2}$ gives (20.65) at once. Moreover, the equality in (20.65) holds if and only if

$$\mathscr{C}(\mathbf{U}) = \mathscr{C}(\mathbf{K} : \mathbf{L}) = \mathscr{C}(\mathbf{V})\,, \tag{20.70}$$

which yields (20.67). □



Photograph 20.1 Shuangzhe Liu (Canberra, 2005).

For further matrix versions of the Kantorovich and Wielandt inequalities, see Alpargu, Drury & Styan (1997), Alpargu & Styan (2000), Baksalary & Puntanen (1991), and Drury, Liu, Lu, Puntanen et al. (2002), Khatri & Rao (1981, 1982), Liu & King (2002), Liu & Neudecker (1995, 1997, 1999), Liu (2000b), Pečarić, Puntanen & Styan (1996), Wang & Shao (1992), and Rao (2007). For matrix trace Wielandt inequalities with statistical applications, see Liu, Lu & Puntanen (2009).

20.8 Exercises

20.1. Prove the Proposition 20.3 (p. 421).

20.2. Prove (20.57) (p. 424):

$$\mathbf{X}'\mathbf{P}_{\mathbf{V}}\mathbf{X}(\mathbf{X}'\mathbf{V}^{+}\mathbf{X})^{-}\mathbf{X}'\mathbf{P}_{\mathbf{V}}\mathbf{X} \leq_{\mathsf{L}} \mathbf{X}'\mathbf{V}\mathbf{X}\,,$$

with equality if and only if $\mathscr{C}(\mathbf{V}\mathbf{X}) = \mathscr{C}(\mathbf{P}_{\mathbf{V}}\mathbf{X})$.

20.3. Choosing $\mathbf{X} = (\mathbf{I}_p : \mathbf{0})'$ in (20.63) (p. 424) confirm that

$$(\mathbf{V}^{11})^{-1} \leq_{\mathsf{L}} \mathbf{V}_{11} \leq_{\mathsf{L}} \frac{(\lambda_1 + \lambda_n)^2}{4\lambda_1\lambda_n}(\mathbf{V}^{11})^{-1},$$

$$\mathbf{V}_{11}^{-1} \leq_{\mathsf{L}} \mathbf{V}^{11} \leq_{\mathsf{L}} \frac{(\lambda_1 + \lambda_n)^2}{4\lambda_1\lambda_n}\mathbf{V}_{11}^{-1},$$

where $\mathbf{V}$ is positive definite and $\mathbf{V}^{-1} = \begin{pmatrix} \mathbf{V}^{11} & \mathbf{V}^{12} \\ \mathbf{V}^{21} & \mathbf{V}^{22} \end{pmatrix}$.

20.4. Confirm that the upper bound in (10.143b) (p. 238), cf. (20.63) (p. 424, i.e., in

$$\operatorname{cov}(\tilde{\boldsymbol{\beta}}) \leq_{\mathsf{L}} \operatorname{cov}(\hat{\boldsymbol{\beta}}) \leq_{\mathsf{L}} \frac{(\lambda_1 + \lambda_n)^2}{4\lambda_1\lambda_n}\operatorname{cov}(\tilde{\boldsymbol{\beta}})\,,$$

is attained, for example, when

$$\mathbf{X} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \\ 1 & 1 \\ -1 & 1 \end{pmatrix}, \quad \mathbf{V} = \begin{pmatrix} 1 & r & 0 & 0 \\ r & 1 & 0 & 0 \\ 0 & 0 & 1 & r \\ 0 & 0 & r & 1 \end{pmatrix}, \quad 0 < r < 1\,.$$

Magness & McGuire (1962).

20.5 (A matrix version of the Kantorovich inequality). Let $\mathbf{V}$ be a nonnegative definite matrix with $\operatorname{rank}(\mathbf{V}) = v > 1$, and $\mathbf{X} \in \mathbb{R}^{n\times p}$. Show that then

$$\mathbf{X}'\mathbf{V}\mathbf{X} \leq_{\mathsf{L}} \frac{(\lambda_1 + \lambda_v)^2}{4\lambda_1\lambda_v}\mathbf{X}'\mathbf{P}_{\mathbf{V}}\mathbf{X}(\mathbf{X}'\mathbf{V}^{+}\mathbf{X})^{-}\mathbf{X}'\mathbf{P}_{\mathbf{V}}\mathbf{X}\,.$$

Notice that in (20.59) (p. 424) we have the lower bound for $\mathbf{X}'\mathbf{V}\mathbf{X}$.

Drury, Liu, Lu, Puntanen et al. (2002, Th. 1).

20.6 (A matrix version of the Wielandt inequality). Let $\mathbf{V}$ be positive definite, and let $\mathbf{X} \in \mathbb{R}^{n\times p}$, $\mathbf{Y} \in \mathbb{R}^{n\times q}$ satisfying $\mathbf{X}'\mathbf{Y} = \mathbf{0}$. Show that then

$$\mathbf{X}'\mathbf{V}\mathbf{Y}(\mathbf{Y}'\mathbf{V}\mathbf{Y})^{-}\mathbf{Y}'\mathbf{V}\mathbf{X} \leq_{\mathsf{L}} \left(\frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n}\right)^2 \mathbf{X}'\mathbf{V}\mathbf{X}\,.$$

Observe that if $n = p + q$ and $\mathbf{X} = (\mathbf{I}_p : \mathbf{0})'$ and $\mathbf{Y} = (\mathbf{0} : \mathbf{I}_{n-p})'$, we get

$$\mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}_{21} \leq_{\mathsf{L}} \left(\frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n}\right)^2 \mathbf{V}_{11}\,.$$

Drury, Liu, Lu, Puntanen et al. (2002, Th. 2), Wang & Ip (1999, Th. 1).

Notation

$\mathbb{R}$	real numbers
$\mathbb{R}^{n\times m}$	set of $n \times m$ real matrices
$\mathbb{R}_r^{n\times m}$	subset of $\mathbb{R}^{n\times m}$ consisting of matrices with rank r
$\mathbb{R}_s^n$	subset of $\mathbb{R}^{n\times n}$ consisting of symmetric matrices
NND_n	subset of $\mathbb{R}_s^n$ consisting of nonnegative definite (nnd) matrices: $\mathbf{A} \in \mathrm{NND}_n \iff \mathbf{A} = \mathbf{LL}'$ for some $\mathbf{L}$; instead of nnd, the term positive semidefinite is often used
PD_n	subset of NND_n consisting of positive definite (pd) matrices: $\mathbf{A} = \mathbf{LL}'$ for some nonsingular $\mathbf{L}$
$\mathbf{0}$	null vector, null matrix; denoted also as $\mathbf{0}_n$ or $\mathbf{0}_{n\times m}$
$\mathbf{1}_n$	column vector of ones, shortened $\mathbf{1}$
$\mathbf{I}_n$	identity matrix, shortened $\mathbf{I}$
$\mathbf{i}_j$	the jth column of $\mathbf{I}$; the jth standard basis vector
$\mathbf{A} = \{a_{ij}\}$	matrix $\mathbf{A}$ with its elements a_{ij}
$\mathbf{A}_{n\times m}$	$n \times m$ matrix $\mathbf{A}$
$\mathbf{a}$	column vector $\mathbf{a} \in \mathbb{R}^n$
$\mathbf{A}'$	transpose of the matrix $\mathbf{A}$
$(\mathbf{A} : \mathbf{B})$	partitioned (augmented) matrix
$\mathbf{A} = (\mathbf{a}_1 : \ldots : \mathbf{a}_m)$	$\mathbf{A}_{n\times m}$ represented columnwise
$\mathbf{A} = \begin{pmatrix} \mathbf{a}'_{(1)} \\ \vdots \\ \mathbf{a}'_{(n)} \end{pmatrix}$	$\mathbf{A}_{n\times m}$ represented row-wise
$\mathbf{A}^{-1}$	inverse of the matrix $\mathbf{A}$
$\mathbf{A}^{-}$	generalized inverse of the matrix $\mathbf{A}$: $\mathbf{AA}^{-}\mathbf{A} = \mathbf{A}$, also called $\{1\}$-inverse, or inner inverse
$\{\mathbf{A}^{-}\}$	the set of generalized inverses of $\mathbf{A}$

S. Puntanen et al., *Matrix Tricks for Linear Statistical Models: Our Personal Top Twenty*,
DOI 10.1007/978-3-642-10473-2, © Springer-Verlag Berlin Heidelberg 2011

$\mathbf{A}_{12}^{-}$	reflexive generalized inverse of $\mathbf{A}$: $\mathbf{A}\mathbf{A}^{-}\mathbf{A} = \mathbf{A}$, $\mathbf{A}^{-}\mathbf{A}\mathbf{A}^{-} = \mathbf{A}^{-}$, also called $\{12\}$-inverse
$\mathbf{A}^{+}$	the Moore–Penrose inverse of $\mathbf{A}$: the unique matrix satisfying the four Moore–Penrose conditions: (mp1) $\mathbf{A}\mathbf{A}^{-}\mathbf{A} = \mathbf{A}$, (mp2) $\mathbf{A}^{-}\mathbf{A}\mathbf{A}^{-} = \mathbf{A}^{-}$, (mp3) $(\mathbf{A}\mathbf{A}^{-})' = \mathbf{A}\mathbf{A}^{-}$, (mp4) $(\mathbf{A}^{-}\mathbf{A})' = \mathbf{A}^{-}\mathbf{A}$
$\mathbf{A}_{ij}^{-}$	generalized inverse of $\mathbf{A}$ satisfying the Moore–Penrose conditions (mp$i$) and (mp$j$)
$\mathbf{A}^{1/2}$	symmetric nnd square root of $\mathbf{A} \in \mathrm{NND}_n$: $\mathbf{A}^{1/2} = \mathbf{T}\boldsymbol{\Lambda}^{1/2}\mathbf{T}'$, where $\mathbf{A} = \mathbf{T}\boldsymbol{\Lambda}\mathbf{T}'$ is the eigenvalue decomposition of $\mathbf{A}$
$\mathbf{A}^{+1/2}$	$(\mathbf{A}^{+})^{1/2}$
$\mathrm{In}(\mathbf{A}) = (\pi, \nu, \delta)$	inertia of the square matrix $\mathbf{A}$: π, ν, and δ are the number of positive, negative, and zero eigenvalues of $\mathbf{A}$, respectively, all counting multiplicities
$\langle \mathbf{a}, \mathbf{b} \rangle$	standard inner product in $\mathbb{R}^n$: $\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a}'\mathbf{b}$; can denote also a general inner product in a vector space
$\langle \mathbf{a}, \mathbf{b} \rangle_{\mathbf{V}}$	inner product $\mathbf{a}'\mathbf{V}\mathbf{b}$; $\mathbf{V}$ is the inner product matrix (ipm)
$\mathbf{a} \perp \mathbf{b}$	vectors $\mathbf{a}$ and $\mathbf{b}$ are orthogonal with respect to a given inner product
$\Vert\mathbf{a}\Vert$	Euclidean norm (standard norm, 2-norm) of vector $\mathbf{a}$, also denoted $\Vert\mathbf{a}\Vert_2$: $\Vert\mathbf{a}\Vert^2 = \mathbf{a}'\mathbf{a}$; can denote also a general vector norm in a vector space
$\Vert\mathbf{a}\Vert_{\mathbf{V}}$	$\Vert\mathbf{a}\Vert_{\mathbf{V}}^2 = \mathbf{a}'\mathbf{V}\mathbf{a}$, norm when the ipm is $\mathbf{V}$ (ellipsoidal norm)
$\langle \mathbf{A}, \mathbf{B} \rangle$	standard matrix inner product between $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times m}$: $\langle \mathbf{A}, \mathbf{B} \rangle = \mathrm{tr}(\mathbf{A}'\mathbf{B}) = \sum_{i,j} a_{ij} b_{ij}$
$\Vert\mathbf{A}\Vert_F$	Euclidean (Frobenius) norm of the matrix $\mathbf{A}$: $\Vert\mathbf{A}\Vert_F^2 = \mathrm{tr}(\mathbf{A}'\mathbf{A}) = \sum_{i,j} a_{ij}^2$
$\Vert\mathbf{A}\Vert_2$	matrix 2-norm of the matrix $\mathbf{A}$ (spectral norm): $\Vert\mathbf{A}\Vert_2 = \max\limits_{\Vert\mathbf{x}\Vert_2 = 1} \Vert\mathbf{A}\mathbf{x}\Vert_2 = \mathrm{sg}_1(\mathbf{A}) = +\sqrt{\mathrm{ch}_1(\mathbf{A}'\mathbf{A})}$
$\Vert\mathbf{A}^{-1}\Vert_2$	matrix 2-norm of nonsingular $\mathbf{A}_{n \times n}$: $\Vert\mathbf{A}^{-1}\Vert_2 = 1/\mathrm{sg}_n(\mathbf{A})$
$\mathrm{cond}(\mathbf{A})$	condition number of nonsingular $\mathbf{A}_{n \times n}$: $\mathrm{cond}(\mathbf{A}) = \Vert\mathbf{A}\Vert_2 \Vert\mathbf{A}^{-1}\Vert_2 = \mathrm{sg}_1(\mathbf{A})/\mathrm{sg}_n(\mathbf{A})$
$\cos(\mathbf{a}, \mathbf{b})$	$\cos\angle(\mathbf{a}, \mathbf{b})$, the cosine of the angle, θ, between the nonzero vectors $\mathbf{a}$ and $\mathbf{b}$: $\cos(\mathbf{a}, \mathbf{b}) = \cos\theta = \cos\angle(\mathbf{a}, \mathbf{b}) = \frac{\langle \mathbf{a}, \mathbf{b} \rangle}{\Vert\mathbf{a}\Vert \Vert\mathbf{b}\Vert}$
$\angle(\mathbf{a}, \mathbf{b})$	the angle, θ, $0 \leq \theta \leq \pi$, between the nonzero vectors $\mathbf{a}$ and $\mathbf{b}$: $\theta = \angle(\mathbf{a}, \mathbf{b}) = \cos^{-1}(\mathbf{a}, \mathbf{b})$
$\mathbf{A}[\alpha, \beta]$	submatrix of $\mathbf{A}_{n \times n}$, obtained by choosing the elements of $\mathbf{A}$ which lie in rows α and columns β; α and β are index sets of the rows and the columns of $\mathbf{A}$, respectively
$\mathbf{A}[\alpha]$	$\mathbf{A}[\alpha, \alpha]$, principal submatrix; same rows and columns chosen

$\mathbf{A}_i^{\mathrm{L}}$	ith leading principal submatrix of $\mathbf{A}_{n\times n}$: $\mathbf{A}_i^{\mathrm{L}} = \mathbf{A}[\alpha,\alpha]$, where $\alpha = \{1,\dots,i\}$
$\mathbf{A}(\alpha,\beta)$	submatrix of $\mathbf{A}$, obtained by choosing the elements of $\mathbf{A}$ which do not lie in rows α and columns β
$\mathbf{A}(i,j)$	submatrix of $\mathbf{A}$, obtained by deleting row i and column j from $\mathbf{A}$
$\operatorname{minor}(a_{ij})$	ijth minor of $\mathbf{A}$ corresponding to a_{ij}: $\operatorname{minor}(a_{ij}) = \det(\mathbf{A}(i,j))$, $i,j \in \{1,\dots,n\}$
$\operatorname{cof}(a_{ij})$	ijth cofactor of $\mathbf{A}$: $\operatorname{cof}(a_{ij}) = (-1)^{i+j}\operatorname{minor}(a_{ij})$
$\det(\mathbf{A})$	determinant of the matrix $\mathbf{A}_{n\times n}$: $\det(a) = a$, $a \in \mathbb{R}$, $\det(\mathbf{A}) = \sum_{j=1}^{n} a_{ij}\operatorname{cof}(a_{ij})$, $i \in \{1,\dots,n\}$: the Laplace expansion by minors along the ith row
$\det(\mathbf{A}[\alpha])$	principal minor
$\det(\mathbf{A}_i^{\mathrm{L}})$	leading principal minor of order i
$\lvert\mathbf{A}\rvert$	determinant of the matrix $\mathbf{A}_{n\times n}$
$\operatorname{diag}(\mathbf{A})$	diagonal matrix formed by the diagonal entries of $\mathbf{A}_{n\times n}$
$\operatorname{diag}(d_1,\dots,d_n)$	$n\times n$ diagonal matrix with listed diagonal entries
$\operatorname{diag}(\mathbf{d})$	$n\times n$ diagonal matrix whose ith diagonal element is d_i
$\mathbf{A}_\delta$	diagonal matrix formed by the diagonal entries of $\mathbf{A}_{n\times n}$
$\operatorname{rk}(\mathbf{A})$	rank of the matrix $\mathbf{A}$
$\operatorname{rank}(\mathbf{A})$	rank of the matrix $\mathbf{A}$
$\operatorname{tr}(\mathbf{A})$	trace of the matrix $\mathbf{A}_{n\times n}$: $\operatorname{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$
$\operatorname{trace}(\mathbf{A})$	trace of the matrix $\mathbf{A}_{n\times n}$
$\operatorname{vec}(\mathbf{A})$	vectoring operation: the vector formed by placing the columns of $\mathbf{A}$ under one another successively
$\mathbf{A}\otimes\mathbf{B}$	Kronecker product of $\mathbf{A}_{n\times m}$ and $\mathbf{B}_{p\times q}$: $\mathbf{A}\otimes\mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & \dots & a_{1m}\mathbf{B} \\ \vdots & \vdots & \vdots \\ a_{n1}\mathbf{B} & \dots & a_{nm}\mathbf{B} \end{pmatrix} \in \mathbb{R}^{np\times mq}$
$\mathbf{A}/\mathbf{A}_{11}$	Schur complement of $\mathbf{A}_{11}$ in $\mathbf{A} = \left(\begin{smallmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{smallmatrix}\right)$: $\mathbf{A}/\mathbf{A}_{11} = \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-}\mathbf{A}_{12}$
$\mathbf{A}_{22\cdot 1}$	$\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-}\mathbf{A}_{12}$
$\mathbf{A} \geq_{\mathsf{L}} \mathbf{0}$	$\mathbf{A}$ is nonnegative definite: $\mathbf{A} = \mathbf{L}\mathbf{L}'$ for some $\mathbf{L}$; $\mathbf{A} \in \mathrm{NND}_n$
$\mathbf{A} >_{\mathsf{L}} \mathbf{0}$	$\mathbf{A}$ is positive definite: $\mathbf{A} = \mathbf{L}\mathbf{L}'$ for some invertible $\mathbf{L}$; $\mathbf{A} \in \mathrm{PD}_n$
$\mathbf{A} \leq_{\mathsf{L}} \mathbf{B}$	$\mathbf{B} - \mathbf{A}$ is nonnegative definite; $\mathbf{B} - \mathbf{A} \in \mathrm{NND}_n$; $\mathbf{A}$ lies below $\mathbf{B}$ with respect to the Löwner ordering
$\mathbf{A} <_{\mathsf{L}} \mathbf{B}$	$\mathbf{B} - \mathbf{A}$ is positive definite; $\mathbf{B} - \mathbf{A} \in \mathrm{PD}_n$
$\mathbf{A} \leq_{\mathsf{rs}} \mathbf{B}$	$\mathbf{A}$ and $\mathbf{B}$ are rank-subtractive; $\operatorname{rk}(\mathbf{B}-\mathbf{A}) = \operatorname{rk}(\mathbf{B}) - \operatorname{rk}(\mathbf{A})$; $\mathbf{A}$ lies below $\mathbf{B}$ with respect to the minus ordering

$\mathrm{Sh}(\mathbf{V} \mid \mathbf{X})$	the shorted matrix of $\mathbf{V} \in \mathrm{NND}_n$ with respect to $\mathbf{X}_{n\times p}$, $\mathrm{Sh}(\mathbf{V} \mid \mathbf{X})$ is the maximal element $\mathbf{U}$ (in the Löwner ordering) in the set $\mathcal{U} = \{\, \mathbf{U} : \mathbf{0} \leq_{\mathsf{L}} \mathbf{U} \leq_{\mathsf{L}} \mathbf{V},\ \mathscr{C}(\mathbf{U}) \subset \mathscr{C}(\mathbf{X}) \,\}$
$\mathbf{P}_{\mathbf{A}}$	orthogonal projector onto $\mathscr{C}(\mathbf{A})$ (w.r.t. $\mathbf{I}$): $\mathbf{P}_{\mathbf{A}} = \mathbf{A}(\mathbf{A}'\mathbf{A})^{-}\mathbf{A}' = \mathbf{A}\mathbf{A}^{+}$
$\mathbf{P}_{\mathbf{A};\mathbf{V}}$	orthogonal projector onto $\mathscr{C}(\mathbf{A})$ w.r.t. $\mathbf{V} \in \mathrm{PD}_n$: $\mathbf{P}_{\mathbf{A};\mathbf{V}} = \mathbf{A}(\mathbf{A}'\mathbf{V}\mathbf{A})^{-}\mathbf{A}'\mathbf{V}$
$\mathbf{P}_{\mathbf{A};\mathbf{V}}$	generalized orthogonal projector onto $\mathscr{C}(\mathbf{A})$ w.r.t. $\mathbf{V} \in \mathrm{NND}_n$: $\mathbf{P}_{\mathbf{A};\mathbf{V}} = \mathbf{A}(\mathbf{A}'\mathbf{V}\mathbf{A})^{-}\mathbf{A}'\mathbf{V} + \mathbf{A}[\mathbf{I} - (\mathbf{A}'\mathbf{V}\mathbf{A})^{-}\mathbf{A}'\mathbf{V}\mathbf{A}]\mathbf{U}$, where $\mathbf{U}$ is arbitrary
$\mathbf{P}_{\mathbf{A}\mid\mathbf{B}}$	projector onto $\mathscr{C}(\mathbf{A})$ along $\mathscr{C}(\mathbf{B})$: $\mathbf{P}_{\mathbf{A}\mid\mathbf{B}}(\mathbf{A} : \mathbf{B}) = (\mathbf{A} : \mathbf{0})$
$\{\mathbf{P}_{\mathbf{A}\mid\mathbf{B}}\}$	set of matrices satisfying: $\mathbf{P}_{\mathbf{A}\mid\mathbf{B}}(\mathbf{A} : \mathbf{B}) = (\mathbf{A} : \mathbf{0})$
$\mathbf{P}_{\mathcal{U}}$	orthogonal projector onto the vector space $\mathcal{U}$ (w.r.t. a given inner product)
$\mathscr{C}(\mathbf{A})$	column space of the matrix $\mathbf{A}_{n\times p}$: $\mathscr{C}(\mathbf{A}) = \{\, \mathbf{y} \in \mathbb{R}^n : \mathbf{y} = \mathbf{A}\mathbf{x} \text{ for some } \mathbf{x} \in \mathbb{R}^p \,\}$
$\mathscr{N}(\mathbf{A})$	null space of the matrix $\mathbf{A}_{n\times p}$: $\mathscr{N}(\mathbf{A}) = \{\, \mathbf{x} \in \mathbb{R}^p : \mathbf{A}\mathbf{x} = \mathbf{0} \,\}$
$\mathscr{C}(\mathbf{A})^{\perp}$	orthocomplement of $\mathscr{C}(\mathbf{A})$ w.r.t. $\mathbf{I}$: $\mathscr{C}(\mathbf{A})^{\perp} = \{\, \mathbf{z} \in \mathbb{R}^n : \mathbf{z}'\mathbf{A}\mathbf{x} = \mathbf{0}\ \forall\, \mathbf{x} \in \mathbb{R}^p \,\} = \mathscr{N}(\mathbf{A}')$
$\mathbf{A}^{\perp}$	matrix whose column space is $\mathscr{C}(\mathbf{A}^{\perp}) = \mathscr{C}(\mathbf{A})^{\perp}$
$\mathscr{C}(\mathbf{A})^{\perp}_{\mathbf{V}}$	orthocomplement of $\mathscr{C}(\mathbf{A})$ w.r.t. $\mathbf{V}$: $\mathscr{C}(\mathbf{A})^{\perp}_{\mathbf{V}} = \{\, \mathbf{z} \in \mathbb{R}^n : \mathbf{z}'\mathbf{V}\mathbf{A}\mathbf{x} = \mathbf{0}\ \forall\, \mathbf{x} \in \mathbb{R}^p \,\} = \mathscr{N}(\mathbf{A}'\mathbf{V})$
$\mathbf{A}^{\perp}_{\mathbf{V}}$	matrix whose column space is $\mathscr{C}(\mathbf{A})^{\perp}_{\mathbf{V}}$
$p_{\mathbf{A}}(x)$	the characteristic polynomial of $\mathbf{A}$: $p_{\mathbf{A}}(x) = \det(\mathbf{A} - x\mathbf{I})$
$\mathcal{U} \subset \mathcal{V}$	$\mathcal{U}$ is a subset of $\mathcal{V}$; possibly $\mathcal{U} = \mathcal{V}$
$\mathcal{U} + \mathcal{V}$	sum of the vector spaces $\mathcal{U}$ and $\mathcal{V}$
$\mathcal{U} \oplus \mathcal{V}$	direct sum of the vector spaces $\mathcal{U}$ and $\mathcal{V}$
$\mathcal{U} \boxplus \mathcal{V}$	direct sum of the orthogonal vector spaces $\mathcal{U}$ and $\mathcal{V}$
$\mathcal{U} \cap \mathcal{V}$	intersection of the vector spaces $\mathcal{U}$ and $\mathcal{V}$
$\mathrm{ch}_i(\mathbf{A}) = \lambda_i$	the ith largest eigenvalue of $\mathbf{A}_{n\times n}$ (all eigenvalues being real)
$\mathrm{ch}(\mathbf{A})$	set of all n eigenvalues of $\mathbf{A}_{n\times n}$, including multiplicities, called also the spectrum of $\mathbf{A}$: $\mathrm{ch}(\mathbf{A}) = \{\mathrm{ch}_1(\mathbf{A}), \dots, \mathrm{ch}_n(\mathbf{A})\}$
$\mathrm{ch}(\mathbf{A}, \mathbf{B})$	set of proper eigenvalues of symmetric $\mathbf{A}_{n\times n}$ with respect to $\mathbf{B} \in \mathrm{NND}_n$; $\lambda \in \mathrm{ch}(\mathbf{A}, \mathbf{B})$ if $\mathbf{A}\mathbf{w} = \lambda\mathbf{B}\mathbf{w}, \mathbf{B}\mathbf{w} \neq \mathbf{0}$
$\mathrm{nzch}(\mathbf{A})$	set of the nonzero eigenvalues of $\mathbf{A}_{n\times n}$: $\mathrm{nzch}(\mathbf{A}) = \{\mathrm{ch}_1(\mathbf{A}), \dots, \mathrm{ch}_r(\mathbf{A})\}$, $r = \mathrm{rank}(\mathbf{A})$
$\mathrm{chv}_i(\mathbf{A})$	eigenvector of $\mathbf{A}_{n\times n}$ with respect to $\lambda_i = \mathrm{ch}_i(\mathbf{A})$: a nonzero vector $\mathbf{t}_i$ satisfying the equation $\mathbf{A}\mathbf{t}_i = \lambda_i\mathbf{t}_i$
$\mathrm{sg}_i(\mathbf{A}) = \delta_i$	the ith largest singular value of $\mathbf{A}_{n\times m}$: $\mathrm{sg}_i(\mathbf{A}) = +\sqrt{\mathrm{ch}_i(\mathbf{A}'\mathbf{A})} = +\sqrt{\mathrm{ch}_i(\mathbf{A}\mathbf{A}')}$

$\mathrm{sg}(\mathbf{A})$	set of the singular values of $\mathbf{A}_{n\times m}$ $(m \leq n)$: $\mathrm{sg}(\mathbf{A}) = \{\mathrm{sg}_1(\mathbf{A}), \dots, \mathrm{sg}_m(\mathbf{A})\}$
$\mathrm{nzsg}(\mathbf{A})$	set of the nonzero singular values of $\mathbf{A}_{n\times m}$: $\mathrm{nzsg}(\mathbf{A}) = \{\mathrm{sg}_1(\mathbf{A}), \dots, \mathrm{sg}_r(\mathbf{A})\}$, $r = \mathrm{rank}(\mathbf{A})$
$\rho(\mathbf{A})$	the spectral radius of $\mathbf{A}_{n\times n}$: the maximum of the absolute values of the eigenvalues of $\mathbf{A}_{n\times n}$
$\mathrm{var_s}(y)$	sample variance of the variable y
$\mathrm{var_d}(\mathbf{y}) = s_y^2$	sample variance: argument is the variable vector $\mathbf{y} \in \mathbb{R}^n$: $\mathrm{var_d}(\mathbf{y}) = \frac{1}{n-1}\mathbf{y}'\mathbf{C}\mathbf{y} = \frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})^2$
$\mathrm{cov_s}(x, y)$	sample covariance between the variables x and y
$\mathrm{cov_d}(\mathbf{x}, \mathbf{y}) = s_{xy}$	sample covariance: arguments are variable vectors $\in \mathbb{R}^n$: $\mathrm{cov_d}(\mathbf{x}, \mathbf{y}) = \frac{1}{n-1}\mathbf{x}'\mathbf{C}\mathbf{y} = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x}_i)(y_i - \bar{y})$
$\mathrm{cor_d}(\mathbf{x}, \mathbf{y}) = r_{xy}$	sample correlation: $r_{xy} = \mathbf{x}'\mathbf{C}\mathbf{y}/\sqrt{\mathbf{x}'\mathbf{C}\mathbf{x} \cdot \mathbf{y}'\mathbf{C}\mathbf{y}} = \cos(\mathbf{C}\mathbf{x}, \mathbf{C}\mathbf{y})$
$\bar{\bar{\mathbf{x}}}$	projection of $\mathbf{x}$ onto $\mathscr{C}(\mathbf{1}_n)$: $\bar{\bar{\mathbf{x}}} = \mathbf{J}\mathbf{x} = \bar{x}\mathbf{1}_n$
$\tilde{\mathbf{x}}$	centered $\mathbf{x}$: $\tilde{\mathbf{x}} = \mathbf{C}\mathbf{x} = \mathbf{x} - \mathbf{J}\mathbf{x} = \mathbf{x} - \bar{x}\mathbf{1}_n$
$\mathbf{U}$	$n \times d$ data matrix of the u-variables:

$$\mathbf{U} = (\mathbf{u}_1 : \dots : \mathbf{u}_d) = \begin{pmatrix} \mathbf{u}'_{(1)} \\ \vdots \\ \mathbf{u}'_{(n)} \end{pmatrix}$$

$\mathbf{u}_1, \dots, \mathbf{u}_d$	"variable vectors" in "variable space" $\mathbb{R}^n$
$\mathbf{u}_{(1)}, \dots, \mathbf{u}_{(n)}$	"observation vectors" in "observation space" $\mathbb{R}^d$
$\bar{\mathbf{u}}$	vector of means of the variables $u_1, \dots, u_d$: $\bar{\mathbf{u}} = (\bar{u}_1, \dots, \bar{u}_d)'$
$\widetilde{\mathbf{U}}$	centered $\mathbf{U}$: $\widetilde{\mathbf{U}} = \mathbf{C}\mathbf{U}$, $\mathbf{C}$ is the centering matrix
$\tilde{\mathbf{u}}_1, \dots, \tilde{\mathbf{u}}_d$	centered variable vectors
$\tilde{\mathbf{u}}_{(1)}, \dots, \tilde{\mathbf{u}}_{(n)}$	centered observation vectors
$\mathrm{var_d}(\mathbf{u}_i) = s_i^2$	sample variance: argument is the variable vector $\mathbf{u}_i \in \mathbb{R}^n$: $\mathrm{var_d}(\mathbf{u}_i) = \frac{1}{n-1}\mathbf{u}'_i\mathbf{C}\mathbf{u}_i = \frac{1}{n-1}\sum_{\ell=1}^n (u_{\ell i} - \bar{u}_i)^2$
$\mathrm{cov_d}(\mathbf{u}_i, \mathbf{u}_j) = s_{ij}$	sample covariance: arguments are variable vectors $\in \mathbb{R}^n$: $s_{ij} = \frac{1}{n-1}\mathbf{u}'_i\mathbf{C}\mathbf{u}_j = \frac{1}{n-1}\sum_{\ell=1}^n (u_{\ell i} - \bar{u}_i)(u_{\ell j} - \bar{u}_j)$
$\mathrm{ssp}(\mathbf{U}) = \{t_{ij}\}$	matrix $\mathbf{T}$ $(d \times d)$ of the sums of squares and products of deviations about the mean: $\mathbf{T} = \mathbf{U}'\mathbf{C}\mathbf{U} = \sum_{i=1}^n (\mathbf{u}_{(i)} - \bar{\mathbf{u}})(\mathbf{u}_{(i)} - \bar{\mathbf{u}})'$
$\mathrm{cov_d}(\mathbf{U}) = \{s_{ij}\}$	sample covariance matrix $\mathbf{S}$ $(d \times d)$ of the data matrix $\mathbf{U}$: $\mathbf{S} = \frac{1}{n-1}\mathbf{T} = \frac{1}{n-1}\sum_{i=1}^n (\mathbf{u}_{(i)} - \bar{\mathbf{u}})(\mathbf{u}_{(i)} - \bar{\mathbf{u}})'$
$\mathrm{cor_d}(\mathbf{u}_i, \mathbf{u}_j) = r_{ij}$	sample correlation: arguments are variable vectors $\in \mathbb{R}^n$
$\mathrm{cor_d}(\mathbf{U}) = \{r_{ij}\}$	sample correlation matrix $\mathbf{R}$ $(d \times d)$ of the data matrix $\mathbf{U}$: $\mathbf{R} = \mathrm{cor_d}(\mathbf{U}) = (\mathrm{diag}\,\mathbf{S})^{-1/2}\mathbf{S}(\mathrm{diag}\,\mathbf{S})^{-1/2}$
$\mathrm{MHLN}^2(\mathbf{u}_{(i)}, \bar{\mathbf{u}}, \mathbf{S})$	sample Mahalanobis distance (squared) of the ith observation from the mean: $\mathrm{MHLN}^2(\mathbf{u}_{(i)}, \bar{\mathbf{u}}, \mathbf{S}) = (\mathbf{u}_{(i)} - \bar{\mathbf{u}})'\mathbf{S}^{-1}(\mathbf{u}_{(i)} - \bar{\mathbf{u}})$

$\mathrm{MHLN}^2(\bar{\mathbf{u}}_i, \bar{\mathbf{u}}_j, \mathbf{S}_*)$	sample Mahalanobis distance (squared) between two mean vectors: $\mathrm{MHLN}^2(\bar{\mathbf{u}}_i, \bar{\mathbf{u}}_j, \mathbf{S}_*) = (\bar{\mathbf{u}}_i - \bar{\mathbf{u}}_j)'\mathbf{S}_*^{-1}(\bar{\mathbf{u}}_i - \bar{\mathbf{u}}_j)$, where $\mathbf{S}_* = \frac{1}{n_1+n_2-2}(\mathbf{U}_1'\mathbf{C}_{n_1}\mathbf{U}_1 + \mathbf{U}_2'\mathbf{C}_{n_2}\mathbf{U}_2)$
$\mathrm{MHLN}^2(\mathbf{u}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$	population Mahalanobis distance squared: $\mathrm{MHLN}^2(\mathbf{u}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\mathbf{u} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{u} - \boldsymbol{\mu})$
$\mathrm{E}(\cdot)$	expectation of a random argument: $\mathrm{E}(x) = p_1x_1 + \cdots + p_kx_k$ if x is a discrete random variable whose values are $x_1, \dots, x_k$ with corresponding probabilities $p_1, \dots, p_k$
$\mathrm{var}(x) = \sigma_x^2$	variance of the random variable x: $\sigma_x^2 = \mathrm{E}(x - \mu_x)^2$, $\mu_x = \mathrm{E}(x)$
$\mathrm{cov}(x, y) = \sigma_{xy}$	covariance between the random variables x and y: $\sigma_{xy} = \mathrm{E}(x - \mu_x)(y - \mu_y)$, $\mu_x = \mathrm{E}(x)$, $\mu_y = \mathrm{E}(y)$
$\mathrm{cor}(x, y) = \varrho_{xy}$	correlation between the random variables x and y: $\varrho_{xy} = \dfrac{\sigma_{xy}}{\sigma_x\sigma_y}$
$\mathrm{cov}(\mathbf{x})$	covariance matrix ($d \times d$) of a d-dimensional random vector $\mathbf{x}$: $\mathrm{cov}(\mathbf{x}) = \boldsymbol{\Sigma} = \mathrm{E}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})'$
$\mathrm{cor}(\mathbf{x})$	correlation matrix ($d \times d$) of the random vector $\mathbf{x}$: $\mathrm{cor}(\mathbf{x}) = \boldsymbol{\rho} = (\mathrm{diag}\,\boldsymbol{\Sigma})^{-1/2}\boldsymbol{\Sigma}(\mathrm{diag}\,\boldsymbol{\Sigma})^{-1/2}$
$\mathrm{cov}(\mathbf{x}, \mathbf{y})$	(cross-)covariance matrix between the random vectors $\mathbf{x}$ and $\mathbf{y}$: $\mathrm{cov}(\mathbf{x}, \mathbf{y}) = \mathrm{E}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}})' = \boldsymbol{\Sigma}_{\mathbf{xy}}$
$\mathrm{cov}(\mathbf{x}, \mathbf{x})$	$\mathrm{cov}(\mathbf{x}, \mathbf{x}) = \mathrm{cov}(\mathbf{x})$
$\mathrm{cor}(\mathbf{x}, \mathbf{y})$	(cross-)correlation matrix between the random vectors $\mathbf{x}$ and $\mathbf{y}$
$\mathrm{cov}\binom{\mathbf{x}}{y}$	partitioned covariance matrix of the random vector $\binom{\mathbf{x}}{y}$: $\mathrm{cov}\begin{pmatrix}\mathbf{x}\\ y\end{pmatrix} = \begin{pmatrix}\boldsymbol{\Sigma}_{\mathbf{xx}} & \boldsymbol{\sigma}_{\mathbf{x}y}\\ \boldsymbol{\sigma}_{\mathbf{x}y}' & \sigma_y^2\end{pmatrix} = \begin{pmatrix}\mathrm{cov}(\mathbf{x}, \mathbf{x}) & \mathrm{cov}(\mathbf{x}, y)\\ \mathrm{cov}(\mathbf{x}, y)' & \mathrm{var}(y)\end{pmatrix}$
$\mathbf{x} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$	$\mathrm{E}(\mathbf{x}) = \boldsymbol{\mu}$, $\mathrm{cov}(\mathbf{x}) = \boldsymbol{\Sigma}$
$\mathbf{x} \sim \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	$\mathbf{x}$ follows the p-dimensional normal distribution $\mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
$n(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$	density for $\mathbf{x} \sim \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma}$ pd: $n(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \dfrac{1}{(2\pi)^{p/2}\lvert\boldsymbol{\Sigma}\rvert^{1/2}}\mathrm{e}^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$
$\mathrm{cc}_i(\mathbf{x}, \mathbf{y})$	ith largest canonical correlation between the random vectors $\mathbf{x}$ and $\mathbf{y}$
$\mathrm{cc}(\mathbf{x}, \mathbf{y})$	set of the canonical correlations between the random vectors $\mathbf{x}$ and $\mathbf{y}$
$\mathrm{cc}_+(\mathbf{x}, \mathbf{y})$	set of the nonzero (necessarily positive) canonical correlations between the random vectors $\mathbf{x}$ and $\mathbf{y}$; square roots of the nonzero eigenvalues of $\mathbf{P}_{\mathbf{A}}\mathbf{P}_{\mathbf{B}}$: $\mathrm{cc}_+(\mathbf{x}, \mathbf{y}) = \mathrm{nzch}^{1/2}(\mathbf{P}_{\mathbf{A}}\mathbf{P}_{\mathbf{B}}) = \mathrm{nzsg}[(\mathbf{A}'\mathbf{A})^{+1/2}\mathbf{A}'\mathbf{B}(\mathbf{B}'\mathbf{B})^{+1/2}]\,,$ $\mathrm{cov}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix} = \begin{pmatrix}\mathbf{A}'\mathbf{A} & \mathbf{A}'\mathbf{B}\\ \mathbf{B}'\mathbf{A} & \mathbf{B}'\mathbf{B}\end{pmatrix}$
$\mathbf{X} = (\mathbf{1} : \mathbf{X}_0)$	in regression context often the model matrix

$\mathbf{X}_0$	$n \times k$ data matrix of the x-variables: $$\mathbf{X}_0 = (\mathbf{x}_1 : \ldots : \mathbf{x}_k) = \begin{pmatrix} \mathbf{x}'_{(1)} \\ \vdots \\ \mathbf{x}'_{(n)} \end{pmatrix}$$
$\mathbf{x}_1, \ldots, \mathbf{x}_k$	variable vectors in the variable space $\mathbb{R}^n$
$\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(n)}$	observation vectors in the observation space $\mathbb{R}^k$
$\operatorname{ssp}(\mathbf{X}_0 : \mathbf{y})$	partitioned matrix of the sums of squares and products of deviations about the mean of data $(\mathbf{X}_0 : \mathbf{y})$: $$\operatorname{ssp}(\mathbf{X}_0 : \mathbf{y}) = \begin{pmatrix} \mathbf{T}_{\mathbf{xx}} & \mathbf{t}_{\mathbf{x}y} \\ \mathbf{t}'_{\mathbf{x}y} & t_{yy} \end{pmatrix} = (\mathbf{X}_0 : \mathbf{y})'\mathbf{C}(\mathbf{X}_0 : \mathbf{y})$$
$\operatorname{cov}_{\mathrm{d}}(\mathbf{X}_0 : \mathbf{y})$	partitioned sample covariance matrix of data $(\mathbf{X}_0 : \mathbf{y})$: $$\operatorname{cov}_{\mathrm{d}}(\mathbf{X}_0 : \mathbf{y}) = \begin{pmatrix} \mathbf{S}_{\mathbf{xx}} & \mathbf{s}_{\mathbf{x}y} \\ \mathbf{s}'_{\mathbf{x}y} & s_y^2 \end{pmatrix}$$
$\operatorname{cor}_{\mathrm{d}}(\mathbf{X}_0 : \mathbf{y})$	partitioned sample correlation matrix of data $(\mathbf{X}_0 : \mathbf{y})$: $$\operatorname{cor}_{\mathrm{d}}(\mathbf{X}_0 : \mathbf{y}) = \begin{pmatrix} \mathbf{R}_{\mathbf{xx}} & \mathbf{r}_{\mathbf{x}y} \\ \mathbf{r}'_{\mathbf{x}y} & 1 \end{pmatrix}$$
$\mathbf{H}$	orthogonal projector onto $\mathscr{C}(\mathbf{X})$, the hat matrix: $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}' = \mathbf{X}\mathbf{X}^{+} = \mathbf{P}_{\mathbf{X}}$
$\mathbf{M}$	orthogonal projector onto $\mathscr{C}(\mathbf{X})^{\perp}$: $\mathbf{M} = \mathbf{I}_n - \mathbf{H}$
$\mathbf{J}$	the orthogonal projector onto $\mathscr{C}(\mathbf{1}_n)$: $\mathbf{J} = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n' = \mathbf{P}_{\mathbf{1}_n}$
$\mathbf{C}$	centering matrix, the orthogonal projector onto $\mathscr{C}(\mathbf{1}_n)^{\perp}$: $\mathbf{C} = \mathbf{I}_n - \mathbf{J}$
$(\mathbf{X}_1 : \mathbf{X}_2)$	partitioned model matrix $\mathbf{X}$
$\mathbf{M}_1$	orthogonal projector onto $\mathscr{C}(\mathbf{X}_1)^{\perp}$: $\mathbf{M}_1 = \mathbf{I}_n - \mathbf{P}_{\mathbf{X}_1}$
$\hat{\boldsymbol{\beta}}$	solution to normal equation $\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}$, $\operatorname{OLSE}(\boldsymbol{\beta})$
$\mathbf{X}\hat{\boldsymbol{\beta}} = \hat{\mathbf{y}}$	$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y} =$ OLS fitted values, $\operatorname{OLSE}(\mathbf{X}\boldsymbol{\beta})$, denoted also $\widehat{\mathbf{X}\boldsymbol{\beta}} = \hat{\boldsymbol{\mu}}$, when $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$
$\tilde{\boldsymbol{\beta}}$	solution to generalized normal equation $\mathbf{X}'\mathbf{W}^{-}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{W}^{-}\mathbf{y}$, where $\mathbf{W} = \mathbf{V} + \mathbf{X}\mathbf{U}\mathbf{X}'$, $\mathscr{C}(\mathbf{W}) = \mathscr{C}(\mathbf{X} : \mathbf{V})$
$\tilde{\boldsymbol{\beta}}$	if $\mathbf{V}$ is positive definite and $\mathbf{X}$ has full column rank, then $\tilde{\boldsymbol{\beta}} = \operatorname{BLUE}(\boldsymbol{\beta}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$
$\mathbf{X}\tilde{\boldsymbol{\beta}}$	$\operatorname{BLUE}(\mathbf{X}\boldsymbol{\beta})$, denoted also $\widetilde{\mathbf{X}\boldsymbol{\beta}} = \tilde{\boldsymbol{\mu}}$
$\bar{y}$	mean of the response variable y: $\bar{y} = (y_1 + \cdots + y_n)/n$
$\bar{\mathbf{x}}$	vector of the means of k regressor variables $x_1, \ldots, x_k$: $\bar{\mathbf{x}} = (\bar{x}_1, \ldots, \bar{x}_k)' \in \mathbb{R}^k$
$\bar{\bar{\mathbf{y}}}$	projection of $\mathbf{y}$ onto $\mathscr{C}(\mathbf{1}_n)$: $\bar{\bar{\mathbf{y}}} = \mathbf{J}\mathbf{y} = \bar{y}\mathbf{1}_n$

$\tilde{\mathbf{y}}$	centered $\mathbf{y}$, $\tilde{\mathbf{y}} = \mathbf{C}\mathbf{y} = \mathbf{y} - \bar{\bar{\mathbf{y}}}$
$\hat{\boldsymbol{\beta}}_{\mathbf{x}}$	$\hat{\boldsymbol{\beta}}_{\mathbf{x}} = \mathbf{T}_{\mathbf{xx}}^{-1}\mathbf{t}_{\mathbf{x}y} = \mathbf{S}_{\mathbf{xx}}^{-1}\mathbf{s}_{\mathbf{x}y}$: the OLS-regression coefficients of x-variables when $\mathbf{X} = (\mathbf{1} : \mathbf{X}_0)$
$\hat{\beta}_0$	$\hat{\beta}_0 = \bar{y} - \hat{\boldsymbol{\beta}}'_{\mathbf{x}}\bar{\mathbf{x}} = \bar{y} - (\hat{\beta}_1\bar{x}_1 + \cdots + \hat{\beta}_k\bar{x}_k)$: OLSE of the constant term (intercept) when $\mathbf{X} = (\mathbf{1} : \mathbf{X}_0)$
$\mathrm{BLP}(\mathbf{y};\mathbf{x})$	the best linear predictor of the random vector $\mathbf{y}$ on the basis of the random vector $\mathbf{x}$
$\mathrm{BLUE}(\mathbf{K}'\boldsymbol{\beta})$	the best linear unbiased estimator of estimable parametric function $\mathbf{K}'\boldsymbol{\beta}$, denoted as $\mathbf{K}'\tilde{\boldsymbol{\beta}}$ or $\widetilde{\mathbf{K}'\boldsymbol{\beta}}$
$\mathrm{BLUP}(\mathbf{y}_f;\mathbf{y})$	the best linear unbiased predictor of a new unobserved $\mathbf{y}_f$
$\mathrm{LE}(\mathbf{K}'\boldsymbol{\beta};\mathbf{y})$	(homogeneous) linear estimator of $\mathbf{K}'\boldsymbol{\beta}$, where $\mathbf{K} \in \mathbb{R}^{p\times q}$: $\{\mathrm{LE}(\mathbf{K}'\boldsymbol{\beta};\mathbf{y})\} = \{\, \mathbf{A}\mathbf{y} : \mathbf{A} \in \mathbb{R}^{q\times n} \,\}$
$\mathrm{LP}(\mathbf{y};\mathbf{x})$	(inhomogeneous) linear predictor of the p-dimensional random vector $\mathbf{y}$ on the basis of the q-dimensional random vector $\mathbf{x}$: $\{\mathrm{LP}(\mathbf{y};\mathbf{x})\} = \{\, f(\mathbf{x}) : f(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{a},\ \mathbf{A} \in \mathbb{R}^{p\times q},\ \mathbf{a} \in \mathbb{R}^{p} \,\}$
$\mathrm{LUE}(\mathbf{K}'\boldsymbol{\beta};\mathbf{y})$	(homogeneous) linear unbiased estimator of $\mathbf{K}'\boldsymbol{\beta}$: $\{\mathrm{LUE}(\mathbf{K}'\boldsymbol{\beta};\mathbf{y})\} = \{\, \mathbf{A}\mathbf{y} : \mathrm{E}(\mathbf{A}\mathbf{y}) = \mathbf{K}'\boldsymbol{\beta} \,\}$
$\mathrm{LUP}(\mathbf{y}_f;\mathbf{y})$	linear unbiased predictor of a new unobserved $\mathbf{y}_f$: $\{\mathrm{LUP}(\mathbf{y}_f;\mathbf{y})\} = \{\, \mathbf{A}\mathbf{y} : \mathrm{E}(\mathbf{A}\mathbf{y} - \mathbf{y}_f) = \mathbf{0} \,\}$
$\mathrm{MSEM}(f(\mathbf{x});\mathbf{y})$	mean squared error matrix of $f(\mathbf{x})$ (= random vector, function of the random vector $\mathbf{x}$) with respect to $\mathbf{y}$ (= random vector or a given fixed vector): $\mathrm{MSEM}[f(\mathbf{x});\mathbf{y}] = \mathrm{E}[\mathbf{y} - f(\mathbf{x})][\mathbf{y} - f(\mathbf{x})]'$
$\mathrm{MSEM}(\mathbf{F}\mathbf{y};\mathbf{K}'\boldsymbol{\beta})$	mean squared error matrix of the linear estimator $\mathbf{F}\mathbf{y}$ under $\{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{V}\}$ with respect to $\mathbf{K}'\boldsymbol{\beta}$: $\mathrm{MSEM}(\mathbf{F}\mathbf{y};\mathbf{K}'\boldsymbol{\beta}) = \mathrm{E}(\mathbf{F}\mathbf{y} - \mathbf{K}'\boldsymbol{\beta})(\mathbf{F}\mathbf{y} - \mathbf{K}'\boldsymbol{\beta})'$
$\mathrm{OLSE}(\mathbf{K}'\boldsymbol{\beta})$	the ordinary least squares estimator of parametric function $\mathbf{K}'\boldsymbol{\beta}$, denoted as $\mathbf{K}'\hat{\boldsymbol{\beta}}$ or $\widehat{\mathbf{K}'\boldsymbol{\beta}}$; here $\hat{\boldsymbol{\beta}}$ is any solution to the normal equation $\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}$
$\mathrm{risk}(\mathbf{F}\mathbf{y};\mathbf{K}'\boldsymbol{\beta})$	quadratic risk of $\mathbf{F}\mathbf{y}$ under $\{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{V}\}$ with respect to $\mathbf{K}'\boldsymbol{\beta}$: $\mathrm{risk}(\mathbf{F}\mathbf{y};\mathbf{K}'\boldsymbol{\beta}) = \mathrm{tr}[\mathrm{MSEM}(\mathbf{F}\mathbf{y};\mathbf{K}'\boldsymbol{\beta})] = \mathrm{E}(\mathbf{F}\mathbf{y} - \mathbf{K}'\boldsymbol{\beta})'(\mathbf{F}\mathbf{y} - \mathbf{K}'\boldsymbol{\beta})$
$\mathscr{M}$	linear model: $\{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta},\, \sigma^2\mathbf{V}\}$: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\mathrm{cov}(\mathbf{y}) = \mathrm{cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{V}$, $\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$
$\mathscr{M}_{\mathrm{mix}}$	mixed linear model: $\mathscr{M}_{\mathrm{mix}} = \{\mathbf{y},\, \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma},\, \mathbf{D},\, \mathbf{R}\}$: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$; $\boldsymbol{\gamma}$ is the vector of the random effects, $\mathrm{cov}(\boldsymbol{\gamma}) = \mathbf{D}$, $\mathrm{cov}(\boldsymbol{\varepsilon}) = \mathbf{R}$, $\mathrm{cov}(\boldsymbol{\gamma}, \boldsymbol{\varepsilon}) = \mathbf{0}$, $\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$
$\mathscr{M}_f$	linear model with new future observations $\mathbf{y}_f$: $\mathscr{M}_f = \left\{ \begin{pmatrix} \mathbf{y} \\ \mathbf{y}_f \end{pmatrix},\ \begin{pmatrix} \mathbf{X}\boldsymbol{\beta} \\ \mathbf{X}_f\boldsymbol{\beta} \end{pmatrix},\ \sigma^2 \begin{pmatrix} \mathbf{V} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{pmatrix} \right\}$

List of Figures and Philatelic Items

Ph.I. 7.1, p. 154	Leonid Vitaliyevich Kantorovich (1912–1986) was a Soviet/Russian mathematician and economist. Abū Abdallāh Muḥammad ibn Mūsā al-Khwārizmī (c. 780–c. 850), a Persian mathematician, astronomer and geographer and scholar in the House of Wisdom in Baghdad. He is considered the founder of algebra, a credit he shares with Diophantus of Alexandria.
Ph.I. 7.2, p. 154	Pierre-Simon marquis de Laplace (1749–1827) was a French mathematician and astronomer whose work was pivotal to the development of mathematical astronomy and statistics.
Ph.I. 9.1, p. 214	Johann Carl Friedrich Gauss (1777–1855) was a German mathematician and scientist who contributed significantly to many fields, including number theory, statistics, analysis, differential geometry, geodesy, geophysics, electrostatics, astronomy and optics.
Ph.I. 11.1, p. 282	John von Neumann (1903–1957) was a Hungarian–American mathematician who made major contributions to a vast range of fields, including set theory, functional analysis, quantum mechanics, ergodic theory, continuous geometry, economics and game theory, computer science, numerical analysis, and statistics.
Ph.I. 11.2, p. 282	Jules Henri Poincaré (1854–1912) was one of France's greatest mathematicians and theoretical physicists and in 1886 he was appointed to the Chair of Mathematical Physics and Probability at the Sorbonne.
Ph.I. 12.1, p. 290	Tadeusz Banachiewicz (1882–1954) was a Polish astronomer, mathematician and geodesist. The French mathematician Baron Augustin-Louis Cauchy (1789–1857) started the project of formulating and proving the theorems of infinitesimal calculus in a rigorous manner.
Ph.I. 17.1, p. 356	Determinants were applied in 1683 by the Japanese mathematician Takakazu Seki Kôwa (1642–1708) in the construction of the resolvent of a system of polynomial equations [but] were independently invented in 1693 by Gottfried Wilhelm von Leibniz (1646–1716). Leibniz was a German mathematician and philosopher who invented infinitesimal calculus independently of Isaac Newton (1643–1727).
Ph.I. 17.2, p. 356	Charles Lutwidge Dodgson (1832–1898), better known by the pen name Lewis Carroll, was an English author, mathematician, Anglican clergyman and photographer. As a mathematician, Dodgson was the author of *Condensation of Determinants* (1866) and *Elementary Treatise on Determinants* (1867). His most famous writings, however, are *Alice's Adventures in Wonderland* and its sequel *Through the Looking-Glass*.
Ph.I. 19.1, p. 414	Paul Anthony Samuelson (1914–2009) was the first American to win the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel (in 1970). Jan Tinbergen (1903–1994) was a Dutch economist, who was awarded the first Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel in 1969.
Ph.I. 19.2, p. 414	John William Strutt, 3rd Baron Rayleigh, OM (1842–1919) was an English physicist who, with Sir William Ramsay, KCB (1852–1916), discovered the element argon, an achievement for which he earned the Nobel Prize for Physics in 1904.

References

Abadir, K. M. & Magnus, J. R. (2005). *Matrix Algebra.* Cambridge University Press. [vii, 53, 280, 303, 316, 389, 398, 410]

Afriat, S. N. (1957). Orthogonal and oblique projectors and the characteristics of pairs of vector spaces. *Proceedings of the Cambridge Philosophical Society*, 53, 800–816. [134]

Aigner, D. J. & Balestra, P. (1988). Optimal experimental design for error components models. *Econometrica*, 56, 955–971. [331]

Aitken, A. C. (1935). On least squares and linear combination of observations. *Proceedings of the Royal Society of Edinburgh, Section A*, 55, 42–49. [43]

Aitken, A. C. (1939). *Determinants and Matrices.* Oliver & Boyd. 2nd–9th editions, 1942–1956; 9th edition, reset & reprinted, 1967. [291]

Alalouf, I. S. & Styan, G. P. H. (1979a). Characterizations of estimability in the general linear model. *The Annals of Statistics*, 7, 194–200. [346]

Alalouf, I. S. & Styan, G. P. H. (1979b). Estimability and testability in restricted linear models. *Mathematische Operationsforschung und Statistik, Series Statistics*, 10, 189–201. [346]

Albert, A. (1969). Conditions for positive and nonnegative definiteness in terms of pseudoinverses. *SIAM Journal on Applied Mathematics*, 17, 434–440. [306]

Albert, A. (1972). *Regression and the Moore–Penrose Pseudoinverse.* Academic Press. [306]

Albert, A. (1973). The Gauss–Markov theorem for regression models with possibly singular covariances. *SIAM Journal on Applied Mathematics*, 24, 182–187. [43, 229, 325]

Alpargu, G., Drury, S. W. & Styan, G. P. H. (1997). Some remarks on the Bloomfield–Watson–Knott Inequality and on some other inequalities related to the Kantorovich Inequality. In *Proceedings of the Conference in Honor of Shayle R. Searle, August 9–10, 1996*, Biometrics Unit, Cornell University, Ithaca, New York, pp. 125–143. [235, 425]

Alpargu, G. & Styan, G. P. H. (2000). Some comments and a bibliography on the Frucht–Kantorovich and Wielandt inequalities. In *Innovations in Multivariate Statistical Analysis: A Festschrift for Heinz Neudecker* (R. D. H. Heijmans, D. S. G. Pollock & A. Satorra, eds.), Kluwer, pp. 1–38. [235, 425]

Anderson, T. W. (1948). On the theory of testing serial correlation. *Skandinavisk Aktuarietidskrift*, 31, 88–116. [215]

Anderson, T. W. (1971). *The Statistical Analysis of Time Series.* Wiley. Reprinted as Wiley Classics Library Edition, 1994. [221]

Anderson, T. W. (1972). Efficient estimation of regression coefficients in time series. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1* (L. M. Le Cam, J. Neyman & E. L. Scott, eds.), University of California Press, Berkeley, pp. 471–482. [221]

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, Third Ed. Wiley. [vii, 19, 26, 132, 184, 195, 198, 213, 291, 366]

Anderson, T. W. & Styan, G. P. H. (1982). Cochran's theorem, rank additivity and tripotent matrices. In *Statistics and Probability: Essays in Honor of C. R. Rao* (G. Kallianpur, P. R. Krishnaiah & J. K. Ghosh, eds.), North-Holland, pp. 1–23. [352, 390]

Anderson, Jr., W. N. (1971). Shorted operators. *SIAM Journal on Applied Mathematics*, 20, 520–525. [313]

Anderson, Jr., W. N. & Trapp, G. E. (1975). Shorted operators, II. *SIAM Journal on Applied Mathematics*, 28, 60–71. [313]

Andrews, D. W. K. & Phillips, P. C. B. (1986). A simplified proof of a theorem on the difference of the Moore–Penrose inverses of two positive semidefinite matrices. *Communications in Statistics: Theory and Methods*, 15, 2973–2975. [312]

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, 27, 17–21. [199, 200, 436]

Arav, M., Hall, F. J. & Li, Z. (2008). A Cauchy–Schwarz inequality for triples of vectors. *Mathematical Inequalities & Applications*, 11, 629–634. [307]

Arnold, B. C. & Groeneveld, R. A. (1974). Bounds for deviations between sample and population statistics. *Biometrika*, 61, 387–389. [421]

Baksalary, J. K. (1984). Nonnegative definite and positive definite solutions to the matrix equation $AXA^* = B$. *Linear and Multilinear Algebra*, 16, 133–139. [265]

Baksalary, J. K. (1987). Algebraic characterizations and statistical implications of the commutativity of orthogonal projectors. In *Proceedings of the Second International Tampere Conference in Statistics* (T. Pukkila & S. Puntanen, eds.), Dept. of Mathematical Sciences, University of Tampere, pp. 113–142. [156, 190]

Baksalary, J. K. (1988). Criteria for the equality between ordinary least squares and best linear unbiased estimators under certain linear models. *The Canadian Journal of Statistics*, 16, 97–102. [241]

Baksalary, J. K. (1990). Solution to Problem 89-7. [Let X, Y, and Z be random variables. If the correlations $\varrho(X,Y)$ and $\varrho(Y,Z)$ are known, what are the sharp lower and upper bounds for $\varrho(X,Z)$?]. *The IMS Bulletin*, 19, 213–214. [307]

Baksalary, J. K. (2004). An elementary development of the equation characterizing best linear unbiased estimators. *Linear Algebra and its Applications*, 388, 3–6. [218]

Baksalary, J. K., Baksalary, O. M. & Szulc, T. (2004). Properties of Schur complements in partitioned idempotent matrices. *Linear Algebra and its Applications*, 379, 303–318. [303]

Baksalary, J. K. & van Eijnsbergen, A. C. (1988). A comparison of two criteria for ordinary-least-squares estimators to be best linear unbiased estimators. *The American Statistician*, 42, 205–208. [227]

Baksalary, J. K., Hauke, J., Liu, X. & Liu, S. (2004). Relationships between partial orders of matrices and their powers. *Linear Algebra and its Applications*, 379, 277–287. [316]

Baksalary, J. K. & Kala, R. (1977). An extension of the rank criterion for the least squares estimator to be the best linear unbiased estimator. *Journal of Statistical Planning and Inference*, 1, 309–312. [221]

Baksalary, J. K. & Kala, R. (1978a). A bound for the Euclidean norm of the difference between the least squares and the best linear unbiased estimators. *The Annals of Statistics*, 6, 1390–1393. [325]

Baksalary, J. K. & Kala, R. (1978b). Relationships between some representations of the best linear unbiased estimator in the general Gauss–Markoff model. *SIAM Journal on Applied Mathematics*, 35, 515–520. [229]

Baksalary, J. K. & Kala, R. (1979). Two relations between oblique and Λ-orthogonal projectors. *Linear Algebra and its Applications*, 24, 99–103. [156]

Baksalary, J. K. & Kala, R. (1980). A new bound for the Euclidean norm of the difference between the least squares and the best linear unbiased estimators. *The Annals of Statistics*, 8, 679–681. [325]

Baksalary, J. K. & Kala, R. (1981a). Linear transformations preserving best linear unbiased estimators in a general Gauss–Markoff model. *The Annals of Statistics*, 9, 913–916. [257]

Baksalary, J. K. & Kala, R. (1981b). Simple least squares estimation versus best linear unbiased prediction. *Journal of Statistical Planning and Inference*, 5, 147–151. [249]

Baksalary, J. K. & Kala, R. (1983a). On equalities between BLUEs, WLSEs, and SLSEs. *The Canadian Journal of Statistics*, 11, 119–123. [Corrigendum: (1984), 12, p. 240]. [149]

Baksalary, J. K. & Kala, R. (1983b). Partial orderings between matrices one of which is of rank one. *Bulletin of the Polish Academy of Sciences, Mathematics*, 31, 5–7. [306]

Baksalary, J. K. & Kala, R. (1986). Linear sufficiency with respect to a given vector of parametric functions. *Journal of Statistical Planning and Inference*, 14, 331–338. [258, 265, 266]

Baksalary, J. K., Kala, R. & Kłaczyński, K. (1983). The matrix inequality $M \geq B^*MB$. *Linear Algebra and its Applications*, 54, 77–86. [316]

Baksalary, J. K., Liski, E. P. & Trenkler, G. (1989). Mean square error matrix improvements and admissibility of linear estimators. *Journal of Statistical Planning and Inference*, 23, 313–325. [312]

Baksalary, J. K. & Markiewicz, A. (1988). Admissible linear estimators in the general Gauss–Markov model. *Journal of Statistical Planning and Inference*, 19, 349–359. [260]

Baksalary, J. K. & Markiewicz, A. (1989). A matrix inequality and admissibility of linear estimators with respect to the mean square error matrix criterion. *Linear Algebra and its Applications*, 112, 9–18. [260]

Baksalary, J. K. & Markiewicz, A. (1990). Admissible linear estimators of an arbitrary vector of parametric functions in the general Gauss–Markov model. *Journal of Statistical Planning and Inference*, 26, 161–171. [260]

Baksalary, J. K., Markiewicz, A. & Rao, C. R. (1995). Admissible linear estimation in the general Gauss–Markov model with respect to an arbitrary quadratic risk function. *Journal of Statistical Planning and Inference*, 44, 341–347. [260]

Baksalary, J. K. & Mathew, T. (1986). Linear sufficiency and completeness in an incorrectly specified general Gauss–Markov model. *Sankhyā, Ser. A*, 48, 169–180. [271]

Baksalary, J. K. & Mathew, T. (1990). Rank invariance criterion and its application to the unified theory of least squares. *Linear Algebra and its Applications*, 127, 393–401. [286, 288]

Baksalary, J. K., Nurhonen, M. & Puntanen, S. (1992). Effect of correlations and unequal variances in testing for outliers in linear regression. *Scandinavian Journal of Statistics*, 19, 91–95. [342]

Baksalary, J. K. & Pukelsheim, F. (1991). On the Löwner, minus, and star partial orderings of nonnegative definite matrices and their squares. *Linear Algebra and its Applications*, 151, 135–141. [316]

Baksalary, J. K., Pukelsheim, F. & Styan, G. P. H. (1989). Some properties of matrix partial orderings. *Linear Algebra and its Applications*, 119, 57–85. [316]

Baksalary, J. K. & Puntanen, S. (1989). Weighted-least-squares estimation in the general Gauss–Markov model. In *Statistical Data Analysis and Inference* (Y. Dodge, ed.), North-Holland, pp. 355–368. [44, 149, 286, 287, 288, 289]

Baksalary, J. K. & Puntanen, S. (1990a). Characterizations of the best linear unbiased estimator in the general Gauss–Markov model with the use of matrix partial orderings. *Linear Algebra and its Applications*, 127, 363–370. [221]

Baksalary, J. K. & Puntanen, S. (1990b). A complete solution to the problem of robustness of Grubbs's test. *The Canadian Journal of Statistics*, 18, 285–287. [342]

Baksalary, J. K. & Puntanen, S. (1991). Generalized matrix versions of the Cauchy–Schwarz and Kantorovich inequalities. *Aequationes Mathematicae*, 41, 103–110. [238, 423, 425]

Baksalary, J. K., Puntanen, S. & Styan, G. P. H. (1990a). On T. W. Anderson's contributions to solving the problem of when the ordinary least-squares estimator is best linear unbiased and to characterizing rank additivity of matrices. In *The Collected Papers of T. W. Anderson: 1943–1985* (G. P. H. Styan, ed.), Wiley, pp. 1579–1591. [216]

Baksalary, J. K., Puntanen, S. & Styan, G. P. H. (1990b). A property of the dispersion matrix of the best linear unbiased estimator in the general Gauss–Markov model. *Sankhyā, Ser. A*, 52, 279–296. [149, 261, 286, 287, 323]

Baksalary, J. K., Puntanen, S. & Yanai, H. (1992). Canonical correlations associated with symmetric reflexive generalized inverses of the dispersion matrix. *Linear Algebra and its Applications*, 176, 61–74. [295, 345, 386, 387]

Baksalary, J. K., Rao, C. R. & Markiewicz, A. (1992). A study of the influence of the "natural restrictions" on estimation problems in the singular Gauss–Markov model. *Journal of Statistical Planning and Inference*, 31, 335–351. [39, 125]

Baksalary, J. K., Schipp, B. & Trenkler, G. (1992). Some further results on Hermitian-matrix inequalities. *Linear Algebra and its Applications*, 160, 119–129. [306]

Baksalary, J. K. & Styan, G. P. H. (2002). Generalized inverses of partitioned matrices in Banachiewicz–Schur form. *Linear Algebra and its Applications*, 354, 41–47. [295]

Baksalary, O. M. & Styan, G. P. H. (2007). Some comments on the life and publications of Jerzy K. Baksalary (1944–2005). *Linear Algebra and its Applications*, 410, 3–53. [ix]

Baksalary, O. M., Styan, G. P. H. & Trenkler, G. (2009). On a matrix decomposition of Hartwig and Spindelböck. *Linear Algebra and its Applications*, 430, 2798–2812. [112]

Baksalary, O. M. & Trenkler, G. (2009a). On angles and distances between subspaces. *Linear Algebra and its Applications*, 431, 2243–2260. [134]

Baksalary, O. M. & Trenkler, G. (2009b). A projector oriented approach to the best linear unbiased estimator. *Statistical Papers*, 50, 721–733. [156, 218]

Balestra, P. (1970). On the efficiency of ordinary least-squares in regression models. *Journal of the American Statistical Association*, 65, 1330–1337. [227]

Baltagi, B. H. (1989). Applications of a necessary and sufficient condition for OLS to be BLUE. *Statistics & Probability Letters*, 8, 457–461. [241]

Baltagi, B. H. (2000). Further evidence on the efficiency of least squares in regression models. In *Panel Data Econometrics: Future Directions* (J. Krisnakumar & E. Ronchetti, eds.), Elsevier, pp. 279–291. [163]

Baltagi, B. H. (2008). *Econometrics*, Fourth Ed. Springer. [163]

Banachiewicz, T. (1937a). Sur l'inverse d'un cracovien et une solution générale d'un système d'équations linéaires. *Comptes Rendus Mensuels des Séances de la Classe des Sciences Mathématiques et Naturelles de l'Académie Polonaise des Sciences et des Lettres*, no. 4, 3–4. [295]

Banachiewicz, T. (1937b). Zur Berechnung der Determinanten, wie auch der Inversen, und zur darauf basierten Auflösung der Systeme lineare Gleichungen. *Acta Astronomica, Série C*, 3, 41–67. [295]

Bapat, R. B. (2000). *Linear Algebra and Linear Models*, Second Ed. Springer. [vii, 52, 283, 348, 352, 360, 386, 389]

Barnard, G. A. (1963). The logic of least squares. *Journal of the Royal Statistical Society, Ser. B*, 25, 124–127. [257]

Bartmann, F. C. & Bloomfield, P. (1981). Inefficiency and correlation. *Biometrika*, 68, 67–71. [244]

Bekker, P. A. & Neudecker, H. (1989). Albert's theorem applied to problems of efficiency and MSE superiority. *Statistica Neerlandica*, 43, 157–167. [306]

Belsley, D. A. (1991). *Conditioning Diagnostics: Collinearity and Weak Data in Regression.* Wiley. [20, 93, 100, 162, 178, 412]

Belsley, D. A., Kuh, E. & Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* Wiley. [181]

Ben-Israel, A. (1992). A volume associated with $m \times n$ matrices. *Linear Algebra and its Applications*, 167, 87–111. [132]

Ben-Israel, A. (2002). The Moore of the Moore–Penrose inverse. *Electronic Journal of Linear Algebra*, 9, 150–157. [108]

Ben-Israel, A. & Greville, T. N. E. (2003). *Generalized Inverses: Theory and Applications*, Second Ed. Springer. [vii, 5, 10, 13, 107, 108, 111, 132, 134, 156, 268, 355, 401, 407]

Ben-Israel, A. & Levin, Y. (2006). The geometry of linear separability in data sets. *Linear Algebra and its Applications*, 416, 75–87. [374, 377]

Benson, W. H. & Jacoby, O. (1976). *New Recreations with Magic Squares.* Dover. [55]

Benzécri, J.-P. (1973). *L'Analyse des Données: L'Analyse des Correspondances.* Dunod. [14]

Bernstein, D. S. (2009). *Matrix Mathematics: Theory, Facts, and Formulas*, Second Ed. Princeton University Press. [vii]

Bertrand, P. V. & Holder, R. L. (1988). A quirk in multiple regression: the whole regression can be greater than the sum of its parts. *The Statistician*, 37, 371–374. [309]

Bhimasankaram, P. (1988). Rank factorization of a matrix and its applications. *The Mathematical Scientist*, 13, 4–14. [350]

Bhimasankaram, P. & Saha Ray, R. (1997). On a partitioned linear model and some associated reduced models. *Linear Algebra and its Applications*, 264, 329–339. [323, 331]

Bhimasankaram, P. & Sengupta, D. (1996). The linear zero functions approach to linear models. *Sankhyā, Ser. B*, 58, 338–351. [331]

Bhimasankaram, P., Shah, K. R. & Saha Ray, R. (1998). On a singular partitioned linear model and some associated reduced models. *Journal of Combinatorics, Information & System Sciences*, 23, 415–421. [331]

Blom, G. (1976). When is the arithmetic mean BLUE? *The American Statistician*, 30, 40–42. [224]

Bloomfield, P. & Watson, G. S. (1975). The inefficiency of least squares. *Biometrika*, 62, 121–128. [238, 239, 240]

Bock, R. D. (2007). Rethinking Thurstone. In *Factor Analysis at 100: Historical Developments and Future Directions* (R. Cudeck & R. C. MacCallum, eds.), Lawrence Erlbaum, pp. 35–45. [viii]

Bouniakowsky, V. Y. (1859). Sur quelques inégalités concernant les intégrales ordinaires et les intégrales aux différences finies. *Mémoires de l'Académie Impériale des Sciences de St.-Pétersbourg, Septième Série*, 1, 1–18. [415]

Bring, J. (1996). Geometric approach to compare variables in a regression model. *The American Statistician*, 50, 57–62. [183]

Bryant, P. (1984). Geometry, statistics, probability: variations on a common theme. *The American Statistician*, 38, 38–48. [156, 183, 213]

Canner, P. L. (1969). Some curious results using minimum variance linear unbiased estimators. *The American Statistician*, 23, 39–40. [263]

Carlson, D. (1986). What are Schur complements, anyway? *Linear Algebra and its Applications*, 74, 257–275. [293]

Casella, G. (2008). *Statistical Design.* Springer. [vii, 188]

Cauchy, A.-L. (1821). *Cours d'Analyse de l'École Royale Polytechnique, Première Partie: Analyse Algébrique*. Debure Frères, Paris. [Cf. pp. 360–377; reprinted in *Œuvres Complètes d'Augustin Cauchy, Seconde Série: Vol. 4*, Gauthier-Villars, Paris, 1897]. [415]

Chatterjee, S. & Firat, A. (2007). Generating data with identical statistics but dissimilar graphics. *The American Statistician*, 61, 248–254. [199]

Chatterjee, S. & Hadi, A. S. (1986). Influential observations, high leverage points, and outliers in linear regression (with discussion). *Statistical Science*, 1, 379–416. [158]

Chatterjee, S. & Hadi, A. S. (1988). *Sensitivity Analysis in Linear Regression*. Wiley. [158]

Chatterjee, S. & Hadi, A. S. (2006). *Regression Analysis by Example*, Fourth Ed. Wiley. [ix]

Cheng, T.-k. (1984). Some Chinese Islamic "magic square" porcelain. In *Studies in Chinese Ceramics* (T.-k. Cheng, ed.), The Chinese University Press, Hong Kong, pp. 133–143. [First published in *Journal of Asian Art*, vol. 1, pp. 146–159.]. [55]

Chipman, J. S. (1964). On least squares with insufficient observations. *Journal of the American Statistical Association*, 59, 1078–1111. [423]

Chipman, J. S. (1968). Specification problems in regression analysis. In *Proceedings of the Symposium on Theory and Application of Generalized Inverses of Matrices* (T. L. Boullion & P. L. Odell, eds.), Texas Tech. Press, pp. 114–176. [315]

Chipman, J. S. (1976). Estimation and aggregation in econometrics: an application of the theory of generalized inverses. In *Generalized Inverses and Applications* (M. Z. Nashed, ed.), Academic Press, pp. 549–769. [315]

Chipman, J. S. (1979). Efficiency of least squares estimation of linear trend when residuals are autocorrelated. *Econometrica*, 47, 115–128. [241]

Chipman, J. S. (1997). "Proofs" and proofs of the Eckart–Young theorem (with an Appendix by Heinz Neudecker). In *Stochastic Processes and Functional Analysis. In Celebration of M. M. Rao's 65th Birthday* (J. A. Goldstein, N. E. Gretsky & J. J. Uhl Jr., eds.), Dekker, pp. 71–83. [401]

Chipman, J. S. (1998). The contributions of Ragnar Frisch to economics and econometrics. In *Econometrics and Economic Theory in the 20th Century: The Ragnar Frisch Centennial Symposium (Oslo, 1995)* (S. Strøm, ed.), Cambridge University Press, pp. 58–108. [331]

Chipman, J. S. & Rao, M. M. (1964). Projections, generalized inverses and quadratic forms. *Journal of Mathematical Analysis and Applications*, 9, 1–11. [156]

Christensen, R. (1990). Comment on Puntanen and Styan (1989) (Letter to the Editor). *The American Statistician*, 44, 191–192. [39, 125]

Christensen, R. (2001). *Advanced Linear Modeling: Multivariate, Time Series, and Spatial Data; Nonparametric Regression and Response Surface Maximization*, Second Ed. Springer. [vii, 203, 206, 230, 231, 232, 234]

Christensen, R. (2002). *Plane Answers to Complex Questions: The Theory of Linear Models*, Third Ed. Springer. [vii, 195, 212, 247, 255, 256]

Christensen, R., Pearson, L. M. & Johnson, W. (1992). Case-deletion diagnostics for mixed models. *Technometrics*, 34, 38–45. [338]

Chu, K. L. (2004). Inequalities and equalities associated with ordinary least squares and generalized least squares in partitioned linear models. Ph.D. Thesis, Dept. of Mathematics & Statistics, McGill University, Montréal. [241]

Chu, K. L., Drury, S. W., Styan, G. P. H. & Trenkler, G. (2010). Magic generalized inverses. Report 2010-02, Dept. of Mathematics & Statistics, McGill University, Montréal. [In preparation]. [54, 355]

Chu, K. L., Isotalo, J., Puntanen, S. & Styan, G. P. H. (2004). On decomposing the Watson efficiency of ordinary least squares in a partitioned weakly singular linear model. *Sankhyā*, 66, 634–651. [236, 333]

Chu, K. L., Isotalo, J., Puntanen, S. & Styan, G. P. H. (2005). Some further results concerning the decomposition of the Watson efficiency in partitioned linear models. *Sankhyā*, 67, 74–89. [333, 384]

Chu, K. L., Isotalo, J., Puntanen, S. & Styan, G. P. H. (2007). The efficiency factorization multiplier for the Watson efficiency in partitioned linear models: some examples and a literature review. *Journal of Statistical Planning and Inference*, 137, 3336–3351. [241, 263, 333]

Chu, K. L., Isotalo, J., Puntanen, S. & Styan, G. P. H. (2008). Inequalities and equalities for the generalized efficiency function in orthogonally partitioned linear models. In *Inequalities and Applications* (T. M. Rassias & D. Andrica, eds.), Cluj University Press, pp. 13–69. [333]

Chu, K. L., Puntanen, S. & Styan, G. P. H. (2009). Some comments on philatelic Latin squares from Pakistan. *The Pakistan Journal of Statistics*, 25, 427–471. [68]

Chu, K. L., Puntanen, S. & Styan, G. P. H. (2010). Rank, eigenvalues and Moore–Penrose inverse of a special symmetric matrix associated with postage stamps: a new problem proposed. *Statistical Papers*, 51, submitted. [390]

Chu, K. L., Puntanen, S. & Styan, G. P. H. (2011). Solution to Problem 1/SP09 "Inverse and determinant of a special symmetric matrix" (Problem proposed by H. Neudecker, G. Trenkler, and S. Liu). *Statistical Papers*, 52, 258–260. [263]

Chu, M. T., Funderlic, R. E. & Golub, G. H. (1995). A rank-one reduction formula and its applications to matrix factorizations. *SIAM Review*, 37, 512–530. [304]

Colbourn, C. J. & Dinitz, J. H., eds. (2007). *Handbook of Combinatorial Designs*, Second Ed. Chapman & Hall/CRC. [vii]

Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics*, 19, 15–18. [Additional correspondence, pp. 348–350]. [181]

Cook, R. D. (1996). Personal communication. (S. P.). [206]

Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regression through Graphics.* Wiley. [199]

Cook, R. D. (2007). Fisher lecture: Dimension reduction in regression (with discussion). *Statistical Science*, 22, 1–26. [199]

Cook, R. D. & Forzani, L. (2008). Principal fitted components for dimension reduction in regression. *Statistical Science*, 23, 485–501. [199]

Cook, R. D., Li, B. & Chiaramonte, F. (2007). Dimension reduction in regression without matrix inversion. *Biometrika*, 94, 569–584. [199]

Cook, R. D. & Weisberg, S. (1999). *Applied Regression Including Computing and Graphics.* Wiley. [199]

Cottle, R. W. (1974). Manifestations of the Schur complement. *Linear Algebra and its Applications*, 8, 189–211. See also *Rendiconti del Seminario Matematico e Fisico di Milano*, 45 (1975), 31–40. [293]

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334. [422]

Cuadras, C. M. (1993). Interpreting an inequality in multiple regression. *The American Statistician*, 47, 256–258. [309]

Cuadras, C. M. (1995). Increasing the correlations with the response variable may not increase the coefficient of determination: a PCA interpretation. In *Multivariate Statistics and Matrices in Statistics: Proceedings of the 5th Tartu Conference, Tartu–Pühajärve, Estonia, 23–28 May 1994* (E.-M. Tiit, T. Kollo & H. Niemi, eds.), TEV & VSP, pp. 75–83. [309]

Das Gupta, S. (1971). Nonsingularity of the sample covariance matrix. *Sankhyā, Ser. A*, 33, 475–478. [129]

Das Gupta, S. (1993). The evolution of the D^2-statistic of Mahalanobis. *Sankhyā, Ser. A*, 55, 442–459. [25]

Das Gupta, S. (1997). Personal communication. (S. P.). [208]

Davidson, R. & MacKinnon, J. G. (1993). *Estimation and Inference in Econometrics.* Oxford University Press. [163]

Davidson, R. & MacKinnon, J. G. (2004). *Econometric Theory and Methods.* Oxford University Press. [163, 251]

DeGroot, M. H. (1986). A conversation with T. W. Anderson. *Statistical Science*, 1, 97–105. [215]

Demidenko, E. (2004). *Mixed Models: Theory and Applications.* Wiley. [256]

Dénes, J. & Keedwell, A. D. (1974). *Latin Squares and their Applications.* Academic Press. [68]

Dey, A., Hande, S. & Tiku, M. L. (1994). Statistical proofs of some matrix results. *Linear and Multilinear Algebra*, 38, 109–116. [viii, 306]

Diderrich, G. T. (1985). The Kalman filter from the perspective of Goldberger–Theil estimators. *The American Statistician*, 39, 193–198. [265]

Draper, N. R. & Pukelsheim, F. (1996). An overview of design of experiments. *Statistical Papers*, 37, 1–32. [vii]

Draper, N. R. & Smith, H. (1998). *Applied Regression Analysis*, Third Ed. Wiley. [ix]

Driscoll, M. F. & Gundberg, Jr., W. R. (1986). A history of the development of Craig's theorem. *The American Statistician*, 40, 65–70. [19]

Drury, S. W., Liu, S., Lu, C.-Y., Puntanen, S. & Styan, G. P. H. (2002). Some comments on several matrix inequalities with applications to canonical correlations: historical background and recent developments. *Sankhyā, Ser. A*, 64, 453–507. [136, 144, 238, 241, 386, 424, 425, 426]

Drygas, H. (1970). *The Coordinate-Free Approach to Gauss–Markov Estimation.* Springer. [41, 221, 325]

Drygas, H. (1983). Sufficiency and completeness in the general Gauss–Markov model. *Sankhyā, Ser. A*, 45, 88–98. [257, 258]

Drygas, H. & Zmyślony, R. (1988). On admissible estimation for parametric functions in linear models. *Statistical Papers*, 113–123. [260]

Dumais, M. F. (2000). The Craig–Sakamoto theorem. MSc Thesis, Dept. of Mathematics & Statistics, McGill University, Montréal. [19]

Duncan, W. J. (1944). Some devices for the solution of large sets of simultaneous linear equations (with an appendix on the reciprocation of partitioned matrices). *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science, Seventh Series*, 35, 660–670. [301]

Eaton, M. L. (1970). Gauss–Markov estimation for multivariate linear models: A coordinate free approach. *The Annals of Mathematical Statistics*, 41, 528–538. [221]

Eaton, M. L. (1976). A maximization problem and its application to canonical correlation. *Journal of Multivariate Analysis*, 6, 422–425. [419]

Eaton, M. L. (1978). A note on the Gauss–Markov Theorem. *Annals of the Institute of Statistical Mathematics*, 30, 181–184. [221]

Eaton, M. L. (1983). *Multivariate Statistics: A Vector Space Approach.* Wiley. Reprinted: IMS Lecture Notes – Monograph Series, Vol. 53, 2007. [221]

Eckart, C. & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, 1, 211–218. [400, 401]

Elian, S. N. (2000). Simple forms of the best linear unbiased predictor in the general linear regression model. *The American Statistician*, 54, 25–28. [249]

Eubank, R. L. & Webster, J. T. (1985). The singular-value decomposition as a tool for solving estimability problems. *The American Statistician*, 39, 64–66. [393]

Farebrother, R. W. (1976). Further results on the mean square error of ridge regression. *Journal of the Royal Statistical Society, Ser. B*, 38, 248–250. [306]

Farebrother, R. W. (1990). Least squares and the Gauss–Markov theorem [comment on Puntanen and Styan (1989)] (Letter to the Editor). *The American Statistician*, 44, 191. [39, 125]

Farebrother, R. W. (1996). Some early statistical contributions to the theory and practice of linear algebra. *Linear Algebra and its Applications*, 237/238, 205–224. [viii]

Farebrother, R. W. (1997). A. C. Aitken and the consolidation of matrix theory. *Linear Algebra and its Applications*, 264, 3–12. [viii]

Farebrother, R. W. (2000). What is a Matrician? *Image: The Bulletin of the ILAS*, 25, 32. [vi]

Farebrother, R. W. (2009). Further results on Samuelson's inequality. In *Statistical Inference, Econometric Analysis and Matrix Algebra: Festschrift in Honour of Götz Trenkler* (B. Schipp & W. Krämer, eds.), Physica-Verlag, pp. 311–316. [421]

Farebrother, R. W., Jensen, S. T. & Styan, G. P. H. (2000). Charles Lutwidge Dodgson: a biographical and philatelic note. *Image: The Bulletin of the ILAS*, 25, 22–23. [356]

Farebrother, R. W., Styan, G. P. H. & Tee, G. J. (2003). Gottfried Wilhelm von Leibniz: 1646–1716. *Image: The Bulletin of the ILAS*, 30, 13–16. [356]

Fiebig, D. G., Bartels, R. & Krämer, W. (1996). The Frisch–Waugh theorem and generalised least squares. *Econometric Reviews*, 15, 431–443. [331]

Flury, B. (1997). *A First Course in Multivariate Statistics*. Springer. [x, 198, 201, 372, 373]

Flury, B. & Riedwyl, H. (1986). Standard distance in univariate and multivariate analysis. *The American Statistician*, 40, 249–251. [372]

Freund, R. J. (1988). When is $R^2 > r_{y1}^2 + r_{y2}^2$ (revisited). *The American Statistician*, 42, 89–90. [309]

Frisch, R. & Waugh, F. V. (1933). Partial time regressions as compared with individual trends. *Econometrica*, 1, 387–401. [163, 331]

Frucht, R. (1943). Sobre algunas desigualdades: Observación relativa a la solución del problema N$^{\text{o}}$ 21, indicada por el Ing. Ernesto M. Saleme (1942). *Mathematicae Notæ, Boletin del Instituto de Matemática "Beppo Levi"*, 3, 41–46. English Translation: Appendix A of Watson et al., 1997. [418]

Fujikoshi, Y., Ulyanov, V. V. & Shimizu, R. (2010). *Multivariate Statistics: High-Dimensional and Large-Sample Approximations*. Wiley. [vii, 411, 412]

Galántai, A. (2004). *Projectors and Projection Methods*. Kluwer. [156]

Galántai, A. (2008). Subspaces, angles and pairs of orthogonal projections. *Linear and Multilinear Algebra*, 56, 227–260. [134, 156]

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute*, 15, 246–263. [201, 436]

Gantmacher, F. R. (1959). *The Theory of Matrices, Volume 2*. Chelsea. Reprint by AMS Chelsea Publishing, 2000. [375]

Gelman, A. (2005). Analysis of variance – why it is more important than ever (with discussion). *The Annals of Statistics*, 33, 1–53. [256]

Gentle, J. E. (2007). *Matrix Algebra: Theory, Computations, and Applications in Statistics*. Springer. [vii]

Gnot, S., Klonecki, W. & Zmyślony, R. (1980). Best unbiased linear estimation, a coordinate free approach. *Probability and Mathematical Statistics*, 1, 1–13. [221]

Goldberger, A. S. (1962). Best linear unbiased prediction in the generalized linear regression model. *Journal of the American Statistical Association*, 57, 369–375. [246]

Golub, G. H. & Styan, G. P. H. (1973). Numerical computations for univariate linear models. *Journal of Statistical Computation and Simulation*, 2, 253–274. [390]

Golub, G. H. & Van Loan, C. F. (1996). *Matrix Computations*, Third Ed. Johns Hopkins University Press. [375, 390, 392, 393]

Gourieroux, C. & Monfort, A. (1980). Sufficient linear structures: econometric applications. *Econometrica*, 48, 1083–1097. [266, 331]

Graham, A. (1981). *Kronecker Products and Matrix Calculus: with Applications*. Ellis Horwood Ltd. [vi]

Grala, J., Markiewicz, A. & Styan, G. P. H. (2000). Tadeusz Banachiewicz: 1882–1954. *Image: The Bulletin of the ILAS*, 25, 24. [290, 295]

Graybill, F. A. (2002). *Matrices with Applications in Statistics*, Second Ed. Cengage Learning/Brooks/Cole. [vii]

Green, B. F. (1952). The orthogonal approximation of an oblique structure in factor analysis. *Psychometrika*, 17, 429–440. [401]

Greenacre, M. (2007). *Correspondence Analysis in Practice*, Second Ed. Chapman & Hall/CRC. [vi, 50, 413]

Groß, J. (1997a). A note on equality of MINQUE and simple estimator in the general Gauss–Markov model. *Statistics & Probability Letters*, 35, 335–339. [340]

Groß, J. (1997b). Special generalized inverse matrices connected with the theory of unified least squares. *Linear Algebra and its Applications*, 264, 325–327. [346]

Groß, J. (2003). *Linear Regression*. Springer. [vii]

Groß, J. (2004). The general Gauss–Markov model with possibly singular dispersion matrix. *Statistical Papers*, 45, 311–336. [218]

Groß, J. & Puntanen, S. (2000a). Estimation under a general partitioned linear model. *Linear Algebra and its Applications*, 321, 131–144. [163, 328, 331, 345]

Groß, J. & Puntanen, S. (2000b). Remark on pseudo-generalized least squares. *Econometric Reviews*, 19, 131–133. [331]

Groß, J. & Puntanen, S. (2005). Extensions of the Frisch–Waugh–Lovell theorem. *Discussiones Mathematicae – Probability and Statistics*, 25, 39–49. [331]

Groß, J., Puntanen, S. & Trenkler, G. (1996). On the equality of OLSE and BLUE in a linear model with partitioned data. In *Proceedings of the A. C. Aitken Centenary Conference (Dunedin, August 1995)* (L. Kavalieris, F. C. Lam, L. A. Roberts & J. A. Shanks, eds.), University of Otago Press, pp. 143–152. [265]

Groß, J. & Trenkler, G. (1997). When do linear transforms of ordinary least squares and Gauss–Markov estimator coincide? *Sankhyā, Ser. A*, 59, 175–178. [265]

Groß, J. & Trenkler, G. (1998). On the product of oblique projectors. *Linear and Multilinear Algebra*, 44, 247–259. [156, 190]

Groß, J., Trenkler, G. & Troschke, S.-O. (1997). Problem no. 10519. *The American Mathematical Monthly*, 103, 347. [351]

Groß, J., Trenkler, G. & Werner, H. J. (2001). The equality of linear transforms of the ordinary least squares estimator and the best linear unbiased estimator. *Sankhyā, Ser. A*, 63, 118–127. [222, 265]

Gustafson, K. (1972). Antieigenvalue inequalities in operator theory. In *Inequalities, III (Proc. Third Sympos., Univ. California, Los Angeles, 1969; Dedicated to the Memory of Theodore S. Motzkin)* (O. Shisha, ed.), Academic Press, pp. 115–119. [237]

Gustafson, K. (1994). Antieigenvalues. *Linear Algebra and its Applications*, 208/209, 437–454. [237]

Gustafson, K. (1999). The geometrical meaning of the Kantorovich-Wielandt inequalities. *Linear Algebra and its Applications*, 296, 143–151. [237, 419]

Gustafson, K. (2000). An extended operator trigonometry. *Linear Algebra and its Applications*, 319, 117–135. [237]

Gustafson, K. (2002). Operator trigonometry of statistics and econometrics. *Linear Algebra and its Applications*, 354, 141–158. [237]

Gustafson, K. (2005). The geometry of statistical efficiency. In *Proceedings of the 14th International Workshop on Matrices and Statistics,* vol. 8 *of Research Letters in the Information and Mathematical Sciences* (P. S. P. Cowpertwait, ed.), Massey University, pp. 105–121. [237]

Gustafson, K. (2006). The trigonometry of matrix statistics. *International Statistical Review*, 74, 187–202. [237]

Gustafson, K. (2007). The geometry of statistical efficiency and matrix statistics. *Journal of Applied Mathematics and Decision Sciences*, 2007, doi:10.1155/2007/94515. [237]

Gustafson, K. (2011). Operator geometry in statistics. In *The Oxford Handbook of Functional Data Analysis* (F. Ferraty & Y. Romain, eds.), Oxford University Press, pp. 355–382. [237]

Gustafson, K. & Rao, D. K. M. (1997). *Numerical Range: The Field of Values of Linear Operators and Matrices*. Springer. [237]

Gustafson, K. & Styan, G. P. H. (2009). Superstochastic matrices and magic Markov chains. *Linear Algebra and its Applications*, 430, 2705–2715. [224]

Guttman, L. (1944). General theory and methods of matric factoring. *Psychometrika*, 9, 1–16. [304]

Guttman, L. (1946). Enlargement methods for computing the inverse matrix. *The Annals of Mathematical Statistics*, 17, 336–343. [299, 304]

Guttman, L. (1952). Multiple group methods for common-factor analysis: their basis, computation and interpretation. *Psychometrika*, 17, 209–222. [304]

Guttman, L. (1957). A necessary and sufficient formula for matric factoring. *Psychometrika*, 22, 79–81. [304]

Haberman, S. J. (1975). How much do Gauss–Markov and least square estimates differ? A coordinate-free approach. *The Annals of Statistics*, 3, 982–990. [325]

Hadi, A. S. (1996). *Matrix Algebra as a Tool*. Duxbury Press. [vii]

Hägele, G. & Pukelsheim, F. (2001). Llull's writings on electoral systems. *Studia Lulliana*, 41, 3–38. [67]

Hager, W. W. (1989). Updating the inverse of a matrix. *SIAM Review*, 31, 221–239. [301]

Hall, F. J. & Meyer, C. D. (1975). Generalized inverses of the fundamental bordered matrix used in linear estimation. *Sankhyā, Ser. A*, 37, 428–438. [Corrigendum (1978), 40, p. 399]. [264]

Halmos, P. R. (1951). *Introduction to Hilbert Space and the Theory of Spectral Multiplicity*. Chelsea Publishing Company. [156]

Halmos, P. R. (1958). *Finite-Dimensional Vector Spaces*, Second Ed. Van Nostrand. Reprinted by Springer, 1974. [156]

Hamilton, D. (1987). Sometimes $R^2 > r^2_{yx_1} + r^2_{yx_2}$. Correlated variables are not always redundant. *The American Statistician*, 41, 129–132. [309]

Härdle, W. & Hlávka, Z. (2007). *Multivariate Statistics: Exercises and Solutions*. Springer. [vii]

Hartwig, R. E. (1980). How to partially order regular elements. *Mathematica Japonica*, 25, 1–13. [315, 316]

Hartwig, R. E. & Spindelböck, K. (1984). Matrices for which A^* and $A^\dagger$ commute. *Linear and Multilinear Algebra*, 14, 241–256. [112]

Hartwig, R. E. & Styan, G. P. H. (1986). On some characterizations of the "star" partial ordering for matrices and rank subtractivity. *Linear Algebra and its Applications*, 82, 145–161. [316]

Harville, D. A. (1976). Extension of the Gauss–Markov theorem to include the estimation of random effects. *The Annals of Statistics*, 4, 384–395. [273]

Harville, D. A. (1979). Some useful representations for constrained mixed-model estimation. *Journal of the American Statistical Association*, 74, 200–206. [273]

Harville, D. A. (1990a). BLUP (best linear unbiased prediction) and beyond. In *Advances in Statistical Methods for Genetic Improvement of Livestock* (D. Gianola & G. Hammond, eds.), Springer, pp. 239–276. [256]

Harville, D. A. (1990b). Comment on Puntanen and Styan (1989) (Letter to the Editor). *The American Statistician*, 44, 192. [39, 125]

Harville, D. A. (1997). *Matrix Algebra From a Statistician's Perspective*. Springer. [vii, 53, 87, 89, 280, 286, 288, 340, 358, 360, 398, 401]

Harville, D. A. (2001). *Matrix Algebra: Exercises and Solutions*. Springer. [vii]

Haslett, S. J. (1996). Updating linear models with dependent errors to include additional data and/or parameters. *Linear Algebra and its Applications*, 237/238, 329–349. [337, 338]

Haslett, S. J. & Govindaraju, K. (2009). Cloning data: generating datasets with exactly the same multiple linear regression fit. *Australian & New Zealand Journal of Statistics*, 51, 499–503. [199]

Haslett, S. J. & Puntanen, S. (2010a). Effect of adding regressors on the equality of the BLUEs under two linear models. *Journal of Statistical Planning and Inference*, 140, 104–110. [333, 337]

Haslett, S. J. & Puntanen, S. (2010b). Equality of BLUEs or BLUPs under two linear models using stochastic restrictions. *Statistical Papers*, 51, 465–475. [256, 273, 275, 277]

Haslett, S. J. & Puntanen, S. (2010c). A note on the equality of the BLUPs for new observations under two linear models. *Acta et Commentationes Universitatis Tartuensis de Mathematica*, 14, 27–33. [279, 280]

Haslett, S. J. & Puntanen, S. (2010d). On the equality of the BLUPs under two linear mixed models. *Metrika*, available online. [256, 278, 279]

Hauke, J., Markiewicz, A. & Puntanen, S. (2011). Comparing the BLUEs under two linear models. *Communications in Statistics: Theory and Methods*, 40, in press. [89, 266]

Hauke, J. & Pomianowska, J. (1987). Correlation relations in the light of the nonnegative definiteness of a partitioned matrix (in Polish). *Przegląd Statystyczny*, 34, 219–224. [307]

Hayes, K. & Haslett, J. (1999). Simplifying general least squares. *The American Statistician*, 53, 376–381. [256, 338]

Haynsworth, E. V. (1968a). Determination of the inertia of a partitioned Hermitian matrix. *Linear Algebra and its Applications*, 1, 73–81. [293, 294]

Haynsworth, E. V. (1968b). On the Schur complement. *Basle Mathematical Notes*, #BMN 20, 17 pages. [293, 294]

Healy, M. J. R. (2000). *Matrices for Statistics*, Second Ed. Oxford University Press. [vii]

Henderson, C. R. (1950). Estimation of genetic parameters. *The Annals of Mathematical Statistics*, 21, 309–310. [255, 273]

Henderson, C. R. (1963). Selection index and expected genetic advance. In *Statistical Genetics and Plant Breeding*, National Academy of Sciences – National Research Council Publication No. 982, pp. 141–163. [255, 273]

Henderson, C. R., Kempthorne, O., Searle, S. R. & von Krosigk, C. M. (1959). The estimation of environmental and genetic trends from records subject to culling. *Biometrics*, 15, 192–218. [255]

Henderson, H. V. & Searle, S. R. (1979). Vec and vech operators for matrices, with some uses in Jacobians and multivariate statistics. *The Canadian Journal of Statistics*, 7, 65–81. [53]

Henderson, H. V. & Searle, S. R. (1981a). On deriving the inverse of a sum of matrices. *SIAM Review*, 23, 53–60. [255, 293, 301]

Henderson, H. V. & Searle, S. R. (1981b). The vec-permutation matrix, the vec operator and Kronecker products: a review. *Linear and Multilinear Algebra*, 9, 271–288. [53]

Herr, D. G. (1980). On the history of the use of geometry in the general linear model. *The American Statistician*, 34, 43–47. [183]

Herzberg, A. M. & Aleong, J. (1985). Further conditions on the equivalence of ordinary least squares and weighted least squares estimators with examples. In *Contributions to Probability and Statistics in Honour of Gunnar Blom* (J. Lanke & G. Lindgren, eds.), University of Lund, pp. 127–142. [241]

Hoaglin, D. C. & Welsch, R. E. (1978). The hat matrix in regression and ANOVA. *The American Statistician*, 32, 17–22. [Corrigendum: 32 (1978), p. 146]. [158]

Hogben, L., ed. (2007). *Handbook of Linear Algebra*. Chapman & Hall/CRC. Associate editors: Richard Brualdi, Anne Greenbaum and Roy Mathias. [vii]

Hogg, R. V. & Craig, A. T. (1958). On the decomposition of certain χ^2 variables. *The Annals of Mathematical Statistics*, 29, 608–610. [355]

de Hoog, F. R., Speed, T. P. & Williams, E. R. (1990). A matrix version of the Wielandt inequality and its applications to statistics. *Linear Algebra and its Applications*, 127, 449–456. [323]

Horn, R. A. (1990). The Hadamard product. In *Matrix Theory and Applications* (C. R. Johnson, ed.), American Mathematical Society, pp. 87–169. [vi]

Horn, R. A. & Johnson, C. R. (1990). *Matrix Analysis*. Cambridge University Press. Corrected reprint of the 1985 original. [52, 358, 393]

Horn, R. A. & Olkin, I. (1996). When does $A^*A = B^*B$ and why does one want to know? *The American Mathematical Monthly*, 103, 470–482. [392, 393, 397]

Hotelling, H. (1933). Analysis of a complex statistical variables into principal components. *Journal of Educational Psychology*, 24, 417–441, 498–520. [206]

Hotelling, H. (1934). Open letter (Reply to Horace Secrist). *Journal of the American Statistical Association*, 29, 198–199. [200]

Householder, A. S. & Young, G. (1938). Matrix approximation and latent roots. *The American Mathematical Monthly*, 45, 165–171. [400]

Hubert, L., Meulman, J. & Heiser, W. (2000). Two purposes for matrix factorization: a historical appraisal. *SIAM Review*, 42, 68–82. [304, 401]

Ipsen, I. C. F. & Meyer, C. D. (1995). The angle between complementary subspaces. *The American Mathematical Monthly*, 102, 904–911. [134]

Isotalo, J. (2007). Linear estimation and prediction in the general Gauss–Markov model. *Acta Universitatis Tamperensis Series A*, 1242. Ph.D. Thesis, Dept. of Mathematics, Statistics and Philosophy, University of Tampere. [ix, 241]

Isotalo, J., Möls, M. & Puntanen, S. (2006). Invariance of the BLUE under the linear fixed and mixed effects models. *Acta et Commentationes Universitatis Tartuensis de Mathematica*, 10, 69–76. [256]

Isotalo, J. & Puntanen, S. (2006a). Linear prediction sufficiency for new observations in the general Gauss–Markov model. *Communications in Statistics: Theory and Methods*, 35, 1011–1024. [247, 251, 252, 258, 259]

Isotalo, J. & Puntanen, S. (2006b). Linear sufficiency and completeness in the partitioned linear model. *Acta et Commentationes Universitatis Tartuensis de Mathematica*, 10, 53–67. [266]

Isotalo, J. & Puntanen, S. (2009). A note on the equality of the OLSE and the BLUE of the parametric function in the general Gauss–Markov model. *Statistical Papers*, 50, 185–193. [222]

Isotalo, J., Puntanen, S. & Styan, G. P. H. (2006). On the role of the constant term in linear regression. In *Festschrift for Tarmo Pukkila on his 60th Birthday* (E. P. Liski, J. Isotalo, J. Niemelä, S. Puntanen & G. P. H. Styan, eds.), Dept. of Mathematics, Statistics and Philosophy, University of Tampere, pp. 243–259. [95]

Isotalo, J., Puntanen, S. & Styan, G. P. H. (2007). Effect of adding regressors on the equality of the OLSE and BLUE. *International Journal of Statistical Sciences*, 6, 193–201. [337, 341]

Isotalo, J., Puntanen, S. & Styan, G. P. H. (2008a). The BLUE's covariance matrix revisited: a review. *Journal of Statistical Planning and Inference*, 138, 2722–2737. [312, 368, 378, 384]

Isotalo, J., Puntanen, S. & Styan, G. P. H. (2008b). Decomposing matrices with Jerzy K. Baksalary. *Discussiones Mathematicae – Probability and Statistics*, 28, 91–111. [ix]

Isotalo, J., Puntanen, S. & Styan, G. P. H. (2008c). Formulas useful for linear regression analysis and related matrix theory, fourth edition. Report A 384, Dept. of Mathematics and Statistics, University of Tampere, Finland. [16]

Isotalo, J., Puntanen, S. & Styan, G. P. H. (2008d). A useful matrix decomposition and its statistical applications in linear regression. *Communications in Statistics: Theory and Methods*, 37, 1436–1457. [318, 419]

Isotalo, J., Puntanen, S. & Styan, G. P. H. (2009). Some comments on the Watson efficiency of the ordinary least squares estimator under the Gauss–Markov model. *Calcutta Statistical Association Bulletin*, 61, 1–15. (Proceedings of the Sixth International Triennial Calcutta Symposium on Probability and Statistics, 29–31 December, 2006, N. Mukhopadhyay and M. Pal, eds.). [333]

Ito, K. & Kunisch, K. (2008). *Lagrange Multiplier Approach to Variational Problems and Applications.* Society for Industrial and Applied Mathematics (SIAM). [vi]

Izenman, A. J. (2008). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning.* Springer. [413]

Jaeger, A. & Krämer, W. (1998). A final twist on the equality of OLS and GLS. *Statistical Papers*, 39, 321–324. [265]

Jensen, B. C. & McDonald, J. B. (1976). A pedagogigal example of heteroskedasticity and autocorrelation. *The American Statistician*, 30, 192–193. [261]

Jensen, S. T. (1999). The Laguerre–Samuelson inequality with extensions and applications in statistics and matrix theory. MSc Thesis, Dept. of Mathematics & Statistics, McGill University, Montréal. [420]

Jensen, S. T. & Styan, G. P. H. (1999). Some comments and a bibliography on the Laguerre–Samuelson inequality with extensions and applications in statistics and matrix theory. In *Analytic and Geometric Inequalities and Applications* (T. M. Rassias & H. M. Srivastava, eds.), Kluwer, pp. 151–181. [420]

Jewell, N. P. & Bloomfield, P. (1983). Canonical correlations of past and future for time series: definitions and theory. *The Annals of Statistics*, 11, 837–847. [387]

Jiang, J. (1997). A derivation of BLUP – best linear unbiased predictor. *Statistics & Probability Letters*, 32, 321–324. [256]

Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications.* Springer. [256]

Johnson, R. A. & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*, 6th Ed. Pearson Prentice Hall. [132]

Jolliffe, I. T. (2002). *Principal Component Analysis*, Second Ed. Springer. [203]

Kala, R. (1981). Projectors and linear estimation in general linear models. *Communications in Statistics: Theory and Methods*, 10, 849–873. [89, 139, 156, 271]

Kala, R. (2008). On commutativity of projectors. *Discussiones Mathematicae – Probability and Statistics*, 28, 157–165. [156]

Kala, R. (2009). On nested block designs geometry. *Statistical Papers*, 50, 805–815. [190]

Kala, R. & Pordzik, P. (2006). Two local operators and the BLUE. *Linear Algebra and its Applications*, 417, 134–139. [156]

Kala, R. & Pordzik, P. (2009). Estimation in singular partitioned, reduced or transformed linear models. *Statistical Papers*, 50, 633–638. [331]

Kanto, A. J. & Puntanen, S. (1983). A connection between the partial correlation coefficient and the correlation coefficient of certain residuals. *Communications in Statistics: Simulation and Computation*, 12, 639–641. [179]

Kanto, A. J. & Puntanen, S. (1985). An interesting feature of the partial correlation coefficient. In *Posters: Fourth International Symposium on Data Analysis and Informatics, Edition Provisoire (Versailles 1985), Editeur*, Centre de Rocquencourt, Centre de Rennes and Centre de Sophia-Antipolis, INRIA, Versailles, pp. 105–108. [179]

Kantorovich, L. V. (1948). Funkcional'nyi analiz i prikladnaya matematika (in Russian). *Uspekhi Matematičeskiĭ Nauk, Novaya Seriya*, 3, 89–185. See pp. 142–144. English translation: Kantorovich, 1952. [418]

Kantorovich, L. V. (1952). *Functional Analysis and Applied Mathematics.* U.S. Dept. of Commerce National Bureau of Standards. Translated by C. D. Benster. [418]

Kempthorne, O. (1976). Best linear unbiased estimation with arbitrary variance matrix. In *Essays in Probability and Statistics: A Volume in Honor of Professor Junjiro Ogawa* (S. Ikeda, ed.), Shinko Tsusho, Tokyo, pp. 203–225. [216]

Kempthorne, O. (1989). Comment [on Puntanen and Styan (1989)]. *The American Statistician*, 43, 161–162. [215]

Khatri, C. G. (1966). A note on a MANOVA model applied to problems in growth curves. *Annals of the Institute of Statistical Mathematics*, 18, 75–86. [323]

Khatri, C. G. (1976). A note on multiple and canonical correlation for a singular covariance matrix. *Psychometrika*, 41, 465–470. [386]

Khatri, C. G. (1978). Some optimization problems with applications to canonical correlations and sphericity tests. *Journal of Multivariate Analysis*, 8, 453–476. [Corrigendum: (1982), 12, p. 612]. [386]

Khatri, C. G. (1981). Study of F-tests under dependent model. *Sankhyā, Ser. A*, 43, 107–110. [342]

Khatri, C. G. (1982). A representation of a matrix and its use in the Gauss–Markoff model. *Journal of the Indian Statistical Association*, 20, 89–98. [323]

Khatri, C. G. (1989). Study of redundancy of vector variables in canonical correlations. *Communications in Statistics: Theory and Methods*, 18, 1425–1440. [386]

Khatri, C. G. (1990). Some properties of BLUE in a linear model and canonical correlations associated with linear transformations. *Journal of Multivariate Analysis*, 34, 211–226. [323, 386]

Khatri, C. G. & Rao, C. R. (1981). Some extensions of the Kantorovich inequality and statistical applications. *Journal of Multivariate Analysis*, 11, 498–505. [239, 425]

Khatri, C. G. & Rao, C. R. (1982). Some generalizations of the Kantorovich inequality. *Sankhyā, Ser. A*, 44, 91–102. [239, 425]

Khattree, R. (2001). On the calculation of antieigenvalues and antieigenvectors. *Journal of Interdisciplinary Mathematics*, 4, 195–199. [237]

Khattree, R. (2002). On generalized antieigenvalue and antieigenmatrix of order r. *American Journal of Mathematical and Management Sciences*, 22, 89–98. [237]

Khattree, R. (2003). Antieigenvalues and antieigenvectors in statistics. *Journal of Statistical Planning and Inference*, 114, 131–144. [237]

Khuri, A. I. (2009). *Linear Model Methodology.* Chapman & Hall/CRC. [vii]

Kloetzel, J. E., ed. (2010). *Scott Standard Postage Stamp Catalogue.* Scott Publishing. [Currently published annually in 6 volumes (on paper)]. [x]

Knott, M. (1975). On the minimum efficiency of least squares. *Biometrika*, 62, 129–132. [238, 239]

Koch, G. G. (1969). A useful lemma for proving the equality of two matrices with applications to least squares type quadratic forms. *Journal of the American Statistical Association*, 64, 969–970. [323, 340]

Kollo, T. & von Rosen, D. (2005). *Advanced Multivariate Statistics with Matrices.* Springer. [vii]

Krämer, W. (1980a). Finite sample efficiency of ordinary least squares in the linear regression model with autocorrelated errors. *Journal of the American Statistical Association*, 75, 1005–1009. [241]

Krämer, W. (1980b). A note on the equality of ordinary least squares and Gauss–Markov estimates in the general linear model. *Sankhyā, Ser. A*, 42, 130–131. [265]

Krämer, W. (1982). Note on estimating linear trend when residuals are autocorrelated. *Econometrica*, 50, 1065–1067. [241]

Krämer, W. (1984). High correlation among errors and the efficiency of ordinary least squares in linear models. *Statistical Papers*, 25, 135–142. [241]

Krämer, W. (1986). Least squares regression when the independent variable follows an ARIMA process. *Journal of the American Statistical Association*, 81, 150–154. [241]

Krämer, W. & Donninger, C. (1987). Spatial autocorrelation among errors and the relative efficiency of OLS in the linear regression model. *Journal of the American Statistical Association*, 82, 577–579. [241]

Krein, M. G. (1947). The theory of self-adjoint extensions of semi-bounded Hermitian transformations and its applications, I. *Matematicheskii Sbornik, Novaya Seria*, 20(62), 431–495. [313]

Kruskal, W. (1968). When are Gauss–Markov and least squares estimators identical? A coordinate-free approach. *The Annals of Mathematical Statistics*, 39, 70–75. [215, 221]

Kruskal, W. (1975). The geometry of generalized inverses. *Journal of the Royal Statistical Society, Ser. B*, 37, 272–283. [Corrigendum: (1986), 48 (2), p. 258]. [114]

LaMotte, L. R. (2007). A direct derivation of the REML likelihood function. *Statistical Papers*, 48, 321–327. [318]

Larocca, R. (2005). Reconciling conflicting Gauss–Markov conditions in the classical linear regression model. *Political Analysis*, 13, 188–207. [227]

Latour, D., Puntanen, S. & Styan, G. P. H. (1987). Equalities and inequalities for the canonical correlations associated with some partitioned generalized inverses of a covariance matrix. In *Proceedings of the Second International Tampere Conference in Statistics* (T. Pukkila & S. Puntanen, eds.), Dept. of Mathematical Sciences, University of Tampere, pp. 541–553. [386, 387]

Laywine, C. F. & Mullen, G. L. (1998). *Discrete Mathematics using Latin Squares.* Wiley. [68]

Ledermann, W. (1983). Issai Schur and his school in Berlin. *Bulletin of the London Mathematical Society*, 15, 97–106. [293]

de Leeuw, J. (1982). Generalized eigenvalue problems with positive semidefinite matrices. *Psychometrika*, 47, 87–93. [368, 377]

Lehman, L. M. & Young, D. M. (1993). A note on on a complete solution to the problem of robustness of Grubbs' test for a single outlier. *The Journal of the Industrial Mathematics Society*, 43, 11–13. [342]

Li, K.-C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, 86, 316–342. [199]

Liski, E. P. & Puntanen, S. (1989). A further note on a theorem on the difference of the generalized inverses of two nonnegative definite matrices. *Communications in Statistics: Theory and Methods*, 18, 1747–1751. [312, 315]

Liski, E. P., Puntanen, S. & Wang, S.-G. (1992). Bounds for the trace of the difference of the covariance matrices of the OLSE and BLUE. *Linear Algebra and its Applications*, 176, 121–130. [240, 261]

Liu, J. (2002). Linear sufficiency in the general Gauss–Markov model. *Applied Mathematics – A Journal of Chinese Universities*, 17, 85–92. [258]

Liu, S. (2000a). Efficiency comparisons between the OLSE and the BLUE in a singular linear model. *Journal of Statistical Planning and Inference*, 84, 191–200. [238, 241]

Liu, S. (2000b). On matrix trace Kantorovich-type inequalities. In *Innovations in Multivariate Statistical Analysis: A Festschrift for Heinz Neudecker* (R. D. H. Heijmans, D. S. G. Pollock & A. Satorra, eds.), Kluwer, pp. 39–50. [238, 241, 425]

Liu, S. & King, M. L. (2002). Two Kantorovich-type inequalities and efficiency comparisons between the OLSE and BLUE. *Journal of Inequalities and Application*, 7, 169–177. [241, 425]

Liu, S., Lu, C.-Y. & Puntanen, S. (2009). Matrix trace Wielandt inequalities with statistical applications. *Journal of Statistical Planning and Inference*, 139, 2254–2260. [425]

Liu, S. & Neudecker, H. (1995). Matrix-trace Cauchy–Schwarz inequalities and applications in canonical correlation analysis. *Statistical Papers*, 36, 287–298. [238, 425]

Liu, S. & Neudecker, H. (1997). Kantorovich inequalities and efficiency comparisons for several classes of estimators in linear models. *Statistica Neerlandica*, 51, 345–355. [238, 425]

Liu, S. & Neudecker, H. (1999). A survey of Cauchy–Schwarz and Kantorovich-type matrix inequalities. *Statistical Papers*, 40, 55–73. [425]

Liu, Y. (2009). On equality of ordinary least squares estimator, best linear unbiased estimator and best linear unbiased predictor in the general linear model. *Journal of Statistical Planning and Inference*, 139, 1522–1529. [249]

Llinarès, A. (1969). *Raymond Llulle: Doctrine d'Enfant*. C. Klincksieck, Paris. ["Version médiévale du ms. fr. 22933 de la B. N. de Paris"]. [67]

Loly, P. D. & Styan, G. P. H. (2010a). Comments on 4 × 4 philatelic Latin squares. *Chance*, 23 (1), 57–62. [67]

Loly, P. D. & Styan, G. P. H. (2010b). Comments on 5 × 5 philatelic Latin squares. *Chance*, 23 (2), 58–62. [68]

Lovell, M. C. (1963). Seasonal adjustment of economic time series and multiple regression analysis. *Journal of the American Statistical Association*, 58, 993–1010. [163, 331]

Lovell, M. C. (2008). A simple proof of the FWL Theorem. *The Journal of Economic Education*, 39, 88–91. [163, 331]

Lowerre, J. (1974). Some relationships between BLUEs, WLSEs, and SLSEs. *Journal of the American Statistical Association*, 69, 223–225. [149]

Löwner, K. (1934). Über monotone Matrixfunktionen. *Mathematische Zeitschrift*, 38, 443–446. [12]

Luati, A. & Proietti, T. (2010). On the equivalence of the weighted least squares and the generalised least squares estimators, with applications to kernel smoothing. *Annals of the Institute of Statistical Mathematics*, 62, in press. [241]

Luskin, R. C. (2008). Wouldn't it be nice . . . ? The automatic unbiasedness of OLS (and GLS). *Political Analysis*, 16, 345–349. [227]

Magness, T. A. & McGuire, J. B. (1962). Comparison of least squares and minimum variance estimates of regression parameters. *The Annals of Mathematical Statistics*, 33, 462–470. [Acknowledgement of Priority: (1963), 34, p. 352]. [238, 426]

Magnus, J. R. & Neudecker, H. (1999). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley. Revised edition of the 1988 original. [vi, vii]

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2, 49–55. [25]

Mäkeläinen, T. (1970a). Extrema for characteristic roots of product matrices. *Commentationes Physico-Mathematicae, Societas Scientiarum Fennica*, 38, 27–53. [398]

Mäkeläinen, T. (1970b). Projections and generalized inverses in the general linear model. *Commentationes Physico-Mathematicae, Societas Scientiarum Fennica*, 38, 13–25. [156]

Mandel, J. (1982). Use of the singular value decomposition in regression analysis. *The American Statistician*, 36, 15–24. [393]

Marcus, M. & Minc, H. (1992). *A Survey of Matrix Theory and Matrix Inequalities*. Dover Publications. Corrected reprint of the 1969 edition. [415]

Margolis, M. S. (1979). Perpendicular projections and elementary statistics. *The American Statistician*, 33, 131–135. [183, 213]

Markiewicz, A. (2001). On dependence structures preserving optimality. *Statistics & Probability Letters*, 53, 415–419. [342]

Markiewicz, A. & Puntanen, S. (2009). Admissibility and linear sufficiency in linear model with nuisance parameters. *Statistical Papers*, 50, 847–854. [185, 260]

Markiewicz, A., Puntanen, S. & Styan, G. P. H. (2010). A note on the interpretation of the equality of OLSE and BLUE. *The Pakistan Journal of Statistics*, 26, 127–134. [272]

Marsaglia, G. (1964). Conditional means and covariances of normal variables with singular covariance matrix. *Journal of the American Statistical Association*, 59, 1203–1204. [194]

Marsaglia, G. & Styan, G. P. H. (1972). When does $\operatorname{rank}(A+B) = \operatorname{rank}(A) + \operatorname{rank}(B)$? *Canadian Mathematical Bulletin*, 15, 451–452. [352]

Marsaglia, G. & Styan, G. P. H. (1974a). Equalities and inequalities for ranks of matrices. *Linear and Multilinear Algebra*, 2, 269–292. [123, 143, 146, 300, 350, 352, 354]

Marsaglia, G. & Styan, G. P. H. (1974b). Rank conditions for generalized inverses of partitioned matrices. *Sankhyā, Ser. A*, 36, 437–442. [294]

Marshall, A. W. & Olkin, I. (1990). Matrix versions of the Cauchy and Kantorovich inequalities. *Aequationes Mathematicae*, 40, 89–93. [237, 424]

Marshall, A. W., Olkin, I. & Arnold, B. C. (2011). *Inequalities: Theory of Majorization and Its Applications*, Second Ed. Springer. [13]

Martin, R. J. (1992). Leverage, influence and residuals in regression models when observations are correlated. *Communications in Statistics: Theory and Methods*, 21, 1183–1212. [338]

Mathew, T. (1983). Linear estimation with an incorrect dispersion matrix in linear models with a common linear part. *Journal of the American Statistical Association*, 78, 468–471. [227]

Mathew, T. (1985). On inference in a general linear model with an incorrect dispersion matrix. In *Linear Statistical Inference: Proceedings of the International Conference held at Poznań, Poland, June 4–8, 1984* (T. Caliński & W. Klonecki, eds.), Springer, pp. 200–210. [342]

Mathew, T. & Bhimasankaram, P. (1983a). On the robustness of LRT in singular linear models. *Sankhyā, Ser. A*, 45, 301–312. [334, 342]

Mathew, T. & Bhimasankaram, P. (1983b). On the robustness of the LRT with respect to specification errors in a linear model. *Sankhyā, Ser. A*, 45, 212–225. [342]

Mathew, T. & Bhimasankaram, P. (1983c). Optimality of BLUEs in a general linear model with incorrect design matrix. *Journal of Statistical Planning and Inference*, 8, 315–329. [271]

Mathew, T., Rao, C. R. & Sinha, B. K. (1984). Admissible linear estimation in singular linear models. *Communications in Statistics: Theory and Methods*, 13, 3033–3045. [260]

McCulloch, C., Hedayat, S. & Wells, M. (2008). Obituary: Walter Federer, 1915–2008. *The IMS Bulletin*, 37, 13. [357]

McCulloch, C. E., Searle, S. R. & Neuhaus, J. M. (2008). *Generalized, Linear, and Mixed Models*, Second Ed. Wiley. [256]

McDonald, R. P., Torii, Y. & Nishisato, S. (1979). Some results on proper eigenvalues and eigenvectors with applications to scaling. *Psychometrika*, 44, 211–227. [368]

McElroy, F. W. (1967). A necessary and sufficient condition that ordinary least-squares estimators be best linear unbiased. *Journal of the American Statistical Association*, 62, 1302–1304. [227]

Meyer, C. D. (2000). *Matrix Analysis and Applied Linear Algebra*. Society for Industrial and Applied Mathematics (SIAM). [vii, 10, 134, 358]

Miao, J. M. & Ben-Israel, A. (1992). On principal angles between subspaces in $\mathbb{R}^n$. *Linear Algebra and its Applications*, 171, 81–98. [132, 134]

Miao, J. M. & Ben-Israel, A. (1996). Product cosines of angles between subspaces. *Linear Algebra and its Applications*, 237/238, 71–81. [134]

Miller, J. (2010). Images of Mathematicians on Postage Stamps. New Port Richey, Florida. http://jeff560.tripod.com/stamps.html. [vi]

Milliken, G. A. & Akdeniz, F. (1977). A theorem on the difference of the generalized inverses of two nonnegative matrices. *Communications in Statistics: Theory and Methods*, 6, 1747–1751. [315]

Mirsky, L. (1960). Symmetric gauge functions and unitarily invariant norms. *The Quarterly Journal of Mathematics, Oxford, Second Series*, 11, 50–59. [401]

Mitra, S. (1988). The relationship between the multiple and the zero-order correlation coefficients. *The American Statistician*, 42, 89. [309]

Mitra, S. K. (1972). Fixed rank solutions of linear matrix equations. *Sankhyā, Ser. A*, 34, 387–392. [316]

Mitra, S. K. (1973a). Statistical proofs of some propositions on nonnegative definite matrices. *Bulletin of the International Statistical Institute*, 45, 206–211. [viii, 315]

Mitra, S. K. (1973b). Unified least squares approach to linear estimation in a general Gauss–Markov model. *SIAM Journal on Applied Mathematics*, 25, 671–680. [44, 346]

Mitra, S. K. (1982). Properties of the fundamental bordered matrix used in linear estimation. In *Statistics and Probability: Essays in Honor of C. R. Rao* (G. Kallianpur, P. R. Krishnaiah & J. K. Ghosh, eds.), North-Holland, pp. 505–509. [264]

Mitra, S. K. (1986). The minus partial order and the shorted matrix. *Linear Algebra and its Applications*, 83, 1–27. [316]

Mitra, S. K. (1989). Block independence in generalized inverse: a coordinate free look. In *Statistical Data Analysis and Inference* (Y. Dodge, ed.), North-Holland, pp. 429–443. [264]

Mitra, S. K. (1993). Transformations of the range space and the direct orthogonal projector. In *Statistics and Probability: A Raghu Raj Bahadur Festschrift* (J. K. Ghosh, S. K. Mitra, K. R. Parthasarathy & B. L. S. Prakasa Rao, eds.), Wiley Eastern, pp. 463–474. [156]

Mitra, S. K., Bhimasankaram, P. & Malik, S. B. (2010). *Matrix Partial Orders, Shorted Operators and Applications*. World Scientific. [vii, 314]

Mitra, S. K. & Moore, B. J. (1973). Gauss–Markov estimation with an incorrect dispersion matrix. *Sankhyā, Ser. A*, 35, 139–152. [144, 269, 270, 271, 280, 281, 368]

Mitra, S. K. & Puntanen, S. (1991). The shorted operator statistically interpreted. *Calcutta Statistical Association Bulletin*, 40, 97–102. [viii, 314]

Mitra, S. K., Puntanen, S. & Styan, G. P. H. (1995). Shorted matrices and their applications in linear statistical models: a review. In *Multivariate Statistics and Matrices in Statistics: Proceedings of the 5th Tartu Conference, Tartu–Pühajärve, Estonia, 23–28 May 1994* (E.-M. Tiit, T. Kollo & H. Niemi, eds.), TEV & VSP, pp. 289–311. [314, 315]

Mitra, S. K. & Puri, M. L. (1979). Shorted operators and generalized inverses of matrices. *Linear Algebra and its Applications*, 25, 45–56. [313]

Mitra, S. K. & Puri, M. L. (1982). Shorted matrices – an extended concept and some applications. *Linear Algebra and its Applications*, 42, 57–79. [313]

Mitra, S. K. & Rao, C. R. (1968). Simultaneous reduction of a pair of quadratic forms. *Sankhyā, Ser. A*, 30, 313–322. [367, 371, 376]

Mitra, S. K. & Rao, C. R. (1969). Conditions for optimality and validity of simple least squares theory. *The Annals of Mathematical Statistics*, 40, 1617–1624. [215]

Mitra, S. K. & Rao, C. R. (1974). Projections under seminorms and generalized Moore–Penrose inverses. *Linear Algebra and its Applications*, 9, 155–167. [75, 87, 156]

Moler, C. (2006). Professor SVD (Gene H. Golub). *The MathWorks News & Notes*, October 2006. Available online. [393]

Möls, M. (2004). Linear mixed models with equivalent predictors. *Dissertationes Mathematicae Universitatis Tartuensis*, 36. Ph.D. Thesis, University of Tartu. [278, 281]

Monahan, J. F. (2008). *A Primer on Linear Models*. Chapman & Hall/CRC. [vii, 347]

Moore, E. H. (1920). On the reciprocal of the general algebraic matrix (Abstract). *Bulletin of the American Mathematical Society*, 26, 890–891. [35, 107]

Moore, E. H. (1935). *General Analysis, Part 1* (with editorial collaboration of R. W. Barnard). Memoirs of the American Philosophical Society, Vol. 1 [Cf. pp. 197–209]. [107]

Morgenthaler, S. & Clerc-Bérod, A. (1997). A close look at the hat matrix. *Student*, 2, 1–12. [158]

Moyé, L. A. (2006). *Statistical Monitoring of Clinical Trials: Fundamentals for Investigators*. Springer. [263]

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*. Wiley. [26, 194, 232]

Mukhopadhyay, N. (2010). When finiteness matters: counterexamples to notions of covariance, correlation, and independence. *The American Statistician*, 64, 231–233. [53]

Müller, J. (1987). Sufficiency and completeness in the linear model. *Journal of Multivariate Analysis*, 21, 312–323. [258]

Müller, J., Rao, C. R. & Sinha, B. K. (1984). Inference on parameters in a linear model: a review of recent results. In *Experimental Design, Statistical Models, and Genetic Statistics* (K. Hinkelmann, ed.), Dekker, pp. 277–295. [258]

Mustonen, S. (1995). *Tilastolliset monimuuttujamenetelmät* [Statistical Multivariate Methods, in Finnish]. Survo Systems. [392, 407]

Mustonen, S. (1997). A measure for total variability in multivariate normal distribution. *Computational Statistics & Data Analysis*, 23, 321–334. [212]

Nelder, J. A. (1985). An alternative interpretation of the singular-value decomposition in regression. *The American Statistician*, 39, 63–64. [393]

Neudecker, H. (1997). Mahalanobis distance for multinomial data, Problem 97.5.4. *Econometric Theory*, 13, 890–891. [50]

Neudecker, H. & Satorra, A. (2003). On best affine prediction. *Statistical Papers*, 44, 257–266. [213]

Neudecker, H., Trenkler, G. & Liu, S. (2009). Problem 1/SP09: Inverse and determinant of a special symmetric matrix. *Statistical Papers*, 50, 221. [263]

Neudecker, H. & van de Velden, M. (2000). Relationship satisfied by two representations of a positive semi-definite matrix, Problem 00.1.3. *Econometric Theory*, 16, 143. [397]

Neudecker, H. & van de Velden, M. (2001). Relationship satisfied by two representations of a positive semi-definite matrix; Solution to Problem 00.1.3. *Econometric Theory*, 17, 281–282. [397]

von Neumann, J. (1937). Some matrix-inequalities and metrization of matrix space. *Tomsk University Review*, 1, 286–300. Reprinted in A. H. Taub (ed.) *John von Neumann, Collected Works, vol. IV: Continuous Geometry and Other Topics*, Pergamon Press, 1962, pp. 205–219. [410]

Nurhonen, M. & Puntanen, S. (1991). Remarks on the invariance of regression diagnostics. In *A Spectrum of Statistical Thought: Essays in Statistical Theory, Economics and Population Genetics in Honour of Johan Fellman* (G. Rosenqvist, K. Juselius, K. Nordström & J. Palmgren, eds.), Swedish School of Economics and Business Administration, Helsinki, No. 46, pp. 189–199. [342]

Nurhonen, M. & Puntanen, S. (1992a). Effect of deleting an observation on the equality of the OLSE and BLUE. *Linear Algebra and its Applications*, 176, 131–136. [337, 341]

Nurhonen, M. & Puntanen, S. (1992b). A property of partitioned generalized regression. *Communications in Statistics: Theory and Methods*, 21, 1579–1583. [331]

Ogasawara, T. & Takahashi, M. (1951). Independence of quadratic quantities in a normal system. *Journal of Science of the Hiroshima University, Series A: Mathematics, Physics, Chemistry*, 15, 1–9. [355]

Ogawa, J. & Olkin, I. (2008). A tale of two countries: The Craig–Sakamoto–Matusita theorem. *Journal of Statistical Planning and Inference*, 138, 3419–3428. [19]

Okamoto, M. (1969). Optimality of principal components. In *Multivariate Analysis – II* (P. R. Krishnaiah, ed.), Academic Press, pp. 673–685. [411]

Okamoto, M. & Kanazawa, M. (1968). Minimization of eigenvalues of a matrix and optimality of principal components. *The Annals of Mathematical Statistics*, 39, 859–863. [411, 412]

Olkin, I. (1981). Range restrictions for product-moment correlation matrices. *Psychometrika*, 46, 469–472. [307]

Olkin, I. (1990). Interface between statistics and linear algebra. In *Matrix Theory and Applications* (C. R. Johnson, ed.), American Mathematical Society, pp. 233–256. [viii]

Olkin, I. (1992). A matrix formulation on how deviant an observation can be. *The American Statistician*, 46, 205–209. [420]

Olkin, I. (1998). Why is matrix analysis part of the statistics curriculum. *Student*, 2, 343–348. [viii]

Olkin, I. & Raveh, A. (2009). Bounds for how much influence an observation can have. *Statistical Methods and Applications*, 18, 1–11. [421]

Ouellette, D. V. (1981). Schur complements and statistics. *Linear Algebra and its Applications*, 36, 187–295. [293, 301]

Pan, J.-X. & Fang, K.-T. (2002). *Growth Curve Models and Statistical Diagnostics*. Springer. [vii]

Patterson, H. D. & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58, 545–554. [318]

Pearson, K. & Lee, A. (1903). On the laws of inheritance in man: I. Inheritance of physical characters. *Biometrika*, 2, 357–462. [201, 436]

Pečarić, J. E., Puntanen, S. & Styan, G. P. H. (1996). Some further matrix extensions of the Cauchy–Schwarz and Kantorovich inequalities, with some statistical applications. *Linear Algebra and its Applications*, 237/238, 455–476. [238, 425]

Penrose, R. (1955). A generalized inverse for matrices. *Proceedings of the Cambridge Philosophical Society*, 51, 406–413. [35, 108]

Phillips, P. C. B. (1990). The geometry of the equivalence of OLS and GLS in the linear model, Problem 90.4.7. *Econometric Theory*, 6, 489–490. [221]

Phillips, P. C. B. (1992). The geometry of the equivalence of OLS and GLS in the linear model; Solution to Problem 90.4.7, proposed by Peter C. B. Phillips. *Econometric Theory*, 8, 158–159. [221]

Pickover, C. A. (2002). *The Zen of Magic Squares, Circles, and Stars: An Exhibition of Surprising Structures Across Dimensions*. Princeton University Press. [Second printing in 2003 contains "Updates and breakthroughs" on pp. 395–401 not included in the first printing in 2002]. [55]

Piegorsch, W. W. & Casella, G. (1989). The early use of matrix diagonal increments in statistical problems. *SIAM Review*, 31, 428–434. [301]

Piziak, R. & Odell, P. L. (1999). Full rank factorization of matrices. *Mathematics Magazine*, 73, 193–201. [350]

Piziak, R. & Odell, P. L. (2007). *Matrix Theory: From Generalized Inverses to Jordan Form*. Chapman & Hall/CRC. [vii]

Poincaré, H. (1890). Sur les équations aux dérivées partielles de la physique mathematiqué. *American Journal of Mathematics*, 12, 211–294. [399]

Preece, D. A. (1990). R. A. Fisher and experimental design: a review. *Biometrics*, 46, 925–935. [68]

Pringle, R. M. & Rayner, A. A. (1970). Expressions for generalized inverses of a bordered matrix with application to the theory of constrained linear models. *SIAM Review*, 12, 107–115. [264]

Pringle, R. M. & Rayner, A. A. (1971). *Generalized Inverse Matrices with Applications to Statistics*. Charles Griffin & Hafner. [264]

Pukelsheim, F. (1977). Equality of two BLUEs and ridge-type estimates. *Communications in Statistics: Theory and Methods*, 6, 603–610. [229]

Pukelsheim, F. (1993). *Optimal Design of Experiments*. Wiley. Reprinted by SIAM in series Classics in Applied Mathematics, 2006. [312]

Puntanen, S. (1987). On the relative goodness of ordinary least squares estimation in the general linear model. *Acta Universitatis Tamperensis Series A*, 216. Ph.D. Thesis, Dept. of Mathematical Sciences, University of Tampere. [ix, 136, 387]

Puntanen, S. (1996). Some matrix results related to a partitioned singular linear model. *Communications in Statistics: Theory and Methods*, 25, 269–279. [331]

Puntanen, S. (1997). Some further results related to reduced singular linear models. *Communications in Statistics: Theory and Methods*, 26, 375–385. [323, 331]

Puntanen, S. (2010a). Personal collection of photographs. Nokia, Finland. [x]

Puntanen, S. (2010b). Solution to Problem 4/SP08 (Problem proposed by Götz Trenkler and Dietrich Trenkler). *Statistical Papers*, 51, available online. [263]

Puntanen, S. & Scott, A. J. (1996). Some further remarks on the singular linear model. *Linear Algebra and its Applications*, 237/238, 313–327. [149, 289, 319, 326]

Puntanen, S., Seber, G. A. F. & Styan, G. P. H. (2007). Multivariate statistical analysis. In *Handbook of Linear Algebra* (L. Hogben, ed.), Chapman & Hall/CRC, Chap. 53, pp. 53.1–53.15. [viii, 16]

Puntanen, S., Šemrl, P. & Styan, G. P. H. (1996). Some remarks on the parallel sum of two matrices. In *Proceedings of the A. C. Aitken Centenary Conference (Dunedin, August 1995)* (L. Kavalieris, F. C. Lam, L. A. Roberts & J. A. Shanks, eds.), University of Otago Press, pp. 243–256. [289]

Puntanen, S. & Styan, G. P. H. (1989). The equality of the ordinary least squares estimator and the best linear unbiased estimator. *The American Statistician*, 43, 153–164. With comments by O. Kempthorne and S. R. Searle and a reply by the authors. [215]

Puntanen, S. & Styan, G. P. H. (1990). Reply [to R. Christensen (1990), R. W. Farebrother (1990), and D. A. Harville (1990)] (Letter to the Editor). *The American Statistician*, 44, 192–193. [39, 125]

Puntanen, S. & Styan, G. P. H. (2005a). Historical introduction: Issai Schur and the early development of the Schur complement. In *The Schur Complement and Its Applications* (F. Zhang, ed.), Springer, Chap. 0 & Bibliography, pp. 1–16, 259–288. [vii, 293, 303]

Puntanen, S. & Styan, G. P. H. (2005b). Schur complements in statistics and probability. In *The Schur Complement and Its Applications* (F. Zhang, ed.), Springer, Chap. 6 & Bibliography, pp. 163–226, 259–288. [vii, 13, 293, 307, 348]

Puntanen, S. & Styan, G. P. H. (2007). Random vectors and linear statistical models. In *Handbook of Linear Algebra* (L. Hogben, ed.), Chapman & Hall/CRC, Chap. 52, pp. 52.1–52.17. [viii, 16]

Puntanen, S. & Styan, G. P. H. (2008a). Foreword for the Special Issue devoted to Jerzy K. Baksalary. *Discussiones Mathematicae – Probability and Statistics*, 28, 85–90. [ix]

Puntanen, S. & Styan, G. P. H. (2008b). Stochastic stamps: a philatelic introduction to chance. *Chance*, 21 (1), 36–41. [vi]

Puntanen, S. & Styan, G. P. H. (2011). Best linear unbiased estimation in a linear model. In *International Encyclopedia of Statistical Science, Part 2* (M. Lovric, ed.), Springer, pp. 141–144. [215]

Puntanen, S., Styan, G. P. H. & Subak-Sharpe, G. E. (1998). Mahalanobis distance for multinomial data; Solution to Problem 97.5.4, proposed by Heinz Neudecker. *Econometric Theory*, 14, 695–698. [50, 120]

Puntanen, S., Styan, G. P. H. & Werner, H. J. (2000a). A comment on the article "Simple forms of the best linear unbiased predictor in the general linear regression model" by Silvia N. Elian (*The American Statistician*, 54, 25–28), Letter to the Editor. *The American Statistician*, 54, 327–328. [249]

Puntanen, S., Styan, G. P. H. & Werner, H. J. (2000b). Two matrix-based proofs that the linear estimator Gy is the best linear unbiased estimator. *Journal of Statistical Planning and Inference*, 88, 173–179. [218]

Puterman, M. L. (1988). Leverage and influence in autocorrelated regression models. *Applied Statistics*, 37, 76–86. [Corrigendum: (1991), 40, p. 237]. [338]

Randles, R. H. (2007). A conversation with Robert V. Hogg. *Statistical Science*, 24, 137–152. [19]

Rao, A. R. & Bhimasankaram, P. (2000). *Linear Algebra*, Second Ed. Hindustan Book Agency. [vii, 123, 146, 352]

Rao, C. R. (1962). A note on a generalized inverse of a matrix with applications to problems in mathematical statistics. *Journal of the Royal Statistical Society, Ser. B*, 24, 152–158. [35]

Rao, C. R. (1964). The use and interpretation of principal component analysis in applied research. *Sankhyā, Ser. A*, 26, 329–358. [203]

Rao, C. R. (1967). Least squares theory using an estimated dispersion matrix and its application to measurement of signals. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability (Berkeley, California, 1965/66), Vol. I: Statistics* (L. M. Le Cam & J. Neyman, eds.), University of California Press, Berkeley, pp. 355–372. [215, 323]

Rao, C. R. (1968). A note on a previous lemma in the theory of least squares and some further results. *Sankhyā, Ser. A*, 30, 259–266. [215, 271]

Rao, C. R. (1971a). Estimation of variance and covariance components – MINQUE theory. *Journal of Multivariate Analysis*, 1, 257–275. [45]

Rao, C. R. (1971b). Minimum variance quadratic unbiased estimation of variance components. *Journal of Multivariate Analysis*, 1, 445–456. [45]

Rao, C. R. (1971c). Unified theory of linear estimation. *Sankhyā, Ser. A*, 33, 371–394. [Corrigenda: (1972), 34, p. 194 and p. 477]. [44, 223, 264, 271]

Rao, C. R. (1972). Some recent results in linear estimation. *Sankhyā, Ser. B*, 34, 369–377. [264]

Rao, C. R. (1973a). *Linear Statistical Inference and its Applications*, Second Ed. Wiley. [vii, 18, 75, 123, 179, 195, 203, 208, 212, 264, 323, 370, 387, 398, 407]

Rao, C. R. (1973b). Representations of best linear unbiased estimators in the Gauss–Markoff model with a singular dispersion matrix. *Journal of Multivariate Analysis*, 3, 276–292. [41, 43, 141, 271, 325]

Rao, C. R. (1974). Projectors, generalized inverses and the BLUE's. *Journal of the Royal Statistical Society, Ser. B*, 36, 442–448. [6, 74, 87, 88, 124, 156]

Rao, C. R. (1976). Estimation of parameters in a linear model. *The Annals of Statistics*, 4, 1023–1037. [Corrigendum: (1979), 7, p. 696]. [259]

Rao, C. R. (1978). Choice of best linear estimators in the Gauss–Markoff model with a singular dispersion matrix. *Communications in Statistics: Theory and Methods*, 7, 1199–1208. [229]

Rao, C. R. (1979). Separation theorems for singular values of matrices and their applications in multivariate analysis. *Journal of Multivariate Analysis*, 9, 362–377. [401]

Rao, C. R. (1980). Matrix approximations and reduction of dimensionality in multivariate statistical analysis. In *Multivariate Analysis – V* (P. R. Krishnaiah, ed.), North-Holland, pp. 3–22. [398, 401, 411]

Rao, C. R. (1981). A lemma on g-inverse of a matrix and computation of correlation coefficients in the singular case. *Communications in Statistics: Theory and Methods*, 10, 1–10. [213, 386]

Rao, C. R. (1985a). The inefficiency of least squares: extensions of the Kantorovich inequality. *Linear Algebra and its Applications*, 70, 249–255. [240]

Rao, C. R. (1985b). A unified approach to inference from linear models. In *Proceedings of the First International Tampere Seminar on Linear Statistical Models and their Applications* (T. Pukkila & S. Puntanen, eds.), Dept. of Mathematical Sciences, University of Tampere, pp. 9–36. [114]

Rao, C. R. (1987). Estimation in linear models with mixed effects: a unified theory. In *Proceedings of the Second International Tampere Conference in Statistics* (T. Pukkila & S. Puntanen, eds.), Dept. of Mathematical Sciences, University of Tampere, pp. 73–98. [256]

Rao, C. R. (1995). A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance. *Qüestiió*, 19, 23–63. [413]

Rao, C. R. (2000). Statistical proofs of some matrix inequalities. *Linear Algebra and its Applications*, 321, 307–320. [viii]

Rao, C. R. (2006). Statistical proofs of some matrix theorems. *International Statistical Review*, 74, 169–185. [viii]

Rao, C. R. (2007). Antieigenvalues and antisingularvalues of a matrix and applications to problems in statistics. *Mathematical Inequalities & Applications*, 10, 471–489. [237, 425]

Rao, C. R. & Mitra, S. K. (1971a). Further contributions to the theory of generalized inverse of matrices and its applications. *Sankhyā, Ser. A*, 33, 289–300. [Corrigendum: (1972), 34, p. 477]. [229]

Rao, C. R. & Mitra, S. K. (1971b). *Generalized Inverse of Matrices and Its Applications*. Wiley. [vii, 106, 111, 114, 116, 137, 156, 222, 268, 280, 281, 284, 289, 342, 367, 371, 376]

Rao, C. R., Mitra, S. K. & Bhimasankaram, P. (1972). Determination of a matrix by its subclasses of generalized inverses. *Sankhyā, Ser. A*, 34, 5–8. [284]

Rao, C. R. & Rao, M. B. (1998). *Matrix Algebra and Its Applications to Statistics and Econometrics*. World Scientific. [vii, 86, 123, 389, 410, 411]

Rao, C. R., Toutenburg, H., Shalabh & Heumann, C. (2008). *Linear Models and Generalizations: Least Squares and Alternatives*, Third Ed. Springer. With contributions by Michael Schomaker. [vii, 273]

Rao, C. R. & Yanai, H. (1979). General definition and decomposition of projectors and some applications to statistical problems. *Journal of Statistical Planning and Inference*, 3, 1–17. [156]

Rao, C. R. & Yanai, H. (1985a). Generalized inverse of linear transformations: a geometric approach. *Linear Algebra and its Applications*, 66, 87–98. [114]

Rao, C. R. & Yanai, H. (1985b). Generalized inverses of partitioned matrices useful in statistical applications. *Linear Algebra and its Applications*, 70, 105–113. [264]

Raveh, A. (1985). On the use of the inverse of the correlation matrix in multivariate data analysis. *The American Statistician*, 39, 39–42. [172]

Rencher, A. C. & Schaalje, G. B. (2008). *Linear Models in Statistics*, Second Ed. Wiley. [vii]

Robinson, G. K. (1991). That BLUP is a good thing: the estimation of random effects (with discussion on pp. 32–51). *Statistical Science*, 6, 15–51. [255, 256]

Rong, J.-Y. & Liu, X.-Q. (2010). On misspecification of the dispersion matrix in mixed linear models. *Statistical Papers*, 51, 445–453. [278]

Ryan, T. P. (2009). *Modern Regression Methods*, Second Ed. Wiley. [vii, 199]

Samuelson, P. A. (1968). How deviant can you be? *Journal of the American Statistical Association*, 63, 1522–1525. [414, 420]

Satorra, A. & Neudecker, H. (2003). A matrix equality useful in goodness-of-fit testing of structural equation models. *Journal of Statistical Planning and Inference*, 114, 63–80. [323]

Schall, R. & Dunne, T. T. (1988). A unified approach to outliers in the general linear model. *Sankhyā, Ser. B*, 50, 157–167. [338]

Schmidt, E. (1907). Zur Theorie der linearen und nichtlinearen Integralgleichungen: I. Teil: Entwicklung willkürlicher Funktionen nach Systemen vorgeschriebener. *Mathematische Annalen*, 63, 433–476. [401]

Schönemann, P. H. (1966). A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, 31, 1–10. [401]

Schott, J. R. (2005). *Matrix Analysis for Statistics*, Second Ed. Wiley. [vii]

Schur, J. (1917). Über Potenzreihen, die im Innern des Einheitskreises beschränkt sind [I]. *Journal für die reine und angewandte Mathematik*, 147, 205–232. [293, 303]

Schwarz, H. A. (1888). Ueber ein die Flächen kleinsten Flächeninhalts betreffendes Problem der Variationsrechnung. *Acta Societatis Scientiarum Fennicæ*, 15, 315–362. [Preface dated 31 October 1885]. [415]

Scott, A. J. & Styan, G. P. H. (1985). On a separation theorem for generalized eigenvalues and a problem in the analysis of sample surveys. *Linear Algebra and its Applications*, 70, 209–224. [368, 398, 411]

Searle, S. R. (1956). Matrix methods in components of variance and covariance analysis. *The Annals of Mathematical Statistics*, 27, 737–748. [viii]

Searle, S. R. (1971). *Linear Models*. Wiley. Reprinted as Wiley Classics Library Edition, 1997. [34, 347]

Searle, S. R. (1978). A univariate formulation of the multivariate linear model. In *Contributions to Survey Sampling and Applied Statistics: Papers in Honor of H. O. Hartley* (H. A. David, ed.), Academic Press, pp. 181–189. [230]

Searle, S. R. (1982). *Matrix Algebra Useful for Statistics*. Wiley. [vii, 171, 358, 370, 387, 393]

Searle, S. R. (1989). Comment [on Puntanen and Styan (1989)]. *The American Statistician*, 43, 162–163. [216]

Searle, S. R. (1994). Extending some results and proofs for the singular linear model. *Linear Algebra and its Applications*, 210, 139–151. [149, 256]

Searle, S. R. (1996). Some follow-up on Aitken's least squares equations. In *Proceedings of the A. C. Aitken Centenary Conference (Dunedin, August 1995)* (L. Kavalieris, F. C. Lam, L. A. Roberts & J. A. Shanks, eds.), University of Otago Press, pp. 299–309. [256]

Searle, S. R. (1997). The matrix handling of BLUE and BLUP in the mixed linear model. *Linear Algebra and its Applications*, 264, 291–311. [255, 256]

Searle, S. R. (1999). Comments from thirty years of teaching matrix algebra to applied statisticians. *Image: The Bulletin of the ILAS*, 22, 3–5. [viii]

Searle, S. R. (2000). The infusion of matrices into statistics. *Image: The Bulletin of the ILAS*, 24, 25–32. [viii, 34]

Searle, S. R. (2005). Recollections from a 50-year random walk midst matrices, statistics and computing. In *Proceedings of the 14th International Workshop on Matrices and Statistics,* vol. 8 *of Research Letters in the Information and Mathematical Sciences* (P. S. P. Cowpertwait, ed.), Massey University, pp. 45–52. [viii]

Searle, S. R., Casella, G. & McCulloch, C. E. (1992). *Variance Components*. Wiley. [195, 212, 255, 323]

Seber, G. A. F. (1980). *The Linear Hypothesis: A General Theory*, Second Ed. Griffin. [vii, 156, 347]

Seber, G. A. F. (1984). *Multivariate Observations*. Wiley. [20, 156, 230, 232, 234, 386, 389, 412]

Seber, G. A. F. (2008). *A Matrix Handbook for Statisticians*. Wiley. [vii, 18, 26, 27, 28, 53, 358]

Seber, G. A. F. & Lee, A. J. (2003). *Linear Regression Analysis*, Second Ed. Wiley. [vii, 156, 168]

Seddighin, M. (2009). Antieigenvalue techniques in statistics. *Linear Algebra and its Applications*, 430, 2566–2580. [237]

SenGupta, A. (1991). Generalized correlations in the singular case. *Journal of Statistical Planning and Inference*, 28, 241–245. [368, 386]

Sengupta, D. & Jammalamadaka, S. R. (2003). *Linear Models: An Integrated Approach*. World Scientific. [vii, 65, 234, 258, 331]

Serre, D. (2007). *Matrices: Theory & Applications, Additional Exercises*. École Normale Supérieure. Translated from the 2001 French original. [vi]

Seshadri, V. & Styan, G. P. H. (1980). Canonical correlations, rank additivity and characterizations of multivariate normality. In *Analytic Function Methods in Probability Theory: Proc. Colloquium on the Methods of Complex Analysis in the Theory of Probability and Statistics held at the Kossuth L. University, Debrecen, Hungary, August 29–September 2, 1977*, vol. 21 of *Colloquia Mathematica Societatis János Bolyai*, North-Holland, pp. 331–344. [386]

Sherman, J. & Morrison, W. J. (1949). Adjustment of an inverse matrix corresponding to changes in the elements of a given column or a given row of the original matrix (abstract). *The Annals of Mathematical Statistics*, 20, 621. [301]

Sherman, J. & Morrison, W. J. (1950). Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics*, 21, 124–127. [301]

Shieh, G. (2001). The inequality between the coefficient of determination and the sum of squared simple correlation coefficients. *The American Statistician*, 55, 121–124. [309]

Sibuya, M. (1970). Subclasses of generalized inverses of matrices. *Annals of the Institute of Statistical Mathematics*, 22, 543–556. [156]

Smith, G. (1997). Do statistics test scores regress toward the mean? *Chance*, 10 (4), 42–45. [201]

Speed, T. P. (1991). Comment on "That BLUP is a good thing: the estimation of random effects", by G. K. Robinson (1991). *Statistical Science*, 6, 42–44. [256]

Speed, T. P. (2008a). Terence's stuff: PCA. *The IMS Bulletin*, 37, 9. [206]

Speed, T. P. (2008b). Terence's stuff: Two by two. *The IMS Bulletin*, 37, 17. [47]

Speed, T. P. (2009). Terence's stuff: Bibliographies. *The IMS Bulletin*, 38, 12. [ix]

Speed, T. P. (2010). Terence's stuff: The multivariate normal. *The IMS Bulletin*, 39, 10. [18]

Srivastava, M. S. (2002). *Methods of Multivariate Statistics.* Wiley. [vii]

Srivastava, M. S. & Khatri, C. G. (1979). *An Introduction to Multivariate Statistics.* North-Holland. [323]

Stapleton, J. H. (2009). *Linear Statistical Models*, Second Ed. Wiley. [vii]

Steele, J. M. (2004). *The Cauchy–Schwarz Master Class: An Introduction to the Art of Mathematical Inequalities.* Cambridge University Press. [4, 75, 415]

Stewart, G. W. (1987). Collinearity and least squares regression. *Statistical Science*, 2, 68–100. With discussion. [162]

Stewart, G. W. (1993). On the early history of the singular value decomposition. *SIAM Review*, 35, 551–566. [393, 400, 401, 411]

Stewart, G. W. (1998). *Matrix Algorithms. Volume I: Basic Decompositions.* Society for Industrial and Applied Mathematics (SIAM). [393, 395]

Stewart, G. W. (2001). *Matrix Algorithms. Volume II: Eigensystems.* Society for Industrial and Applied Mathematics (SIAM). [358, 395]

Stewart, G. W. & Sun, J. G. (1990). *Matrix Perturbation Theory.* Academic Press. [375, 401]

Stigler, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900.* The Belknap Press of Harvard University Press. [200]

Stigler, S. M. (1999). *Statistics on the Table: The History of Statistical Concepts and Methods.* Harvard University Press. [199, 200, 306]

Stępniak, C. (1985). Ordering of nonnegative definite matrices with application to comparison of linear models. *Linear Algebra and its Applications*, 70, 67–71. [312]

Styan, G. P. H. (1969). Multivariate Normal Inference with Correlation Structure. Ph.D. Thesis, Dept. of Mathematical Statistics, Columbia University, New York City. [ix]

Styan, G. P. H. (1970). Notes on the distribution of quadratic forms in singular normal variables. *Biometrika*, 57, 567–572. [355]

Styan, G. P. H. (1973a). Hadamard products and multivariate statistical analysis. *Linear Algebra and its Applications*, 6, 217–240. [vi]

Styan, G. P. H. (1973b). When does least squares give the best linear unbiased estimate? In *Proceedings of the Research Seminar at Dalhousie University, Halifax, Nova Scotia, March 23–25, 1972* (D. G. Kabe & R. P. Gupta, eds.), North-Holland, pp. 241–246. [221]

Styan, G. P. H. (1983). On some inequalities associated with ordinary least squares and the Kantorovich inequality. In *Festschrift for Eino Haikala on his Seventieth Birthday*

(P. Huuhtanen, E. P. Liski, T. Pukkila & S. Puntanen, eds.), Acta Universitatis Tamperensis, University of Tampere, pp. 158–166. [240]

Styan, G. P. H. (1985). Schur complements and linear statistical models. In *Proceedings of the First International Tampere Seminar on Linear Statistical Models and their Applications* (T. Pukkila & S. Puntanen, eds.), Dept. of Mathematical Sciences, University of Tampere, pp. 37–75. [293, 348, 387]

Styan, G. P. H. (2007). A philatelic introduction to magic squares and Latin squares for Euler's 300th birthyear. In *Proceedings of the Canadian Society for History and Philosphy of Mathematics: Volume 20* (A. Cupillari, ed.), CSHPM, pp. 306–319. [55]

Styan, G. P. H. (2010). Personal collection of philatelic items. Verdun (Québec), Canada. [x]

Styan, G. P. H., Boyer, C. & Chu, K. L. (2009). Some comments on Latin squares and on Graeco-Latin squares, illustrated with postage stamps and old playing cards. *Statistical Papers*, 50, 917–941. [67]

Styan, G. P. H. & Subak-Sharpe, G. E. (1997). Inequalities and equalities associated with the Campbell–Youla generalized inverse of the indefinite admittance matrix of resistive networks. *Linear Algebra and its Applications*, 250, 349–370. [120]

Takane, Y. (2004). Matrices with special reference to applications in psychometrics. *Linear Algebra and its Applications*, 388, 341–361. [401]

Takane, Y. & Yanai, H. (1999). On oblique projectors. *Linear Algebra and its Applications*, 289, 297–310. [156]

Takane, Y. & Yanai, H. (2005). On the Wedderburn–Guttman theorem. *Linear Algebra and its Applications*, 410, 267–278. [304]

Takane, Y. & Yanai, H. (2007). Alternative characterizations of the extended Wedderburn–Guttman theorem. *Linear Algebra and its Applications*, 422, 701–711. [304]

Takeuchi, K., Yanai, H. & Mukherjee, B. N. (1982). *The Foundations of Multivariate Analysis: A Unified Approach by Means of Projection onto Linear Subspaces*. Wiley, Halsted Press. [156]

Tanabe, K. & Sagae, M. (1992). An exact Cholesky decomposition and the generalized inverse of the variance-covariance matrix of the multinomial distribution, with applications. *Journal of the Royal Statistical Society, Ser. B*, 54, 211–219. [120]

Tee, G. J. (2003). Up with determinants! *Image: The Bulletin of the ILAS*, 30, 7–11. [300, 356]

Theil, H. (1968). On the use of incomplete information in regression analysis. *Journal of the American Statistical Association*, 58, 401–414. [265]

Theil, H. & Goldberger, A. S. (1961). On pure and mixed statistical estimation in economics. *International Economic Review*, 2, 65–78. [265]

Thomas, D. H. (1968). When do minimum variance estimators coincide? [Abstract]. *The Annals of Mathematical Statistics*, 39, 1365. [89]

Thompson, W. R. (1935). On a criterion for the rejection of observations and the distribution of the ratio of deviation to sample standard deviation. *The Annals of Mathematical Statistics*, 6, 214–219. [420]

Tian, Y. (2007). Some decompositions of OLSEs and BLUEs under a partitioned linear model. *International Statistical Review*, 75, 224–248. [87, 123]

Tian, Y. (2009a). On an additive decomposition of the BLUE in a multiple-partitioned linear model. *Journal of Multivariate Analysis*, 100, 767–776. [87, 123]

Tian, Y. (2009b). On equalities for BLUEs under misspecified Gauss–Markov models. *Acta Mathematica Sinica, English Series*, 25, 1907–1920. [271]

Tian, Y. (2010). Weighted least-squares estimators of parametric functions of the regression coefficients under a general linear model. *Annals of the Institute of Statistical Mathematics*, 62, 929–941. [87]

Tian, Y., Beisiegel, M., Dagenais, E. & Haines, C. (2008). On the natural restrictions in the singular Gauss–Markov model. *Statistical Papers*, 49, 553–564. [39]

Tian, Y. & Puntanen, S. (2009). On the equivalence of estimations under a general linear model and its transformed models. *Linear Algebra and its Applications*, 430, 2622–2641. [259]

Tian, Y. & Styan, G. P. H. (2001). Rank equalities for idempotent and involutory matrices. *Linear Algebra and its Applications*, 335, 101–117. [390]

Tian, Y. & Takane, Y. (2008a). On sum decompositions of weighted least-squares estimators for the partitioned linear model. *Communications in Statistics: Theory and Methods*, 37, 55–69. [87, 123]

Tian, Y. & Takane, Y. (2008b). Some properties of projectors associated with the WLSE under a general linear model. *Journal of Multivariate Analysis*, 99, 1070–1082. [87, 89, 123]

Tian, Y. & Takane, Y. (2009a). More on generalized inverses of partitioned matrices with Banachiewicz–Schur forms. *Linear Algebra and its Applications*, 430, 1641–1655. [295]

Tian, Y. & Takane, Y. (2009b). On V-orthogonal projectors associated with a seminorm. *Annals of the Institute of Statistical Mathematics*, 61, 517–530. [87, 89, 123]

Tian, Y. & Tian, Z. (2010). On additive and block decompositions of WLSEs under a multiple partitioned regression model. *Statistics*, 44, 361–379. [87]

Tian, Y. & Wiens, D. P. (2006). On equality and proportionality of ordinary least squares, weighted least squares and best linear unbiased estimators in the general linear model. *Statistics & Probability Letters*, 76, 1265–1272. [87, 123]

Toutenburg, H. & Shalabh (2009). *Statistical Analysis of Designed Experiments*, Third Ed. Springer. [vii, 51]

Trenkler, D. & Trenkler, G. (2001). Magic squares, melancholy and the Moore–Penrose inverse. *Image: The Bulletin of the ILAS*, 27, 3–10. [54]

Trenkler, G. (1994). Characterizations of oblique and orthogonal projectors. In *Proceedings of the International Conference on Linear Statistical Inference LINSTAT '93 (Poznań, 1993)* (T. Caliński & R. Kala, eds.), Kluwer, pp. 255–270. [156, 325]

Trenkler, G. (1995). On the singularity of the sample covariance matrix. *Journal of Statistical Computation and Simulation*, 52, 172–173. [129]

Trenkler, G. (2006). On oblique and orthogonal projectors. In *Contributions to Probability and Statistics: Applications and Challenges – Proceedings of the International Statistics Workshop held at University of Canberra, 4–5 April 2005* (P. Brown, S. Liu & D. Sharma, eds.), World Scientific, pp. 178–191. [112, 156]

Trenkler, G. & Puntanen, S. (2005). A multivariate version of Samuelson's inequality. *Linear Algebra and its Applications*, 410, 143–149. [421]

Trenkler, G. & Trenkler, D. (2008). Problem 4/SP08. *Statistical Papers*, 49, 803. [263]

Vehkalahti, K. (2000). Reliability of measurement scales: Tarkkonen's general method supersedes Cronbach's alpha. Ph.D. Thesis, Dept. of Statistics, University of Helsinki, Statistical Research Reports 17, Finnish Statistical Society. [422]

Vehkalahti, K., Puntanen, S. & Tarkkonen, L. (2007). Effects on measurement errors in predictor selection of linear regression model. *Computational Statistics & Data Analysis*, 52, 1183–1195. [401]

Vehkalahti, K., Puntanen, S. & Tarkkonen, L. (2009). Implications of dimensionality on measurement reliability. In *Statistical Inference, Econometric Analysis and Matrix Algebra: Festschrift in Honour of Götz Trenkler* (B. Schipp & W. Krämer, eds.), Physica-Verlag, pp. 143–160. [422]

Vik, K. (1924). Bedømmelse av feilene på forsøksfelter med og uten målestokk. *Meldinger fra Norges Landbrukshøgskole*, 4, 129–181. [68]

Vinograde, B. (1950). Canonical positive definite matrices under internal linear transformations. *Proceedings of the American Mathematical Society*, 1, 159–161. [397]

Wachsmuth, A., Wilkinson, L. & Dallal, G. E. (2003). Galton's bend: a previously undiscovered nonlinearity in Galton's family stature regression data. *The American Statistician*, 57, 190–192. [201]

Waller, N. G. & Jones, J. A. (2010). Correlation weights in multiple regression. *Psychometrika*, 75, 58–69. [390]

Wang, S.-G. & Chow, S.-C. (1987). Some results on canonical correlations and measures of multivariate association. *Communications in Statistics: Theory and Methods*, 16, 339–351. [386]

Wang, S.-G. & Chow, S.-C. (1994). *Advanced Linear Models: Theory and Applications*. Dekker. [vii]

Wang, S.-G. & Ip, W.-C. (1999). A matrix version of the Wielandt inequality and its applications to statistics. *Linear Algebra and its Applications*, 296, 171–181. [426]

Wang, S.-G. & Shao, J. (1992). Constrained Kantorovich inequalities and relative efficiency of least squares. *Journal of Multivariate Analysis*, 11, 284–298. [241, 425]

Wang, S.-G., Wu, M.-X. & Ma, W.-Q. (2003). Comparison of MINQUE and simple estimate of the error variance in the general linear model. *Acta Mathematica Sinica, English Series*, 19, 13–18. [340]

Watkins, D. S. (2007). Unsymmetric matrix eigenvalue techniques. In *Handbook of Linear Algebra* (L. Hogben, ed.), Chapman & Hall/CRC, Chap. 43, pp. 43.1–43.12. [375]

Watson, G. S. (1951). Serial correlation in regression analysis. Ph.D. Thesis, Dept. of Experimental Statistics, North Carolina State College, Raleigh. [238]

Watson, G. S. (1955). Serial correlation in regression analysis, I. *Biometrika*, 42, 327–341. [238, 239]

Watson, G. S. (1967). Linear least squares regression. *The Annals of Mathematical Statistics*, 38, 1679–1699. [215]

Watson, G. S. (1972). Prediction and the efficiency of least squares. *Biometrika*, 59, 91–98. [229, 249]

Watson, G. S. (1996a). Some singular value decompositions of statistical interest. In *Proceedings of the A. C. Aitken Centenary Conference (Dunedin, August 1995)* (L. Kavalieris, F. C. Lam, L. A. Roberts & J. A. Shanks, eds.), University of Otago Press, pp. 359–366. [253]

Watson, G. S. (1996b). Spectral decomposition of the covariance matrix of a multinomial. *Journal of the Royal Statistical Society, Ser. B*, 58, 289–291. [120]

Watson, G. S., Alpargu, G. & Styan, G. P. H. (1997). Some comments on six inequalities associated with the inefficiency of ordinary least squares with one regressor. *Linear Algebra and its Applications*, 264, 13–54. [235, 418]

Wedderburn, J. H. M. (1934). *Lectures on Matrices, Colloquium Publications, Vol. 17.* American Mathematical Society. Reprinted by Dover, 1964. [304]

Weisberg, S. (2005). *Applied Linear Regression*, Third Ed. Wiley. [ix, 164, 187, 199]

Wells, M. T. (2009). A conversation with Shayle R. Searle. *Statistical Science*, 24, 244–254. [viii]

Werner, H. J. (1997). Solution 18-3.2 [to Problem 18-3: A further matrix version of the Cauchy–Schwarz inequality]. *Image: The Bulletin of the ILAS*, 19, 26–27. [216]

Werner, H. J. & Yapar, C. (1995). More on partitioned possibly restricted linear regression. In *Multivariate Statistics and Matrices in Statistics: Proceedings of the 5th Tartu Conference, Tartu–Pühajärve, Estonia, 23–28 May 1994* (E.-M. Tiit, T. Kollo & H. Niemi, eds.), TEV & VSP, pp. 57–66. [331]

Werner, H. J. & Yapar, C. (1996). A BLUE decomposition in the general linear regression model. *Linear Algebra and its Applications*, 237/238, 395–404. [337]

Wilson, R. J. (2001). *Stamping Through Mathematics*. Springer. [vi]

Wolkowicz, H. & Styan, G. P. H. (1979). Extensions of Samuelson's inequality. *The American Statistician*, 33, 143–144. [420]

Wong, Y. K. (1935). An application of orthogonalization process to the theory of least squares. *The Annals of Mathematical Statistics*, 6, 53–75. [viii]

Woodbury, M. A. (1950). *Inverting Modified Matrices*. Princeton University Press. [301]

Yanai, H. (1990). Some generalized forms of least squares g-inverse, minimum norm g-inverse, and Moore–Penrose inverse matrices. *Computational Statistics & Data Analysis*, 10, 251–260. [149]

Yanai, H. (2003). Vectors and matrices in psychometrics with special emphasis on generalized inverse and projection matrices. In *New Developments in Psychometrics: Proceedings of the International Meeting of the Psychometric Society IMPS2001* (H. Yanai, A. Okada, K. Shigemasu, Y. Kano & J. J. Meulman, eds.), Springer, pp. 19–32. [13, 307]

Yanai, H. & Puntanen, S. (1993). Partial canonical correlations associated with the inverse and some generalized inverses of a partitioned dispersion matrix. In *Statistical Sciences and Data Analysis: Proceedings of the Third Pacific Area Statistical Conference* (K. Matusita, M. L. Puri & T. Hayakawa, eds.), VSP, pp. 345–366. [386]

Yanai, H. & Takane, Y. (2007). Matrix methods and their applications to factor analysis. In *Handbook of Latent Variable and Related Models* (S.-Y. Lee, ed.), Elsevier, pp. 345–366. [307, 314, 401]

Yang, H. & Wang, L. (2009). An alternative form of the Watson efficiency. *Journal of Statistical Planning and Inference*, 139, 2767–2774. [241]

Young, D. M., Odell, P. L. & Hahn, W. (2000). Nonnegative-definite covariance structures for which the BLU, WLS, and LS estimators are equal. *Statistics & Probability Letters*, 49, 271–276. [340]

Young, D. M., Scariano, S. M. & Hallum, C. R. (2005). Estimation-equivalent covariance structures for the least squares and minque estimators of the linear model variance. *Communications in Statistics: Theory and Methods*, 34, 625–629. [340]

Young, M. M. & Young, D. M. (2004). An alternative proof of the relationship between F and G when $FF^T = GG^T$. *The Mathematical Scientist*, 29, 33–35. [397]

Zehna, P. W. (1991). On proving that $\bar{X}$ and s are independent. *The American Statistician*, 45, 121–122. [184]

Zhang, B. X., Liu, B. S. & Lu, C.-Y. (2004). A study of the equivalence of the BLUEs between a partitioned singular linear model and its reduced singular linear models. *Acta Mathematica Sinica, English Series*, 20, 557–568. [331]

Zhang, F. (1999). *Matrix Theory: Basic Results and Techniques*. Springer. [vii, 316]

Zhang, F. (2005a). Block matrix techniques. In *The Schur Complement and Its Applications* (F. Zhang, ed.), Springer, Chap. 3, pp. 83–110. [386]

Zhang, F., ed. (2005b). *The Schur Complement and Its Applications*. Springer. [vii, 293]

Zhang, F. (2009). *Linear Algebra: Challenging Problems for Students*, Second Ed. Johns Hopkins University Press. [vii, 253]

Zmyślony, R. (1980). A characterization of best linear unbiased estimators in the general linear model. In *Mathematical Statistics and Probability Theory: Proceedings of the Sixth International Conference (Wisla, Poland, 1978)* (W. Klonecki, A. Kozek & J. Rosiński, eds.), Springer, pp. 365–373. [221]

Zyskind, G. (1967). On canonical forms, non-negative covariance matrices and best and simple least squares linear estimators in linear models. *The Annals of Mathematical Statistics*, 38, 1092–1109. [215, 219]

Zyskind, G. (1969). Parametric augmentations and error structures under which certain simple least squares and analysis of variance procedures are also best. *Journal of the American Statistical Association*, 64, 1353–1368. [215, 224, 227, 234]

Zyskind, G. (1975). Error structures, projections and conditional inverses in linear model theory. In *A Survey of Statistical Design and Linear Model* (J. N. Srivastava, ed.), North-Holland, pp. 647–663. [215]

Zyskind, G. & Martin, F. B. (1969). On best linear estimation and general Gauss–Markov theorem in linear models with arbitrary nonnegative covariance structure. *SIAM Journal on Applied Mathematics*, 17, 1190–1202. [123, 149, 215, 324]

Author Index

Subject Index