Kohei Adachi

Matrix-Based Introduction to Multivariate Data Analysis

Second Edition

Springer

Matrix-Based Introduction to Multivariate Data Analysis

Kohei Adachi

Matrix-Based Introduction to Multivariate Data Analysis

Second Edition

Springer

Kohei Adachi
Graduate School of Human Sciences
Osaka University
Suita, Osaka, Japan

ISBN 978-981-15-4102-5 ISBN 978-981-15-4103-2 (eBook)
https://doi.org/10.1007/978-981-15-4103-2

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface to the Second Edition

In this second edition, I have added new six chapters (Chaps. 17–22) and three Appendices (A.7–A.9) to the first edition (which spanned Chaps. 1–16 and Appendices A.1–A.6), together with correcting all known misprints and other errors in the first edition. Furthermore, I have made minor modifications to some parts of the first edition, in line with the additional chapters and appendices.

The chapters added in this second edition are as follows:

17. Advanced Matrix Operations
18. Exploratory Factor Analysis (Part 2)
19. Principal Component Analysis versus Factor Analysis
20. Three-way Principal Component Analysis
21. Sparse Regression Analysis
22. Sparse Factor Analysis

which form Part V (*Advance Procedures*) following Parts I–IV.

Chapter 17 serves as a mathematical preparation for the following chapters. In Chap. 17, the Moore–Penrose (MP) inverse in particular is covered in detail, emphasizing its definition through singular value decomposition (SVD). I believe that the MP inverse is of secondary importance among matrix operations, with SVD being of primary importance, as the SVD-based definition of the MP inverse allows us to easily derive its properties and various matrix operations. In this chapter, we also introduce an orthogonal complement matrix, as it is foreseeable that the need for this matrix will increase in multivariate analysis procedures.

Chapter 18 is titled "Exploratory Factor Analysis (Part 2)", while "(Part 1)" was added to the title of Chap. 12 in the first edition. The contents of Chap. 12 remain unchanged in this second edition, but the exploratory factor analysis (EFA) in Chap. 18 is of a new type, i.e., the EFA procedure formulated as a matrix decomposition problem. This differs from EFA based on the latent variable model in Chap. 12. To emphasize the difference, the former (new) EFA is referred to as matrix decomposition FA (MDFA), while the latter is called latent variable FA

(LVFA) in Chap. 18. Its addition owes to recent developments after the publication of the first edition, as studies of MDFA advanced rapidly. I believe that MDFA is generally superior to LVFA in that the former makes the essence of FA more transparent.

In Chap. 19, answers are given to the question of how solutions from principal component analysis (PCA) and FA differ. No clear answer to this question is found in other books, to the best of my knowledge. The answers in Chap. 19 also are owing to advances in MDFA studies, with the MDFA formulation allowing for straightforward comparisons to be made between FA and PCA.

Three-way principal component analysis (3WPCA) is treated in Chap. 20. 3WPCA refers to a specially modified PCA designed for three-way data sets. The given example is a data array of inputs $\times$ outputs $\times$ boxes, whose elements are the magnitudes of output signals elicited by input signals for multiple black boxes. Three-way data are often encountered in various areas of sciences, and as such 3WPCA is a useful dimension reduction methodology. Its algorithms are very matrix-intensive and suitably treated in this book.

Sparse estimation procedures are introduced in Chaps. 21 and 22. Here, sparse estimation refers to estimating a number of parameters as zeros. Such procedures are popular topics in the field of machine learning. This field can be defined as learning attained by machines (in particular computers) as opposed to humans or living organisms. Statistical analysis procedures are useful methodologies for machine learning. Sparse estimation is also I believe a key property of human learning: our perception performs sparse estimation too in that usually we only cognize useful signals, neglecting useless ones as "zeros". In this respect, it is very important to enable machines to perform sparse estimation, as a complement to humans. In Chap. 21, sparse regression analysis procedures are described, including Tibshirani's (1996) procedure called *lasso* which spurred the developments in sparse estimation. Finally, sparse factor analysis (FA) procedures are introduced in Chap. 22.

The Appendices added in this second edition are as follows:

A.7. Scale Invariance of Covariance Structure Analysis
A.8. Probability Densities and Expected Values with EM Algorithm
A.9. EM Algorithm for Factor Analysis.

Though the scale invariance in A.7 had been described with short notes in the first edition, the notes were found too short and insufficient. Thus, the scale invariance is described in more detail in Appendix A.7: Notes 9.3 and 10.2 in the first edition have been expanded and moved to A.7 in this edition. The new Appendix A.9 is necessary for explaining one of the two sparse FA procedures in Chap. 22 and is also useful for deepening the understanding of the confirmatory and exploratory FA treated in Chaps. 10 and 12. The foundations of the algorithm in A.9 are introduced in the preceding new Appendix A.8. Further, this A.8 serves to deepen the understanding of the treatment in Chap. 8.

In the first edition, some parts of the bibliographical notes and exercises were provided to allow readers to extend their understanding beyond the scope covered in that edition. Such parts have become unnecessary in the second edition, as the advanced contents are now described in the additional chapters. Hence, sections of the bibliographical notes and exercises related to the new chapters (Chaps. 17–22) have been deleted or moved to the relevant chapters in the second edition.

Yutaka Hirachi of Springer has encouraged me for publishing this revised version, as well as when I prepared the drafts for the first edition. I am most grateful to him. I am also thankful to the reviewers who read through drafts of this book.

Kyoto, Japan Kohei Adachi
February 2020

Preface to the First Edition

A set of multivariate data can be expressed as a table, i.e., a matrix, of individuals (rows) by variables (columns), with the variables interrelated. Statistical procedures for analyzing such data sets are generally referred to as multivariate data analysis. The demand for this kind of analysis is increasing in a variety of fields. Each procedure in multivariate data analysis features a special purpose. For example, predicting future performance, classifying individuals, visualizing inter-individual relationships, finding a few factors underlying a number of variables, and examining causal relationships among variables are included in the purposes for the procedures.

The aim of this book is to enable readers who may not be familiar with matrix operations to understand major multivariate data analysis procedures in matrix forms. For that aim, this book begins with explaining fundamental matrix calculations and the matrix expressions of elementary statistics, followed by an introduction to popular multivariate procedures, with chapter-by-chapter advances in the levels of matrix algebra. The organization of this book allows readers without knowledge of matrices to deepen their understanding of multivariate data analysis.

Another feature of this book is its emphasis on the model that underlies each procedure and the objective function that is optimized for fitting the model to data. The author believes that the matrix-based learning of such models and objective functions is the shortest way to comprehend multivariate data analysis. This book is also arranged so that readers can intuitively capture for what purposes multivariate analysis procedures are utilized; plain explanations of the purposes with numerical examples precede mathematical descriptions in almost all chapters.

The preceding paragraph featured three key words: purpose, model, and objective function. The author considers that capturing those three points for each procedure suffices to understand it. This consideration implies that the mechanisms behind how objective functions are optimized must not necessarily be understood. Thus, the mechanisms are only described in appendices and some exercises.

This book is written with the following guidelines in mind:

(1) Not using mathematics except matrix algebra
(2) Emphasizing singular value decomposition (SVD)
(3) Preferring a simultaneous solution to a successive one.

Although the exceptions to (1) are found in Appendix A.6, where differential calculus is used, and in some sections of Part III and Chap. 15, where probabilities are used, those exceptional parts only occupy a limited number of pages; the majority of the book is matrix-intensive. Matrix algebra is also exclusively used for formulating the optimization of objective functions in Appendix A.4. For matrix-intensive formulations, ten Berge's (1983, 1993) theorem is considered to be the best starting fact, as found in Appendix A.4.1.

Guideline (2) is due to the fact that SVD can be defined for any matrix, and a number of important properties of matrices are easily derived from SVD. In the former point, SVD is more general than eigenvalue decomposition (EVD), which is only defined for symmetric matrices. Thus, EVD is only mentioned in Sect. 6.2. Further, SVD takes on an important role in optimizing trace and least squares functions of matrices: The optimization problems are formulated with the combination of SVD and ten Berge's (1983, 1993) theorem, as found in Appendix A.4.2 and Appendix A.4.3.

Guideline (3) is particularly concerned with principal component analysis (PCA), which can be formulated as minimizing $\|\mathbf{X} - \mathbf{F}\mathbf{A}'\|^2$ over PC score matrix $\mathbf{F}$ and loading matrix $\mathbf{A}$ for a data matrix $\mathbf{X}$. In some of the literature, PCA is described as obtaining the first component, the second, and the remaining components in turn (i.e., per column of $\mathbf{F}$ and $\mathbf{A}$). This can be called a successive solution. On the other hand, PCA can be described as obtaining $\mathbf{F}$ and $\mathbf{A}$ matrix-wise, which can be called a simultaneous solution. This is preferred in this book, as the above formulation is actually made matrix-wise, and the simultaneous solution facilitates understanding PCA as a reduced rank approximation of $\mathbf{X}$.

This book is appropriate for undergraduate students who have already learned introductory statistics, as the author has used preliminary versions of the book in a course for such students. It is also useful for graduate students and researchers who are not familiar with the matrix-intensive formulations of multivariate data analysis.

I owe this book to the people who can be called the "matricians" in statistics, more exactly, the ones taking matrix-intensive approaches for formulating and developing data analysis procedures. Particularly, I have been influenced by the Dutch psychometricians, as found above, in that I emphasize the theorem by Jos M. F. ten Berge (Professor Emeritus, University of Groningen). Yutaka Hirachi of Springer has been encouraging me since I first considered writing this book. I am

most grateful to him. I am also thankful to the reviewers who read through drafts of this book. Finally, I must show my gratitude to Yoshitaka Shishikura of the publisher Nakanishiya Shuppan, as he readily agreed to the use of the numerical examples in this book, which I had originally used in that publisher's book.

Kyoto, Japan Kohei Adachi
May 2016

Contents

Part I
Elementary Statistics with Matrices

This part begins with introducing elementary matrix operations, followed by explanations of fundamental statistics with their matrix expressions. These initial chapters serve as preparation for learning the multivariate data analysis procedures that are described in Part II and thereafter.

Chapter 1
Elementary Matrix Operations

The mathematics for studying the properties of matrices is called *matrix algebra* or *linear algebra*. This first chapter treats the introductory part of matrix algebra required for learning multivariate data analysis. We begin by explaining what a matrix is, in order to describe elementary matrix operations.

In later chapters, more advanced properties of matrices are described, where necessary, with references to Appendices for more detailed explanations.

1.1 Matrices

Let us note that Table 1.1 is a 6 teams × 4 items table. When such a table (i.e., a two-way array) is treated as a unit entity and expressed as

$$\mathbf{X} = \begin{bmatrix} 0.617 & 731 & 140 & 3.24 \\ 0.545 & 680 & 139 & 4.13 \\ 0.496 & 621 & 143 & 3.68 \\ 0.493 & 591 & 128 & 4.00 \\ 0.437 & 617 & 186 & 4.80 \\ 0.408 & 615 & 184 & 4.80 \end{bmatrix},$$

this is called a 6 (rows) × 4 (columns) *matrix*, or a matrix of 6 rows by 4 columns. "*Matrices*" is the plural of "matrix". Here, a horizontal array and a vertical one are called a *row* and a *column*, respectively. For example, the fifth row of **A** is "0.437, 617, 0.260, 4.80", while the third column is "140, 139, 143, 128, 186, 184". Further, the cell at which the fifth row and third column intersect is occupied by 186, which is called "the (5,3) *element*". Rewriting the rows of a matrix as columns (or its columns as rows) is referred to as a *transpose*. The transpose of **X** is denoted as $\mathbf{X}'$:

K. Adachi, *Matrix-Based Introduction to Multivariate Data Analysis*,
https://doi.org/10.1007/978-981-15-4103-2_1

Table 1.1 Averages of the six-teams in Japanese Central Baseball League 2005

Team	Item			
	Win %	Runs	HR	ERA
Tigers	0.617	731	140	3.24
Dragons	0.545	680	139	4.13
BayStars	0.496	621	143	3.68
Swallows	0.493	591	128	4.00
Giants	0.437	617	186	4.80
Carp	0.408	615	184	4.80

$$\mathbf{X}' = \begin{bmatrix} 0.617 & 0.545 & 0.496 & 0.493 & 0.437 & 0.408 \\ 731 & 680 & 621 & 591 & 617 & 615 \\ 140 & 139 & 143 & 128 & 186 & 184 \\ 3.24 & 4.13 & 3.68 & 4.00 & 4.80 & 4.80 \end{bmatrix}$$

Let us describe a matrix in a generalized setting. The array of a_{ij} ($i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, m$) arranged in n rows and m columns, i.e.,

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix}, \tag{1.1}$$

is called an $n \times m$ *matrix* with a_{ij} its (i, j) *element*. The transpose of $\mathbf{A}$ is an $m \times n$ matrix

$$\mathbf{A}' = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{12} & a_{22} & \cdots & a_{n2} \\ \vdots & \vdots & \vdots & \vdots \\ a_{1m} & a_{2m} & \cdots & a_{nm} \end{bmatrix}. \tag{1.2}$$

The transpose of a transposed matrix is obviously the original matrix, with $(\mathbf{A}')' = \mathbf{A}$.

The expression of matrix $\mathbf{A}$ as the right-hand side in (1.1) takes a large amount of space. For economy of space, the matrix $\mathbf{A}$ in (1.1) is also expressed as

$$\mathbf{A} = (a_{ij}), \tag{1.3}$$

using the general expression a_{ij} for the elements of $\mathbf{A}$. The statement "We define an $n \times m$ matrix as $\mathbf{A} = (a_{ij})$" stands for the matrix $\mathbf{A}$ being expressed as (1.1).

1.2 Vectors

A vertical array,

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}, \tag{1.4}$$

is called a *column vector* or simply a *vector*. In exactness, (1.4) is said to be an $n \times 1$ *vector*, since it contains n elements. Vectors can be viewed as a special case of matrices; (1.4) can also be called an $n \times 1$ matrix. Further, a *scalar* is a 1×1 matrix. The right side of (1.4) is vertically long, and for the sake of the economy of space, (1.4) is often expressed as

$$\mathbf{a} = [a_1, a_2, \ldots, a_n]' \text{ or } \mathbf{a}' = [a_1, a_2, \ldots, a_n], \tag{1.5}$$

using a transpose. A horizontal array as $\mathbf{a}'$ is called a *row vector*.

We can use vectors to express a matrix: by using $n \times 1$ vectors $\mathbf{a}_j = [a_{1j}, a_{2j}, \ldots, a_{nj}]'$, $j = 1, 2, \ldots, m$, and $m \times 1$ vectors $\tilde{\mathbf{a}}_i = [a_{i1}, a_{i2}, \ldots, a_{im}]'$, i = 1, 2, …, n, and the matrix (1.1) or (1.3) is expressed as

$$\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m] = \begin{bmatrix} \tilde{\mathbf{a}}_1' \\ \tilde{\mathbf{a}}_2' \\ \vdots \\ \tilde{\mathbf{a}}_n' \end{bmatrix} = [\tilde{\mathbf{a}}_1, \tilde{\mathbf{a}}_2, \ldots, \tilde{\mathbf{a}}_n]' = (a_{ij}). \tag{1.6}$$

In this book, a **bold uppercase** letter such as $\mathbf{X}$ is used for denoting a *matrix*, a **bold lowercase** letter such as $\mathbf{x}$ is used for a *vector*, and an *italic* letter (not bold) such as x is used for a *scalar*. Though a *series of integers* has so far been expressed as i = 1, 2, …, n, this may be rewritten as i = 1, …, n, omitting 2 when it obviously follows 1. With this notation, (1.1) or (1.6) is rewritten as

$$\mathbf{A} = \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \vdots & \vdots \\ a_{n1} & \cdots & a_{nm} \end{bmatrix} = [\mathbf{a}_1, \ldots, \mathbf{a}_m] = [\tilde{\mathbf{a}}_1, \ldots \tilde{\mathbf{a}}_n]'.$$

1.3 Sum of Matrices and Their Multiplication by Scalars

The sum of matrices can be defined when they are of the *same size*. Let matrices **A** and **B** be equivalently $n \times m$. Their *sum* **A** + **B** yields the $n \times m$ matrix, each of whose *elements is the sum of the corresponding ones* of $\mathbf{A} = (a_{ij})$ and $\mathbf{B} = (b_{ij})$: The sum is defined as

$$\mathbf{A} + \mathbf{B} = (a_{ij} + b_{ij}), \tag{1.7}$$

using the notation in (1.3). For example, when $\mathbf{X} = \begin{bmatrix} 3 & -2 & 6 \\ 8 & 0 & -2 \end{bmatrix}$ and $\mathbf{Y} = \begin{bmatrix} 2 & 1 & -9 \\ -7 & 2 & -3 \end{bmatrix}$,

$$\mathbf{X} + \mathbf{Y} = \begin{bmatrix} 3+2 & -2+1 & 6-9 \\ 8-7 & 0+2 & -2-3 \end{bmatrix} = \begin{bmatrix} 5 & -1 & -3 \\ 1 & 2 & -5 \end{bmatrix}.$$

The multiplication of matrix $\mathbf{A} = (a_{ij})$ by scalar s is defined as *all elements* of **A** being *multiplied* by s:

$$s\mathbf{A} = \left(s \times a_{ij}\right), \tag{1.8}$$

using the notation in (1.3). For example, when $\mathbf{Z} = \begin{bmatrix} 8 & -2 & 6 \\ -5 & 0 & -3 \end{bmatrix}$

$$\begin{aligned} -0.1\mathbf{Z} &= \begin{bmatrix} -0.1 \times 8 & -0.1 \times (-2) & -0.1 \times 6 \\ -0.1 \times (-5) & -0.1 \times 0 & -0.1 \times (-3) \end{bmatrix} \\ &= \begin{bmatrix} -0.8 & 0.2 & -0.6 \\ 0.5 & 0 & 0.3 \end{bmatrix}. \end{aligned}$$

The *sum of the matrices multiplied by scalars* is defined simply as the combination of (1.7) and (1.8):

$$v\mathbf{A} + w\mathbf{B} = \left(va_{ij} + wb_{ij}\right). \tag{1.9}$$

For example, when $\mathbf{X} = \begin{bmatrix} 4 & -2 & 6 \\ 8 & 0 & -2 \end{bmatrix}$ and $\mathbf{Y} = \begin{bmatrix} 2 & 1 & -9 \\ -7 & 2 & -3 \end{bmatrix}$,

$$0.5\mathbf{X} + (-2)\mathbf{Y} = \begin{bmatrix} 2-4 & -1-2 & 3+18 \\ 4+14 & 0-4 & -1+6 \end{bmatrix} = \begin{bmatrix} -2 & -3 & 21 \\ 18 & -4 & 5 \end{bmatrix}.$$

Obviously, setting $v = 1$ and $w = -1$ in (1.9) leads to the definition of the matrix difference $\mathbf{A} - \mathbf{B}$.

The above definition is generalized as

$$\sum_{k=1}^{K} v_k \mathbf{A}_k = v_1 \mathbf{A}_1 + \cdots + v_K \mathbf{A}_K = \left(\sum_{k=1}^{K} v_k a_{ijk} \right), \tag{1.10}$$

where $\mathbf{A}_1, \ldots, \mathbf{A}_K$ are of the same size and a_{ijk} is the (i, j) element of $\mathbf{A}_k$ $(k = 1, \ldots, K)$.

1.4 Inner Product and Norms of Vectors

The *inner product* of the vectors $\mathbf{a} = [a_1, \ldots, a_m]'$ and $\mathbf{b} = [b_1, \ldots, b_m]'$ is defined as

$$\mathbf{a}'\mathbf{b} = \mathbf{b}'\mathbf{a} = [a_1, \ldots, a_m] \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix} = a_1 b_1 + \cdots + a_m b_m = \sum_{k=1}^{m} a_k b_k. \tag{1.11}$$

Obviously, this can be defined only for the vectors of the same size. The inner product is expressed as $\mathbf{a}'\mathbf{b}$ or $\mathbf{b}'\mathbf{a}$, i.e., the form of a *transposed column vector* (i.e., *row vector*) followed by a *column vector*, so as to be congruous to the matrix product introduced in the next section.

The inner product of the identical vectors $\mathbf{a}$ and $\mathbf{a}$ is in particular called the *squared norm* of $\mathbf{a}$ and denoted as $\|\mathbf{a}\|^2$:

$$\|\mathbf{a}\|^2 = \mathbf{a}'\mathbf{a} = [a_1, \ldots, a_m]' \begin{bmatrix} a_1 \\ \vdots \\ a_m \end{bmatrix} = a_1^2 + \cdots + a_m^2 = \sum_{k=1}^{m} a_k^2. \tag{1.12}$$

The *square root* of $\|\mathbf{a}\|^2$, that is, $\|\mathbf{a}\|$ is simply called the *norm* of the vector $\mathbf{a} = [a_1, \ldots, a_m]'$ with

$$\|\mathbf{a}\| = \sqrt{a_1^2 + \cdots + a_m^2}. \tag{1.13}$$

It is also called the *length* of $\mathbf{a}$, for the following reason. If $m = 3$ with $\mathbf{a} = [a_1, a_2, a_3]'$ and $\mathbf{a}$ is viewed as the line extending from the origin to the point whose coordinate is $[a_1, a_2, a_3]'$, as illustrated in Fig. 1.1: (1.13) expresses the length of the line. It also holds for $m = 1, 2$. If $m > 3$, the line cannot be depicted or seen by those of us (i.e., the human beings living in three-dimensional world), but the length of $\mathbf{a}$ is also defined as (1.13) for $m > 3$ in mathematics (in which the entities that do not exist in the real world are also considered if they are treated logically).

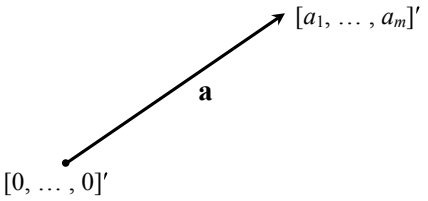


Fig. 1.1 Graphical representation of a vector

1.5 Product of Matrices

Let $n \times m$ and $m \times p$ matrices be defined as

$$\mathbf{A}\begin{bmatrix}\mathbf{a}_1' \\ \vdots \\ \mathbf{a}_n'\end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \vdots & \vdots \\ a_{n1} & \cdots & a_{nm}\end{bmatrix} \text{ and } \mathbf{B} = [\mathbf{b}_1 \cdots \mathbf{b}_p] = \begin{bmatrix} b_{11} & \cdots & b_{1p} \\ \vdots & \vdots & \vdots \\ b_{m1} & \cdots & b_{mp}\end{bmatrix},$$

respectively, with $\mathbf{a}_i' = [a_{i1}, \ldots, a_{im}] (i = 1, \ldots, n)$ and $\mathbf{b}_j = \begin{bmatrix} b_{1j} \\ \vdots \\ b_{mj}\end{bmatrix} (j = 1, \ldots, p)$.

Then, the *post-multiplication* of $\mathbf{A}$ by $\mathbf{B}$ is defined as

$$\mathbf{AB} = \begin{bmatrix} \mathbf{a}_1'\mathbf{b}_1 & \cdots & \mathbf{a}_1'\mathbf{b}_p \\ \vdots & \ldots & \vdots \\ \mathbf{a}_n'\mathbf{b}_1 & \cdots & \mathbf{a}_n'\mathbf{b}_p\end{bmatrix} = \left(\mathbf{a}_i'\mathbf{b}_j\right), \tag{1.14}$$

using the *inner products* of the *row vectors of the preceding matrix* $\mathbf{A}$ and *the column vectors of the following matrix* $\mathbf{B}$. The resulting matrix $\mathbf{AB}$ is the $n \times p$ matrix whose (i, j) element is the inner product of the ith row of $\mathbf{A}$ and the jth column of $\mathbf{B}$:

$$\mathbf{a}_i'\mathbf{b}_j = [a_{i1}, \ldots, a_{im}]\begin{bmatrix} b_{1j} \\ \vdots \\ b_{mj}\end{bmatrix} = a_{i1}b_{1j} + \cdots + a_{im}b_{mi} = \sum_{k=1}^{m} a_{ik}b_{kj}. \tag{1.15}$$

For example, if $\mathbf{A} = \begin{bmatrix}\mathbf{a}_1' \\ \mathbf{a}_2'\end{bmatrix} = \begin{bmatrix} 2 & -4 \\ 1 & 7\end{bmatrix}, \mathbf{B} = [\mathbf{b}_1\ \mathbf{b}_2] = \begin{bmatrix} -3 & 1 \\ 2 & -5\end{bmatrix}$, then

$$\mathbf{AB} = \begin{bmatrix} \mathbf{a}_1'\mathbf{b}_1 & \mathbf{a}_1'\mathbf{b}_2 \\ \mathbf{a}_2'\mathbf{b}_1 & \mathbf{a}_2'\mathbf{b}_2\end{bmatrix} = \begin{bmatrix} 2 \times (-3) + (-4) \times 2 & 2 \times 1 + (-4) \times (-5) \\ 1 \times (-3) + 7 \times 2 & 1 \times 1 + 7 \times (-5)\end{bmatrix}$$
$$= \begin{bmatrix} -14 & 22 \\ 11 & -34\end{bmatrix}.$$

As found above, the matrix product **AB** is defined only when the following holds:

$$\text{the number of columns in } \mathbf{A} = \text{the number of rows in } \mathbf{B}. \tag{1.16}$$

The resulting matrix **AB** is

$$(\text{the number of rows in } \mathbf{A}) \times (\text{the number of columns in } \mathbf{B}). \tag{1.17}$$

Thus, the product is sometimes expressed as

$$\underset{n\times m}{\mathbf{A}}\ \underset{m\times p}{\mathbf{B}} = \underset{n\times p}{\mathbf{C}}, \text{ or, more simply, } {}_n\mathbf{A}_m\mathbf{B}_p = {}_n\mathbf{C}_p, \tag{1.18}$$

with which we can easily verify (1.16) and (1.17). If $n = p$, we can define products **AB** and **BA**. Here, we should note

$$\mathbf{AB} \neq \mathbf{BA}, \tag{1.19}$$

except for special **A** and **B**, which is different from the product of scalars with $st = ts$, the inner product (1.11), and that of scalar s and matrix **A** with

$$s\mathbf{A} = \mathbf{A} \times s. \tag{1.20}$$

For this reason, we call **AB** "the *post-multiplication* of **A** by **B**" or "the *pre-multiplication* of **B** by **A**", so as to clarify the order of the matrices.

Here, four examples of matrix products are presented as follows:

Ex. 1. For $\mathbf{X} = \begin{bmatrix} 2 & 3 & -1 \\ -2 & 0 & 4 \end{bmatrix}$ and $\mathbf{Y} = \begin{bmatrix} 3 & 5 & 4 \\ -1 & 0 & -2 \\ 0 & 6 & 0 \end{bmatrix}$,

$$\mathbf{XY} = \begin{bmatrix} 2\times 3+3\times(-1)+(-1)\times 0 & 2\times 5+3\times 0+(-1)\times 6 & 2\times 4+3\times(-2)+(-1)\times 0 \\ -2\times 3+0\times(-1)+4\times 0 & -2\times 5+0\times 0+4\times 6 & -2\times 4+0\times(-2)+4\times 0 \end{bmatrix} = \begin{bmatrix} 3 & 4 & 2 \\ -6 & 14 & -8 \end{bmatrix}.$$

Ex. 2. For $\mathbf{F} = \begin{bmatrix} 2 & -1 \\ -3 & 0 \\ 1 & 3 \\ -2 & -3 \end{bmatrix}$ and $\mathbf{A} = \begin{bmatrix} -4 & 1 \\ 6 & -3 \\ 2 & 5 \end{bmatrix}$,

$$\mathbf{FA}' = \begin{bmatrix} 2 & -1 \\ -3 & 0 \\ 1 & 3 \\ -2 & -3 \end{bmatrix} \begin{bmatrix} -4 & 6 & 2 \\ 1 & -3 & 5 \end{bmatrix}$$

$$= \begin{bmatrix} 2\times(-4)+(-1)\times 1 & 2\times 6+(-1)\times(-3) & 2\times 2+(-1)\times 5 \\ -3\times(-4)+0\times 1 & -3\times 6+0\times(-3) & -3\times 2+0\times 5 \\ 1\times(-4)+3\times 1 & 1\times 6+3\times(-3) & 1\times 2+3\times 5 \\ -2\times(-4)+(-3)\times 1 & -2\times 6+(-3)\times(-3) & -2\times 2+(-3)\times 5 \end{bmatrix}$$

$$= \begin{bmatrix} -9 & 15 & -1 \\ 12 & -18 & -6 \\ -1 & -3 & 17 \\ 5 & -3 & -19 \end{bmatrix},$$

where it should be noted that $\mathbf{A}$ has been transposed in the product.

Ex. 3. In statistics, the product of a matrix and its transpose is often used.

For $\mathbf{A} = \begin{bmatrix} -4 & 1 \\ 6 & -3 \\ 2 & 5 \end{bmatrix}$, the post-multiplication of $\mathbf{A}$ by $\mathbf{A}'$, which we denote by $\mathbf{S}$, is

$$\mathbf{S} = \mathbf{AA}' = \begin{bmatrix} (-4)^2+1^2 & -4\times 6+1\times(-3) & -4\times 2+1\times 5 \\ 6\times(-4)+(-3)\times 1 & 6^2+(-3)^2 & 6\times 2+(-3)\times 5 \\ 2\times(-4)+5\times 1 & 2\times 6+5\times(-3) & 2^2+5^2 \end{bmatrix}$$

$$= \begin{bmatrix} 17 & -27 & -3 \\ -27 & 45 & -3 \\ -3 & -3 & 29 \end{bmatrix}.$$

The pre-multiplication of $\mathbf{A}$ by $\mathbf{A}'$, which we denote by $\mathbf{T}$, is

$$\mathbf{T} = \mathbf{A}'\mathbf{A} = \begin{bmatrix} (-4)^2+6^2+2^2 & (-4)\times 1+6\times(-3)+2\times 5 \\ 1\times(-4)+(-3)\times 6+5\times 2 & 1^2+(-3)^2+5^2 \end{bmatrix}$$

$$= \begin{bmatrix} 56 & -12 \\ -12 & 35 \end{bmatrix}.$$

Ex. 4. The product of vectors is a special case of that of matrices:

For $\mathbf{u} = \begin{bmatrix} 2 \\ -1 \\ 3 \end{bmatrix}$ and $\mathbf{v} = \begin{bmatrix} -2 \\ 3 \\ -4 \end{bmatrix}$,

the inner product yields a scalar as

$$\mathbf{u}'\mathbf{v} = 2 \times (-2) + (-1) \times 3 + 3 \times (-4) = -19,$$

but the post-multiplication of 3×1 vector $\mathbf{u}$ by 1×3 $\mathbf{v}'$ gives a 3×3 matrix with

$$\mathbf{u}\mathbf{v}' = \begin{bmatrix} 2 \\ -1 \\ 3 \end{bmatrix} [-2 \quad 3 \quad -4] = \begin{bmatrix} 2 \times (-2) & 2 \times 3 & 2 \times (-4) \\ (-1) \times (-2) & (-1) \times 3 & (-1) \times (-4) \\ 3 \times (-2) & 3 \times 3 & 3 \times (-4) \end{bmatrix}$$
$$= \begin{bmatrix} -4 & 6 & -8 \\ 2 & -3 & 4 \\ -6 & 9 & -12 \end{bmatrix}.$$

1.6 Two Properties of Matrix Products

The *transposed product* of matrices satisfies

$$(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'; \ (\mathbf{ABC})' = \mathbf{C}'\mathbf{B}'\mathbf{A}' \tag{1.21}$$

Let $\mathbf{A}$ and $\mathbf{B}$ be matrices of size $n \times m$; let $\mathbf{C}$ and $\mathbf{D}$ be those of $m \times l$. Then, the *product of their sums multiplied by scalars* s, t, u, and v satisfies

$$(s\mathbf{A} + t\mathbf{B})(u\mathbf{C} + v\mathbf{D}) = su\mathbf{AC} + sv\mathbf{AD} + tu\mathbf{BC} + tv\mathbf{BD}. \tag{1.22}$$

1.7 Trace Operator and Matrix Norm

A matrix with the number of rows equivalent to that of columns is said to be *square*. For a square matrix $\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1n} \\ s_{21} & s_{22} & \cdots & s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & \cdots & s_{nn} \end{bmatrix}$, the elements on the diagonal, i.e., $s_{11}, \ldots, s_{nn}$, are called the *diagonal elements* of $\mathbf{S}$. Their sum is called a *trace* and is denoted as

$$\mathrm{tr}\mathbf{S} = s_{11} + s_{22} + \cdots + s_{nn}. \tag{1.23}$$

Obviously,

$$\mathrm{tr}\mathbf{S}' = \mathrm{tr}\mathbf{S} \tag{1.24}$$

The trace fulfills important roles when it is defined for the product of matrices. Let us consider $\mathbf{A} = [\mathbf{a}_1 \cdots \mathbf{a}_m] = \begin{bmatrix} \tilde{\mathbf{a}}_1' \\ \vdots \\ \tilde{\mathbf{a}}_n' \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \vdots & \vdots \\ a_{n1} & \cdots & a_{nm} \end{bmatrix}$ and $\mathbf{B} = [\mathbf{b}_1 \cdots \mathbf{b}_n] = \begin{bmatrix} \tilde{\mathbf{b}}_1' \\ \vdots \\ \tilde{\mathbf{b}}_m \end{bmatrix} = \begin{bmatrix} b_{11} & \cdots & b_{1n} \\ \vdots & \vdots & \vdots \\ b_{m1} & \cdots & b_{mn} \end{bmatrix}$. Then, **AB** and **BA** are $n \times n$ and $m \times m$ square matrices, respectively, for which traces can be defined, with

$$\mathbf{AB} = \begin{bmatrix} \tilde{\mathbf{a}}_1'\mathbf{b}_1 & & \# \\ & \ddots & \\ \# & & \tilde{\mathbf{a}}_n'\mathbf{b}_n \end{bmatrix} \text{ and } \mathbf{BA} = \begin{bmatrix} \tilde{\mathbf{b}}_1'\mathbf{a}_1 & & \# \\ & \ddots & \\ \# & & \tilde{\mathbf{b}}_m'\mathbf{a}_m \end{bmatrix}.$$

Here, # is used for all elements other than the diagonal ones. In this book, the matrix product precedes the trace operation:

$$\mathrm{tr}\mathbf{AB} = \mathrm{tr}(\mathbf{AB}). \tag{1.25}$$

Thus,

$$\mathrm{tr}\mathbf{AB} = \sum_{i=1}^{n} \tilde{\mathbf{a}}_i'\mathbf{b}_i = \sum_{i=1}^{n} (a_{i1}b_{1i} + \cdots + a_{im}b_{mi}) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}b_{ji}, \tag{1.26}$$

$$\mathrm{tr}\mathbf{BA} = \sum_{j=1}^{m} \tilde{\mathbf{b}}_j'\mathbf{a}_j = \sum_{j=1}^{m} \left(b_{j1}a_{1j} + \cdots + b_{jn}a_{nj}\right) = \sum_{j=1}^{m}\sum_{i=1}^{n} b_{ji}a_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}b_{ji}. \tag{1.27}$$

Both are found to be equivalent, i.e.,

$$\mathrm{tr}\mathbf{AB} = \mathrm{tr}\mathbf{BA}, \tag{1.28}$$

and express the sum of $a_{ij}b_{ji}$ over all pairs of i and j.

It is an important property of the trace that (1.28) implies

$$\mathrm{tr}\mathbf{ABC} = \mathrm{tr}\mathbf{CAB} = \mathrm{tr}\mathbf{BCA}; \mathrm{tr}\mathbf{ABCD} = \mathrm{tr}\mathbf{BCDA} = \mathrm{tr}\mathbf{CDAB} = \mathrm{tr}\mathbf{DABC}. \tag{1.29}$$

Using (1.21), (1.28), and (1.29), we also have

$$\mathrm{tr}(\mathbf{AB})' = \mathrm{tr}\mathbf{B}'\mathbf{A}' = \mathrm{tr}\mathbf{A}'\mathbf{B}'; \mathrm{tr}(\mathbf{ABC})' = \mathrm{tr}\mathbf{C}'\mathbf{B}'\mathbf{A}' = \mathrm{tr}\mathbf{A}'\mathbf{C}'\mathbf{B}' = \mathrm{tr}\mathbf{B}'\mathbf{A}'\mathbf{C}'. \quad (1.30)$$

Substituting $\mathbf{A}'$ into $\mathbf{B}$ in (1.25), we have $\mathrm{tr}\mathbf{A}\mathbf{A}' = \mathrm{tr}\mathbf{A}'\mathbf{A} = \sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}^2$ which is the sum of the squared elements of $\mathbf{A}$. This is called the *squared norm* of $\mathbf{A}$, i.e., the matrix version of (1.12), and is denoted as $\|\mathbf{A}\|^2$:

$$\|\mathbf{A}\|^2 = \mathrm{tr}\mathbf{A}\mathbf{A}' = \mathrm{tr}\mathbf{A}'\mathbf{A} = \sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}^2. \quad (1.31)$$

This is also referred to as the *squared Frobenius norm* of $\mathbf{A}$, with Frobenius (1849–1917) a German mathematician. The squared norm of the sum of matrices weighted by scalars is expanded as

$$\begin{aligned}\|s\mathbf{X}+t\mathbf{Y}\|^2 &= \mathrm{tr}(s\mathbf{X}+t\mathbf{Y})'(s\mathbf{X}+t\mathbf{Y}) \\ &= \mathrm{tr}\left(s^2\mathbf{X}'\mathbf{X}+st\mathbf{X}'\mathbf{Y}+ts\mathbf{Y}'\mathbf{X}+t^2\mathbf{Y}'\mathbf{Y}\right) \\ &= s^2\mathrm{tr}\mathbf{X}'\mathbf{X}+st\mathrm{tr}\mathbf{X}'\mathbf{Y}+st\mathrm{trX}'\mathbf{Y}+t^2\mathrm{trY}'\mathbf{Y} \\ &= s^2\mathrm{tr}\mathbf{X}'\mathbf{X}+2st\mathrm{tr}\mathbf{X}'\mathbf{Y}+t^2\mathrm{tr}\mathbf{Y}'\mathbf{Y} \\ &= s^2\|\mathbf{X}\|^2+2st\mathrm{tr}\mathbf{X}'\mathbf{Y}+t^2\|\mathbf{Y}\|^2.\end{aligned} \quad (1.32)$$

1.8 Vectors and Matrices Filled with Ones or Zeros

A *zero vector* refers to a vector filled with zeros. The $p \times 1$ zero vector is denoted as $\mathbf{0}_{p,}$ using the boldfaced zero:

$$\mathbf{0}_p = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \quad (1.33)$$

A zero matrix refers to a matrix whose elements are all zeros. In this book, the $n \times p$ zero matrix is denoted as ${}_n\mathbf{O}_p$, using the boldfaced "O":

$${}_n\mathbf{O}_p = \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ldots & \vdots \\ 0 & \cdots & 0 \end{bmatrix}. \quad (1.34)$$

A *vector of ones* refers to a vector filled with ones. The $n \times 1$ vector of ones is denoted as $\mathbf{1}_n$, with the boldfaced one:

$$1_n = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}. \tag{1.35}$$

The $n \times p$ matrix filled with ones is given by

$$1_n 1_p' = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ldots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}. \tag{1.36}$$

1.9 Special Square Matrices

A square matrix $\mathbf{S} = (s_{ij})$ satisfying

$$\mathbf{S} = \mathbf{S}', \text{ i.e., } s_{ij} = s_{ji} \tag{1.37}$$

is said to be *symmetric*. An example of a 3×3 symmetric matrix is $\mathbf{S} = \begin{bmatrix} 2 & -4 & 9 \\ -4 & 6 & -7 \\ 9 & -7 & 3 \end{bmatrix}$. The products of a matrix $\mathbf{A}$ and its transpose, $\mathbf{AA}'$ and $\mathbf{A}'\mathbf{A}$, are symmetric; using (1.21), we have

$$(\mathbf{AA}')' = (\mathbf{A}')'\mathbf{A}' = \mathbf{AA}' \text{ and } (\mathbf{A}'\mathbf{A})' = \mathbf{A}'(\mathbf{A}')' = \mathbf{A}'\mathbf{A}. \tag{1.38}$$

This has already been exemplified in Ex. 3 (Sect. 1.5).

The elements of $\mathbf{A} = (a_{ij})$ with $i \neq j$ are called the off-diagonal elements of $\mathbf{A}$. A square matrix $\mathbf{D}$ whose off-diagonal elements are all zeros,

$$\mathbf{D} = \begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & d_p \end{bmatrix}, \tag{1.39}$$

is called a *diagonal matrix*. The products of two diagonal matrices are easily obtained as

$$\begin{bmatrix} c_1 & 0 & \cdots & 0 \\ 0 & c_2 & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & c_p \end{bmatrix} \begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & d_p \end{bmatrix} = \begin{bmatrix} c_1 d_1 & 0 & \cdots & 0 \\ 0 & c_2 d_2 & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & c_p d_p \end{bmatrix}, \tag{1.40}$$

$$\mathbf{D}^t = \mathbf{D}\mathbf{D}\ldots\mathbf{D} = \begin{bmatrix} d_1^t & 0 & \cdots & 0 \\ 0 & d_2^t & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & d_p^t \end{bmatrix}, \tag{1.41}$$

where $\mathbf{D}^t$ denotes the matrix obtained by multiplying $\mathbf{D}$ t times. Thus, we use the following expression:

$$\mathbf{D}^{-t} = \begin{bmatrix} d_1^{-t} & 0 & \cdots & 0 \\ 0 & d_2^{-t} & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & d_p^{-t} \end{bmatrix}. \tag{1.42}$$

When $t = 1/2$, (1.42) shows $\mathbf{D}^{-1/2}$ whose diagonal elements are $d_1^{-1/2}$, …, $d_p^{-1/2}$.

The *identity matrix* refers to the diagonal matrix whose diagonal elements are all ones. The $p \times p$ identity matrix is denoted as $\mathbf{I}_p$, using the boldfaced "I":

$$\mathbf{I}_p = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{bmatrix}. \tag{1.43}$$

For example, $\mathbf{I}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$. An important property of the identity matrix is

$$\mathbf{A}\mathbf{I}_p = \mathbf{A} \text{ and } \mathbf{I}_p\mathbf{B} = \mathbf{B}. \tag{1.44}$$

The identity matrix of 1×1 is $\mathbf{I}_1 = 1$, with $s \times 1 = 1 \times s = s$. That is, $\mathbf{I}_p$ is a generalization of one (or unit).

1.10 Bibliographical Notes

Matrix operations, which are necessary for describing multivariate data analysis, but have not been treated in this chapter, are introduced in the following chapters. As in the present book, introductory matrix operations are treated intelligibly in Carroll, Green, and Chaturvedi (1997), where geometric illustrations are emphasized. Banerjee and Roy (2014) and Schott (2005) are among the textbooks recommended for those who finished reading Chaps. 1–3 in the present book. Formulas for matrix operations are exhaustively listed in Lütkepohl (1996).

Exercises

1.1. Let $\mathbf{X} = (x_{ij})$ be an $n \times p$ matrix. Express $\mathbf{X}$ using $n \times 1$ vectors $\mathbf{x}_j = [x_{1j}, \ldots, x_{nj}]', j = 1, \ldots, p$, and express $\mathbf{X}$ using $p \times 1$ vectors $\tilde{\mathbf{x}}_i = [x_{i1}, \ldots, x_{ip}]', i = 1, \ldots, n$.

1.2. Let $\mathbf{A} = \begin{bmatrix} -2 & 3 & 9 \\ 1 & -6 & -5 \\ 8 & 2 & 0 \\ -4 & 6 & -3 \end{bmatrix}$ and $\mathbf{B} = \begin{bmatrix} 5 & 7 \\ 6 & -8 \\ -2 & 1 \end{bmatrix}$. Compute $\mathbf{AB}$, $\mathbf{B}'\mathbf{B}$, $\mathbf{BB}'$, $\mathbf{A}'\mathbf{A}$, and $\mathbf{AA}'$.

1.3. Let $\mathbf{A}_1 = \begin{bmatrix} -2 & 3 & 9 \\ 1 & -6 & -5 \\ 8 & 2 & 0 \\ -4 & 6 & -3 \end{bmatrix}$, $\mathbf{A}_2 = \begin{bmatrix} 7 & -1 & -5 \\ -2 & -2 & 3 \\ 0 & 3 & 9 \\ 6 & -4 & 0 \end{bmatrix}$, $\mathbf{B}_1 = \begin{bmatrix} 2 & -3 \\ -9 & 6 \\ 1 & -7 \end{bmatrix}$, $\mathbf{B}_2 = \begin{bmatrix} 5 & 7 \\ 6 & -8 \\ -2 & 1 \end{bmatrix}$, $s_1 = -5$, $s_2 = 7$, $t_1 = 3$, and $t_2 = -2$. Compute $(s_1\mathbf{A}_1 + s_2\mathbf{A}_2)(t_1\mathbf{B}_1 + t_2\mathbf{B}_2)$.

1.4. Let $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_m]$. Show $\mathbf{AB} = [\mathbf{Ab}_1, \ldots, \mathbf{Ab}_m]$.

1.5. Prove $\mathrm{tr}\mathbf{ABCDE} = \mathrm{tr}\mathbf{C}'\mathbf{B}'\mathbf{A}'\mathbf{E}'\mathbf{D}'$.

1.6. Let $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_p]$. Show that the (j, k) element of $\mathbf{W}'\mathbf{X}'\mathbf{XW}$ is $\mathbf{w}_j'\mathbf{X}'\mathbf{X}\mathbf{w}_k$ and $\mathrm{tr}\mathbf{W}'\mathbf{X}'\mathbf{XW} = \sum_{j=1}^{p} \mathbf{w}_j'\mathbf{X}'\mathbf{X}\mathbf{w}_j$.

1.7. Let a matrix $\mathbf{F}$ satisfy $\frac{1}{n}\mathbf{F}'\mathbf{F} = \mathrm{I}_m$. Show $\|\mathbf{X} - \mathbf{FA}'\|^2 = \|\mathbf{X}\|^2 - 2\mathrm{tr}\mathbf{F}'\mathbf{XA} + n\|\mathbf{A}\|^2$.

1.8. Compute $\mathbf{1}_4'[4, 2, 6, 1]'$ and $\mathbf{1}_4[4, 2, 6, 1]$.

1.9. Prove $\left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\right)'\left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\right) = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'$.

1.10. Show that $\frac{1}{n}\mathbf{F}'\mathbf{F} = \mathbf{I}_3$, if $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3]$ with $\|\mathbf{f}_j\|^2 = n (j = 1, 2, 3)$ and $\mathbf{f}_j'\mathbf{f}_k = 0$ for $j \neq k$.

Chapter 2
Intra-variable Statistics

This chapter begins with expressing data sets by matrices. Then, we introduce two statistics (statistical indices), average and variance, where the *average* is an index value that represents scores and the *variance* stands for how widely scores disperse. Further, how the original scores are transformed into *centered* and *standard scores* using the average and variance is described.

As the statistics in this chapter summarize the scores *within* a variable, the chapter is named *intra*-variable statistics, in contrast to the immediately following chapter entitled *inter*-variable statistics, where the statistics *between* variables would be treated.

2.1 Data Matrices

A *multivariate data* set refers to a set of values arranged in a table whose rows and columns are individuals and variables, respectively. This is illustrated in each panel of Table 2.1. Here, the term "*individuals*" implies the sources from which data are obtained; for example, individuals are examinees, cities, and baseball teams, respectively, in Panels (A), (B), and (C) of Table 2.1. On the other hand, the term "*variables*" refers to the indices or items for which individuals are measured; for example, variables are Japanese, mathematics, English, and sciences in Table 2.1(A). By attaching "multi-" to "variate", which is a synonym of "variable", we use the adjective "*multivariate*" for the data sets with multiple variables, as in Table 2.1. On the other hand, data with a single variable are called *univariate data*.

K. Adachi, *Matrix-Based Introduction to Multivariate Data Analysis*,
https://doi.org/10.1007/978-981-15-4103-2_2

Table 2.1 Three examples of multivariate data

(A) Test scores (artificial example)

Participant	Item			
	Japan	Mathematics	English	Science
1	82	70	70	76
2	96	65	67	71
3	84	41	54	65
4	90	54	66	80
5	93	76	74	77
6	82	85	60	89

(B) Weather in cities in January (http://www2m.biglobe.ne.jp/ZenTech/world/kion/Japan/Japan.htm)

City	Weather		
	Min °C	Max °C	Precipitation
Sapporo	−7.7	−0.9	110.7
Tokyo	2.1	9.8	48.6
•	•	•	•
•	•	•	•
•	•	•	•
Naha	14.3	19.1	114.5

(C) Team scores (2005 in Japan) (http://npb.jp/bis/2005/stats/)

Team	Averages				
	Win %	Runs	HR	Avg.	ERA
Tigers	0.617	731	140	0.274	3.24
Dragons	0.545	680	139	0.269	4.13
BayStars	0.496	621	143	0.265	3.68
Swallows	0.493	591	128	0.276	4.00
Giants	0.437	617	186	0.260	4.80
Carp	0.408	615	184	0.275	4.80

Let us express a data set as an n-individuals $\times$ p-variables matrix

$$\mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix} = \left[\mathbf{x}_1, \ldots, \mathbf{x}_j, \ldots, \mathbf{x}_p\right], \tag{2.1}$$

whose jth column

$$\mathbf{x}_j = \begin{bmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{bmatrix} = \left[x_{1j}, \ldots, x_{nj}\right]' \tag{2.2}$$

stands for the jth *variable*. Examples of (2.1) have been given in Table 2.1(A), (B), and (C).

A different example is presented in Table 2.2(A), where n individuals and p variables are six students and two items, respectively, with x_{ij} the score of student i for test j and $\mathbf{x}_j$ the 6×1 vector containing the scores on the jth test:

$$\underset{6\times 2}{\mathbf{X}} = \begin{bmatrix} 66 & 74 \\ 72 & 98 \\ \vdots & \vdots \\ 56 & 84 \end{bmatrix} \text{ with } \mathbf{x}_1 = \begin{bmatrix} 66 \\ 72 \\ \vdots \\ 56 \end{bmatrix} \text{ and } \mathbf{x}_2 = \begin{bmatrix} 74 \\ 98 \\ \vdots \\ 84 \end{bmatrix}.$$

The scores in Table 2.2(B) and (C) will be explained later, in Sects. 2.4 and 2.6.

2.2 Distributions

The distribution of the six students' scores for each variable in Table 2.2(A) is graphically depicted in Fig. 2.1, where those scores are plotted on lines extending from 0 to 100. The distributions allow us to intuitively recognize that [1] their scores in history are lower on *average* than those in mathematics, and [2] the scores *disperse* more widely in mathematics than in history. The statistics related to [1] and [2] are introduced in Sects. 2.3 and 2.5, respectively.

Table 2.2 Raw, centered, and standard scores of tests with their averages, variances, and standard deviations (SD) (artificial example)

Student	(A) Raw		(B) Centered		(C) Standard	
	History	Math	History	Math	History	Math
1	66	74	5	−3	0.52	−0.20
2	72	98	11	21	1.15	1.43
3	44	62	−17	−15	−1.78	−1.02
4	58	88	−3	11	−0.31	0.75
5	70	56	9	−21	0.94	−1.43
6	56	84	−5	7	−0.52	0.48
Average	61.0	77.0	0	0	0	0
Variance	91.67	214.33	91.67	214.33	1.00	1.00
SD	9.57	14.64	9.57	14.64	1.00	1.00

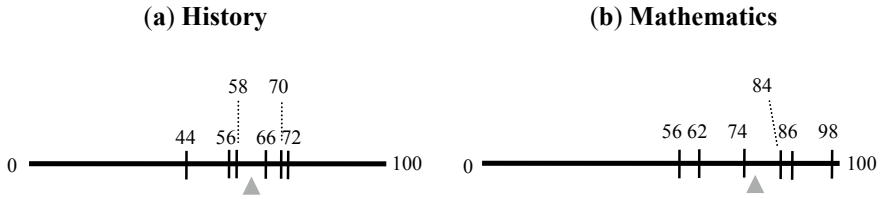


Fig. 2.1 Distributions of the test scores in Table 2.2(A)

2.3 Averages

Let us consider summarizing n scores into a single statistic. The most popular statistic for the summary is the *average*, which is defined as

$$\bar{x}_j = \frac{1}{n}\left(x_{1j} + \cdots + x_{nj}\right) = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \tag{2.3}$$

for variable j, i.e., the jth column of $\mathbf{X}$. For example, the average score in mathematics ($j = 2$) in Table 2.2(A) is $\bar{x}_2 = (74 + 98 + 62 + 88 + 56 + 84)/6 = 77.0$. The average can be rewritten, using the $n \times 1$ *vector of ones* $\mathbf{1}_n = [1, 1, \ldots, 1]'$ defined in (1.35): The *sum* $x_{1j} + \cdots + x_{nj}$ is expressed as

$$\mathbf{1}_n'\mathbf{x}_j = [1, \ldots, 1]\begin{bmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{bmatrix}, \tag{2.4}$$

thus, the *average* (2.3) is also simply expressed as

$$\bar{x}_j = \frac{1}{n}\mathbf{1}_n'\mathbf{x}_j, \tag{2.5}$$

without using the complicated "Sigma" symbol. For example, the average score in history ($j = 1$) in Table 2.2(A) is expressed as $6^{-1}\mathbf{1}_6'\mathbf{x}_1$ with $\mathbf{x}_1 = [66, 72, 44, 58, 70, 56]'$. The resulting average is $6^{-1}\mathbf{1}_6'\mathbf{x}_1 = 61.0$.

2.4 Centered Scores

The raw scores minus their average are called *centered scores* or *deviations from average*. Let the centered score vector for variable j be denoted as $\mathbf{y}_j = [y_{1j}, \ldots, y_{nj}]'$ ($n \times 1$), which is expressed as

$$\mathbf{y}_i = \begin{bmatrix} y_{1j} \\ \vdots \\ y_{nj} \end{bmatrix} = \begin{bmatrix} x_{1j} - \bar{x}_j \\ \vdots \\ x_{nj} - \bar{x}_j \end{bmatrix} = \begin{bmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{bmatrix} - \begin{bmatrix} \bar{x}_j \\ \vdots \\ \bar{x}_j \end{bmatrix} = \mathbf{x}_i - \begin{bmatrix} \bar{x}_j \\ \vdots \\ \bar{x}_j \end{bmatrix}. \tag{2.6}$$

In Table 2.2(B), the centered data for (A) are shown: The centered scores [5, 11, …, −5]′ for history are given by subtracting 61 from all elements of [66, 72, …, 56]′ and the centered scores for mathematics are given by subtracting 77 in a parallel manner.

Here, we rewrite (2.6) in a simpler form. First, let us note that all elements of the subtracted vector $\left[\bar{x}_j, \ldots, \bar{x}_j\right]'$ in (2.6) are equal to an average $\bar{x}_j$, thus, that vector can be written as

$$\begin{bmatrix} \bar{x}_j \\ \vdots \\ \bar{x}_j \end{bmatrix} = \bar{x}_j \mathbf{1}_n = \mathbf{1}_n \times \bar{x}_j, \tag{2.7}$$

where we have used (1.20). Substituting (2.5) into $\bar{x}_j$ in (2.7), this is rewritten as

$$\begin{bmatrix} \bar{x}_j \\ \vdots \\ \bar{x}_j \end{bmatrix} = \mathbf{1}_n \times \left(\frac{1}{n}\mathbf{1}_n'\mathbf{x}_j\right) = \frac{1}{n}\mathbf{1}_n\left(\mathbf{1}_n'\mathbf{x}_j\right) = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\mathbf{x}_j. \tag{2.8}$$

Here, we have made use of the fact that "× scalar (n^{-1})" can be moved and $\mathbf{A}(\mathbf{BC}) = \mathbf{ABC}$ generally holds for matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$, which implies $\mathbf{1}_n\left(\mathbf{1}_n'\mathbf{x}_j\right) = \mathbf{1}_n\mathbf{1}_n'\mathbf{x}_j$. Using (2.8) in (2.6) and noting property (1.44) for an identity matrix, the *centered score vector* (2.6) can be rewritten as

$$\mathbf{y}_j = \begin{bmatrix} x_{1j} - \bar{x}_j \\ \vdots \\ x_{nj} - \bar{x}_j \end{bmatrix} = \mathbf{x}_j - \begin{bmatrix} \bar{x}_j \\ \vdots \\ \bar{x}_j \end{bmatrix} = \mathbf{I}_n\mathbf{x}_j - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\mathbf{x}_j = \left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\right)\mathbf{x}_j = \mathbf{J}\mathbf{x}_j, \tag{2.9}$$

where $\mathbf{J} = \mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n'$, and we have made use of the fact that $\mathbf{BC} + \mathbf{EC} = (\mathbf{B} + \mathbf{E})\mathbf{C}$ holds for matrices $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{E}$. The matrix $\mathbf{J}$ has a special name and important properties:

Note 2.1. Centering Matrix
This is defined as

$$\mathbf{J} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'. \tag{2.10}$$

The centering matrix has the following properties:

$$\mathbf{J} = \mathbf{J}'(\text{symmetric}), \tag{2.11}$$

$$\mathbf{J}^2 = \mathbf{JJ} = \mathbf{J}(\textit{idempotent}), \tag{2.12}$$

$$\mathbf{1}_n'\mathbf{J} = \mathbf{0}_n', \tag{2.13}$$

Here, an *idempotent* matrix is defined as follows: $\mathbf{S}$ is said to be idempotent if $\mathbf{SS} = \mathbf{S}$.

Equation (2.11) can easily be found. Equations (2.12) and (2.13) can be proved as follows:

$$\begin{aligned}\mathbf{JJ} &= \left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\right)\left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\right) = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n' - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n' + \frac{1}{n^2}\mathbf{1}_n\mathbf{1}_n'\mathbf{1}_n\mathbf{1}_n' \\ &= \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n' - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n' + \frac{1}{n^2}\mathbf{1}_n(n)\mathbf{1}_n' = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n' = \mathbf{J}, \\ \mathbf{1}_n'\mathbf{J} &= \mathbf{1}_n'\left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\right) = \mathbf{1}_n' - \frac{1}{n}\mathbf{1}_n'\mathbf{1}_n\mathbf{1}_n' = \mathbf{1}_n' - \frac{1}{n}(n)\mathbf{1}_n' = \mathbf{0}_n',\end{aligned}$$

where $\mathbf{1}_n'\mathbf{1}_n = n$ has been used.

Equations (2.12) and (2.13) further lead to the following important facts:

Note 2.2. Matrices Pre-multiplied by the Centering Matrix
A matrix $s\mathbf{JA}$ with $\mathbf{A}$ an $n \times p$ matrix and s a scalar satisfies

$$\mathbf{1}_n'(s\mathbf{JA}) = s\mathbf{1}_n'\mathbf{JA} = \mathbf{0}_p', \tag{2.14}$$

$$\mathbf{J}(s\mathbf{JA}) = s\mathbf{JJA} = s\mathbf{JA}. \tag{2.15}$$

When $\mathbf{A}$ is an $n \times 1$ vector $\mathbf{a}$, those equations are rewritten as $\mathbf{1}_n'(s\mathbf{Ja}) = 0$ and $\mathbf{J}(s\mathbf{Ja}) = s\mathbf{Ja}$, respectively.

Comparing (2.9) with (2.14), we can find that the sum and average of centered scores are always *zero*:

$$\mathbf{1}_n'\mathbf{y}_j = \frac{1}{n}\mathbf{1}_n'\mathbf{y}_j = 0. \tag{2.16}$$

This is shown in the row named "Average" in Table 2.2(B). Figure 2.3(B) (*on a later page*) illustrates (2.16); the centered scores are distributed with their average being the zero which is a *center between negative and positive values*. This property provides the name "centered scores", and the transformation of raw scores into centered ones is called *centering*. Comparing (2.9) with (2.15), we also find

$$\mathbf{J}\mathbf{y}_j = \mathbf{y}_j. \tag{2.17}$$

The centered score vector, pre-multiplied by the centering matrix, remains unchanged.

2.5 Variances and Standard Deviations

The locations of averages in the distributions of scores are indicated by triangles in Fig. 2.1(A), which do not stand for how widely scores disperse. The most popular statistic for indicating dispersion is *variance*. It is defined using the *sum of squared distances* between *scores* and *their average*, which is illustrated in Fig. 2.2. The division of the *sum* by the number of scores gives the *variance*. Denoting the variance for variable j as v_{jj}, it is formally expressed as

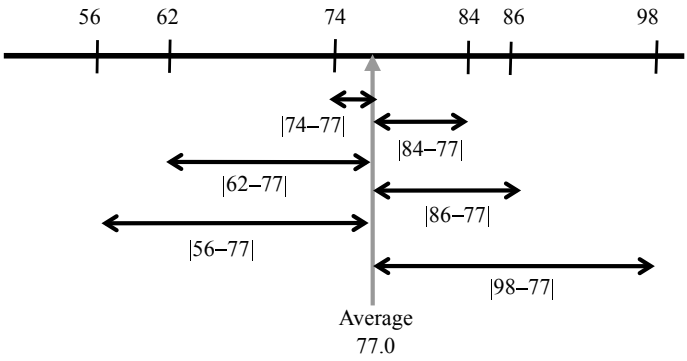


Fig. 2.2 Distances of scores to their average, which are squared, summed, and divided by n, to give the variance of the mathematics scores in Table 2.2(A)

$$\begin{aligned} v_{jj} &= \frac{1}{n}\left\{\left|x_{1j}-\bar{x}_j\right|^2+\cdots+\left|x_{ni}-\bar{x}_j\right|^2\right\} \\ &= \frac{1}{n}\left\{\left(x_{1j}-\bar{x}_j\right)^2+\cdots+\left(x_{ni}-\bar{x}_j\right)^2\right\} = \frac{1}{n}\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_j\right)^2, \end{aligned} \tag{2.18}$$

where the same subscript j is used twice as v_{jj}, for the sake of accordance with the related statistic introduced in the next chapter. The variance of the scores for mathematics in Table 2.2(A) is obtained as $6^{-1}\{(74-77)^2+(98-77)^2+\cdots+(84-77)^2\} = 214.33$, for example.

To express (2.18) in *vector* form, we should note that it can be rewritten as

$$v_{jj} = \frac{1}{n}\left[x_{1j}-\bar{x}_j,\ldots,x_{nj}-\bar{x}_j\right]\begin{bmatrix} x_{1j}-\bar{x}_j \\ \vdots \\ x_{nj}-\bar{x}_j \end{bmatrix}. \tag{2.19}$$

Comparing (2.19) with $\begin{bmatrix} x_{1j}-\bar{x}_j \\ \vdots \\ x_{nj}-\bar{x}_j \end{bmatrix} = \mathbf{J}\mathbf{x}_j$ in (2.9), the *variance* (2.18) or (2.19) is expressed as

$$v_{jj} = \frac{1}{n}\left(\mathbf{J}\mathbf{x}_j\right)'\mathbf{J}\mathbf{x}_j = \frac{1}{n}\left\|\mathbf{J}\mathbf{x}_j\right\|^2 = \frac{1}{n}\mathbf{x}_j'\mathbf{J}'\mathbf{J}\mathbf{x}_j = \frac{1}{n}\mathbf{x}_j'\mathbf{J}\mathbf{x}_j = \frac{1}{n}\mathbf{x}_j'\mathbf{J}\mathbf{x}_j, \tag{2.20}$$

where (1.12), (2.11), and (2.12) have been used. Further, we can use (2.9) in (2.20) to rewrite it as

$$v_{jj} = \frac{1}{n}\mathbf{x}_j'\mathbf{J}\mathbf{x}_j = \frac{1}{n}\mathbf{x}_j'\mathbf{J}'\mathbf{J}\mathbf{x}_j = \frac{1}{n}\mathbf{y}_j'\mathbf{y}_j = \frac{1}{n}\left\|\mathbf{y}_j\right\|^2. \tag{2.21}$$

The variance of raw scores is expressed using their *centered score vector* simply as $n^{-1}\|\mathbf{y}_j\|^2$. We can also find in (2.20) and (2.21) that the variance is the squared length of vector $\mathbf{y}_j = \mathbf{J}\mathbf{x}_j$ divided by n.

How is the variance of the centered scores (rather than raw scores) expressed? To find this, we substitute the centered score vector $\mathbf{y}_j$ for $\mathbf{x}_j$ in the variance (2.20). Then, we use (2.17) and (2.9) to get

$$\frac{1}{n}\mathbf{y}_j'\mathbf{J}'\mathbf{J}\mathbf{y}_j = \frac{1}{n}\mathbf{y}_j'\mathbf{y}_j = \frac{1}{n}\mathbf{x}_j'\mathbf{J}'\mathbf{J}\mathbf{x}_j, \tag{2.22}$$

which is equal to (2.20); the variance of the centered scores equals that for their raw scores.

The square root of variance (20.20), (2.21), or (2.22)

$$\sqrt{v_{jj}} = \sqrt{\frac{1}{n}\mathbf{x}_j'\mathbf{J}\mathbf{x}_j} = \frac{1}{\sqrt{n}}\left\|\mathbf{J}\mathbf{x}_j\right\| = \frac{1}{\sqrt{n}}\left\|\mathbf{y}_j\right\| \tag{2.23}$$

is called the *standard deviation*, which is also used for reporting the dispersion of data. We can find in (2.23) that the standard deviation is the length of vector $\mathbf{y}_j = \mathbf{J}\mathbf{x}_j$ divided by $n^{1/2}$.

2.6 Standard Scores

The centered scores (i.e., the raw scores minus their average) divided by their standard deviation are called *standard scores* or *z scores*. Let the standard score vector for variable j be denoted by $\mathbf{z}_j = \left[z_{1j}, \ldots, z_{nj}\right]'$, which is expressed as

$$\mathbf{z}_j = \begin{bmatrix} \left(x_{1j} - \bar{x}_j\right)/\sqrt{v_{jj}} \\ \vdots \\ \left(x_{nj} - \bar{x}_j\right)/\sqrt{v_{jj}} \end{bmatrix} = \frac{1}{\sqrt{v_{jj}}}\begin{bmatrix} x_{1j} - \bar{x}_j \\ \vdots \\ x_{nj} - \bar{x}_j \end{bmatrix} = \frac{1}{\sqrt{v_{jj}}}\mathbf{J}\mathbf{x}_i = \frac{1}{\sqrt{v_{jj}}}\mathbf{y}_j, \tag{2.24}$$

where we have used (2.9). In Table 2.2(C), the standard scores for (A) are shown; the standard scores $[-0.20, \ldots, 0.48]'$ for mathematics are given by dividing its centered scores (B) by 14.64. Transforming raw scores into standard ones is called *standardization*.

Standard scores have two important properties. One is that the sum and average of standard scores are always *zero*, as are those of centered scores:

$$\mathbf{1}_n'\mathbf{z}_j = \frac{1}{n}\mathbf{1}_n'\mathbf{z}_j = 0, \tag{2.25}$$

which follows from (2.16) and (2.24). The other property is that the variance of standard scores is always *one*, which is shown as follows: The substitution of $\mathbf{z}_j$ into $\mathbf{x}_j$ in (2.20) leads to the variance of standard scores being expressed as $n^{-1}\mathbf{z}_j'\mathbf{J}'\mathbf{J}\mathbf{z}_j = n^{-1}\mathbf{z}_j'\mathbf{z}_j = n^{-1}\left\|\mathbf{z}_j\right\|^2$, where we have used $\mathbf{z}_j = \mathbf{J}\mathbf{z}_j$, following from the use of (2.17) in (2.24). Further, we can substitute (2.24) in $n^{-1}\|\mathbf{z}_j\|^2$ and use (2.21) to rewrite the variance of standard scores as

$$\frac{1}{n}\left\|\mathbf{z}_j\right\|^2 = \frac{1}{nv_{jj}}\left\|\mathbf{y}_j\right\|^2 = \frac{nv_{jj}}{nv_{jj}} = 1 \tag{2.26}$$

This also implies that the length of every standard score vector is always $\|\mathbf{z}_j\| = n^{1/2}$.

2.7 What Centering and Standardization Do for Distributions

The properties of centered and standard scores shown with (2.16), (2.22), (2.25), and (2.26) are summarized in Table 2.3. Figure 2.3 illustrates the roles that centering and standardization (i.e., transforming raw scores into centered and standard ones) perform for the distributions of data: *Centering* simply *moves the distributions* of raw scores so that the average of the moved distributions is zero, and *standardization* further *accommodates the scale* of the moved distributions so that their variances are equal to one. The standard scores are unified among different variables so that the averages and variances are zero and one, respectively, thus, the greatness/smallness of the standard scores can be compared reasonably between variables.

Table 2.3 Averages and variances of centered and standard scores

	Average	Variance
Centered scores	0	Variance of raw scores
Standard scores	0	1

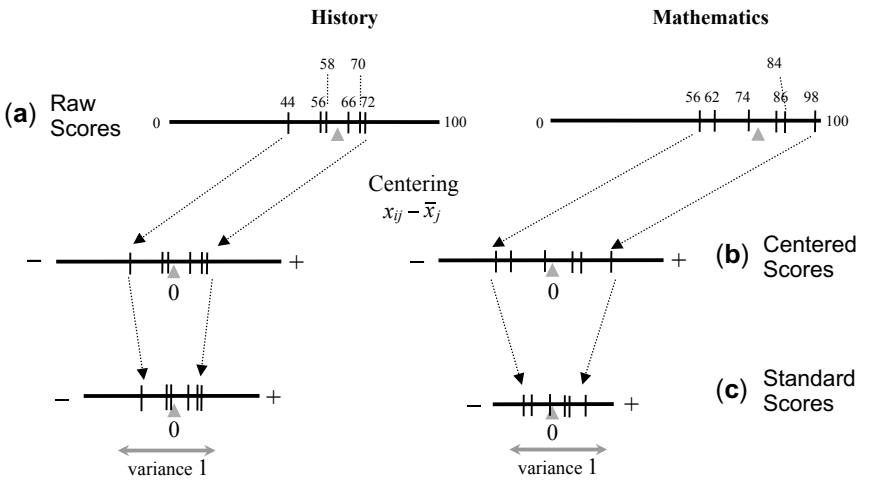


Fig. 2.3 Distributions of raw, centered, and standard scores in Table 2.2

2.8 Matrix Representation

We will now introduce a basic formula in matrix algebra:

Note 2.3. A Property of Matrix Product
If $\mathbf{A}$ is a matrix of $n \times m$ and $\mathbf{b}_1, \ldots, \mathbf{b}_K$ are $m \times 1$ vectors, then

$$[\mathbf{A}\mathbf{b}_1, \ldots, \mathbf{A}\mathbf{b}_K] = \mathbf{A}[\mathbf{b}_1, \ldots, \mathbf{b}_K]. \tag{2.27}$$

Using this and (2.5), the $1 \times p$ row vector containing the *averages* of p variables is expressed as

$$\left[\bar{x}_1, \ldots, \bar{x}_p\right] = \left[\frac{1}{n}\mathbf{1}_n'\mathbf{x}_1, \ldots, \frac{1}{n}\mathbf{1}_n'\mathbf{x}_p\right] = \frac{1}{n}\mathbf{1}_n'\left[\mathbf{x}_1, \ldots, \mathbf{x}_p\right] = \frac{1}{n}\mathbf{1}_n'\mathbf{X}. \tag{2.28}$$

For example, when $\mathbf{X}$ consists of the six students' scores in Table 2.2(A), $6^{-1}\mathbf{1}_6'\mathbf{X} = [61.0, 77.0]$.

Let $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_p]$ denote the $n \times p$ matrix of *centered scores* whose jth column is defined as (2.9) for the corresponding column of $\mathbf{X}$. Then, we can use (2.9) and (2.27) to express $\mathbf{Y}$ as

$$\mathbf{Y} = \left[\mathbf{J}\mathbf{x}_1, \ldots, \mathbf{J}\mathbf{x}_p\right] = \mathbf{J}\left[\mathbf{x}_1, \ldots, \mathbf{x}_p\right] = \mathbf{J}\mathbf{X}, \tag{2.29}$$

an example of which is presented in Table 2.2(B).

Let $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_p]$ denote the $n \times p$ matrix of *standard scores* whose jth column is defined as (2.24) for the corresponding columns of $\mathbf{X}$ and $\mathbf{Y}$. Then, $\mathbf{Z}$ is expressed as

$$\mathbf{Z} = \left[\frac{1}{\sqrt{v_{11}}}\mathbf{y}_1, \ldots, \frac{1}{\sqrt{v_{pp}}}\mathbf{y}_p\right] = \left[\mathbf{y}_1, \ldots, \mathbf{y}_p\right]\begin{bmatrix} \frac{1}{\sqrt{v_{11}}} & & \\ & \ddots & \\ & & \frac{1}{\sqrt{v_{pp}}} \end{bmatrix} = \mathbf{Y}\mathbf{D}^{-1}. \tag{2.30}$$

Here, the blanks in $\begin{bmatrix} \frac{1}{\sqrt{v_{11}}} & & \\ & \ddots & \\ & & \frac{1}{\sqrt{v_{pp}}} \end{bmatrix}$ stand for the corresponding elements being zeros and $\mathbf{D} = \begin{bmatrix} \sqrt{v_{11}} & & \\ & \ddots & \\ & & \sqrt{v_{pp}} \end{bmatrix}$ is the $p \times p$ diagonal matrix whose diagonal elements are the standard deviations for p variables: We should recall (1.42) to

notice that $\mathbf{D}^{-1}$ is the diagonal matrix whose diagonal elements are the reciprocals of the standard deviations. Those readers who have difficulties in understanding (2.30) should note the following simple example with $\mathbf{Y}$ being 3×2:

$$\mathbf{YD}^{-1} = \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \\ y_{31} & y_{32} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{v_{11}}} & \\ & \frac{1}{\sqrt{v_{22}}} \end{bmatrix} = \begin{bmatrix} y_{11}/\sqrt{v_{11}} & y_{12}/\sqrt{v_{22}} \\ y_{21}/\sqrt{v_{11}} & y_{22}/\sqrt{v_{22}} \\ y_{31}/\sqrt{v_{11}} & y_{23}/\sqrt{v_{22}} \end{bmatrix}, \tag{2.31}$$

which illustrates the equalities in (2.30) in the reverse order. The standard score matrix $\mathbf{Z}$ can also be expressed as

$$\mathbf{Z} = \mathbf{JXD}^{-1}, \tag{2.32}$$

using (2.29) in (2.30).

2.9 Bibliographical Notes

Carroll, Green, and Chaturvedi (1997, Chap. 3) and Reyment and Jöreskog (2002, Chap. 2) are among the literature in which the matrix expressions of intra-variable statistics are intelligibly treated.

Exercises

2.1. Compute $\mathbf{J} = \mathbf{I}_5 - 5^{-1}\mathbf{1}_5\mathbf{1}_5'$ and obtain the centered score matrix $\mathbf{Y} = \mathbf{JX}$ for the 5×3 matrix $\mathbf{X}$ in Table 2.3.

2.2. Compute the variance $v_{jj} = 5^{-1}\mathbf{x}_j'\mathbf{J}\mathbf{x}_j$ $(j = 1, 2, 3)$, the diagonal matrix $\mathbf{D}^{-1} = \begin{bmatrix} \frac{1}{\sqrt{v_{11}}} & & \\ & \frac{1}{\sqrt{v_{22}}} & \\ & & \frac{1}{\sqrt{v_{33}}} \end{bmatrix}$, and the standard score matrix $\mathbf{Z} = \mathbf{JXD}^{-1}$ for $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3]$ (5×3) in Table 2.4.

2.3. Discuss the benefits of standardizing the data in Table 2.4.

2.4. Show that if the average for each column of $\mathbf{Y}$ $(n \times p)$ is zero, then the average for each column of $\mathbf{YA}$ is also zero.

Table 2.4 Data matrix $\mathbf{X}$ of 5 persons $\times$ 3 variables

Person	Height	Weight	Sight
Bill	172	64	0.8
Brian	168	70	1.4
Charles	184	80	1.2
Keith	176	64	0.2
Michael	160	62	1.0

2.5. Let $\mathbf{Z}$ be an n-individuals $\times$ p-variables matrix containing standard scores. Show that $\|\mathbf{Z}\|^2 = \text{tr}\mathbf{Z}'\mathbf{Z} = \text{tr}\mathbf{Z}\mathbf{Z}' = np$.

2.6. Let $\mathbf{x}$ be an $n \times 1$ vector with $v = n^{-1}\mathbf{x}'\mathbf{J}\mathbf{x}$ the variance of the elements in $\mathbf{x}$. Show that the variance of the elements in $b\mathbf{x} + c\mathbf{1}_n$ is b^2v.

2.7. Let $\mathbf{y} = [y_1, \ldots, y_n]'$ contain centered scores. Show that the average of the elements in $-\mathbf{y} + c\mathbf{1}_n = [-y_1 + c, \ldots, -y_n + c]'$ is c and their variance is equivalent to that for $\mathbf{y}$.

2.8. Let $\mathbf{z} = [z_1, \ldots, z_n]'$ contain standard scores. Show that the average of the elements in $b\mathbf{z} + c\mathbf{1}_n = [bz_1 + c, \ldots, bz_n + c]'$ is c and their standard deviation is b.

Chapter 3
Inter-variable Statistics

In the previous chapter, we described the two statistics, average and variance, which summarize the distribution of scores *within* a variable. In this chapter, we introduce *covariance* and the *correlation coefficient*, which are the *inter*-variable statistics indicating the relationships *between* two variables. Finally, the *rank* of a matrix, an important notion in linear algebra, is introduced.

3.1 Scatter Plots and Correlations

As in the previous chapter, we consider an n-individuals $\times$ p-variables data matrix (2.1), i.e., $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_j, \ldots, \mathbf{x}_p]$. An example of $\mathbf{X}$ is presented in Table 3.1(A). There, n individuals are 10 kinds of *foods* ($n = 10$) and p variables are their *sweetness*, degree of *spice*, and *sales* ($p = 3$). The relationship between two variables, j and k ($j, k = 1, \ldots, p$), which is called a *correlation*, can be captured by the *scatter plot* in which n individuals are plotted as points with their coordinates $[x_{ij}, x_{ik}]$, $i = 1, \ldots, n$, where x_{ij} and x_{ik} are the scores of individual i for variables j and k, respectively. The plots for Table 3.1(A) are shown in Fig. 3.1a–c. For example, (c) is the scatter plot for *sweetness* and *sales*, where 10 foods are plotted as points with their coordinates $[x_{i1}, x_{i3}]$, $i = 1, \ldots, 10$, i.e., [32, 62], [28, 83], … , [22, 63].

In Fig. 3.1, distributions are easily captured by the ellipses roughly surrounding the points. The slope of the ellipse in Fig. 3.1c shows that *sales* are positively proportional to *sweetness*. Two variables with such a proportional relation are said to have a *positive correlation*. The inverse relationship is found between *spice* and *sweetness* in Fig. 3.1a; the former tends to decrease with an increase in the latter, which is expressed as the variables having a *negative correlation*. On the other hand, the ellipse in Fig. 3.1b is not sloped; *no correlation* is found between *spice* and *sales*.

K. Adachi, *Matrix-Based Introduction to Multivariate Data Analysis*,
https://doi.org/10.1007/978-981-15-4103-2_3

Table 3.1 Data matrices of 10 individuals × 3 variables (artificial example)

Food	(A) Raw score: **X**			(B) Centered Scores: **Y**			(C) Standard Scores: **Z**		
	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_3$	$\mathbf{z}_1$	$\mathbf{z}_2$	$\mathbf{z}_3$
	Sweet	Spice	Sales	Sweet	Spice	Sales	Sweet	Spice	Sales
1	32	10	62	3.5	−7.7	−5.5	0.69	−1.77	−0.35
2	28	20	83	−0.5	2.3	15.5	−0.10	0.53	0.98
3	20	19	34	−8.5	1.3	−33.5	−1.68	0.30	−2.11
4	34	21	91	5.5	3.3	23.5	1.09	0.76	1.48
5	25	16	53	−3.5	−1.7	−14.5	−0.69	−0.39	−0.91
6	35	14	70	6.5	−3.7	2.5	1.28	−0.85	0.16
7	25	20	62	−3.5	2.3	−5.5	−0.69	0.53	−0.35
8	30	18	73	1.5	0.3	5.5	0.30	0.07	0.35
9	34	13	84	5.5	−4.7	16.5	1.09	−1.08	1.04
10	22	26	63	−6.5	8.3	−4.5	−1.28	1.90	−0.28
Average	28.5	17.7	67.5	0.00	0.00	0.00	0.00	0.00	0.00
Variance	25.65	19.01	251.45	25.65	19.01	251.45	1.00	1.00	1.00
SD	5.06	4.36	15.86	5.06	4.36	15.86	1.00	1.00	1.00

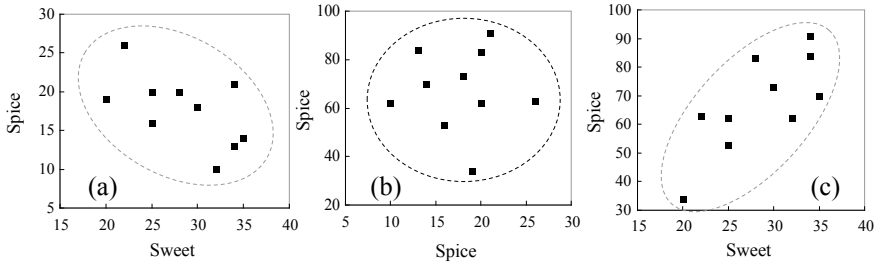


Fig. 3.1 Scatter plots for the pairs of the variables in Table 3.1(A)

3.2 Covariances

The correlation between two variables j and k can be indicated by a *covariance*, which is defined as

$$v_{jk} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)(x_{ik} - \overline{x}_k), \tag{3.1}$$

the average of the product $(x_{ij} - \overline{x}_j) \times (x_{ik} - \overline{x}_k)$ over $i = 1, \ldots, n$, with $x_{ij} - \overline{x}_j$ and $x_{ik} - \overline{x}_k$ being the centered scores for variables j and k, respectively. It takes a *positive* value when variables j and k have a positive correlation, while v_{jk} shows a *negative* value when the variables have a negative correlation, and v_{jk} is close to

Table 3.2 Covariance matrix for Table 3.1(A)

Variable	Sweet	Spice	Sales
Sweet	25.65	−12.65	60.15
Spice	−12.65	19.01	0.15
Sales	60.15	0.15	251.45

zero when j and k have no correlation. This property can be verified as follows: The covariance between the variables *sweet* and *spice* in Table 3.1(A) is computed as

$$v_{12} = \frac{1}{10}\{3.5 \times (-7.7) + (-0.5) \times 2.3 + \cdots + (-6.5) \times 8.3\} = -12.65, \quad (3.2)$$

using the centered scores in (B). Those variables are found negatively correlated in Fig. 3.1a and their covariance (3.2) is also negative. In a parallel manner, we can have the positive $v_{13} = 60.15$, which is the covariance between *sweet* and *sales* correlated positively, as in Fig. 3.1b, while we can find $v_{23} = 0.15$ closing to zero, which is the covariance between *spices* and *sales* uncorrelated, as in Fig. 3.1c. Those covariances are summarized in Table 3.2.

To express (3.1) in a *vector* form, (3.1) can be rewritten as

$$v_{jk} = \frac{1}{n}\left[x_{1j} - \overline{x}_j, \quad \ldots, \quad x_{nj} - \overline{x}_j\right] \begin{bmatrix} x_{1k} - \overline{x}_k \\ \vdots \\ x_{nk} - \overline{x}_k \end{bmatrix}. \quad (3.3)$$

Here, (2.9) should be recalled, noticing that the right vector in (3.3) can be expressed as $\begin{bmatrix} x_{1k} - \overline{x}_k \\ \vdots \\ x_{nk} - \overline{x}_k \end{bmatrix} = \mathbf{y}_k = \mathbf{J}\mathbf{x}_k$ by replacing the subscript j in (2.9) by k, with $\mathbf{y}_k$ the $n \times 1$ vector of centered scores corresponding to the raw scores $\mathbf{x}_k$ and $\mathbf{J}$ the $n \times n$ centering matrix defined in (2.10). Thus, (3.3) is rewritten as

$$v_{jk} = \frac{1}{n}\left(\mathbf{J}\mathbf{x}_j\right)'\mathbf{J}\mathbf{x}_k = \frac{1}{n}\mathbf{x}_j'\mathbf{J}'\mathbf{J}\mathbf{x}_k = \frac{1}{n}\mathbf{x}_j'\mathbf{J}\mathbf{J}\mathbf{x}_k = \frac{1}{n}\mathbf{x}_j'\mathbf{J}\mathbf{x}_k = \frac{1}{n}\mathbf{y}_j'\mathbf{y}_k, \quad (3.4)$$

in which (2.9), (2.11), and (2.12) have been used. That is, the covariance between variables j and k is the *inner product* of the *centered score vectors* $\mathbf{y}_j = \mathbf{J}\mathbf{x}_j$ and $\mathbf{y}_k = \mathbf{J}\mathbf{x}_k$ divided by n.

A p-variables $\times$ p-variables matrix containing covariances, as in Table 3.2, is called a *covariance matrix*. Each of its diagonal elements expresses the covariance for the same variable.

$$v_{jj} = \frac{1}{n}\mathbf{x}_j'\mathbf{J}\mathbf{x}_j = \frac{1}{n}\mathbf{y}_j'\mathbf{y}_j, \quad (3.5)$$

which equals (2.21), i.e., the *variance* for that variable. This implies that covariance is an extension of the concept of variance for two variables. We should thus consider *covariance* as including *variance* as a *special case*.

Let us substitute the centered score vector $\mathbf{y}_j$ for $\mathbf{x}_j$ in covariance (3.4). Then, we have

$$v_{jk} = \frac{1}{n}\left(\mathbf{J}\mathbf{y}_j\right)'\mathbf{J}\mathbf{y}_k = \frac{1}{n}\mathbf{y}_j'\mathbf{J}\mathbf{y}_k = \frac{1}{n}\mathbf{y}_j'\mathbf{y}_k = \frac{1}{n}\mathbf{x}_j'\mathbf{J}'\mathbf{J}\mathbf{x}_k, \tag{3.6}$$

which equals (3.4); the covariance of centered scores equals that of their raw scores.

Though the covariance is a theoretically important statistic, an inconvenient property of the covariance is that its value does not allow us to easily capture *how strong* the positive/negative *correlations* between variables are. For example, (3.2) shows that *sweet* and *spice* are negatively correlated with its sign (negative), but its absolute value of 12.65 does not show to what degree those variables are negatively correlated. This problem can easily be dealt with by modifying the covariance into a correlation coefficient, as described next.

3.3 Correlation Coefficients

A *correlation coefficient* between variables j and k is given by dividing the covariance (3.1) or (3.4) by the square roots of the variances of variables j and k (i.e., by the standard deviations of j and k). That is, the coefficient is defined using (2.23) as

$$r_{jk} = \frac{v_{jk}}{\sqrt{v_{jj}}\sqrt{v_{kk}}} = \frac{\frac{1}{n}\mathbf{x}_j'\mathbf{J}\mathbf{x}_k}{\sqrt{\frac{1}{n}\mathbf{x}_j'\mathbf{J}\mathbf{x}_j}\sqrt{\frac{1}{n}\mathbf{x}_k'\mathbf{J}\mathbf{x}_k}}. \tag{3.7}$$

Here, it should be noted that n^{-1} and the two square roots of n^{-1} in the right-hand side can be canceled out; (3.7) is rewritten as

$$r_{jk} = \frac{\mathbf{x}_j'\mathbf{J}\mathbf{x}_k}{\sqrt{\mathbf{x}_j'\mathbf{J}\mathbf{x}_j}\sqrt{\mathbf{x}_k'\mathbf{J}\mathbf{x}_k}} = \frac{\mathbf{x}_j'\mathbf{J}\mathbf{x}_k}{\sqrt{(\mathbf{J}\mathbf{x}_j)'\mathbf{J}\mathbf{x}_j}\sqrt{(\mathbf{J}\mathbf{x}_k)'\mathbf{J}\mathbf{x}_k}} = \frac{(\mathbf{J}\mathbf{x}_j)'\mathbf{J}\mathbf{x}_k}{\|\mathbf{J}\mathbf{x}_j\|\|\mathbf{J}\mathbf{x}_k\|} = \frac{\mathbf{y}_j'\mathbf{y}_k}{\|\mathbf{y}_j\|\|\mathbf{y}_k\|} \tag{3.8}$$

which shows that the correlation coefficient is defined as the *inner product* of the *centered score* vectors $\mathbf{y}_j = \mathbf{J}\mathbf{x}_j$ and $\mathbf{y}_k = \mathbf{J}\mathbf{x}_k$ divided by their *lengths*. The coefficient (3.7) or (3.8) is also called *Pearson's product-moment correlation coefficient*, named after Karl Pearson (1857–1936, British statistician), who derived the coefficient.

The correlation coefficient r_{jk} between variables j and k has the following properties:

[1] Its absolute value cannot exceed one with $-1 \leq r_{jk} \leq 1$.
[2] It takes a *positive value* when j and k have a *positive correlation*.
[3] It takes a *negative value* when j and k have a *negative correlation*.
[4] It approximates *zero* when j and k have *no correlation*.

Property [1], which is not possessed by covariances, allows us to easily capture the strength of the correlation, as illustrated in the following paragraph. Before that, we will show some numerical examples. The coefficient between *sweet* and *spice* can be obtained as

$$r_{12} = \frac{v_{12}}{\sqrt{v_{11}}\sqrt{v_{22}}} = \frac{-12.65}{\sqrt{25.65}\sqrt{19.01}} = -0.57, \tag{3.9}$$

using (3.2) and v_{jj} (Table 3.2) in (3.7). The value from (3.9) shows that *sweetness* is negatively correlated with *spice*. In a parallel manner, the coefficient between *spice* and *sales* is computed as

$$r_{23} = \frac{v_{23}}{\sqrt{v_{22}}\sqrt{v_{33}}} = \frac{0.15}{\sqrt{19.01}\sqrt{251.45}} = 0.00, \tag{3.10}$$

indicating that *spice* and *sales* have no correlation, while r_{13} is found to be 0.75, which shows that *sweetness* is positively correlated to *sales*.

The *upper limit* $r_{jk} = 1$, shown in Property [1], is attained for $\mathbf{y}_j = a\mathbf{y}_k$ with $a > 0$; its substitution in (3.8) leads to $r_{jk} = a\mathbf{y}_j'\mathbf{y}_j/(\|\mathbf{y}_j\| \times a\|\mathbf{y}_j\|) = 1$. On the other hand, the *lower limit* $r_{jk} = -1$ is attained when $\mathbf{y}_j = b\mathbf{y}_k$ with $b < 0$. The scatter plots of the variables with $r_{jk} = 1$ and $r_{jk} = -1$ are presented at the far left and right in Fig. 3.2, respectively. In each plot, all points are located on a straight line. Any r_{jk} takes a value between the two extremes ± 1, as shown in Fig. 3.2. There, we can find that the *strength of a positive or negative correlation* is captured by noting to what degree r_{jk} is *far from the* 0 *point* corresponding to no correlation and *close to*

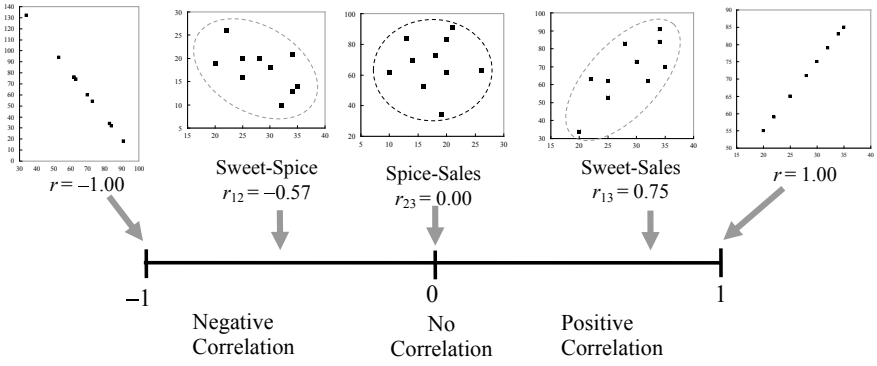


Fig. 3.2 Scatter plots and the corresponding correlation coefficients with their locations on the scale ranging from −1 to 1

Table 3.3 The correlation matrix for Table 3.1

Variable	Sweet	Spice	Sales
Sweet	1.00	−0.57	0.75
Spice	−0.57	1.00	0.00
Sales	0.75	0.00	1.00

±1. For example, $r_{13} = 0.75$ is close to 1, which indicates *sweetness* and *sales* are highly positively correlated. On the other hand, $r_{12} = -0.57$ is located a little to the left of the midpoint between 0 and −1, which indicates that *sweetness* and *spice* have a fairly negative correlation, though the correlation might not be said to be high.

The correlation coefficients among the variables in Table 3.1(A) are shown in Table 3.3. Such a p-variables × p-variables matrix is called a *correlation matrix*. As found in its diagonal elements, the correlation for the same variable is always one: $r_{jj} = \mathbf{y}_j'\mathbf{y}_j/(\|\mathbf{y}_j\|\|\mathbf{y}_j\|) = 1$.

3.4 Variable Vectors and Correlations

In this section, vector $\mathbf{y}_j = [y_{1j}, \ldots, y_{nj}]'$ is regarded as the line extending from $[0, \ldots, 0]'$ to $[y_{1j}, \ldots, y_{nj}]'$. As explained in Appendix A.1.1 with (A.1.3), the cosine of the angle between two vectors is equal to the division of their inner product by the product of their lengths. Thus, (3.8) shows that the correlation coefficient equals the *cosine of the angle* θ_{jk} between vectors $\mathbf{y}_j = \mathbf{J}\mathbf{x}_j$ and $\mathbf{y}_k = \mathbf{J}\mathbf{x}_k$ with

$$r_{jk} = \frac{\mathbf{y}_j'\mathbf{y}_k}{\|\mathbf{y}_j\|\|\mathbf{y}_k\|} = \cos\theta_{jk} \begin{cases} = 1 \text{ if } \theta_{jk} = 0^\circ \\ > 0 \text{ if } \theta_{jk} < 90^\circ \\ = 0 \text{ if } \theta_{jk} = 90^\circ \\ < 0 \text{ if } \theta_{jk} > 90^\circ \\ = -1 \text{ if } \theta_{jk} = 180^\circ \end{cases}. \tag{3.11}$$

Here, the right-hand side shows the relationships of θ_{jk} to $\cos\theta_{jk}$. In (3.11), we can find that the angles between the vectors of *positively correlated* variables are *less than 90°*, while the angles between *negatively correlated* ones are *more than 90°*, and the angle between *uncorrelated* variable vectors is 90°. When the angles between two vectors are 90°, they are said to be *orthogonal*. Using (3.11), we can find that $r_{12} = -0.57$, $r_{13} = 0.75$, and $r_{23} = 0.00$ lead to $\theta_{12} = 125^\circ$, $\theta_{13} = 41^\circ$, and $\theta_{23} = 90^\circ$, respectively, which allows us to visually illustrate the variable vectors as in Fig. 3.3.

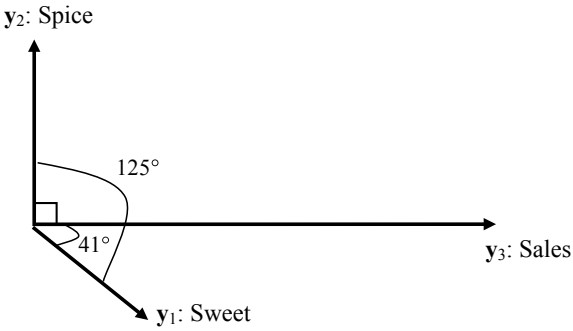


Fig. 3.3 Illustration of correlations with variable vectors in a three-dimensional space

3.5 Covariances and Correlations for Standard Scores

Let us recall (2.24), i.e., the definition of standard scores. By substituting (2.24) into $\mathbf{x}_j$ in (3.4), the covariance between standard score vectors $\mathbf{z}_j$ and $\mathbf{z}_k$ is expressed as

$$v_{jk}^{[z]} = \frac{1}{n}\left(\mathbf{J}\mathbf{z}_j\right)'\mathbf{J}\mathbf{z}_k = \frac{1}{n}\mathbf{z}_j'\mathbf{z}_k = \frac{1}{n}\left(\frac{1}{\sqrt{v_{jj}}}\mathbf{y}_j\right)'\frac{1}{\sqrt{v_{kk}}}\mathbf{y}_k = \frac{\frac{1}{n}\mathbf{y}_j'\mathbf{y}_k}{\sqrt{v_{ij}}\sqrt{v_{kk}}} = \frac{v_{jk}}{\sqrt{v_{ij}}\sqrt{v_{kk}}}. \tag{3.12}$$

Here, $\mathbf{z}_j = \mathbf{J}\mathbf{z}_j$ has been used. We can find (3.12) be equal to (3.7); the *covariance of standard scores* is equivalent to the *correlation coefficient of raw scores*.

The correlation coefficient between standard score vectors $\mathbf{z}_j$ and $\mathbf{z}_k$ is expressed as $r_{jk}^{[z]} = v_{jk}^{[z]} / \left(\sqrt{v_{jj}^{[z]}}\sqrt{v_{kk}^{[z]}}\right)$ by replacing v_{jk} in (3.7) by the covariance (3.12). Here, the variances $v_{jj}^{[z]}$ and $v_{kk}^{[z]}$ for standard scores are equal to one, as found in Table 2.3, thus, $r_{jk}^{[z]} = v_{jk}^{[z]}$ or (3.12); the *correlation coefficient of standard scores* equals the *correlation coefficients of the raw scores*.

Table 3.4 summarizes the properties of the covariances and correlation coefficients for standard and centered scores. The *correlation coefficients for centered/standard scores* and the *covariances for standard scores* equal the *correlation coefficients of their original raw scores*, and the *covariances for centered scores* equal *those for the raw scores*. We can regard the *correlation coefficient* as a *standardized* version of the *covariance*, as the covariances for standard scores equal the correlation coefficients of the raw scores.

Table 3.4 Covariances and correlations of centered and standard scores

	Covariance	Correlation coefficient
Centered scores	Covariance for raw scores	Correlation coefficient for raw scores
Standard scores	Correlation coefficient for raw scores	

3.6 Matrix Expressions of Covariances and Correlations

Using (3.6) with (2.27), the p-variables $\times$ p-variables *covariance matrix* $\mathbf{V} = (v_{jk})$ for the data matrix $\mathbf{X}$ (2.1) can be expressed as

$$\begin{aligned}\mathbf{V} &= \frac{1}{n}\begin{bmatrix} \mathbf{x}_1'\mathbf{J}'\mathbf{J}\mathbf{x}_1 & \cdots & \mathbf{x}_1'\mathbf{J}'\mathbf{J}\mathbf{x}_k & \cdots & \mathbf{x}_1'\mathbf{J}'\mathbf{J}\mathbf{x}_p \\ \vdots & \cdots & \cdots & \cdots & \vdots \\ \mathbf{x}_j'\mathbf{J}'\mathbf{J}\mathbf{x}_1 & \cdots & \mathbf{x}_j'\mathbf{J}'\mathbf{J}\mathbf{x}_k & \cdots & \mathbf{x}_j'\mathbf{J}'\mathbf{J}\mathbf{x}_p \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \mathbf{x}_p'\mathbf{J}'\mathbf{J}\mathbf{x}_1 & \cdots & \mathbf{x}_p'\mathbf{J}'\mathbf{J}\mathbf{x}_k & \cdots & \mathbf{x}_p'\mathbf{J}'\mathbf{J}\mathbf{x}_p \end{bmatrix} \\ &= \frac{1}{n}\begin{bmatrix} \mathbf{x}_1'\mathbf{J}' \\ \vdots \\ \mathbf{x}_j'\mathbf{J}' \\ \vdots \\ \mathbf{x}_p'\mathbf{J}' \end{bmatrix}\begin{bmatrix}\mathbf{J}\mathbf{x}_1 \cdots \mathbf{J}\mathbf{x}_k \cdots \mathbf{J}\mathbf{x}_p\end{bmatrix} \\ &= \frac{1}{n}\begin{bmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_j' \\ \vdots \\ \mathbf{x}_p' \end{bmatrix}\mathbf{J}'\mathbf{J}\begin{bmatrix}\mathbf{x}_1 \cdots \mathbf{x}_k \cdots \mathbf{x}_p\end{bmatrix} = \frac{1}{n}\mathbf{X}'\mathbf{J}'\mathbf{J}\mathbf{X} = \frac{1}{n}\mathbf{X}'\mathbf{J}\mathbf{X},\end{aligned} \tag{3.13}$$

in which (2.11) and (2.12) have been used. We can use (2.29) to rewrite the covariance matrix (3.13) simply as

$$\mathbf{V} = \frac{1}{n}\mathbf{Y}'\mathbf{Y}. \tag{3.14}$$

Let $\mathbf{R} = (r_{jk})$ denote the p-variables $\times$ p-variables *correlation matrix* $\mathbf{R} = (r_{jk})$ for $\mathbf{X}$. Since the covariance for the standard scores is equal to the correlation coefficient for the raw scores as shown in Table 3.4, the (j, k) element $\mathbf{R} = (r_{jk})$ is expressed as (3.12):

$$\mathbf{R} = \frac{1}{n}\begin{bmatrix} \mathbf{z}_1'\mathbf{J}\mathbf{z}_1 & \cdots & \mathbf{z}_1'\mathbf{J}\mathbf{z}_k & \cdots & \mathbf{z}_1'\mathbf{J}\mathbf{z}_p \\ \vdots & \cdots & \cdots & \cdots & \vdots \\ \mathbf{z}_j'\mathbf{J}\mathbf{z}_1 & \cdots & \mathbf{z}_j'\mathbf{J}\mathbf{z}_k & \cdots & \mathbf{z}_j'\mathbf{J}\mathbf{z}_p \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \mathbf{z}_p'\mathbf{J}\mathbf{z}_1 & \cdots & \mathbf{z}_p'\mathbf{J}\mathbf{z}_k & \cdots & \mathbf{z}_p'\mathbf{J}\mathbf{z}_p \end{bmatrix} = \frac{1}{n}\begin{bmatrix} \mathbf{z}_1'\mathbf{z}_1 & \cdots & \mathbf{z}_1'\mathbf{z}_k & \cdots & \mathbf{z}_1'\mathbf{z}_p \\ \vdots & \cdots & \cdots & \cdots & \vdots \\ \mathbf{z}_j'\mathbf{z}_1 & \cdots & \mathbf{z}_j'\mathbf{z}_k & \cdots & \mathbf{z}_j'\mathbf{z}_p \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \mathbf{z}_p'\mathbf{z}_1 & \cdots & \mathbf{z}_p'\mathbf{z}_k & \cdots & \mathbf{z}_p'\mathbf{z}_p \end{bmatrix} = \frac{1}{n}\mathbf{Z}'\mathbf{Z}, \tag{3.15}$$

where $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_p]$ is the n-individuals $\times$ p-variables matrix of *standard scores* matrix. Using (2.32) in (3.15), $\mathbf{R}$ is also expressed as

$$\mathbf{R} = \frac{1}{n}\mathbf{D}^{-1}\mathbf{X}'\mathbf{J}'\mathbf{J}\mathbf{X}\mathbf{D}^{-1} = \frac{1}{n}\mathbf{D}^{-1}\mathbf{X}'\mathbf{J}\mathbf{X}\mathbf{D}^{-1}, \tag{3.16}$$

with $\mathbf{D} = \begin{bmatrix} \sqrt{v_{11}} & & \\ & \ddots & \\ & & \sqrt{v_{pp}} \end{bmatrix}$, as defined in Sect. 2.8. Further, if we compare (3.16) with (3.13), we have

$$\mathbf{R} = \mathbf{D}^{-1}\mathbf{V}\mathbf{D}^{-1}. \tag{3.17}$$

3.7 Unbiased Covariances

For covariances (and variances), a definition exists that is different from (3.4). In this definition, $\mathbf{x}_j'\mathbf{J}\mathbf{x}_k$ is divided by $n - 1$ in place of n; the covariance matrix for $\mathbf{X}$ may be defined as

$$\mathbf{V}^{\#} = \frac{1}{n-1}\mathbf{X}'\mathbf{J}\mathbf{X} = \frac{n}{n-1}\mathbf{V}. \tag{3.18}$$

This is called an *unbiased covariance* matrix. Its off-diagonal and diagonal elements are called *unbiased covariances* and *unbiased variances*, respectively, for distinguishing (3.18) from (3.13); one may use either equation. In this book, we refrain from explaining why two types of definition exist, and (3.13) is used throughout. For example, see Hogg, McKean, and Craig (2005) for the statistical theory about the adjective "*unbiased*" and its antonym "*biased*".

Though the covariance is defined in the two above manners, the correlation coefficient is defined uniquely, i.e., in a single way, as follows: If we use covariance (3.18), the correlation in (3.7) is rewritten as

$$r_{jk} = \frac{v_{jk}}{\sqrt{v_{ij}}\sqrt{v_{kk}}} = \frac{\frac{1}{n-1}\mathbf{x}_j'\mathbf{J}\mathbf{x}_k}{\sqrt{\frac{1}{n-1}\mathbf{x}_j'\mathbf{J}\mathbf{x}_j}\sqrt{\frac{1}{n-1}\mathbf{x}_k'\mathbf{J}\mathbf{x}_k}} = \frac{\mathbf{x}_j'\mathbf{J}\mathbf{x}_k}{\sqrt{\mathbf{x}_j'\mathbf{J}\mathbf{x}_j}\sqrt{\mathbf{x}_k'\mathbf{J}\mathbf{x}_k}}. \tag{3.19}$$

Here, $n - 1$ in the numerator and denominator are canceled out so that (3.19) becomes equivalent to (3.8), i.e., (3.7).

3.8 Centered Matrices

When a data matrix $\mathbf{X}$ contains centered scores, i.e., is already centered, with

$$\mathbf{1}_n'\mathbf{X} = \mathbf{0}_p', \text{ or, equivalently, } \mathbf{X} = \mathbf{J}\mathbf{X}, \tag{3.20}$$

$\mathbf{X}$ is said to be *centered*, where $\mathbf{J} = \mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n'$ is the centering matrix (2.10). The equivalence of the two equations in (3.20) will now be proved.

Note 3.1. Two Expressions of Zero Average
The sum and average of the elements in each column of an $n \times m$ vector $\mathbf{F}$ being zero are equivalent to the pre-multiplication of $\mathbf{F}$ by the centering matrix being $\mathbf{F}$:

$$\mathbf{1}_n'\mathbf{F} = \frac{1}{n}\mathbf{1}_n'\mathbf{F} = \mathbf{0}_m' \quad \text{is equivalent to} \quad \mathbf{J}\mathbf{F} = \mathbf{F}. \tag{3.21}$$

This is proved by showing both [1] $\mathbf{J}\mathbf{F} = \mathbf{F} \Rightarrow \mathbf{1}_n'\mathbf{F} = \mathbf{0}_m$ and [2] $\mathbf{1}_n'\mathbf{F} = \mathbf{0}_m' \Rightarrow \mathbf{J}\mathbf{F} = \mathbf{F}$. [1] is easily found by using (2.13) in $\mathbf{1}_n'\mathbf{F} = \mathbf{1}_n'\mathbf{J}\mathbf{F}$, while [2] follows from pre-multiplying of both sides of $\mathbf{1}_n'\mathbf{F} = \mathbf{0}_m'$ by $-n^{-1}\mathbf{1}_n$ yields $-n^{-1}\mathbf{1}_n\mathbf{1}_n'\mathbf{F} = {}_n\mathbf{O}_m$, to whose both sides we can add $\mathbf{F}$ so as to have $\mathbf{F} - n^{-1}\mathbf{1}_n\mathbf{1}_n'\mathbf{F} = \mathbf{F}$, i.e., $(\mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n')\mathbf{F} = \mathbf{F}$.

When $\mathbf{X}$ is centered, (3.13) and (3.16) are expressed as

$$\mathbf{V} = \frac{1}{n}\mathbf{X}'\mathbf{X}, \tag{3.22}$$

$$\mathbf{R} = \frac{1}{n}\mathbf{D}^{-1}\mathbf{X}'\mathbf{X}\mathbf{D}^{-1}, \tag{3.23}$$

respectively, simply without using $\mathbf{J}$.

The covariance matrix, which has been treated so far, contains covariances among the variables in a single data matrix $\mathbf{X}$. Now, let us consider the $p \times m$ matrix containing covariances between the variables in $\mathbf{X}$ ($n \times p$) and those corresponding to the columns of $\mathbf{F}$ ($n \times m$). The covariance matrix is expressed as

$$\mathbf{V}_{\mathrm{XF}} = \frac{1}{n}\mathbf{X}'\mathbf{J}\mathbf{F}. \tag{3.24}$$

If both $\mathbf{X}$ and $\mathbf{F}$ are centered with $\mathbf{X} = \mathbf{J}\mathbf{X}$ and $\mathbf{F} = \mathbf{J}\mathbf{F}$, (3.24) is simplified as

$$\mathbf{V}_{\mathrm{XF}} = \frac{1}{n}\mathbf{X}'\mathbf{F}. \tag{3.25}$$

Further, when both $\mathbf{X}$ and $\mathbf{F}$ contain standard scores, (3.25) also expresses a correlation matrix.

3.9 Ranks of Matrices: Intuitive Introduction

For every matrix, its *rank* is given as an integer. It is an important number that stands for a property of the matrix and is used in the following chapters. In this section, we introduce rank so that it can be *intuitively* captured using the four 5×3 matrices in Table 3.5.

First, note the matrix in Table 3.5(A). The values seem to be different among the three columns. Indeed, no relationships exist between $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$. That is, those three columns are regarded as, respectively, conveying three different kinds of information. Such a matrix is said to be the one whose *rank* is *three*. Next, note (B), whose third column is the same as the first one; though the matrix has three columns, it conveys to us only two kinds of information. The *rank* of this matrix is said to be *two*.

The third column in Table 3.5(C) is different from the first one, but multiplication of the latter by −3 gives the third column. Its elements can be considered as expressing the same information as those in the first column, except that the signs of the values are reversed and their scales differ. The *rank* of this matrix is also said to be *two*, not three.

Table 3.5 Four matrices for illustrating their ranks

	(A)			(B)			(C)			(D)		
	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
1	2	3	−2	2	3	2	2	3	−6	2	9	−2
2	4	5	9	4	5	4	4	5	−12	4	−21	9
3	−1	7	3	−1	7	−1	−1	7	3	−1	−10.5	3
4	−5	0	3	−5	0	−5	−5	0	15	−5	−16.5	3
5	7	5	2	7	5	7	7	5	−21	7	4.5	2

Finally, let us note Table 3.5(D). Though the three columns seem to mutually differ, $\mathbf{x}_2 = 1.5\mathbf{x}_1 - 3\mathbf{x}_3$. The *rank* of this matrix is also said to be *two*, in that the information conveyed by the second column can be found by knowing that found in the other two.

In the next section, the *rank* of a matrix is precisely defined.

3.10 Ranks of Matrices: Mathematical Definition

A sum of the vectors $\mathbf{h}_1, \ldots, \mathbf{h}_p$ multiplied by scalars

$$b_1\mathbf{h}_1 + \cdots + b_p\mathbf{h}_p = \mathbf{Hb} \tag{3.26}$$

is called the *linear combination* of $\mathbf{h}_1, \ldots, \mathbf{h}_p$. Here, $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_p]$ is an $n \times p$ matrix, and $\mathbf{b} = [b_1, \ldots, b_p]'$ is a $p \times 1$ vector. Before defining the rank of $\mathbf{H}$, we introduce the following two notions:

Note 3.2. Linear Independence
The set of vectors $\mathbf{h}_1, \ldots, \mathbf{h}_p$ is said to be *linearly independent*, if

$$b_1\mathbf{h}_1 + \cdots + b_p\mathbf{h}_p = \mathbf{Hb} = \mathbf{0}_p \text{ implies } \mathbf{b} = \mathbf{0}_p. \tag{3.27}$$

The inverse of the above is defined as follows:

Note 3.3. Linear Dependence
The set of vectors $\mathbf{h}_1, \ldots, \mathbf{h}_p$ is said to be *linearly dependent*, if $\mathbf{Hb} = \mathbf{0}_p$ does not imply $\mathbf{b} = \mathbf{0}_p$, that is, if

$$\begin{aligned} &b_1\mathbf{h}_1 + \cdots + b_p\mathbf{h}_p = \mathbf{Hb} = \mathbf{0}_p \text{ holds,} \\ &\quad \text{with at least } b_J (1 \le J \le p) \text{ not being zero.} \end{aligned} \tag{3.28}$$

This implies that $b_J\mathbf{h}_J = -\sum_{j \neq J} b_j\mathbf{h}_j$ and we can divide both sides by b_J to have

$$\mathbf{h}_J = -\sum_{j \neq J} \frac{b_j}{b_J}\mathbf{h}_j: \tag{3.29}$$

the vector $\mathbf{h}_J$ is a *linear combination* of the other vectors with coefficients $-b_j/b_J$. Here, $\sum_{j \neq J} a_j$ denotes the sum of a_j over j excluding a_J. When $j = 1$, 2, 3, $\sum_{j \neq 2} a_j = a_1 + a_3$, for example.

The *rank* of $\mathbf{H}$, which we denote as rank($\mathbf{H}$), is defined as follows:

Note 3.4. Rank of a Matrix

$$\text{rank}(\mathbf{H}) = \text{the maximum number of linearly independent columns in } \mathbf{H}. \tag{3.30}$$

For illustrating the definition of rank, we present the following three examples:

[1] Let $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3] = \begin{bmatrix} 1 & 1 & 9 \\ 2 & 2 & 6 \\ 1 & 3 & 4 \\ 2 & 4 & 7 \end{bmatrix}$. Then, rank($\mathbf{P}$) = 3,

since $b_1\mathbf{p}_1 + b_2\mathbf{p}_2 + b_3\mathbf{p}_3 = \mathbf{0}_3$ implies $b_1 = b_2 = b_3 = 0$; we cannot find nonzero b_J with $\mathbf{p}_J = -\Sigma_{j \neq J} b_j / b_J \, \mathbf{p}_j$.

[2] Let $\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, \mathbf{q}_3] = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 1 & 2 & 3 \\ 2 & 4 & 6 \end{bmatrix}$. Then, rank($\mathbf{Q}$) = 1,

since $\mathbf{q}_2 = 2\mathbf{q}_1$ and $\mathbf{q}_3 = 3\mathbf{q}_1$; the linearly independent vector sets are $\{\mathbf{q}_1\}$, $\{\mathbf{q}_2\}$, and $\{\mathbf{q}_3\}$, each of which consists of a single vector.

[3] Let $\mathbf{R} = [\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3] = \begin{bmatrix} 1 & 1 & 3 \\ 2 & 2 & 6 \\ 1 & 3 & 5 \\ 2 & 4 & 8 \end{bmatrix}$. Then, rank($\mathbf{R}$) = 2,

since $\mathbf{r}_3 = 2\mathbf{r}_1 + \mathbf{r}_2$, thus, rank($\mathbf{R}$) < 3, but the set of $\mathbf{r}_1$ and $\mathbf{r}_2$ is linearly independent.

It is difficult to find the rank of a matrix by glancing at it, but we can easily find the rank through the extended version of the *singular value decomposition* introduced in Appendix A3.1.

Here, we introduce properties of the rank without proof. The rank of an $n \times p$ matrix $\mathbf{A}$ satisfies

$$\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{A}'), \tag{3.31}$$

which implies that the "columns" in (3.30) may be replaced by "rows". Further, (3.31) implies

$$\text{rank}(\mathbf{A}) \leq \min(n, p). \tag{3.32}$$

The following properties are also used in the remaining chapters:

$$\text{rank}(\mathbf{BA}) \leq \min(\text{rank}(\mathbf{A}), \text{rank}(\mathbf{B})) \tag{3.33}$$

$$\mathbf{A}'\mathbf{A} = \mathbf{I}_p \text{ implies } \text{rank}(\mathbf{A}) = p. \tag{3.34}$$

3.11 Bibliographical Notes

Carroll, Green, and Chaturvedi (1997, Chap. 3), Rencher and Christensen (2012, Chap. 3), and Reyment and Jöreskog (2002, Chap. 2) are among the literature in which matrix expressions of inter-variable statistics are intelligibly treated. The rank of a matrix is detailed in those books introduced in Sect. 1.9.

Exercises

3.1. Prove the *Cauchy–Schwarz inequality*

$$(\mathbf{a}'\mathbf{b})^2 \leq \|\mathbf{a}\|^2 \|\mathbf{b}\|^2 \tag{3.35}$$

by defining matrix $\mathbf{C} = \mathbf{ab}' - \mathbf{ba}'$ and using $\|\mathbf{C}\|^2 \geq 0$.

3.2. Use (3.35) to show

$$\mathbf{x}_1'\mathbf{J}\mathbf{x}_2 \leq \|\mathbf{J}\mathbf{x}_1\| \|\mathbf{J}\mathbf{x}_2\|, \tag{3.36}$$

with $\mathbf{x}_1$ and $\mathbf{x}_2$ $n \times 1$ vectors and $\mathbf{J} = \mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n'$ the centering matrix.

3.3. Use (3.36) to show that the correlation coefficient takes a value within the range from -1 to 1.

3.4. Let $\mathbf{x} = [x_1, \ldots, x_n]'$ and $\mathbf{y} = [y_1, \ldots, y_n]'$, with $v = n^{-1}\mathbf{x}'\mathbf{J}\mathbf{y}$ the covariance between $\mathbf{x}$ and $\mathbf{y}$. Show that the covariance between $a\mathbf{x} + c\mathbf{1}_n = \begin{bmatrix} ax_1 + c \\ \vdots \\ ax_n + c \end{bmatrix}$ and $b\mathbf{y} + d\mathbf{1}_n = \begin{bmatrix} by_1 + d \\ \vdots \\ by_n + d \end{bmatrix}$ is given by $abv = n^{-1}ab\mathbf{x}'\mathbf{J}\mathbf{y}$.

3.5. Let r denote the correlation coefficient between vectors $\mathbf{x} = [x_1, \ldots, x_n]'$ and $\mathbf{y} = [y_1, \ldots, y_n]'$. Show that the correlation coefficient between $a\mathbf{x} + c\mathbf{1}_n$ and $b\mathbf{y} + d\mathbf{1}_n$ is also r for $ab > 0$, but is $-r$ for $ab < 0$, with the coefficient not defined for $ab = 0$.

3.6. Let $\mathbf{X}$ and $\mathbf{Y}$ be the matrices containing n rows, with $\mathbf{V}_{XY} = n^{-1}\mathbf{X}'\mathbf{J}\mathbf{Y}$ the covariance matrix between the columns of $\mathbf{X}$ and those of $\mathbf{Y}$. Show that $\mathbf{A}'\mathbf{V}_{XY}\mathbf{B}$ gives the covariance matrix between the columns of $\mathbf{XA}$ and those of $\mathbf{YB}$.

3.7. Let $\mathbf{X}$ and $\mathbf{Y}$ be the matrices containing n rows, $\mathbf{D}_X$ be the diagonal matrix whose jth diagonal element is the standard deviation of the elements in the jth column of $\mathbf{X}$, and $\mathbf{D}_Y$ be defined for $\mathbf{Y}$ in a parallel manner. Show that $\mathbf{R}_{XY} = n^{-1}\mathbf{D}_X^{-1}\mathbf{X}'\mathbf{J}\mathbf{Y}\mathbf{D}_Y^{-1}$ gives the correlation matrix between the columns of $\mathbf{X}$ and those of $\mathbf{Y}$.

3.8. Consider the matrices defined in Exercise 3.7. Show that $\mathbf{R}_{XY} = n^{-1}\mathbf{D}_X^{-1}\mathbf{X}'\mathbf{Y}\mathbf{D}_Y^{-1}$ gives the correlation matrix between the columns of $\mathbf{X}$ and $\mathbf{Y}$, when they are centered.

3.9. Let $\mathbf{A} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 2 & 2 \\ 3 & 0 & 3 \\ 0 & 4 & 4 \end{bmatrix}$. Show $\text{rank}(\mathbf{A}) = 2$ by noting the columns of $\mathbf{A}$ and $\text{rank}(\mathbf{A}') = 2$ by noting the rows of $\mathbf{A}$.

3.10. Let $\mathbf{G}$ be $p \times q$ and $\mathbf{H}$ be $q \times r$, with $q \geq p \geq r$. Show $\text{rank}(\mathbf{GH}) \leq r$.

3.11. Let $\mathbf{F}$ be $n \times m$ and $\mathbf{A}$ be $p \times m$, with $m \leq \min(n, p)$. Show $\text{rank}(\mathbf{FA}') \leq m$.

3.12. Show $\text{rank}(\mathbf{I}_n) = n$, with $\mathbf{I}_n$ the $n \times n$ identity matrix.

3.13. Show that $\text{rank}(\mathbf{JX}) \leq \min(n - 1, p)$, with $\mathbf{X}$ an $n \times p$ matrix and $\mathbf{J} = \mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n'$ the centering matrix.

Part II
Least Squares Procedures

Regression, principal component, and cluster analyses are introduced as least squares procedures. Here, principal component analysis is treated in two chapters, as it can be described in various ways. The three analysis procedures are formulated as minimizing least squares functions, though other formulations are also possible.

Chapter 4
Regression Analysis

In the previous two chapters, we expressed elementary statistics in matrix form as preparation for introducing multivariate analysis procedures. The introduction to those procedures begins in this chapter. Here, we treat *regression analysis*, whose purpose is to *predict* or *explain* a variable from a set of other variables. The origin of regression analysis is found in the studies of Francis Golton (1822–1911, British scientist) on heredity stature in the mid-1880s. The history of developments in regression analysis is well summarized in Izenman (2008, pp. 107–108).

4.1 Prediction of a Dependent Variable by Explanatory Variables

In Table 4.1, we show a 50-products × 4-variables (*quality*, *price*, *appearance*, and *sales*) data matrix. Let us consider predicting or explaining the *sales* of products by their *quality*, *price*, and *appearance*, with the formula

$$sales = b_1 \times quality + b_2 \times price + b_3 \times appearance + c + error. \tag{4.1}$$

Here, the term "*error*" must be attached to the right-hand side, because a perfectly exact prediction of *sales* is impossible.

Let us use x_{i1}, x_{i2}, x_{i3}, and y_i for the *quality*, *price*, *appearance*, and *sales* of the i-th product in Table 4.1, respectively. Then, (4.1) is rewritten as

$$y_i = b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + c + e_i, \tag{4.2}$$

with e_i the error value for product i. Since (4.2) is supposed for all products, $i = 1, \ldots, 50$ in Table 4.1. Thus, we have

K. Adachi, *Matrix-Based Introduction to Multivariate Data Analysis*,
https://doi.org/10.1007/978-981-15-4103-2_4

Table 4.1 Data matrix for the *quality*, *price*, *appearance*, and *sales* of products, which is an artificial example found in Adachi (2006)

Product	Quality	Price	Appearance	Sales
1	10	1800	2.6	48
2	5	1550	4.2	104
3	5	1250	3.0	122
4	5	1150	1.0	104
5	6	1700	7.0	125
6	6	1550	4.0	105
7	5	1200	3.6	135
8	3	1000	1.8	128
9	3	1300	5.8	145
10	5	1300	3.0	124
11	6	1550	5.8	99
12	9	1800	4.2	102
13	8	1400	4.4	146
14	6	1300	3.0	138
15	5	1400	3.8	122
16	10	1950	3.0	13
17	4	1550	5.2	103
18	2	1300	4.0	86
19	7	1800	6.8	109
20	4	1300	3.4	103
21	6	1350	4.0	113
22	9	1450	1.8	100
23	5	1300	4.2	111
24	6	1450	4.0	138
25	8	1750	4.0	101
26	4	1500	4.2	126
27	3	1700	4.6	29
28	6	1500	2.2	73
29	4	1250	3.4	129
30	9	1650	3.2	77
31	5	1500	3.4	84
32	4	1350	3.8	103
33	4	1350	3.8	112
34	3	1550	4.6	77
35	3	1200	3.6	135
36	1	1450	6.0	112
37	4	1600	4.8	106
38	5	1600	3.8	99
39	1	1100	4.2	143
40	6	1600	3.8	54
41	4	1450	6.6	139

(continued)

Table 4.1 (continued)

Product	Quality	Price	Appearance	Sales
42	2	1300	1.6	90
43	4	1200	5.2	203
44	3	1150	2.4	96
45	7	1350	3.2	125
46	7	1200	1.2	107
47	5	1550	5.0	130
48	5	1600	4.2	72
49	7	1400	3.8	137
50	7	1600	5.4	106

$$\begin{bmatrix} 48 \\ 104 \\ \vdots \\ 106 \end{bmatrix} = b_1 \begin{bmatrix} 10 \\ 5 \\ \vdots \\ 7 \end{bmatrix} + b_2 \begin{bmatrix} 1800 \\ 1550 \\ \vdots \\ 1600 \end{bmatrix} + b_3 \begin{bmatrix} 2.6 \\ 4.2 \\ \vdots \\ 5.4 \end{bmatrix} + \begin{bmatrix} c \\ c \\ \vdots \\ c \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_{50} \end{bmatrix}. \tag{4.3}$$

Further, it is rewritten as

$$\begin{bmatrix} 48 \\ 104 \\ \vdots \\ 106 \end{bmatrix} = \begin{bmatrix} 10 & 1800 & 2.6 \\ 5 & 1550 & 4.2 \\ \vdots & \vdots & \vdots \\ 7 & 1600 & 5.4 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} + c \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_{50} \end{bmatrix} \tag{4.4}$$

by summarizing the vectors for *quality*, *price*, and *appearance* into a matrix. Expressing this matrix as $\mathbf{X}$ and using $\mathbf{y}$ for the *sales* vector, (4.4) can expressed as

$$\mathbf{y} = \mathbf{X}\mathbf{b} + c\mathbf{1}_n + \mathbf{e}, \tag{4.5}$$

with $\mathbf{b} = [b_1, b_2, b_3]'$, $\mathbf{e} = [e_1, \ldots, e_{50}]'$, and $\mathbf{1}_n$ the $n \times 1$ vector of ones ($n = 50$ in this example). *Regression analysis* refers to a procedure for obtaining the *optimal* values of c and the elements of $\mathbf{b}$ from data $\mathbf{y}$ and $\mathbf{X}$. Though $\mathbf{y}$ was used for a centered score vector in the last two chapters, it is not so in this chapter.

Hereafter, we generally describe $\mathbf{X}$ as an n-individuals $\times$ p-variables matrix, which implies that $\mathbf{y}$ and $\mathbf{e}$ are $n \times 1$ vectors and $\mathbf{b}$ is a $p \times 1$ vector. The model (4.5) for regression analysis is thus expressed as

$$\underset{\mathbf{y}}{\begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix}} = \underset{\mathbf{X}}{\begin{bmatrix} x_{11} & \cdots & x_{1p} \\ & \vdots & \\ x_{i1} & \cdots & x_{ip} \\ & \vdots & \\ x_{n1} & \cdots & x_{np} \end{bmatrix}} \underset{\mathbf{b}}{\begin{bmatrix} b_1 \\ \vdots \\ b_p \end{bmatrix}} + c \underset{\mathbf{1}_n}{\begin{bmatrix} 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{bmatrix}} + \underset{\mathbf{e}}{\begin{bmatrix} e_1 \\ \vdots \\ e_i \\ \vdots \\ e_n \end{bmatrix}} \tag{4.6}$$

The term *model* refers to a formula that expresses the idea underlying an analysis procedure.

In this paragraph, we explain the terms used in regression analysis. The predicted or explained vector, i.e., $\mathbf{y}$, is called a *dependent variable*, while the columns of $\mathbf{X}$ are called *explanatory variables*. On the other hand, the elements of $\mathbf{b}$ are referred to as *regression coefficients*, and c is called an *intercept*. In particular, regression analysis with $p = 1$, i.e., a single exploratory variable, is called *simple regression* analysis, while the procedure with $p \geq 2$ is called *multiple regression* analysis; (4.6) is its model.

The terms generally for describing analysis procedures are summarized next:

Note 4.1. Data Versus Parameters
In contrast to $\mathbf{y}$ and $\mathbf{X}$ given as *data* beforehand, the values of $\mathbf{b}$ and c are unknown before regression analysis is performed. Such entities as $\mathbf{b}$ and c, whose values are estimated by an analysis procedure, are generally called *parameters*. When one sees symbols in equations, it is very important to note whether the symbols express *data* or *parameters*.

Besides the data and parameters, *errors* ($\mathbf{e}$) are included in (4.6). So as to minimize their amount, the parameter values are estimated, as described in the next section.

4.2 Least Squares Method

Parameters $\mathbf{b}$ and c can be estimated using a *least squares method*. It generally refers to a class of the procedures for obtaining parameter values that *minimize* the *sum of squared errors*. This sum for (4.5) is expressed as

$$\|\mathbf{e}\|^2 = e_1^2 + \cdots + e_n^2 = \|\mathbf{y} - \mathbf{X}\mathbf{b} - c\mathbf{1}_n\|^2, \tag{4.7}$$

since (4.5) is rewritten as $\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b} - c\mathbf{1}_n$. Thus, regression analysis is formulated as

$$\text{minimizing } f(\mathbf{b}, c) = \|\mathbf{e}\|^2 = \|\mathbf{y} - \mathbf{X}\mathbf{b} - c\mathbf{1}_n\|^2 \text{ over } \mathbf{b} \text{ and } c, \tag{4.8}$$

which can be restated as obtaining the optimal $\mathbf{b}$ and c (i.e., their solutions) that minimize (4.7). Let us express the *solutions* of $\mathbf{b}$ and c for (4.8) as $\hat{\mathbf{b}}$ and $\hat{c}$, respectively, which are given as described in the following paragraphs.

It is known that $\hat{c}$ must satisfy

$$\hat{c} = \frac{1}{n}\mathbf{1}_n'\mathbf{y} - \frac{1}{n}\mathbf{1}_n'\mathbf{X}\mathbf{b}. \tag{4.9}$$

This result can be derived as follows: We can define $\mathbf{h} = \mathbf{y} - \mathbf{X}\mathbf{b}$ to rewrite (4.7) as $\|\mathbf{h} - c\mathbf{1}_n\|^2$, which is minimized for $c = n^{-1}\mathbf{1}_n'\mathbf{h}$, as shown with (A.2.2) in Appendix A.2.1. The use of $\mathbf{h} = \mathbf{y} - \mathbf{X}\mathbf{b}$ in $\hat{c} = n^{-1}\mathbf{1}_n'\mathbf{h}$ leads to (4.9).

Substituting (4.9) into c in $\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b} - \mathbf{1}_n \times c$, which follows from (4.5), we have

$$\begin{aligned}\mathbf{e} &= \mathbf{y} - \mathbf{X}\mathbf{b} - \left(\frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\mathbf{y} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\mathbf{X}\mathbf{b}\right) \\ &= \left(\mathbf{y} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\mathbf{y}\right) - \left(\mathbf{X}\mathbf{b} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\mathbf{X}\mathbf{b}\right) = \mathbf{J}\mathbf{y} - \mathbf{J}\mathbf{X}\mathbf{b},\end{aligned} \tag{4.10}$$

with $\mathbf{J} = \mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n'$ the *centering matrix* defined in (2.10). Thus, (4.7) is rewritten as

$$\|\mathbf{e}\|^2 = \|\mathbf{J}\mathbf{y} - \mathbf{J}\mathbf{X}\mathbf{b}\|^2. \tag{4.11}$$

This is minimized when $\mathbf{b}$ is

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{J}\mathbf{X})^{-1}\mathbf{X}'\mathbf{J}\mathbf{y}, \tag{4.12}$$

as shown with (A.2.16) in Appendix A.2.2. Here, $(\mathbf{X}'\mathbf{J}\mathbf{X})^{-1}$ is the *inverse matrix* of $\mathbf{X}'\mathbf{J}\mathbf{X}$, which is introduced below.

Note 4.2. Inverse Matrix
A $p \times p$ square matrix $\mathbf{H}$ is said to be *nonsingular* if

$$\text{rank}(\mathbf{H}) = p; \tag{4.13}$$

otherwise, $\mathbf{H}$ is said to be singular. If $\mathbf{H}$ is nonsingular, the $p \times p$ matrix $\mathbf{H}^{-1}$ exists that satisfies

$$\mathbf{H}^{-1}\mathbf{H} = \mathbf{H}\mathbf{H}^{-1} = \mathbf{I}_p. \tag{4.14}$$

The matrix $\mathbf{H}^{-1}$ is called the *inverse matrix* of $\mathbf{H}$. For example,

$$\mathbf{H} = \begin{bmatrix} 3 & -1 & 2 \\ -4 & 6 & -3 \\ 1 & 0 & 5 \end{bmatrix} \text{ is nonsingular, and } \mathbf{H}^{-1} = \begin{bmatrix} 0.49 & 0.08 & -0.15 \\ 0.28 & 0.21 & 0.02 \\ -0.10 & -0.02 & 0.23 \end{bmatrix}.$$

We can find in (4.14) that $\mathbf{H}$ is the inverse matrix of $\mathbf{H}^{-1}$ with $\mathbf{H} = (\mathbf{H}^{-1})^{-1}$. The inverse matrix $\mathbf{H}^{-1}$ does not exist if $\mathbf{H}$ is singular.

Two basic properties of inverse matrices are

$$(\mathbf{H}')^{-1} = \left(\mathbf{H}^{-1}\right)', \tag{4.15}$$

$$(\mathbf{G}\mathbf{H})^{-1} = \mathbf{H}^{-1}\mathbf{G}^{-1}, \tag{4.16}$$

which includes $(s\mathbf{H})^{-1} = s^{-1}\mathbf{H}^{-1}$ as a special case with $s \neq 0$ a scalar. The inverse matrix of a symmetric matrix $\mathbf{S}$ is also symmetric:

$$\mathbf{S}^{-1} = \mathbf{S}^{-1'}. \tag{4.17}$$

As found in the note, we suppose that $\mathbf{X}'\mathbf{J}\mathbf{X}$ is nonsingular in (4.12). Actually, the data set in Table 4.1 gives such a $\mathbf{X}'\mathbf{J}\mathbf{X}$.

Thus, the solution of regression analysis is given by obtaining (4.12) and substituting $\hat{\mathbf{b}}$ into $\mathbf{b}$ in (4.9). The solution (4.12) for $\mathbf{b}$ is also geometrically derived, as explained later, in Sect. 4.7.

4.3 Predicted and Error Values

The solutions (4.9) and (4.12) for the data set in Table 4.1 are shown in Table 4.2 (A); $\hat{\mathbf{b}} = [7.61, -0.18, 18.23]'$ and $\hat{c} = 256.4$. Substituting these values into (4.1), we have

Table 4.2 Results of regression analysis for the data in Table 4.1

Solution	Regression coefficient			Intercept $\hat{c}$	Variance explained	Multiple correlation
	$\hat{b}_1$: quality	$\hat{b}_2$: price	$\hat{b}_3$: appearance			
(A) Unstandardized	7.61	−0.18	18.23	256.4	0.73	0.85
(B) Standardized	0.51	−1.17	0.77	0.0		

$$sales = 7.61 \times quality - 0.18 \times price + 18.23 \times appearance + 256.4 + error. \tag{4.18}$$

This equation is useful for *predicting* the *future sales* of a product, which is not included in Table 4.1. For example, let us suppose that a product has not yet been sold, but its *quality*, *price*, and *appearance* have been found to be 6, 1500, and 4. We can substitute those values into (4.18) to predict the *sales* as follows:

$$sales = 7.61 \times 6 - 0.18 \times 1500 + 18.23 \times 4 + 256.4 + error = 105 + error. \tag{4.19}$$

That is, *future sales* can be counted as 105, although any *future error* is unknown. However, *existent errors* can be assessed as described in the following paragraph.

Let us consider substituting the solutions $\hat{\mathbf{b}}$ and $\hat{c}$ into (4.5), giving $\mathbf{y} = \mathbf{X}\hat{\mathbf{b}} + \hat{c}\mathbf{1}_n + \hat{\mathbf{e}}$, which is rewritten as

$$\mathbf{y} = \hat{\mathbf{y}} + \hat{\mathbf{e}}, \text{ i.e., } \hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}} \tag{4.20}$$

by defining a *predicted value vector* as

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{b}} + \hat{c}\mathbf{1}_n. \tag{4.21}$$

In (4.20), we have attached the "hat" mark to the $\mathbf{e}$ in (4.5), i.e., replaced it with $\hat{\mathbf{e}}$, in order to emphasize that the error vector $\mathbf{e}$, which had been unknown before analysis, becomes known, as shown next: Using $\hat{\mathbf{b}} = [7.61, -0.18, 18.23]'$, $\hat{c} = 256.4$, and $\mathbf{X}$ in Table 4.1, the values in (4.21) are given by

$$\hat{\mathbf{y}} = \begin{bmatrix} 10 & 1800 & 2.6 \\ 5 & 1550 & 4.2 \\ \vdots & \vdots & \vdots \\ 7 & 1600 & 5.4 \end{bmatrix} \begin{bmatrix} 7.61 \\ -0.18 \\ 18.23 \end{bmatrix} + 256.4 \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = \begin{bmatrix} 56.0 \\ 92.1 \\ \vdots \\ 120.2 \end{bmatrix}, \tag{4.22}$$

while $\mathbf{y} = [48, 104, \ldots, 106]'$ as seen in Table 4.1. This vector and (4.22) are used in (4.20) to provide

$$\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}} = \begin{bmatrix} 48 \\ 104 \\ \vdots \\ 106 \end{bmatrix} - \begin{bmatrix} 56.0 \\ 92.1 \\ \vdots \\ 120.2 \end{bmatrix} = \begin{bmatrix} -8.0 \\ 11.9 \\ \vdots \\ -14.2 \end{bmatrix}. \tag{4.23}$$

Its squared norm

$$\|\hat{\mathbf{e}}\|^2 = (-8.0)^2 + 11.9^2 + \cdots + (-14.2)^2 \tag{4.24}$$

indicates the largeness of errors.

4.4 Proportion of Explained Variance and Multiple Correlation

The purpose of this section is to introduce a statistic that indicates how *successful* the results of regression analysis are, using (4.24) and the three properties for $\hat{\mathbf{y}}$ and $\hat{\mathbf{e}}$ described in the following paragraph.

The first property is

$$\mathbf{J}\hat{\mathbf{y}} = \mathbf{J}\mathbf{X}\hat{\mathbf{b}}, \tag{4.25}$$

which is derived as follows: (4.21) implies $\mathbf{J}\hat{\mathbf{y}} = \mathbf{J}\mathbf{X}\hat{\mathbf{b}} + \mathbf{J}(\hat{c}\mathbf{1}_n)$, with $\mathbf{J}(\hat{c}\mathbf{1}_n) = \hat{c}(\mathbf{1}_n{}'\mathbf{J})' = \mathbf{0}_n$ following from (2.11) and (2.13). The second property is

$$\mathbf{J}\hat{\mathbf{e}} = \hat{\mathbf{e}}, \tag{4.26}$$

which follows from the use of (2.12) in (4.10). Property (4.26) shows that the *average of an error vector* is always *zero*; $n^{-1}\mathbf{1}_n'\hat{\mathbf{e}} = n^{-1}\mathbf{1}_n'\mathbf{J}\hat{\mathbf{e}} = 0$, because of (2.13). The third property is that *errors* are *uncorrelated* to *predicted values* with their covariance $n^{-1}\hat{\mathbf{e}}'\mathbf{J}\hat{\mathbf{y}} = 0$, i.e.,

$$\hat{\mathbf{e}}'\mathbf{J}\hat{\mathbf{y}} = 0. \tag{4.27}$$

Readers interested in the proof of (4.27) should see the following note:

Note 4.3. No Correlation between Errors and Predictive Values
The use of (4.21) and (4.25) in (4.20) leads to $\mathbf{J}\hat{\mathbf{e}} = \mathbf{J}\mathbf{y} - \mathbf{J}\mathbf{X}\hat{\mathbf{b}}$. Substituting this and (4.25) in $\hat{\mathbf{e}}'\mathbf{J}\hat{\mathbf{y}} = \hat{\mathbf{e}}'\mathbf{J}'\mathbf{J}\hat{\mathbf{y}}$, this is rewritten as

$$\hat{\mathbf{e}}'\mathbf{J}\hat{\mathbf{y}} = \left(\mathbf{J}\mathbf{y} - \mathbf{J}\mathbf{X}\hat{\mathbf{b}}\right)'\mathbf{J}\mathbf{X}\hat{\mathbf{b}} = \mathbf{y}'\mathbf{J}\mathbf{X}\hat{\mathbf{b}} - \hat{\mathbf{b}}'\mathbf{X}'\mathbf{J}\mathbf{X}\hat{\mathbf{b}},$$

where (2.11) and (2.12) have been used. We can further substitute (4.12) in the above equation to have

$$\begin{aligned}\hat{\mathbf{e}}'\mathbf{J}\hat{\mathbf{y}} &= \mathbf{y}'\mathbf{J}\mathbf{X}(\mathbf{X}'\mathbf{J}\mathbf{X})^{-1}\mathbf{X}'\mathbf{J}\mathbf{y} - \mathbf{y}'\mathbf{J}'\mathbf{X}(\mathbf{X}'\mathbf{J}\mathbf{X})^{-1}\mathbf{X}'\mathbf{J}\mathbf{X}(\mathbf{X}'\mathbf{J}\mathbf{X})^{-\mathbf{I}}\mathbf{X}'\mathbf{J}\mathbf{y} \\ &= \mathbf{y}'\mathbf{J}\mathbf{X}(\mathbf{X}'\mathbf{J}\mathbf{X})^{-1}\mathbf{X}'\mathbf{J}\mathbf{y} - \mathbf{y}'\mathbf{J}\mathbf{X}(\mathbf{X}'\mathbf{J}\mathbf{X})^{-1}\mathbf{X}'\mathbf{J}\mathbf{y} = 0\end{aligned} \tag{4.28}$$

The pre-multiplication of the first equality in (4.20) by $\mathbf{J}$ leads to $\mathbf{J}\mathbf{y} = \mathbf{J}\hat{\mathbf{y}} + \mathbf{J}\hat{\mathbf{e}} = \mathbf{J}\hat{\mathbf{y}} + \hat{\mathbf{e}}$ because of (4.26). Further, the angle between $\mathbf{J}\hat{\mathbf{y}}$ and $\hat{\mathbf{e}}$ being 90° is found in (4.27). This fact implies that $\mathbf{J}\mathbf{y}$, $\mathbf{J}\hat{\mathbf{y}}$, and $\hat{\mathbf{e}}$ form the right triangle illustrated in Fig. 4.1. We can thus use the *Pythagorean theorem* to have

$$\|\mathbf{J}\mathbf{y}\|^2 = \|\mathbf{J}\hat{\mathbf{y}}\|^2 + \|\hat{\mathbf{e}}\|^2. \tag{4.29}$$

From (4.29) we can derive a statistic indicating how *successful* the results of regression analysis are, as follows: The division of both sides of (4.29) by $\|\mathbf{J}\mathbf{y}\|^2$ leads to

$$1 = \frac{\|\mathbf{J}\hat{\mathbf{y}}\|^2}{\|\mathbf{J}\mathbf{y}\|^2} + \frac{\|\hat{\mathbf{e}}\|^2}{\|\mathbf{J}\mathbf{y}\|^2}. \tag{4.30}$$

Here, the proportion $\|\hat{\mathbf{e}}\|^2/\|\mathbf{J}\mathbf{y}\|^2$ taking a value within the range [0, 1] stands for the relative largeness of errors; equivalently, one minus that proportion,

$$\frac{\|\mathbf{J}\hat{\mathbf{y}}\|^2}{\|\mathbf{J}\mathbf{y}\|^2} = 1 - \frac{\|\hat{\mathbf{e}}\|^2}{\|\mathbf{J}\mathbf{y}\|^2}, \tag{4.31}$$

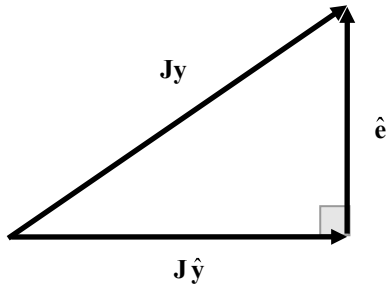


Fig. 4.1 Geometric relationship among **e** (errors), **Jy** (centered dependent variable), and **Jŷ** (centered predicted values)

taking a value within the range [0, 1] indicates the smallness of errors, i.e., *successfulness* of regression analysis. Statistic (4.31) is called the *proportion of explained variance*, as it can be rewritten as:

$$\frac{\|\mathbf{J}\hat{\mathbf{y}}\|^2}{\|\mathbf{J}\mathbf{y}\|^2} = \frac{n^{-1}\|\mathbf{J}\hat{\mathbf{y}}\|^2}{n^{-1}\|\mathbf{J}\mathbf{y}\|^2} = \frac{n^{-1}\left\|\mathbf{J}\mathbf{X}\hat{\mathbf{b}}\right\|^2}{n^{-1}\|\mathbf{J}\mathbf{y}\|^2} \tag{4.32}$$

using (4.25). That is, the denominator of (4.32), $n^{-1}\|\mathbf{J}\mathbf{y}\|^2$, is the variance of a dependent variable, while the numerator, $n^{-1}\left\|\mathbf{J}\mathbf{X}\hat{\mathbf{b}}\right\|^2$, is the variance of predicted values based on the explanatory variables in $\mathbf{X}$, which implies that (4.32) indicates the *proportion of the variance explained by explanatory variables* in the variance of the dependent variable. The resulting proportion of the explained variance in Table 4.2 is found to be 0.73, which is interpreted to mean that 73% of the variance of the dependent variable (i.e., how much more/less *sales* are) is explained by *quality*, *price*, and *appearance*. Statistic (4.32) is also called a *coefficient of determination*.

Let us consider the square root of (4.32). This can be rewritten as

$$\frac{\|\mathbf{J}\hat{\mathbf{y}}\|}{\|\mathbf{J}\mathbf{y}\|} = \frac{\|\mathbf{J}\hat{\mathbf{y}}\|}{\|\mathbf{J}\mathbf{y}\|} \times \frac{\|\mathbf{J}\hat{\mathbf{y}}\|}{\|\mathbf{J}\hat{\mathbf{y}}\|} = \frac{\hat{\mathbf{y}}'\mathbf{J}\hat{\mathbf{y}}}{\|\mathbf{J}\mathbf{y}\|\|\mathbf{J}\hat{\mathbf{y}}\|} = \frac{\mathbf{y}'\mathbf{J}\hat{\mathbf{y}}}{\|\mathbf{J}\mathbf{y}\|\|\mathbf{J}\hat{\mathbf{y}}\|} = \frac{(\mathbf{J}\mathbf{y})'\mathbf{J}\hat{\mathbf{y}}}{\|\mathbf{J}\mathbf{y}\|\|\mathbf{J}\hat{\mathbf{y}}\|} \tag{4.33}$$

where we have used $\mathbf{y}'\mathbf{J}\hat{\mathbf{y}} = (\hat{\mathbf{y}} + \hat{\mathbf{e}})'\mathbf{J}\hat{\mathbf{y}} = \hat{\mathbf{y}}'\mathbf{J}\hat{\mathbf{y}}$ because of (4.27). Comparing (4.33) with (3.8), we can find (4.33) to be the *correlation coefficient* between $\mathbf{y}$ and $\hat{\mathbf{y}}$. In particular, (4.33) is called the *multiple correlation coefficient* between dependent and explanatory variables, as we can use (4.25) to rewrite (4.33) as

$$\frac{(\mathbf{J}\mathbf{y})'\mathbf{J}\hat{\mathbf{y}}}{\|\mathbf{J}\mathbf{y}\|\|\mathbf{J}\hat{\mathbf{y}}\|} = \frac{(\mathbf{J}\mathbf{y})'\mathbf{J}\mathbf{X}\hat{\mathbf{b}}}{\|\mathbf{J}\mathbf{y}\|\left\|\mathbf{J}\mathbf{X}\hat{\mathbf{b}}\right\|} \tag{4.34}$$

which stands for the relationship of $\mathbf{y}$ to the *multiple* variables in $\mathbf{X}$. Its value, $0.85 = \sqrt{0.73}$ in Table 4.2, is near the upper limit of 1 and indicates a close relationship of *sales* to *quality*, *price*, and *appearance*.

4.5 Interpretation of Regression Coefficients

Simple regression analysis with a single explanatory variable can be formulated in the same manner as the multiple regression analysis described so far, except that p is restricted to one, i.e., $\mathbf{X}$ is set to an $n \times 1$ vector $\mathbf{x}$. *Simple regression* analysis with the model "*sales* = $b \times$ *quality* + c + *error*" for the data in Table 4.1 produces the result

Table 4.3 Covariances and correlations among the four variables in Table 4.1

Variable	V: covariance matrix				R: correlation matrix			
	Quality	Price	Appear[a]	Sales	Quality	Price	Appear[a]	Sales
Quality	4.5				1			
Price	245.5	41,801			0.57	1		
Appear[a]	−0.4	104.6	1.8		−0.16	0.39	1	
Sales	−18	−3748	10	985	−0.27	−0.58	0.24	1

[a]Appearance

$$sales = -4.02 \times quality + 128.73 + error. \tag{4.35}$$

Here, it should be noted that the *regression coefficient* for *quality* is *negative*. The covariance and correlation coefficients between *sales* and *quality* in Table 4.3 are also *negative*. Those negative values show that the products of *lower quality* tend to *sell better*. This seems to be *unreasonable*. This is due to the fact that the above coefficients are obtained only from a pair of two variables (*quality* and *sales*) without using the other variables (*prices* and *appearance*), as explained next.

Let us note the positive correlation of *quality* to *price*, which tends to *decrease sales*. That is, a *third variable*, *price, intermediates* between *quality* and *sales*. These may also be intermediated by *appearance*. The effects of *intermediate variables* cannot be considered by the statistics obtained for two variables excluding the intermediate ones.

On the other hand, we can find in Table 4.2 that the coefficient for quality, $\hat{b}_1 = 7.61$, resulting from *multiple regression*, is reasonably *positive*. This is because the other variables are included in the model, as found in (4.1). The coefficient value $\hat{b}_1 = 7.61$ is interpreted as indicating the following relationship: The *sales* increase by 7.61 on average with a unit increase in *quality*, while the values of the other variables are kept fixed. Why this interpretation is derived should be understood in a rather subjective manner: A unit increase in quality with price and appearance fixed in (4.1) can be expressed as

$$sales^* = b_1 \times (quality + 1) + b_2 \times price + b_3 \times appearance + c + error^*, \tag{4.36}$$

where asterisks have been attached to *sales* and *error*, since an increase in *quality* changes their values from those of the *sales* and *error* in (4.1), i.e.,

$$sales = b_1 \times quality + b_2 \times price + b_3 \times appearance + c + error.$$

The subtraction of both sides of (4.1) from those of (4.36) gives

$$sales^* - sales = b_1 + (error^* - error), \tag{4.37}$$

whose average equals b_1, since the average of errors is zero, i.e., (4.26) leads to $n^{-1}\mathbf{1}_n{}'\hat{\mathbf{e}} = 0$, implying that the average of $error^* - error$ in (4.37) is zero.

The other coefficients are also interpreted in the same manner. For example, $\hat{b}_2 = -0.18$ in Table 4.2 allows us to consider the following: *sales* tend to *decrease* by 0.18 on average with a unit increase in *price*, while *quality* and *appearance* are fixed.

4.6 Standardization

It is senseless to compare the largeness of the three regression coefficients in Table 4.2(A) ($\hat{b}_1 = 7.61$, $\hat{b}_2 = -0.18$, $\hat{b}_3 = 18.23$), since they are obtained from the raw scores in which the variances (i.e., how widely the values range) differ across variables. For the comparison of coefficients to make sense, regression analysis must be carried out for the standardized data in which the values in all variables have been transformed into standard scores, so that the variances are equivalent among the variables (i.e., all unity). The solutions for standard scores are called *standardized solutions*, while those for raw scores, which we have seen so far, are called *unstandardized solutions*. However, the standardized and unstandardized solutions of regression analysis for the same data set can be regarded as the *two different expressions of the same solution*, as shown next.

The standard score vector for $\mathbf{y}$ is expressed as

$$\mathbf{y}_S = \frac{1}{\sqrt{v_y}}\mathbf{J}\mathbf{y} \tag{4.38}$$

by substituting $\mathbf{y}$ for $\mathbf{x}_j$ and v_y for v_{jj} in (2.24), where v_y is the variance of the dependent variable; it should be noticed that $\mathbf{y}_j$ in (2.24) is different from $\mathbf{y}$ in this chapter. The standard score matrix for $\mathbf{X}$ is expressed as (2.32), i.e.,

$$\mathbf{Z} = \mathbf{J}\mathbf{X}\mathbf{D}^{-1}. \tag{4.39}$$

Here, $\mathbf{D} = \begin{bmatrix} \sqrt{v_{11}} & & \\ & \ddots & \\ & & \sqrt{v_{pp}} \end{bmatrix}$ is the $p \times p$ diagonal matrix, with its jth diagonal element $v_{jj}^{1/2}$ being the standard deviation of the jth explanatory variable, implying that its variance is v_{jj}. Substituting (4.38) and (4.39) into $\mathbf{y}$ and $\mathbf{X}$ in (4.12), respectively, we have the *standardized solution of the regression coefficient vector*

$$\begin{aligned} \hat{\mathbf{b}}_S &= (\mathbf{Z}'\mathbf{J}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{J}\mathbf{y}_S = \left(\mathbf{D}^{-1}\mathbf{X}'\mathbf{J}\mathbf{X}\mathbf{D}^{-1}\right)^{-1}\left(\mathbf{D}^{-1}\mathbf{X}'\mathbf{J}'\right)\left(\frac{1}{\sqrt{v_y}}\mathbf{J}\mathbf{y}\right) \\ &= \frac{1}{\sqrt{v_y}}\mathbf{D}(\mathbf{X}'\mathbf{J}\mathbf{X})^{-1}\mathbf{D}\mathbf{D}^{-1}\mathbf{X}'\mathbf{J}\mathbf{y} = \frac{1}{\sqrt{v_y}}\mathbf{D}(\mathbf{X}'\mathbf{J}\mathbf{X})^{-1}\mathbf{X}'\mathbf{J}\mathbf{y} = \frac{1}{\sqrt{v_y}}\mathbf{D}\hat{\mathbf{b}}. \end{aligned} \tag{4.40}$$

Here, (4.16) has been used: $(\mathbf{D}^{-1}\mathbf{X}'\mathbf{J}\mathbf{X}\mathbf{D}^{-1})^{-1} = (\mathbf{X}'\mathbf{J}\mathbf{X}\mathbf{D}^{-1})^{-1}\mathbf{D} = \mathbf{D}(\mathbf{X}'\mathbf{J}\mathbf{X})^{-1}\mathbf{D}$. Formula (4.40) shows that $\hat{\mathbf{b}}_S$ is easily transformed from the unstandardized solution $\hat{\mathbf{b}}$, i.e., the pre-multiplication of $\hat{\mathbf{b}}$ by $v_y^{-1/2}\mathbf{D}$. Further, the substitution of (4.38) and (4.39) into $\mathbf{y}$ and $\mathbf{X}$ in (4.9) leads to the *standardized solution of the intercept* simply being zero:

$$\hat{c}_S = 0. \tag{4.41}$$

Let us substitute (4.38), (4.39), and (4.40) for $\mathbf{y}$, $\mathbf{X}$, and $\hat{\mathbf{b}}$ in (4.32). Then, we have

$$\frac{n^{-1}\left\|\mathbf{J}\mathbf{Z}\hat{\mathbf{b}}_S\right\|^2}{n^{-1}\|\mathbf{J}\mathbf{y}_S\|^2} = \frac{n^{-1}\left\|\mathbf{J}\mathbf{X}\mathbf{D}^{-1}\left(\frac{1}{\sqrt{v_y}}\mathbf{D}\hat{\mathbf{b}}\right)\right\|^2}{n^{-1}\left\|\frac{1}{\sqrt{v_y}}\mathbf{J}\mathbf{y}\right\|^2} = \frac{n^{-1}\left\|\mathbf{J}\mathbf{X}\hat{\mathbf{b}}\right\|^2}{n^{-1}\|\mathbf{J}\mathbf{y}\|^2}. \tag{4.42}$$

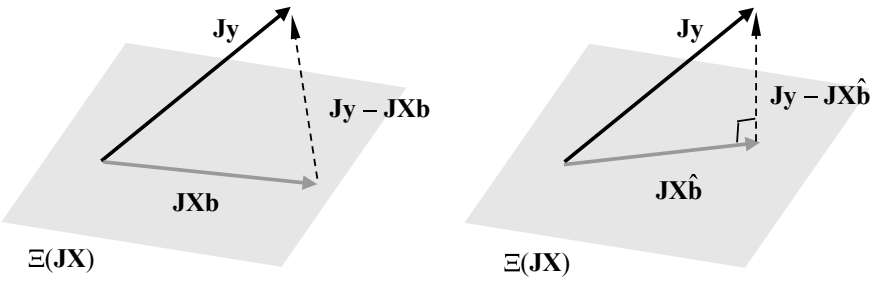
This shows that the *proportion of explained variance* remains *equal* to (4.32) and its square root (4.33) or (4.34) (*multiple correlation coefficient*) also remains unchanged, even if the data set is standardized. That is, the index value for the successfulness of regression analysis is equivalent between unstandardized and standardized solutions.

Let us see the regression coefficient $\hat{\mathbf{b}}_S$ in the standardized solution, which is called the *standardized regression coefficient*, in Table 4.2(B). A comparison of their values makes sense. We can find that the absolute value of the coefficient for *price* is the largest among the three exploratory variables, showing that the effect of *price* on *sales* is the largest among the three. Further, the coefficient of *price* is negative, implying that *sales* tend to decrease with an increase in *price*. The effect of *quality* is found to be the least among the three variables.

4.7 Geometric Derivation of Regression Coefficients

This section can deepen our understanding of regression analysis, with knowledge of the vector space explained in Appendix A.1.3 being necessary here.

The minimization of (4.11) over $\mathbf{b}$ is also restated as minimizing the squared length of the vector (4.10), i.e., $\mathbf{e} = \mathbf{J}\mathbf{y} - \mathbf{J}\mathbf{X}\mathbf{b}$. To solve this problem, let us consider what $\mathbf{J}\mathbf{X}\mathbf{b}$ geometrically stands for when the elements of $\mathbf{b}$ take any real values. It can be represented as the grayed plane in Fig. 4.2a. Though it has been depicted as a two-dimensional plane in the figure, the grayed plane indeed represents a *p-dimensional space*, which is formally expressed as



(a) When **b** is arbitrary (b) When **b** is the solution $\hat{\mathbf{b}}$

Fig. 4.2 **JXb** in space Ξ(**JX**) with vector **Jy**

$$\begin{aligned}\Xi(\mathbf{JX}) = \{\mathbf{JX}^* : \mathbf{Jx}^* = \mathbf{JXb} = [\mathbf{Jx}_1, \ldots, \mathbf{Jx}_p]\mathbf{b} \\ = b_1\mathbf{J}_{\mathbf{X}1} + \cdots + b_p\mathbf{Jx}_p; -\infty < b_j < \infty, j = 1, \ldots, p\},\end{aligned} \tag{4.43}$$

with $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_p]$ and $\mathbf{b} = [b_1, \ldots, b_p]$; (4.43) is equivalent to (A.1.12) in Appendix A.1.3, with $\mathbf{h}^*$ and **H** in (A.1.12) replaced by $\mathbf{Jx}^*$ and **JX** in (4.43), respectively. We can set $b_1, \ldots, b_p$ to any real values so that the terminus of the vector $\mathbf{Jx}^* = \mathbf{JXb}$ moves in the space (4.43), i.e., on the grayed plane in Fig. 4.2a.

The function (4.11) to be minimized is the squared length of the difference vector $\mathbf{e} = \mathbf{Jy} - \mathbf{JXb}$, which is depicted as a dotted line in Fig. 4.2a. It is found to be the shortest, i.e., the minimum, when $\mathbf{e} = \mathbf{Jy} - \mathbf{JXb}$ is orthogonal to **JXb**, as in Fig. 4.2b, that is, when

$$(\mathbf{JXb})'(\mathbf{Jy} - \mathbf{JXb}) = \mathbf{b}'\mathbf{X}'\mathbf{Jy} - \mathbf{b}'\mathbf{X}'\mathbf{JXb} = 0, \tag{4.44}$$

which holds for **b** equaling (4.12). This is shown by the fact that the substitution of (4.12) into **b** in $\mathbf{b}'\mathbf{X}'\mathbf{Jy} - \mathbf{b}'\mathbf{X}'\mathbf{JXb}$ [i.e., the middle side of (4.44)] leads to the second and last equalities in (4.28). We should also note that the right triangle found in Fig. 4.2b is the one in Fig. 4.1.

4.8 Bibliographical Notes

There are a number of books in which regression analysis is exhaustively detailed. Among them are Montgomery, Peck, and Vining (2012) and Fahrmeir, Kneib, Lang, and Marx (2013).

Multivariate data analysis procedures including regression analysis are exhaustively introduced in Lattin et al. (2003) with a number of real data examples. Izenman (2008) and Koch (2014) are examples of advanced books on multivariate

data analysis procedures recommended for those who have finished reading the present book.

One topic that has not been mentioned in this chapter is *variable selection*, i.e., the problem of selecting useful exploratory variables and discarding useless ones among the initial set of variables. A modern approach to this problem is treated in Chap. 21.

Exercises

4.1. Show that

$$\mathbf{A}^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix} \tag{4.45}$$

is the inverse matrix of $\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$.

4.2. Let us consider the system of equations $\begin{cases} -6x_1 + 2x_2 = 7 \\ 3x_1 + 9x_2 = -12 \end{cases}$, i.e., $\mathbf{Ax} = \mathbf{c}$, with $\mathbf{A} = \begin{bmatrix} -6 & 2 \\ 3 & 9 \end{bmatrix}$ and $\mathbf{c} = \begin{bmatrix} 7 \\ -12 \end{bmatrix}$. Compute the solution of $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ for the system using $\mathbf{A}^{-1}$ in (4.45).

4.3. Show $(\mathbf{AB})^{-1\prime} = \mathbf{A}'^{-1}\mathbf{B}'^{-1}$ with $\mathbf{A}$ and $\mathbf{B}$ being nonsingular.

4.4. Consider the model $y_i = c + e_i$ ($i = 1, \ldots, n$), i.e., $\mathbf{y} = c\mathbf{1}_n + \mathbf{e}$, for a data vector $\mathbf{y} = [y_1, \ldots, y_n]'$, with $\mathbf{e} = [e_1, \ldots, e_n]'$ containing errors and c the parameter to be obtained. Show that the average $\bar{y} = n^{-1}\sum_{i=1}^{n} y_i$ is the least squares solution of c in the model, i.e., that $f(c) = \|\mathbf{y} - c\mathbf{1}_n\|^2$ is minimized for $c = \bar{y}$, using the facts in Appendix A.2.1.

4.5. Show that the solution of intercept c in (4.9) is zero if $\mathbf{y}$ and each column of $\mathbf{X}$ contain centered scores.

4.6. Show that the regression model (4.5) can be rewritten as

$$\mathbf{y} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \mathbf{e}, \tag{4.46}$$

with $\tilde{\mathbf{X}} = \begin{bmatrix} x_{11} & \cdots & x_{1p} & 1 \\ & \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} & 1 \end{bmatrix}$ an $n \times (p + 1)$ matrix and $\boldsymbol{\beta} = \begin{bmatrix} b_1 \\ \vdots \\ b_p \\ c \end{bmatrix}$ a $(p+1) \times 1$ vector.

4.7. Show that $\hat{\boldsymbol{\beta}} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{y}$ is the least squares solution of $\boldsymbol{\beta}$ for (4.46), i.e., $\left\|\mathbf{y} - \tilde{\mathbf{X}}\boldsymbol{\beta}\right\|^2$ is minimized for $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, using the facts in Appendix A.2.2.

4.8. Show that (4.12) can be rewritten as $\hat{\mathbf{b}} = \mathbf{V}_{XX}^{-1}\mathbf{v}_{Xy}$. Here, $\mathbf{V}_{XX} = n^{-1}\mathbf{X}'\mathbf{J}\mathbf{X}$ is the covariance matrix among explanatory variables, and $\mathbf{v}_{Xy} = n^{-1}\mathbf{X}'\mathbf{J}\mathbf{y}$ is the vector containing the covariances between explanatory and dependent variables, with n the number of individuals.

4.9. Show that (4.40) can be rewritten as $\hat{\mathbf{b}}_S = \mathbf{R}_{XX}^{-1}\mathbf{r}_{Xy}$. Here, $\mathbf{R}_{XX} = n^{-1}\mathbf{D}^{-1}\mathbf{X}'\mathbf{J}\mathbf{X}\mathbf{D}^{-1}$ is the correlation matrix among explanatory variables, and $\mathbf{r}_{Xy} = n^{-1}d^{-1}\mathbf{D}^{-1}\mathbf{X}'\mathbf{J}\mathbf{y}$ is the vector containing the correlation coefficients between explanatory and dependent variables, with n the number of individuals, $\mathbf{D}$ the diagonal matrix whose jth diagonal element is the standard deviation for the jth variable in $\mathbf{X}$, and d the standard deviation of the elements in $\mathbf{y}$.

4.10. Discuss how $\mathbf{J}\mathbf{X}\hat{\mathbf{b}}$ in Fig. 4.2b is the image of a pencil reflected in a mirror, when $\mathbf{J}\mathbf{y}$ and $\Xi(\mathbf{J}\mathbf{X})$ stand for the pencil and mirror, respectively, with $p = 2$.

4.11. In some procedures, a combination of function $f(\boldsymbol{\theta})$ and another one $g(\boldsymbol{\theta})$, i.e.,

$$f(\theta) + \tau g(\theta), \tag{4.47}$$

is minimized, where $\boldsymbol{\theta}$ is a parameter vector, τ is a given nonnegative scalar value, and $g(\boldsymbol{\theta})$ is called a *penalty function* in that it penalizes $\boldsymbol{\theta}$ for increasing $g(\boldsymbol{\theta})$. In a special version of regression analysis (Hoerl & Kennard, 1970), function $f(\boldsymbol{\theta})$ is defined as $f(\mathbf{b}) = \|\mathbf{J}\mathbf{y} - \mathbf{J}\mathbf{X}\mathbf{b}\|^2$ for a dependent variable vector $\mathbf{y}$ ($n \times 1$) and explanatory variable matrix $\mathbf{X}$ ($n \times p$), while a penalty function is defined as $g(\mathbf{b}) = \|\mathbf{b}\|^2$ which penalizes $\mathbf{b}$ for having a large squared norm. That is, $\|\mathbf{J}\mathbf{y} - \mathbf{J}\mathbf{X}\mathbf{b}\|^2 + \tau\|\mathbf{b}\|^2$ is minimized over $\mathbf{b}$ for a given τ. Show that the solution is given by $\mathbf{b} = (\mathbf{X}'\mathbf{J}\mathbf{X} + \tau\mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{J}\mathbf{y}$.

Chapter 5
Principal Component Analysis (Part 1)

In regression analysis (Chap. 4), variables are classified as dependent and explanatory variables. Such a distinction does not exist in *principal component analysis* (*PCA*), which is introduced in this chapter. A single data matrix $\mathbf{X}$ is analyzed in PCA. This was originally conceived by Pearson (1901) and formulated by Hotelling (1933) who named the procedure PCA. As found in this chapter and the next, PCA can be formulated *apparently* in *different* manners. In some textbooks, PCA is firstly formulated as in Sect. 6.3 (in the next chapter), or the formulation found in this chapter is not described. However, the author believes that the latter formulation should precede the former one, in order to comprehend what PCA is. According to ten Berge and Kiers (1996), in which the formulations of PCA are classified into types based on Hotelling (1933), Pearson (1901), and Rao (1973), the formulation in this chapter is based on Pearson, while the next chapter is based on Hotelling.

5.1 Reduction of Variables into Components

PCA is usually used for an n-individuals $\times$ p-variables centered data matrix $\mathbf{X}$, with (3.20), i.e., $\mathbf{1}_n'\mathbf{X} = \mathbf{0}_p'$. Table 5.1(B) shows an example of $\mathbf{X}$ which is a 6-students $\times$ 5-courses matrix of the centered scores transformed from the test scores in Table 5.1(A).

For such a data matrix $\mathbf{X}$, PCA can be formulated with

$$\mathbf{X} = \mathbf{F}\mathbf{A}' + \mathbf{E}. \tag{5.1}$$

Here, $\mathbf{F}$ is an n-individuals $\times$ *m-components* matrix whose elements are called *principal component* (*PC*) *scores*, $\mathbf{A}$ is a p-variables $\times$ *m-components* matrix whose elements are called component *loadings*, and $\mathbf{E}$ contains errors, with

K. Adachi, *Matrix-Based Introduction to Multivariate Data Analysis*,
https://doi.org/10.1007/978-981-15-4103-2_5

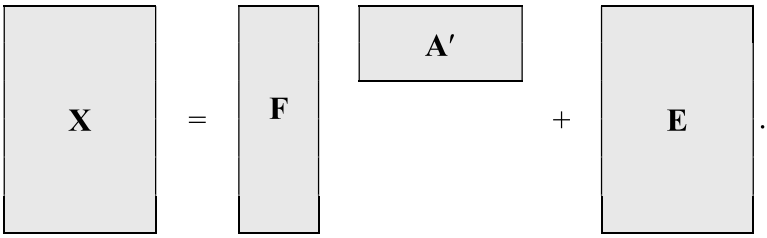
Table 5.1 Test scores for four courses, M (mathematics), P (physics), C (chemistry), and B (biology), with their averages and standard deviations (SD) (artificial example)

Student	(A) Raw scores				(B) Centered scores				(C) Standard scores			
	M	P	C	B	M	P	C	B	M	P	C	B
S1	69.0	66.4	77.0	74.1	−4.9	−10.6	0.3	5.3	−0.45	−0.70	0.02	0.38
S2	67.2	53.6	53.9	58.7	−6.7	−23.4	−22.8	−10.1	−0.61	−1.54	−1.75	−0.73
S3	78.6	96.9	97.3	96.2	4.7	19.9	20.6	27.4	0.43	1.31	1.58	1.97
S4	84.4	87.7	83.9	69.8	10.5	10.7	7.2	1.0	0.96	0.70	0.55	0.07
S5	56.3	68.7	72.1	56.8	−17.6	−8.3	−4.6	−12.0	−1.62	−0.55	−0.35	−0.86
S6	87.9	88.8	76.0	57.2	14.0	11.8	−0.7	−11.6	1.29	0.78	−0.05	−0.83
Average	73.9	77.0	76.7	68.8	0.0	0.0	0.0	0.0	0.00	0.00	0.00	0.00
SD	10.9	15.2	13.0	13.9	10.9	15.2	13.0	13.9	1.00	1.00	1.00	1.00

$$m \leq \mathrm{rank}(\mathbf{X}) \leq \min(n, p). \tag{5.2}$$

The term "*components*" roughly means those entities into which p variables are *summarized* or *reduced*. The kth columns of $\mathbf{F}$ and $\mathbf{A}$ are called the kth components.

Inequality (5.2) implies that (5.1) takes the form



(5.3)

That is, $\mathbf{X}$ is assumed to be approximated by the product of unknown matrices $\mathbf{F}$ and transposed $\mathbf{A}$, with the number of columns (components) in $\mathbf{F}$ and $\mathbf{A}$ being smaller than that of $\mathbf{X}$, as illustrated by the rectangles in (5.3).

The matrices to be obtained in PCA are PC score matrix $\mathbf{F}$ and loading matrix $\mathbf{A}$. For obtaining them, a *least squares method* is used; the sum of the squares of the errors in $\mathbf{E} = \mathbf{X} - \mathbf{F}\mathbf{A}'$,

$$f(\mathbf{F}, \mathbf{A}) = \|\mathbf{E}\|^2 = \|\mathbf{X} - \mathbf{F}\mathbf{A}'\|^2, \tag{5.4}$$

is minimized over $\mathbf{F}$ and $\mathbf{A}$. When $\mathbf{X}$ is the 6×4 matrix in Table 5.1(B) and m is set to 2, the function (5.4) is minimized for the matrices $\mathbf{F}$ and $\mathbf{A}$ shown in Table 5.2 (whose $\mathbf{W}$ is introduced later). There, it should be noticed that $\mathbf{A}$ is of variables $\times$ components, i.e., not transposed as in (5.1) or (5.3). As found in the table, the *students* (individuals), which have been assessed by four kinds of scores (variables) in $\mathbf{X}$, are described by the *two kinds of PC scores* in $\mathbf{F}$, while the *relationships of the*

Table 5.2 Matrices **F**, **A**, and **W** obtained for the centered data matrix in Table 5.1(B)

F (PC scores)			**A** (loadings)			**W** (weights)		
	F1	F2		A1	A2		W1	W2
S1	−0.23	−0.93	M	7.16	6.87	M	0.01	0.05
S2	−1.46	−0.18	P	14.28	4.34	P	0.03	0.03
S3	1.65	−1.04	C	12.52	−1.96	C	0.02	−0.01
S4	0.62	0.73	B	11.02	−7.86	B	0.02	−0.06
S5	−0.81	−0.41						
S6	0.25	1.82						

PC scores to the original four variables are described in **A**. How **F** and **A** are interpreted is explained in Sect. 5.4. Currently, readers need only keep in mind that the *original four variables have been reduced to two components*, which implies that variables are explained by the components whose number is smaller than that of variables. Such a reduction is called *reduced rank approximation*, which is detailed in Appendix A.4.3.

5.2 Singular Value Decomposition

PCA solutions are given through the *singular value decomposition* (*SVD*) introduced in the note below. As SVD is one of the most important properties of matrices, carefully memorizing the following note as *absolute truth* is strongly recommended.

Note 5.1. Singular Value Decomposition (SVD)
Any $n \times p$ matrix $\mathbf{X}$ with $\mathrm{rank}(\mathbf{X}) = r$ can be decomposed as

$$\mathbf{X} = \mathbf{K}\mathbf{\Lambda}\mathbf{L}'. \tag{5.5}$$

Here, $\mathbf{K}$ ($n \times r$) and $\mathbf{L}$ ($p \times r$) satisfy

$$\mathbf{K}'\mathbf{K} = \mathbf{L}'\mathbf{L} = \mathbf{I}_r \tag{5.6}$$

and

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_r \end{bmatrix} \tag{5.7}$$

is an $r \times r$ diagonal matrix whose diagonal elements are positive and arranged in decreasing order:

$$\lambda_1 \geq \cdots \geq \lambda_r > 0. \tag{5.8}$$

This decomposition is called *singular value decomposition* (*SVD*) and λ_k (the kth diagonal element of $\mathbf{\Lambda}$) is called the kth largest *singular value* of $\mathbf{X}$.

In (5.7) the blank elements of $\mathbf{\Lambda}$ represent their being zero, with this expression used in the remaining parts of this book. SVD is described in more detail in Appendix A.3, where the theorem in the above note is presented as a compact form of SVD in Theorem A.3.2.

Another expression of the SVD explained in Note 5.1 is given next:

Note 5.2. Another Expression of SVD (1)
Let us express the matrices $\mathbf{K}$ and $\mathbf{L}$ in Note 5.1 as $\mathbf{K} = [\mathbf{k}_1, \ldots, \mathbf{k}_m, \mathbf{k}_{m+1}, \ldots, \mathbf{k}_r] = [\mathbf{K}_m, \mathbf{K}_{[m]}]$ and $\mathbf{L} = [\mathbf{l}_1, \ldots, \mathbf{l}_m, \mathbf{l}_{m+1}, \ldots, \mathbf{l}_r] = [\mathbf{L}_m, \mathbf{L}_{[m]}]$, with

$$\mathbf{K}_m = [\mathbf{k}_1, \ldots, \mathbf{k}_m] \text{ and } \mathbf{L}_m = [\mathbf{l}_1, \ldots, \mathbf{l}_m] (\text{the } \textit{first } m \text{ columns}), \tag{5.9}$$

$$\mathbf{K}_{[m]} = [\mathbf{k}_{m+1}, \ldots, \mathbf{k}_r] \text{ and } \mathbf{L}_{[m]} = [\mathbf{l}_{m+1}, \ldots, \mathbf{l}_r] \text{ (the } \textit{remaining} \text{ columns)}. \tag{5.10}$$

Then, (5.6) can be rewritten as $\mathbf{k}_u'\mathbf{k}_u = \mathbf{l}_u'\mathbf{l}_u = 1$ and $\mathbf{k}_u'\mathbf{k}_v = \mathbf{l}_u'\mathbf{l}_v = 0$ for $u \neq v$ ($u = 1, \ldots, r$; $v = 1, \ldots, r$). Further, (5.5) can be rewritten as $\mathbf{X} = \lambda_1\mathbf{k}_1\mathbf{l}_1' + \cdots + \lambda_m\mathbf{k}_m\mathbf{l}_m' + \lambda_{m+1}\,\mathbf{k}_{m+1}\mathbf{l}_{m+1}' + \cdots + \lambda_r\mathbf{k}_r\mathbf{l}_r'$, which is expressed in matrix form as

$$\mathbf{X} = \mathbf{K}\mathbf{\Lambda}\mathbf{L}' = \mathbf{K}_m\mathbf{\Lambda}_m\mathbf{L}_m{}' + \mathbf{K}_{[m]}\mathbf{\Lambda}_{[m]}\mathbf{L}_{[m]}{}', \tag{5.11}$$

with

$$\mathbf{\Lambda}_m = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_m \end{bmatrix} \text{ and } \mathbf{\Lambda}_{[m]} = \begin{bmatrix} \lambda_{m+1} & & \\ & \ddots & \\ & & \lambda_r \end{bmatrix}; \quad \text{i.e., } \mathbf{\Lambda} = \begin{bmatrix} \mathbf{\Lambda}_m & \\ & \mathbf{\Lambda}_{[m]} \end{bmatrix}. \tag{5.12}$$

Further, SVD has the following important property, which is directly related to the PCA solution minimizing (5.4):

Note 5.3. SVD and Least Squares Solution
Let $\mathbf{X}$ be an $n \times p$ matrix whose SVD is defined as in Notes 5.1 and 5.2, $\mathbf{F}$ be an $n \times m$ matrix, and $\mathbf{A}$ be a $p \times m$ matrix, with $m \leq r = \text{rank}(\mathbf{X}) \leq \min(n, p)$. Then,

$$f(\mathbf{FA}') = \|\mathbf{X} - \mathbf{FA}'\|^2 \tag{5.13}$$

is minimized for

$$\mathbf{FA}' = \mathbf{K}_m \mathbf{\Lambda}_m \mathbf{L}_{m'}, \tag{5.14}$$

with $\mathbf{K}_m$, $\mathbf{L}_m$, and $\mathbf{\Lambda}_m$ defined as in (5.9) and (5.12).

The fact in the above note is proved by Theorem A.4.5 with (A.4.17) in Appendix A.4.3. The theorem is referred to as Eckart and Young's (1936) theorem in some of the literature.

Let us illustrate the SVD in Note 5.1 and the solution in Note 5.3. The SVD (5.5) for the $\mathbf{X}$ in Table 5.1(B) is given as

$$\underset{\mathbf{X}}{\begin{bmatrix} -4.9 & -10.6 & 0.3 & 5.3 \\ -6.7 & -23.4 & -22.8 & -10.1 \\ 4.7 & 19.9 & 20.6 & 27.4 \\ 10.5 & 10.7 & 7.2 & 1.0 \\ -17.6 & -8.3 & -4.6 & -12.0 \\ 14.0 & 11.8 & -0.7 & -11.6 \end{bmatrix}} = \underset{\mathbf{K}}{\begin{bmatrix} -0.10 & -0.38 & 0.19 & -0.71 \\ -0.60 & -0.07 & 0.52 & 0.42 \\ 0.67 & -0.42 & 0.07 & 0.44 \\ 0.25 & 0.30 & 0.05 & -0.33 \\ -0.33 & -0.17 & -0.82 & 0.11 \\ 0.10 & 0.74 & -0.02 & 0.09 \end{bmatrix}} \underset{\mathbf{\Lambda}}{\begin{bmatrix} 56.57 & & & \\ & 28.10 & & \\ & & 15.72 & \\ & & & 5.16 \end{bmatrix}} \underset{\mathbf{L}'}{\begin{bmatrix} 0.31 & 0.62 & 0.54 & 0.48 \\ 0.60 & 0.38 & -0.17 & -0.68 \\ 0.68 & -0.37 & -0.40 & 0.49 \\ -0.29 & 0.58 & -0.72 & 0.25 \end{bmatrix}} \tag{5.15}$$

Note 5.3 thus shows that the solution of $\mathbf{FA}'$ for minimizing (5.13) is given by

$$\mathbf{FA}' = \underset{\mathbf{K}_m}{\begin{bmatrix} -0.10 & -0.38 \\ -0.60 & -0.07 \\ 0.67 & -0.42 \\ 0.25 & 0.30 \\ -0.33 & -0.17 \\ 0.10 & 0.74 \end{bmatrix}} \underset{\mathbf{\Lambda}_m}{\begin{bmatrix} 56.57 & \\ & 28.10 \end{bmatrix}} \underset{\mathbf{L}_{m'}}{\begin{bmatrix} 0.31 & 0.62 & 0.54 & 0.48 \\ 0.60 & 0.38 & -0.17 & -0.68 \end{bmatrix}} \tag{5.16}$$

We should note that SVD provides the solution of $\mathbf{FA}'$ in function (5.4) for PCA, but not each of $\mathbf{F}$ and $\mathbf{A}$ is given. Their solutions are generally expressed as

$$\mathbf{F} = \mathbf{K}_m \mathbf{\Lambda}_m{}^{\alpha} \mathbf{S}, \tag{5.17}$$

$$\mathbf{A} = \mathbf{L}_m \boldsymbol{\Lambda}_m^{1-\alpha} \mathbf{S}^{-1\prime}, \tag{5.18}$$

with α and $\mathbf{S}$ being arbitrary scalar and nonsingular matrices, respectively. We can easily verify that (5.17) and (5.18) meet (5.14). That is, the solution is *not unique*: There are *infinitely many solutions* for $\{\mathbf{F}, \mathbf{A}\}$. One of the solutions has been shown in Table 5.2, as explained in Sect. 5.4.

5.3 Formulation with a Weight Matrix

Notes 5.1 and 5.2 lead to the following facts:

Note 5.4. Another Expression of SVD (2)
Let $\mathbf{K}_m$, $\mathbf{L}_m$, and $\boldsymbol{\Lambda}_m$ be the matrices defined in Note 5.2. The post-multiplication of $\mathbf{K}_m$ and $\mathbf{L}_m$ by $\boldsymbol{\Lambda}_m$ can be expressed as

$$\mathbf{K}_m \boldsymbol{\Lambda}_m = \mathbf{X}\mathbf{L}_m, \tag{5.19}$$

$$\mathbf{L}_m \boldsymbol{\Lambda}_m = \mathbf{X}'\mathbf{K}_m. \tag{5.20}$$

The facts in the above note are proved in Appendix A.3.3.

By comparing (5.19) with (5.17), we can rewrite the latter as $\mathbf{F} = \mathbf{K}_m \boldsymbol{\Lambda}_m \boldsymbol{\Lambda}_m^{\alpha-1} \mathbf{S} = \mathbf{X}\mathbf{L}_m \boldsymbol{\Lambda}_m^{\alpha-1} \mathbf{S}$, i.e.,

$$\mathbf{F} = \mathbf{X}\mathbf{W} \tag{5.21}$$

with

$$\mathbf{W} = \mathbf{L}_m \boldsymbol{\Lambda}_m^{\alpha-1} \mathbf{S} \tag{5.22}$$

a p-variables $\times$ m-components matrix that we refer to as a *weight matrix*. Equation (5.21) shows that the PC score matrix $\mathbf{F}$ is expressed as the *data matrix post-multiplied by the weight matrix*.

This fact shows that PCA may be formulated, by using (5.21) in (5.4), as minimizing

$$f(\mathbf{W}, \mathbf{A}) = \|\mathbf{X} - \mathbf{X}\mathbf{W}\mathbf{A}'\|^2 \tag{5.23}$$

over $\mathbf{W}$ and $\mathbf{A}$. This minimization is equivalent to minimizing (5.4) over $\mathbf{F}$ and $\mathbf{A}$. Some authors have first presented (5.23) rather than (5.4) as the loss function for

PCA, where the term *loss function* refers to the one to be minimized; its examples are (5.23), (5.4), and (4.7).

Equation (5.21) implies that the resulting PC scores are *centered* ones with

$$\mathbf{1}_n'\mathbf{F} = \mathbf{0}_m', \text{ i.e., } \mathbf{JF} = \mathbf{F}, \tag{5.24}$$

when PCA is performed for a data matrix of centered scores with $\mathbf{1}_n'\mathbf{X} = \mathbf{0}_p'$, since $\mathbf{1}_n'\mathbf{F} = \mathbf{1}_n'\mathbf{XW} = \mathbf{0}_p'\mathbf{W} = \mathbf{0}_m'$, and it is equivalent to $\mathbf{JF} = \mathbf{F}$, as proved in Note 3.1.

5.4 Constraints for Components

For selecting a single set of $\mathbf{F}$ and $\mathbf{A}$ from the multiple solutions satisfying (5.17) and (5.18), we must impose *constraints* onto $\mathbf{F}$ and $\mathbf{A}$. There are various types of constraints, and one of them is that

$$\frac{1}{n}\mathbf{F}'\mathbf{F} = \mathbf{I}_m, \tag{5.25}$$

$$\begin{aligned}&\mathbf{A}'\mathbf{A} \text{ is a diagonal matrix whose}\\ &\quad\text{diagonal elements are arranged in decreasing order.}\end{aligned} \tag{5.26}$$

The solution that satisfies this constraint is

$$\mathbf{F} = \sqrt{n}\mathbf{K}_m, \tag{5.27}$$

$$\mathbf{A} = \frac{1}{\sqrt{n}}\mathbf{L}_m\mathbf{\Lambda}_m, \tag{5.28}$$

which are derived from (5.17) and (5.18) by setting $\alpha = 0$ and $\mathbf{S} = n^{1/2}\mathbf{I}_m$. We can verify that (5.27) and (5.28) satisfy (5.25) and (5.26) by noting that (5.6) and (5.9) imply $\mathbf{K}_m'\mathbf{K}_m = \mathbf{L}_m'\mathbf{L}_m = \mathbf{I}_m$. Under (5.25) and (5.26), the weight matrix is expressed as

$$\mathbf{W} = \sqrt{n}\mathbf{L}_m\mathbf{\Lambda}_m^{-1}, \tag{5.29}$$

which is derived from (5.22) by setting $\alpha = 0$ and $\mathbf{S} = n^{1/2}\mathbf{I}_m$. Table 5.2 shows the solutions of (5.27), (5.28), and (5.29) for the data in Table 5.1(B).

To consider the implications of constraints (5.25) and (5.26), we express the columns of $\mathbf{F}$ and $\mathbf{A}$ as $\mathbf{F} = [\mathbf{f}_1, \ldots, \mathbf{f}_m]$ and $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_m]$, where the elements of $\mathbf{f}_k$ are called *the k*th *PC scores* and those of $\mathbf{a}_k$ are called *the k*th *loadings* ($k = 1, \ldots, m$). Let us note (5.24) and recall (3.22). They show that the left-hand side $n^{-1}\mathbf{F}'\mathbf{F}$ in (5.25) is

the *inter-component covariance matrix of PC scores*, whose diagonal elements $n^{-1}\mathbf{f}'_k\mathbf{f}_k$ are variances, and whose off-diagonal elements $n^{-1}\mathbf{f}'_k\mathbf{f}_l$ $(k \neq l)$ are covariances. The variances and covariances are constrained to be one and zero, respectively, in (5.25). This implies the following:

[1] PC scores are *standardized*.
[2] The kth PC scores are *uncorrelated* with the lth PC ones with $\mathbf{f}'_k\mathbf{f}_l = 0$ for $k \neq l$.

Similarly, the constraint of $\mathbf{A}'\mathbf{A}$ being a diagonal matrix in (5.26) is rewritten as $\mathbf{a}_k'\mathbf{a}_l = 0$ for $k \neq l$, which does *not* imply that $\mathbf{a}_k$ is uncorrelated to $\mathbf{a}_l$, since $\mathbf{1}_p'\mathbf{A} \neq \mathbf{0}_m$, in general, but allows the loadings to have the following property:

[3] The kth loading vector $\mathbf{a}_k$ is *orthogonal* to the lth one $\mathbf{a}_l$.

The properties are desirable in that [2] and [3] allow different components to be distinct and [1] makes it easier to compare PC scores between different components. Further, [1] leads to the following property:

[4] $\mathbf{A}$ $(p \times m)$ is the *covariance matrix* between p *variables* and m *components*, in particular, the *correlation matrix* when $\mathbf{X}$ is *standardized*.

It is proved as follows: We can use (5.20) and (5.27) to rewrite (5.28) as

$$\mathbf{A} = \frac{1}{\sqrt{n}}\mathbf{L}_m\mathbf{\Lambda}_m = \frac{1}{\sqrt{n}}\mathbf{X}'\mathbf{K}_m = \frac{1}{n}\mathbf{X}'\mathbf{F}, \tag{5.30}$$

which equals (3.25) and is the covariance matrix for $\mathbf{X}$ and $\mathbf{F}$, since of $\mathbf{1}_n'\mathbf{X} = \mathbf{0}_p'$ and (5.24). Further, if $\mathbf{X}$ is standardized, (5.30) is the correlation matrix, because of property [1] and (3.25).

Note that the loading matrix $\mathbf{A}$ in Table 5.2 is the covariance matrix for $\mathbf{X}$ and $\mathbf{F}$, but is not their correlation matrix, since it is the result for the data set which is not standardized. On the other hand, Table 5.3 shows the PCA solution for the *standard scores* in Table 5.1(C), where the constraints (5.25) and (5.26) are imposed. The $\mathbf{A}$ in Table 5.3 is the correlation matrix between variables and components. The solution has been given through SVD:

Table 5.3 Matrices $\mathbf{F}$, $\mathbf{A}$, $\mathbf{W}$ obtained for the standard scores in Table 5.1(C)

F (PC scores)			**A** (loadings)			**W** (weights)		
	F1	F2		A1	A2		W1	W2
S1	−0.23	−0.78	M	0.70	0.66	M	0.24	0.79
S2	−1.43	0.13	P	0.94	0.20	P	0.33	0.24
S3	1.58	−1.10	C	0.94	−0.24	C	0.33	−0.28
S4	0.66	0.72	B	0.77	−0.56	B	0.27	−0.66
S5	−0.92	−0.74						
S6	0.33	1.76						

$$\underset{\mathbf{X}}{\begin{bmatrix} -0.45 & -0.70 & 0.02 & 0.38 \\ -0.61 & -1.54 & -1.75 & -0.73 \\ 0.43 & 1.31 & 1.58 & 1.97 \\ 0.96 & 0.70 & 0.55 & 0.07 \\ -1.62 & -0.55 & -0.35 & -0.86 \\ 1.29 & 0.78 & -0.05 & -0.83 \end{bmatrix}} = \underset{\mathbf{K}}{\begin{bmatrix} -0.09 & -0.32 & -0.26 & 0.72 \\ -0.58 & 0.05 & -0.55 & -0.41 \\ 0.65 & -0.45 & -0.13 & -0.44 \\ 0.27 & 0.29 & 0.02 & 0.33 \\ -0.38 & -0.30 & 0.77 & -0.10 \\ 0.14 & 0.72 & 0.16 & -0.10 \end{bmatrix}} \underset{\Lambda}{\begin{bmatrix} 4.15 & & & \\ & 2.24 & & \\ & & 1.23 & \\ & & & 0.38 \end{bmatrix}} \underset{\mathbf{L}'}{\begin{bmatrix} 0.41 & 0.56 & 0.56 & 0.46 \\ 0.72 & 0.22 & -0.26 & -0.61 \\ -0.51 & 0.47 & 0.40 & -0.60 \\ 0.23 & -0.65 & 0.68 & -0.25 \end{bmatrix}} \tag{5.31}$$

with $\mathbf{X}$ being the matrix in Table 5.1(C).

5.5 Interpretation of Loadings

Let us define the columns of matrices as $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_p]$, $\mathbf{A}' = [\mathbf{a}_1, \ldots, \mathbf{a}_p]$, and $\mathbf{E} = [\mathbf{e}_1, \ldots, \mathbf{e}_p]$, with $\mathbf{x}_j$, $\mathbf{a}_j$, and $\mathbf{e}_j$ ($j = 1, \ldots, p$) corresponding to variable j (i.e., $\mathbf{a}_j'$ the jth row vector of $\mathbf{A}$). Then, the PCA model (5.1) is rewritten as

$$\mathbf{x}_i = \mathbf{F}\mathbf{a}_j + \mathbf{e}_j \quad (j = 1, \ldots, p). \tag{5.32}$$

This takes the same form as (4.5) except that (5.32) does not include an intercept. That is, PCA can be regarded as the *regression* of $\mathbf{x}_j$ onto $\mathbf{F}$. When $m = 2$, as in Table 5.3, (5.32) is expressed as

$$\mathbf{x}_i = a_{j1}\mathbf{f}_1 + a_{j2}\mathbf{f}_2 + \mathbf{e}_j \quad (j = 1, \ldots, p), \tag{5.33}$$

with $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2]$ and $\mathbf{a}_j = [a_{j1}, a_{j2}]'$. That is, $\mathbf{f}_1$ and $\mathbf{f}_2$ can be viewed as the *explanatory* variables for a *dependent* variable $\mathbf{x}_j$, with loadings a_{j1} and a_{j2} as the *regression coefficients*. The equation is further rewritten as

$$x_{ij} = a_{j1}f_{i1} + a_{j2}f_{i2} + e_{ij} \quad (i = 1, \ldots, p; j = 1, \ldots, p), \tag{5.34}$$

using $\mathbf{X} = (x_{ij})$, $\mathbf{F} = (f_{ik})$, and $\mathbf{A} = (a_{jk})$.

On the basis of (5.34), we can interpret the loadings in Table 5.3 as follows:

[A1] All a_{j1} show fairly large positive values for all variables (courses), which implies that students with higher values of f_{i1} (the 1st PC score) tend to show higher scores x_{ij} for all courses ($j = 1, \ldots, p$). The *1st component* can thus be interpreted as standing for a *general ability* common to M, P, C, and B.

[A2] a_{j2} show positive loadings for M and P, but negative ones for C and B. The *2nd component* can be interpreted as standing for a *specific ability* advantageous for M and P, but disadvantageous for C and B.

As described with (5.30) in Sect. 5.4, the loadings in Table 5.3 can also be regarded as the *correlation coefficients* of variables to components. For example, courses P and C are very highly correlated with Component 1, since the corresponding coefficient 0.94 is close to the upper limit.

5.6 Interpretation of Weights

The role of weight matrix $\mathbf{W}$ is easily understood by rewriting (5.21) as

$$\mathbf{f}_k = \mathbf{X}\mathbf{w}_k = w_{1k}\mathbf{x}_1 + \cdots + w_{pk}\mathbf{x}_p \quad (k = 1, \ldots, m), \tag{5.35}$$

with $\mathbf{w}_k = [w_{1k}, \ldots, w_{pk}]'$ the kth column of $\mathbf{W} = (w_{jk})$. In (5.35), we find that the elements in $\mathbf{W}$ provide the *weights* by which variables are multiplied to form the PC scores in $\mathbf{F}$. We can further rewrite (5.35) as

$$f_{ik} = w_{1k}x_{i1} + \cdots + w_{pk}x_{ip} \quad (i = 1, \ldots, n; k = 1, \ldots, m). \tag{5.36}$$

This allows us to interpret the $\mathbf{W}$ in Table 5.3 as follows:

[W1] All w_{j1} show positive values for all variables (courses), which shows that the *1st PC score* $f_{i1} = w_{11}x_{i1} + w_{21}x_{i2} + w_{31}x_{i3} + w_{41}x_{i4}$ is the sum of all variables positively weighted. Thus, the score can be interpreted as standing for a *general ability* common to M, P, C, B.

[W2] w_{j2} show positive values for M and P, but negative ones for C and B; the *2nd PC scores* $f_{i2} = w_{12}x_{i1} + w_{22}x_{i2} + w_{32}x_{i3} + w_{42}x_{i4}$ are higher for students who are superior in M and P, while the scores are lower for those who are superior in C and B. The scores can thus be interpreted as standing for a *specific ability* advantageous for M and P, but disadvantageous for C and B.

Those interpretations are congruous with [A1] and [A2] in the last section.

5.7 Percentage of Explained Variance

In this section, we consider assessing the amount of the errors for the resulting solutions. Substituting SVD (5.5) and the solution (5.14) into $\mathbf{X}$ and $\mathbf{F}\mathbf{A}'$, respectively, in the squared sum of errors (5.4), its resulting value can be expressed as

$$\begin{aligned} \|\mathbf{E}\|^2 &= \left\|\mathbf{K}\boldsymbol{\Lambda}\mathbf{L}' - \mathbf{K}_m\boldsymbol{\Lambda}_m\mathbf{L}_m'\right\|^2 \\ &= \mathrm{tr}\mathbf{L}\boldsymbol{\Lambda}\mathbf{K}'\mathbf{K}\boldsymbol{\Lambda}\mathbf{L}' - 2\mathrm{tr}\mathbf{L}\boldsymbol{\Lambda}\mathbf{K}'\mathbf{K}_m\boldsymbol{\Lambda}_m\mathbf{L}_m' + \mathrm{tr}\mathbf{L}_m\boldsymbol{\Lambda}_m\mathbf{K}_m'\mathbf{K}_m\boldsymbol{\Lambda}_m\mathbf{L}_m' \\ &= \mathrm{tr}\boldsymbol{\Lambda}^2 - \mathrm{tr}\boldsymbol{\Lambda}_m^2 \geq 0. \end{aligned} \tag{5.37}$$

Here, we have used $\mathrm{tr}\mathbf{L}\boldsymbol{\Lambda}\mathbf{K}'\mathbf{K}_m\boldsymbol{\Lambda}_m\mathbf{L}_{m'} = \mathrm{tr}(\mathbf{L}_{m'}\mathbf{L}\boldsymbol{\Lambda})(\mathbf{K}'\mathbf{K}_m\boldsymbol{\Lambda}_m) = \mathrm{tr}(\boldsymbol{\Lambda}_m)(\boldsymbol{\Lambda}_m) = \mathrm{tr}\boldsymbol{\Lambda}_m^2\boldsymbol{\Lambda}_m$. This result follows from

$$\mathbf{K}'\mathbf{K}_m = \mathbf{L}'\mathbf{L}_m = \begin{bmatrix} 1 & & & \\ & \ddots & & \\ & & & 1 \\ 0 & \cdots & & 0 \\ & \vdots & & \\ 0 & \cdots & & 0 \end{bmatrix} :$$

$\mathbf{K}'\mathbf{K}_m = \mathbf{L}'\mathbf{L}_m$ equals the $r \times m$ matrix whose first m rows are those of $\mathbf{I}_m$ and the remaining rows are filled with zeros. Dividing (5.37) by $\mathrm{tr}\boldsymbol{\Lambda}^2$ leads to

$$\frac{\|\mathbf{E}\|^2}{\mathrm{tr}\,\boldsymbol{\Lambda}^2} = 1 - PEV_m \geq 0, \tag{5.38}$$

with

$$PEV_m = \frac{\mathrm{tr}\,\boldsymbol{\Lambda}_m^2}{\mathrm{tr}\,\boldsymbol{\Lambda}^2} = \frac{\mathrm{tr}\,\boldsymbol{\Lambda}_m^2}{\|\mathbf{X}\|^2}. \tag{5.39}$$

Here, we have used

$$\|\mathbf{X}\|^2 = \mathrm{tr}\mathbf{X}'\mathbf{X} = \mathrm{tr}\mathbf{L}\boldsymbol{\Lambda}\mathbf{K}'\mathbf{K}\boldsymbol{\Lambda}\mathbf{L}' = \mathrm{tr}\boldsymbol{\Lambda}^2. \tag{5.40}$$

Since (5.38) expresses the *largeness of errors* with taking a *nonnegative* value, (5.39) indicates the *smallness of errors*, i.e., how well $\mathbf{FA}'$ approximates $\mathbf{X}$, by taking a value within the range [0, 1]. Some different terms are used for proportion (5.39). One of them is the *proportion of explained variance* (*PEV*), since (5.39) can be rewritten as

$$PEV_m = \frac{\frac{1}{n}\mathrm{tr}(\mathbf{K}_m\boldsymbol{\Lambda}_m\mathbf{L}'_m)'\mathbf{K}_m\boldsymbol{\Lambda}_m\mathbf{L}'_m}{\frac{1}{n}\mathrm{tr}\,\mathbf{X}'\mathbf{X}} = \frac{\frac{1}{n}\mathrm{tr}(\mathbf{FA}')'\mathbf{FA}'}{\mathrm{tr}\,\mathbf{V}}, \tag{5.41}$$

with $\mathbf{V} = n^{-1}\mathbf{X}'\mathbf{X}$ the covariance matrix given in (3.22); the denominator of (5.41) is the sum of the variances of p variables, while the numerator is the sum of the variances of the columns of $\mathbf{FA}'$, i.e., (5.14), since (5.24) implies that $\mathbf{FA}'$ is centered with $\mathbf{1}_n'\mathbf{FA}' = \mathbf{0}_p'$.

The PEV for the solution with $m = 2$ in Table 5.3 is obtained as

$$PEV_2 = \frac{4.15^2 + 2.24^2}{4.15^2 + 2.24^2 + 1.23^2 + 0.38^2} = \frac{22.24}{23.90} = 0.93, \tag{5.42}$$

using (5.31). This implies that 93% of the data variances are accounted for by two components; in other words, 7% (=100–93) of the variances remain unexplained. If we adopt the $m = 3$ solution, the PEV is

$$PEV_3 = \frac{4.15^2 + 2.24^2 + 1.23^2}{4.15^2 + 2.24^2 + 1.23^2 + 0.38^2} = \frac{23.75}{23.90} = 0.99.$$

The PEV for the solution with $m = 2$ in Table 5.2 (i.e., the solution for the centered data matrix without standardization) is obtained as

$$PEV_2 = \frac{56.57^2 + 28.10^2}{56.57^2 + 28.10^2 + 15.72^2 + 5.16^2} = \frac{3989.78}{4263.52} = 0.94, \tag{5.43}$$

using (5.15) and (5.16). This differs from (5.42); the difference is not due to a round-off error. This shows that the PCA solution for a centered data matrix without standardization differs from that for the standard scores for the same data matrix. The latter solution cannot be straightforwardly transformed from the former, which differs from the regression analysis in the last chapter.

5.8 High-Dimensional Data Analysis

Recently, we have often encountered data sets with much more variables than individuals, i.e., an $n \times p$ data matrix $\mathbf{X}$ with $p \gg n$. Such a data set is said to be *high-dimensional* (e.g., Kock, 2014). In order to find a few components underlying a number of variables, PCA is useful. In this section, we illustrate PCA for high-dimensional data using Yeung and Ruzzo's (2001) gene expression data with $n = 17$ time points and $p = 384$ genes. The data matrix is publicly available at http://faculty.washington.edu/kayee/pca.

We performed PCA for the data set with $m = 4$. The solution shows $PEV_4 = 0.81$, which implies that 81% of the variances in 384 variables are explained by only four components. For the resulting *loading* matrix, we performed a *varimax rotation*, which is described in Chap. 13, for the following reason:

Note 5.5. Rotation of Components
If constraint (5.26) is removed and only (5.25) is considered, (5.1) can be rewritten as

$$\mathbf{X} = \mathbf{F}\mathbf{A}' + \mathbf{E} = \mathbf{F}\mathbf{T}\mathbf{T}'\mathbf{A}' + \mathbf{E} = \mathbf{F}_\mathrm{T}\mathbf{A}_\mathrm{T}' + \mathbf{E}. \tag{5.44}$$

Here,

$$\mathbf{F}_{\mathrm{T}} = \mathbf{FT} \quad \text{and} \quad \mathbf{A}_{\mathrm{T}} = \mathbf{AT}, \tag{5.45}$$

with $\mathbf{T}$ a special matrix satisfying $\mathbf{T}'\mathbf{T} = \mathbf{T}\mathbf{T}' = \mathbf{I}_m$, which is detailed in Appendix A.1.2. If $\mathbf{F}$ meets (5.25), $\mathbf{F}_{\mathrm{T}}$ also satisfies it:

$$\frac{1}{n}\mathbf{F}_{\mathrm{T}}{}'\mathbf{F}_{\mathrm{T}} = \frac{1}{n}\mathbf{T}'\mathbf{F}'\mathbf{F}\mathbf{T} = \mathbf{T}'\left(\frac{1}{n}\mathbf{F}'\mathbf{F}\right)\mathbf{T} = \mathbf{T}'(\mathbf{I}_m)\mathbf{T} = \mathbf{T}'\mathbf{T} = \mathbf{I}_m. \tag{5.46}$$

Equations (5.44) and (5.46) imply that if $\mathbf{F}$ and $\mathbf{A}$ are the PCA solution that minimizes (5.4) subject to (5.25), so are $\mathbf{F}_{\mathrm{T}}$ and $\mathbf{A}_{\mathrm{T}}$.

The above $\mathbf{T}$ can be chosen by the *rotation* techniques in Chap. 13, so that the resulting $\mathbf{A}_{\mathrm{T}}$ is easily interpreted.

The resulting loading matrix $\mathbf{A}_{\mathrm{T}}$ is of 384×4, which is too big to capture the values of its elements. Such a matrix can be effectively presented by a *heat map*, in which the largeness of the absolute values of each element is represented as the *depth of color* in the cell corresponding to each element. Figure 5.1 shows a heat map for the resulting loadings, block-wise. There, the blocks correspond to the five groups, into which the 384 genes are known to be categorized; each block is a matrix whose rows and columns are occupied by the genes in the corresponding group and the four components (C1–C4), respectively, though the genes in Group 2 are divided into two blocks. The solution is considered to be reasonable, as each phase has a unique feature of the loadings: the genes in Groups 1, 2, 4, and 5 positively load Components 1, 2, 3, and 4, respectively, while those in Group 3 positively load both Components 2 and 3.

5.9 Bibliographical Notes

Jolliffe (2002) exhaustively details various aspects of PCA. A subject that has not been treated in this book is the graphical *biplot* methodology for jointly representing the PC scores of individuals and the loadings of variables in a single configuration (Gower, Lubbe, & le Roux, 2011). The author of the present book has proposed a modified PCA procedure for easily capturing the biplot (Adachi, 2011).

A *three-way data array* is often observed whose element can be expressed as x_{ijk}, with i, j, and k standing for an individual, a variable, and an occasion, respectively, for example. The PCA formulation in this chapter can be modified to the approximation of the three-way data array by the reduced components, as introduced in Chap. 20.

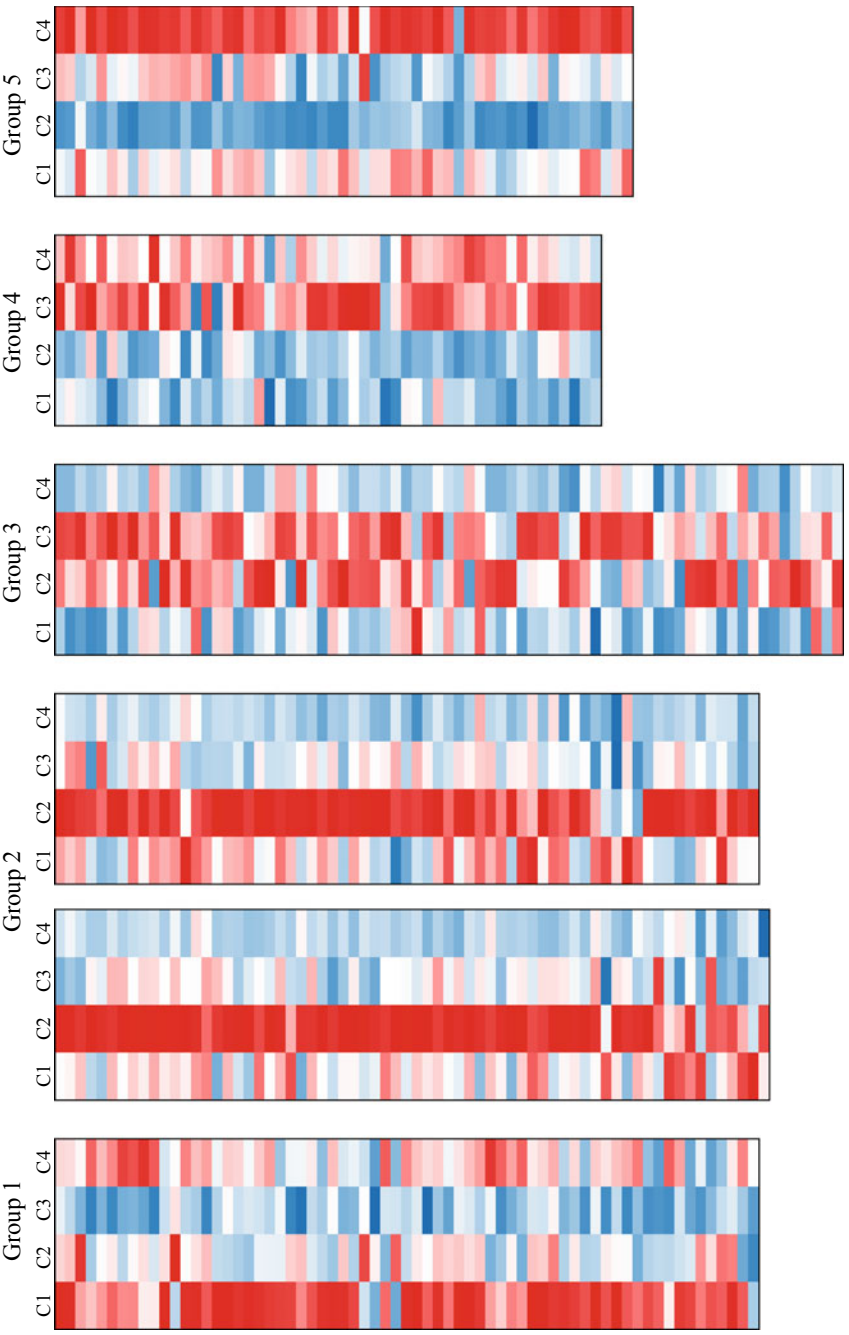


Fig. 5.1 Loadings for gene expression data with red and blue expressing positive and negative values, respectively

Exercises

5.1. Write Eqs. (5.5)–(5.8) ten times, using different characters for matrices, in order to *learn SVD by heart*.

5.2. Show that (5.5) can be rewritten as $\mathbf{X} = \tilde{\mathbf{K}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{L}}'$ for $n \geq p$. Here, $\tilde{\mathbf{K}}$ and $\tilde{\mathbf{L}}$ are the $n \times p$ and $p \times p$ matrices, respectively, satisfying $\tilde{\mathbf{K}}'\tilde{\mathbf{K}} = \tilde{\mathbf{L}}'\tilde{\mathbf{L}} = \mathbf{I}_p$, while $\tilde{\mathbf{\Lambda}}$ is a $p \times p$ diagonal matrix with $\tilde{\mathbf{\Lambda}} = \begin{bmatrix} \lambda_1 & & & & & \\ & \ddots & & & & \\ & & \lambda_r & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix}$.
This is the extended version of SVD in Appendix A.3.1.

5.3. Show that the error matrix $\mathbf{E} = \mathbf{X} - \mathbf{FA}'$ resulting in the minimization of (5.4) is expressed as $\mathbf{E} = \mathbf{K}_{[m]}\mathbf{\Lambda}_{[m]}\mathbf{L}_{[m]}'$, with its right-hand side defined as in (5.11) and the resulting PC scores are uncorrelated to the errors with $\mathbf{F}'\mathbf{E} = {}_m\mathbf{O}_p$.

5.4. Show that (5.14) can be rewritten as $\mathbf{FA}' = \lambda_1\mathbf{k}_1\mathbf{l}_1' + \cdots + \lambda_m\mathbf{k}_m\mathbf{l}_m'$ using (5.9).

5.5. Show that the problem in Note 5.3 would be trivial with its solution $\mathbf{FA}' = \mathbf{X}$, if $m \leq \mathrm{rank}(\mathbf{X})$ were not supposed.

5.6. Show that (5.27) and (5.28) must be replaced by $\mathbf{F} = \mathbf{K}_m\mathbf{\Lambda}_m^{1/2}$ and $\mathbf{A} = \mathbf{L}_m\mathbf{\Lambda}_m^{1/2}$, respectively, if constraint (5.25) was replaced by $\mathbf{F}'\mathbf{F} = \mathbf{A}'\mathbf{A}$,

5.7. Show that the SVD in Notes 5.1, 5.2, and 5.4 implies $\mathbf{K}_m = \mathbf{X}\mathbf{L}_m{\mathbf{\Lambda}_m}^{-1}$.

5.8. Discuss the similarities and differences between the loading matrix $\mathbf{A}$ and the weight matrix $\mathbf{W}$.

5.9. Show $PEV_m \leq PEV_{m+1}$ for (5.39).

5.10. Let us define

$$\mathbf{X}^+ = \mathbf{L}\mathbf{\Lambda}^{-1}\mathbf{K}' \tag{5.47}$$

for the matrix $\mathbf{X}$ whose SVD is defined in Note 5.1. Show that $\mathbf{X}^+$ satisfies $\mathbf{XX}^+\mathbf{X} = \mathbf{X}$, $\mathbf{X}^+\mathbf{XX}^+ = \mathbf{X}^+$, $(\mathbf{XX}^+)' = \mathbf{XX}^+$, and $(\mathbf{X}^+\mathbf{X})' = \mathbf{X}^+\mathbf{X}$. Matrix (5.47) is called the *Moore–Penrose inverse* of $\mathbf{X}$, as introduced in Chap. 17.

5.11. If $\mathbf{X}$ is nonsingular, show that its inverse matrix $\mathbf{X}^{-1}$ is a special case of the Moore–Penrose inverse $\mathbf{X}^+$ (5.47).

5.12. Show that the Moore–Penrose inverse (5.47) is defined for every matrix.

5.13. As with SVD and the Moore–Penrose inverse, *QR decomposition* is also defined for every matrix. Here, the QR decomposition of $\mathbf{A}$ $(p \times m)$ is expressed as $\mathbf{A} = \mathbf{QR}$, with $\mathbf{Q}'\mathbf{Q} = \mathbf{I}_m$ and the elements of $\mathbf{R} = (r_{jk})$ $(m \times m)$ being zero for $j > k$. Verify that

A			
–0.27	–1.74	1.24	1.58
–1.74	1.46	–2.57	–6.48
1.95	–4.00	4.85	0.09
0.81	1.20	–0.37	4.18
–1.14	–1.04	–0.02	–2.14
0.42	4.04	–3.04	2.82

=

Q			
–0.09	–0.32	–0.26	0.72
–0.58	0.05	–0.55	–0.41
0.65	–0.45	–0.13	–0.44
0.27	0.29	0.02	0.33
–0.38	–0.30	0.77	–0.10
0.14	0.72	0.16	–0.10

R			
3	–2	4	6
0	6	–5	3
0	0	0	2
0	0	0	5

represents a QR decomposition.

Chapter 6
Principal Component Analysis (Part 2)

In this chapter, principal component analysis (PCA) is reformulated. The loss function to be minimized is the same as that in the previous chapter, but the constraints for the matrices are different. This reformulation gives two purposes of PCA that were not found in the previous chapter. They are [1] forming a *weighted composite score* with the *maximum variance* and [2] *visualizing* a *high-dimensional invisible distribution* of individuals. In Sects. 6.1 and 6.2, the reformulation of PCA is mathematically described, followed by illustrations of the two purposes in Sects. 6.3, 6.4, and 6.5. Finally, a subject parallel to that in Sect. 5.7 is treated in Sect. 6.6.

6.1 Reformulation with Different Constraints

Let $\mathbf{X}$ denote an n-individuals $\times$ p-variables centered data matrix with $\mathbf{1}_n'\mathbf{X} = \mathbf{0}_p'$, as in the last chapter. As described there, PCA is formulated as minimizing (5.4), which is equivalent to (5.23), i.e., minimizing

$$f(\mathbf{W}, \mathbf{A}) = \|\mathbf{X} - \mathbf{F}\mathbf{A}'\|^2 = \|\mathbf{X} - \mathbf{X}\mathbf{W}\mathbf{A}'\|^2 \tag{6.1}$$

over *weight* matrix $\mathbf{W}$ and *loading* matrix $\mathbf{A}$ with $\mathbf{F} = \mathbf{X}\mathbf{W}$ containing PC scores. Using the *singular value decomposition* (*SVD*) in Notes 5.1 and 5.2, the solutions for $\mathbf{W}$ and $\mathbf{A}$ are expressed as (5.18) and (5.22), which are presented again here:

$$\mathbf{A} = \mathbf{L}_m \mathbf{\Lambda}_m^{1-\alpha} \mathbf{S}^{-1\prime}, \tag{6.2}$$

$$\mathbf{W} = \mathbf{L}_m \mathbf{\Lambda}_m^{\alpha-1} \mathbf{S}. \tag{6.3}$$

Here, α and $\mathbf{S}$ are arbitrary scalar and nonsingular matrices, respectively, which show that *infinitely many solutions* exist.

K. Adachi, *Matrix-Based Introduction to Multivariate Data Analysis*,
https://doi.org/10.1007/978-981-15-4103-2_6

To select a single solution among them, we consider the following constraints:

$$\mathbf{F}'\mathbf{F} = \mathbf{W}'\mathbf{X}'\mathbf{X}\mathbf{W} = \text{a diagonal matrix whose diagonal elements are arranged in descending order,} \tag{6.4}$$

$$\mathbf{W}'\mathbf{W} = \mathbf{I}_m, \tag{6.5}$$

which differ from constraints (5.25) and (5.26) in the last chapter. Then, the solution for $\mathbf{W}$ and $\mathbf{A}$ is expressed as

$$\mathbf{W} = \mathbf{A} = \mathbf{L}_m. \tag{6.6}$$

Both matrices are identical, which are given by (6.2) and (6.3) with $\alpha = 0$ and $\mathbf{S} = \mathbf{\Lambda}_m$. Obviously, (6.5) is satisfied by (6.6). This also allows $\mathbf{F} = \mathbf{X}\mathbf{W}$ to satisfy (6.4) as follows: (6.6) and (5.19) lead to $\mathbf{F} = \mathbf{X}\mathbf{W} = \mathbf{X}\mathbf{L}_m = \mathbf{K}_m\mathbf{\Lambda}_m$. This fact and $\mathbf{K}_m'\mathbf{K}_m = \mathbf{I}_m$ imply

$$\mathbf{F}'\mathbf{F} = \mathbf{W}'\mathbf{X}'\mathbf{X}\mathbf{W} = \mathbf{\Lambda}_m{}^2. \tag{6.7}$$

where the diagonal elements of $\mathbf{\Lambda}_m{}^2$ are in descending order, because of (5.8) and (5.12).

The identity of $\mathbf{W}$ to $\mathbf{A}$ in (6.6) shows that we may rewrite (6.1) as:

$$f(\mathbf{W}) = \|\mathbf{X} - \mathbf{X}\mathbf{W}\mathbf{W}'\|^2 \tag{6.8}$$

without $\mathbf{A}$.

6.2 Maximizing the Sum of Variances

Minimization of (6.8) subject to (6.4) and (6.5) is equivalent to maximizing

$$g(\mathbf{W}) = \mathrm{tr}\frac{1}{n}\mathbf{F}'\mathbf{F} = \frac{1}{n}\mathrm{tr}\mathbf{W}'\mathbf{X}'\mathbf{X}\mathbf{W} \tag{6.9}$$

subject to the same constraints. The equivalence is shown by expanding (6.8) as

$$\begin{aligned} f(\mathbf{W}) &= \mathrm{tr}\mathbf{X}'\mathbf{X} - 2\mathrm{tr}\mathbf{X}'\mathbf{X}\mathbf{W}\mathbf{W}' + \mathrm{tr}\mathbf{W}\mathbf{W}'\mathbf{X}'\mathbf{X}\mathbf{W}\mathbf{W}' \\ &= \mathrm{tr}\mathbf{X}'\mathbf{X} - 2\mathrm{tr}\mathbf{W}'\mathbf{X}'\mathbf{X}\mathbf{W} + \mathrm{tr}\mathbf{W}'\mathbf{X}'\mathbf{X}\mathbf{W}\mathbf{W}'\mathbf{W}. \end{aligned} \tag{6.10}$$

Using (6.5), the function (6.10) can be further rewritten as

$$f(\mathbf{W}) = \mathrm{tr}\mathbf{X}'\mathbf{X} - 2\mathrm{tr}\,\mathbf{W}'\mathbf{X}'\mathbf{X}\mathbf{W} + \mathrm{tr}\,\mathbf{W}'\mathbf{X}'\mathbf{X}\mathbf{W} = \mathrm{tr}\mathbf{X}'\mathbf{X} - \mathrm{tr}\mathbf{W}'\mathbf{X}'\mathbf{X}\mathbf{W}. \quad (6.11)$$

Here, we should note that only $-\mathrm{tr}\mathbf{W}'\mathbf{X}'\mathbf{X}\mathbf{W}$ is a function of $\mathbf{W}$ in the right-hand side. This implies that the minimization of $f(\mathbf{W})$ over $\mathbf{W}$ is equivalent to minimizing $-\mathrm{tr}\mathbf{W}'\mathbf{X}'\mathbf{X}\mathbf{W}$ or maximizing $\mathrm{tr}\mathbf{W}'\mathbf{X}'\mathbf{X}\mathbf{W}$. Further, this maximization is equivalent to $\mathrm{tr}\mathbf{W}'\mathbf{X}'\mathbf{X}\mathbf{W}$ divided by n, i.e., (6.9).

Thus, PCA can also be formulated as maximizing (6.9) subject to (6.4) and (6.5). Here, the matrix $n^{-1}\mathbf{F}'\mathbf{F}$ in (6.9) is the covariance matrix of PC scores between components, since $\mathbf{F} = \mathbf{X}\mathbf{W}$ is centered: $\mathbf{1}_n'\mathbf{X} = \mathbf{0}_p'$ leads to $\mathbf{1}_n'\mathbf{F} = \mathbf{0}_m'$. Thus, the diagonal elements of $n^{-1}\mathbf{F}'\mathbf{F}$ are the variances of m PC scores, implying that (6.9) is the sum of the *variances* of the 1st, … , mth *PC scores*:

$$g(\mathbf{W}) = \frac{1}{n}\mathbf{f}_1'\mathbf{f}_1 + \cdots + \frac{1}{n}\mathbf{f}_m'\mathbf{f}_m = \sum_{k=1}^{m}\left(\frac{1}{n}\mathbf{f}_k'\mathbf{f}_k\right), \quad (6.12)$$

with $\mathbf{F} = [\mathbf{f}_1, \ldots, \mathbf{f}_m]$.

We can also rewrite (6.9) as

$$g(\mathbf{W}) = \mathrm{tr}\mathbf{W}'\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)\mathbf{W} = \mathrm{tr}\mathbf{W}'\mathbf{V}\mathbf{W}, \quad (6.13)$$

where

$$\mathbf{V} = \frac{1}{n}\mathbf{X}'\mathbf{X} \quad (6.14)$$

is the covariance matrix for centered $\mathbf{X}$. In some books, PCA is introduced with the following decomposition:

Note 6.1. Eigenvalue Decomposition of a Covariance Matrix
The *singular value decomposition* $\mathbf{X} = \mathbf{K}\boldsymbol{\Lambda}\mathbf{L}'$ in (5.5) with (5.6) and (5.7) leads to $\mathbf{X}'\mathbf{X} = \mathbf{L}\boldsymbol{\Lambda}^2\mathbf{L}'$. Comparing it with $\mathbf{X}'\mathbf{X} = n\mathbf{V}$ following from (6.14), we have $n\mathbf{V} = \mathbf{L}\boldsymbol{\Lambda}^2\mathbf{L}'$. This equation can be rewritten as

$$\mathbf{V} = \mathbf{L}\boldsymbol{\Delta}\mathbf{L}'. \quad (6.15)$$

Here,

$$\boldsymbol{\Delta} = \begin{bmatrix} \delta_1 & & \\ & \ddots & \\ & & \delta_r \end{bmatrix} = \frac{1}{n}\boldsymbol{\Lambda}^2, \quad (6.16)$$

with $r = \mathrm{rank}(\mathbf{X})$ and $\delta_1 \geq \cdots \geq \delta_r \geq 0$. Decomposition (6.15) is referred to as the *eigenvalue decomposition* (*EVD*) or *spectral decomposition* of $\mathbf{V}$, δ_k ($k = 1, \ldots, r$) is called the kth largest *eigenvalue* of $\mathbf{V}$, and the kth column of $\mathbf{L}$ is called the *eigenvector* of $\mathbf{V}$ corresponding to δ_k.

6.3 Weighted Composite Scores with Maximum Variance

Let us express the columns of a data matrix as $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_p]$. An example of $\mathbf{X}$ with $n = 9$ (examinees) and $p = 3$ (tests) is given in Table 6.1(B), which contains the centered scores of the raw ones in (A). They are the scores of the entrance examinations for a company. The examinations consist of the following items:

ES: essay,

IN: interview,

PR: presentation,

which define the three variables in $\mathbf{X}$.

We perform PCA for this data set with the number of components m equaling one, i.e., $\mathbf{W} = \mathbf{w}_1$ ($p \times 1$) and $\mathbf{F} = \mathbf{f}_1 = \mathbf{X}\mathbf{w}_1$ ($n \times 1$) being vectors. By defining $\mathbf{w}_1 = [w_{11}, \ldots, w_{p1}]'$, the PC score vector $\mathbf{f}_1$ is written as

$$\mathbf{f}_1 = \mathbf{X}\mathbf{w}_1 = w_{11}\mathbf{x}_1 + \cdots + w_{p1}\mathbf{x}_p = w_{11}\mathrm{ES} + w_{21}\mathrm{IN} + w_{31}\mathrm{PR}, \tag{6.17}$$

Table 6.1 Scores for an entrance examination and its PCA scores, which are artificial examples found in Adachi (2006)

Examinee	(A) Raw scores			(B) Centered scores			(C) PC scores		
	ES	IN	PR	ES	IN	PR	1st	2nd	3rd
1	88	70	65	21.2	4.3	−3.0	10.8	−19.0	0.6
2	52	78	88	−14.8	12.3	20.0	13.3	24.3	−1.8
3	77	87	89	10.2	21.3	21.0	31.3	4.7	−1.4
4	35	40	43	−31.8	−25.7	−25.0	−46.5	10.9	−3.3
5	60	43	40	−6.8	−22.7	−28.0	−34.8	−11.4	−0.7
6	97	95	91	30.2	29.3	23.0	46.9	−10.3	−1.1
7	48	62	83	−18.8	−3.7	15.0	−1.8	23.4	6.5
8	66	66	65	−0.8	0.3	−3.0	−2.0	−0.9	−2.2
9	78	50	48	11.2	−15.7	−20.0	−17.1	−21.6	3.4
Average	66.8	65.7	68.0	0.0	0.0	0.0	0.0	0.0	0.0
Variance	358.0	324.2	380.2	358.0	324.2	380.2	793.3	260.4	8.5

with the abbreviations for the variables in Table 6.1 used in the right-hand side. Here, $\mathbf{f}_1$ is found to contain the *weighted composite scores* for the examinees, i.e., the sum of the data in $\mathbf{x}_j$ weighted by w_{j1} over $j = 1, \ldots, p$.

Using $\mathbf{W} = \mathbf{w}_1$ and $\mathbf{F} = \mathbf{f}_1 = \mathbf{X}\mathbf{w}_1$, the function (6.9) or (6.12) is rewritten as

$$g(\mathbf{w}_1) = \mathbf{w}_1' \left(\frac{1}{n} \mathbf{X}'\mathbf{X} \right) \mathbf{w}_1 = \mathbf{w}_1' \mathbf{V} \mathbf{w}_1 = \frac{1}{n} \mathbf{f}_1' \mathbf{f}_1. \tag{6.18}$$

This stands for the variance of the weighted composite scores in (6.17); their variance is defined as $n^{-1}\mathbf{f}'_1\mathbf{J}\mathbf{f}_1 = \mathbf{w}_1'(n^{-1}\mathbf{X}'\mathbf{J}\mathbf{X})\mathbf{w}_1 = \mathbf{w}_1'(n^{-1}\mathbf{X}'\mathbf{X})\mathbf{w}_1 = n^{-1}\mathbf{f}_1'\mathbf{f}_1$, since $\mathbf{X}$ is centered with $\mathbf{X} = \mathbf{J}\mathbf{X}$. This variance is to be maximized subject to (6.5), i.e., $\mathbf{w}_1'\mathbf{w}_1 = 1$ for $m = 1$ (where (6.4) may not be considered for $m = 1$, since $\mathbf{F}'\mathbf{F} = \mathbf{f}_1'\mathbf{f}_1$ is a single scalar). That is, the PC scores in $\mathbf{f}_1$ are the composite scores obtained by *weighting the variables* so that the *variance of the scores is maximized*, in other words, so that *individuals are best distinguished*.

PCA for the data set in Table 6.1(B) provides

$$\mathbf{w}_1 = [0.47, 0.63, 0.62]', \tag{6.19}$$

which implies that

$$\text{PC score} = 0.47\,\text{ES} + 0.63\,\text{IN} + 0.62\,\text{PR} \tag{6.20}$$

is to be obtained for each examinee. For example, the centered scores for the second examinee are −14.8, 12.3, and 20.0, thus, that examinee's first PC score is obtained as

$$0.47 \times (-14.8) + 0.63 \times 12.3 + 0.62 \times 20.0 = 13.3. \tag{6.21}$$

The PC scores computed for all examinees in this way are shown in the first column of Table 6.1(C).

In everyday life, we often use a composite score:

$$\text{Simple Sum Score} = \mathbf{x}_1 + \cdots + \mathbf{x}_p = \text{ES} + \text{IN} + \text{PR}, \tag{6.22}$$

i.e., the sum of the equally weighted variables. As compared to this score, the PC score (6.20) is more useful for distinguishing individuals.

6.4 Projecting Three-Dimensional Vectors onto Two-Dimensional Ones

Though the maximization of (6.9) (for $m = 1$) was considered in the last section, the purpose of this section is to explain that the minimization of (6.8) implies *projecting* a three-dimensional (3D) space onto one that is two-dimensional (2D), for $p = 3$, as in Table 6.1(B), and $m = 2$. For that purpose, let us use $\mathbf{F} = \mathbf{XW}$ in (6.8) to rewrite it as

$$f(\mathbf{W}) = \|\mathbf{X} - \mathbf{XWW}'\|^2 = \|\mathbf{X} - \mathbf{FW}'\|^2. \tag{6.8'}$$

Further, we use row vector $\tilde{\mathbf{x}}_i'(1 \times p)$ for the data vector of individual i, and $\tilde{\mathbf{f}}_i'(1 \times m)$ for the PC score vector of i:

$$\mathbf{X} = \begin{bmatrix} \tilde{\mathbf{x}}_1' \\ \vdots \\ \tilde{\mathbf{x}}_i' \\ \vdots \\ \tilde{\mathbf{x}}_n' \end{bmatrix}, \quad \mathbf{F} = \begin{bmatrix} \tilde{\mathbf{f}}_1' \\ \vdots \\ \tilde{\mathbf{f}}_i' \\ \vdots \\ \tilde{\mathbf{f}}_n' \end{bmatrix} = \mathbf{XW} = \begin{bmatrix} \tilde{\mathbf{x}}_1'\mathbf{W} \\ \vdots \\ \tilde{\mathbf{x}}_i'\mathbf{W} \\ \vdots \\ \tilde{\mathbf{x}}_n'\mathbf{W} \end{bmatrix} \tag{6.23}$$

with $\mathbf{W}$ being $p \times m$. Then, the rows of $\mathbf{FW}' = \mathbf{XWW}'$ in (6.8′) are expressed as

$$\mathbf{FW}' = \begin{bmatrix} \tilde{\mathbf{f}}_1'\mathbf{W}' \\ \vdots \\ \tilde{\mathbf{f}}_i'\mathbf{W}' \\ \vdots \\ \tilde{\mathbf{f}}_n'\mathbf{W}' \end{bmatrix} = \mathbf{XWW}' = \begin{bmatrix} \tilde{\mathbf{x}}_1'\mathbf{WW}' \\ \vdots \\ \tilde{\mathbf{x}}_i'\mathbf{WW}' \\ \vdots \\ \tilde{\mathbf{x}}_n'\mathbf{WW}' \end{bmatrix}. \tag{6.24}$$

Using (6.23) and (6.24) in (6.8′), this is rewritten as

$$f(\mathbf{W}) = \left\| \begin{bmatrix} \tilde{\mathbf{x}}_1' \\ \vdots \\ \tilde{\mathbf{x}}_i' \\ \vdots \\ \tilde{\mathbf{x}}_n' \end{bmatrix} - \begin{bmatrix} \tilde{\mathbf{x}}_1'\mathbf{WW}' \\ \vdots \\ \tilde{\mathbf{x}}_i'\mathbf{WW}' \\ \vdots \\ \tilde{\mathbf{x}}_n'\mathbf{WW}' \end{bmatrix} \right\|^2 = \left\| \begin{bmatrix} \tilde{\mathbf{x}}_1' \\ \vdots \\ \tilde{\mathbf{x}}_i' \\ \vdots \\ \tilde{\mathbf{x}}_n' \end{bmatrix} - \begin{bmatrix} \tilde{\mathbf{f}}_1'\mathbf{W}' \\ \vdots \\ \tilde{\mathbf{f}}_i'\mathbf{W}' \\ \vdots \\ \tilde{\mathbf{f}}_n'\mathbf{W}' \end{bmatrix} \right\|^2. \tag{6.8''}$$

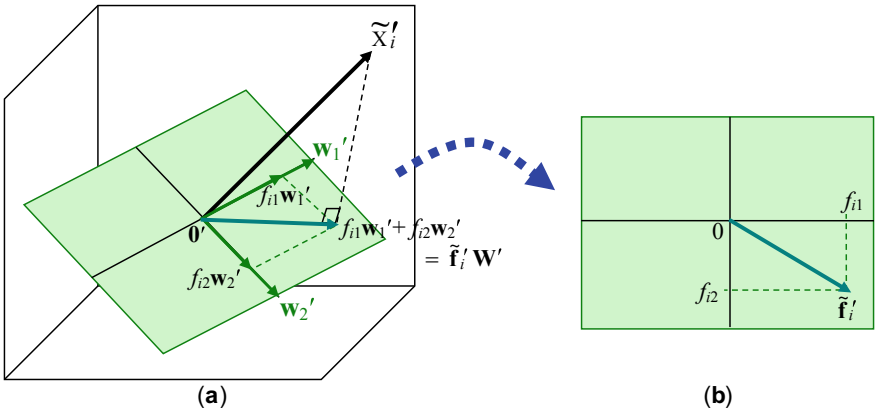


Fig. 6.1 The projection of a data vector onto a plane (**a**) with its front view (**b**)

When $p = 3$ and $m = 2$, the minimization of (6.8″) amounts to matching individuals' data vectors $\tilde{\mathbf{x}}_i' = [x_{i1}, x_{i2}, x_{i3}]$ to $\tilde{\mathbf{x}}_i'\mathbf{W}\mathbf{W}' = \tilde{\mathbf{f}}_i'\mathbf{W}'(1 \times 3)$, which can be expressed as

$$\tilde{\mathbf{x}}_i'\mathbf{W}\mathbf{W}' = \tilde{\mathbf{f}}_i'\mathbf{W}' = [f_{i1}, f_{i2}]\begin{bmatrix}\mathbf{w}_1' \\ \mathbf{w}_2'\end{bmatrix} = f_{i1}\mathbf{w}_1' + f_{i2}\mathbf{w}_2', \tag{6.25}$$

with $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2]$ (3×2) and $\tilde{\mathbf{f}}_i' = [f_{i1}, f_{i2}]$. A key point involves capturing what (6.25) geometrically stands for. This is explained in the following two paragraphs.

As the data vector $\tilde{\mathbf{x}}_i' = [x_{i1}, x_{i2}, x_{i3}]$ is 1×3, $\tilde{\mathbf{x}}_i'$ can be depicted in a 3D space, as in Fig. 6.1a; $\tilde{\mathbf{x}}_i'$ is the line extending to the coordinate $[x_{i1}, x_{i2}, x_{i3}]$. There, we can also depict a *plane* whose direction in the 3D space is defined by vectors $\mathbf{w}_1'$ and $\mathbf{w}_2'$. As found there, the *projection* of $\tilde{\mathbf{x}}_i'$ on the *plane* is expressed as (6.25), where the *projection* refers to the vector that extends to the *intersection of the plane and the line drawn from* $\tilde{\mathbf{x}}_i'$, *vertical to the plane*. Further, the PC scores in $\tilde{\mathbf{f}}_i' = [f_{i1}, f_{i2}] = \tilde{\mathbf{x}}_i'\mathbf{W}$ stand for the *coordinates* of the projection within the *plane*. Why this fact holds is explained in Appendix A.1.4. The *plane* seen head-on is shown in Fig. 6.1b. There, the first and second PC scores in $[f_{i1}, f_{i2}]'$ are the *coordinates* on the horizontal and vertical axes of the plane. Below, we note the difference in this plane compared with that used in Chap. 4:

Note 6.2. Differences from Fig. 4.2
The plane in Fig. 4.2 differs from the one in Fig. 6.1a and the remaining ones in this chapter, in that *variable* vectors extend on the plane in Fig. 4.2, while *individuals*' vectors extend/are distributed on the planes in the figures appearing in this chapter.

We can freely define spaces (i.e., planes); which spaces are to be considered depends on one's research interests.

Now, let us recall function (6.8″), which is minimized over $\mathbf{W} = [\mathbf{w}_1', \mathbf{w}_2']'$ in PCA. This minimization implies bringing the projection (6.25) as close to $\tilde{\mathbf{x}}_i'$ as possible for all $i = 1, \ldots, n$. In other words, one purpose of PCA is to find the matrix $\mathbf{W} = [\mathbf{w}_1', \mathbf{w}_2']'$ that defines the direction of the plane so that the *projections* (6.25) are *closest* to the *original data vectors* $\tilde{\mathbf{x}}_i'$. The plane obtained by PCA is thus called a *best-fitting plane*, since it is closest, i.e., the best fitted to the data vectors

We illustrate the above case with the data in Table 6.1(B), whose data vectors $\tilde{\mathbf{x}}_i'(i = 1, \ldots, 9)$ can be depicted as in Fig. 6.2a. Here, the endpoints of the vectors $\tilde{\mathbf{x}}_i'$ have been indicated by circles (not by lines as in Fig. 6.1a) for the sake of ease in viewing. For the data set, PCA provides the solution of $\mathbf{W} = [\mathbf{w}_1', \mathbf{w}_2']'$ with $\mathbf{w}_1$ given by (6.19) and

$$\mathbf{w}_2 = [-0.84, 0.11, 0.53]'. \tag{6.26}$$

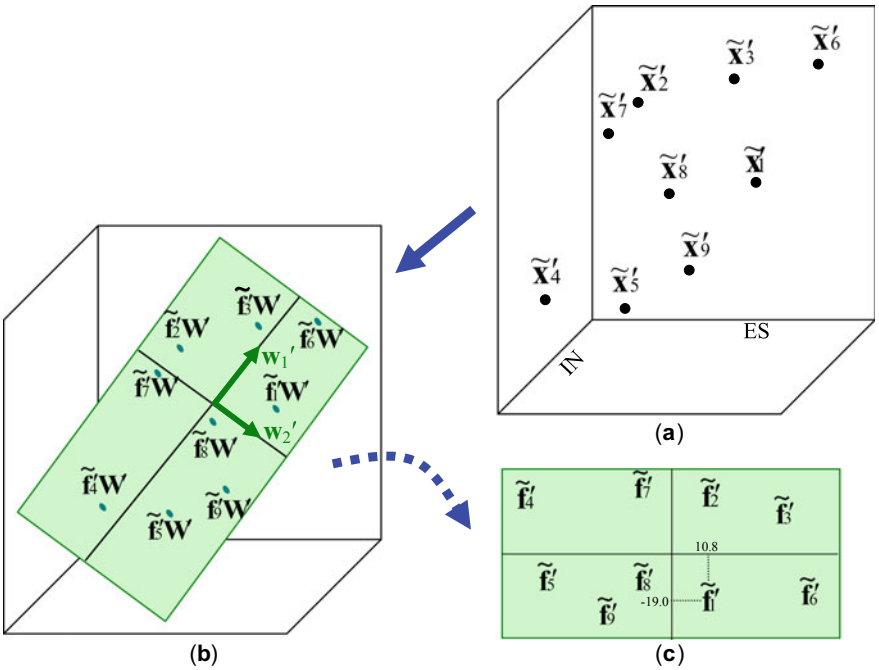


Fig. 6.2 Projections of data vectors (**a**) on a plane (**b**) with its front view (**c**)

These vectors define the *best-fitting plane* in Fig. 6.2b, on which the projections $\tilde{\mathbf{f}}_i' \mathbf{W}' (i = 1, \ldots, 9)$ for data vectors $\tilde{\mathbf{x}}_i'$ exist. A head-on view of the plane is shown in Fig. 6.2c. Here, the coordinates of the points are the first and second PC scores in $\tilde{\mathbf{f}}_i' = [f_{i1}, f_{i2}]$, whose values are presented in Table 6.1(C). For example, the PC score vector for Examinee 1 is found to be $\tilde{\mathbf{f}}_1' = [10.8, -19.0]$ in the table, and it is located at the point with the coordinates [10.8, −19.0] in Fig. 6.2c. Here, the second PC score, −19.0, has been obtained as $f_{12} = -0.84 \times 21.2 + 0.11 \times 4.3 + 0.53 \times (-3.0) = -19.0$ using her/his centered scores [21.2, 4.3, −3.0] and the weights in (6.26).

This section deals with the logic in PCA by which the *original 3D* data distributions (as in Fig. 6.2a) are *projected* on a *2D* plane (as in (b)), whose front view is a scatter plot (as in (c)). This 2D plot is useful, in that it is easier to capture than the original 3D plot. However, this section is merely a preparation for the one that follows, where the distributions in the space of a higher dimension can be projected onto a lower-dimensional space in the same way as in this section. It is one of the most important benefits gained by using PCA.

6.5 Visualization of Invisible Distributions

We consider a new data set in Table 6.2 (Committee for Guiding Psychological Experiments, 1985). It contains the results of the rating by participants for to what extent 12 adjectives characterize 14 occupational categories. The table shows the average rating values for the categories on a scale of 1–5. For example, let us note the final column "busy": the busyness of "bank clerk" is rated at 4.2, while that of "professor" is 3.0, that is, people think that bank clerks are busier than professors.

Let $\mathbf{X} = (x_{ij})$ (14 × 12) contain the centered scores of the data in Table 6.2. For example, x_{32} is 3.2–3.7 = −0.5 (the usefulness of "cartoonist" minus the average of usefulness). Can we depict the distribution of the 14 categories' scores on the 12 variables? That would require a 12-dimensional (12D) space with its 12 coordinate axes orthogonally intersected. Unfortunately, a space of dimensionality $m > 3$ can *neither be drawn nor seen* by us, as *we live in a 3D world!* However, such a *high-dimensional space* can be considered in logic, i.e., mathematically, regardless of how high the dimensionality is.

Let us suppose that $\tilde{\mathbf{x}}_i' = [x_{i1}, \ldots, x_{i,12}]$ (i = 1, … , 14; categories) are distributed in a 12D space as depicted in Fig. 6.3a. PCA for $\mathbf{X}$ yields the weight matrix $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2]$ in Table 6.3(A). It defines the *best-fitting plane* on which the projections $\tilde{\mathbf{f}}_i' \mathbf{W}' (i = 1, \ldots, 14)$ are located, as illustrated in Fig. 6.3b. This plane can be seen head-on, as shown in Fig. 6.4. There, the 14 categories are plotted, with their coordinates being the PC scores $\tilde{\mathbf{f}}_i' = [f_{i1}, f_{i2}]$, whose values are obtained as in Table 6.3(B) using the centered scores for Table 6.2 and the weights in Table 6.3(A).

Table 6.2 Impressions of 14 occupational categories rated for 12 adjectives

Category	Noble	Useful	Good	Big	Powerful	Strong	Quick	Noisy	Young	Faithful	Strict	Busy
Monk	3.2	2.7	3.7	2.8	2.6	2.6	2.2	1.4	1.7	3.3	3.8	1.8
Bank clerk	3.4	3.5	3.4	2.5	2.2	2.6	3.2	2.1	3.6	4.1	4.7	4.2
Cartoonist	3.0	3.2	3.5	2.2	2.1	2.2	3.3	3.4	4.1	3.4	1.3	4.3
Designer	3.2	3.2	3.5	2.6	2.5	2.6	3.6	2.9	4.2	3.2	1.5	4.0
Nurse[a]	4.2	4.6	4.5	3.1	3.0	3.2	2.8	3.3	4.1	4.5	2.3	4.9
Professor	4.0	4.0	3.8	3.4	3.2	3.1	2.4	1.5	1.6	3.7	3.9	3.0
Doctor[b]	4.0	4.8	3.9	3.5	3.8	3.7	3.2	2.1	2.6	3.7	3.6	4.5
Policeman	3.7	4.6	4.1	3.4	4.0	4.1	4.3	3.4	3.5	4.2	4.4	4.0
Journalist	3.6	4.3	3.7	2.9	3.5	3.6	4.7	4.2	4.1	3.9	3.7	5.0
Sailor	3.6	3.6	3.5	3.5	4.2	4.2	3.5	3.5	3.7	3.5	2.5	3.5
Athlete	3.7	3.2	3.7	3.9	4.7	4.7	4.9	3.5	4.2	3.7	2.8	4.1
Novelist	3.4	3.7	3.5	3.1	2.7	2.4	2.3	1.8	2.3	3.3	2.9	3.3
Actor	3.2	3.2	3.6	2.9	2.2	2.5	3.3	3.3	3.4	2.8	1.8	4.3
Stewardess	3.2	3.8	3.8	2.8	2.3	2.4	3.9	2.5	4.7	3.9	2.3	4.3
Average[c]	3.5	3.7	3.7	3.0	3.1	3.1	3.4	2.8	3.4	3.7	3.0	3.9

[a]In nursing school
[b]Medical doctor
[c]Column average

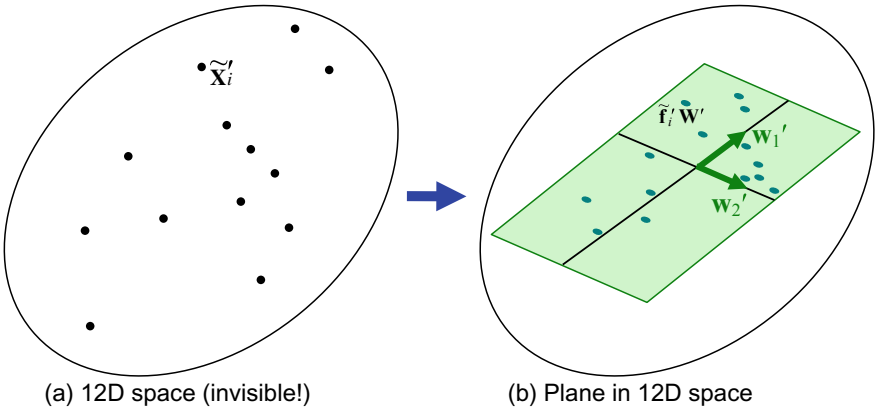


Fig. 6.3 Projecting the distributions in a 12D space (**a**) on a 2D plane (**b**)

Table 6.3 Weights and PC scores obtained for the centered scores transformed from the data in Table 6.2

(A) W (weights)			(B) **F** (PC scores)		
	$\mathbf{w}_1$	$\mathbf{w}_2$		$\mathbf{f}_1$	$\mathbf{f}_2$
Noble	0.03	0.18	Monk	−3.46	0.27
Useful	0.12	0.23	Bank clerk	−0.91	0.19
Good	0.04	0.07	Cartoonist	0.43	−2.57
Big	0.06	0.25	Designer	0.43	−1.93
Powerful	0.22	0.46	Nurse	1.06	−0.25
Strong	0.26	0.42	Professor	−2.41	1.40
Quick	0.44	0.09	Doctor	−0.24	1.60
Noisy	0.48	−0.07	Policeman	1.13	2.11
Young	0.50	−0.27	Journalist	2.16	0.72
Faithful	0.09	0.15	Sailor	0.95	0.67
Strict	−0.19	0.59	Athlete	2.24	1.30
Busy	0.39	−0.09	Novelist	−2.06	−0.29
			Actor	0.04	−1.78
			Stewardess	0.65	−1.43

Although the original distribution of $\tilde{\mathbf{x}}_i'$ in Fig. 6.3a was *invisible*, the projection of $\tilde{\mathbf{x}}_i'$ on the best-fitting 2D plane is *visible*, as found in Fig. 6.4. This shows that a benefit of PCA is the *visualization* of a *high-dimensional invisible space*. The resulting plot in Fig. 6.4 can be captured in the same manner as for a usual *map*; two objects close to each other can be viewed as similar, while those that are distant

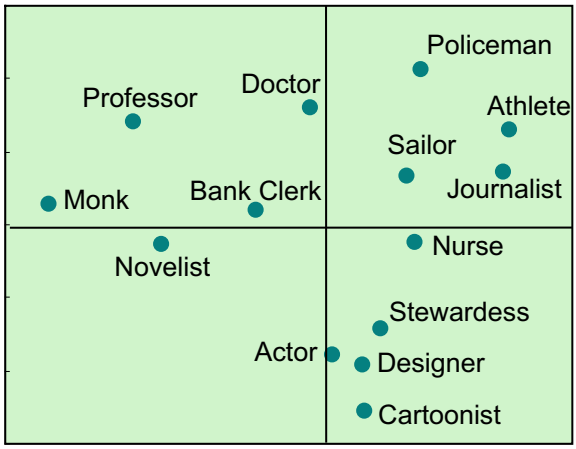


Fig. 6.4 Front view of the plane in Fig. 6.3b

can be regarded as dissimilar. For example, Fig. 6.4 shows that "designer" and "cartoonist" are similar occupations, while "monk" and "journalist" are very different.

6.6 Goodness of Projection

It should be noticed that the original distribution in Fig. 6.3a is not perfectly reflected on the plane in Fig. 6.3b, which in turn gives Fig. 6.4; some information in the original distribution has been *lost* in Figs. 6.3b and 6.4. The *amount of the loss* can be assessed by the resulting value of loss function (6.8) or (6.8″), since it expresses the differences between the data vectors $\tilde{\mathbf{x}}_i'(i = 1, \ldots, n)$ in Fig. 6.3a and their projections $\tilde{\mathbf{f}}_i' \mathbf{W}'(1 \times 3)$ in (b).

The resulting value of (6.8), into which solution (6.6) is substituted, is expressed as

$$\left\|\mathbf{X} - \mathbf{X}\mathbf{L}_m\mathbf{L}_m'\right\|^2 = \left\|\mathbf{X} - \mathbf{K}_m\mathbf{\Lambda}_m\mathbf{L}_m'\right\|^2 = \left\|\mathbf{K}\mathbf{\Lambda}\mathbf{L}' - \mathbf{K}_m\mathbf{\Lambda}_m\mathbf{L}_m'\right\|^2, \tag{6.27}$$

where we have used (5.5) and (5.19). It can be found that (6.27) is equivalent to (5.37), which implies that the *proportion of explained variance* (5.39), i.e., $PEV_m = \mathrm{tr}\mathbf{\Lambda}_m{}^2/\mathrm{tr}\mathbf{\Lambda}^2 = \mathrm{tr}\mathbf{\Lambda}_m{}^2/\|\mathbf{X}\|^2$, is also an index for the *goodness of projection*.

For the centered scores of the data in Table 6.2, PCA gives $\mathrm{tr}\mathbf{\Lambda}_2{}^2 = 64.1$ and $\mathrm{tr}\mathbf{\Lambda}^2 = \|\mathbf{X}\|^2 = 86.6$, thus $PEV_2 = 64.1/86.6 = 0.74$. This implies that 74% (=0.74 × 100%) of the information of the distribution in Fig. 6.3a is reflected in Fig. 6.4; the former invisible distribution is visualized in the latter and, furthermore, we can see 74% of the former. This demonstrates the benefit of PCA.

6.7 Bibliographical Notes

As described in Sect. 5.9, various aspects of PCA are exhaustively detailed in Jolliffe (2002). *Visualization* as a benefit of PCA in the natural sciences has been illustrated in Izenman (2008) and Koch (2014), which are among the advanced books recommended for a deeper understanding of multivariate analysis, though the term visualization is not used in those books.

Here, we must mention *sparse PCA* for obtaining the sparse weight matrix $\mathbf{W}$ (Jolliffe et al., 2003; Zou et al., 2006). Here, a sparse matrix refers to a matrix including a number of zero elements. That is, sparse PCA refers to the modified PCA procedures in which the elements of $\mathbf{W}$ to be zero are computationally chosen jointly with estimating the values of the nonzero elements. The resulting $\mathbf{W}$ can be easily interpreted, as only nonzero elements may be noted. The procedures related to the sparse PCA would be treated in Chap. 22.

Exercises

6.1. Show $\mathrm{tr}\mathbf{V} = \mathrm{tr}\boldsymbol{\Delta}$ for $\mathbf{V}$ and $\boldsymbol{\Delta}$ in Note 6.1.

6.2. Show that the eigenvalue decomposition (EVD) in Note 6.1 implies $\mathbf{V}\mathbf{l}_k = \delta_k \mathbf{l}_k$ ($k = 1, \ldots, r$) with $\mathbf{l}_k$ the kth column of $\mathbf{L}$. The equation is called the *eigen equation* for $\mathbf{V}$.

6.3. Show that the EVD in Note 6.1 can be rewritten as $\mathbf{X}'\mathbf{X} = \mathbf{L}\boldsymbol{\Lambda}^2\mathbf{L}'$ and post-multiplying its both sides by $\mathbf{L}\mathbf{L}'$ leads to $\mathbf{X}'\mathbf{X}\mathbf{L}\mathbf{L}' = \mathbf{X}'\mathbf{X}$, i.e.,

$$\mathbf{X}'\mathbf{X}\left(\mathbf{I}_p - \mathbf{L}\mathbf{L}'\right) =_p \mathbf{O}_p. \tag{6.28}$$

6.4. Show that (6.28) leads to $(\mathbf{I}_p - \mathbf{L}\mathbf{L}')\mathbf{X}'\mathbf{X}(\mathbf{I}_p - \mathbf{L}\mathbf{L}') = {}_p\mathbf{O}_p$, which implies

$$\mathbf{X}\left(\mathbf{I}_p - \mathbf{L}\mathbf{L}'\right) =_n \mathbf{O}_p, \tag{6.29}$$

using the fact that $\mathbf{M}'\mathbf{M} = {}_p\mathbf{O}_p$ implies $\mathbf{M} = {}_n\mathbf{O}_p$ for $\mathbf{M}$ being $n \times p$.

6.5. Show that the *SVD* in Note 5.1 can be derived from the *EVD* in Note 6.1, noting the fact that (6.29) implies $\mathbf{X} = \mathbf{X}\mathbf{L}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}\mathbf{L}'$ and $\mathbf{X}\mathbf{L}\boldsymbol{\Lambda}^{-1}$ can be regarded as $\mathbf{K}$ in Note 5.1.

6.6. A square matrix $\mathbf{N}$ is said to be *nonnegative definite* if $f(\mathbf{w}) = \mathbf{w}'\mathbf{N}\mathbf{w} \geq 0$ for any vector $\mathbf{w}$. It is known that $\mathbf{S}$ being nonnegative definite and symmetric is equivalent to the property of $\mathbf{S}$ that it can be rewritten as $\mathbf{S} = \mathbf{B}\mathbf{B}'$. Show that the covariance matrix $\mathbf{V} = n^{-1}\mathbf{X}'\mathbf{J}\mathbf{X}$ is nonnegative definite.

6.7. A square matrix $\mathbf{P}$ is said to be *positive definite*, if $f(\mathbf{w}) = \mathbf{w}'\mathbf{P}\mathbf{w} > 0$ for any vector $\mathbf{w}$ other than the zero vector. Show that a diagonal matrix $\mathbf{D}$ being positive definite is equivalent to all diagonal elements of $\mathbf{D}$ being positive.

6.8. Let $v(\mathbf{f}_k)$ denote the variance of the kth PC scores, i.e., the elements in $\mathbf{f}_k = \mathbf{X}\mathbf{w}_k$. Show that $v(\mathbf{f}_k)$ equals δ_k, i.e., the kth eigenvalue of $\mathbf{V}$ defined in Note 6.1.

6.9. Show that the vectors $\tilde{\mathbf{f}}_i'\mathbf{W}'$ and $\tilde{\mathbf{x}}_i' - \tilde{\mathbf{f}}_i'\mathbf{W}'$ intersect orthogonally, as in Fig. 6.1, i.e., $\tilde{\mathbf{f}}_i'\mathbf{W}'(\tilde{\mathbf{x}}_i - \mathbf{W}\tilde{\mathbf{f}}_i) = 0$.

6.10. Show that (6.6) is replaced by $\mathbf{W} = \mathbf{A} = \mathbf{L}_m\mathbf{T}$, with $\mathbf{T}$ the orthonormal matrix satisfying (A.1.6) in Appendix A.1.2, if constraint (6.4) is removed and only (6.5) is imposed in PCA.

Chapter 7
Cluster Analysis

The term "*cluster*" is synonymous with both "*group*" as a noun and "*classify*" as a verb. *Cluster analysis*, which is also simply called *clustering*, generally refers to the procedures for computationally classifying (i.e., clustering) individuals into groups (i.e., clusters) so that similar individuals are classified into the same group and mutually dissimilar ones are allocated to different groups. There are various procedures for performing cluster analysis. One of the most popular of these, called *k-means clustering* (*KMC*), which was first presented by MacQueen (1967), is introduced here.

7.1 Membership Matrices

An example of a membership matrix is given here:

$\mathbf{G} = (g_{ik}) =$

	Australia	UK	USA
Mick		1	
Kieth	1		
Ronny			1
Charly		1	
Bill			1

It indicates the nationalities of individuals, and the blank cells stand for the elements taking zero. In general, a *membership matrix* $\mathbf{G} = (g_{ik})$ is defined as the matrix of *n individuals* $\times$ *K-clusters* satisfying

$$g_{ik} = \begin{cases} 1 & \text{if individual } i \text{ belongs to cluster } k \\ 0 & \text{otherwise} \end{cases}, \tag{7.1}$$

$$\mathbf{G}\mathbf{1}_K = \mathbf{1}_n. \tag{7.2}$$

K. Adachi, *Matrix-Based Introduction to Multivariate Data Analysis*,
https://doi.org/10.1007/978-981-15-4103-2_7

These equations imply that each row of **G** has only one element taking 1, i.e., each individual belongs to only one cluster. Such a matrix is also called an *indicator matrix* or a *design matrix*. A major purpose of clustering procedures including *k-means clustering* (*KMC*) is to obtain **G** from an n-individuals $\times$ p-variables data matrix **X**.

7.2 Example of Clustering Results

For a data matrix **X**, KMC provides a membership matrix **G** together with a K-clusters $\times$ p-variables *cluster feature matrix* **C**, which expresses how each cluster is characterized by variables.

Before explaining how to obtain **G** and **C**, we show the KMC solution for the 14-occupations $\times$ 12-adjectives data matrix $\mathbf{X} = (x_{ij})$ in Table 6.2 in the last chapter. It describes to what extent the occupations are characterized by the adjectives. For the data matrix, KMC with K set at 4 provides the solutions of **G** and **C** shown in Tables 7.1 and 7.2. First, let us note the resulting *membership* matrix **G** in Table 7.1. The cluster numbers 1, 2, 3, and 4 are merely for the purpose of distinguishing different clusters; **G** simply shows that the occupations having 1 in the same column belong to the same cluster. For example, *monk*, *professor*, and *novelist* are members of a cluster, while *policeman*, *journalist*, *sailor*, and *athlete* are members of another cluster. Next, let us note Table 7.2. There, the resulting *cluster feature* matrix **C** is shown, which describes the values of variables characterizing each cluster. For example, Cluster 2, whose members include *bank clerk* and *doctor*, are found to be very *useful*, *strict*, and *busy*.

Table 7.1 Membership matrix **G** obtained for the data in Table 6.2

Occupation	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Monk	1			
Bank clerk		1		
Cartoonist			1	
Designer			1	
Nurse			1	
Professor	1			
Doctor		1		
Policeman				1
Journalist				1
Sailor				1
Athlete				1
Novelist	1			
Actor			1	
Stewardess			1	

Table 7.2 Cluster feature matrix **C** obtained for the data in Table 6.2

Cluster	Noble	Useful	Good	Big	Powerful	Strong	Quick	Noisy	Young	Faithful	Strict	Busy
1	3.5	3.5	3.7	3.1	2.8	2.7	2.3	1.6	1.9	3.4	3.5	2.7
2	3.7	4.2	3.7	3.0	3.0	3.2	3.2	2.1	3.1	3.9	4.2	4.4
3	3.4	3.6	3.8	2.7	2.4	2.6	3.4	3.1	4.1	3.6	1.8	4.4
4	3.7	3.9	3.8	3.4	4.1	4.2	4.4	3.7	3.9	3.8	3.4	4.2

7.3 Formulation

KMC is underlain by the model

$$\mathbf{X} = \mathbf{GC} + \mathbf{E}, \tag{7.3}$$

with $\mathbf{E}$ containing errors. To obtain $\mathbf{G}$ and $\mathbf{C}$, a *least squares method* is used; the sum of squared errors

$$f(\mathbf{G}, \mathbf{C}) = \|\mathbf{E}\|^2 = \|\mathbf{X} - \mathbf{GC}\|^2 \tag{7.4}$$

is minimized over $\mathbf{G}$ and $\mathbf{C}$ subject to $\mathbf{G}$ satisfying (7.1) and (7.2).

For the sake of ease in understanding (7.3) and (7.4), we use the example of $\mathbf{X}$ in Fig. 7.1, which is more compact than the data set in Table 6.2. In Fig. 7.1, a 10×2 data matrix $\mathbf{X}$ is shown together with a scatter plot of the 10 row vectors in $\mathbf{X}$. For this data matrix, KMC with $K = 3$ gives the solution expressed as follows:

$$\underset{\mathbf{X}}{\begin{bmatrix} 1 & 4 \\ 7 & 3 \\ 6 & 1 \\ 8 & 6 \\ 3 & 5 \\ 5 & 7 \\ 9 & 2 \\ 2 & 3 \\ 8 & 4 \\ 6 & 9 \end{bmatrix}} = \underset{\mathbf{G}}{\begin{bmatrix} 1 & & \\ & & 1 \\ & & 1 \\ & 1 & \\ 1 & & \\ & 1 & \\ & & 1 \\ 1 & & \\ & & 1 \\ & 1 & \end{bmatrix}} \underset{\mathbf{C}}{\begin{bmatrix} 2.0 & 4.0 \\ 6.3 & 7.3 \\ 7.5 & 2.5 \end{bmatrix}} + \underset{\mathbf{E}}{\begin{bmatrix} -1.0 & 0.0 \\ -0.5 & 0.5 \\ -1.5 & -1.5 \\ 1.7 & -1.3 \\ 1.0 & 1.0 \\ -1.3 & -0.3 \\ 1.5 & -0.5 \\ 0.0 & -1.0 \\ 0.5 & 1.5 \\ -0.3 & 1.7 \end{bmatrix}} . \tag{7.5}$$

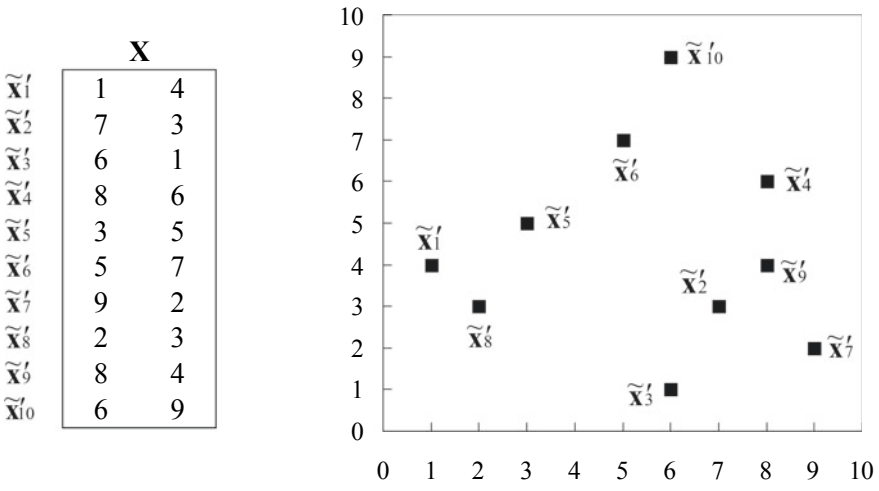


Fig. 7.1 Data matrix $\mathbf{X}$ and the scatter plot of the row vectors in $\mathbf{X}$

Here, model (7.3) is shown, into which the data and the resulting solution were substituted. Further, we can obtain the *product* of **G** and **C** to rewrite (7.5) as

X			**GC**			**E**	
1	4	=	2.0	4.0	+	-1.0	0.0
7	3		7.5	2.5		-0.5	0.5
6	1		7.5	2.5		-1.5	-1.5
8	6		6.3	7.3		1.7	-1.3
3	5		2.0	4.0		1.0	1.0
5	7		6.3	7.3		-1.3	-0.3
9	2		7.5	2.5		1.5	-0.5
2	3		2.0	4.0		0.0	-1.0
8	4		7.5	2.5		0.5	1.5
6	9		6.3	7.3		-0.3	1.7

, (7.6)

where white, light gray, and dark gray have been used for the background colors of the rows corresponding to Clusters 1, 2, and 3, respectively. In (7.6), we find that the ith row of **X** is matched to the row of $\mathbf{C} = [\tilde{\mathbf{c}}_1, \tilde{\mathbf{c}}_2, \tilde{\mathbf{c}}_3]'$ associated with the cluster into which individual i is classified; for example, $\tilde{\mathbf{x}}_3' = [8, 6]$ is matched to $\tilde{\mathbf{c}}_2' = [6.3, 7.3]$.

Solution (7.6) can be illustrated graphically, as in Fig. 7.2, in which the rows of **X** and **C** are plotted. There, we can find that $\tilde{\mathbf{c}}_k'$ (the kth row of **C**) expresses the *representative point* of cluster k that is located at the *center* of the individuals ($\tilde{\mathbf{x}}_i'$ s) belonging to that cluster. For this reason, **C** is also called a *cluster center matrix*. In Fig. 7.2, each of the *lines* connects $\tilde{\mathbf{x}}_i'$ for individual i and $\tilde{\mathbf{c}}_k'$ for the cluster including i. The lines in the figure indicate the row vectors of error matrix $\mathbf{E} = \begin{bmatrix} \tilde{\mathbf{e}}_1' \\ \vdots \\ \tilde{\mathbf{e}}_n' \end{bmatrix}$. For example, the line extending from center $\tilde{\mathbf{c}}_3'$ to $\tilde{\mathbf{x}}_8'$ indicates $\tilde{\mathbf{e}}_8' = \tilde{\mathbf{x}}_8' - \tilde{\mathbf{c}}_3'$ with $\tilde{\mathbf{e}}_i'$ the ith row of **E**. Here, we should note that the function (7.4) to be

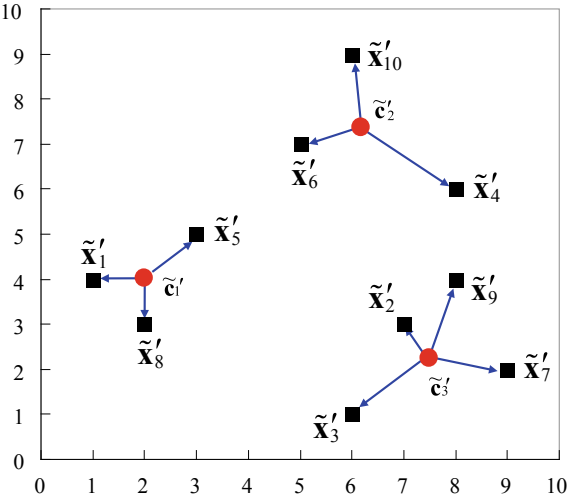


Fig. 7.2 Joint plot of the rows of **X** in Fig. 7.1 with those of **C**

minimized is rewritten as $\|\mathbf{E}\|^2 = \|\tilde{\mathbf{e}}_1\|^2 + \cdots + \|\tilde{\mathbf{e}}_n\|^2$ Its minimization is restated as minimizing the *sum of the squared lengths of the lines* in Fig. 7.2, which implies making each individual vector ($\tilde{\mathbf{x}}_i'$) close to the center of the cluster ($\tilde{\mathbf{c}}_k'$) including the individual.

7.4 Iterative Algorithm

Let us remember that the PCA solution is obtained through (5.14) and the solution for regression analysis is given by (4.9) and (4.12); those solutions are expressed explicitly as formulas. On the other hand, the KMC solution minimizing (7.4) *cannot* be given explicitly by a formula. In general, statistical analysis procedures can be classified into the following two types:

[1] those *with explicit solutions* (as regression analysis and PCA)
[2] those *without explicit solutions* (as KMC)

How are solutions for [2] obtained? They can be attained with *iterative algorithms*, where steps are iterated for finding the solution. There are some types of iterative algorithms, as described in Appendix A.6.1.

The algorithm for KMC is formed using the following fact: although the $\mathbf{G}$ and $\mathbf{C}$ minimizing (7.4), i.e., $f(\mathbf{G}, \mathbf{C})$, cannot be expressed as formulas, the matrices

$$\mathbf{C} \text{ that minimizes } f(\mathbf{G}, \mathbf{C}) \text{ while } \mathbf{G} \text{ is fixed at a specified matrix} \tag{7.7}$$

and

$$\mathbf{G} \text{ that minimizes } f(\mathbf{G}, \mathbf{C}) \text{ while } \mathbf{C} \text{ is fixed at a specified matrix} \tag{7.8}$$

can be explicitly given, as shown in the next sections. This fact allows us to form the iterative *algorithm for KMC*, described by the following steps:

Step 1. Set $\mathbf{G}$ and $\mathbf{C}$ to specified matrices $\mathbf{G}_{[t]}$ and $\mathbf{C}_{[t]}$, respectively, with $t = 0$.
Step 2. Obtain $\mathbf{C}$ defined as (7.7) with $\mathbf{G}$ being fixed at $\mathbf{G}_{[t]}$, and express the resulting $\mathbf{C}$ as $\mathbf{C}_{[t+1]}$.
Step 3. Obtain $\mathbf{G}$ defined as (7.8) with $\mathbf{C}$ fixed at $\mathbf{C}_{[t+1]}$, and express the resulting $\mathbf{G}$ as $\mathbf{G}_{[t+1]}$.
Step 4. Finish and regard $\mathbf{C}_{[t+1]}$ and $\mathbf{G}_{[t+1]}$ as the solution, if convergence is reached; otherwise, go back to Step 2 with increasing t by one.

Here, t stands for the number of iterations, and the convergence in Step 4 is explained later. The central part of the algorithm is the *alternate iteration of Steps 2 and 3*. With this iteration, the value of function (7.4) *decreases monotonically* (or remains unchanged), regardless of what is used for the specified matrices in Step 1, as described in the following paragraphs.

Let us consider the value of (7.4) at Step 1, i.e., $f(\mathbf{G}_{[0]}, \mathbf{C}_{[0]}) = f(\mathbf{G}_{[t]}, \mathbf{C}_{[t]})$ for $t = 0$, which is followed by Steps 2 and 3, providing $f(\mathbf{G}_{[0]}, \mathbf{C}_{[1]})$ and $f(\mathbf{G}_{[1]}, \mathbf{C}_{[1]})$, respectively. They are found to satisfy

$$f\left(\mathbf{G}_{[0]}, \mathbf{C}_{[0]}\right) \geq f\left(\mathbf{G}_{[0]}, \mathbf{C}_{[1]}\right) \geq f\left(\mathbf{G}_{[1]}, \mathbf{C}_{[1]}\right). \tag{7.9}$$

Here, the first inequality $f(\mathbf{G}_{[0]}, \mathbf{C}_{[0]}) \geq f(\mathbf{G}_{[0]}, \mathbf{C}_{[1]})$ follows from the fact that $\mathbf{C}_{[1]}$ is the matrix $\mathbf{C}$ that minimizes $f(\mathbf{G}, \mathbf{C})$ with $\mathbf{G}$ fixed to $\mathbf{G}_{[0]}$ as found in (7.7), and the second inequality $f(\mathbf{G}_{[0]}, \mathbf{C}_{[1]}) \geq f(\mathbf{G}_{[1]}, \mathbf{C}_{[1]})$ follows from (7.8), i.e., $\mathbf{G}_{[1]}$ being the matrix $\mathbf{G}$ that minimizes $f(\mathbf{G}, \mathbf{C}_{[1]})$ with $\mathbf{C}$ fixed to $\mathbf{C}_{[1]}$.

As described in Step 4, unless convergence is reached, the algorithm must go back to Step 2, with an increase in t from one to two. Then, Steps 2 and 3 are performed again to have $\mathbf{C}_{[2]}$ and $\mathbf{G}_{[2]}$, which allows (7.9) to be followed by two inequalities $\geq f(\mathbf{G}_{[1]}, \mathbf{C}_{[2]}) \geq f(\mathbf{G}_{[2]}, \mathbf{C}_{[2]})$: it leads to

$$f\left(\mathbf{G}_{[0]}, \mathbf{C}_{[0]}\right) \geq f\left(\mathbf{G}_{[0]}, \mathbf{C}_{[1]}\right) \geq f\left(\mathbf{G}_{[1]}, \mathbf{C}_{[1]}\right) \geq f\left(\mathbf{G}_{[1]}, \mathbf{C}_{[2]}\right) \geq f\left(\mathbf{G}_{[2]}, \mathbf{C}_{[2]}\right). \tag{7.10}$$

We can generalize (7.9) and (7.10) as

$$f\left(\mathbf{G}_{[t]}, \mathbf{C}_{[t]}\right) \geq f\left(\mathbf{G}_{[t]}, \mathbf{C}_{[t+1]}\right) \geq f\left(\mathbf{G}_{[t+1]}, \mathbf{C}_{[t+1]}\right) \tag{7.11}$$

for $t = 0, 1, 2, \ldots$, where $\mathbf{C}_{[t+1]}$ denotes the matrix $\mathbf{C}$ obtained in Step 2 and $\mathbf{G}_{[t+1]}$ denotes the matrix $\mathbf{G}$ obtained in Step 3 at the tth iteration. That is, the value of $f(\mathbf{G}, \mathbf{C})$ *decreases monotonically* with an *increase* in t so that the value is expected to *converge* to the minimum. Convergence can be defined as having a difference in the value of (7.4) from the previous round of iteration that is small enough to be ignored, i.e.,

$$f\left(\mathbf{G}_{[t]}, \mathbf{C}_{[t]}\right) - f\left(\mathbf{G}_{[t+1]}, \mathbf{C}_{[t+1]}\right) \leq \varepsilon, \tag{7.12}$$

with ε being a small value, such as 0.1^6 or 0.1^5.

Figure 7.3 shows the change in $f(\mathbf{G}, \mathbf{C})$ with the iterative KMC algorithm for the data in Fig. 7.1, where the elements of the specified matrices in Step 1 were

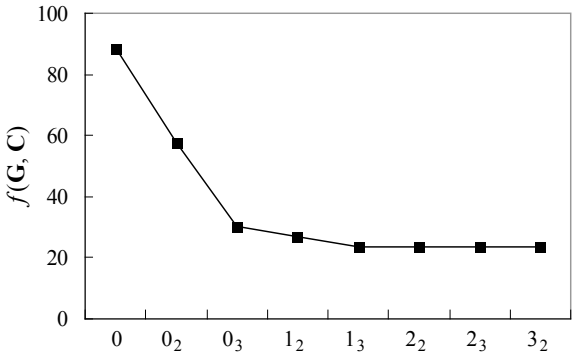


Fig. 7.3 Values of $f(\mathbf{G}, \mathbf{C})$ at steps in the t-iteration ($t = 0$, 1, 2, 3) with subscripts 2 and 3 indicating Steps 2 and 3

randomly chosen. In Fig. 7.3 we find the monotonic decrease in the $f(\mathbf{G}, \mathbf{C})$ value with t, and the value is unchanged from $t = 2$ to 3, i.e., convergence is reached at $t = 3$. The matrices $\mathbf{C}$ and $\mathbf{G}$ at this time are their solution in (7.5). The computations that were used in Steps 2 and 3 are described in the following two sections.

7.5 Obtaining Cluster Features

In this section, we consider Step 2 from the previous section, i.e., obtaining the *cluster feature matrix* $\mathbf{C}$ defined as (7.7). The matrix $\mathbf{C}$ to be obtained is the one minimizing $f(\mathbf{C}) = \left\|\mathbf{X} - \mathbf{G}_{[t]}\mathbf{C}\right\|^2$, i.e., the function (7.4) with $\mathbf{G}$ fixed at $\mathbf{G}_{[t]}$. This $\mathbf{C}$ is given by

$$\mathbf{C}_{[t+1]} = \left(\mathbf{G}'_{[t]}\mathbf{G}_{[t]}\right)^{-1}\mathbf{G}'_{[t]}\mathbf{X} = \mathbf{D}^{-1}\mathbf{G}'_{[t]}\mathbf{X}, \tag{7.13}$$

with $\mathbf{D} = \mathbf{G}_{[t]}{}'\mathbf{G}_{[t]}$, as explained in Appendix A.2.2. There, we can compare (A.2.11) with (7.4) to find that (A.2.12) leads to (7.13).

Let us consider what matrix$\mathbf{D} = \mathbf{G}_{[t]}{}'\mathbf{G}_{[t]}$ is, with a simple example of $\mathbf{G}_{[t]}$:

$$\text{If } \mathbf{G}_{[t]} = \begin{bmatrix} & 1 & \\ 1 & & \\ & & 1 \\ 1 & & \\ & 1 & \\ & 1 & \end{bmatrix}, \text{ then } \mathbf{D} = \begin{bmatrix} & 1 & & 1 & & \\ 1 & & & & 1 & 1 \\ & & 1 & & & \end{bmatrix} \begin{bmatrix} & 1 & \\ 1 & & \\ & & 1 \\ 1 & & \\ & 1 & \\ & 1 & \end{bmatrix} = \begin{bmatrix} 2 & & \\ & 3 & \\ & & 1 \end{bmatrix}.$$

In general, $\mathbf{D} = \mathbf{G}_{[t]}{}'\mathbf{G}_{[t]}$ is a $K \times K$ diagonal matrix, with its kth diagonal element is n_k which denotes the number of individuals belonging to cluster k. Thus, the inverse matrix $\mathbf{D}^{-1}$ is found to be the diagonal matrix whose kth diagonal element is $1/n_k$. Further, in the above example, $\mathbf{D}^{-1}\mathbf{G}_{[t]}{}' = \begin{bmatrix} & 1/2 & 1/2 & & & \\ 1/3 & & & 1/3 & 1/3 & \\ & & 1 & & & \end{bmatrix}$. This is post-multiplied by 6×2 $\mathbf{X}$ to give an example of (7.13):

$$\mathbf{C}_{[t+1]} = \mathbf{D}^{-1}\mathbf{G}'_{[t]}\mathbf{X} = \begin{bmatrix} & 1/2 & & 1/2 & & \\ 1/3 & & 1 & & 1/3 & 1/3 \end{bmatrix} \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \\ x_{41} & x_{42} \\ x_{51} & x_{52} \\ x_{61} & x_{62} \end{bmatrix}$$

$$= \begin{bmatrix} \bar{x}_{11} & \bar{x}_{12} \\ \bar{x}_{21} & \bar{x}_{22} \\ \bar{x}_{31} & \bar{x}_{32} \end{bmatrix}. \tag{7.14}$$

Here,

$$\bar{x}_{kj} = \frac{1}{n_k} \sum_{i \in \text{cluster } k} x_{ij}, \tag{7.15}$$

with $\sum_{i \in \text{cluster } k} x_{ij}$ denoting the summation of x_{ij} over the individuals belonging to cluster k. That is, (7.15) is the *average* of the data within *cluster k* for variable j, and such cluster averages are the elements of (7.13) and its example (7.14). The term "*k-means*" originates in the fact that "mean" which is a synonym of "average"plays an important role in the algorithm.

7.6 Obtaining Memberships

In this section, Step 3 from Sect. 7.4 is considered; it is shown how the *membership* matrix $\mathbf{G}$ is obtained that minimizes $f(\mathbf{G}) = \left\|\mathbf{X} - \mathbf{G}\mathbf{C}_{[t+1]}\right\|^2$, i.e., the function (7.4) with $\mathbf{C}$ fixed at $\mathbf{C}_{[t+1]}$.

Using $\tilde{\mathbf{g}}'_i$ for the ith row of $\mathbf{G}$, the function $f(\mathbf{G})$ can be rewritten as

$$\left\|\mathbf{X} - \mathbf{G}\mathbf{C}_{[t+1]}\right\|^2 = \left\| \begin{bmatrix} \tilde{\mathbf{x}}'_1 \\ \vdots \\ \tilde{\mathbf{x}}'_n \end{bmatrix} - \begin{bmatrix} \tilde{\mathbf{g}}'_1 \\ \vdots \\ \tilde{\mathbf{g}}'_n \end{bmatrix} \mathbf{C}_{[t+1]} \right\|^2 = \sum_{i=1}^{n} \left\|\tilde{\mathbf{x}}'_i - \tilde{\mathbf{g}}'_i\mathbf{C}_{[t+1]}\right\|^2, \tag{7.16}$$

which is the sum of the least squares function of $\tilde{\mathbf{g}}'_i$,

$$f_i(\tilde{\mathbf{g}}'_i) = \left\|\tilde{\mathbf{x}}'_i - \tilde{\mathbf{g}}'_i\mathbf{C}_{[t+1]}\right\|^2, \tag{7.17}$$

over $i = 1, \ldots, n$. Here, it should be noted that $\tilde{\mathbf{g}}'_i$ appears only in $f_i(\tilde{\mathbf{g}}'_i)$, i.e., not in the other functions $f_u(\tilde{\mathbf{g}}'_u)$ with $u \neq i$. This implies that the optimal $\tilde{\mathbf{g}}'_i$, which minimizes (7.17), can be obtained independently of $\tilde{\mathbf{g}}'_u$ with $u \neq i$; the repetition of

obtaining the optimal $\tilde{\mathbf{g}}_i'$ over $i = 1, \ldots, n$ provides the rows of the membership matrix $\mathbf{G}$ that minimizes (7.16).

Let us recall (7.1) and (7.2), i.e., that $\tilde{\mathbf{g}}_i'$ is filled with zeros except for one element taking 1. For example, if $K = 3$ and individual i belongs to cluster 2, then $\tilde{\mathbf{g}}_i' = [0, 1, 0]$, thus, $\tilde{\mathbf{g}}_i'\mathbf{C}_{[t+1]} = \tilde{\mathbf{c}}_2'$ and (7.17) can be rewritten as $\left\|\tilde{\mathbf{x}}_i' - \tilde{\mathbf{c}}_2'\right\|^2$, with $\tilde{\mathbf{c}}_k'$ the kth row of $\mathbf{C}_{[t+1]}$. This example allows us to find that (7.17) takes one of K distinct values as

$$f_i(\tilde{\mathbf{g}}_i') = \left\|\tilde{\mathbf{x}}_i' - \tilde{\mathbf{g}}_i'\mathbf{C}\right\|^2 = \begin{cases} \left\|\tilde{\mathbf{x}}_i' - \tilde{\mathbf{c}}_1'\right\|^2 & \text{if } \tilde{\mathbf{g}}_i' = [1, 0, 0, \ldots, 0] \\ \left\|\tilde{\mathbf{x}}_i' - \tilde{\mathbf{c}}_2'\right\|^2 & \text{if } \tilde{\mathbf{g}}_i' = [0, 1, 0, \ldots, 0] \\ \vdots & \\ \left\|\tilde{\mathbf{x}}_i' - \tilde{\mathbf{c}}_K'\right\|^2 & \text{if } \tilde{\mathbf{g}}_i' = [0, 0, 0, \ldots, 1] \end{cases}. \tag{7.18}$$

Therefore, we can compare the largeness of $\left\|\tilde{\mathbf{x}}_i' - \tilde{\mathbf{c}}_k'\right\|^2$ across $k = 1, \ldots, K$ to select the vector $\tilde{\mathbf{g}}_i'$ corresponding to the minimal one among $\left\|\tilde{\mathbf{x}}_i' - \tilde{\mathbf{c}}_k'\right\|^2, k = 1, \ldots, K$. This selection is formally expressed as

$$g_{ik} = \begin{cases} 1 & \text{if } \left\|\tilde{\mathbf{x}}_i - \tilde{\mathbf{c}}_k'\right\|^2 = \min\limits_{1 \le l \le K} \left\|\tilde{\mathbf{x}}_i - \tilde{\mathbf{c}}_l'\right\|^2 \\ 0 & \text{otherwise} \end{cases}. \tag{7.19}$$

The selected vector is the optimal $\tilde{\mathbf{g}}_i' = [g_{i1}, \ldots, g_{iK}]$ minimizing (7.17). Repeating the selection (7.19) over $i = 1, \ldots, n$ provides the vectors $\tilde{\mathbf{g}}_1', \ldots, \tilde{\mathbf{g}}_n'$, that form the rows of $\mathbf{G}_{[t+1]}$ to be obtained.

7.7 Brief Description of Algorithm

The steps of the KMC algorithm in Sect. 7.4 can be rewritten in a simpler manner (without using the subscript t indicating the number of iteration) as follows:

Step 1. Initialize $\mathbf{G}$.
Step 2. Obtain $\mathbf{C} = \mathbf{D}^{-1}\mathbf{G}'\mathbf{X}$
Step 3. Update $\mathbf{G}$ with (7.19)
Step 4. Finish if convergence is reached; otherwise, go back to Step 2.

Here, the facts in Sects. 7.5 and 7.6 have been used in Steps 2 and 3. The phrase "*initialize* $\mathbf{G}$" in Step 1 refers to "set $\mathbf{G}$ to a matrix", as the elements of the latter matrix are called *initial values*. It should be noted that $\mathbf{C}$ may not be initialized in Step 1, since $\mathbf{C}$ is obtained in the next step.

Another version of the KMC algorithm can be formed in which rather $\mathbf{C}$ is initialized with Steps 2 and 3 interchanged. The version can be listed as follows:

Step 1. Initialize $\mathbf{C}$.
Step 2. Update $\mathbf{G}$ with (7.19).
Step 3. Obtain $\mathbf{C} = \mathbf{D}^{-1}\mathbf{G}'\mathbf{X}$.
Step 4. Finish if convergence is reached; otherwise, go back to Step 2.

7.8 Bibliographical Notes

Everitt (1993) has intelligibly treated the cluster analysis procedures, including *hierarchal clustering*, which was not introduced in the present book. The recent developments in clustering have been exhaustively detailed in Gan et al. (2007). Hartigan and Wang (1979) have proposed a modified version of the KMC algorithm described in this chapter.

Exercises

7.1. Show that (7.2) could not be satisfied if two or more elements took one with the other elements being zero in a row of $\mathbf{G}$.
7.2. Let $\mathbf{D} = \mathbf{G}'\mathbf{G}$. Show that $\mathbf{D}$ is a $K \times K$ diagonal matrix and $\mathbf{G}'\mathbf{1}_n = \mathbf{D}\mathbf{1}_K$ with the kth element of $\mathbf{D}\mathbf{1}_K$ and the kth diagonal element of $\mathbf{D}$ being the number of the individuals in group k.
7.3. Show that (7.4) can be rewritten as $\sum_{i=1}^{n} \left\|\tilde{\mathbf{x}}_i - \tilde{\mathbf{c}}_{y_i}\right\|^2$, with y_i the index representing the cluster to which individual i belongs and $\tilde{\mathbf{c}}'_{y_i}$ being the y_ith row of $\mathbf{C}$.
7.4. Show that (7.4) can be rewritten as $\sum_{i=1}^{n} \sum_{k=1}^{K} g_{ik} \|\tilde{\mathbf{x}}_i - \tilde{\mathbf{c}}_k\|^2$.
7.5. One drawback of the k-means clustering (KMC) is that it tends to give *local* minima, but not the *global* minimum. Here, the *global minimum* is defined as the minimum of $f(\boldsymbol{\theta})$ for all possible $\boldsymbol{\theta}$, using $f(\boldsymbol{\theta})$ for the loss function of $\boldsymbol{\theta}$ (parameter vector or a parameter) to be minimized. On the other hand, a *local minimum* is defined as the minimum of $f(\boldsymbol{\theta})$ for the $\boldsymbol{\theta}$ value within a restricted range. Those minima are illustrated in Fig. 7.4. To avoid the selection of $\boldsymbol{\theta}$ for a local minimum as the solution, the algorithm is run multiple times by starting with different initial values, in the procedures sensitive to local minima. Let us use $f(\boldsymbol{\theta}_l)$ for the loss function value resulting in the lth run of the algorithm with $l = 1, \ldots, L$. Then, $\boldsymbol{\theta}_{l^*}$ is selected as the solution with $f(\theta_{l*}) = \min_{1 \le l \le L} f(\theta_l)$. Describe why this *multi-run procedure* decreases the possibility of selecting $\boldsymbol{\theta}$ for a local minimum as the solution.
7.6. The iterative algorithm in Sects. 7.4–7.7 is included in a family of algorithms generally called *alternating least squares* (*ALS*) *algorithms*, as described in Appendix A.6.1. In this exercise, let us consider an ALS algorithm for a problem different from KMC. The problem is the minimization of $f(a, b) = \|\mathbf{y} - a\mathbf{x}_1 - ab\mathbf{x}_2\|^2$ over a and b for $n \times 1$ data vectors $\mathbf{y}$, $\mathbf{x}_1$, and $\mathbf{x}_2$. Here, it should be noted that the coefficient of $\mathbf{x}_2$ is the product of a and b. Show that

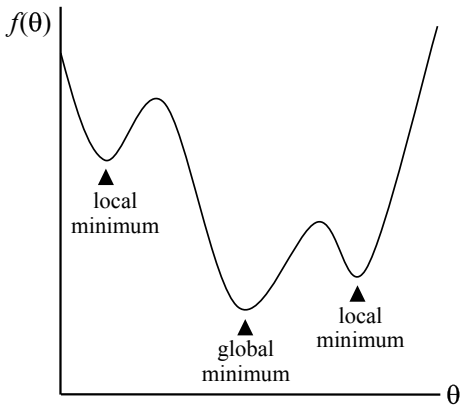


Fig. 7.4 Illustration of local minima and the global minimum

$f(a, b)$ can be rewritten as $\|\mathbf{y} - a\mathbf{x}\|^2$ with $\mathbf{x} = \mathbf{x}_1 + b\mathbf{x}_2$ and also as $\|\mathbf{y}^* - b\mathbf{x}^*\|^2$ with $\mathbf{y}^* = \mathbf{y} - a\mathbf{x}_1$ and $\mathbf{x}^* = a\mathbf{x}_2$, leading to an ALS algorithm in which the minimization can be attained by the following steps:

Step 1. Initialize b.
Step 2. Obtain $\mathbf{x} = \mathbf{x}_1 + b\mathbf{x}_2$ to update a with $a = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y}$.
Step 3. Obtain $\mathbf{y}^* = \mathbf{y} - a\mathbf{x}_1$ and $\mathbf{x}^* = a\mathbf{x}_2$ to update b with $b = (\mathbf{x}^{*\prime}\mathbf{x}^*)^{-1}\mathbf{x}^{*\prime}\mathbf{y}^*$.
Step 4. Finish if convergence is reached; otherwise, go back to Step 2.

Hints are found in Appendix A.2.2.

7.7. Show that (7.4) can be decomposed as

$$\|\mathbf{X} - \mathbf{GC}\|^2 = \left\|\mathbf{X} - \mathbf{GD}^{-1}\mathbf{G}'\mathbf{X}\right\|^2 + \left\|\mathbf{GD}^{-1}\mathbf{G}'\mathbf{X} - \mathbf{GC}\right\|^2, \tag{7.20}$$

with $\mathbf{D} = \mathbf{G}'\mathbf{G}$, by noting

$$\begin{aligned}\|\mathbf{X} - \mathbf{GC}\|^2 &= \left\|\mathbf{X} - \mathbf{GD}^{-1}\mathbf{G}'\mathbf{X} + \mathbf{GD}^{-1}\mathbf{G}'\mathbf{X} - \mathbf{GC}\right\|^2 \\ &= \left\|\mathbf{X} - \mathbf{GD}^{-1}\mathbf{G}'\mathbf{X}\right\|^2 + \left\|\mathbf{GD}^{-1}\mathbf{G}'\mathbf{X} - \mathbf{GC}\right\|^2 \\ &\quad + 2\mathrm{tr}\left(\mathbf{X} - \mathbf{GD}^{-1}\mathbf{G}'\mathbf{X}\right)'\left(\mathbf{GD}^{-1}\mathbf{G}'\mathbf{X} - \mathbf{GC}\right).\end{aligned}$$

7.8. Show that $\left\|\mathbf{GD}^{-1}\mathbf{G}'\mathbf{X} - \mathbf{GC}\right\|^2$ in (7.20) can be rewritten as $\left\|\mathbf{D}^{-1/2}\mathbf{G}'\mathbf{X} - \mathbf{D}^{1/2}\mathbf{C}\right\|^2$, i.e., (7.4) can be decomposed as

$$\|\mathbf{X} - \mathbf{GC}\|^2 = \left\|\mathbf{X} - \mathbf{GD}^{-1}\mathbf{G}'\mathbf{X}\right\|^2 + \left\|\mathbf{D}^{-1/2}\mathbf{G}'\mathbf{X} - \mathbf{D}^{1/2}\mathbf{C}\right\|^2. \tag{7.21}$$

7.9. De Soete and Carroll (1994) have proposed *reduced k-means analysis* (*RKM*) in which clustering is performed simultaneously with principal component analysis. In RKM, the matrix $\mathbf{C}$ ($K \times p$) in (7.4) is constrained as $\mathbf{C} = \mathbf{FA}'$.

Here, $\mathbf{F}$ is $K \times m$, $\mathbf{A}$ is $p \times m$, $\mathbf{A}'\mathbf{A} = \mathbf{I}_m$, and $\mathbf{F}'\mathbf{D}\mathbf{F}$ being a diagonal matrix whose diagonal elements are arranged in descending order, with $m \leq \min(K, p)$ and $\mathbf{D} = \mathbf{G}'\mathbf{G}$. That is, RKM is formulated as minimizing $\|\mathbf{X} - \mathbf{G}\mathbf{F}\mathbf{A}'\|^2$ over $\mathbf{G}$, $\mathbf{F}$, and $\mathbf{A}$ subject to the above constraints, (7.1) and (7.2). Show that an ALS algorithm for RKM can be formed by the following steps:

Step 1. Initialize $\mathbf{C} = \mathbf{F}\mathbf{A}'$.
Step 2. Obtain $\mathbf{G}$ with (7.19).
Step 3. Perform SVD of $\mathbf{D}^{-1/2}\mathbf{G}'\mathbf{X}$, defined as $\mathbf{D}^{-1/2}\mathbf{G}'\mathbf{X} = \mathbf{K}\mathbf{\Lambda}\mathbf{L}'$.
Step 4. Obtain $\mathbf{C} = \mathbf{D}^{-1/2}\mathbf{K}_m\mathbf{\Lambda}_m\mathbf{L}_m{}'$ with $\mathbf{K}_m$ ($K \times m$) and $\mathbf{L}_m$ ($p \times m$) containing the first m columns of $\mathbf{K}$ and $\mathbf{L}$, respectively, and $\mathbf{\Lambda}_m$ ($m \times m$) the diagonal matrix whose lth diagonal element is that of $\mathbf{\Lambda}$.
Step 5. Set $\mathbf{F} = \mathbf{D}^{-1/2}\mathbf{K}_m\mathbf{\Lambda}_m$ and $\mathbf{A} = \mathbf{L}_m$ to finish if convergence is reached; otherwise, go back to Step 2.

Here, (7.21) has been used in Steps 3 and 4 with the hints for those steps found in Note 5.3.

7.10. Show that the algorithm in Sects. 7.4–7.7 can give a $\mathbf{G}$ whose columns include $\mathbf{0}_n$ during iteration, which implies that $\mathbf{D}^{-1} = (\mathbf{G}'\mathbf{G})^{-1}$ does not exist and stops the algorithm, i.e., makes KMC fail.

Part III
Maximum Likelihood Procedures

This part starts with the introduction of the principle underlying the maximum likelihood method. This is followed by introductions to path analysis, factor analysis, and structural equation modeling, whose solutions are estimated by the maximum likelihood method. Their solutions can also be obtained by least squares methods, and the procedures in Part II can also be formulated with the maximum likelihood method. However, the latter are often introduced with the least squares methods, while the maximum likelihood method is often used for the procedures discussed in this part.

Chapter 8
Maximum Likelihood and Multivariate Normal Distribution

In the analysis procedures introduced in the last four chapters, parameters are estimated by the *least squares* (*LS*) *method*, as reviewed in Sect. 8.1. The remaining sections in this chapter serve to prepare readers for the following chapters, in which a *maximum likelihood* (*ML*) *method*, which differs from LS, is used for estimating parameters. That is, the ML method is introduced in Sect. 8.2, which is followed by describing the notion of *probability density function* and the ML method with *multivariate normal distribution*. Finally, ML-based *model selection* with *information criteria* is introduced.

8.1 Model, Parameter, Objective Function, and Optimization

This section deals with a very big subject: We discuss a general framework in which almost all statistical analysis procedures can be formulated; namely, any procedure is underlain by a *model* that can be expressed as

$$\text{Data} \cong \phi(\boldsymbol{\Theta}) \quad \text{or} \quad \text{Data} = \phi(\boldsymbol{\Theta}) + \text{Errors}, \tag{8.1}$$

with $\boldsymbol{\Theta}$ standing for the *parameters* to be obtained. For example, in *K*-means clustering (Chap. 6), $\boldsymbol{\Theta}$ is $\{\mathbf{G}, \mathbf{C}\}$ and $\phi(\boldsymbol{\Theta}) = \phi(\mathbf{G}, \mathbf{C}) = \mathbf{GC}$, as found in (7.3). Another example is regression analysis (Chap. 4). In its model (4.5), the "Data" in (8.1) are denoted as dependent variable vector $\mathbf{y}$, while $\Theta = \{\mathbf{b}, c\}$ and $\phi(\Theta) = \phi(\mathbf{b}, c) = \mathbf{Xb} + c\mathbf{1}$, with $\mathbf{X}$ containing explanatory variables.

An analysis procedure modeled as (8.1) obtains or estimates parameter $\boldsymbol{\Theta}$ values. This is formulated as "Obtaining $\boldsymbol{\Theta}$ that *optimizes* an *objective function obj*($\boldsymbol{\Theta}$) subject to a *constraint* on $\boldsymbol{\Theta}$". This phrase is rewritten as

K. Adachi, *Matrix-Based Introduction to Multivariate Data Analysis*,
https://doi.org/10.1007/978-981-15-4103-2_8

$$\text{Optimizing } obj(\boldsymbol{\Theta}) \text{over } \boldsymbol{\Theta} \text{ subject to a constraint on } \boldsymbol{\Theta}. \tag{8.2}$$

Here, the term "*optimizes*" refers to either "*minimizes*" or "*maximizes*", and some function can be used as $obj(\boldsymbol{\Theta})$. In Chaps. 4, 5, 6, and 7, *least squares* functions are used as $obj(\boldsymbol{\Theta})$, which are generally expressed as $\|\text{Data} - \phi(\boldsymbol{\Theta})\|^2$, i.e., the sum of the squared Errors = Data $- \phi(\boldsymbol{\Theta})$, with "optimizes" referring to "minimizes". The phrase "subject to a *constraint* on $\boldsymbol{\Theta}$" in (8.2) is not indispensable; whether the phrase is necessary or not depends on analysis procedures. For example, it is necessary in the *k*-means clustering in which $\mathbf{G}$ in $\boldsymbol{\Theta} = \{\mathbf{G}, \mathbf{C}\}$ is constrained to satisfy (7.1) and (7.2), while the phrase is unnecessary in the regression analysis, in which $\boldsymbol{\Theta} = \{\mathbf{b}, c\}$ is unconstrained.

A methodology formulated by rephrasing "Optimizing $obj(\boldsymbol{\Theta})$ over $\boldsymbol{\Theta}$" in (8.2) as "*minimizing a least squares function*" is generally called a *least squares* (*LS*) *method*, which is used for the procedures in Part 2. Another methodology, which is as important as the LS method, is introduced next.

8.2 Maximum Likelihood Method

A *maximum likelihood* (*ML*) *method* can be formulated by rephrasing "*optimizing*" and "*an objective function*" in (8.2) as "*maximizing*" and "*probability*", respectively. One feature of the ML method is that it uses the notion of probabilities, which are not used in the LS method. In this section, we introduce the ML method using a simple example.

We suppose that a black box contains *black* and *white* balls, where the total number of the balls is known to be 100, but the number of *black*/*white* balls is unknown. We use θ for the number of black ones. Let us consider a case illustrated in Fig. 8.1: In order to estimate θ, a ball was chosen from the box and returned five times, which gave the data set

$$\mathbf{d} = [1, 0, 0, 1, 0]'. \tag{8.3}$$

Here, $d_i = 1$ and $d_i = 0$ indicate black and white balls chosen, respectively, with d_i the ith element of $\mathbf{d}$.

Let us consider the probability of the data set in (8.3) being observed. On the supposition of a ball randomly chosen, $P(d_i = 1|\theta)$ and $P(d_i = 0|\theta)$, which denote the probability of $d_i = 1$ observed (i.e., a *black* ball chosen) and that of $d_i = 0$ (i.e., a *white* one chosen), respectively, are expressed as

$$P(d_i = 1|\theta) = \frac{\theta}{100} \quad \text{and} \quad P(d_i = 0|\theta) = 1 - \frac{\theta}{100}. \tag{8.4}$$

Further, we suppose the balls were chosen mutually independently. Then, the probability of the data set $\mathbf{d} = [1, 0, 0, 1, 0]'$ observed in (8.3), i.e., $d_1 = 1$, $d_2 = 0$,

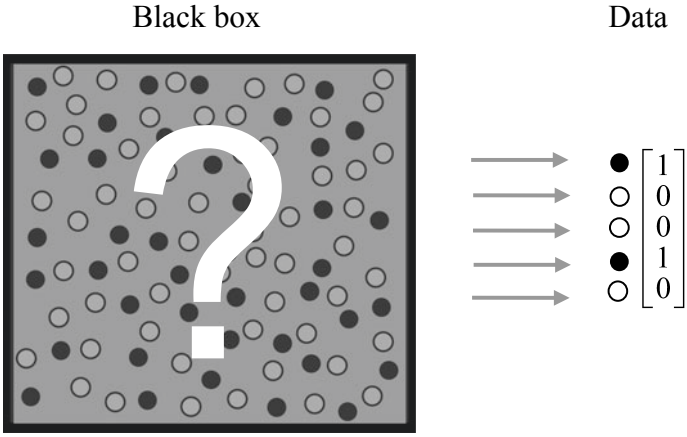


Fig. 8.1 Data of balls chosen from a black box that contains white and black balls with their numbers unknown

$d_3 = 0$, $d_4 = 1$, and $d_5 = 0$ is given by the product $P(d_1 = 1|\theta) \times P(d_2 = 0|\theta) \times P(d_3 = 0|\theta) \times P(d_4 = 1|\theta) \times P(d_5 = 0|\theta)$:

$$\begin{aligned} P(\mathbf{d}|\theta) &= \frac{\theta}{100} \times \left(1 - \frac{\theta}{100}\right) \times \left(1 - \frac{\theta}{100}\right) \times \frac{\theta}{100} \times \left(1 - \frac{\theta}{100}\right) \\ &= \left(\frac{\theta}{100}\right)^2 \left(1 - \frac{\theta}{100}\right)^3. \end{aligned} \tag{8.5}$$

For estimating the value of θ, the ML method can be used. Without using mathematics, the idea of the method can be stated as "Obtaining the *parameter* value such that the occurrence of an *event* is the *most likely*", which can be rephrased as

$$\begin{aligned} &\text{Obtaining the } \textit{parameter} \text{ value which} \\ &\quad \textit{maximizes} \text{ how likely it is that the } \textit{event} \text{ will occur.} \end{aligned} \tag{8.6}$$

Here, the "*event*" refers to the *observation* of a *data* set, i.e., observing **d** in (8.3), and "*how likely* it is that the *event* will occur" is measured by its *probability*. That is, we can use statistical terms to rephrase (8.6) as:

$$\begin{aligned} &\text{Obtaining the } \textit{parameter} \text{ value that } \textit{maximizes} \\ &\quad \text{the probability of the } \textit{data} \text{ being observed.} \end{aligned} \tag{8.7}$$

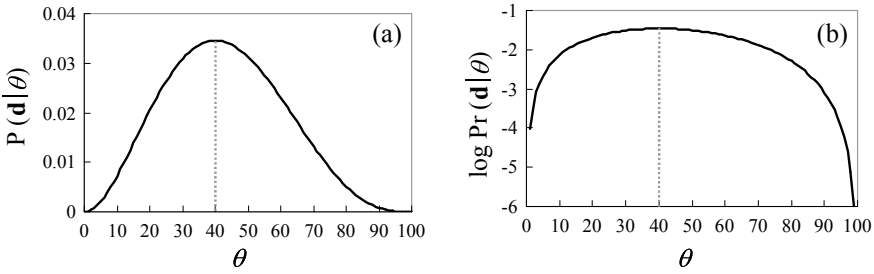


Fig. 8.2 Probability values (**a**) and their logarithms (**b**) against θ

Therefore, the ML method for the data set in (8.3) is to obtain the value of θ that maximizes (8.5). Figure 8.2a shows that the values of $\theta = 0, 1, \ldots, 100$. There, we can find that the solution of θ that maximizes the probability is 40. The ML method is similar to a human *psychological* process; *most people seem to think in a manner similar to that in the ML method*. For example, in order to determine who caused an event as "James caused it? Jim? Or Miller did?", one would consider the person most likely to cause the event is the person to be found!

Let us note that $P(\mathbf{d}|\theta)$ is treated as a *function of parameter* θ for a fixed $\mathbf{d}$ in the ML method (Fig. 8.2a), in contrast to cases where $P(\mathbf{d}|\theta)$ is regarded as expressing how probable it is that data set $\mathbf{d}$ occurs for a fixed value of θ. As in Fig. 8.2a, the probability, if it is treated as a function of parameters, is rephrased as *likelihood*, from which the name *maximum likelihood method* originates.

For the sake of ease in mathematical operation, the parameter value that maximizes the *logarithm* of probability (*log likelihood*) rather than the probability (likelihood) is often obtained in the ML method, since a function, $f(y)$, and its logarithm, $\log f(y)$, take their maximums at the same value of y. The log of (8.5) is given by

$$\log P(\mathbf{d}|\theta) = 2\log\frac{\theta}{100} + 3\log\left(1 - \frac{\theta}{100}\right). \tag{8.8}$$

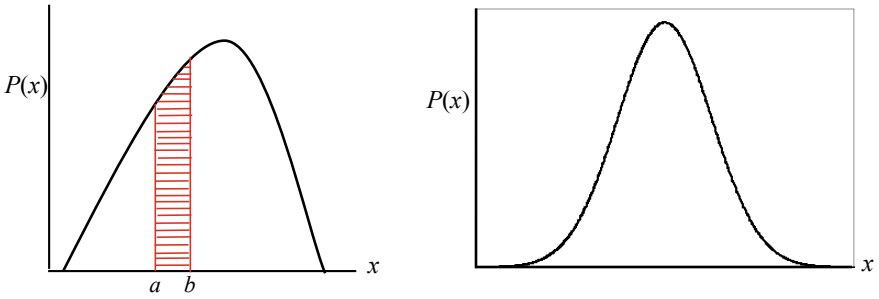
Figure 8.2b shows the change in (8.8), where it is also found to attain its maximum for $\theta = 40$.

A solution in the ML method is called a *maximum likelihood estimate* (*MLE*). The MLE $\theta = 40$ divided by 100 gives 0.4, which equals the proportion of black balls in (8.3). Thus, one may only glance at (8.3) to intuitively conjecture that θ is about 40, without using the ML method. However, when solutions cannot be intuitively conjectured, the ML method serves as an effective parameter estimation methodology.

8.3 Probability Density Function

In the last section, we used an example of cases where a variable can only take *discrete values* as 1 and 0. In the remaining sections of this chapter, we do not treat such discrete variables, but rather those variables taking continuous or almost continuous values.

The probability of a genuinely *continuous variable* being a specific value cannot reasonably be defined. For example, "the probability of a person's stature being exactly 170.0 cm" stands for "the probability of it completely equaling 170.0 cm", which have to be said to be zero. However, the probability can reasonably be defined for the *intervals* of a continuous variable by letting $P(x \pm \delta)$ be the probability of variable x taking the values within the interval of $x - \delta$ to $x + \delta$ with $\delta > 0$. The density of the probability is given by dividing $P(x \pm \delta)$ by the width of interval $\delta - (-\delta) = 2\delta$ as $P(x \pm \delta)/(2\delta)$. The density $P(x) = P(x \pm \delta)/(2\delta)$, in which the width 2δ is reduced to be small enough to be ignored, can be used to express how likely x is to take a specific value, and $P(x)$ is called a *probability density* or the *probability density function* (*PDF*) of variable x. An example of PDF is given in Fig. 8.3a. Its horizontal axis shows the values that x can take and its vertical axis indicates the value of PDF $P(x)$. The following two points should be known about PDF:



(a) An example of functions (b) Normal distribution

Fig. 8.3 Probability density functions

Note 8.1 Probability Density
A probability density stands for how likely it is that a value will be observed; an x value with a *greater* value of probability density $P(x)$ is *more likely* to be observed. For example, $P(a) < P(b)$ in Fig. 8.3a implies that $x = b$ is more likely to occur than $x = a$.

The probability density also has the following property: The *area* below PDF $P(x)$ expresses a probability. In Fig. 7.3a, the probability of x taking the values within the interval $[a, b]$ is indicated by the area with the horizontal lines.

For a variable taking *almost* continuous values, its probability of being a specific value can be reasonably considered. For example, it makes sense to consider "the probability of a test score being 8" for a test whose scores take the integers from 0 to 10. However, such a variable is also usually *treated as a continuous variable* for which a *probability density* is defined, as it is more efficiently analyzed than in cases where it is treated as a discrete variable.

Among a variety of PDFs, the symmetric bell-shaped function shown in Fig. 8.3b is used in a number of univariate statistical procedures. The distribution of x with this PDF is called the *normal distribution* or *Gaussian distribution*, the latter name originating from the German mathematician Gauss (1777–1855), who derived the function. Its generalization is introduced next.

8.4 Multivariate Normal Distribution

For multivariate analysis, a PDF for multiple variables is needed, for example, in order to express how likely a person's stature, weight, and waist measurement are to show the values 170.6 cm, 65.3 kg, and 80.7 cm, respectively. As such a PDF,

$$P(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \tag{8.9}$$

is very often used, where $\mathbf{x} = [x_1, \ldots, x_p]'$ is the $p \times 1$ vectors of p variables, $\boldsymbol{\mu}$ is a $p \times 1$ vector containing fixed values, π ($\cong$3.14) denotes the circle ratio, $\exp\{\bullet\} = e^{\{\bullet\}}$ with e ($\cong$2.72) the base of the natural logarithm, $\boldsymbol{\Sigma}$ is not the symbol of summation but a $p \times p$ *positive-definite* matrix containing fixed values, and $|\Sigma|$ denotes the *determinant* of $\boldsymbol{\Sigma}$. The positive-definiteness and determinant are explained in the next notes.

Note 8.2 Nonnegative and Positive Definite Matrices
A $p \times p$ square matrix $\mathbf{S}$ is said to be *nonnegative definite* if $f(\mathbf{w}) = \mathbf{w}'\mathbf{S}\mathbf{w} \geq 0$ for any vector $\mathbf{w}$. It is known that $\mathbf{S}$ being nonnegative definite and symmetric is equivalent to the property of $\mathbf{S}$ that it can be rewritten as $\mathbf{S} = \mathbf{B}\mathbf{B}'$.

Nonnegative matrix $\mathbf{S}$ is particularly said to be *positive definite* if $f(\mathbf{w}) = \mathbf{w}'\mathbf{S}\mathbf{w} \neq 0$, i.e., $f(\mathbf{w}) > \mathbf{w}'\mathbf{S}\mathbf{w}$ for any vector $\mathbf{w} \neq \mathbf{0}_p$. It is known that any positive definite matrix is *nonsingular*; i.e., its inverse matrix exists, and this matrix is also positive definite.

A *determinant* is defined for any *square* matrix to yield a scalar as a function of the matrix. However, only the determinants of positive-definite matrices are treated in this book, which can be obtained as follows:

Note 8.3 Determinants
Let $\mathbf{S}$ be a $p \times p$ positive-definite matrix whose singular values are $\lambda_1, \ldots, \lambda_p$; the determinant of $\mathbf{S}$ is given as

$$|\mathbf{S}| = \lambda_1 \times \lambda_2 \times \cdots \times \lambda_p. \tag{8.10}$$

The determinant has the following properties:

$$|\mathbf{S}\mathbf{U}| = |\mathbf{S}| \times |\mathbf{U}|, \tag{8.11}$$

$$\left|\mathbf{S}^{-1}\right| = |\mathbf{S}|^{-1}. \tag{8.12}$$

The distribution of $\mathbf{x}$ whose PDF is (8.9) is called a *multivariate normal* (*MVN*) *distribution*. The value of (8.9) for a specified $\mathbf{x}$ can be obtained, with $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ given. We next describe cautions for notations:

Note 8.4 Three Types of Vector Expressions for Data
Until the last chapter,

$$\mathbf{X} = \left[\mathbf{x}_1, \ldots, \mathbf{x}_j, \ldots, \mathbf{x}_p\right] = \begin{bmatrix} \tilde{\mathbf{x}}_1' \\ \vdots \\ \tilde{\mathbf{x}}_i' \\ \vdots \\ \tilde{\mathbf{x}}_p' \end{bmatrix}$$

had been used for an n-individuals $\times$ p-variables data matrix; we had expressed the $p \times 1$ vector for individual i as $\tilde{\mathbf{x}}_i$ with the *tilde* symbol ($\sim$) attached to $\mathbf{x}_i$, in order to distinguish $\tilde{\mathbf{x}}_i$ from the vector $\mathbf{x}_j$ ($n \times 1$) standing for variable j.

The $p \times 1$ vector $\mathbf{x}$ in (8.9) is associated with $\tilde{\mathbf{x}}_i$. However, a *tilde* is not used in (8.9) for the sake of simplicity. We do not attach the *tilde* to the vectors standing for individuals from this chapter; they are expressed as $\mathbf{X} = \begin{bmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_i' \\ \vdots \\ \mathbf{x}_p \end{bmatrix}$. This is the same for the other vectors. Thus, readers should be careful about whether vectors stand for the rows of matrices or their columns.

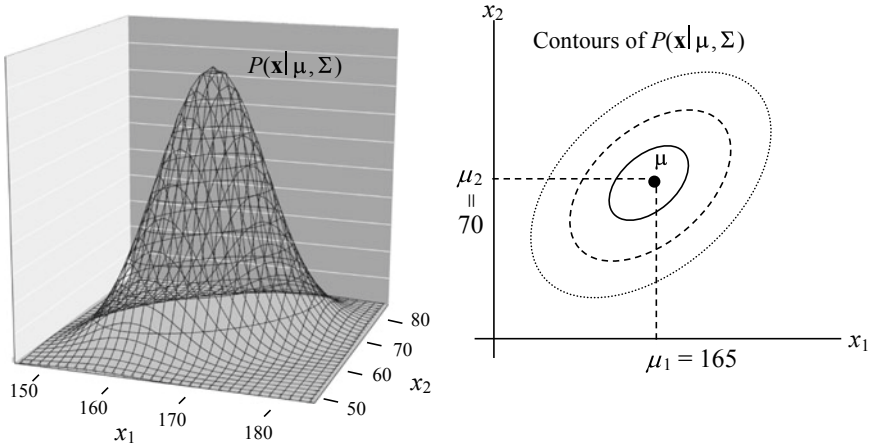
The reason for vector $\mathbf{x}$ in (8.9) not having a subscript is that $\mathbf{x}$ is a *random vector*. This term means that the elements of that vector can take arbitrary values. Thus, $\mathbf{x}_i$ for any i can be substituted into $\mathbf{x}$, with the probability density of $\mathbf{x} = \mathbf{x}_i$ expressed as $P(\mathbf{x}_i)$. An element of a random vector is called a *random variable*.

An expected value for a random variable is introduced next:

Note 8.5 Expected Value
Let us consider a trial in which an infinite number of random vector $\mathbf{x} = [x_1, \ldots, x_p]'$ ($p \times 1$) are observed. The average of those $\mathbf{x}$ is called the *expected vector* of $\mathbf{x}$ and denoted as $E[\mathbf{x}]$, with "E" the abbreviation for "expected". The expected vector $E[\mathbf{x}]$ is $p \times 1$ and expressed as $E[\mathbf{x}] = E[[x_1, \ldots, x_p]'] = [E[x_1], \ldots, E[x_p]]'$. Its jth element $E[x_j]$ is called the *expected value* of random variable x_j. The *expected value* and *vector* are described in more detail in Appendix 8.2.

In particular, it is known that if $\mathbf{x}$ follows the *MVN* distribution with its PDF (8.9), $E[\mathbf{x}]$ and the corresponding inter-variable covariance matrix ($p \times p$) equal $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in (8.9), respectively. This fact is described more exactly in Appendix A.8.4, following Appendices A.8.1–A.8.3 which serve as preparations for A.8.4.

Thus, vector $\mathbf{x}$ with its PDF (8.9) is said to *have* (or *follow*) *the MVN distribution with its mean vector* $\boldsymbol{\mu}$ *and covariance matrix* $\boldsymbol{\Sigma}$. This statement is denoted as



(a) Probability density function (PDF) (b) Bird's-eye view of PDF

Fig. 8.4 Illustration of a multivariate normal distribution for $p = 2$, i.e., $\mathbf{x} = [x_1, x_2]'$

$$\mathbf{x} \sim N_p(\boldsymbol{\mu}, \Sigma), \tag{8.13}$$

where N and its subscript p stand for "*normal distribution*" and the number of variables, respectively. The term "mean" in "mean vector $\boldsymbol{\mu}$" is a synonym of "average" (as mentioned in Sect. 7.5).

The PDF (8.9) with $p = 2$, $\boldsymbol{\mu} = [165, 70]$, and $\Sigma = \begin{bmatrix} 150 & 136 \\ 136 & 159 \end{bmatrix}$ is drawn in Fig. 8.4a. It resembles a *bell*. The vector $\mathbf{x}$ closer to the place corresponding to the top of $P(\mathbf{x}|\boldsymbol{\mu}, \Sigma)$ is more likely to be observed. A bird's-eye view of the distribution in (a) is shown in Fig. 8.4b. There, we can find that the *center* corresponding to the *top* of $P(\mathbf{x}|\boldsymbol{\mu}, \Sigma)$ is the *mean vector* $\boldsymbol{\mu}$. It is surrounded by *ellipses* which express the *contours* of $P(\mathbf{x}|\boldsymbol{\mu}, \Sigma)$; that is, each of the ellipses stands for the terminus of the vector $\mathbf{x}$ providing an equivalent value of $P(\mathbf{x}|\boldsymbol{\mu}, \Sigma)$. The shapes of those ellipses are known to be determined by the *covariance matrix* $\boldsymbol{\Sigma}$. If p is reduced to one, the shape of $P(\mathbf{x}|\boldsymbol{\mu}, \Sigma)$ is equal to that drawn in Fig. 8.3b. If $p \geq 3$, then we need graphs of more than three dimensions, which *cannot be drawn or seen*. But, we can imagine "a bell in a multidimensional space".

8.5 Maximum Likelihood Method for Normal Variables

In Fig. 8.4, the PDF $P(\mathbf{x}|\boldsymbol{\mu}, \Sigma)$ for MVN distribution is illustrated on the assumption that $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are known. But, in practical situations, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are often unknown and $\mathbf{x}$ is observed as specific vectors $\mathbf{x}_i$ ($i = 1, \ldots, n$), for example, as the rows of $\mathbf{X}$ in

Table 8.1 Data matrix showing scores of 11 students × 3 subject tests, with the first *five* and the remaining *six* students belonging to two different classes (artificial example)

		Physics	Chemistry	Biology
$\mathbf{X}$ =	$\mathbf{x}_1'$	80	77	68
	$\mathbf{x}_2'$	65	46	70
	$\mathbf{x}_3'$	82	57	76
	$\mathbf{x}_4'$	66	61	60
	$\mathbf{x}_5'$	73	72	76
	$\mathbf{x}_6'$	79	84	89
	$\mathbf{x}_7'$	89	74	78
	$\mathbf{x}_8'$	67	60	61
	$\mathbf{x}_9'$	91	87	85
	$\mathbf{x}_{10}'$	81	64	72
	$\mathbf{x}_{11}'$	71	73	75

Table 8.1. In this section, we consider *estimating parameters* $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ from an $n \times p$ *data* matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]'$ on the assumption that their row vectors follow the MVN distribution with its average vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$:

$$\mathbf{x}_i \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad (i = 1, \ldots, n). \tag{8.14}$$

For this estimation, we can use the *ML method* introduced in Sect. 7.2. The ML method for continuous variables can be expressed simply by attaching "density" to "probability" in (8.7), as

$$\begin{aligned} &\text{Obtaining the parameter value that } \textit{maximizes} \\ &\text{the } \textit{probability density} \text{ of the } \textit{data} \text{ being observed.} \end{aligned} \tag{8.15}$$

It is because both a probability density and a probability stand for how likely it is that a value will be observed, as described in Sect. 8.2 and Note 8.1.

By substituting $\mathbf{x}_i$ for $\mathbf{x}$ in (8.9), the probability density of $\mathbf{x} = \mathbf{x}_i$ is expressed as

$$P(\mathbf{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right\}. \tag{8.16}$$

For example, the probability density of $\mathbf{x} = \mathbf{x}_1$ in Table 8.1 is

$$P(\mathbf{x}_1|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}\left([80, 77, 68]' - \boldsymbol{\mu}\right)'\boldsymbol{\Sigma}^{-1}\left([80, 77, 68]' - \boldsymbol{\mu}\right)\right\}. \tag{8.17}$$

Mathematical operations for *probabilities* also hold for *probability densities* (e.g., Hogg, McKean, & Craig, 2005). On the supposition that the rows of $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]'$ are observed mutually independently, the *probability density* of the n rows in $\mathbf{X}$ being jointly observed is given by the *product* of (8.16) over $i = 1, \ldots, n$:

$$\begin{aligned} P(\mathbf{X}|\boldsymbol{\mu}, \Sigma) &= \prod_{i=1}^{n} \left\{ \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{ -\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) \right\} \right\} \\ &= \frac{1}{(2\pi)^{np/2}|\boldsymbol{\Sigma}|^{n/2}} \prod_{i=1}^{n} \exp\left\{ -\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) \right\} \\ &= \frac{1}{(2\pi)^{np/2}|\boldsymbol{\Sigma}|^{n/2}} \exp\left\{ -\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) \right\}, \end{aligned} \tag{8.18}$$

with the operator $\prod$ defined as follows:

Note 8.6 Repetition of Products

$$\prod_{i=1}^{m} a_i = a_1 \times a_2 \times \cdots \times a_m$$

The probability density, if it is treated as a function of parameters, is also rephrased as the *likelihood*. That is, (8.18) can be called the likelihood of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ for the data matrix $\mathbf{X}$.

8.6 Maximum Likelihood Estimates of Means and Covariances

The $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ values are obtained in the ML method, such that the data matrix $\mathbf{X}$ is the *most likely* to be observed. That is, the *maximum likelihood estimates* (*MLE*) of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are estimated that *maximizes* (8.18) or its logarithm. This is given by

$$\log P(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{np}{2}\log 2\pi - \frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}). \tag{8.19}$$

Here, $-(np/2)\log 2\pi$ is a *constant irrelevant* to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Thus, the maximization of (8.19) is equivalent to maximizing the function

$$l(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}), \tag{8.20}$$

with the constant term deleted from (8.19). We refer to (8.20) as the *log likelihood* below. As shown in Appendix A.5.1, the MLE of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is given by

$$\hat{\boldsymbol{\mu}} = \overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i, \tag{8.21}$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})' = \mathbf{V}. \tag{8.22}$$

Here, $\mathbf{V}$ is the matrix defined in (3.13) as shown next:

Note 8.7 Another Expression of V
Let us recall (3.13). It can be rewritten as $\mathbf{V} = n^{-1}\mathbf{X}'\mathbf{J}\mathbf{X} = n^{-1}(\mathbf{J}\mathbf{X})'\mathbf{J}\mathbf{X}$, where $\mathbf{J}\mathbf{X}$ contains the centered scores: $\mathbf{J}\mathbf{X} = \begin{bmatrix} (\mathbf{x}_1 - \overline{\mathbf{x}})' \\ \vdots \\ (\mathbf{x}_n - \overline{\mathbf{x}})' \end{bmatrix}$. Thus, $n^{-1}(\mathbf{J}\mathbf{X})'\mathbf{J}\mathbf{X}$ is found to equal $n^{-1}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})'$ in (8.22).

In (8.21) and (8.22), we find that the MLE of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is found to equal the *average vector* and *covariance matrix* obtained from the *data* set, respectively.

Though both $\boldsymbol{\mu}$ and $\overline{\mathbf{x}}$ are referred to as average/mean vectors, and both $\boldsymbol{\Sigma}$ and $\mathbf{V}$ are called covariance matrices, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ differ from $\overline{\mathbf{x}}$ and $\mathbf{V}$, in that the former are the parameters determining $N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$, while $\overline{\mathbf{x}}$ and $\mathbf{V}$ are the statistics obtained from $\mathbf{X}$. However, the MLE of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ equals $\overline{\mathbf{x}}$ and $\mathbf{V}$, respectively, as shown in (8.21) and (8.22) on the assumption of the rows of $\mathbf{X}$ following $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ mutually independently. For distinguishing $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ from $\overline{\mathbf{x}}$ and $\mathbf{V}$, the latter statistics are called a *sample average vector* and a *sample covariance matrix*, respectively.

By substituting MLE (8.21) and (8.22) into the log likelihood (8.20), its maximum is expressed as

$$
\begin{aligned}
l(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) &= -\frac{n}{2}\log|\mathbf{V}| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{V}^{-1}(\mathbf{x}_i - \overline{\mathbf{x}}) \\
&= -\frac{n}{2}\log|\mathbf{V}| - \frac{1}{2}\mathrm{tr}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{V}^{-1}(\mathbf{x}_i - \overline{\mathbf{x}}) \\
&= -\frac{n}{2}\log|\mathbf{V}| - \frac{1}{2}\mathrm{tr}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{V}^{-1} \\
&= -\frac{n}{2}\log|\mathbf{V}| - \frac{n}{2}\mathrm{tr}\mathbf{V}\mathbf{V}^{-1} = -\frac{n}{2}\log|\mathbf{V}| - \frac{np}{2}.
\end{aligned}
\tag{8.23}
$$

8.7 Model Selection

Cases exist for which several models are considered to explain a single data set, as illustrated in Fig. 8.5. *Model selection* refers to comparing models and selecting the *model best fitted* to a data set. An advantage of the ML method is that its MLE can be used for model selection with statistics generally called *information criteria*.

One statistic included in such information criteria was first derived by the Japanese statistician Hirotsugu Akaike (1927–2009). The statistic is known as *Akaike's* (1974) *information criterion* (*AIC*), which is defined as

$$
\mathrm{AIC} = -2l(\hat{\boldsymbol{\Theta}}) + 2\eta \tag{8.24}
$$

for a model in which η is the *number of parameters* to be estimated in the model, $\hat{\boldsymbol{\Theta}}$ stands for a set of MLEs of parameters, and $l(\hat{\boldsymbol{\Theta}})$ expresses the value of the *log likelihood* $l(\boldsymbol{\Theta})$ for $\boldsymbol{\Theta} = \hat{\boldsymbol{\Theta}}$. AIC is defined for each of the models considered for a data set, and the model with a *smaller* AIC value is regarded as the *better* model.

Following AIC, similar statistics have been proposed. Among them, a popular one is Schwarz's (1978) *Bayesian information criterion* (*BIC*), defined as

$$
\mathrm{BIC} = -2l(\hat{\boldsymbol{\Theta}}) + \eta\log n, \tag{8.25}
$$

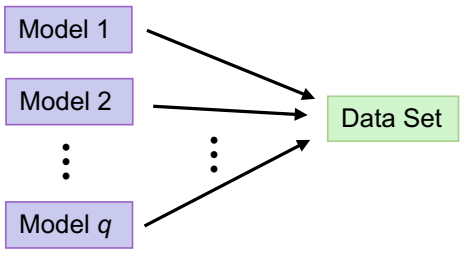


Fig. 8.5 Several models for a data set

with n the number of individuals in a data set. As with AIC, (8.25) is defined for each model, and a smaller value implies that the model is better. It should be noted that both (8.24) and (8.25) *penalize* a model for having *more parameters*, which can be related to the philosophy of science, as in the following note:

Note 8.8 Information Criteria and Philosophy
How information criteria such as (8.24) and (8.25) are derived is beyond the scope of this book. However, the following arguments are to be born in mind:

Let us view Fig. 8.5 by replacing "*model*" with "*theory*" and "*data set*" with "*phenomenon*". In the *philosophy of science*, it had been argued that the *goodness of a theory* should be evaluated by

[1] how *well* it *explains* a phenomenon;
[2] how *simple* (*parsimonious*, in philosophical terms) it is.

(e.g., Hempel, 1966). We can reasonably consider that [1] corresponds to the attained value of the *log likelihood* $l(\hat{\boldsymbol{\Theta}})$ and [2] is associated with the *smallness* of η (*the number of parameters*). Thus, [1] and [2] are found to correspond to smaller values of (8.24) and (8.25). In this sense, *information criteria* can be viewed as a *mathematical validation* of the *philosophical argument*.

Sometimes, the model chosen by AIC is different from that by BIC. For such a case, the model must be chosen by users' subjective consideration. This shows that no absolute index exists for model selection, which should be kept in mind.

8.8 Assessment of Between-Group Heterogeneity

In order to illustrate model selection by information criteria, we consider two models for the data matrix $\mathbf{X}$ in Table 8.1. *Model 1* is expressed as (8.14); all row vectors of $\mathbf{X}$ are assumed to follow an *identical MVN* distribution, $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, in Model 1. On the other hand, let us consider *Model 2* expressed as

$$\mathbf{x}_i \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \text{ for } i = 1, \ldots, 5 \text{ and } \mathbf{x}_i \sim N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) \text{ for } i = 6, \ldots, 11 \tag{8.26}$$

The row vectors for the first five students and those for the remaining six students are assumed to follow *different MVN* distributions in Model 2, where the former and the latter students belong to two different classes.

The MLEs for *Model 1* are given by (8.21) and (8.22) with n = 11, and their values are obtained as in Table 8.2(A). As found there, η (*the number of parameters*) is 3 + 6 = 9, where 3 is the number of elements in $\boldsymbol{\mu}$ and 6 is the number of

Table 8.2 MLE, η (the number of parameters), AIC, and BIC for the data in Table 8.1, with the lower triangular elements omitted in symmetric covariance matrices

Model	(A) Model 1				(B) Model 2							
Parameter	$\hat{\boldsymbol{\mu}}$	$\hat{\boldsymbol{\Sigma}}$			$\hat{\boldsymbol{\mu}}_1$	$\hat{\boldsymbol{\Sigma}}_1$			$\hat{\boldsymbol{\mu}}_2$	$\hat{\boldsymbol{\Sigma}}_2$		
Physics	76.7	73.7	63.9	48.4	73.2	48.6	38.9	22.0	79.7	75.6	52.2	50.9
Chemistry	68.6		136.8	64.6	62.6		121.0	2.0	73.7		94.2	83.2
Biology	73.6			71.9	70.0			35.2	76.7			82.2
η	9				18							
AIC	588.54				189.45							
BIC	592.12				196.61							

different covariances in $\boldsymbol{\Sigma}$; it has 3×3 elements, but the number of different ones is 6, since $\boldsymbol{\Sigma}$ is symmetric. Substituting those MLEs into (8.23), we have $l(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = -\frac{11}{2}\log(202139.9) - \frac{33}{2} = -285.268$. Further, this is substituted into $l(\hat{\boldsymbol{\Theta}})$ in (8.24) and (8.25) to give AIC = $-2\times(-285.268) + 2 \times 9 = 588.54$ and BIC = $-2 \times (-285.268) + 9\log 11 = 592.12$, as shown in Table 8.2(A).

Note 8.9 Base of Logarithm
In this book, "log x" stands for "$\log_e x$", with $e \cong 2.72$ the base of the natural logarithm.

Next, let us obtain the MLE, AIC, and BIC for *Model 2*. On the supposition that the rows of $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]'$ are observed mutually independently, the *probability density* of the n rows in $\mathbf{X}$ being observed jointly is expressed as

$$P(\mathbf{X}|\boldsymbol{\mu}_1, \Sigma_1, \boldsymbol{\mu}_2, \Sigma_2) = \frac{1}{(2\pi)^{5p/2}|\Sigma_1|^{5/2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{5}(\mathbf{x}_i - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_1)\right\} \times \frac{1}{(2\pi)^{6p/2}|\boldsymbol{\Sigma}_2|^{6/2}} \exp\left\{-\frac{1}{2}\sum_{i=6}^{11}(\mathbf{x}_i - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_2)\right\}, \tag{8.27}$$

because of (8.26), where

$$\frac{1}{(2\pi)^{5p/2}|\boldsymbol{\Sigma}_1|^{5/2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{5}(\mathbf{x}_i - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_1)\right\}$$

stands for the probability density of $\mathbf{x}_1, \ldots, \mathbf{x}_5$ being jointly observed, while

$$\frac{1}{(2\pi)^{6p/2}|\boldsymbol{\Sigma}_2|^{6/2}}\exp\left\{-\frac{1}{2}\sum_{i=6}^{11}(\mathbf{x}_i-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_2)\right\}$$

is the probability density for $\mathbf{x}_6, \ldots, \mathbf{x}_{11}$.

The *log likelihood* corresponding to (8.27) can be expressed as

$$l(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1,\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)=l_1(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)+l_2(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2), \tag{8.28}$$

where

$$l_1(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)=-\frac{5}{2}\log|\boldsymbol{\Sigma}_1|-\frac{1}{2}\sum_{i=1}^{5}(\mathbf{x}_i-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_2), \tag{8.29}$$

$$l_2(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)=-\frac{6}{2}\log|\boldsymbol{\Sigma}_2|-\frac{1}{2}\sum_{i=6}^{11}(\mathbf{x}_i-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_2), \tag{8.30}$$

with the constants irrelevant to parameters being deleted.

As found in (8.28), the log likelihood is decomposed into $l_1(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $l_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$. Since they are functions of different sets of parameters, the sets $\{\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1\}$ maximizing $l_1(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\{\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2\}$ maximizing $l_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ are found to be the MLEs that maximize (8.28). By comparing (8.29) and (8.30) with (8.20), we can find that (8.29) or (8.30) is equivalent to log likelihood (8.20), in which $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ have the subscript 1 or 2, and the series $i = 1, \ldots, n$ is replaced by $i = 1, \ldots, 5$ or $i = 6, \ldots, 11$. This fact, along with (8.21) and (8.22), shows that $\boldsymbol{\mu}_1$ and $\boldsymbol{\Sigma}_1$ maximizing $l_1(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ are given by

$$\hat{\boldsymbol{\mu}}_1=\frac{1}{5}\sum_{i=1}^{5}\mathbf{x}_i,\quad \hat{\boldsymbol{\Sigma}}_1=\frac{1}{5}\sum_{i=1}^{5}(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_1)(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_1)'$$

while $\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}_2$ maximizing $l_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ are given by

$$\hat{\boldsymbol{\mu}}_2=\frac{1}{6}\sum_{i=6}^{11}\mathbf{x}_i,\quad \hat{\boldsymbol{\Sigma}}_2=\frac{1}{6}\sum_{i=6}^{11}(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_2)(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_2)'.$$

Those values are shown in Table 8.2(B), whose substitution into (8.28) gives the value of the *maximum log likelihood*:

$$\begin{aligned}\log l\left(\hat{\boldsymbol{\mu}}_1,\hat{\boldsymbol{\Sigma}}_1,\hat{\boldsymbol{\mu}}_2,\hat{\boldsymbol{\Sigma}}_2\right)&=-\frac{5}{2}\log(98{,}328.73)-\frac{15}{2}-\frac{6}{2}\log(36{,}140.64)-\frac{18}{2}\\&=-76.73,\end{aligned}$$

with $\eta = 18$ for Model 2 and $n = 5 + 6 = 11$. Using them in (8.24) and (8.25), we get the AIC and BIC values in Table 8.2(B).

In Table 8.2, both the AIC and BIC are found to show that *Model 2* is better; the 11 students are classified into the two groups characterized by different MVN distributions. It should be kept in mind that comparing the AIC and BIC values is senseless; the comparison is to be made *within the same index*. Comparing AIC values for different data sets as well as BIC values for different data sets is also senseless. A model comparison must be made *for a single data set* (Fig. 8.5).

8.9 Bibliographical Notes

This chapter can serve as a preliminary stage before learning statistical inferences, which are not treated in the present book. *Statistical inferences* refer to the theories in which the relationships of the estimates of parameters to their true values are discussed on the basis of probabilities. One of the established books on elementary statistical inferences was written by Hogg, McKean, and Craig (2019). Books on multivariate statistical inferences include Anderson (2003), Rao (2001), Rencher and Christensen (2012), Seber (1984), and Timm (2002). Searl and Khuri's (2017) book is among the ones in which the matrix algebra for probabilities and statistical inferences is introduced. Detailed treatments of information criteria are found in Konishi (2014) and Konishi and Kitagawa (2008). In those books, properties of maximum likelihood estimates are detailed.

Exercises

8.1. Let $\mathbf{d}$ be an $n \times 1$ data vector whose m elements take one and whose remaining elements are zero, with ω the probability of an element in $\mathbf{d}$ taking one. The likelihood of parameter ω for the data set $\mathbf{d}$ is expressed as

$$P(\omega) = \omega^m (1-\omega)^{n-m}, \tag{8.31}$$

on the supposition that the elements in $\mathbf{d}$ are mutually independently observed. Show that the MLE of ω is given by m/n, using the fact that $d\log P(\omega)/d\omega$, i.e., the differentiation of the logarithm of (8.31) with respect to ω, is zero for the ω value being MLE, with $d\log\omega/d\omega = 1/\omega$ and $d\log(1-\omega)/d\omega = -1/(1-\omega)$.

8.2. The function

$$\phi(x|b,c) = \frac{1}{1+\exp(-bx+c)} \tag{8.32}$$

is called a *logistic function* and is used for relating a continuous variable x to probability $\phi(x|b,c)$. Verify that the function $\phi(x|b,c)$ takes the forms in Fig. 8.6 with $0 \leq \phi(x|b,c) \leq 1$, by substituting some values into x with b and c fixed at specific values.

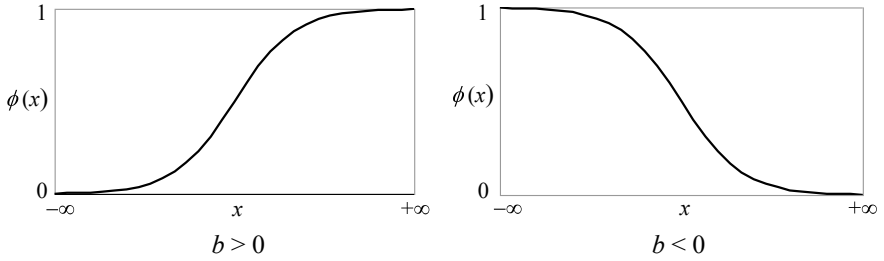


Fig. 8.6 Illustrations of logistic functions

8.3. Let us suppose that the probability of engine i (= 1, …, n) having trouble is expressed as $1/\{1 + \exp(-bx_i + c)\}$ with x_i the value of the variable for i explaining the trouble probability. Show that the likelihood of b and c can be expressed as

$$\prod_{i=1}^{n} \left(\frac{1}{1+\exp(-bx_i+c)}\right)^{d_i} \left(\frac{\exp(-bx_i+c)}{1+\exp(-bx_i+c)}\right)^{1-d_i} \tag{8.33}$$

for observed data x_i and d_i, i = 1, …, n, with d_i = 1 if i has trouble; d_i = 0 otherwise. Here, d_1, …, d_n are assumed to be mutually independently observed.

8.4. Show that the logarithm of (8.33) can be written as

$$\sum_{i=1}^{n} \{(1-d_i)(-bx_i+c) - \log[1+\exp(-bx_i+c)]\}. \tag{8.34}$$

8.5. Let us consider another model for engine trouble in which the probability of engine i (= 1, …, n) having trouble is expressed as $1/\{1 + \exp(-\alpha z_i - \beta x_i + \gamma)\}$, with x_i the one in (8.33) and z_i the value of another explanatory variable for i. The likelihood for this model is expressed as

$$\prod_{i=1}^{n} \left(\frac{1}{1+\exp(-\alpha z_i-\beta x_i+\gamma)}\right)^{d_i} \left(\frac{\exp(-\alpha z_i-\beta x_i+\gamma)}{1+\exp(-\alpha z_i-\beta x_i+\gamma)}\right)^{1-d_i}. \tag{8.35}$$

Let $\hat{b}$ and $\hat{c}$ denote the MLE of b and c for maximizing (8.33) or (8.34). On the other hand, let $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\gamma}$ be the MLE of α, β, and γ for maximizing (8.35). Show that BIC is expressed as

$$-2\sum_{i=1}^{n}\{(1-d_i)(-\hat{b}x_i+\hat{c})-\log[1+\exp(-\hat{b}x_i+\hat{c})]\}+2\log n$$

for (8.33) or (8.34), while it is expressed as

$$-2\sum_{i=1}^{n}\left\{(1-d_i)(-\hat{\alpha}z_i-\hat{\beta}x_i+\hat{\gamma})-\log[1+\exp(-\hat{\alpha}z_i-\hat{\beta}x_i+\hat{\gamma})]\right\}+3\log n$$

for (8.35).

8.6. The similarity of the ML method to a human psychological process was mentioned with an example in Sect. 8.2. Present another example for illustrating the similarity.

8.7. If $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, it is known that $\mathbf{x}+\mathbf{a} \sim N_p(\boldsymbol{\mu}+\mathbf{a}, \boldsymbol{\Sigma})$ for fixed $\mathbf{a}$. Use this fact to show the equivalence between $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{x}=\boldsymbol{\mu}+\mathbf{e}$ with $\mathbf{e} \sim N_p(\mathbf{0}_p, \boldsymbol{\Sigma})$.

8.8. Use the fact that $\left|\sigma^2\mathbf{I}_p\right| = \sigma^{2p}$ to show that the probability density function (PDF) of $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \sigma^2\mathbf{I}_p)$ is the product of $P\left(x_j|\mu_j,\sigma^2\right) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2}\frac{(x_j-\mu_j)^2}{\sigma^2}\right\}$ over $j = 1, \ldots, p$, with x_j the jth element of $\mathbf{x}$, μ_j that of $\boldsymbol{\mu}$, and $P(x_j \mid \mu_j, \sigma^2)$ being the PDF of the (univariate) normal distribution with its mean μ_j and variance σ^2.

8.9. Show that the MLE of σ^2 is given by $\hat{\sigma}^2 = \frac{1}{np}\Sigma_{i=1}^{n}\|\mathbf{x}_i-\overline{\mathbf{x}}\|^2$, if $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \sigma^2\mathbf{I}_p)$, where $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are the $p \times 1$ observed vectors for $\mathbf{x}$ and $\overline{\mathbf{x}} = n^{-1}\Sigma_{i=1}^{n}\mathbf{x}_i$.

8.10. Let us consider the model $\mathbf{x}_i = \mathbf{f}_i(\boldsymbol{\Theta}) + \mathbf{e}_i$ for $p \times 1$ data vectors $\mathbf{x}_i$, $i = 1, \ldots, n$, observed mutually independently, with $\mathbf{e}_i \sim N(\mathbf{0}_p, \sigma^2\mathbf{I}_p)$ and $\mathbf{f}_i(\boldsymbol{\Theta})$ a function of parameter $\boldsymbol{\Theta}$ yielding a $p \times 1$ data vector. Show the equivalence between the MLE of $\boldsymbol{\Theta}$ and the least squares estimate of $\boldsymbol{\Theta}$ minimizing $\Sigma_{i=1}^{n}\|\mathbf{x}_i-\mathbf{f}_i(\Theta)\|^2$, using the facts in Exercises 8.7 to 8.9.

8.11. For n_k vectors $\mathbf{x}_{ki}$ $(p \times 1)$, $i = 1, \ldots, n_k$, observed mutually independently, for group $k = 1, 2, 3$, let us consider the following models:
Model 1. $\mathbf{x}_{ki} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$: All observations follow an identical distribution.
Model 2. $\mathbf{x}_{ki} \sim N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$: Each group has a specific distribution.
Model 3. $\mathbf{x}_{1i} \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathbf{x}_{ki} \sim N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ for $k = 2, 3$: Group 1 differs from 2 and 3.
Express AIC for the models as functions of $\mathbf{x}_{ki}$ and the number of parameters.

8.12. For n_k vectors $\mathbf{x}_{ki}$ $(p \times 1)$, $i = 1, \ldots, n_k$ observed mutually independently, for group $k = 1, \ldots, K$, let us consider the following models:
Model 1. $\mathbf{x}_{ki} \sim N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$: The covariances are homogeneous among groups.
Model 2. $\mathbf{x}_{ki} \sim N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$: The covariances are heterogeneous across groups.

Express BIC for the models as functions of $\mathbf{x}_{ki}$ and the number of parameters, using the facts described in Appendix A5.2.

8.13. For n vectors $\mathbf{x}_i$ $(p \times 1)$, $i = 1, \ldots, n$, observed mutually independently, let us consider the models:
Model 1. $\mathbf{x}_i \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$: The covariances are unconstrained.
Model 2. $\mathbf{x}_i \sim N_p(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_p)$: The covariances are constrained.
Express AIC for the models as functions of $\mathbf{x}_{ki}$ and the number of parameters.

Chapter 9
Path Analysis

Let us assume three variables, A, B, and C, to be analyzed. The regression analysis for predicting C from A and B is based on the causal model, with A and B causes and C the result. However, this model is not guaranteed to indicate the true relationships among A, B, and C. The true causal model might be "A causes B which causes C" or "A causes B and C". *Path analysis* is a procedure in which *users form causal models* by themselves and *select the model* fitted well to a data set. The origins of path analysis can be found in Wright's (1918, 1960) biometric studies and Haavelmo's (1943) econometric ones (Kaplan, 2000).

9.1 From Multiple Regression Analysis to Path Analysis

In this chapter, we use the data set of 60 students by 5 variables in Table 9.1(A). The five variables concern a lecture:

IN: to what extent students were **in**terested in the lecture

KN: the amount of prior **kn**owledge of the lecture subjects

AB: how often students were **ab**sent from the lecture

SH: **s**tudy **h**ours that students took at home for the lecture

RE: **re**cords that students were finally given.

For this data set, the *regression analysis* for predicting RE is modeled as

$$\mathrm{RE} = b_1 \times \mathrm{IN} + b_2 \times \mathrm{KN} + b_3 \times \mathrm{AB} + b_4 \times \mathrm{SH} + c + \mathrm{error}. \tag{9.1}$$

This model can be expressed as the *path diagram* in Fig. 9.1a. There, *double-headed arrows* indicate linked variables being merely *correlated*, and *single-headed arrows* indicate the *causal* relationships; they extend from causes

K. Adachi, *Matrix-Based Introduction to Multivariate Data Analysis*,
https://doi.org/10.1007/978-981-15-4103-2_9

Table 9.1 A data set for five variables for a lecture (artificial example)

	IN	KN	AB	SH	RE
(A) *Raw data*					
1	4	54	13.8	120	82
2	7	68	0	150	96
3	4	66	19.6	90	82
4	4	68	17.5	90	80
5	4	68	35.1	60	70
6	4	66	24	90	58
7	3	76	26.1	30	82
8	3	66	32.2	60	66
9	2	58	41.2	0	40
10	6	70	1.1	150	90
11	6	98	10.6	60	90
12	2	48	48	60	44
13	4	70	11.9	150	98
14	6	76	13.7	120	90
15	3	50	39.7	90	70
16	5	62	11.8	120	96
17	3	52	25.2	60	60
18	2	74	34	0	54
19	3	52	33.1	90	64
20	5	70	13	150	86
21	5	80	9.5	150	88
22	1	56	39.7	0	48
23	7	74	11.5	180	84
24	4	60	15.5	90	80
25	1	64	53.6	0	52
26	5	60	23.4	150	80
27	5	50	16.7	180	74
28	5	66	13.9	90	74
29	5	76	26.2	120	80
30	5	62	10.4	120	88
31	3	64	25.5	60	78
32	3	62	27.4	60	68
33	3	72	37	30	64
34	3	74	22.8	90	90
35	6	68	24.2	180	94
36	3	64	35.8	60	76
37	5	70	16.8	90	94
38	4	58	17.5	90	90
39	2	56	25.2	0	58

(continued)

Table 9.1 (continued)

	IN	KN	AB	SH	RE
40	5	64	9.4	120	90
41	5	66	6.2	120	86
42	3	52	38	30	48
43	6	66	5.8	150	86
44	5	62	19.4	90	86
45	5	82	9.9	30	92
46	3	60	36.4	60	62
47	4	58	24	120	82
48	2	56	32.1	60	60
49	4	58	38.8	60	56
50	2	40	30.7	90	64
51	4	50	31.9	90	72
52	4	64	10.5	120	78
53	3	44	19.8	60	66
54	5	70	9.4	150	82
55	4	50	24.5	120	74
56	4	66	25.6	120	76
57	5	62	26	120	86
58	5	74	15.8	60	90
59	5	64	4.8	90	94
60	4	52	43.3	90	58
Av	4.03	63.47	22.78	90.50	75.77
SD	1.35	9.99	12.08	46.63	14.61
(B) *Centered data* = X					
1	−0.03	−9.47	−8.97	29.50	6.23
2	2.97	4.53	−22.78	59.50	20.23
3	−0.03	2.53	−3.17	−0.50	6.23
4	−0.03	4.53	−5.28	−0.50	4.23
5	−0.03	4.53	12.33	−30.50	−5.77
6	−0.03	2.53	1.23	−0.50	−17.77
7	−1.03	12.53	3.33	−60.50	6.23
8	−1.03	2.53	9.43	−30.50	−9.77
9	−2.03	−5.47	18.43	−90.50	−35.77
10	1.97	6.53	−21.68	59.50	14.23
11	1.97	34.53	−12.18	−30.50	14.23
12	−2.03	−15.47	25.23	−30.50	−31.77
13	−0.03	6.53	−10.88	59.50	22.23
14	1.97	12.53	−9.08	29.50	14.23
15	−1.03	−13.47	16.93	−0.50	−5.77
16	0.97	−1.47	−10.98	29.50	20.23

(continued)

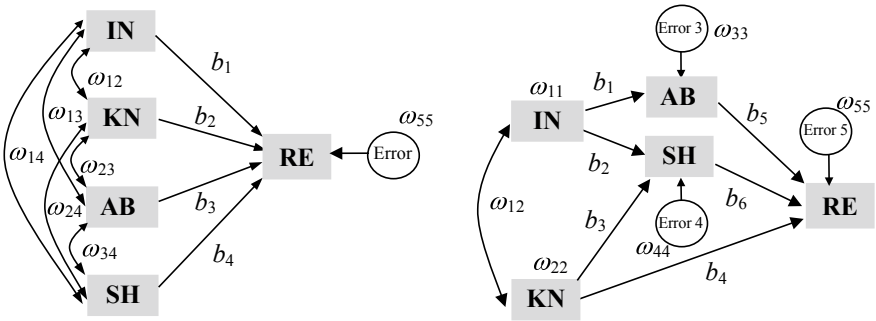
Table 9.1 (continued)

	IN	KN	AB	SH	RE
17	−1.03	−11.47	2.43	−30.50	−15.77
18	−2.03	10.53	11.23	−90.50	−21.77
19	−1.03	−11.47	10.33	−0.50	−11.77
20	0.97	6.53	−9.78	59.50	10.23
21	0.97	16.53	−13.28	59.50	12.23
22	−3.03	−7.47	16.93	−90.50	−27.77
23	2.97	10.53	−11.28	89.50	8.23
24	−0.03	−3.47	−7.28	−0.50	4.23
25	−3.03	0.53	30.83	−90.50	−23.77
26	0.97	−3.47	0.63	59.50	4.23
27	0.97	−13.47	−6.08	89.50	−1.77
28	0.97	2.53	−8.87	−0.50	−1.77
29	0.97	12.53	3.43	29.50	4.23
30	0.97	−1.47	−12.38	29.50	12.23
31	−1.03	0.53	2.73	−30.50	2.23
32	−1.03	−1.47	4.63	−30.50	−7.77
33	−1.03	8.53	14.23	−60.50	−11.77
34	−1.03	10.53	0.03	−0.50	14.23
35	1.97	4.53	1.43	89.50	18.23
36	−1.03	0.53	13.03	−30.50	0.23
37	0.97	6.53	−5.97	−0.50	18.23
38	−0.03	−5.47	−5.28	−0.50	14.23
39	−2.03	−7.47	2.43	−90.50	−17.77
40	0.97	0.53	−13.38	29.50	14.23
41	0.97	2.53	−16.58	29.50	10.23
42	−1.03	−11.47	15.23	−60.50	−27.77
43	1.97	2.53	−16.98	59.50	10.23
44	0.97	−1.47	−3.38	−0.50	10.23
45	0.97	18.53	−12.88	−60.50	16.23
46	−1.03	−3.47	13.63	−30.50	−13.77
47	−0.03	−5.47	1.23	29.50	6.23
48	−2.03	−7.47	9.33	−30.50	−15.77
49	−0.03	−5.47	16.03	−30.50	−19.77
50	−2.03	−23.47	7.93	−0.50	−11.77
51	−0.03	−13.47	9.13	−0.50	−3.77
52	−0.03	0.53	−12.28	29.50	2.23
53	−1.03	−19.47	−2.97	−30.50	−9.77
54	0.97	6.53	−13.38	59.50	6.23
55	−0.03	−13.47	1.73	29.50	−1.77
56	−0.03	2.53	2.83	29.50	0.23

(continued)

Table 9.1 (continued)

	IN	KN	AB	SH	RE
57	0.97	−1.47	3.23	29.50	10.23
58	0.97	10.53	−6.97	−30.50	14.23
59	0.97	0.53	−17.98	−0.50	18.23
60	−0.03	−11.47	20.53	−0.50	−17.77
Av	0.00	0.00	0.00	0.00	0.00
SD	1.35	9.99	12.08	46.63	14.61



(a) Multiple Regression (b) Path Analysis

Fig. 9.1 Multiple regression model and an example of path analysis models for the data in Table 9.1

(*explanatory* variables) to a result (*dependent* variable). That is, regression analysis is based on the causal model with *multiple causes* and *a single result*.

But, *other causal models* may better describe the relationships of variables. An example of other models is shown by the path diagram in Fig. 9.1b, in which it is considered that IN influences RE by way of AB and SH, while KN influences RE directly and by way of SH. In other words, [1] AB is influenced by IN; [2] SH is influenced by IN and KN; and [3] RE is influenced by KN, AB, and SH. These causal relationships are expressed as a set of regression analysis models:

$$\begin{aligned} \text{AB} &= b_1 \times \text{IN} + c_3 + e_3, \\ \text{SH} &= b_2 \times \text{IN} + b_3 \times \text{KN} + c_4 + e_4, \\ \text{RE} &= b_4 \times \text{KN} + b_5 \times \text{AB} + b_6 \times \text{SH} + c_5 + e_5, \end{aligned} \tag{9.2}$$

with c_j and e_j (j = 3, 4, 5) intercepts and errors, respectively. Here, the subscripts 3, 4, and 5 attached to c and e merely correspond to AB, SH, and RE being the third, fourth, and fifth variables. The set of the three equations is equivalent to the path diagram in Fig. 9.1b, where the intercepts are omitted. Parameters b_1, …, b_6 in (9.2) are called *path coefficients*.

In path analysis, the variables are classified into explanatory and dependent variables as follows:

[1] An *explanatory* variable is one to which no single-headed arrow extends in a path diagram; in Fig. 9.1b, IN and KN are explanatory variables. The *errors* e_3, e_4, and e_5 are also included in explanatory variables.
[2] A *dependent* variable is one to which at least a single-headed arrow extends; AB, SH, and RE are dependent variables in Fig. 9.1b.

Explanatory and dependent variables are also called *exogenous* and *endogenous* variables, respectively.

9.2 Matrix Expression

Table 9.1(B) contains the centered scores transformed from the raw scores in (A). It is known that the path analysis for (A) and that for (B) give an identical solution, except for the resulting *intercepts* (c_3, c_4, c_5) being *zero* in the *latter* analysis. We thus *omit* the *intercepts* in the models for path analysis, for the sake of simplicity, supposing that a data set to be analyzed contains *centered scores*. Thus, (9.2) is simplified without c_3, c_4, and c_5 as

$$\begin{aligned} \mathrm{AB} &= b_1 \times \mathrm{IN} + e_3, \\ \mathrm{SH} &= b_2 \times \mathrm{IN} + b_3 \times \mathrm{KN} + e_4, \\ \mathrm{RE} &= b_4 \times \mathrm{KN} + b_5 \times \mathrm{AB} + b_6 \times \mathrm{SH} + e_3. \end{aligned} \tag{9.3}$$

Using matrices and vectors, the three equations in (9.3) are summarized in the following single equation:

$$\begin{array}{c|c|c|c|c} \mathbf{x} & & \mathbf{B} & \mathbf{x} & \mathbf{u} \\ \hline \mathrm{IN} & & & \mathrm{IN} & \mathrm{IN} \\ \mathrm{KN} & & & \mathrm{KN} & \mathrm{KN} \\ \mathrm{AB} & = & b_1 & \mathrm{AB} & + \quad e_3 \\ \mathrm{SH} & & b_2 \quad b_3 & \mathrm{SH} & e_4 \\ \mathrm{RE} & & b_4 \quad b_5 \quad b_6 & \mathrm{RE} & e_5 \end{array} , \tag{9.4}$$

with the blank cells in **B** occupied by zeros. Here, the first and second rows in the left- and right-hand sides of (9.4) stand for "IN = IN" and "KN = KN", which obviously hold true, and the remaining rows are found to equal (9.3). Any model for path analysis can be expressed in the form of (9.4), i.e.,

$$\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{u}, \tag{9.5}$$

for a $p \times 1$ random vector $\mathbf{x}$ for p variables, with the *expected vector* $E[\mathbf{x}]$ for $\mathbf{x}$ supposed to be $\mathbf{0}_p$. This corresponds to the above supposition that a data set to be

analyzed contains centered. The $p \times 1$ $\mathbf{u}$ in (9.5) is a random vector containing p *explanatory* variables, and $\mathbf{B}$ is the $p \times p$ *path coefficient* matrix, in which the (i, j) element being nonzero implies that variable i is influenced by variable j.

Equation (9.5) is rewritten as $\mathbf{x} - \mathbf{Bx} = \mathbf{u}$, thus, $(\mathbf{I}_p - \mathbf{B})\mathbf{x} = \mathbf{u}$. We can further rewrite it as

$$\mathbf{x} = \left(\mathbf{I}_p - \mathbf{B}\right)^{-1}\mathbf{u}, \tag{9.6}$$

supposing the existence of $(\mathbf{I}_p - \mathbf{B})^{-1}$, i.e., that the inverse matrix of $\mathbf{I}_p - \mathbf{B}$ can be obtained. For the model in Fig. 9.1b, (9.6) is expressed in the concrete form:

$$\begin{array}{c|c} \mathbf{x} \\ \hline \text{IN} \\ \text{KN} \\ \text{AB} \\ \text{SH} \\ \text{RE} \end{array} = \begin{array}{|ccccc|} \multicolumn{5}{c}{(\mathbf{I}_p - \mathbf{B})^{-1}} \\ \hline 1 & & & & \\ & 1 & & & \\ -b_1 & & 1 & & \\ -b_2 & -b_3 & & 1 & \\ & -b_4 & -b_5 & -b_6 & 1 \\ \hline \end{array}^{-1} \begin{array}{c|} \mathbf{u} \\ \hline \text{IN} \\ \text{KN} \\ e_3 \\ e_4 \\ e_5 \end{array}. \tag{9.7}$$

9.3 Distributional Assumptions

Let us assume that *explanatory variable* vector $\mathbf{u}$ follows the *multivariate normal* (*MVN*) *distribution* with its mean vector $\mathbf{0}_p$ and covariance matrix $\mathbf{\Omega}$:

$$\mathbf{u} \sim N_p\left(\mathbf{0}_p, \mathbf{\Omega}\right). \tag{9.8}$$

The elements of the covariance matrix are described as

$$\mathbf{\Omega} = \begin{array}{c|ccccc} & \text{IN} & \text{KN} & e_3 & e_4 & e_5 \\ \hline \text{IN} & \omega_{11} & \omega_{12} & & & \\ \text{KN} & \omega_{12} & \omega_{22} & & & \\ e_3 & & & \omega_{33} & & \\ e_4 & & & & \omega_{44} & \\ e_5 & & & & & \omega_{55} \end{array} \tag{9.9}$$

for the model in Fig. 9.1b. Here, the blank cells indicate zero elements, which implies that *errors* are assumed to be *uncorrelated* with *explanatory variables* and that *errors* are assumed to be mutually *uncorrelated*. Those assumptions are found in Fig. 9.1b; they are not linked by paths there.

Here, we introduce a property of MVN variables without its proof:

Note 9.1. A Property of MVN Distribution
If $\mathbf{u}$ is a *random* vector with $\mathbf{u} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Omega})$, then

$$\mathbf{A}\mathbf{u} + \mathbf{c} \sim N_p(\mathbf{A}\boldsymbol{\mu} + \mathbf{c}, \mathbf{A}\boldsymbol{\Omega}\mathbf{A}') \tag{9.10}$$

for fixed $\mathbf{A}$ ($p \times p$) and $\mathbf{c}$ ($p \times 1$).
Here, the difference of random $\mathbf{u}$ to fixed $\mathbf{A}$ and $\mathbf{c}$ should be noted; the elements of $\mathbf{u}$ take a variety of values as $\mathbf{x}$ in Note 8.4, while the elements in $\mathbf{A}$ and $\mathbf{c}$ are constant.

Because of (9.6), (9.8), and (9.10), variable vector $\mathbf{x}$ is found to follow an MVN distribution as follows:

$$\mathbf{x} \sim N_p\left(\mathbf{0}_p, \boldsymbol{\Sigma}\right), \tag{9.11}$$

with its covariance matrix

$$\boldsymbol{\Sigma} = \left(\mathbf{I}_p - \mathbf{B}\right)^{-1} \boldsymbol{\Omega} \left(\mathbf{I}_p - \mathbf{B}\right)^{-1\prime}. \tag{9.12}$$

9.4 Likelihood for Covariance Structure Analysis

Let a 60 (students) $\times$ 5 (variables) data matrix $\mathbf{X}$ contain the centered scores in Table 9.1(B) and $\mathbf{x}_i'$ denote the ith row of $\mathbf{X}$. If $\mathbf{x}_i \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the log likelihood for $\mathbf{X}$ is expressed as (8.20) in Chap. 8. However, in the path analysis model, $\boldsymbol{\mu}$ is restricted to $\mathbf{0}_p$, as in (9.11), with $\boldsymbol{\Sigma}$ constrained as (9.12).

The substitution of $\mathbf{0}_p$ into $\boldsymbol{\mu}$ in (8.20) leads to the *log likelihood* of $\boldsymbol{\Sigma}$ for path analysis:

$$\begin{aligned} l(\boldsymbol{\Sigma}) &= -\frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n} \mathbf{x}_i'\boldsymbol{\Sigma}^{-1}\mathbf{x}_i \\ &= -\frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\mathrm{tr}\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i' \\ &= -\frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{n}{2}\mathrm{tr}\boldsymbol{\Sigma}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\right) = -\frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{n}{2}\mathrm{tr}\boldsymbol{\Sigma}^{-1}\mathbf{V}, \end{aligned} \tag{9.13}$$

where $\boldsymbol{\Sigma}$ is constrained as (9.12) and

$$\mathbf{V} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' = \frac{1}{n}\mathbf{X}'\mathbf{X} \tag{9.14}$$

is the inter-variable covariance matrix for the centered score matrix $\mathbf{X}$.

Let us note that the matrix $\boldsymbol{\Sigma}$ maximizing (9.13) is equivalent to the one maximizing

$$l^*(\boldsymbol{\Sigma}) = \frac{n}{2}\log\left|\boldsymbol{\Sigma}^{-1}\mathbf{V}\right| - \frac{n}{2}\mathrm{tr}\boldsymbol{\Sigma}^{-1}\mathbf{V}, \tag{9.15}$$

since we can use (8.11) and (8.12) in (9.15) to rewrite this as

$$l^*(\boldsymbol{\Sigma}) = \frac{n}{2}\log\left(\left|\boldsymbol{\Sigma}^{-1}\right| \times |\mathbf{V}|\right) - \frac{n}{2}\mathrm{tr}\boldsymbol{\Sigma}^{-1}\mathbf{V} = -\frac{n}{2}\log|\boldsymbol{\Sigma}| + \frac{n}{2}\log|\mathbf{V}| - \frac{n}{2}\mathrm{tr}\boldsymbol{\Sigma}^{-1}\mathbf{V}. \tag{9.15'}$$

Its parts relevant to $\boldsymbol{\Sigma}$ are the same as in (9.13); that is, (9.15) can be regarded as the log likelihood equivalent to (9.13). The former is easier to treat than (9.13) in that the same matrix $\boldsymbol{\Sigma}^{-1}\mathbf{V}$ appears in the determinant and trace. We thus use (9.15) for the log likelihood of $\boldsymbol{\Sigma}$ from here. The log likelihoods for the procedures in Chaps. 10, 11, and 12 are also written in the form of (9.15).

Note 9.2. Covariance Structure Analysis
Likelihood (9.15) is a function of the covariance matrices $\mathbf{V}$ and $\boldsymbol{\Sigma}$ that are obtained from data and derived from a model as in (9.6), respectively. To distinguish the two matrices from one another, the data-based $\mathbf{V}$ is called a *sample covariance matrix*, while the model-based $\boldsymbol{\Sigma}$ is called a *covariance structure*. Further, the path analysis and the procedures in Chaps. 10, 11, and 12 are generally called *covariance structure analysis*, as those procedures share in common log likelihoods that are written in the form of (9.15) and differ only in the covariance structure; it is constrained as (9.12) in the path analysis, while constraints different from (9.12) are imposed upon $\boldsymbol{\Sigma}$ in the other procedures.

9.5 Maximum Likelihood Estimation

Substituting (9.12) into $\boldsymbol{\Sigma}$ in (9.15) leads to the *log likelihood* of *parameters* $\mathbf{B}$ and $\boldsymbol{\Omega}$ for the data matrix $\mathbf{X}$:

$$l^*(\mathbf{B}, \mathbf{\Omega}) = -\frac{n}{2}\log\left|(\mathbf{I}_p - \mathbf{B})'\mathbf{\Omega}^{-1}(\mathbf{I}_p - \mathbf{B})\mathbf{V}\right| - \frac{n}{2}\mathrm{tr}(\mathbf{I}_p - \mathbf{B})'\mathbf{\Omega}^{-1}(\mathbf{I}_p - \mathbf{B})\mathbf{V}. \tag{9.16}$$

Here, we have used the fact that the inverse matrix of (9.12) is expressed as

$$\mathbf{\Sigma}^{-1} = \left\{(\mathbf{I}_p - \mathbf{B})^{-1}\mathbf{\Omega}(\mathbf{I}_p - \mathbf{B})'^{-1}\right\}^{-1} = (\mathbf{I}_p - \mathbf{B})'\mathbf{\Omega}^{-1}(\mathbf{I}_p - \mathbf{B})$$

because of (4.16).

In path analysis, log likelihood (9.16) is maximized; in other words, its negative $-l^*(\mathbf{B}, \mathbf{\Omega})$ is minimized, over $\mathbf{B}$ and $\mathbf{\Omega}$. In model (9.7), the number of parameters to be obtained is 12, since the distinct nonzero elements in $\mathbf{B}$ and $\mathbf{\Omega}$ are $b_1, \ldots, b_6$ and $\omega_{11}, \omega_{22}, \omega_{33}, \omega_{44}, \omega_{55}, \omega_{12}$, respectively, with $\mathbf{\Omega}$ symmetric, i.e., its (1, 2) and (2, 1) elements being the same. Since the solution is not explicitly given, the minimization of $-l^*(\mathbf{B}, \mathbf{\Omega})$ is attained by iterative algorithms. A popular approach is the one using a *gradient algorithm*, which is illustrated in Appendix A.6.3. Setting the vector $\boldsymbol{\theta}$ in A6.3 to $[b_1, \ldots, b_6, \omega_{11}, \omega_{22}, \omega_{33}, \omega_{44}, \omega_{55}, \omega_{12}]'$, the solution can be obtained. We express the solution of $\mathbf{B}$, $\mathbf{\Omega}$, and (9.12) as $\hat{\mathbf{B}}$, $\hat{\mathbf{\Omega}}$, and $\hat{\Sigma}$, respectively.

9.6 Estimated Covariance Structure

For the data set in Table 9.1(B), the *solution* of the path analysis with its model (9.7) is given by

$$\hat{\mathbf{B}} = \begin{array}{c|ccccc} & \text{IN} & \text{KN} & \text{AB} & \text{SH} & \text{RE} \\ \hline \text{IN} & & & & & \\ \text{KN} & & & & & \\ \text{AB} & -6.90 & & & & \\ \text{SH} & 31.94 & -1.65 & & & \\ \text{RE} & & 0.40 & -0.62 & 0.10 & \end{array}, \quad \hat{\mathbf{\Omega}} = \begin{array}{c|ccccc} & \text{IN} & \text{KN} & e_3 & e_4 & e_5 \\ \hline \text{IN} & 1.83 & 6.38 & & & \\ \text{KN} & 6.38 & 99.72 & & & \\ e_3 & & & 58.55 & & \\ e_4 & & & & 706.86 & \\ e_5 & & & & & 51.36 \end{array} \tag{9.17}$$

Figure 9.2a presents a path diagram with the values in the above solution shown in the corresponding parts. The *GFI statistic*, defined as

$$\mathrm{GFI} = 1 - \frac{\mathrm{tr}(\hat{\mathbf{\Sigma}}^{-1}\mathbf{V} - \mathbf{I}_p)^2}{\mathrm{tr}(\hat{\mathbf{\Sigma}}^{-1}\mathbf{V})^2}, \tag{9.18}$$

is convenient for assessing whether a solution is satisfactory or not. Index (9.18) indicates the closeness of the sample covariance matrix $\mathbf{V}$ and the estimated covariance structure

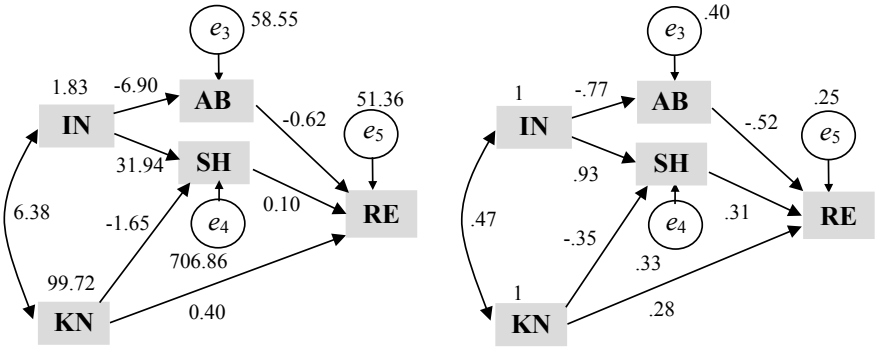


Fig. 9.2 Solution of the model in Fig. 9.1b for the data in Table 9.1

$$\hat{\boldsymbol{\Sigma}} = \left(\mathbf{I}_p - \hat{\mathbf{B}}\right)^{-1} \hat{\boldsymbol{\Omega}} \left(\mathbf{I}_p - \hat{\mathbf{B}}\right)^{-1'}, \tag{9.19}$$

i.e., (9.12) in which the solutions $\hat{\mathbf{B}}$ and $\hat{\boldsymbol{\Omega}}$ have been substituted. If $\hat{\boldsymbol{\Sigma}} = \mathbf{V}$, which implies that a model is fitted completely to a data set with $\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{V} = \mathbf{I}_p$, then (9.18) attains the one at the upper limit; the largeness of the GFI stands for how well solution-based covariances approximate sample covariances. A value of 0.9 is sometimes used as a benchmark with a GFI $\geq$ 0.9 showing a satisfactory model, though selecting 0.9 does not have any theoretical rationale.

The sample covariance matrix for the data in Table 9.1 and the estimated covariance structure for the solution in Fig. 9.2a are given as

$\mathbf{V} =$

	IN	KN	AB	SH	RE
IN	1.83				
KN	6.38	99.72		*Sym*	
AB	-12.65	-51.10	145.87		
SH	47.98	39.27	-350.54	2174.75	
RE	15.34	75.14	-144.42	443.12	213.48

, $\hat{\boldsymbol{\Sigma}} =$

	IN	KN	AB	SH	RE
IN	1.83				
KN	6.38	99.72		*Sym*	
AB	-12.65	-44.08	145.87		
SH	47.98	39.27	-331.26	2174.75	
RE	15.01	70.81	-139.75	431.23	207.71

respectively, where $\hat{\boldsymbol{\Sigma}}$ has been obtained by substituting solution (9.17) in (9.19), and the upper triangular elements in $\mathbf{V}$ and $\hat{\boldsymbol{\Sigma}}$ have been omitted by writing "*Sym*", since they are *symmetric*. The above $\mathbf{V}$ and $\hat{\boldsymbol{\Sigma}}$ are substituted in (9.18) to give a GFI of 0.984, which is higher than 0.9, suggesting that the solution in Fig. 9.2a is satisfactory.

9.7 Unstandardized and Standardized Solutions

The result in Fig. 9.2a is called *unstandardized solution*, as it is obtained from unstandardized data with variables having different variances. Thus, it is senseless to compare the largeness of the resulting parameter values. For the comparison to make sense, we must obtain the *standardized solution* obtained for the standard scores transformed from the raw data.

The unstandardized and standardized solutions can be considered two *different expressions of the same solution*, as the maximum value of log likelihood (9.16) is equivalent between unstandardized and standardized solutions, which is shown in Appendix A.7. This equivalence is called *scale invariance*: The path analysis solution can be said to be *scale invariant*. This property leads to the equivalence of the value of GFI (9.18) between both solutions, as shown by (A.7.10) in Appendix A.7. Furthermore, the standardized solutions of $\mathbf{B}$ and $\mathbf{\Omega}$, which we denote $\tilde{\mathbf{B}}$ and $\tilde{\mathbf{\Omega}}$, can be obtained straightforwardly from the unstandardized solutions $\hat{\mathbf{B}}$ and $\hat{\mathbf{\Omega}}$ with the simple transformations:

$$\tilde{\mathbf{B}} = \mathbf{D}^{-1}\hat{\mathbf{B}}\mathbf{D} \text{ and } \tilde{\mathbf{\Omega}} = \mathbf{D}^{-1}\hat{\Omega}\mathbf{D}^{-1}. \tag{9.20}$$

This fact is also shown in Appendix A.7 with (A.7.20).

The standardized solutions transformed from (9.17) by (9.20) are shown in Fig. 9.2b. There, it makes sense to compare the parameter values. For example, we can find AB to be the most influential for RE among the three explanatory variables AB, SH, and KN that extend paths to RE, since the absolute value of the coefficient (−0.52) attached to the path from AB to RE is the largest. Further, the sign of that coefficient is negative, implying that AB tends to considerably decrease RE. The covariance $\tilde{\omega}_{12} = 0.47$ in the standardized solution $\tilde{\mathbf{\Omega}} = (\tilde{\omega}_{kl})$ is viewed as the correlation coefficient, since all variables are standard scores.

9.8 Other and Extreme Models

Let us refer to the model in Fig. 9.1b as Model 1. Although this model was regarded as satisfactory, with a GFI exceeding 0.9, a model may exist that is better fitted to the data set in Table 9.1(B). This suggests that *other models* should be considered and *compared*; that is, the *model selection* illustrated in Fig. 8.5 (Sect. 8.7) is to be performed. Figure 9.3 shows two examples of other models, which we call Models 2 and 3. In Model 2, one path is added to Model 1 in Fig. 9.1b. On the other hand, in Model 3, one path deleted from Model 1. For *Model 2*, (9.6) is expressed as

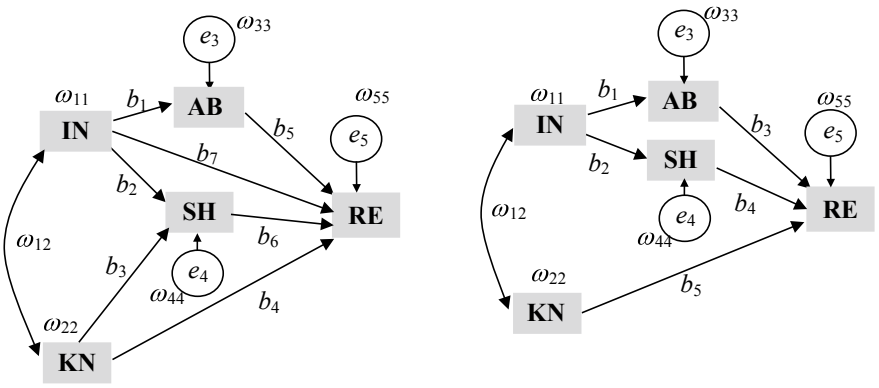


Fig. 9.3 Examples of models that differ from the one in Fig. 9.1b

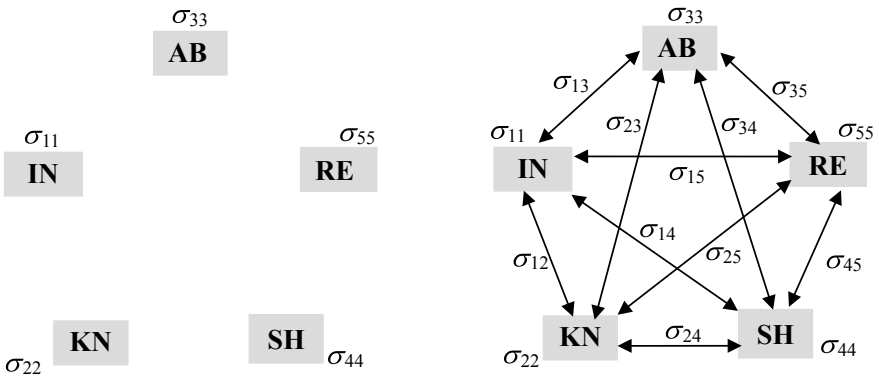
$$\begin{array}{c|c} \mathbf{x} & \\ \hline \text{IN} & \\ \text{KN} & \\ \text{AB} & \\ \text{SH} & \\ \text{RE} & \end{array} = \begin{array}{|ccccc|} \multicolumn{5}{c}{(\mathbf{I}-\mathbf{B})^{-1}} \\ \hline 1 & & & & \\ & 1 & & & \\ -b_1 & & 1 & & \\ -b_2 & -b_3 & & 1 & \\ -b_7 & -b_4 & -b_5 & -b_6 & 1 \\ \hline \end{array}^{-1} \begin{array}{c|} \mathbf{u} \\ \hline \text{IN} \\ \text{KN} \\ e_3 \\ e_4 \\ e_5 \end{array} . \tag{9.21}$$

Here, a parameter, b_7, is added to (9.7): that model has one more parameters than Model 1. The covariance matrix among explanatory variables is the same as that in (9.9). Except for the difference between (9.7) and (9.21), the same procedure is performed for Model 2: The *maximum likelihood method* gives the solutions for Model 2 and other possible models.

Now, let us consider two types of extreme models. One is the *independent model* shown in Fig. 9.4a, where we find that no variable is linked to the others. It implies that all variables are assumed to be mutually independent. This model is the most restrictive, with its *number of parameters* the *least* among possible models. That number is p, i.e., the *number of variables*; only their variances are to be estimated, which are denoted as σ_{jj} ($j = 1, \ldots, p$) in Fig. 9.4a.

The other extreme type is called the *saturated model*, whose number of parameters equals $p(p + 1)/2$, the number of the *distinctive covariances* in $\mathbf{V}$ or $\mathbf{\Sigma} = (\sigma_{jk})$; this is 15 for the data set in Table 9.1(B). This number is the *maximum* among those for all possible models. The saturated models contain several ones, and a typical saturated model is shown in Fig. 9.4b, where all variables are connected by double-headed arrows, implying that *all variables* are assumed to be *merely correlated*. That is, the model in Fig. 9.4b *states nothing* for the causal relationships among the variables.

The covariance structures of the independent and saturated models are expressed as



(a) Independent model (b) Saturated model

Fig. 9.4 Two extreme models with the least and the most parameters

$$\boldsymbol{\Sigma} = \begin{array}{c|ccccc} & \text{IN} & \text{KN} & \text{AB} & \text{SH} & \text{RE} \\ \text{IN} & \sigma_{11} & & & & \\ \text{KN} & & \sigma_{22} & & & \\ \text{AB} & & & \sigma_{33} & & \\ \text{SH} & & & & \sigma_{44} & \\ \text{RE} & & & & & \sigma_{55} \end{array}, \quad \boldsymbol{\Sigma} = \begin{array}{c|ccccc} & \text{IN} & \text{KN} & \text{AB} & \text{SH} & \text{RE} \\ \text{IN} & \sigma_{11} & & & & \\ \text{KN} & \sigma_{21} & \sigma_{22} & & \textit{Sym} & \\ \text{AB} & \sigma_{31} & \sigma_{32} & \sigma_{33} & & \\ \text{SH} & \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} & \\ \text{RE} & \sigma_{51} & \sigma_{52} & \sigma_{53} & \sigma_{54} & \sigma_{55} \end{array}$$

respectively. The former is a *diagonal* matrix, while the latter is a simple unconstrained covariance matrix *without a special structure*.

9.9 Model Selection

So far, we have Models 1, 2, and 3, and two extreme models. For comparing those five models with respect to the *goodness of fit* to the data set, we *cannot use GFI*, since the GFI values *increase* with the *number of parameters* in the models, and the GFI for the *saturated models* always attains the upper limit. This can be found in Table 9.2, where the models are arranged according to their numbers of parameters.

Table 9.2 Number of parameters (NP) and the resulting index values for each model

Model	NP	GFI	AIC	BIC
Saturated	15	1.000	30.000	61.415
Model 2	13	0.987	28.035	55.262
Model 1	12	0.984	26.364	51.496
Model 3	11	0.908	39.792	62.830
Independent	5	0.389	231.480	241.952

This property of GFI is due to the fact that the number of parameters is not considered for defining the GFI, as found in (9.18). GFI is thus *only* useful for assessing whether *a considered model* is satisfactory or not.

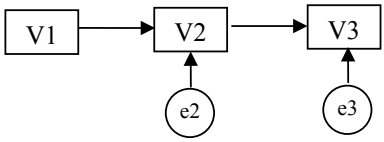
The *information criteria* introduced in Sect. 8.7 are useful for *comparing models*, since the number of parameters is considered in the criteria. The typical information criteria *AIC* and *BIC* for path analysis are obtained by substituting the maximum (9.16) value $l^*(\hat{\mathbf{B}}, \hat{\Omega})$ into $l(\hat{\boldsymbol{\Theta}})$ in (8.24) and (8.25). The AIC and BIC values for each model are shown in Table 9.2. Since smaller values of the information criteria indicate better models, Model 1, for which both the AIC and BIC show the lowest values, is found to be the best of the five models. Different from this example, cases often arise when AIC and BIC indicate different models are best.

9.10 Bibliographical Notes

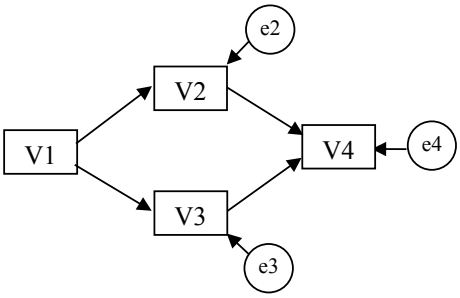
It is difficult to find books in which path analysis is exclusively treated. It is, however, detailed in chapters of books for structural equation modeling, which include Bollen (1989) and Kaplan (2000).

Exercises

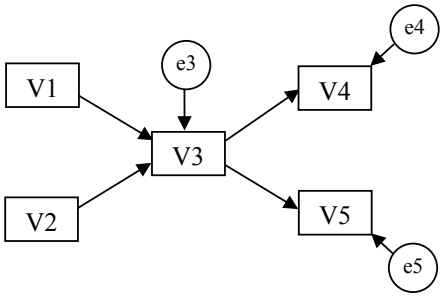
9.1. Present an example of a set of the variables V1, V2, and V3 whose relationships are represented as the following path diagram:



9.2. Present an example of a set of the variables V1, V2, V3, and V4 whose relationships are represented as the following path diagram:



9.3. Present an example of a set of the variables V1, …, V5 whose relationships are represented as the following path diagram:



9.4. Let the elements of the $p \times 1$ vector $\mathbf{x}$ in (9.5) denoted as $\mathbf{x} = [z_1, \ldots, z_q, y_1, \ldots, y_r]' = [\mathbf{z}', \mathbf{y}']'$, with $\mathbf{z} = [z_1, \ldots, z_q]'$ the $q \times 1$ vector containing explanatory variables, $\mathbf{y} = [y_1, \ldots, y_r]'$ the $r \times 1$ vector consisting of dependent variables, and $p = q + r$. Show that the path analysis model (9.5) can be rewritten as

$$\mathbf{y} = \mathbf{A}\mathbf{y} + \mathbf{C}\mathbf{z} + \mathbf{e}, \tag{9.27}$$

with $\mathbf{A}$ $(r \times r)$ and $\mathbf{C}$ $(r \times q)$ containing path coefficients.

9.5. Rewrite the diagram in Fig. 9.1b using the elements of $\mathbf{A}$ and $\mathbf{C}$ in (9.27).

9.6. Let the path analysis model be expressed as (9.27) with $\mathbf{z} \sim N_q(\mathbf{0}_q, \mathbf{\Phi})$, $\mathbf{e} \sim N_r(\mathbf{0}_r, \mathbf{\Psi})$, and no correlation found between $\mathbf{z}$ and $\mathbf{e}$, where $\mathbf{\Psi}$ is an $r \times r$ diagonal matrix. Then, the fact is known that the $q \times r$ covariance matrix between the q explanatory variables in $\mathbf{z}$ and the r dependent ones in $\mathbf{y}$ is given by $\mathbf{\Phi}\mathbf{C}'(\mathbf{I}_r - \mathbf{A})'^{-1}$. Using this fact, show that the covariance structure (9.12) can be rewritten as

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Phi} & \mathbf{\Phi}\mathbf{C}'(\mathbf{I}_r - \mathbf{A})'^{-1} \\ (\mathbf{I}_r - \mathbf{A})^{-1}\mathbf{C}\mathbf{\Phi} & (\mathbf{I}_r - \mathbf{A})^{-1}(\mathbf{C}\mathbf{\Phi}\mathbf{C}' + \mathbf{\Psi})(\mathbf{I}_r - \mathbf{A})'^{-1} \end{bmatrix}, \tag{9.28}$$

where (9.28) is one of the block matrices which are detailed in Sect. 14.1.

9.7. For an $n \times p$ centered data matrix, the independent model can formally be expressed as $\mathbf{x} \sim N_p(\mathbf{0}_p, \mathbf{\Sigma})$ with $\mathbf{\Sigma} = (\sigma_{jk})$ being constrained as a diagonal matrix. Show that the PDF of $\mathbf{x} = [x_1, \ldots, x_p]'$ in this model is expressed as

$$P(\mathbf{x}|\mathbf{\Sigma}) = \prod_{j=1}^{p} \frac{1}{\sqrt{2\pi\sigma_{jj}}} \exp\left\{-\frac{x_j^2}{2\sigma_{jj}}\right\}, \tag{9.29}$$

using the fact that $|\mathbf{\Sigma}| = \prod_{j=1}^{p} \sigma_{jj}$ if $\mathbf{\Sigma} = (\sigma_{jk})$ is diagonal.

9.8. Show that the MLE of $\mathbf{\Sigma}$ in the independent model treated in Exercise 9.7 is given by the diagonal matrix whose diagonal elements are those of

$\mathbf{V} = n^{-1}\mathbf{XX}$, with $\mathbf{X}$ the $n \times p$ centered data matrix whose rows are filled with $\mathbf{x}'$ for n individuals.

9.9. Let us consider model (9.27) with $\mathbf{z} \sim N_q(\mathbf{0}_q, \mathbf{\Phi})$, $\mathbf{e} \sim N_r(\mathbf{0}_r, \mathbf{\Psi})$, and no correlation found between $\mathbf{z}$ and $\mathbf{e}$, where $\mathbf{\Psi}$ is an $r \times r$ diagonal matrix. If the jth variable in $\mathbf{y}$ cannot be a cause for the (j-1)th variable, the log likelihood of the parameters in (9.27) for the $n \times p$ centered data matrix $\mathbf{X} = [\mathbf{Z}, \mathbf{Y}]$, whose rows are filled with the observations of $\mathbf{x}' = [\mathbf{z}', \mathbf{y}']'$, is known to be given by

$$\begin{aligned} \log l(\mathbf{H}, \mathbf{\Psi}, \mathbf{\Phi}) &= -\frac{n}{2}\left\{\log|\mathbf{\Psi}| + \mathrm{tr}\mathbf{M}\mathbf{\Psi}^{-1} + \log|\mathbf{\Phi}| + \mathrm{tr}\mathbf{V}_{ZZ}\mathbf{\Phi}^{-1}\right\} \\ &= -\frac{n}{2}\left\{\log|\mathbf{\Psi}| + \mathrm{tr}(\mathbf{H}\mathbf{V}_{XX}\mathbf{H}' - 2\mathbf{V}_{XX}\mathbf{H}')\mathbf{\Psi}^{-1} + \mathrm{tr}\mathbf{V}_{YY}\mathbf{\Psi}^{-1} + \log|\mathbf{\Phi}| + \mathrm{tr}\mathbf{V}_{ZZ}\mathbf{\Phi}^{-1}\right\} \end{aligned} \tag{9.28}$$

(Adachi, 2014). Here, $\mathbf{H} = [\mathbf{C}, \mathbf{A}]$ ($r \times p$), $\mathbf{V}_{YY} = n^{-1}\mathbf{Y}'\mathbf{Y}$, $\mathbf{V}_{YX} = n^{-1}\mathbf{Y}'\mathbf{X}$, $\mathbf{V}_{XX} = n^{-1}\mathbf{X}'\mathbf{X}$, $\mathbf{V}_{ZZ} = n^{-1}\mathbf{Z}'\mathbf{Z}$, and $\mathbf{M} = \mathbf{H}\mathbf{V}_{XX}\mathbf{H}' - 2\mathbf{V}_{YX}\mathbf{H}' + \mathbf{V}_{YY}$, with $\mathbf{Z}$ ($n \times q$) and $\mathbf{Y}$ ($n \times r$) the blocks of $\mathbf{X}$. Show that the two terms $\mathrm{tr}(\mathbf{H}\mathbf{V}_{XX}\mathbf{H}' - 2\mathbf{V}_{YX}\mathbf{H}')\mathbf{\Psi}^{-1}$ and $\log|\mathbf{\Psi}| + \mathrm{tr}\mathbf{M}\mathbf{\Psi}^{-1}$ in (9.28) can be rewritten as

$$\mathrm{tr}(\mathbf{H}\mathbf{V}_{XX}\mathbf{H}' - 2\mathbf{V}_{YX}\mathbf{H}')\mathbf{\Psi}^{-1} = \sum_{i=1}^{r}\frac{1}{\psi_i}\left(\sum_{j=1}^{p} v_{jj}h_{ij}^2 + 2\sum_{j=1}^{p}\sum_{k\neq j}^{p} v_{jk}h_{ij}h_{ik} - 2\sum_{j=1}^{p} w_{ij}h_{ij}\right) \tag{9.29}$$

$$\log|\mathbf{\Psi}| + \mathrm{tr}\mathbf{M}\mathbf{\Psi}^{-1} = \sum_{i=1}^{r}(\log\psi_i + \frac{m_{ii}}{\psi_i}), \tag{9.30}$$

with $\mathbf{V}_{XX} = (v_{jk})$, $\mathbf{V}_{YX} = (w_{ij})$, $\mathbf{H} = (h_{ij})$, m_{ii} the ith diagonal element of $\mathbf{M}$, and ψ_i that of $\mathbf{\Psi}$. For (9.30), use the fact that $|\mathbf{D}| = d_1 \times \cdots \times d_r$ if $\mathbf{D}$ is the $r \times r$ diagonal matrix whose diagonal elements are $d_1, \ldots, d_r$.

9.10. Let us consider maximizing (9.28). The MLE of $\mathbf{\Phi}$ is explicitly given by $\mathbf{\Phi} = \mathbf{V}_{ZZ}$, but the MLE of the nonzero elements in $\mathbf{H}$ and $\mathbf{\Psi}$ must be obtained by an iterative algorithm. Use (9.29) and (9.30) to show that the algorithm can be formed by the following steps:

Step 1. Initialize the nonzero elements of $\mathbf{H}$.
Step 2. Set $\psi_i = m_{ii}$ for $i = 1, \ldots, r$.
Step 3. Repeat updating h_{ij} as $h_{ij} = \frac{1}{v_{jj}}(w_{ij} - \sum_{k\neq j}^{p} v_{jk}h_{ik})$ over all indexes i and j for the nonzero elements in $\mathbf{H}$.
Step 4. Finish if convergence is reached; otherwise, go back to Step 2.

The hint for Step 2 can be found in Exercise 8.1.

9.11. Show that the minimization of $\|\mathbf{XD} - \mathbf{FA}'\|^2$ over $\mathbf{F}$ and $\mathbf{A}$ gives an essentially different solution from that of (5.4), which implies that the solutions of principal component analysis do not have scale invariance. Here, $\mathbf{X}$ is an $n \times p$ centered data matrix the number of the column of $\mathbf{F}$ is not greater than

$\min(n, p)$, and $\mathbf{D}$ is a $p \times p$ diagonal matrix whose diagonal elements are all positive and take mutually different values.

9.12. Show that the k-means clustering (KMC) for an $n \times p$ data matrix $\mathbf{X}$ gives an essentially different solution from that for $\mathbf{XD}$, which implies that the KMC solutions do not have scale invariance, with $\mathbf{D}$ a diagonal matrix whose $p \times p$ diagonal elements are all positive and take mutually different values.

Chapter 10
Confirmatory Factor Analysis

Let the positive correlations be observed among the test scores for physics, chemistry, and biology. In order to investigate the causal relationships among the three variables, we can use the *path analysis* from the previous chapter. For example, we can evaluate the model in which a person's ability in physics influences his/her scores in chemistry and biology; ability in physics is a cause, while the scores in chemistry and biology are the results. However, it may be rather reasonable to assume that *all* of the scores for physics, chemistry, and biology are *the results of a single factor*, namely "an aptitude for the natural sciences". This is the idea underlying *factor analysis* (*FA*). British psychologist Spearman (1904) had such a conception in his studies of human intelligence, which is the origin of FA. Its key point is that all *observed variables* are regarded as the *results* caused by a few *unobserved latent variables* called *common factors*, in contrast to path analysis, in which causal relationships among observed variables are modeled.

FA can be classified into *exploratory FA* (*EFA*), *confirmatory FA* (*CFA*), and sparse *FA*, where sparse FA is beyond the scope of Part III and treated in Part V. EFA refers to the FA procedures for exploring common factors underlying observed variables for cases without prior knowledge of underlying common factors (Thurstone, 1935, 1947). In contrast, CFA refers to the procedures for confirming a model describing the relationships of common factors to variables (Jöreskog, 1969). Historically, the development of EFA preceded that of CFA, and EFA is often simply called "factor analysis". However, *CFA* is dealt with in this chapter, as introducing CFA before EFA suits the context of this book and CFA is easier to understand than EFA.

K. Adachi, *Matrix-Based Introduction to Multivariate Data Analysis*,
https://doi.org/10.1007/978-981-15-4103-2_10

10.1 Example of Confirmatory Factor Analysis Model

We use the data set of 100 (participants) by 8 (behavioral features) in Table 10.1a containing the self-ratings evaluating to what extent participants' behaviors are characterized by eight variables (features): A (Aggressive), C (Cheerful), I (Initiative), B (Blunt), T (Talkative), V (Vigor), H (tendency to Hesitate), and P (being Popular). For these eight variables, we consider the model with the assumption that A, I, V, and H are caused by *a common factor* (*Factor*_1) interpreted as an *activity*, while C, B, T, and P are caused by *another common factor* (*Factor*_2) that stands for *sociability*. The model is expressed as a set of eight equations:

$$\begin{aligned}
\mathrm{A} &= a_1 \times Factor_1 + c_1 + e_1 \\
\mathrm{I} &= a_2 \times Factor_1 + c_2 + e_2 \\
\mathrm{V} &= a_3 \times Factor_1 + c_3 + e_3 \\
\mathrm{H} &= a_4 \times Factor_1 + \mathrm{c}_4 + e_4 \\
\mathrm{C} &= a_5 \times Factor_2 + c_5 + e_5 \\
\mathrm{B} &= a_6 \times Factor_2 + c_6 + e_6 \\
T &= a_7 \times Factor_2 + c_7 + e_7 \\
P &= a_8 \times Factor_2 + c_8 + e_8
\end{aligned} \tag{10.1}$$

Here, c_j and e_j (j = 1, …, 8) express an intercept and an error, respectively. Each equation in (10.1) is a model for *regression analysis*, though the *factor* is not an observed but rather an unobserved *latent* random variable. The model in (10.1) can be represented as the path diagram in Fig. 10.1.

10.2 Matrix Expression

Table 10.1b shows the centered scores for the raw data in (a). As in path analysis (Chap. 9), CFA for (a) and (b) produces the same solution, except for the resulting *intercepts* (c_1, …, c_8) being *zero* in the *latter* analysis. We thus *omit* the *intercepts* in CFA models, for the sake of simplicity, on the supposition that a data matrix to be analyzed contains *centered scores*. The model in (10.1) without intercepts can be expressed in the matrix form

Table 10.1 Personality data: self-ratings for behavioral features (artificial example)

	A	C	I	B	T	V	H	P		A	C	I	B	T	V	H	P
(a) Raw data																	
1	9	7	9	2	9	8	3	8	51	6	6	7	3	6	5	6	7
2	2	3	5	8	1	3	7	3	52	6	6	8	4	8	8	3	9
3	5	6	7	6	8	4	6	6	53	7	6	8	4	8	6	4	7
4	4	6	6	3	8	5	7	7	54	7	6	7	5	7	6	5	7
5	6	5	7	6	6	5	6	6	55	4	4	6	8	2	3	7	3
6	4	5	5	5	6	3	5	5	56	7	6	8	3	7	8	2	7
7	6	7	6	5	8	3	6	8	57	7	5	8	6	7	6	2	6
8	6	6	7	4	8	7	6	7	58	6	6	6	3	6	4	3	4
9	7	6	8	5	6	5	3	4	59	6	5	7	6	5	5	3	7
10	4	4	6	8	4	3	6	3	60	4	4	7	6	4	5	5	4
11	5	6	6	4	6	4	5	7	61	4	6	6	5	6	5	9	6
12	6	4	6	5	5	6	4	6	62	6	4	5	7	4	3	5	4
13	7	5	8	5	5	6	7	6	63	4	5	6	6	6	4	7	5
14	4	5	6	7	5	4	7	5	64	2	6	5	5	5	2	9	5
15	3	6	5	6	4	2	6	5	65	5	6	6	6	7	6	5	6
16	5	6	7	3	9	5	7	7	66	4	5	4	7	5	3	7	5
17	4	5	8	4	8	5	5	6	67	4	6	6	4	7	4	6	6
18	7	6	8	4	6	6	4	6	68	5	4	7	6	6	8	4	4
19	5	7	7	4	9	5	4	7	69	6	6	7	4	7	4	5	8
20	5	5	7	5	4	4	6	8	70	4	6	5	7	5	4	6	6
21	6	7	7	4	8	6	6	6	71	6	5	7	5	5	6	4	6
22	4	6	6	4	5	3	7	5	72	6	7	8	5	7	7	4	7
23	3	6	5	7	4	3	8	2	73	6	5	8	5	6	6	2	5

(continued)

Table 10.1 (continued)

	A	C	I	B	T	V	H	P		A	C	I	B	T	V	H	P
24	7	7	8	5	7	8	5	6	74	5	6	6	6	6	2	8	5
25	4	6	7	6	5	4	5	6	75	6	6	8	5	6	6	3	6
26	5	6	5	4	8	5	4	7	76	4	4	5	6	5	4	5	3
27	3	5	4	6	5	4	6	5	77	7	5	7	6	5	7	3	6
28	4	6	6	5	5	4	7	5	78	8	6	9	6	6	8	1	7
29	6	5	9	5	8	6	5	6	79	8	5	8	4	6	9	2	7
30	5	6	8	5	6	6	6	6	80	4	5	3	6	5	3	6	5
31	4	3	7	7	4	5	5	5	81	4	6	5	6	5	4	6	5
32	3	3	6	9	3	4	7	3	82	5	6	7	6	5	4	6	5
33	5	5	6	7	5	6	6	5	83	6	6	7	6	4	4	3	5
34	5	6	7	5	4	4	8	8	84	8	6	6	7	6	8	2	7
35	2	5	5	5	5	3	6	5	85	4	6	6	6	6	5	6	6
36	4	6	7	7	5	5	6	5	86	5	7	7	4	8	6	5	9
37	7	7	9	5	6	8	3	5	87	4	4	5	9	2	2	8	4
38	5	7	5	3	8	5	5	5	88	3	5	6	6	4	3	7	4
39	6	5	7	4	7	6	2	7	89	4	6	6	5	6	4	9	6
40	5	5	4	7	5	3	8	4	90	5	6	7	6	5	4	6	6
41	3	4	6	9	4	3	6	4	91	3	5	6	7	5	4	9	5
42	5	7	7	6	6	4	6	7	92	3	3	6	7	3	3	7	4
43	6	4	7	7	5	5	5	4	93	6	6	8	5	5	7	3	7
44	5	7	5	4	9	6	5	6	94	4	5	5	6	6	4	8	5
45	5	7	9	3	8	7	3	7	95	5	7	6	4	6	6	3	5
46	2	5	4	7	3	4	6	5	96	4	4	6	7	3	5	6	5
47	5	9	6	4	9	4	4	8	97	3	7	6	4	7	3	7	5

(continued)

Table 10.1 (continued)

	A	C	I	B	T	V	H	P		A	C	I	B	T	V	H	P
48	5	5	7	8	4	5	3	5	98	4	7	5	6	7	3	7	4
49	7	6	8	5	7	7	2	6	99	5	4	5	9	4	5	6	7
50	6	5	6	7	4	6	5	4	100	4	5	7	7	3	5	4	3
(b) Centered data																	
1	4.01	1.46	2.53	−3.51	3.28	3.12	−2.29	2.38	51	1.01	0.46	0.53	−2.51	0.28	0.12	0.71	1.38
2	−2.99	−2.54	−1.47	2.49	−4.72	−1.88	1.71	−2.62	52	1.01	0.46	1.53	−1.51	2.28	3.12	−2.29	3.38
3	0.01	0.46	0.53	0.49	2.28	−0.88	0.71	0.38	53	2.01	0.46	1.53	−1.51	2.28	1.12	−1.29	1.38
4	−0.99	0.46	−0.47	−2.51	2.28	0.12	1.71	1.38	54	2.01	0.46	0.53	−0.51	1.28	1.12	−0.29	1.38
5	1.01	−0.54	0.53	0.49	0.28	0.12	0.71	0.38	55	−0.99	−1.54	−0.47	2.49	−3.72	−1.88	1.71	−2.62
6	−0.99	−0.54	−1.47	−0.51	0.28	−1.88	−0.29	−0.62	56	2.01	0.46	1.53	−2.51	1.28	3.12	−3.29	1.38
7	1.01	1.46	−0.47	−0.51	2.28	−1.88	0.71	2.38	57	2.01	−0.54	1.53	0.49	1.28	1.12	−3.29	0.38
8	1.01	0.46	0.53	−1.51	2.28	2.12	0.71	1.38	58	1.01	0.46	−0.47	−2.51	0.28	−0.88	−2.29	−1.62
9	2.01	0.46	1.53	−0.51	0.28	0.12	−2.29	−1.62	59	1.01	−0.54	0.53	0.49	−0.72	0.12	−2.29	1.38
10	−0.99	−1.54	−0.47	2.49	−1.72	−1.88	0.71	−2.62	60	−0.99	−1.54	0.53	0.49	−1.72	0.12	−0.29	−1.62
11	0.01	0.46	−0.47	−1.51	0.28	−0.88	−0.29	1.38	61	−0.99	0.46	−0.47	−0.51	0.28	0.12	3.71	0.38
12	1.01	−1.54	−0.47	−0.51	−0.72	1.12	−1.29	0.38	62	1.01	−1.54	−1.47	1.49	−1.72	−1.88	−0.29	−1.62
13	2.01	−0.54	1.53	−0.51	−0.72	1.12	1.71	0.38	63	−0.99	−0.54	−0.47	0.49	0.28	−0.88	1.71	−0.62
14	−0.99	−0.54	−0.47	1.49	−0.72	−0.88	1.71	−0.62	64	−2.99	0.46	−1.47	−0.51	−0.72	−2.88	3.71	−0.62
15	−1.99	0.46	−1.47	0.49	−1.72	−2.88	0.71	−0.62	65	0.01	0.46	−0.47	0.49	1.28	1.12	−0.29	0.38
16	0.01	0.46	0.53	−2.51	3.28	0.12	1.71	1.38	66	−0.99	−0.54	−2.47	1.49	−0.72	−1.88	1.71	−0.62
17	−0.99	−0.54	1.53	−1.51	2.28	0.12	−0.29	0.38	67	−0.99	0.46	−0.47	−1.51	1.28	−0.88	0.71	0.38
18	2.01	0.46	1.53	−1.51	0.28	1.12	−1.29	0.38	68	0.01	−1.54	0.53	0.49	0.28	3.12	−1.29	−1.62
19	0.01	1.46	0.53	−1.51	3.28	0.12	−1.29	1.38	69	1.01	0.46	0.53	−1.51	1.28	−0.88	−0.29	2.38
20	0.01	−0.54	0.53	−0.51	−1.72	−0.88	0.71	2.38	70	−0.99	0.46	−1.47	1.49	−0.72	−0.88	0.71	0.38

(continued)

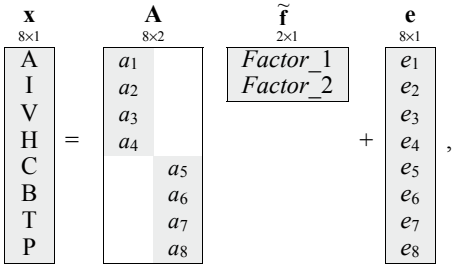
Table 10.1 (continued)

	A	C	I	B	T	V	H	P		A	C	I	B	T	V	H	P
21	1.01	1.46	0.53	−1.51	2.28	1.12	0.71	0.38	71	1.01	−0.54	0.53	−0.51	−0.72	1.12	−1.29	0.38
22	−0.99	0.46	−0.47	−1.51	−0.72	−1.88	1.71	−0.62	72	1.01	1.46	1.53	−0.51	1.28	2.12	−1.29	1.38
23	−1.99	0.46	−1.47	1.49	−1.72	−1.88	2.71	−3.62	73	1.01	−0.54	1.53	−0.51	0.28	1.12	−3.29	−0.62
24	2.01	1.46	1.53	−0.51	1.28	3.12	−0.29	0.38	74	0.01	0.46	−0.47	0.49	0.28	−2.88	2.71	−0.62
25	−0.99	0.46	0.53	0.49	−0.72	−0.88	−0.29	0.38	75	1.01	0.46	1.53	−0.51	0.28	1.12	−2.29	0.38
26	0.01	0.46	−1.47	−1.51	2.28	0.12	−1.29	1.38	76	−0.99	−1.54	−1.47	0.49	−0.72	−0.88	−0.29	−2.62
27	−1.99	−0.54	−2.47	0.49	−0.72	−0.88	0.71	−0.62	77	2.01	−0.54	0.53	0.49	−0.72	2.12	−2.29	0.38
28	−0.99	0.46	−0.47	−0.51	−0.72	−0.88	1.71	−0.62	78	3.01	0.46	2.53	0.49	0.28	3.12	−4.29	1.38
29	1.01	−0.54	2.53	−0.51	2.28	1.12	−0.29	0.38	79	3.01	−0.54	1.53	−1.51	0.28	4.12	−3.29	1.38
30	0.01	0.46	1.53	−0.51	0.28	1.12	0.71	0.38	80	−0.99	−0.54	−3.47	0.49	−0.72	−1.88	0.71	−0.62
31	−0.99	−2.54	0.53	1.49	−1.72	0.12	−0.29	−0.62	81	−0.99	0.46	−1.47	0.49	−0.72	−0.88	0.71	−0.62
32	−1.99	−2.54	−0.47	3.49	−2.72	−0.88	1.71	−2.62	82	0.01	0.46	0.53	0.49	−0.72	−0.88	0.71	−0.62
33	0.01	−0.54	−0.47	1.49	−0.72	1.12	0.71	−0.62	83	1.01	0.46	0.53	0.49	−1.72	−0.88	−2.29	−0.62
34	0.01	0.46	0.53	−0.51	−1.72	−0.88	2.71	2.38	84	3.01	0.46	−0.47	1.49	0.28	3.12	−3.29	1.38
35	−2.99	−0.54	−1.47	−0.51	−0.72	−1.88	0.71	−0.62	85	−0.99	0.46	−0.47	0.49	0.28	0.12	0.71	0.38
36	−0.99	0.46	0.53	1.49	−0.72	0.12	0.71	−0.62	86	0.01	1.46	0.53	−1.51	2.28	1.12	−0.29	3.38
37	2.01	1.46	2.53	−0.51	0.28	3.12	−2.29	−0.62	87	−0.99	−1.54	−1.47	3.49	−3.72	−2.88	2.71	−1.62
38	0.01	1.46	−1.47	−2.51	2.28	0.12	−0.29	−0.62	88	−1.99	−0.54	−0.47	0.49	−1.72	−1.88	1.71	−1.62
39	1.01	−0.54	0.53	−1.51	1.28	1.12	−3.29	1.38	89	−0.99	0.46	−0.47	−0.51	0.28	−0.88	3.71	0.38
40	0.01	−0.54	−2.47	1.49	−0.72	−1.88	2.71	−1.62	90	0.01	0.46	0.53	0.49	−0.72	−0.88	0.71	0.38
41	−1.99	−1.54	−0.47	3.49	−1.72	−1.88	0.71	−1.62	91	−1.99	−0.54	−0.47	1.49	−0.72	−0.88	3.71	−0.62
42	0.01	1.46	0.53	0.49	0.28	−0.88	0.71	1.38	92	−1.99	−2.54	−0.47	1.49	−2.72	−1.88	1.71	−1.62
43	1.01	−1.54	0.53	1.49	−0.72	0.12	−0.29	−1.62	93	1.01	0.46	1.53	−0.51	−0.72	2.12	−2.29	1.38
44	0.01	1.46	−1.47	−1.51	3.28	1.12	−0.29	0.38	94	−0.99	−0.54	−1.47	0.49	0.28	−0.88	2.71	−0.62

(continued)

Table 10.1 (continued)

	A	C	I	B	T	V	H	P		A	C	I	B	T	V	H	P
45	0.01	1.46	2.53	−2.51	2.28	2.12	−2.29	1.38	95	0.01	1.46	−0.47	−1.51	0.28	1.12	−2.29	−0.62
46	−2.99	−0.54	−2.47	1.49	−2.72	−0.88	0.71	−0.62	96	−0.99	−1.54	−0.47	1.49	−2.72	0.12	0.71	−0.62
47	0.01	3.46	−0.47	−1.51	3.28	−0.88	−1.29	2.38	97	−1.99	1.46	−0.47	−1.51	1.28	−1.88	1.71	−0.62
48	0.01	−0.54	0.53	2.49	−1.72	0.12	−2.29	−0.62	98	−0.99	1.46	−1.47	0.49	1.28	−1.88	1.71	−1.62
49	2.01	0.46	1.53	−0.51	1.28	2.12	−3.29	0.38	99	0.01	−1.54	−1.47	3.49	−1.72	0.12	0.71	1.38
50	1.01	−0.54	−0.47	1.49	−1.72	1.12	−0.29	−1.62	100	−0.99	−0.54	0.53	1.49	−2.72	0.12	−1.29	−2.62



(10.2)

with the blank cells in **A** occupied by zeros.

In any CFA model, a $p \times 1$ random variable vector **x**, whose expected vector $E[\mathbf{x}] = \mathbf{0}_p$, is expressed as

$$\mathbf{x} = \mathbf{A}\mathbf{f} + \mathbf{e}. \tag{10.3}$$

Here, **A** is the p variables $\times$ m-factors matrix whose elements are called *factor loadings* (or *path coefficients*), **f** is an $m \times 1$ vector whose elements are called *common factor scores* or simply *common factors*, **e** contains errors, and $E[\mathbf{x}] = \mathbf{0}_p$ corresponds to the above supposition that a data set to be analyzed contains *centered scores*.

10.3 Distributional Assumptions for Common Factors

The common factor vector **f** is assumed to follow the multivariate normal (MVN) distribution whose average vector is $\mathbf{0}_m$ and whose covariance matrix is $\mathbf{\Phi} = \mathbf{\Phi}'$, respectively:

$$\mathbf{f} \sim N_m(\mathbf{0}_m, \mathbf{\Phi}). \tag{10.4}$$

Here, the covariance matrix $\mathbf{\Phi}$ ($m \times m$) is constrained to be a *correlation matrix* with

$$\mathbf{\Phi} = \begin{bmatrix} 1 & \phi_{12} & \cdots & \phi_{1m} \\ \phi_{12} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \phi_{m-1,m} \\ \phi_{1m} & \cdots & \phi_{m-1,m} & 1 \end{bmatrix}. \tag{10.5}$$

It is equivalent to the assumption that common factor scores are standard scores with their variances ones.

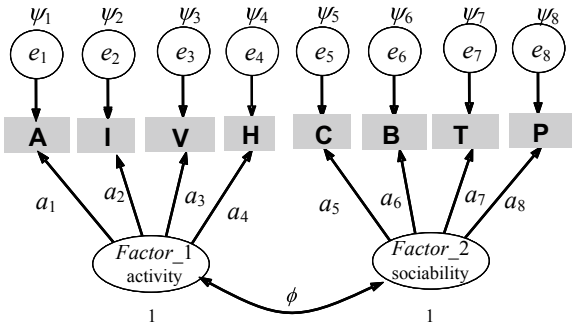


Fig. 10.1 Example of CFA models for the personality data

Let us consider the rationale of the above assumptions for averages and covariances. The average vector $\mathbf{0}_m$ is matched by supposing that a data set to be analyzed being centered scores. The reason for assuming the factor scores to be standard ones is that factors are unobserved latent variables, thus, their variances can be freely determined; we may consider the values of a factor to be distributed over the range [−100, 90], [−50, 60], or [−0.01, 0.01]. For this reason, the variance is usually set to one, as it is a comprehensible value. This implies that the common factor scores are standardized and the covariance matrix between factors is their correlation matrix. Thus, $\mathbf{\Phi}$ in (10.5) is called a *factor correlation* matrix.

10.4 Distributional Assumptions for Errors

The error vector $\mathbf{e}$ is assumed to follow the MVN distribution whose average vector is $\mathbf{0}_p$ and whose covariance matrix is $\mathbf{\Psi}$, respectively:

$$\mathbf{e} \sim N_p\left(\mathbf{0}_p, \mathbf{\Psi}\right), \tag{10.6}$$

with $\mathbf{\Psi}$ the $p \times p$ diagonal matrix, i.e.,

$$\mathbf{\Psi} = \begin{bmatrix} \psi_1 & & \\ & \ddots & \\ & & \psi_p \end{bmatrix}. \tag{10.7}$$

Assumption (10.7) implies that the errors for different variables are mutually uncorrelated, as found in Fig. 10.1, where each of the errors is only linked to the corresponding variable. This is an important feature of factor analysis. In contrast to the common factors in vector $\mathbf{f}$ being the common causes for multiple variables, each error in $\mathbf{e}$ can be viewed as the factor that exclusively or uniquely contributes

to the corresponding variable. For addressing this contrast, the errors in $\mathbf{e}$ are called *unique factors*. Further, the diagonal elements of $\boldsymbol{\Psi}$ are called *unique variances*, as they are the variances of the unique factors.

10.5 Maximum Likelihood Method

We start with a property of the MVN distribution without its proof:

Note 10.1. A Property of MVN Distribution
If $\mathbf{u}_1 \sim N_r(\boldsymbol{\mu}_1, \boldsymbol{\Omega}_1)$, $\mathbf{u}_2 \sim N_r(\boldsymbol{\mu}_2, \boldsymbol{\Omega}_2)$, and $\mathbf{u}_1$ is distributed independently of $\mathbf{u}_2$, then

$$\mathbf{B}_1\mathbf{u}_1 + \mathbf{B}_2\mathbf{u}_2 \sim N_r\left(\mathbf{B}_1\boldsymbol{\mu}_1 + \mathbf{B}_2\boldsymbol{\mu}_2, \mathbf{B}_1\Omega_1\mathbf{B}_1' + \mathbf{B}_2\Omega_2\mathbf{B}_2'\right) \tag{10.8}$$

for fixed matrices $\mathbf{B}_1$ and $\mathbf{B}_2$.

The common and unique factor vectors, $\mathbf{f}$ and $\mathbf{e}$, are assumed to be distributed mutually independently. Using this assumption and (10.8) in (10.3), (10.4) and (10.6), the observation vector $\mathbf{x}$ in (10.3) is found to follow an MVN distribution, as follows:

$$\mathbf{x} \sim N_p\left(\mathbf{0}_p, \boldsymbol{\Sigma}\right), \tag{10.9}$$

with its covariance matrix

$$\boldsymbol{\Sigma} = \mathbf{A}\boldsymbol{\Phi}\mathbf{A}' + \boldsymbol{\Psi}, \tag{10.10}$$

which is called a *covariance structure*, as described in Note 9.2.

Let $\mathbf{X}$ denote the centered data matrix and $\mathbf{V} = n^{-1}\mathbf{X}'\mathbf{X}$ be the sample covariance matrix. As explained in Sect. 9.4, the log likelihood for CFA can be written in the form of (9.15), i.e., $l^*(\boldsymbol{\Sigma}) = (n/2)\log\left|\boldsymbol{\Sigma}^{-1}\mathbf{V}\right| - (n/2)\mathrm{tr}\boldsymbol{\Sigma}^{-1}\mathbf{V}$. Substituting (10.10) into $l^*(\boldsymbol{\Sigma})$, we have

$$l^*(\mathbf{A}, \boldsymbol{\Psi}, \Phi) = \frac{n}{2}\log\left|(\mathbf{A}\boldsymbol{\Phi}\mathbf{A}' + \boldsymbol{\Psi})^{-1}\mathbf{V}\right| - \frac{n}{2}\mathrm{tr}(\mathbf{A}\boldsymbol{\Phi}\mathbf{A}' + \boldsymbol{\Psi})^{-1}\mathbf{V}. \tag{10.11}$$

This is maximized over $\mathbf{A}$, $\boldsymbol{\Psi}$, and $\boldsymbol{\Phi}$, i.e., the 17 parameters $a_1, \ldots, a_8, v_{11}, \ldots, v_{88}$, ϕ for the model in Fig. 10.1.

Since the solution is not explicitly given, the maximization is attained by iterative algorithms. An approach is the one using a *gradient algorithm*, which is illustrated in Appendix A.6.3. Setting the vector $\boldsymbol{\theta}$ in A.6.3 to $[a_1, \ldots, a_8, v_{11}, \ldots, v_{88}, \phi]'$, the solution can be obtained. We express the resulting $\mathbf{A}$, $\boldsymbol{\Psi}$, and $\boldsymbol{\Phi}$ as $\hat{\mathbf{A}}$, $\hat{\boldsymbol{\Psi}}$, and $\hat{\boldsymbol{\Phi}}$, respectively. Another approach is the one using an *EM algorithm* (Dempster et al., 1977). The EM algorithm specialized for CFA (Rubin & Thayer, 1982; Adachi, 2013) is detailed in Appendix 9.

10.6 Solutions

The solution given by the maximum likelihood method is shown in Fig. 10.2a, where the estimated parameter values are presented at the corresponding parts. As in path analysis, the *GFI statistic* defined as (9.18) can be used for assessing whether a solution is satisfactory or not. A value of 0.9 is used as a benchmark, with GFI $\geq$ 0.9 indicating that a model is satisfactory. The GFI value for the solution in Fig. 10.2a was 0.953, which shows that the solution is to be accepted.

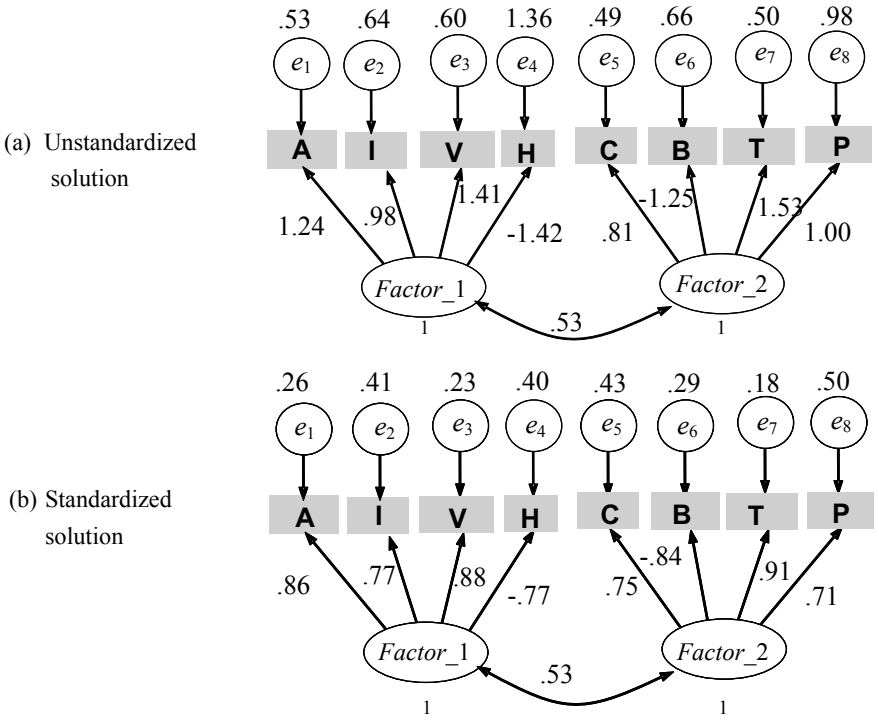


Fig. 10.2 Solution of the model in Fig. 10.1 for the data in Table 10.1

The result in Fig. 10.2a is the *unstandardized solution* obtained from variables with different variances. Thus, it is senseless to compare the largeness of the resulting parameter values. For the comparison to make sense, we must note the *standardized solution* obtained for the standard scores transformed from the original data set. The unstandardized and standardized solutions in CFA can be considered as two *different expressions of the same solution*, since CFA is *scale invariant* with the attained (10.11) and GFI values are the same between both solutions, as is path analysis. This property is shown in Appendix 7. There, the fact is also proved that $\tilde{\mathbf{A}}$, $\tilde{\mathbf{\Phi}}$, and $\tilde{\mathbf{\Psi}}$, which denote the standardized solutions of the loading, factor correlation, and unique variance matrices, respectively, are transformed from the unstandardized solutions $\hat{\mathbf{A}}$, $\hat{\mathbf{\Psi}}$, and $\hat{\mathbf{\Phi}}$, with

$$\tilde{\mathbf{A}} = \mathbf{D}^{-1}\hat{\mathbf{A}}, \tilde{\mathbf{\Phi}} = \hat{\mathbf{\Phi}}, \text{and } \tilde{\mathbf{\Psi}} = \mathbf{D}^{-1}\hat{\mathbf{\Psi}}\mathbf{D}^{-1}. \tag{10.12}$$

The standardized version of the solution in Fig. 10.2a is shown b.

10.7 Other and Extreme Models

Let us refer to the model in Fig. 10.1 as "Two-factor Model 1". Though this model is regarded as satisfactory, with a GFI exceeding the benchmark value of 0.9, a model may exist that is better fitted to the data set in Table 10.1b. This suggests that *other models* should be considered and *compared*, that is, the *model selection* illustrated in Fig. 8.5 should be performed. Figure 10.3 shows two examples of other models. Figure 10.3a presents the *one-factor model* in which only one factor underlies the eight observed variables. For this model, the $\mathbf{A}$ and $\mathbf{f}$ in (10.3) are a vector and a scalar, respectively. Figure 10.3b shows the "*Two-factor Model 2*" in which the variables "Initiative" and "Cheerful" load both factors. This model is written as

$$\underset{8\times1}{\mathbf{x}} = \underset{8\times2}{\mathbf{\Lambda}} \underset{2\times1}{\mathbf{f}} + \underset{8\times1}{\mathbf{e}}. \tag{10.13}$$

$$\begin{bmatrix} A \\ I \\ V \\ H \\ C \\ B \\ T \\ P \end{bmatrix} = \begin{bmatrix} a_1 & \\ a_2 & a_9 \\ a_3 & \\ a_4 & \\ a_{10} & a_5 \\ & a_6 \\ & a_7 \\ & a_8 \end{bmatrix} \begin{bmatrix} Factor\_1 \\ Factor\_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \\ e_8 \end{bmatrix}.$$

As in path analysis, the two types of extreme models are the *independent* and *saturated models*. In the former, all variables are mutually *independent*, without any factor. This is represented as the path diagram in which only eight variables are

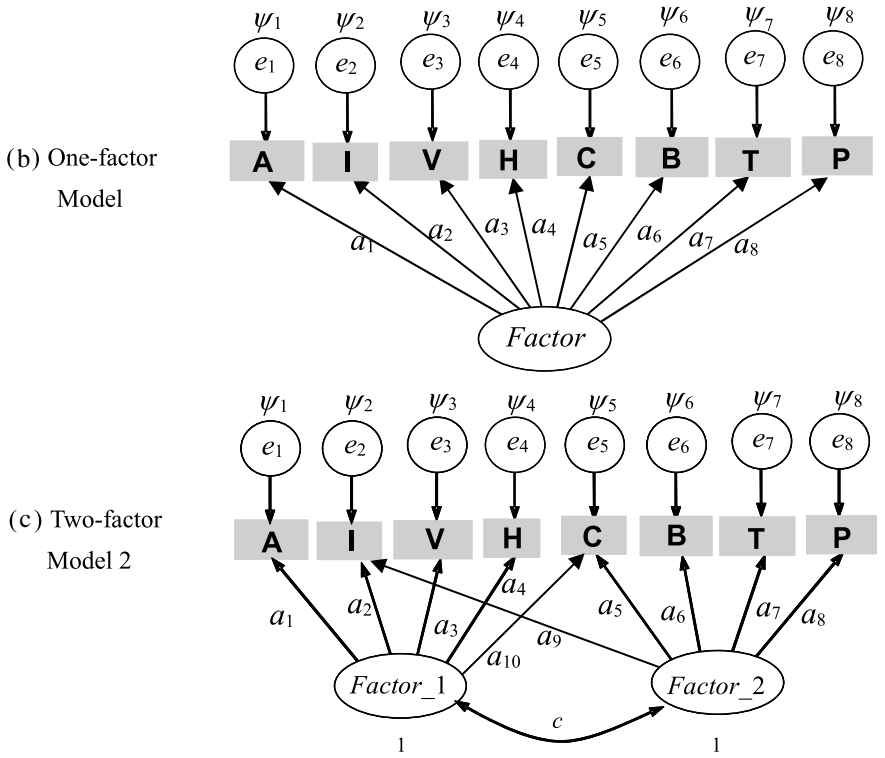


Fig. 10.3 Example of other CFA models

depicted, without any link among them. On the other hand, one of the *saturated models* is represented in the path diagram in which each of the eight variables are linked to the other seven by double-headed arrows, without any factors. This implies that *all variables are merely correlated*.

10.8 Model Selection

So far, we have the two-factor models (1 and 2), the one-factor model, and two extreme ones. For comparing those models with respect to the *goodness of fit* to the data set, we *cannot use the GFI* (9.18) for the reason explained in Sect. 9.9. On the other hand, the *information criteria* introduced in Sect. 8.7 are useful for *comparing models* since the number of parameters is considered in the criteria. The typical information criteria *AIC* and *BIC* for CFA are obtained by substituting the maximum (10.11) value $l^*(\hat{\mathbf{A}}, \hat{\mathbf{\Psi}}, \hat{\mathbf{\Phi}})$ into $l(\hat{\mathbf{\Theta}})$ in (8.24) and (8.25). The AIC and BIC values for each model are shown in Table 10.2. There, the BIC shows that Two-factor Model 1 is the best, while it is found to be slightly worse than

Table 10.2 Number of parameters (NP) and the resulting index values for each model

Model	NP	GFI	AIC	BIC
Saturated	36	1.000	72.000	165.786
Two-factor model 2	19	0.964	*54.168*	103.666
Two-factor model 1	17	0.953	55.464	*99.751*
One-factor mode l	16	0.642	193.494	235.176
Independent	8	0.354	504.322	525.163

Italic font refers to the least AIC and BIC values

Two-factor Model 2 in the AIC values. This demonstrates that model selection statistics indicate different models as the best. For such a case, the model must be chosen by users' subjective consideration. This shows that no absolute index exists for model selection, which should be kept in mind.

10.9 Bibliographical Notes

It is difficult to find books in which CFA is exclusively treated. CFA is, however, described in chapters of books on structural equation modeling or factor analysis, which include Kaplan (2000), Mulaik (2010), and Wang and Wang (2012).

One drawback of CFA is that the model, i.e., the elements that are set to be zero in $\mathbf{A}$, must be selected by users in a subjective manner. Such a drawback can be dealt with by the *sparse factor analysis* treated in Chap. 22.

Exercises

10.1. Let us consider the model $x = t + e$, with x an observed variable, e an error, and t a true score which is an unobserved latent variable. For example, t stands for the ability of mathematics possessed by an examinee, while x is the test score on mathematics shown by the examinee, and an error e must be considered, since t (ability) cannot be perfectly exactly measured by x (score). Present another example for a set of x, t, and e in the model.

10.2. Spearman (1904) hit upon the idea of factor analysis, by considering the scores of achievement tests as variables, and personality test scores have been used as an example in this chapter. Present an example of a data set that is not related to such tests and for which factor analysis is useful.

10.3. Consider another two-factor model for the data in Table 10.1.

10.4. Depict the path diagram of a saturated model for the data in Table 10.1 without a factor and a single-headed path.

10.5. Present an example of the CFA model for 15 observed variables with three factors.

10.6. Eq. (10.3) can be rewritten as $\mathbf{x} = \mathbf{Af} + \mathbf{e} = \mathbf{A}^*\mathbf{f}^* + \mathbf{e}$ with $\mathbf{f}^* = \mathbf{S}^{-1}\mathbf{f}$ and $\mathbf{A}^* = \mathbf{AS}$. It suggests that $\mathbf{f}^*$ and $\mathbf{A}^*$ could also be regarded as a factor score vector and a loading matrix, respectively, with $\mathbf{S}$ an $m \times m$ arbitrary non-singular matrix. However, in CFA, except for special cases, it is not possible

to regard $\mathbf{f}^*$ and $\mathbf{A}^*$ as above. Show the reason for this, noting that $\mathbf{A}$ is constrained in CFA.

10.7. Model (10.3) can be rewritten as $\mathbf{x} = \mathbf{Hg}$, with $\mathbf{H} = [\mathbf{A}, \mathbf{I}_p]$ being $p \times (m + p)$ and $\mathbf{g} = \begin{bmatrix} \mathbf{f} \\ \mathbf{e} \end{bmatrix}$ $(m + p) \times 1$. If $\mathbf{A}$ and $\mathbf{x}$ are given, $\mathbf{x} = \mathbf{Hg}$ is regarded as a system of equations with $\mathbf{g}$ unknown. The necessary and sufficient condition of the system having the solutions of $\mathbf{g}$ is known to be $\mathbf{HH}^+\mathbf{x} = \mathbf{x}$. If this equation holds true, show that the solution of $\mathbf{g}$ satisfying $\mathbf{x} = \mathbf{Hg}$ is expressed as

$$\mathbf{g} = \mathbf{H}^+\mathbf{x} + \left(\mathbf{I}_{m+p} - \mathbf{H}^+\mathbf{H}\right)\mathbf{q}, \tag{10.14}$$

with $\mathbf{H}^+$ the Moore–Penrose inverse of $\mathbf{H}$ defined in Exercise 5.10 and $\mathbf{q}$ an arbitrary $(m + p) \times 1$ vector. This inverse and the solution of a system of equations are also detailed in Chap. 17.

10.8. Show the following: (10.14) implies that factor score vector $\mathbf{f}$ cannot be uniquely determined, i.e., we cannot select a single vector as $\mathbf{f}$ for given $\mathbf{A}$ and $\mathbf{x}$.

10.9. Let us consider the CFA model with intercept vector $\mathbf{c}$: $\mathbf{x} = \mathbf{Af} + \mathbf{c} + \mathbf{e}$, $\mathbf{f} \sim N_m(\mathbf{0}_m, \mathbf{\Phi})$, and $\mathbf{e} \sim N_p(\mathbf{0}_p, \mathbf{\Psi})$. Show that the MLE of transposed intercept vector $\mathbf{c}'$ is given by $n^{-1}\mathbf{1}_n'\mathbf{X}$ for the $n \times p$ data matrix $\mathbf{X}$ whose rows are the observations of $\mathbf{x}'$ for individuals $i = 1, \ldots, n$.

10.10. Let us consider a *confirmatory principal component analysis* (PCA) procedure formulated as minimizing $\|\mathbf{X} - \mathbf{FA}'\|^2$ over $\mathbf{F}$ and $\mathbf{A}$ subject to $n^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_m$ and some elements of $\mathbf{A}$ constrained to be zero. Show that the function can be decomposed as

$$\|\mathbf{X} - \mathbf{FA}'\|^2 = \|\mathbf{X} - \mathbf{FB}'\|^2 + n\|\mathbf{B} - \mathbf{A}\|^2, \tag{10.15}$$

with $\mathbf{B} = n^{-1}\mathbf{X}'\mathbf{F}$ (Adachi & Trendafilov, 2016).

10.11. Show that an algorithm for the confirmatory PCA in Exercise 10.10 can be formed by the following steps:

Step 1. Initialize $\mathbf{F}$.
Step 2. Set the unconstrained elements of $\mathbf{A}$ to the corresponding ones of $n^{-1}\mathbf{X}'\mathbf{F}$.
Step 3. Obtain the SVD $\mathbf{XA} = \mathbf{K\Lambda L}'$ and set $\mathbf{F} = n^{1/2}\mathbf{KL}'$.
Step 4. Finish if convergence is reached; otherwise, go back to Step 2.

The hints for Steps 2 and 3 can be found in (10.15) and Theorem A.4.2 (Appendix A.4.2), respectively.

Chapter 11
Structural Equation Modeling

In confirmatory factor analysis (CFA), introduced in the previous chapter, all factors (latent variables) were causes (explanatory variables). An extended *variant of CFA* is *structural equation modeling* (*SEM*), in which the *causal relationships among factors* are considered, i.e., *factors* appear that are *dependent variables*.

To the best of the author's knowledge, SEM was first presented by the Swedish statistician Jöreskog (1970), who *combined path analysis* and *factor analysis* to formulate SEM. This has been elaborated on and popularized, particularly with the developments of computer software, by the efforts of psychometricians including Bentler (1985).

11.1 Causality Among Factors

We will introduce structural equation modeling (SEM) by starting with the formation of a model, which is followed by the description of the data to be observed.

Let us consider a *model of the causality among four factors*, depicted in Fig. 11.1, with the factors as follows:

[F1] Prior achievements before PostGraduate School (PGS);
[F2] Adaptation to PGS;
[F3] Achievements in PGS;
[F4] Satisfaction with having gone to PGS.

The path diagram in the figure is expressed as a set of formulas

$$\begin{aligned} \mathrm{F3} &= b_1\mathrm{F1} + b_2\mathrm{F2} + e_{\mathrm{F3}}, \\ \mathrm{F4} &= b_3\mathrm{F2} + b_4\mathrm{F3} + e_{\mathrm{F4}}, \end{aligned} \tag{11.1}$$

K. Adachi, *Matrix-Based Introduction to Multivariate Data Analysis*,
https://doi.org/10.1007/978-981-15-4103-2_11

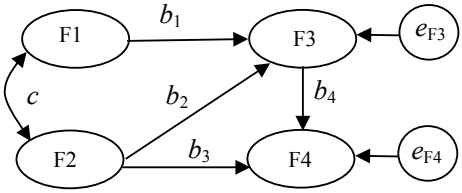


Fig. 11.1 Path diagram for a structural equation model for latent variables

with e_{F3} and e_{F4} being errors. Here, *intercepts* are *omitted*, since it is known that they may be zero, assuming that the averages of the factors are zeros. The set of formulas is a kind of *path analysis* model, though the variables are not observed but rather latent *factors*, which differs from the ordinary path analysis in Chap. 9. A model such as (11.1) is called a *structural equation model* for latent variables.

11.2 Observed Variables as Indicator of Factors

It is reasonable to consider that the above four *factors are difficult to measure directly*, but each of them (F1, F2, F3, F4) *can be measured with several indices* (*observed variables*). Then, let us suppose that each factor can be measured by the four variables shown in Table 11.1. For example, we suppose that X9 (scores for lecture courses), X10 (scores for practice courses), X11 (evaluation of the thesis for

Table 11.1 Variables indicating factors

F	Variable	
F1	X1	Scores for languages when one was a student in a faculty
	X2	Scores for sciences when one was a student in a faculty
	X3	Scores for the entrance examination for a postgraduate school
	X4	Evaluation of a graduation thesis
F2	X5	Goodness of fit to the education in the postgraduate school
	X6	Goodness of fit to the atmosphere in the postgraduate school
	X7	Goodness of fit to the facilities in the postgraduate school
	X8	Inconvenience found in the systems of the postgraduate school
F3	X9	Scores for lecture courses in the postgraduate school
	X10	Scores for practice courses in the postgraduate school
	X11	Evaluation of the thesis for master degree
	X12	Self-rating of achievement
F4	X13	Fulfillment felt from life in the postgraduate school
	X14	How well one enjoyed life in the postgraduate school
	X15	Regret that one went to the postgraduate school
	X16	Hope for the future

master degree), and X12 (self-rating of achievement) can be used as the *indicators* for F3 (achievements in PGS).

The four path diagrams in Fig. 11.2 represent the fact that F1, F2, F3, and F4 are indicated by the variables in Table 11.1. Each diagram can be expressed by a set of formulas; for example, the third diagram is expressed as the set of four equations

$$\begin{aligned} \mathrm{X9} &= a_9\mathrm{F3} + e_9, \\ \mathrm{X10} &= a_{10}\mathrm{F3} + e_{10}, \\ \mathrm{X11} &= a_{11}\mathrm{F}_3 + e_{11}, \\ \mathrm{X12} &= a_{12}\mathrm{F3} + e12. \end{aligned} \tag{11.2}$$

This is just a *factor analysis* model, which is also called a *measurement equation model*, as (11.2) stands for how an *unobserved common factor* (F3), which cannot be measured directly, is *measured* using *several observed variables* as the *indicators of the common factor*.

Let 300 × 16 data matrix $\mathbf{X}$ contain the centered scores of 300 postgraduate students for the 16 items in Table 11.1 with covariance matrix $\mathbf{V} = n^{-1}\mathbf{X}'\mathbf{X}$ for the 16 variables shown in Table 11.2. The data matrix $\mathbf{X}$ is too big to be presented; in place of it, the sample covariance matrix $\mathbf{V}$ is presented here. As described in Note 9.2, the procedures in Chaps. 9–12 can be feasible only with $\mathbf{V}$, even if $\mathbf{X}$ is not available.

11.3 SEM Model

The *structural equation model* in Fig. 11.1 and the four *measurement equation models* in Fig. 11.2 are *integrated into a single model* in Fig. 11.3. This is a *SEM model* for the covariance matrix in Table 11.2. The outer parts of the diagram in Fig. 11.3 are a–d in Fig. 11.2, while the inner part in Fig. 11.3 is the diagram in Fig. 11.1. In other words, the *outer* parts stand for *measurement* equation models (i.e., *factor analysis* models), while the *inner* part represents a *structural* equation model (i.e., a *path analysis* model with latent factors). That is, SEM is an analysis procedure with a model into which *structural* and *measurement* equation models are *integrated*. However, the procedure is called *structural equation modeling*, without the use of the term "measurement".

11.4 Matrix Expression

The path diagram in Fig. 11.3 is formally expressed using the two equations in (11.1) and the four sets of measurement equations, with an example of a set presented in (11.2). Those equations can be written as a single equation in matrix form:

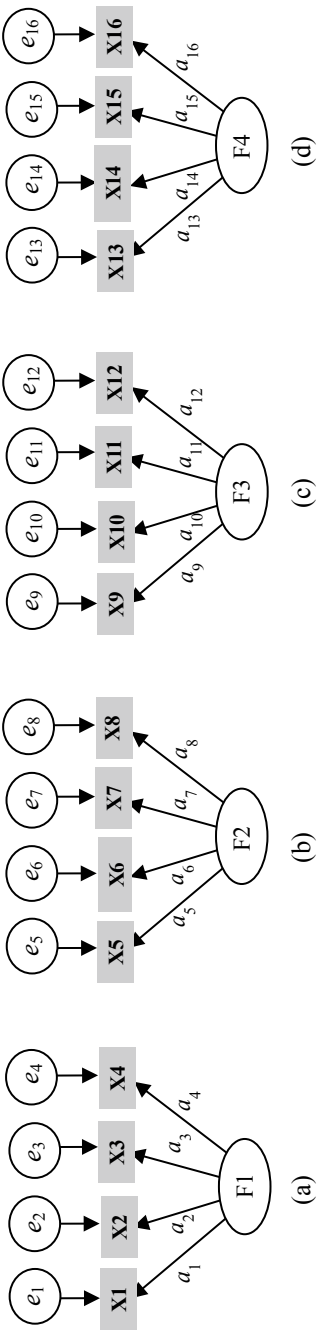


Fig. 11.2 Path diagrams for four measurement equation models

Table 11.2 Data set observed for 300 postgraduate students, which is an artificial example found in Adachi (2006), with the upper triangular covariances omitted

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16
X1	40.323															
X2	27.475	55.696														
X3	21.883	28.439	52.305													
X4	2.669	3.104	2.662	0.620												
X5	1.685	1.320	0.792	0.155	0.810											
X6	1.405	1.469	1.119	0.132	0.484	0.961										
X7	1.478	1.410	0.549	0.164	0.454	0.468	1.128									
X8	–0.820	–0.822	–0.502	–0.109	–0.439	–0.477	–0.458	1.011								
X9	15.131	14.911	12.839	1.557	2.026	2.108	2.177	–1.597	44.213							
X10	19.270	17.290	12.088	2.272	2.437	2.698	2.545	–2.363	28.670	58.632						
X11	2.506	3.065	2.551	0.294	0.359	0.374	0.380	–0.356	4.163	4.513	1.176					
X12	2.757	2.602	1.977	0.268	0.247	0.305	0.273	–0.262	3.809	4.123	0.703	1.077				
X13	2.161	2.816	2.040	0.284	0.371	0.436	0.407	–0.385	3.566	4.567	0.747	0.575	1.377			
X14	1.686	2.091	1.847	0.162	0.293	0.344	0.277	–0.262	2.500	2.349	0.417	0.389	0.573	0.924		
X15	–1.685	–2.117	–1.889	–0.204	–0.311	–0.414	–0.368	0.331	–3.144	–3.595	–0.600	–0.506	–0.727	–0.473	1.141	
X16	1.306	1.726	1.415	0.163	0.305	0.374	0.286	–0.270	3.072	3.083	0.553	0.542	0.672	0.551	–0.592	1.349

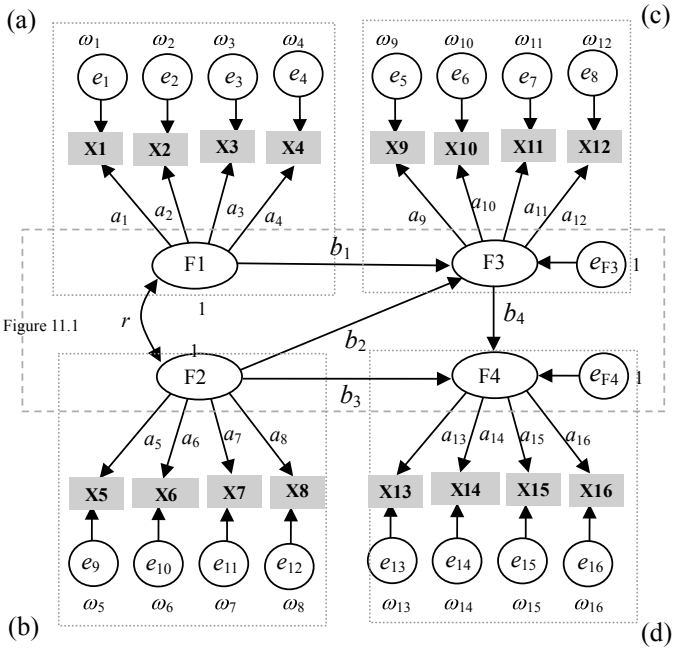


Fig. 11.3 A SEM model

$$
\begin{array}{c|c} \mathbf{t} & \mathbf{B} \\ \hline
\begin{array}{c} \text{F1} \\ \text{F2} \\ \text{F3} \\ \text{F4} \\ \text{X1} \\ \text{X2} \\ \text{X3} \\ \text{X4} \\ \text{X5} \\ \text{X6} \\ \text{X7} \\ \text{X8} \\ \text{X9} \\ \text{X10} \\ \text{X11} \\ \text{X12} \\ \text{X13} \\ \text{X14} \\ \text{X15} \\ \text{X16} \end{array} &
\begin{array}{cccccccccccccccccccc}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
b_1 & b_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & b_3 & b_4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
a_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
a_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
a_3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
a_4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & a_5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & a_6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & a_7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & a_8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & a_9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & a_{10} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & a_{11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & a_{12} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & a_{13} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & a_{14} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & a_{15} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & a_{16} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}
\end{array}
=
\begin{array}{c} \mathbf{t} \\ \hline
\begin{array}{c} \text{F1} \\ \text{F2} \\ \text{F3} \\ \text{F4} \\ \text{X1} \\ \text{X2} \\ \text{X3} \\ \text{X4} \\ \text{X5} \\ \text{X6} \\ \text{X7} \\ \text{X8} \\ \text{X9} \\ \text{X10} \\ \text{X11} \\ \text{X12} \\ \text{X13} \\ \text{X14} \\ \text{X15} \\ \text{X16} \end{array}
\end{array}
+
\begin{array}{c} \mathbf{u} \\ \hline
\begin{array}{c} \text{F1} \\ \text{F2} \\ e_{\text{F3}} \\ e_{\text{F4}} \\ e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \\ e_8 \\ e_9 \\ e_{10} \\ e_{11} \\ e_{12} \\ e_{13} \\ e_{14} \\ e_{15} \\ e_{16} \end{array}
\end{array}
\cdot \qquad (11.3)
$$

Here, $\mathbf{t}$ is the random vector whose first elements are *common factors* and the remaining ones are *observed variables*, while $\mathbf{u}$ is the vector whose first elements are the common factors being explanatory variables and the remaining ones are the errors for the dependent common factors and observed variables. Matrix $\mathbf{B}$ is filled with zeros except for the *path coefficients* corresponding to the links between common factors and the links of common factors to observed variables. The first and second rows in the left- and right-hand sides of (11.3) stand for F1 = F1 and F2 = F2, which obviously hold true; the third and fourth rows express (11.1), and the remaining ones stand for the measurement equation models (Fig. 11.2), with the rows for X9 to X12 corresponding to (11.2).

Any *SEM model* is expressed as

$$\mathbf{t} = \mathbf{B}\mathbf{t} + \mathbf{u}. \tag{11.4}$$

Here, $\mathbf{B}$ is an $(m + p) \times (m + p)$ path coefficient matrix, with m and p being the numbers of common factors and observed variables, respectively. Vector $\mathbf{t}$ is $(m + p) \times 1$ with

$$\mathbf{t} = \begin{bmatrix} \mathbf{f} \\ \mathbf{x} \end{bmatrix}. \tag{11.5}$$

Its first m elements are those of an $m \times 1$ *common factor* vector $\mathbf{f}$ and the $(m + 1)$th, …, $(m + p)$th elements of $\mathbf{t}$ are the 1st, …, pth *observed variables* in $\mathbf{x}$. Vector $\mathbf{u}$ is $(m + p) \times 1$ with

$$\mathbf{u} = \begin{bmatrix} \mathbf{f}_E \\ \mathbf{e}_D \\ \mathbf{e}_X \end{bmatrix}. \tag{11.6}$$

Its first m_E elements are those of the $m_E \times 1$ vector $\mathbf{f}_E$ containing common factors being explanatory variables; the next m_D elements are those of the $m_D \times 1$ vector $\mathbf{e}_D$ consisting of the errors for dependent common factors; and the remaining p ones are the elements of the $p \times 1$ vector $\mathbf{e}_X$ containing the errors for $\mathbf{x}$.

Equation (11.4) can be rewritten as $(\mathbf{I}_{m+p} - \mathbf{B})\mathbf{t} = \mathbf{u}$ with $\mathbf{I}_{m+p}$ the $(m+p) \times (m+p)$ identity matrix. It can be further rewritten as

$$\mathbf{t} = \left(\mathbf{I}_{m+p} - \mathbf{B}\right)^{-1}\mathbf{u}, \quad \text{i.e.,} \begin{bmatrix} \mathbf{f} \\ \mathbf{x} \end{bmatrix} = \left(\mathbf{I}_{m+p} - \mathbf{B}\right)^{-1}\mathbf{u}, \tag{11.7}$$

where we have supposed the existence of $(\mathbf{I}_{m+p} - \mathbf{B})^{-1}$. Now, let us define a $p \times (m + p)$ matrix as

$$\mathbf{H} = \begin{bmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 1 & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}, \tag{11.8}$$

whose first m columns are filled with zeros and whose remaining p columns are the columns of $\mathbf{I}_p$. We find that

$$\mathbf{x} = \mathbf{H}\begin{bmatrix} \mathbf{f} \\ \mathbf{x} \end{bmatrix}, \quad \text{i.e.,} \quad \mathbf{x} = \mathbf{H}\mathbf{t}. \tag{11.9}$$

Using (11.7) in (11.9), it is expressed as

$$\mathbf{x} = \mathbf{H}\left(\mathbf{I}_{m+p} - \mathbf{B}\right)^{-1}\mathbf{u}. \tag{11.10}$$

This is the SEM model for the observation vector $\mathbf{x}$.

11.5 Distributional Assumptions

Let us assume that vector $\mathbf{u}$ is distributed according to the multivariate normal (MVN) distribution, with its mean vector $\mathbf{0}_{m+p}$ and covariance matrix $\mathbf{\Omega}$:

$$\mathbf{u} \sim N_{m+p}(\mathbf{0}_{m+p}, \mathbf{\Omega}). \tag{11.11}$$

The elements of the covariance matrix are described as

Ω =

	F1	F2	e_{F3}	e_{F4}	e_1	...	e_{16}
F1	1	r					
F2	r	1					
e_{F3}			1				
e_{F4}				1			
e_1					ω_1		
...						...	
e_{16}							ω_{16}

(11.12)

for the model in Fig. 11.3, where the blanks (=zeros) indicate *no correlation between errors* and *no correlation* of *errors* to *explanatory variables*. They are not linked by paths, as found in the figure.

In (11.12), we should note the following constraints:

$$\mathrm{V}(\mathrm{F1}) = \mathrm{V}(\mathrm{F2}) = \mathrm{V}(e_{\mathrm{F3}}) = \mathrm{V}(e_{\mathrm{F4}}) = 1, \tag{11.13}$$

with V(F1) denoting the variance of F1. The reason for constraining the variances of factors to be one with V(F1) = V(F2) = 1 is the same in factor analysis (Sect. 10.3); the variances can be set to one, as the common factors are unobserved latent variables and their variances can be freely determined. The errors e_{F3} and e_{F4} for factors F3 and F4, respectively, are also unobserved and their variances can be freely determined. Thus, V(e_{F3}) and V(e_{F4}) can be set to one. The constraint V (F1) = V(F2) = 1 implies that factors F1 and F2 are standardized; thus, their covariance r is a correlation coefficient.

Because of (9.10), (10.8), (11.10), and (11.11), observed variable vector $\mathbf{x}$ is found to follow an MVN distribution as

$$\mathbf{x} \sim N_p\left(\mathbf{0}_p, \Sigma\right), \tag{11.14}$$

with the covariance matrix

$$\Sigma = \mathbf{H}\left(\mathbf{I}_{m+p} - \mathbf{B}\right)^{-1}\mathbf{\Omega}\left(\mathbf{I}_{m+p} - \mathbf{B}\right)^{-1'}\mathbf{H}'. \tag{11.15}$$

11.6 Maximum Likelihood Method

Let $\mathbf{X}$ denote the centered data matrix and $\mathbf{V} = n^{-1}\mathbf{X}'\mathbf{X}$ be the sample covariance matrix. As explained in Sect. 9.4, the log likelihood for SEM can be written in the form of (9.15), i.e., $l^*(\Sigma) = (n/2)\log\left|\Sigma^{-1}\mathbf{V}\right| - (n/2)\mathrm{tr}\Sigma^{-1}\mathbf{V}$. Substituting (11.15) into $l^*(\mathbf{\Sigma})$, we have the *log likelihood* of *parameter* matrices $\mathbf{B}$ and $\mathbf{\Omega}$:

$$\begin{aligned} l^*(\mathbf{B}, \mathbf{\Omega}) = &\frac{n}{2}\log\left|\left\{\mathbf{H}\left(\mathbf{I}_{m+p} - \mathbf{B}\right)^{-1}\mathbf{\Omega}\left(\mathbf{I}_{m+p-\mathbf{B}}\right)^{\prime-1}\mathbf{H}'\right\}^{-1}\mathbf{V}\right| \\ &- \frac{n}{2}\mathrm{tr}\{\mathbf{H}(\mathbf{I}_{m+p} - \mathbf{B})^{-1}\mathbf{\Omega}(\mathbf{I}_{m+p} - \mathbf{B})^{\prime-1}\mathbf{H}'\}^{-1}\mathbf{V}. \end{aligned} \tag{11.16}$$

This is maximized over $\mathbf{B}$ and $\mathbf{\Omega}$, that is, the 37 parameters $a_1, \ldots, a_{16}, b_1, b_2, b_3, b_4, \omega_1, \ldots, \omega_{16}, r$. The maximization of (11.16) is equivalent to minimizing $-l^*(\mathbf{B}, \mathbf{\Omega})$. Since the solution is not explicitly given, the minimization is attained by iterative algorithms. A popular approach is the one using a *gradient algorithm*, illustrated in Appendix A6.3; setting the vector $\boldsymbol{\theta}$ in A.6.3 to [a_1, …, a_{16}, b_1, b_2, b_3, b_4, ω_1, …, ω_{16}, r]′, the solution can be obtained. We express the resulting $\mathbf{B}$ and $\mathbf{\Omega}$ as $\hat{\mathbf{B}}$ and $\hat{\mathbf{\Omega}}$, respectively.

11.7 Solutions

The solution given by the maximum likelihood method is shown in Fig. 11.4; the estimated parameter values are presented at the corresponding parts. As in path analysis and confirmatory factor analysis, the *GFI statistic* defined in (9.18) can be used for assessing whether a solution is satisfactory or not. A value of 0.9 is used as a benchmark with a GFI $\geq$ 0.9 showing that a model is satisfactory. The GFI value for the solution was 0.96, which shows that the solution is to be accepted.

The result in Fig. 11.4 is the *unstandardized solution* obtained from those variables with different variances. Thus, it is senseless to compare the largeness of the resulting parameter values. For the comparison to make sense, we must obtain the *standardized* solution obtained for the standard scores transformed from the original data set. The standardized solution is shown in Fig. 11.5. In this solution, the variances of the dependent common factors are also adjusted to be unity.

As in path analysis and confirmatory factor analysis, SEM also has scale invariance, (though its proof is too complicated to be treated in this book). Thus, the attained value of the maximum of log likelihood (11.15) is equivalent for unstandardized and standardized solutions, and so is the GFI. Further, the standardized

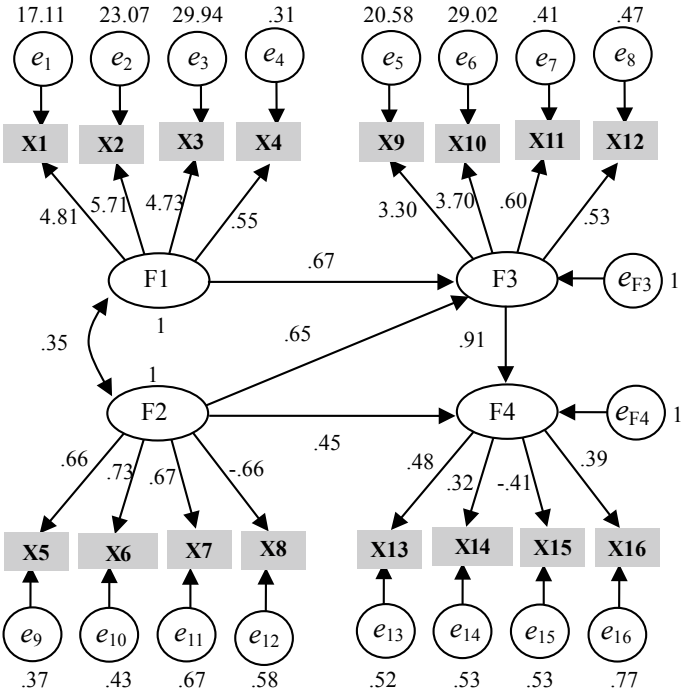


Fig. 11.4 Unstandardized solution

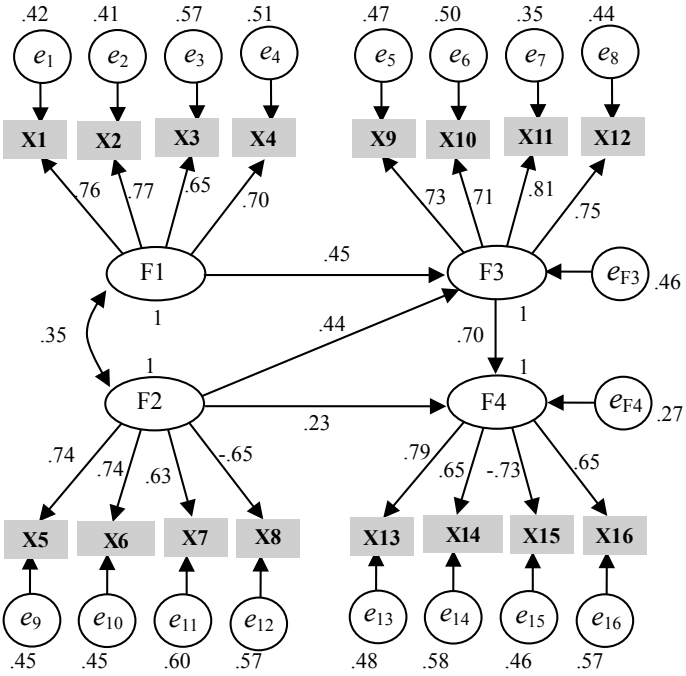


Fig. 11.5 Standardized solution

solution is easily transformed from the unstandardized one. We may thus consider unstandardized and standardized solutions to be two *different expressions of the same solution*.

11.8 Model Selection

As in path analysis and confirmatory factor analysis, several models, including two extreme (independent and saturated) models, should be compared in SEM. For the comparison, information criteria such as the AIC and BIC are useful for selecting a good model, although the GFI cannot be used. The AIC and BIC values for SEM are obtained by substituting the maximum (11.16) value $l^*(\hat{\mathbf{B}}, \hat{\mathbf{\Omega}})$ into $l(\hat{\mathbf{\Theta}})$ in (8.24) and (8.25).

An example of SEM models differing from the model in Fig. 11.3 is shown in Fig. 11.5, where the path connecting F2 and F4 in Fig. 11.3 has been deleted. In Table 11.3, the AIC and BIC for the model in Fig. 11.3 are found to be the least, which shows that model to be the best among the four considered.

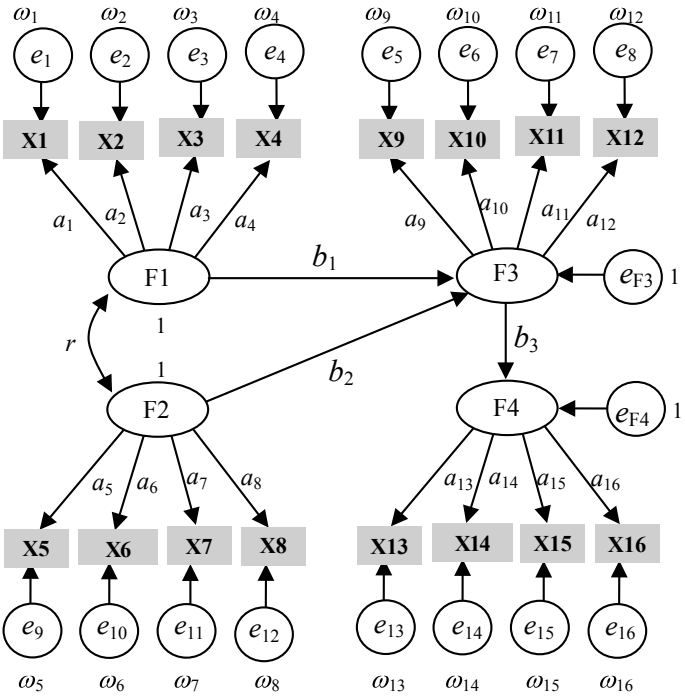


Fig. 11.6 Another SEM model

Table 11.3 Number of parameters (NP) and the resulting index values for each model

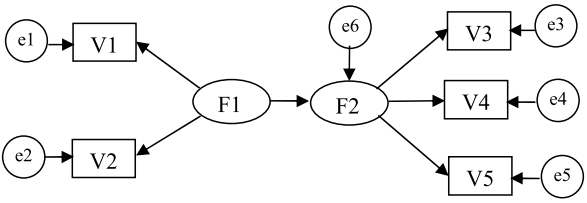
Model	NP	GFI	AIC	BIC
Saturated	136	1.000	272.000	775.714
Figure 11.3	37	0.960	*175.052*	*312.092*
Figure 11.6	36	0.957	183.703	317.039
Independent	16	0.332	2034.899	2036.828

11.9 Bibliographical Notes

The books in which SEM is exhaustively detailed include Bollen (1989), Kaplan (2000), and Wang and Wang (2012). SEM is also illustrated in a chapter of Lattin, Carroll, and Green (2003). The formulation of SEM in this chapter is based on Toyoda (1998), which is a very excellent book, but written in Japanese.

Exercises

11.1. Present an example of a set of the variables (V1–V5) and common factors (F1 and F2) whose relationships are represented as the following path diagram:



11.2. The above diagram can be changed into the one for CFA by changing a few parts. Show those changes.

11.3. Present another SEM model for the covariance matrix in Table 11.2.

11.4. Describe what is implied by removing the double-headed path between F1 and F2 from Fig. 11.3.

11.5. Show that the structural equation model describing the causal relationships among factors can be expressed as

$$\mathbf{f}_D = \mathbf{C}\mathbf{f}_D + \mathbf{Q}\mathbf{f}_E + \mathbf{e}_D. \tag{11.17}$$

Here, $\mathbf{f}_D$ ($m_D \times 1$) contains the common factors as dependent variables, $\mathbf{f}_E$ ($m_E \times 1$) contains the common factors as explanatory variables, and $\mathbf{e}_D$ ($m_D \times 1$) consists of the errors for $\mathbf{f}_D$, as defined in Sect. 11.4, with $\mathbf{C}$ and $\mathbf{Q}$ path coefficient matrices.

11.6. Discuss how the elements of $\mathbf{C}$ and $\mathbf{Q}$ in (11.17) correspond to the diagram in Fig. 11.1.

11.7. Show that the measurement equations describing the relationships of the common factor vectors $\mathbf{f}_D$ and $\mathbf{f}_E$ in (11.17) to observed variables can be expressed as

$$\mathbf{y} = \mathbf{A}_Y\mathbf{f}_D + \mathbf{e}_Y, \tag{11.18}$$

$$\mathbf{z} = \mathbf{A}_Z\mathbf{f}_E + \mathbf{e}_Z. \tag{11.19}$$

Here, $\mathbf{y}$ ($p_D \times 1$) and $\mathbf{z}$ ($p_E \times 1$) are the observed variable vectors corresponding to $\mathbf{f}_D$ and $\mathbf{f}_E$, respectively; $\mathbf{y}$, $\mathbf{z}$, $\mathbf{e}_Y$, and $\mathbf{e}_Z$ form the $p \times 1$ vectors $\mathbf{x} = [\mathbf{z}', \mathbf{y}']'$ and $\mathbf{e}_X = [\mathbf{e}_Z', \mathbf{e}_Y']'$ in (11.5) and (11.6), respectively, with $p = p_D + p_E$; $\mathbf{A}_Y$ and $\mathbf{A}_Z$ are path coefficient matrices.

11.8. Show that model (11.4) is equivalent to a set of (11.17), (11.18), and (11.19).

Chapter 12
Exploratory Factor Analysis (Part 1)

As described in Chap. 10, factor analysis (FA) is classified into *exploratory FA* (*EFA*) and confirmatory FA (CFA), except the sparse FA treated in Chap. 22. EFA refers to the procedures for exploring factors underlying observed variables for cases without prior knowledge of what factors explain the variables. EFA is introduced in this chapter. Two features of *EFA* are that [1] *all common factors* are assumed to be linked to *all variables*, and [2] *multiple solutions* exist for a data set.

The FA model conceived by Spearman (1904), the originator of FA, was restricted to one common factor. In the single-factor case, CFA is not distinguished from EFA, as only that model can be considered in which the common factor is linked to all variables. Spearman's single-factor FA was extended to FA with *multiple common factors* by Thurstone (1935, 1947). Then, he chose the EFA approach with all common factors linked to all variables. That was the origin of EFA.

Part 1 follows EFA in the title of this chapter, as the next part for EFA is introduced in Chap. 18. The formulation of EFA in this chapter is prevalent currently, while the formulation in Chap. 18 is the one established recently.

12.1 Example of Exploratory Factor Analysis Model

We use the same data set as that in Chap. 10, the 100 (participants) × 8 (behavioral features) data matrix in Table 10.1(A). It contains the self-ratings regarding to what extent participants' behaviors are characterized by eight variables (features): A (Aggressive), C (Cheerful), I (Initiative), B (Blunt), T (Talkative), V (Vigor), H (tendency to Hesitate), and P (being Popular).

Let us suppose that two common factors underlie the eight variables, though it is unknown what variables are related to each factor. Thus, those links are considered which connect all variables to all common factors, as illustrated in Fig. 12.1. This is

K. Adachi, *Matrix-Based Introduction to Multivariate Data Analysis*,
https://doi.org/10.1007/978-981-15-4103-2_12

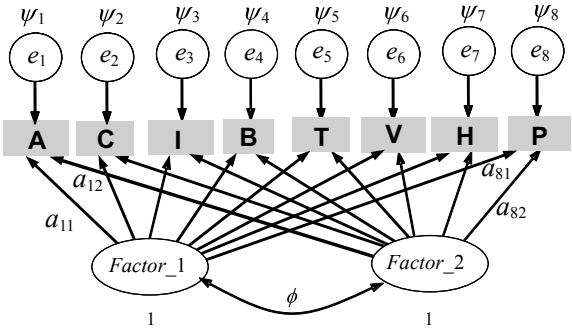


Fig. 12.1 EFA model with two factors for personality data

a key point in EFA. The model in Fig. 12.1 can be written as the set of eight equations:

$$\begin{aligned}
\mathrm{A} &= a_{11} \times Factor_1 + a_{12} \times Factor_2 + c_1 + e_1 \\
\mathrm{C} &= a_{21} \times Factor_1 + a_{22} \times Factor_2 + c_2 + e_2 \\
\mathrm{I} &= a_{31} \times Factor_1 + a_{32} \times Factor_2 + c_3 + e_3 \\
\mathrm{B} &= a_{41} \times Factor_1 + a_{42} \times Factor_2 + c_4 + e_4 \\
\mathrm{T} &= a_{51} \times Factor_1 + a_{52} \times Factor_2 + c_5 + e_5 \\
\mathrm{V} &= a_{61} \times Factor_1 + a_{62} \times Factor_2 + c_6 + e_6 \\
\mathrm{H} &= a_{71} \times Factor_1 + a_{72} \times Factor_2 + c_7 + e_7 \\
\mathrm{P} &= a_{81} \times Factor_1 + a_{82} \times Factor_2 + c_8 + e_8.
\end{aligned} \tag{12.1}$$

Here, c_j and e_j (j = 1,…, 8) express an intercept and an error, respectively; the first subscript j and the second k in a_{jk} indicate a variable and a common factor, respectively. The path coefficients a_{jk} are also called *factor loadings*.

In Fig. 12.1, we can find that each *error* is a *cause* for *a single* variable, in contrast to the *common factors* which are a *common cause* for all variables. For this contrast, an error is also called a *unique factor* (a factor *uniquely* influencing a single variable) with its variance called a *unique variance*, as already mentioned in Chap. 10.

12.2 Matrix Expression

Table 10.1(B) shows the centered scores of the raw data in (A). EFA for (A) and that for (B), on the assumption of the averages of factors being zeros, produces the same solution except for the resulting *intercepts* being *zero* in the *latter* analysis. We thus *omit the intercepts* in EFA models, for the sake of simplicity, by supposing that a data matrix to be analyzed contains *centered scores*. Model (12.1) without intercepts can be expressed in matrix form:

$$\underset{8\times1}{\mathbf{x}} = \underset{8\times2}{\mathbf{A}} \underset{2\times1}{\mathbf{f}} + \underset{8\times1}{\mathbf{e}}$$

$$\begin{bmatrix} \text{A} \\ \text{C} \\ \text{I} \\ \text{B} \\ \text{T} \\ \text{V} \\ \text{H} \\ \text{P} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \\ a_{41} & a_{42} \\ a_{51} & a_{52} \\ a_{61} & a_{62} \\ a_{71} & a_{72} \\ a_{81} & a_{82} \end{bmatrix} \begin{bmatrix} \textit{Factor\_1} \\ \textit{Factor\_2} \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \\ e_8 \end{bmatrix}. \tag{12.2}$$

In any EFA model, a $p \times 1$ random variable vector $\mathbf{x}$, whose expected vector $E[\mathbf{x}]$ is $\mathbf{0}_p$, is expressed as

$$\mathbf{x} = \mathbf{A}\mathbf{f} + \mathbf{e}, \tag{12.3}$$

where $\mathbf{A} = (a_{jk})$ is the p (variables) $\times$ m (common factor) matrix containing *factor loadings*, $\mathbf{f}$ is an $m \times 1$ vector whose elements are called *common factor scores*, and $\mathbf{e}$ contains errors, in other words, *unique factor scores*. This is the same as the CFA model in Chap. 10 except that $\mathbf{A}$ is unconstrained in EFA.

12.3 Distributional Assumptions

The error or unique factor vector $\mathbf{e}$ is assumed to be distributed according to the multivariate normal (MVN) distribution whose average vector and covariance matrix are $\mathbf{0}_p$ and $\mathbf{\Psi}$, respectively:

$$\mathbf{e} \sim N_p\left(\mathbf{0}_p, \mathbf{\Psi}\right), \tag{12.4}$$

with $\mathbf{\Psi}$ the diagonal matrix including *unique variances*, i.e.,

$$\mathbf{\Psi} = \begin{bmatrix} \psi_1 & & \\ & \ddots & \\ & & \psi_p \end{bmatrix}. \tag{12.5}$$

The common factor vector $\mathbf{f}$ is supposed to be distributed according to the MVN distribution whose average vector and covariance matrix are $\mathbf{0}_m$ and $\mathbf{I}_m$, respectively:

$$\mathbf{f} \sim N_m(\mathbf{0}_m, \mathbf{I}_m). \tag{12.6}$$

This differs from assumption (10.4) for CFA in Chap. 10 in that the covariance matrix is the identity matrix. However, this can be transformed into a *factor correlation* matrix $\mathbf{\Phi}$, as in (10.4), for the reason described in Sect. 12.5.

The common factor vector $\mathbf{f}$ and error vector $\mathbf{e}$ are assumed to be distributed mutually independently. Because of (10.8), the independence of $\mathbf{f}$ from $\mathbf{e}$, (12.4), and (12.6) imply that the observed variable vector $\mathbf{x}$ is distributed according to the following MVN distribution:

$$\mathbf{x} \sim N_p(\mathbf{0}_p, \mathbf{\Sigma}), \tag{12.7}$$

with its covariance matrix

$$\mathbf{\Sigma} = \mathbf{A}\mathbf{A}' + \mathbf{\Psi}. \tag{12.8}$$

12.4 Maximum Likelihood Method

Let $\mathbf{X}$ denote the centered data matrix in Table 10.1(B) and $\mathbf{V} = n^{-1}\mathbf{X}'\mathbf{X}$ be the sample covariance matrix. As explained in Sect. 9.4, the *log likelihood* is written in the form of (9.15). By substituting (12.8) in (9.15), we have

$$l^*(\mathbf{A}, \mathbf{\Psi}) = \frac{n}{2}\log\left|(\mathbf{A}\mathbf{A}' + \mathbf{\Psi})^{-1}\mathbf{V}\right| - \frac{n}{2}\mathrm{tr}(\mathbf{A}\mathbf{A}' + \mathbf{\Psi})^{-1}\mathbf{V}. \tag{12.9}$$

This is maximized over $\mathbf{A}$ and $\mathbf{\Psi}$. Since the solution is not explicitly given, the maximization is attained by iterative algorithms. One of the algorithms is a *gradient algorithm*, which is illustrated in Appendix A.6.3. Another approach is the one using an *EM algorithm* (Dempster, Laird, & Rubin, 1977). The EM algorithm specialized for EFA (Rubin & Thayer, 1982, Adachi, 2013) is detailed in Appendix A.9, following Appendix A.8 which serves as a preparation for Appendix A.9. We express the resulting solutions of $\mathbf{A}$ and $\mathbf{\Psi}$ as $\hat{\mathbf{A}}$ and $\hat{\mathbf{\Psi}}$, respectively.

12.5 Indeterminacy of EFA Solutions

The property is called *indeterminacy* that the solution of a procedure is *not unique*, i.e., is not a single; in other words, multiple solutions exist. This property is possessed by EFA. Infinitely many solutions exist in EFA. This is true because the FA model (12.3) can be rewritten as

$$\mathbf{x} = \mathbf{Af} + \mathbf{e} = \mathbf{ATT}'\mathbf{f} + \mathbf{e} = \mathbf{A}_\mathrm{T}\mathbf{f}_\mathrm{T} + \mathbf{e}. \tag{12.10}$$

Here,

$$\mathbf{A}_\mathrm{T} = \mathbf{AT} \text{ and } \mathbf{f}_\mathrm{T} = \mathbf{T}'\mathbf{f}, \tag{12.11}$$

with $\mathbf{T}$ an $m \times m$ matrix satisfying

$$\mathbf{T}'\mathbf{T} = \mathbf{TT}' = \mathbf{I}_m. \tag{12.12}$$

This $\mathbf{T}$ is called an *orthonormal* matrix, which is detailed in Appendix A.1.2. Because of (9.10) and (12.12), (12.6) leads to

$$\mathbf{f}_\mathrm{T} = \mathbf{T}'\mathbf{f} \sim N_m(\mathbf{0}_m, \mathbf{I}_m) : \tag{12.13}$$

$\mathbf{f}_\mathrm{T}$ follows the same distribution as that for $\mathbf{f}$. That is, (12.11) satisfies the assumptions of EFA, which implies that $\hat{\mathbf{A}}_\mathrm{T} = \hat{\mathbf{A}}\mathbf{T}$ is also the solution of $\mathbf{A}$ if $\hat{\mathbf{A}}$ is the solution.

We can also *relax* the condition (12.12) for $\mathbf{T}$ as

$$\mathbf{T}'\mathbf{T} = \begin{bmatrix} 1 & & \# \\ & \ddots & \\ \# & & 1 \end{bmatrix}, \quad \text{or equivalently, } \mathrm{diag}(\mathbf{T}'\mathbf{T}) = \mathrm{I}_m. \tag{12.14}$$

Here, $\begin{bmatrix} 1 & & \# \\ & \ddots & \\ \# & & 1 \end{bmatrix}$ stands for a square matrix whose diagonal elements are restricted to one and diag() is defined next:

Note 12.1. Operator diag(M)
For an $m \times m$ square matrix $\mathbf{M}$, diag($\mathbf{M}$) expresses the $m \times m$ diagonal matrix whose diagonal elements are those of $\mathbf{M}$. For example, if $\mathbf{M} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, then $\mathrm{diag}(\mathbf{M}) = \begin{bmatrix} a & 0 \\ 0 & d \end{bmatrix}$.

When $\mathbf{T}$ is defined as (12.14), the EFA model (12.3) can be rewritten as

$$\mathbf{x} = \mathbf{Af} + \mathbf{e} = \mathbf{AT}'^{-1}\mathbf{T}'\mathbf{f} + \mathbf{e} = \mathbf{A}_\mathrm{T}\mathbf{f}_\mathrm{T} + \mathbf{e}. \tag{12.15}$$

with

$$\mathbf{A}_{\mathrm{T}} = \mathbf{A}\mathbf{T}'^{-1} \text{ and } \mathbf{f}_{\mathrm{T}} = \mathbf{T}'\mathbf{f}. \tag{12.16}$$

Because of (9.10), (12.6) and (12.14) imply

$$\mathbf{f}_{\mathrm{T}} = \mathbf{T}'\mathbf{f} \sim N_m(\mathbf{0}_m, \boldsymbol{\Phi}) \text{ with } \boldsymbol{\Phi} = \mathbf{T}'\mathbf{T}. \tag{12.17}$$

Though this differs from (12.6), (12.17) is a reasonable assumption, if factors are assumed to be correlated, since (12.14) implies that $\boldsymbol{\Phi} = \mathbf{T}'\mathbf{T}$ is a *factor correlation* matrix with its diagonal elements ones. This shows that $\hat{\mathbf{A}}\mathbf{T}'^{-1}$ is also the solution of $\mathbf{A}$ with (12.14) providing the corresponding factor correlation matrix.

As discussed above, the EFA solution of $\mathbf{A}$ is not unique. But, the solution of the diagonal matrix $\boldsymbol{\Psi}$ is *uniquely determined*; the solution of $\boldsymbol{\Psi}$ is single.

12.6 Two-Stage Procedure

As described in the last section, if $\hat{\mathbf{A}}$ is the solution of $\mathbf{A}$, $\hat{\mathbf{A}}\mathbf{T}$ is that with (12.12), and, further, $\hat{\mathbf{A}}\mathbf{T}'^{-1}$ is also a solution with (12.14). Thus, EFA involves the following *two-stage* procedure:

Stage 1. A set of solutions for $\mathbf{A}$ and Ψ, i.e., $\hat{\mathbf{A}}$ and $\hat{\Psi}$, is obtained.
Stage 2. A suitable $\mathbf{T}$ is found to have a solution $\hat{\mathbf{A}}\mathbf{T}$ with (12.12) or $\hat{\mathbf{A}}\mathbf{T}'^{-1}$ (and $\boldsymbol{\Phi} = \mathbf{T}'\mathbf{T}$) with (12.14).

Indeed, the procedure in Sect. 12.4 corresponds to Stage 1. On the other hand, the procedure in Stage 2, which is called *rotation*, is not treated in this chapter, but is detailed in the next chapter. In the next two sections, we illustrate the interpretation of the solution after Stage 2.

12.7 Interpretation of Loadings

The EFA solutions obtained with the above procedures also have *scale invariance*, as explained in Appendix A.7. Thus, the *unstandardized* and *standardized* solutions of EFA can be viewed as *two expressions of the same solution*. In this chapter, only the standardized one is shown. For the data in Table 10.1(B), Stage 1 in the last section, the EFA procedure with $m = 2$ in Sect. 12.4, provides the solution in Table 12.1(A), where the inter-factor correlation is found to be zero, as shown in (12.6). For this solution, Stage 2 in the last section provides the result in Table 12.1 (B), where a rotation technique called "*oblique geomin rotation*", which will be

Table 12.1 Standardized solutions for the data in Table 10.1, with ϕ_{12} the correlation between the first and second factors

	(A) Before rotation			(B) After rotation		
	$\hat{\mathbf{A}}$		ψ_j	$\hat{\mathbf{A}}_{\mathrm{T}}$		ψ_j
A	0.77	−0.38	0.26	**0.82**	0.08	0.26
C	0.61	0.50	0.38	−0.13	**0.84**	0.38
I	0.67	−0.36	0.41	**0.74**	0.04	0.41
B	−0.74	−0.40	0.30	−0.04	**−0.82**	0.30
T	0.79	0.43	0.18	0.04	**0.88**	0.18
V	0.76	−0.44	0.22	**0.87**	0.01	0.22
H	−0.63	0.46	0.39	**−0.82**	0.08	0.39
P	0.70	0.18	0.47	0.23	**0.58**	0.47
ϕ_{12}	0.00			0.48		

detailed in Sect. 13.5, has been used for finding **T**. In this rotation technique, **T** with (12.14) is obtained, and thus an inter-factor factor correlation is also provided.

Let us consider interpreting what each factor in Table 12.1(B) stands for. For facilitating the interpretation, bold font is used for the loadings of large absolute values in Table 12.1(B). Further, the factor loadings can be visually captured in Fig. 12.2, where the widths of paths are proportional to the absolute values of the corresponding loadings and their signs are distinguished by solid and dotted lines. In the figure, we can find that the variables A, I, and V load *Factor*_1 heavily and positively, while H loads that factor negatively; i.e., the higher *Factor*_1 leads to less H (tendency to hesitate). This result allows us to call *Factor*_1 "*the factor of activity*". On the other hand, the variables C, T, and P load *Factor*_2 heavily and positively, while B loads that factor negatively. This allows us to interpret *Factor*_2 as "*the factor of sociability*". Table 12.1(B) shows that the correlation between those two factors is 0.48, which implies the factor of activity is *positively correlated* with the sociability factor.

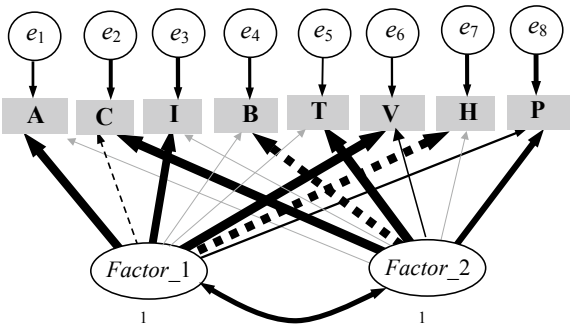


Fig. 12.2 Path diagram in which the widths of paths are proportional to the absolute values of loadings and unique variances, with solid and dotted lines indicating positive and negative values, respectively

12.8 Interpretations of Unique Variances

Unique variances are *uniquely determined*, as they are unrelated to the indeterminacy discussed in Sect. 12.5: their values are equivalent between Table 12.1(A) and (B). The table shows that the unique variance for A (aggressive) is 0.26. This implies that 26% of the variance of variable A remains unexplained by the two common factors; in other words, 74% (=[1 − 0.26] × 100%) of the aggressiveness (A) of individuals are accounted for by the two common factors. That proportion (one minus a unique variance in the standardized solution) is called *communality*. It makes sense to compare the largeness of unique variances among variables in Table 12.1, since the solution is standardized. The largest is that of P (popular) (0.47). It is least explained by the common factors; in other words, P is characterized by a feature unique to that variable beside the two common factors.

12.9 Selecting the Number of Factors

When EFA is used, the suitable number of factors (m) is often unknown for a data set. In order to select m, *information criteria* such as AIC and BIC can be used. Those values for EFA are obtained by substituting the maximum (12.9) value $l^*(\hat{\mathbf{A}}, \hat{\mathbf{\Psi}})$ into $l(\hat{\mathbf{\Theta}})$ in (8.24) and (8.25). For a data set, we can carry out EFA with m set at some candidate numbers, so as to choose the solution with the least AIC or BIC as the solution with the suitable m. In EFA, the number of parameters η in (8.24) and (8.25), which is used for obtaining AIC and BIC, is given by

$$\eta = p + pm - \frac{m(m-1)}{2}. \tag{12.18}$$

Here, $p + pm$ is the number of unique variances and loadings, from which $m(m-1)/2$ must be *subtracted* for the reason described next:

Note 12.2. Loadings Set to Zero
EFA loadings have indeterminacy shown by (12.10) to (12.13). An orthonormal matrix $\mathbf{T}_0$ is known to exist, which can be substituted into $\mathbf{T}$ in (12.12) and leads to

$$\mathbf{A}_0 = \hat{\mathbf{A}}\mathbf{T}_0 \tag{12.19}$$

with

$$\frac{m(m-1)}{2} \quad \text{elements in } \mathbf{A}_0 \text{ being zero.} \tag{12.20}$$

This is illustrated in Table 12.2. The post-multiplication of the left $\hat{\mathbf{A}}$ by $\mathbf{T}_0 = \begin{bmatrix} 0.47 & 0.20 & 0.86 \\ -0.40 & 0.92 & 0.00 \\ -0.79 & -0.35 & 0.51 \end{bmatrix}$ leads to the right $\mathbf{A}_0 = \hat{\mathbf{A}}\mathbf{T}_0$ whose upper left $3(3-1)/2$ elements are zero.

The above fact implies the following: "If $\hat{\mathbf{A}}$ is a solution, then $\mathbf{A}_0$ is also so". This can be rewritten as "If $\mathbf{A}_0$ is a solution, $\mathbf{A}_0\mathbf{T}$ is so, with $\mathbf{T}$ satisfying (12.12)". That is, once the $\{pm - m(m-1)/2\}$ nonzero elements in $\mathbf{A}_0$ are estimated, we can obtain any solution of the loading matrix for a data set. This leads to (12.18).

Table 12.2 Two solutions of loading matrices for the correlation matrix in Table 12.4

$\hat{\mathbf{A}}$: maximum likelihood estimate			$\mathbf{A}_0 = \hat{\mathbf{A}}\mathbf{T}_0'$ with 3 zero elements		
0.60961	0.00029	0.36018	0.00000	0.00000	0.70806
0.38404	−0.03143	0.24305	0.00000	−0.03454	0.45426
0.42288	−0.13882	0.37412	−0.04198	−0.17035	0.55433
0.49556	−0.09901	0.22700	0.09179	−0.06804	0.54208
0.68417	−0.28993	−0.31198	0.68124	−0.01740	0.43022
0.67711	−0.40867	−0.24367	0.67191	−0.15122	0.45884
0.66746	−0.38456	−0.32760	0.72385	−0.10202	0.40785
0.67352	−0.17645	−0.12146	0.48046	0.01826	0.51801
0.68557	−0.45757	−0.25595	0.70521	−0.19000	0.45986
0.48125	0.58511	−0.46732	0.35695	0.79602	0.17686
0.55191	0.30992	−0.12420	0.23023	0.43959	0.41212
0.48094	0.51620	−0.10070	0.09563	0.60577	0.36305
0.60801	0.24713	0.03432	0.15671	0.33861	0.54103
0.40242	0.00397	0.00947	0.17839	0.08261	0.35128
0.37857	0.03413	0.08833	0.09299	0.07800	0.37088
0.50229	0.03928	0.32167	−0.03533	0.02711	0.59609
0.44799	0.13903	0.00273	0.15057	0.21790	0.38714
0.51477	0.26950	0.16531	0.00104	0.29462	0.52739
0.44212	0.06012	0.11279	0.09286	0.10630	0.43804
0.61658	−0.13514	0.11094	0.25417	−0.03614	0.58722
0.60193	0.23149	0.06913	0.13274	0.31099	0.55349
0.61152	−0.12704	0.11541	0.24503	−0.03131	0.58515
0.69611	−0.05113	0.12739	0.24444	0.05132	0.66410
0.65299	0.18337	−0.20214	0.38966	0.37142	0.45944

Table 12.3 AIC and BIC values as functions of m for the data in Table 12.4 with the number of parameters

m	1	2	3	4	5	6	7
η	48	71	93	114	134	153	171
AIC	4247.5	4077.3	3987.9	3956.1	3968.3	3962.9	3972.4
BIC	4390.4	4288.6	4264.7	4295.4	4367.2	4418.3	4481.5

The AIC and BIC values for some m are shown in Table 12.3, which were obtained by the EFA solutions for the correlation matrix in Table 12.4. This is a famous data set known as the 24 psychological test data (Holzinger and Swineford, 1939). The least AIC and BIC are found for $m = 4$ and $m = 3$, which suggest that the best number of factors is 3 or 4. The loading matrices with $m = 3$ in Table 12.2 have been obtained by EFA for the correlations in Table 12.4.

12.10 Difference to Principal Component Analysis

Let the ith rows of $\mathbf{X}$ ($n \times p$), $\mathbf{F}$ ($n \times m$), and $\mathbf{E}$ ($n \times p$) be $\mathbf{x}'$, $\mathbf{f}'$, and $\mathbf{e}'$ observed for individual i, respectively. Then, the EFA model (12.3) can be rewritten in matrix form as $\mathbf{X} = \mathbf{F}\mathbf{A}' + \mathbf{E}$. This takes the same form as the model (5.1) for *principal component analysis* (*PCA*). This begs the question "*In what points does EFA differ from PCA*?" One might answer that EFA is a maximum likelihood (ML) procedure, while PCA is a least squares (LS) one. But, this is incorrect, since EFA can be formulated as LS procedures through the approach in Chap. 18 and the other ones (Harman, 1976; Mulaik, 2011), while PCA can be formulated as an ML procedure (Bishop, 2006; Tipping & Bishop, 1999). Clear answers for the question are found in Chap. 19. In this section, we describe only answers that can be given within the scopes of this and fifth chapters.

A crucial difference between EFA and PCA is found in the *errors*. *No assumption* is made for $\mathbf{E}$ in PCA. Thus, it can be formulated simply as minimizing (5.4), i.e., $\|\mathbf{E}\|^2 = \|\mathbf{X} - \mathbf{F}\mathbf{A}'\|^2$. In contrast, the covariance matrix for errors in EFA is constrained to be a *diagonal matrix* $\mathbf{\Psi}$, as in (12.4). That is, the error for a variable is assumed to be *uncorrelated* with those for the other variables. Thus, errors are called *unique factors*, and its variances (i.e., the diagonal elements of $\mathbf{\Psi}$) are called *unique variances* in *EFA*. On the other hand, the error for a variable in PCA are not unique to that variable; the *correlations* are found among variables, i.e., among the columns of the resulting $\mathbf{E} = \mathbf{X} - \mathbf{F}\mathbf{A}'$.

Table 12.5 shows the EFA and PCA solutions for the correlation matrix in Table 12.4 with $m = 3$. Here, the PCA solution for loadings has been given by (5.28), which can be obtained if only a covariance or correlation matrix is available, as found in Note 6.1. The varimax rotation has been performed for the EFA and

Table 12.4 The correlation matrix for 24 psychological test data with its upper triangular elements omitted

	1	2	3	4	5	6	7	8	9	10	11	12
1	1											
2	0.318	1										
3	0.403	0.317	1									
4	0.468	0.23	0.305	1								
5	0.321	0.285	0.247	0.227	1							
6	0.335	0.234	0.268	0.327	0.622	1						
7	0.304	0.157	0.223	0.335	0.656	0.722	1					
8	0.332	0.157	0.382	0.391	0.578	0.527	0.619	1				
9	0.326	0.195	0.184	0.325	0.723	0.714	0.685	0.532	1			
10	0.116	0.057	−0.075	0.099	0.311	0.203	0.246	0.285	0.17	1		
11	0.308	0.15	0.091	0.11	0.344	0.353	0.232	0.3	0.28	0.484	1	
12	0.314	0.145	0.14	0.16	0.215	0.095	0.181	0.271	0.113	0.585	0.428	1
13	0.489	0.239	0.321	0.327	0.344	0.309	0.345	0.395	0.28	0.408	0.535	0.512
14	0.125	0.103	0.177	0.066	0.28	0.292	0.236	0.252	0.26	0.172	0.35	0.131
15	0.238	0.131	0.065	0.127	0.229	0.251	0.172	0.175	0.248	0.154	0.24	0.173
16	0.414	0.272	0.263	0.322	0.187	0.291	0.18	0.296	0.242	0.124	0.314	0.119
17	0.176	0.005	0.177	0.187	0.208	0.273	0.228	0.255	0.274	0.289	0.362	0.278
18	0.368	0.255	0.211	0.251	0.263	0.167	0.159	0.25	0.208	0.317	0.35	0.349
19	0.27	0.112	0.312	0.137	0.19	0.251	0.226	0.274	0.274	0.19	0.29	0.11
20	0.365	0.292	0.297	0.339	0.398	0.435	0.451	0.427	0.446	0.173	0.202	0.246
21	0.369	0.306	0.165	0.349	0.318	0.263	0.314	0.362	0.266	0.405	0.399	0.355
22	0.413	0.232	0.25	0.38	0.441	0.386	0.396	0.357	0.483	0.16	0.304	0.193
23	0.474	0.348	0.383	0.335	0.435	0.431	0.405	0.501	0.504	0.262	0.251	0.35
24	0.282	0.211	0.203	0.248	0.42	0.433	0.437	0.388	0.424	0.531	0.412	0.414

(continued)

Table 12.4 (continued)

	13	14	15	16	17	18	19	20	21	22	23	24
1												
2												
3												
4												
5												
6												
7												
8												
9												
10												
11												
12												
13	1											
14	0.195	1										
15	0.139	0.37	1									
16	0.281	0.412	0.325	1								
17	0.194	0.341	0.345	0.324	1							
18	0.323	0.201	0.334	0.344	0.448	1						
19	0.263	0.206	0.192	0.258	0.324	0.358	1					
20	0.241	0.302	0.272	0.388	0.262	0.301	0.167	1				
21	0.425	0.183	0.232	0.348	0.173	0.357	0.331	0.413	1			
22	0.279	0.243	0.246	0.283	0.273	0.317	0.342	0.463	0.374	1		
23	0.382	0.242	0.256	0.36	0.287	0.272	0.303	0.509	0.451	0.503	1	
24	0.358	0.304	0.165	0.262	0.326	0.405	0.374	0.366	0.448	0.375	0.434	1

Table 12.5 Solutions obtained by the varimax rotation for the data in Table 12.4

	EFA				PCA			
	1	2	3	ψ_j	1	2	3	Var($\mathbf{e}_j$)
Visual perception	0.64	0.14	0.26	0.50	0.69	0.17	0.23	0.44
Cubes	0.41	0.06	0.18	0.79	0.55	0.01	0.14	0.67
Paper form board	0.54	−0.06	0.22	0.66	0.64	−0.08	0.20	0.54
Flags	0.46	0.06	0.30	0.69	0.56	0.01	0.30	0.60
General information	0.11	0.22	0.77	0.35	0.10	0.20	0.80	0.30
Paragraph comprehension	0.15	0.09	0.81	0.32	0.16	0.10	0.83	0.28
Sentence completion	0.08	0.14	0.82	0.30	0.07	0.12	0.86	0.24
Word classification	0.27	0.23	0.61	0.50	0.26	0.20	0.67	0.44
Word meaning	0.14	0.06	0.85	0.26	0.14	0.09	0.86	0.23
Addition	−0.06	0.87	0.16	0.21	−0.12	0.82	0.19	0.28
Code	0.24	0.55	0.23	0.58	0.13	0.71	0.22	0.44
Counting dots	0.23	0.67	0.05	0.49	0.15	0.72	0.05	0.45
Straight-curved capitals	0.39	0.47	0.25	0.57	0.36	0.53	0.23	0.54
Word recognition	0.24	0.19	0.27	0.84	0.20	0.32	0.27	0.79
Number recognition	0.29	0.17	0.20	0.85	0.29	0.30	0.16	0.80
Figure recognition	0.55	0.14	0.18	0.64	0.62	0.22	0.13	0.55
Object-number	0.27	0.32	0.22	0.78	0.22	0.49	0.18	0.68
Number-figure	0.45	0.39	0.12	0.64	0.46	0.52	0.04	0.52
Figure-word	0.35	0.21	0.22	0.79	0.36	0.31	0.18	0.74
Deduction	0.43	0.14	0.45	0.59	0.46	0.14	0.48	0.53
Numerical puzzles	0.42	0.44	0.24	0.58	0.42	0.47	0.23	0.55
Problem reasoning	0.43	0.14	0.44	0.60	0.44	0.17	0.47	0.56
Series completion	0.50	0.23	0.45	0.50	0.51	0.23	0.47	0.46
Arithmetic problems	0.22	0.53	0.41	0.50	0.17	0.57	0.43	0.45

PCA loading matrices. The PCA loading matrix can be rotated if constraint (5.26) is removed, as explained in Note 5.5. In Table 12.4, var($\mathbf{e}_j$) for PCA is the variance of the resulting error values for variable j, i.e., the jth diagonal elements of the $n^{-1}\mathbf{E}'\mathbf{E}$, while the unique variance ψ_j for EFA can be associated with var($\mathbf{e}_j$) for PCA. There, we can find the similarity between EFA and PCA solutions. The difference is the interpretation for errors. For example, $\psi_1 = 0.5$ for EFA can be interpreted as that 50% of the variance in "visual perception" being *uniquely* and *exclusively* explained

by the corresponding unique factor, but var($\mathbf{e}_1$) = 0.44 for PCA cannot be interpreted so: it is interpreted simply as that 44% of the variance in "visual perception" remains unexplained by the three principal components.

12.11 Bibliographical Notes

Various subjects on EFA are exhaustively detailed in Bartholomew et al. (2011), Harman (1976), and Mulaik (2011). Papers reviewing EFA well include Yanai and Ichikawa (2007). Approaches to the noniterative estimation of parameters have also been found in EFA studies (e.g., Ihara & Kano, 1986; Kano, 1990).

Exercises

12.1. Present another example of a set of the variables for which EFA is useful.

12.2. Show that EFA model (12.3) with (12.4) can be rewritten as $\mathbf{x} = \mathbf{A}\mathbf{f} + \boldsymbol{\Psi}^{1/2}\mathbf{u}$ with $\mathbf{u} \sim N_p(\mathbf{0}_p, \mathbf{I}_p)$.

12.3. In model (12.3), factor vector $\mathbf{f}$ is regarded as a random vector. In contrast to this, the EFA model also exists in which $\mathbf{f}$ is regarded as a fixed parameter vector (Anderson & Rubin, 1956). This model is called a *fixed factor model*, which is expressed as

$$\mathbf{x}_i = \mathbf{A}\mathbf{f}_i + \mathbf{e}_i \tag{12.21}$$

by attaching subscript i to $\mathbf{x}$, $\mathbf{f}$, and $\mathbf{e}$ in (12.3) for explicitly showing that they are related to individual i. Show that if $\mathbf{A}$ is given, the squared norm of the error for i, $\|\mathbf{e}_i\|^2 = \|\mathbf{x}_i - \mathbf{A}\mathbf{f}_i\|^2$, is minimized for $\mathbf{f}_i = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{x}_i$.

12.4. Model (12.21) with $\mathbf{e}_i \sim N_p(\mathbf{0}_m, \boldsymbol{\Psi})$ implies $\mathbf{x}_i \sim N_p(\mathbf{A}\mathbf{f}_i, \boldsymbol{\Psi})$. Show that it leads to the log likelihood for $\mathbf{X} = \begin{bmatrix} \mathbf{x}'_1 \\ \vdots \\ \mathbf{x}'_n \end{bmatrix} = [\mathbf{y}_1, \ldots, \mathbf{y}_p]$ being expressed as

$$\begin{aligned} \log l(\mathrm{F}, \mathrm{A}, \boldsymbol{\Psi}) &= -\frac{n}{2}\log|\boldsymbol{\Psi}| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \mathbf{A}\mathbf{f}_i)'\boldsymbol{\Psi}^{-1}(\mathbf{x}_i - \mathbf{A}\mathbf{f}_i) \\ &= -\frac{n}{2}\log|\boldsymbol{\Psi}| - \frac{1}{2}\mathrm{tr}(\mathbf{X} - \mathbf{F}\mathbf{A}')\boldsymbol{\Psi}^{-1}(\mathbf{X} - \mathbf{F}\mathbf{A}')' \\ &= -\frac{1}{2}\left(n\sum_{j=1}^{p}\log\psi_j + \sum_{j=1}^{p}\frac{1}{\psi_j}\left\|\mathbf{X}_j - \mathbf{F}\mathbf{a}_j\right\|^2\right), \end{aligned} \tag{12.22}$$

with $\mathbf{F} = [\mathbf{f}_1, \ldots, \mathbf{f}_n]', \mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_p]'$, ψ_j the jth diagonal element of $\boldsymbol{\Psi}$, and the term irrelevant to $\mathbf{X}$, $\mathbf{F}$, $\mathbf{A}$, and $\boldsymbol{\Psi}$ omitted.

12.5. Show that $\mathbf{f}_i = \left(\mathbf{A}'\boldsymbol{\Psi}^{-1}\mathbf{A}\right)^{-1}\mathbf{A}'\boldsymbol{\Psi}^{-1}\mathbf{x}_i (i = 1, \ldots, n)$ maximizes (12.22) for given $\mathbf{A}$ and $\boldsymbol{\Psi}$, using $\left\|\boldsymbol{\Psi}^{-1/2}\mathbf{x}_i - \boldsymbol{\Psi}^{-1/2}\mathbf{A}\mathbf{f}_i\right\|^2 = (\mathbf{x}_i - \mathbf{A}\mathbf{f}_i)'\boldsymbol{\Psi}^{-1}(\mathbf{x}_i - \mathbf{A}\mathbf{f}_i)$.

12.6. Show that $\psi_j = n^{-1}\left\|\mathbf{x}_j - \mathbf{F}\mathbf{a}_j\right\|^2$ maximizes (12.22) for given $\mathbf{F}$ and $\mathbf{A}$, by noting the fact in Exercise 8.1.

12.7. Show that the MLE of $\mathbf{F}$, $\mathbf{A}$, and $\boldsymbol{\Psi}$ does not exist for (12.22), since this diverges to infinity when $\mathbf{F}$, $\mathbf{A}$, and $\boldsymbol{\Psi}$ are jointly estimated. A hint is found in the fact that $\mathbf{x}_j$ can be equal to $\mathbf{F}\mathbf{a}_j$.

12.8. Let us consider the model (12.21) with $\mathbf{e}_i \sim N_p(\mathbf{A}\mathbf{f}_i, \psi\mathbf{I}_m)$, i.e., the error variance for every variable constrained to equal ψ. Show that its log likelihood for the data matrix $\mathbf{X}$ is expressed as

$$\log l(\mathbf{F}, \mathbf{A}, \psi) = -\frac{np}{2}\log\psi - \frac{1}{2\psi}\|\mathbf{X} - \mathbf{F}\mathbf{A}'\|^2. \tag{12.23}$$

12.9. The maximization of log likelihood (12.23) has been introduced as a maximum likelihood estimation for principal component analysis (PCA) in Bishop (2006, p. 571). Show that maximizing (12.23) over $\mathbf{F}$, $\mathbf{A}$, and ψ is equivalent to minimizing (5.4) over $\mathbf{F}$ and $\mathbf{A}$, i.e., PCA.

12.10. A least squares method for EFA is formulated as minimizing

$$\|\mathbf{R} - (\mathbf{A}\mathbf{A}' + \boldsymbol{\Psi})\|^2 \tag{12.24}$$

over $p \times m$ loading matrix $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_m]'$ and $p \times p$ diagonal matrix $\boldsymbol{\Psi} = \begin{bmatrix} \psi_1 & & \\ & \ddots & \\ & & \psi_p \end{bmatrix}$ for the $p \times p$ inter-variable correlation coefficient matrix $\mathbf{R} = (r_{jk})$ obtained from a data set. Discuss how this method is rational.

12.11. Let $\mathbf{r}_j$ be the $(p - 1) \times 1$ vector obtained by deleting r_{jj} from the jth column of the correlation matrix $\mathbf{R}$ in (12.24). Show that the minimization of (12.24) over $\mathbf{A}$ and $\boldsymbol{\Psi}$ can be attained by the following algorithm (Harman & Jones, 1966):

Step 1. Initialize $\mathbf{A}$.

Step 2. Repeat the update of the ith row of $\mathbf{A}$ by the transpose of $\mathbf{a}_j = \left(\mathbf{A}_j'\mathbf{A}_j\right)^{-1}\mathbf{A}_j'\mathbf{r}_j$ for $j = 1, \ldots, m$, where $\mathbf{A}_j$ is the $(p - 1) \times m$ matrix obtained by deleting $\mathbf{a}_j'$ from the current $\mathbf{A}$

Step 3. Set $\psi_j = 1 - \mathbf{a}_j'\mathbf{a}_j$ to finish if convergence is reached; otherwise, go back to Step 2.

12.12. *Independent component analysis* (*ICA*) refers to a class of procedures, the most general form of whose models can be expressed as $\mathbf{x} = \mathbf{f}(\mathbf{s}) + \mathbf{e}$ (Izenman, 2008, p. 558). Here, $\mathbf{x}$ is a $p \times 1$ observed variable vector, $\mathbf{e}$ is an error vector, $\mathbf{s}$ is an $m \times 1$ vector containing unobserved signals originating from m mutually independent sources, and $\mathbf{f}(\mathbf{s})$ is a function of $\mathbf{s}$ providing a $p \times 1$ vector. Discuss relationships of ICA to EFA.

Part IV
Miscellaneous Procedures

The types of matrices to be analyzed by the procedures in this part differ from those in Parts II and III. The techniques in Chap. 13 are not procedures for analyzing data, but rather for transforming solutions. The data sets to be analyzed by the procedures in Chap. 14 are given as block and categorical data matrices. In Chap. 15, data sets are treated in which individuals are classified into some groups, while data are considered which describe the quasi-distances among objects in Chap. 16.

Chapter 13
Rotation Techniques

In some analysis procedures, the solution for a data set is *not uniquely determined*; multiple solutions exist. An example of such procedures is exploratory factor analysis (EFA). In this procedure, one of the solutions is first found, and then it is transformed into a useful solution that is included in multiple solutions. A family of such transformations is the *rotation* treated in this section. The rotation for EFA solutions in particular is called *factor rotation*, although the rotation can be used for solutions of procedures other than EFA. This chapter starts with illustrating why the term "rotation" is used, before explaining which solutions are useful in Sect. 13.3. This is followed by the introduction of some rotation techniques.

13.1 Geometric Illustration of Factor Rotation

As discussed with (12.16) in Sect. 12.5, when loading matrix $\hat{\mathbf{A}}$ is an EFA solution of a loading matrix, its transformed version,

$$\mathbf{A}_{\mathrm{T}} = \hat{\mathbf{A}}\mathbf{T}'^{-1}, \tag{13.1}$$

is *also a solution*. Here, $\mathbf{T}$ is an $m \times m$ matrix that satisfies (12.14), which is written again here:

$$\mathbf{T}'\mathbf{T} = \begin{bmatrix} 1 & & \# \\ & \ddots & \\ \# & & 1 \end{bmatrix}, \text{ or equivalently, } \quad \mathrm{diag}(\mathbf{T}'\mathbf{T}) = \mathbf{I}_m. \tag{13.2}$$

where diag() is defined in Note 12.1. In this section, we geometrically illustrate the transformation of $\hat{\mathbf{A}}$ into $\mathbf{A}_{\mathrm{T}} = \hat{\mathbf{A}}\mathbf{T}'^{-1}$, supposing that $\mathbf{T}$ is given.

K. Adachi, *Matrix-Based Introduction to Multivariate Data Analysis*,
https://doi.org/10.1007/978-981-15-4103-2_13

Let us use $\mathbf{a}_j'$ for the jth row of the original matrix $\hat{\mathbf{A}}$ and use $\mathbf{a}_j^{(\mathrm{T})\prime}$ for that of the transformed $\mathbf{A}_\mathrm{T}$. Then, $\mathbf{A}_\mathrm{T} = \hat{\mathbf{A}}\mathbf{T}'^{-1}$ is rewritten as

$$\mathbf{a}_j^{(\mathrm{T})\prime} = \mathbf{a}_i'\mathbf{T}'^{-1} \quad (j = 1, \ldots, p). \tag{13.3}$$

Post-multiplying both sides of (13.3) by $\mathbf{T}'$ leads to $\mathbf{a}_j^{(\mathrm{T})\prime}\mathbf{T}' = \mathbf{a}_j'$, i.e.,

$$\mathbf{a}_j' = \mathbf{a}_j^{(\mathrm{T})\prime}\mathbf{T}' \quad (j = 1, \ldots, p), \tag{13.4}$$

which shows that the original loading vector $\mathbf{a}_j'$ for variable j is expressed by the post-multiplication of the transformed $\mathbf{a}_j^{(\mathrm{T})\prime}$ by $\mathbf{T}'$. We suppose $m = 2$ and define the columns of $\mathbf{T}$ as

$$\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2], \text{with} \quad \|\mathbf{t}_1\| = \|\mathbf{t}_2\| = 1 \tag{13.5}$$

which satisfies (13.2). Using (13.5) and $\mathbf{a}_j^{(\mathrm{T})\prime} = [a_{j1}^{(\mathrm{T})}, a_{j2}^{(\mathrm{T})}]$, (13.4) is rewritten as

$$\mathbf{a}_j' = a_{j1}^{(\mathrm{T})}\mathbf{t}_1' + a_{j2}^{(\mathrm{T})}\mathbf{t}_1'. \tag{13.6}$$

It shows that the *original loading vector* for variable j is equal to the *sum* of $\mathbf{t}_k$ (k = 1, 2) *multiplied* by the *transformed loadings*. Its geometric implications are illustrated in the next two paragraphs.

In Table 13.1(A), we again show the original loading matrix $\hat{\mathbf{A}}$ in Table 12.1(A) obtained by EFA. Its row vectors $\mathbf{a}_j'$ (j = 1, …, 8) corresponding to variables are shown in Fig. 13.1a; the vector $\mathbf{a}_7'$ for H is depicted by the line extending to [−0.63, 0.46], and the other vectors are done in parallel manners. Now, let us consider transforming $\hat{\mathbf{A}}$ into $\mathbf{A}_\mathrm{T} = \hat{\mathbf{A}}\mathbf{T}'^{-1}$ by

Table 13.1 A solution obtained with EFA (Table 12.1A) and an example of its rotated version

	(A) Before rotation			(B) After rotation		
	$\hat{\mathbf{A}}$		ψ_j	$\mathbf{A}_\mathrm{T}$		ψ_j
A	0.77	−0.38	0.26	1.03	−0.76	0.26
C	0.61	0.50	0.38	0.56	0.32	0.38
I	0.67	−0.36	0.41	0.90	−0.69	0.41
B	−0.74	−0.40	0.30	−0.75	−0.15	0.30
T	0.79	0.43	0.18	0.80	0.16	0.18
V	0.76	−0.44	0.22	1.04	−0.82	0.22
H	−0.63	0.46	0.39	−0.89	0.79	0.39
P	0.70	0.18	0.47	0.77	−0.09	0.47
ϕ_{12}	0.00			0.57		

$$\mathbf{T}'^{-1} = \begin{bmatrix} 10.18 & -0.42 \\ -0.32 & 1.14 \end{bmatrix}, \text{following from } \mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2] = \begin{bmatrix} 0.94 & 0.26 \\ 0.34 & 0.97 \end{bmatrix}. \quad (13.7)$$

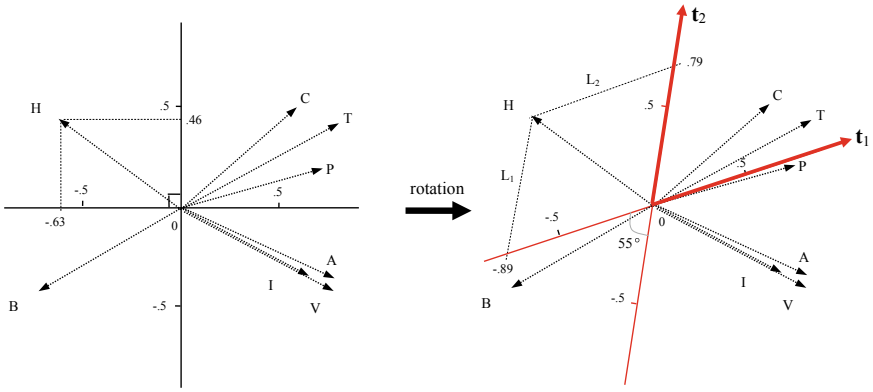
This $\mathbf{T}'^{-1}$ leads to $\mathbf{A}_T = \hat{\mathbf{A}}\mathbf{T}'^{-1}$ in Table 13.1(B). There, we find that the vector for H is $\mathbf{a}_7^{(T)\prime} = \mathbf{a}_7'\mathbf{T}'^{-1} = [-0.89, 0.79]$, transformed from $\mathbf{a}_7' = [-0.63, 0.46]$ in (A). Those two vectors satisfy the relationship in (13.6):

$$[-0.63, 0.46] = -0.89\mathbf{t}_1' + 0.79\mathbf{t}_2', \quad (13.8)$$

with $\mathbf{t}_1' = [0.94, 0.34]$ and $\mathbf{t}_2' = [0.26, 0.97]$.

The geometric implication of (13.8), which is an example of (13.6), is illustrated in Fig. 13.1b. There, the axes extending in the directions of $\mathbf{t}_1'$ = [0.94, 0.34], $\mathbf{t}_2'$ = [0.26, 0.98] are depicted, together with the original loading vectors $\mathbf{a}_1'$, …, $\mathbf{a}_8'$ whose locations are the same as in (A). Let us note that vector $\mathbf{a}_7'$ for H satisfies (13.8); i.e., the −0.89 times of $\mathbf{t}_1'$ plus the 0.79 times of $\mathbf{t}_2'$ is equivalent to $\mathbf{a}_7'$ = [−0.63, 0.64]. Here, the transformed loadings −0.89 and 0.79 can be viewed as the coordinates of point H on $\mathbf{t}_1$ and $\mathbf{t}_2$ axes, as shown by the dotted lines L_1 and L_2 in Fig. 13.1b, where L_1 and L_2 extend in *parallel* to $\mathbf{t}_2$ and $\mathbf{t}_1$, respectively. This relationship holds for the other loading vectors.

In summary, transformation (13.1) implies the rotation of the original horizontal and vertical axes in Fig. 13.1a to the *new axes* extending in the direction of the column vectors of $\mathbf{T}$ as in Fig. 13.1b, where the transformed loadings are the *coordinates* on the new axes. The reason why (13.1) is called *rotation* is found above.



(a) Variable vectors for Table 13.1(A)

(b) Rotated axes, the coordinates which give the loadings in Table 13.1(B)

Fig. 13.1 Illustration of rotation as that of axes

13.2 Oblique and Orthogonal Rotation

Rotation is classified into oblique and orthogonal. The transformation illustrated in the last section is *oblique rotation*, since the new axes are intersected obliquely, as in Fig. 13.1b. On the other hand, *orthogonal rotation* refers to the rotation of *axes* by keeping their *orthogonal intersection*, whose example is described later in Fig. 13.2a. In orthogonal rotation, constraint (13.2) is strengthened so that it is the $m \times m$ identity matrix:

$$\mathbf{T}'\mathbf{T} = \mathbf{I}_m. \tag{13.9}$$

The matrix $\mathbf{T}$ satisfying (13.9) is said to be *orthonormal,* and its properties are detailed in Appendix A.1.2. Customarily, the rotation made by orthonormal $\mathbf{T}$ is not called orthonormal rotation, but rather *orthogonal rotation*. Using (13.9), (13.1) is simplified as

$$\mathbf{A}_\mathrm{T} = \hat{\mathbf{A}}\mathbf{T} \tag{13.10}$$

in orthogonal rotation.

In summary, rotation is classified into two types:

[1] *Oblique* rotation (13.1) with $\mathbf{T}$ constrained as (13.2)
[2] *Orthogonal* rotation (13.10) with $\mathbf{T}$ constrained as (13.9)

Orthogonal rotation can be viewed as a special case of oblique rotation in which (13.2) is strengthened as (13.9).

13.3 Rotation to Simple Structure

The transformed loading matrix in Table 13.1(B) is not a useful one. That matrix is merely an example for illustrating rotation. A "good rotation procedure" is one that gives a useful matrix. Here, we have the question: "What matrix is *useful*?" A variety of answers exist; which answer is right varies from case to case.

When a matrix is a variables $\times$ factors loading matrix, usefulness can be defined as "*interpretability*", i.e., being easily interpreted. What matrix is interpretable? An ideal example is shown in Table 13.2(A), where # indicates a nonzero (positive or negative) value. This matrix has two features:

[1] *Sparse*, i.e., a number of elements are *zero*
[2] *Well classified*, i.e., different variables load different factors

Feature [1] allows us to focus on the nonzero elements to capture the relationships of variables to factors. Feature [2] clarifies the differences between factors. The matrix in Table 13.2(A) is said to have a *simple structure* (Thurstone, 1947).

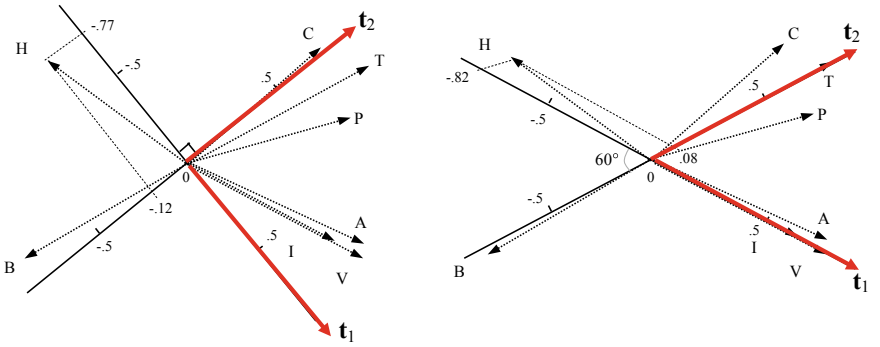
Table 13.2(A) shows an ideally simple structure, but it is almost impossible to have such a matrix; $\mathbf{T}$ cannot be chosen so that some elements of $\mathbf{A}_T = \hat{\mathbf{A}}\mathbf{T}'^{-1}$ are exactly zero as in (A). However, it is feasible to obtain $\mathbf{A}_T = \hat{\mathbf{A}}\mathbf{T}'^{-1}$ that approximates the ideal. It is illustrated in Table 13.2(B). There, "Small" stands for a value *close to zero*, but not exactly being zero, while "Large" expresses a value with a large absolute value. A matrix, which is not ideal but *approximates ideal structure*, is also said to have a *simple structure* in the literature for psychometrics (statistics for psychology).

Let us remember that $\mathbf{A}_T = \hat{\mathbf{A}}\mathbf{T}'^{-1}$ can be viewed as the coordinates on rotated axes. How should the axes be rotated so as to make the loading matrix $\mathbf{A}_T$ be of a simple structure? One answer is found in Fig. 13.2, where the useful orthogonal and oblique rotation for the variable vectors in Fig. 13.1a is illustrated. First, let us note the axes of $\mathbf{t}_1$ and $\mathbf{t}_2$ in Fig. 13.2b. The former axis is approximately *parallel* to the vectors for a group of variables {A, V, I, H} (Group 1), while the latter is almost *parallel* to those for another group {C, T, B, P} (Group 2). Thus, Group 1 has the coordinates of large absolute values on the $\mathbf{t}_1$ axis, but shows those of small absolutes on the $\mathbf{t}_2$ axis. On the other hand, Group 2 shows the coordinates of large and small absolutes for $\mathbf{t}_2$ and $\mathbf{t}_1$ axes, respectively. The resulting loading matrix is presented in Table 13.3(C). There, the matrix successfully attains the simple structure as in Table 13.2(B). Orthogonal rotation is illustrated in Fig. 13.2a, where $\mathbf{t}_1$ and $\mathbf{t}_2$ are orthogonally intersected; (13.9) is satisfied. On the other hand, the axes are obliquely intersected in Fig. 13.2b. Also in (A), the $\mathbf{t}_1$ and $\mathbf{t}_2$ axes are almost parallel to Groups 1 and 2, respectively, which provides the matrix having a simple structure in Table 13.3(B).

In the above paragraph, we visually illustrated how $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2]$ is set to be parallel to groups of variable vectors so that $\mathbf{A}_T = \hat{\mathbf{A}}\mathbf{T}'^{-1}$ has a simple structure. But, this task can only be attained by human vision and is impossible even by that when m exceeds three-dimensions. Indeed, the optimal $\mathbf{T}$ is obtained not visually but *computationally* with

Table 13.2 Simple structure in a matrix of variables × factors

(A) Ideally simple			(B) Simple	
Variable	F1	F2	F1	F2
1	**#**	0	**Large**	Small
2	0	**#**	Small	**Large**
3	0	**#**	Small	**Large**
4	**#**	0	**Large**	Small
5	0	**#**	Small	**Large**
6	**#**	0	**Large**	Small
7	**#**	0	**Large**	Small
8	0	**#**	Small	**Large**



(a) Orthogonal Rotation (b) Oblique Rotation

Fig. 13.2 Illustrations of rotation to a simple structure

Table 13.3 A solution obtained with EFA (Table 12.1A) and its rotated versions

(A) Before rotation				(B) After varimax rotation			(C) After geomin rotation		
	A		ψ_j	$\mathbf{A}_T$		ψ_j	$\mathbf{A}_T$		ψ_j
A	**0.77**	−0.38	0.26	**0.81**	0.28	0.26	**0.82**	0.08	0.26
C	**0.61**	0.50	0.38	0.07	**0.78**	0.38	−0.13	**0.84**	0.38
I	**0.67**	−0.36	0.41	**0.73**	0.22	0.41	**0.74**	0.04	0.41
B	**−0.74**	−0.40	0.30	−0.24	**−0.80**	0.30	−0.04	−0.82	0.30
T	**0.79**	0.43	0.18	0.25	**0.87**	0.18	0.04	0.88	0.18
V	**0.76**	−0.44	0.22	**0.85**	0.23	0.22	**0.87**	0.01	0.22
H	**−0.63**	0.46	0.39	**−0.77**	−0.12	0.39	**−0.82**	0.08	0.39
P	**0.70**	0.18	0.47	0.37	**0.63**	0.47	0.23	**0.58**	0.47
ϕ_{12}	0.00			0.00			0.48		

$$\text{maximize } \text{Simp}(\mathbf{A}_T) = \text{Simp}(\hat{\mathbf{A}}\mathbf{T}'^{-1}) \text{ over } \mathbf{T} \text{ subject to (13.2) or (13.9).} \quad (13.11)$$

Here, $\text{Simp}(\hat{\mathbf{A}}\mathbf{T}'^{-1})$ is the abbreviation for the *simplicity* of $\hat{\mathbf{A}}\mathbf{T}'^{-1}$ and is a function of $\mathbf{T}$ that stands for how well $\mathbf{A}_T = \hat{\mathbf{A}}\mathbf{T}'^{-1}$ approximates the ideal simple structure, that is, how simple the structure in $\mathbf{A}_T$ is. The procedures formulated as (13.11) are generally called (algebraic) *rotation techniques*. In exactness, we should call them *simple structure rotation techniques* in order to distinguish them from the rotation that does not involve a simple structure. A number of simple structure rotation techniques have been proposed so far, which differ in terms of how to define $\text{Simp}(\mathbf{A}_T) = \text{Simp}(\hat{\mathbf{A}}\mathbf{T}'^{-1})$. Two popular techniques are introduced in the next two sections.

13.4 Varimax Rotation

The rotation techniques with (13.9) chosen as the constraint in (13.11) are called *orthogonal* rotation techniques. Among them, the *varimax* rotation method presented by Kaiser (1958) is well known. In this method, the simplicity of $\mathbf{A}_{\mathrm{T}} = \hat{\mathbf{A}}\mathbf{T}$ is defined as

$$\mathrm{Simp}(\mathbf{A}_{\mathrm{T}}) = \mathrm{Simp}(\hat{\mathbf{A}}\mathbf{T}) = \sum_{k=1}^{m} \mathrm{var}\left(a_{1k}^{(\mathrm{T})\,2} \cdots a_{pk}^{(\mathrm{T})\,2}\right) \tag{13.12}$$

to be maximized. Here, we have used the fact that (13.1) is simplified as (13.10) and $\mathrm{var}(a_{1k}^{(T)\,2} \cdots a_{pk}^{(\mathrm{T})\,2})$ stands for the *variance of the squared elements* in the kth column of $\mathbf{A}_{\mathrm{T}} = (a_{jk}^{(\mathrm{T})})$:

$$\mathrm{var}\left(a_{1k}^{(\mathrm{T})\,2} \cdots a_{pk}^{(\mathrm{T})\,2}\right) = \frac{1}{p}\sum_{j=1}^{p}\left(a_{jk}^{(\mathrm{T})\,2} - \frac{1}{p}\sum_{l=1}^{p} a_{lk}^{(\mathrm{T})\,2}\right)^2. \tag{13.13}$$

That is, the varimax rotation is formulated as

$$\text{maximize}\ \mathrm{simp}(\hat{\mathbf{A}}\mathbf{T}) = \frac{1}{p}\sum_{k=1}^{m}\sum_{j=1}^{p}\left(a_{jk}^{(\mathrm{T})2} - \frac{1}{p}\sum_{l=1}^{p} a_{lk}^{(\mathrm{T})2}\right)^2 \ \text{over}\ \mathbf{T}\ \text{subject}\ \mathbf{T}'\mathbf{T} = \mathbf{I}_m. \tag{13.14}$$

For this maximization, an iterative algorithm is needed. One of the algorithms can be included in the gradient methods introduced in Appendix A.6.3 (Jennrich, 2001). However, that is out of the scope of this book.

We should note that variance (13.13) is not defined for loadings $a_{jk}^{(\mathrm{T})}$ but for its squares $a_{jk}^{(\mathrm{T})}$; they are irrelevant to whether $a_{jk}^{(\mathrm{T})}$ are positive or negative, but are relevant to the absolute values of $a_{jk}^{(\mathrm{T})}$. If variance (13.13) is larger, the *absolute values* of the loadings in each column of $\mathbf{A}_{\mathrm{T}}$ would take a *variety* of values so that

$$\text{some absolute values are larger, but others are small,} \tag{13.15}$$

as illustrated in Table 13.2(B).

The sum of the above variances over m columns defines the simplicity as in (13.12). By maximizing the sum, all m columns can have loadings with (13.15). This allows us to consider the two different $\mathbf{A}_{\mathrm{T}}$ results illustrated in Table 13.4(A) and (B). There, we find that (A) is equivalent to Table 13.2(B); i.e., it shows a simple structure, while Table 13.4(B) is not simple, in that the same variables heavily load two factors. However, (13.14) hardly provides a loading matrix $\mathbf{A}_{\mathrm{T}}$, as

Table 13.4 Variables × factors matrices with and without a simple structure

Variable	(A) Simple		(B) Not simple	
	F1	F2	F1	F2
1	**Large**	Small	**Large**	**Large**
2	Small	**Large**	Small	Small
3	Small	**Large**	Small	Small
4	**Large**	Small	**Large**	**Large**
5	Small	**Large**	Small	Small
6	**Large**	Small	**Large**	**Large**
7	**Large**	Small	**Large**	**Large**
8	Small	**Large**	Small	Small

in Table 13.4(B), since it necessitates $\mathbf{t}_1$ and $\mathbf{t}_2$ extending almost in parallel, which contradicts constraint (13.9).

The varimax rotation for loading matrix $\hat{\mathbf{A}}$ in Table 13.3(A) provides the rotation matrix

$$\mathbf{T} = \begin{bmatrix} 0.705 & 0.710 \\ -0.711 & 0.704 \end{bmatrix}, \tag{13.16}$$

which is the solution for (13.14). Post-multiplication of $\hat{\mathbf{A}}$ in Table 13.3(A) by (13.16) yields the matrix $\mathbf{A}_\mathrm{T} = \hat{\mathbf{A}}\mathbf{T}$ in Table 13.3(B) that shows a simple structure. Indeed, Fig. 13.2a has been depicted according to (13.16).

Let us compare $\hat{\mathbf{A}}$ in Table 13.3(A) and $\mathbf{A}_\mathrm{T}$ in (B). It is difficult to reasonably interpret the former loadings in (A), as all variables show the loadings of large absolute values for Factor 1 and those of rather small absolutes for Factor 2. It obliges one to consider that Factor 1 explains all variables, while Factor 2 is irrelevant to all variables, which implies that Factor 2 is trivial. On the other hand, $\mathbf{A}_\mathrm{T} = \hat{\mathbf{A}}\mathbf{T}$ can be reasonably interpreted in the same manner as described in Sect. 12.7.

13.5 Geomin Rotation

The phrase "*maximize* Simp($\mathbf{A}_\mathrm{T}$)" in (13.11) is equivalent to "*minimize* $-1 \times$ Simp($\mathbf{A}_\mathrm{T}$)". Here, $-1 \times$ Simp($\mathbf{A}_\mathrm{T}$) can be rewritten as Comp($\mathbf{A}_\mathrm{T}$) which abbreviates the *complexity* of $\mathbf{A}_\mathrm{T}$ and represents to what extent $\mathbf{A}_\mathrm{T}$ deviates from a simple structure. Some rotation techniques are formulated as substituting "minimize Comp($\mathbf{A}_\mathrm{T}$)" for "maximize Simp($\mathbf{A}_\mathrm{T}$)" in (13.11). One of them is Yates's (1987) *geomin* rotation method, in which complexity is defined as

$$\mathrm{Comp}(\mathbf{A}_\mathrm{T}) = \mathrm{Comp}(\hat{\mathbf{A}}\mathbf{T}'^{-1}) = \sum_{j=1}^{p}\left\{\prod_{k=1}^{m}\left(a_{jk}^{(\mathrm{T})2}+\varepsilon\right)\right\}^{1/m}, \tag{13.17}$$

with ε a specified small positive value such as 0.01. The geomin rotation method has orthogonal and oblique versions. In this section, we treat the latter, i.e., the *oblique geomin* rotation, which is formulated as

$$\text{minimize } \mathrm{Comp}(\hat{\mathbf{A}}\mathbf{T}'^{-1}) = \sum_{j=1}^{p}\left\{\prod_{k=1}^{m}\left(a_{jk}^{(\mathrm{T})2}+\varepsilon\right)\right\}^{1/m} \text{ over } \mathbf{T} \text{ subject to (13.2)}. \tag{13.18}$$

For this minimization, an iterative algorithm is needed. One of the algorithms can be included in the gradient methods introduced in Appendix A.6.3 (Jennrich, 2002). However, that is beyond the scope of this book.

Let us note the parenthesized part in the right-hand side of (13.17):

$$\prod_{k=1}^{m}\left(a_{jk}^{(\mathrm{T})2}+\varepsilon\right) = \left(a_{j1}^{(\mathrm{T})2}+\varepsilon\right)\times\cdots\times\left(a_{jm}^{(\mathrm{T})2}+\varepsilon\right). \tag{13.19}$$

It is close to zero, if some of $a_{jk}^{(\mathrm{T})}$ are close to zero, which would give a matrix approximating that in Table 13.2(A). The sum of (13.19) over p variables is minimized as in (13.18). This minimization for $\hat{\mathbf{A}}$ in Table 13.3(A) provides the rotation matrix

$$\mathbf{T}'^{-1} = \begin{bmatrix} 0.581 & 0.582 \\ -0.979 & 0.979 \end{bmatrix}. \tag{13.20}$$

Post-multiplication of $\hat{\mathbf{A}}$ in Table 13.3(A) by (13.20) yields $\mathbf{A}_\mathrm{T} = \hat{\mathbf{A}}\mathbf{T}'^{-1}$ in Table 13.3(C). This has also been presented in Table 12.1(B), as described in Sect. 12.7.

The reason for adding a small positive constant ε to loadings, as in (13.19), is as follows: (13.19) would be $\prod_{k=1}^{m} a_{jk}^{(\mathrm{T})2} = a_{j1}^{(\mathrm{T})2}\times\cdots\times a_{jm}^{(\mathrm{T})2}$ without ε. Then, the solution which allows $\prod_{k=1}^{m} a_{jk}^{(\mathrm{T})2}$ to attain the lower bound 0 is not uniquely determined; multiple solutions could exist. For example, let m be 2. If $a_{j1}^{(\mathrm{T})} = 0$, then $a_{j1}^{(\mathrm{T})2}\times a_{j2}^{(\mathrm{T})2} = 0$ whatever value $a_{j2}^{(\mathrm{T})}$ takes. This existence of multiple solutions is avoided by adding ε as in (13.19).

13.6 Orthogonal Procrustes Rotation

In this section, we introduce *Procrustes* rotation, whose purpose is *different* from the procedures treated so far. Procrustes rotation generally refers to a class of rotation techniques to rotate $\hat{\mathbf{A}}$ so that the resulting $\mathbf{A}_{\mathrm{T}}$ is *matched* with a *target* matrix $\mathbf{B}$. The rotation was originally conceived by Mosier (1939) and named by Hurley and Cattell (1962) after a figure appearing in Greek mythology.

Let us consider *orthogonal Procrustes rotation* with (13.9), i.e., $\mathbf{T}$ ($m \times m$) constrained to be orthonormal. This is formulated as

$$\text{minimize} f(\mathbf{T}) = \left\|\mathbf{B} - \hat{\mathbf{A}}\mathbf{T}\right\|^2 \text{ over } \mathbf{T} \text{ subject to } \mathbf{T}'\mathbf{T} = \mathbf{I}_m. \tag{13.21}$$

This is useful for every case, in which one wishes to match $\hat{\mathbf{A}}\mathbf{T}$ to target $\mathbf{B}$ and examine how *similar* the resulting matrix $\mathbf{A}_{\mathrm{T}} = \hat{\mathbf{A}}\mathbf{T}$ is to the target, under constraint (13.9).

The function $f(\mathbf{T})$ in (13.21) can be expanded as

$$f(\mathbf{T}) = \|\mathbf{B}\|^2 - 2\mathrm{tr}\mathbf{B}'\hat{\mathbf{A}}\mathbf{T} + \mathrm{tr}\mathbf{T}'\hat{\mathbf{A}}'\hat{\mathbf{A}}\mathbf{T} = \|\mathbf{B}\|^2 - 2\mathrm{tr}\mathbf{B}'\hat{\mathbf{A}}\mathbf{T} + \|\mathbf{A}\|^2, \tag{13.22}$$

where we have used $\mathbf{TT}' = \mathbf{I}_m$ following from (13.9). In the right-hand side of (13.22), only $-2\mathrm{tr}\mathbf{T}'\hat{\mathbf{A}}'\mathbf{B}$ is relevant to $\mathbf{T}$. Thus, the minimization of (13.22) amounts to

$$\text{maximize } g(\mathbf{T}) = \mathrm{tr}\mathbf{B}'\hat{\mathbf{A}}\mathbf{T} \text{ over } \mathbf{T} \text{ subject to } \mathbf{T}'\mathbf{T} = \mathbf{I}_m. \tag{13.23}$$

This problem is equivalent to the one in Theorem A.4.2 (Appendix A.4.2). As found there, the solution of $\mathbf{T}$ is given through the singular value decomposition of $\hat{\mathbf{A}}'\mathbf{B}$.

A numerical example is given in Table 13.5. The matrices $\mathbf{B}$ and $\hat{\mathbf{A}}$ presented there seem to be very different. The orthogonal Procrustes rotation for them provide $\mathbf{T} = \begin{bmatrix} 0.53 & 0.85 \\ -0.85 & 0.53 \end{bmatrix}$. The resulting $\hat{\mathbf{A}}\mathbf{T}$ is shown in the right-hand side of Table 13.5, where $\hat{\mathbf{A}}\mathbf{T}$ is found to be very similar to $\mathbf{B}$.

Table 13.5 Example of orthogonal Procrustes rotation

$\mathbf{B}$		$\hat{\mathbf{A}}$		$\hat{\mathbf{A}}\,\mathbf{T}$	
0.0	0.8	0.6	0.4	−0.02	0.72
0.3	0.7	0.8	0.1	0.34	0.73
0.6	0.6	0.8	−0.2	0.59	0.57
0.8	0.1	0.5	−0.6	0.77	0.11
0.9	0.0	0.5	−0.8	0.94	0.00

13.7 Bibliographical Notes

Simple structure rotation techniques are exhaustively described in Browne (2001) and Mulaik (2011). Procrustes rotation techniques are detailed in Gower and Dijksterhuis (2004), with its special extended version presented by Adachi (2009). The simple structure rotation can be related to the sparse estimation, as discussed in Sect. 22.9 and other literature (e.g., Trendafilov, 2014).

Exercises

13.1. Show that $\mathbf{T} = \mathbf{S}\mathrm{diag}(\mathbf{S}'\mathbf{S})^{-1/2}$ satisfies (13.2), where $\mathrm{diag}(\mathbf{S}'\mathbf{S})$ denotes the $m \times m$ diagonal matrix whose diagonal elements $d_1, \ldots, d_m$ are those of $\mathbf{S}'\mathbf{S}$ (Note 12.1) and $\mathrm{diag}(\mathbf{S}'\mathbf{S})^{-1/2}$ is the $m \times m$ diagonal matrix whose diagonal elements are $1/d_1^{1/2}, \ldots, 1/d_m^{1/2}$.

13.2. Show that a 2×2 orthonormal matrix $\mathbf{T}$ is expressed as $\mathbf{T} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$.

13.3. Thurstone (1947) defined simple structure with provisions, which have been rewritten more clearly by Browne (2001, p. 115) as follows:

[1] Each row should contain at least one zero.
[2] Each column should contain at least m zeros, with m the number of factors.
[3] Every pair of columns should have several rows with a zero in one column but not the other.
[4] If $m \geq 4$, every pair of columns should have several rows with zeros in both columns.
[5] Every pair of columns should have a few rows with nonzero loadings in both columns.

Present an example of a 20×4 matrix meeting provisions [1]–[5].

13.4. Minimizing $\frac{1}{m}\sum_{k=1}^{m-1}\sum_{l=k+1}^{m}\sum_{j=1}^{p}(a_{jk}^{(\mathrm{T})2} - \bar{a}_{.k}^{(\mathrm{T})2})(a_{jl}^{(\mathrm{T})2} - \bar{a}_{.l}^{(\mathrm{T})2})$ over $\mathbf{T}$ subject to $\mathrm{diag}(\mathbf{T}'\mathbf{T}) = \mathbf{I}_m$ is included in a family of oblique rotation called *oblimin rotation* (Jennrich & Sampson, 1966), where $a_{jk}^{(\mathrm{T})}$ is the (j, k) element of the rotated loading matrix $\hat{\mathbf{A}}\mathbf{T}^{\prime-1}$. Discuss the purpose of the above minimization.

13.5. Oblique rotation tends to give a matrix of a *simpler* structure than orthogonal rotation. Explain its reason.

13.6. Show that orthogonal rotation is feasible for the $p \times m$ matrix $\mathbf{A}$ that minimizes $\|\mathbf{V} - \mathbf{A}\mathbf{A}'\|^2$ subject to $\mathbf{A}'\mathbf{A} = \mathbf{I}_m$ for given $\mathbf{V}$.

13.7. Show that oblique rotation is feasible for the solution of principal component analysis, if constraint (5.25) is relaxed as $n^{-1}\mathrm{diag}(\mathbf{F}'\mathbf{F}) = \mathbf{I}_m$ without (5.26). Here, diag() defined in Note 12.1.

13.8. Show the objective function (13.12) in the varimax rotation can be rewritten as

$$f = \frac{1}{n}\mathrm{tr}\mathbf{T}'\hat{\mathbf{A}}'\{(\hat{\mathbf{A}}\mathbf{T}) \odot (\hat{\mathbf{A}}\mathbf{T}) \odot (\hat{\mathbf{A}}\mathbf{T})\} - \frac{1}{n^2}\mathrm{tr}\mathbf{T}'\hat{\mathbf{A}}'\hat{\mathbf{A}}\mathbf{T}\{\mathrm{diag}(\mathbf{T}'\hat{\mathbf{A}}'\hat{\mathbf{A}}\mathbf{T})\}.$$

(ten Berge, Knol, & Kiers, 1988). Here, diag() is defined in Note 12.1, and $\odot$ denotes the element-wise product called the *Hadamard product* and defined as (17.69):

$$\mathbf{X} \odot \mathbf{Y} = \begin{bmatrix} x_{11}y_{11} & \cdots & x_{1p}y_{1p} \\ & \vdots & \\ x_{n1}y_{n1} & \cdots & x_{np}y_{np} \end{bmatrix} = \left(x_{ij}y_{ij}\right)(n \times p) \text{ for } n \times p \text{ matrices}$$

$$\mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ & \vdots & \\ x_{n1} & \cdots & x_{np} \end{bmatrix} \text{and } \mathbf{Y} = \begin{bmatrix} y_{11} & \cdots & y_{1p} \\ & \vdots & \\ y_{n1} & \cdots & y_{np} \end{bmatrix}.$$

13.9. *Generalized orthogonal rotation* is formulated as minimizing $\sum_{k=1}^{K}\|\mathbf{H} - \mathbf{A}_k\mathbf{T}_k\|^2$ over $\mathbf{H}, \mathbf{T}_1, \ldots, \mathbf{T}_\mathrm{K}$ subject to $\mathbf{T}_k'\mathbf{T}_k = \mathbf{T}_k\mathbf{T}_k' = \mathbf{I}_m$, $k = 1, \ldots, K$, for given $p \times m$ matrices $\mathbf{A}_1$, …, $\mathbf{A}_K$. Show that the minimization can be attained by the following algorithm:

Step 1. Initialize $\mathbf{T}_1, \ldots, \mathbf{T}_K$.
Step 2. Set $\mathbf{H} = K^{-1}\sum_{k=1}^{K}\mathbf{A}_k\mathbf{T}_k$.
Step 3. Compute the SVD $\mathbf{A}_k'\mathbf{H} = \mathbf{K}_k\Lambda_k\mathbf{L}_k'$ to set $\mathbf{T}_k = \mathbf{K}_k\mathbf{L}_k'$ for $k = 1, \ldots, K$.
Step 4. Finish if convergence is reached; otherwise, go back to Step 2.

13.10. Show

$$K = \sum_{k=1}^{K}\|\mathbf{H} - \mathbf{A}_k\mathbf{T}_k\|^2 = \sum_{k=1}^{K-1}\sum_{l=k+1}^{K}\|\mathbf{A}_k\mathbf{T}_k - \mathbf{A}_l\mathbf{T}_l\|^2$$

for $\mathbf{H}$ in Step 2 described in Exercise 13.9.

13.11. Let us consider the minimization of $\|[\mathbf{M}, \mathbf{c}] - \mathbf{A}\mathbf{T}\|^2$ over $\mathbf{T}(m \times m)$ and $\mathbf{c}$ $(p \times 1)$ subject to $\mathbf{T}'\mathbf{T} = \mathbf{T}\mathbf{T}' = \mathbf{I}_m$ for given $\mathbf{M}(p \times (m-1))$ and $\mathbf{A}(p \times m)$. Here, $[\mathbf{M}, \mathbf{c}]$ is the $p \times m$ matrix whose final column $\mathbf{c}$ is unknown. Show that the minimization can be attained by the following algorithm:

Step 1. Initialize $\mathbf{T}$.
Step 2. Set $\mathbf{c}$ to the final column of $\mathbf{A}\mathbf{T}$.
Step 3. Compute the SVD $\mathbf{A}'[\mathbf{M}, \mathbf{c}] = \mathbf{K}\Lambda\mathbf{L}'$ to set $\mathbf{T} = \mathbf{K}\mathbf{L}'$.
Step 4. Finish if convergence is reached; otherwise, go back to Step 2.

13.12. Kier's (1994) *simplimax rotation*, which is used for having a matrix of simple structure, is a generalization of the Procrustes rotation introduced in Sect. 13.6. In the simplimax rotation, target matrix $\mathbf{B}$ is unknown except for that $\mathbf{B}$ is constrained to have a specified number of zero elements: $\left\|\mathbf{B} - \hat{\mathbf{A}}\mathbf{T}^{\prime-1}\right\|^2$ is minimized over $\mathbf{B}$ and $\mathbf{T}$ subject to (13.2) or (13.9) and s elements being zero in $\mathbf{B}$, though the locations of the s zero elements are unknown. Show that, for fixed $\mathbf{T}$, the optimal $\mathbf{B} = (b_{jk})$ is given by $b_{jk} = \begin{cases} 0 & \text{if } a_{jk}^{[\mathrm{T}]2} \leq a_{<s>}^{[\mathrm{T}]2} \\ a_{jk}^{[\mathrm{T}]} & \text{otherwise} \end{cases}$, where $a_{jk}^{[\mathrm{T}]}$ is the (j,k) element of $\hat{\mathbf{A}}\mathbf{T}^{\prime-1}$ and $a_{<s>}^{[\mathrm{T}]2}$ is the sth smallest value among the squares of the elements in $\hat{\mathbf{A}}\mathbf{T}^{\prime-1}$.

Chapter 14
Canonical Correlation and Multiple Correspondence Analyses

In this chapter, we treat procedures for the data set in which *variables* are classified into some *groups*. Such a data set is expressed as a *block matrix*, introduced in Sect. 14.1. Then, we describe *canonical correlation analysis* (*CCA*) for data with two groups of variables, which is followed by the introduction of *generalized CCA* (*GCCA*) for more than two groups of variables in Sect. 14.3. GCCA provides a foundation for a procedure, in which the *multivariate categorical data* described in Sect. 14.4 are analyzed. This procedure is called *multiple correspondence analysis* (*MCA*), whose purpose is to *quantify un-numerical categories*, i.e., finding the optimal scores to be given to the categories, as shown in Sect. 14.5.

CCA was originally formulated by Hotelling (1936), and some types of GCCA have been presented (Gifi, 1990; Kettenring, 1984; van de Geer, 1984), among which Gifi's (1990) approach is chosen for describing GCCA and MCA in this chapter.

14.1 Block Matrices

We start with introducing the blocks of a matrix by the following note:

Note 14.1. Blocks of a Matrix
We can rewrite a 5×4 matrix $\mathbf{Y}$ as follows:

$$\mathbf{Y} = \left[\begin{array}{cc|cc} y_{11} & y_{12} & y_{13} & y_{14} \\ y_{21} & y_{22} & y_{23} & y_{24} \\ y_{31} & y_{32} & y_{33} & y_{34} \\ \hline y_{41} & y_{42} & y_{43} & y_{44} \\ y_{51} & y_{52} & y_{53} & y_{54} \end{array}\right] = \left[\begin{array}{cc} \mathbf{Y}_{11} & \mathbf{Y}_{12} \\ \mathbf{Y}_{21} & \mathbf{Y}_{22} \end{array}\right]$$

K. Adachi, *Matrix-Based Introduction to Multivariate Data Analysis*,
https://doi.org/10.1007/978-981-15-4103-2_14

where

$$\mathbf{Y}_{11} = \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \\ y_{31} & y_{32} \end{bmatrix}, \quad \mathbf{Y}_{12} = \begin{bmatrix} y_{13} & y_{14} \\ y_{23} & y_{24} \\ y_{33} & y_{34} \end{bmatrix}, \quad \mathbf{Y}_{21} = \begin{bmatrix} y_{41} & y_{42} \\ y_{51} & y_{52} \end{bmatrix}, \quad \mathbf{Y}_{22} = \begin{bmatrix} y_{43} & y_{44} \\ y_{53} & y_{54} \end{bmatrix}.$$

$\mathbf{Y}_{11}$, $\mathbf{Y}_{12}$, $\mathbf{Y}_{21}$, and $\mathbf{Y}_{22}$ are called the *blocks* of $\mathbf{Y}$, while $\mathbf{Y}$ is called a *block matrix* consisting of $\mathbf{Y}_{11}$, $\mathbf{Y}_{12}$, $\mathbf{Y}_{21}$, and $\mathbf{Y}_{22}$.

This example is generalized as follows: an $n \times p$ matrix $\mathbf{Y}$ can be rewritten as

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_{11} & \cdots & \mathbf{Y}_{1j} & \cdots & \mathbf{Y}_{1J} \\ & & \vdots & & \\ \mathbf{Y}_{i1} & \cdots & \mathbf{Y}_{ij} & \cdots & \mathbf{Y}_{iJ} \\ & & \vdots & & \\ \mathbf{Y}_{I1} & \cdots & \mathbf{Y}_{Ij} & \cdots & \mathbf{Y}_{IJ} \end{bmatrix}. \tag{14.1}$$

Here, $\mathbf{Y}_{ij}$ is called the (i, j) *block* of $\mathbf{Y}$, while $\mathbf{Y}$ is called a *block matrix* containing $\mathbf{Y}_{ij}$ ($i = 1, \ldots, I$; $j = 1, \ldots, J$). If $\mathbf{Y}_{ij}$ is $n_i \times p_j$, then $n = \sum_{i=1}^{I} n_i$ and $p = \sum_{j=1}^{J} p_j$.

In this chapter, a block matrix of data is treated in which blocks $\mathbf{X}_1, \ldots, \mathbf{X}_J$ are arranged horizontally:

$$\mathbf{X} = \left[\mathbf{X}_1, \ldots, \mathbf{X}_j, \ldots, \mathbf{X}_J\right], \tag{14.2}$$

while a block matrix of parameters is considered in which $\mathbf{C}_1, \ldots, \mathbf{C}_J$ are stacked vertically:

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_1 \\ \vdots \\ \mathbf{C}_j \\ \vdots \\ \mathbf{C}_J \end{bmatrix}. \tag{14.3}$$

Here, $\mathbf{X}_j$ and $\mathbf{C}_j$ are called the jth block of $\mathbf{X}$ and $\mathbf{C}$, respectively.

A weighted sum of matrices can be expressed block-wise as follows:

Note 14.2. Weighted Sum of Block Matrices

Let the block matrices $\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \cdots & \mathbf{A}_{1J} \\ & \vdots & \\ \mathbf{A}_{I1} & \cdots & \mathbf{A}_{IJ} \end{bmatrix}$ and $\mathbf{B} = \begin{bmatrix} \mathbf{B}_{11} & \cdots & \mathbf{B}_{1J} \\ & \vdots & \\ \mathbf{B}_{I1} & \cdots & \mathbf{B}_{IJ} \end{bmatrix}$ be of the same order and their blocks $\mathbf{A}_{ij}$ and $\mathbf{B}_{ij}$ ($i = 1, \ldots, I; j = 1, \ldots, J$) also be so. Then, the sum of $\mathbf{A}$ and $\mathbf{B}$ multiplied by scalars s and t is defined as

$$s\mathbf{A} + t\mathbf{B} = \begin{bmatrix} s\mathbf{A}_{11} + t\mathbf{B}_{11} & \cdots & s\mathbf{A}_{1J} + t\mathbf{B}_{1J} \\ & \vdots & \\ s\mathbf{A}_{I1} + t\mathbf{B}_{I1} & \cdots & s\mathbf{A}_{IJ} + t\mathbf{B}_{IJ} \end{bmatrix}, \tag{14.4}$$

whose (i, j) block is $s\mathbf{A}_{ij} + t\mathbf{B}_{ij}$.

The product of the matrices can also be expressed block-wise:

Note 14.3. Product of Block Matrices

Let us define $n \times p$ and $p \times m$ block matrices as $\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \cdots & \mathbf{A}_{1J} \\ & \vdots & \\ \mathbf{A}_{I1} & \cdots & \mathbf{A}_{IJ} \end{bmatrix}$ and $\mathbf{Q} = \begin{bmatrix} \mathbf{Q}_{11} & \cdots & \mathbf{Q}_{1K} \\ & \vdots & \\ \mathbf{Q}_{J1} & \cdots & \mathbf{Q}_{JK} \end{bmatrix}$, respectively, with $\mathbf{A}_{ij}$ being the (i, j) block of $\mathbf{A}$, $\mathbf{Q}_{jk}$ the (j, k) one of $\mathbf{Q}$ and the number of the columns of $\mathbf{A}_{ij}$ equaling the number of rows of $\mathbf{Q}_{jk}$. Post-multiplication of $\mathbf{A}$ by $\mathbf{Q}$ provides the $n \times m$ matrix

$$\mathbf{V} = \mathbf{AQ} = \begin{bmatrix} \mathbf{V}_{11} & \cdots & \mathbf{V}_{1K} \\ & \vdots & \\ \mathbf{V}_{I1} & \cdots & \mathbf{V}_{IK} \end{bmatrix}, \tag{14.5}$$

whose (i, k) block is

$$\mathbf{V}_{ik} = \sum_{j=1}^{J} \mathbf{A}_{ij}\mathbf{Q}_{jk} = \mathbf{A}_{i1}\mathbf{Q}_{1k} + \mathbf{A}_{i2}\mathbf{Q}_{2k} + \cdots + \mathbf{A}_{iJ}\mathbf{Q}_{Jk}. \tag{14.6}$$

In this chapter, the special case of (14.5),

$$\mathbf{XC} = \sum_{j=1}^{J} \mathbf{X}_j\mathbf{C}_j = \mathbf{X}_1\mathbf{C}_1 + \mathbf{X}_2\mathbf{C}_2 + \cdots + \mathbf{X}_J\mathbf{C}_J, \tag{14.7}$$

is often used with $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_J]$ and $\mathbf{C} = \begin{bmatrix} \mathbf{C}_1 \\ \vdots \\ \mathbf{C}_J \end{bmatrix}$.

14.2 Canonical Correlation Analysis

Let us consider an n-individuals $\times$ p-variables data matrix $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$ consisting of the *two blocks* $\mathbf{X}_1 = \left[\mathbf{x}_{11}, \ldots, \mathbf{x}_{1p_1}\right](n \times p_1)$ and $\mathbf{X}_2 = \left[\mathbf{x}_{21}, \ldots, \mathbf{x}_{2p_2}\right](n \times p_2)$. That is, the p variables in $\mathbf{X}$ are classified into a group of p_1 variables and into a group of p_2 variables. We suppose that $\mathbf{X}$ is centered with $\mathbf{1}_n'\mathbf{X} = \mathbf{0}_p'$. An example of such data is presented in Table 14.1.

For $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$, *canonical correlation analysis* (*CCA*) is formulated as minimizing

$$f(\mathbf{C}_1, \mathbf{C}_2) = \|\mathbf{X}_1\mathbf{C}_1 - \mathbf{X}_2\mathbf{C}_2\|^2 \tag{14.8}$$

over $p_1 \times m$ coefficient matrix $\mathbf{C}_1$ and $p_2 \times m$ coefficient matrix $\mathbf{C}_2$ subject to the constraints

$$\frac{1}{n}\mathbf{C}_1'\mathbf{X}_1'\mathbf{X}_1\mathbf{C}_1 = \frac{1}{n}\mathbf{C}_2'\mathbf{X}_2'\mathbf{X}_2\mathbf{C}_2 = \mathbf{I}_m, \tag{14.9}$$

Table 14.1 Standard scores for strength and athletic test data (Tanaka & Tarumi, 1995)

Ind.	$\mathbf{X}_1$: strength test*							$\mathbf{X}_2$: athletic test*				
	RJ	VJ	DM	GP	SM	DB	BW	SP	LJ	LT	CE	MA
1	−0.42	−0.68	0.74	1.19	0.56	1.62	1.51	−0.96	1.13	−0.30	0.11	−0.22
2	1.36	−0.68	−1.24	−0.49	0.99	0.45	−1.20	0.18	0.54	0.79	−0.68	−0.60
3	−0.42	1.35	−0.48	−1.24	2.15	0.45	1.75	−0.96	−0.26	1.52	0.38	0.62
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
37	1.36	0.91	0.99	0.44	−1.45	−0.23	−1.81	−0.96	1.51	0.07	0.90	−0.38
38	0.17	1.20	−0.92	0.07	−0.92	1.29	1.26	0.18	1.91	0.07	0.38	−0.47

*Variable names are abbreviated as follows: *RJ* repetition of jump, *VJ* vertical jump, *DM* dorsal muscles, *GP* grasping power, *SM* step motion, *DB* deep forward bow, *BW* body warping, *SP* sprint, *LJ* long jump, *LT* long throw, *CE* chinning exercises, *MA* marathon

with $m \leq \mathrm{rank}\left(\mathbf{X}_1'\mathbf{X}_2\right)$. That is, the purpose of CCA is to obtain the coefficient matrices $\mathbf{C}_1$ and $\mathbf{C}_2$ that allow $\mathbf{X}_1\mathbf{C}_1$ and $\mathbf{X}_2\mathbf{C}_2$ to be mutually best *matched*. Loss function (14.8) can be rewritten using (14.9) as $\mu = \mathrm{tr}\mathbf{C}_1'\mathbf{X}_1'\mathbf{X}_1\mathbf{C}_1 + \mathrm{tr}\mathbf{C}_2'\mathbf{X}_2'\mathbf{X}_2\mathbf{C}_2 - 2\mathrm{tr}\mathbf{C}_1'\mathbf{X}_1'\mathbf{X}_1\mathbf{C}_1 = 2m - 2\mathrm{tr}\mathbf{C}_1'\mathbf{X}_1'\mathbf{X}_2\mathbf{C}_2$, whose minimization is equivalent to maximizing

$$\frac{1}{n}\mathrm{tr}\mathbf{C}_1'\mathbf{X}_1'\mathbf{X}_2\mathbf{C}_2. \tag{14.10}$$

This maximization subject to (14.9) is attained as in Theorem A.4.8 (Appendix A.4.5; where we can set $\mathbf{V}_{11} = n^{-1}\mathbf{X}_1'\mathbf{X}_1$, $\mathbf{V}_{22} = n^{-1}\mathbf{X}_2'\mathbf{X}_2$, and $\mathbf{V}_{12} = n^{-1}\mathbf{X}_1'\mathbf{X}_2$ to find the solution for the above CCA problem).

We illustrate CCA by performing it to the data set in Table 14.1, setting $m = 1$. In this unidimensional case, $\mathbf{C}_1$ and $\mathbf{C}_2$ are simplified as vectors $\mathbf{c}_1 = [c_{11}, \ldots, c_{1p_1}]'$ and $\mathbf{c}_2 = [c_{21}, \ldots, c_{2p_2}]'$, respectively; $\mathbf{X}_1\mathbf{C}_1$ and $\mathbf{X}_2\mathbf{C}_2$ are expressed as $\mathbf{X}_1\mathbf{c}_1 = c_{11}\mathbf{x}_{11} + \cdots + c_{1p_1}\mathbf{x}_{1p_1}$ and $\mathbf{X}_2\mathbf{c}_2 = c_{21}\mathbf{x}_{21} + \cdots + c_{2p_2}\mathbf{x}_{2p_2}$, respectively. The CCA for the data set gives the following solution:

$$\begin{aligned}\mathbf{X}_1\mathbf{c}_1 = {} & 0.442 \times \mathrm{RJ} + 0.267 \times \mathrm{VJ} + 0.588 \times \mathrm{DM} \\ & + 0.061 \times \mathrm{GP} + 0.222 \times \mathrm{SM} + 0.091 \times \mathrm{DB} + 0.014 \times \mathrm{BW},\end{aligned} \tag{14.11}$$

$$\mathbf{X}_2\mathbf{c}_2 = -0.426 \times \mathrm{SP} + 0.233 \times \mathrm{L} + 0.370 \times \mathrm{LT} + 0.004 \times \mathrm{CE} - 0.356 \times \mathrm{MA}, \tag{14.12}$$

where the resulting coefficient for each variable is followed by the abbreviation of its name in Table 14.1. The solutions in (14.11) and (14.12) stand for the weighted sums of strength and athletic test scores that are best matched.

Since $\mathbf{1}_n'\mathbf{X} = \mathbf{0}_p'$, the correlation coefficient between $\mathbf{X}_1\mathbf{c}_1$ and $\mathbf{X}_2\mathbf{c}_2$ is expressed as

$$\frac{n^{-1}\mathbf{c}_1\mathbf{X}'_1\mathbf{X}_2\mathbf{c}_2}{\sqrt{n^{-1}\mathbf{c}_1\mathbf{X}'_1\mathbf{X}_1\mathbf{c}_1}\sqrt{n^{-1}\mathbf{c}_2\mathbf{X}'_2\mathbf{X}_2\mathbf{c}_2}}, \tag{14.13}$$

whose denominator equals one because of (14.9): (14.10) with $m = 1$ is equivalent to (14.13). This particular coefficient is called a *canonical correlation coefficient* between the variables in $\mathbf{X}_1$ and those in $\mathbf{X}_2$. The CCA solution for the data set in

Table 14.1 gives the (14.13) value equaling 0.85, which shows that the items in the strength test are strongly related to those in the athletic test.

14.3 Generalized Canonical Correlation Analysis

Let us compare the CCA loss function (14.8) and the function

$$f(\mathbf{F}, \mathbf{C}_1, \mathbf{C}_2) = \|\mathbf{F} - \mathbf{X}_1\mathbf{C}_1\|^2 + \|\mathbf{F} - \mathbf{X}_2\mathbf{C}_2\|^2 \tag{14.14}$$

with a new matrix $\mathbf{F}$ ($n \times m$) whose rows correspond to individuals. The minimization of (14.8) is equivalent to minimizing (14.14) over $\mathbf{F}$, $\mathbf{C}_1$, and $\mathbf{C}_2$. It follows from the fact that the solution of $\mathbf{F}$ must satisfy $\mathbf{F} = 2^{-1}(\mathbf{X}_1\mathbf{C}_1 + \mathbf{X}_2\mathbf{C}_2)$, as shown with (A.2.6) in Appendix A.2.1. Substituting the equation for $\mathbf{F}$ in (14.14), it is rewritten as

$$\begin{aligned} f(\mathbf{F}, \mathbf{C}_1, \mathbf{C}_2) &= \left\| \frac{1}{2}(\mathbf{X}_1\mathbf{C}_1 + \mathbf{X}_2\mathbf{C}_2) - \mathbf{X}_1\mathbf{C}_1 \right\|^2 + \left\| \frac{1}{2}(\mathbf{X}_1\mathbf{C}_1 + \mathbf{X}_2\mathbf{C}_2) - \mathbf{X}_2\mathbf{C}_2 \right\|^2 \\ &= \left\| \frac{1}{2}\mathbf{X}_2\mathbf{C}_2 - \frac{1}{2}\mathbf{X}_1\mathbf{C}_1 \right\|^2 + \left\| \frac{1}{2}\mathbf{X}_1\mathbf{C}_1 - \frac{1}{2}\mathbf{X}_2\mathbf{C}_2 \right\|^2 \\ &= \frac{1}{2}\|\mathbf{X}_1\mathbf{C}_1 - \mathbf{X}_2\mathbf{C}_2\|^2, \end{aligned} \tag{14.15}$$

which equals half of (14.8).

Generalized canonical correlation analysis (*GCCA*) can be formulated through the extension of (14.14) to the cases when the $n \times p$ data matrix $\mathbf{X}$ is expressed as $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_j, \ldots, \mathbf{X}_J]$ with $J \geq 2$. Here, $\mathbf{X}_j$ is $n \times p_j$ and $p = \sum_{j=1}^{J} p_j$. For the data set $\mathbf{X}$, the loss function of GCCA is defined as

$$\eta(\mathbf{F}, \mathbf{C}) = \sum_{j=1}^{J} \left\|\mathbf{F} - \mathbf{X}_j\mathbf{C}_j\right\|^2, \tag{14.16}$$

with $\mathbf{C} = \begin{bmatrix} \mathbf{C}_1 \\ \vdots \\ \mathbf{C}_J \end{bmatrix}$. In GCCA, $\mathbf{F}$ rather than $\mathbf{X}_j\mathbf{C}_j$ is constrained as

$$\frac{1}{n}\mathbf{F}'\mathbf{F} = \mathbf{I}_m, \tag{14.17}$$

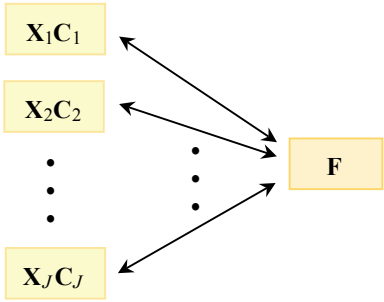


Fig. 14.1 Illustration of generalized canonical correlation analysis

with $m \leq r = \text{rank}(\mathbf{X})$. That is, GCCA can be formulated as minimizing (14.16) over $\mathbf{F}$ and $\mathbf{C}$ subject to (14.17). The implication of this minimization is illustrated in Fig. 14.1, where a *single* $\mathbf{F}$ and *multiple* $\mathbf{X}_j\mathbf{C}_j$ ($j = 1, \ldots, J$) are depicted. The double-headed arrows in the figure express the deviations of $\mathbf{X}_j\mathbf{C}_j$ from $\mathbf{F}$. The deviations are expressed as squared differences and summated as in (14.16), which is minimized so that $\mathbf{X}_j\mathbf{C}_j$ are well *matched* with $\mathbf{F}$. As a result, $\mathbf{X}_j\mathbf{C}_j$ ($j = 1, \ldots, J$) becomes *similar* across different j, and $\mathbf{X}_j\mathbf{C}_j$ is *summarized* into a single matrix $\mathbf{F}$.

As explained later, the matrix $\mathbf{X}\mathbf{D}_x^{-1/2}$ plays an important role in GCCA with

$$\mathbf{D}_X = \begin{bmatrix} \mathbf{X}'_1\mathbf{X}_1 & & & \\ & \mathbf{X}'_2\mathbf{X}_2 & & \\ & & \ddots & \\ & & & \mathbf{X}'_J\mathbf{X}_J \end{bmatrix} \tag{14.18}$$

a $p \times p$ *block diagonal matrix* in which the blank cells are filled with zeros. We explain the term block diagonal matrix and the superscript $-1/2$ in $\mathbf{D}_x^{-1/2}$ in the following two notes.

Note 14.4. Block Diagonal Matrices
A matrix $\mathbf{B}$ whose (i, j) block is a zero matrix for $i \neq j$, i.e.,

$$\mathbf{B} = \begin{bmatrix} \mathbf{B}_1 & & & \\ & \mathbf{B}_2 & & \\ & & \ddots & \\ & & & \mathbf{B}_I \end{bmatrix} \tag{14.19}$$

is called a *block diagonal matrix* and $\mathbf{B}_i$ ($i = 1, \ldots, I$) is called the ith *diagonal block* of $\mathbf{B}$.

The products of block matrices are given as

$$\begin{bmatrix} \mathbf{B}_1 & & & \\ & \mathbf{B}_2 & & \\ & & \ddots & \\ & & & \mathbf{B}_I \end{bmatrix} \begin{bmatrix} \mathbf{H}_1 & & & \\ & \mathbf{H}_2 & & \\ & & \ddots & \\ & & & \mathbf{H}_I \end{bmatrix} = \begin{bmatrix} \mathbf{B}_1\mathbf{H}_1 & & & \\ & \mathbf{B}_2\mathbf{H}_2 & & \\ & & \ddots & \\ & & & \mathbf{B}_I\mathbf{H}_I \end{bmatrix}, \tag{14.20}$$

$$[\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_J] \begin{bmatrix} \mathbf{C}_1 & & & \\ & \mathbf{C}_2 & & \\ & & \ddots & \\ & & & \mathbf{C}_J \end{bmatrix} = [\mathbf{X}_1\mathbf{C}_1, \mathbf{X}_2\mathbf{C}_j, \ldots, \mathbf{X}_J\mathbf{C}_J], \tag{14.21}$$

$$\begin{bmatrix} \mathbf{B}_1 & & & \\ & \mathbf{B}_2 & & \\ & & \ddots & \\ & & & \mathbf{B}_I \end{bmatrix} \begin{bmatrix} \mathbf{Q}_1 \\ \mathbf{Q}_2 \\ \vdots \\ \mathbf{Q}_I \end{bmatrix} = \begin{bmatrix} \mathbf{B}_1\mathbf{Q}_1 \\ \mathbf{B}_2\mathbf{Q}_2 \\ \vdots \\ \mathbf{B}_I\mathbf{Q}_I \end{bmatrix}. \tag{14.22}$$

Here, we have supposed that the products of the blocks are defined. If (14.19) and $\mathbf{B}_1$, …, $\mathbf{B}_I$ are square and nonsingular, the inverse matrix of (14.19) is expressed as

$$\mathbf{B}^{-1} \begin{bmatrix} \mathbf{B}_1^{-1} & & & \\ & \mathbf{B}_2^{-1} & & \\ & & \ddots & \\ & & & \mathbf{B}_I^{-1} \end{bmatrix}. \tag{14.23}$$

Note 14.5. Square and Square Root of a Matrix
The *square* of an $n \times n$ matrix $\mathbf{V}$ is expressed as

$$\mathbf{V}^2 = \mathbf{V}\mathbf{V}. \tag{14.24}$$

The *square root* of $\mathbf{V}$, denoted as $\mathbf{V}^{1/2}$, is the matrix satisfying

$$\mathbf{V}^{1/2}\mathbf{V}^{1/2} = \mathbf{V} \tag{14.25}$$

and the inverse matrix of $\mathbf{V}^{1/2}$, denoted as $\mathbf{V}^{-1/2}$, satisfies

$$\mathbf{V}^{-1/2}\mathbf{V}^{-1/2} = \mathbf{V}^{-1}. \tag{14.26}$$

Thus, $\mathbf{D}_{\mathrm{X}}^{-1/2}$ is the matrix satisfying $\mathbf{D}_{\mathrm{X}}^{-1/2}\mathbf{D}_{\mathrm{X}}^{-1/2} = \mathbf{D}_{\mathrm{X}}^{-1}$. Comparing (14.18) with (14.23) and (14.26), we find

$$\mathbf{D}_{\mathrm{X}}^{-1/2}\begin{bmatrix} (\mathbf{X}'_1\mathbf{X}_1)^{-1/2} & & & \\ & (\mathbf{X}'_2\mathbf{X}_2)^{-1/2} & & \\ & & \ddots & \\ & & & (\mathbf{X}'_J\mathbf{X}_J)^{-1/2} \end{bmatrix}, \tag{14.27}$$

and use (14.21) to get

$$\mathbf{X}\mathbf{D}_{\mathrm{X}}^{-1/2} = \left[\mathbf{X}_1(\mathbf{X}_1\mathbf{X}_1)^{-1/2}, \ldots, \mathbf{X}_J(\mathbf{X}_J\mathbf{X}_J)^{-1/2}\right]. \tag{14.28}$$

How to obtain $\left(\mathbf{X}'_j\mathbf{X}_j\right)^{-1/2}$ is described in Appendix A.4.6.

As described in Theorem A.4.6 (Appendix A.4.4), the GCCA problem, i.e., the minimization of (14.16) subject to (14.17), is equivalent to minimizing

$$f(\mathbf{F}, \mathbf{C}) = \left\| \mathbf{X}\mathbf{D}_{\mathrm{X}}^{-1/2} - \frac{1}{n}\mathbf{F}\mathbf{C}'\mathbf{D}_{\mathrm{X}}^{1/2} \right\|^2 \tag{14.29}$$

over $\mathbf{F}$ and $\mathbf{C}$ subject to (14.17), which can be viewed as the reduced rank approximation of $\mathbf{X}\mathbf{D}_{\mathrm{X}}^{-1/2}$ with $\mathrm{rank}(\mathbf{X}\mathbf{D}_{\mathrm{X}}^{-1/2}) = r$ as explained in Appendix A.4.4. The solution of $\mathbf{F}$ and $\mathbf{C}$ is given by

$$\mathbf{F} = \sqrt{n}\mathbf{N}_m\mathbf{T}, \tag{14.30}$$

$$\mathbf{C} = \sqrt{n}\mathbf{D}_{\mathrm{X}}^{-1/2}\mathbf{M}_m\mathbf{\Phi}_m\mathbf{T}, \tag{14.31}$$

as found in Theorem A.4.6. Here, $\mathbf{T}$ is an $m \times m$ orthonormal matrix, and $\mathbf{N}_m$, $\mathbf{M}_m$, and $\mathbf{\Phi}_m$ are obtained through the singular value decomposition (SVD) of $\mathbf{X}\mathbf{D}_{\mathrm{X}}^{-1/2}$ defined as

$$\mathbf{X}\mathbf{D}_{\mathrm{X}}^{-1/2} = \mathbf{N}\boldsymbol{\Phi}\mathbf{M}', \tag{14.32}$$

with $\mathbf{N}'\mathbf{N} = \mathbf{M}'\mathbf{M} = \mathbf{I}_r$ and $\boldsymbol{\Phi}$ a diagonal matrix whose diagonal elements are ordered in descending order; $\mathbf{N}_m$ and $\mathbf{M}_m$ contain the first m columns of $\mathbf{N}$ and those of $\mathbf{M}$, respectively, with $\boldsymbol{\Phi}_m$ the first $m \times m$ diagonal block. The matrix $\mathbf{T}$ appearing in (14.30) and (14.31) implies that the solution can be rotated as in EFA.

The importance of GCCA may not be its usefulness in real data analysis, but rather that it leads to multiple correspondence analysis for the categorical data described in the next sections.

14.4 Multivariate Categorical Data

An example of *multivariate categorical data* is given by a 10-individuals $\times$ 3-variables matrix $\mathbf{Y} = \left(y_{ij}\right)$ in Table 14.2a, where the variables are

[V1] Faculty to which each individual belongs,
[V2] Subject at which she/he is best,
[V3] Sciences, basic or applied, to which she/he is oriented.

We should note that the elements of $\mathbf{Y}$ are *not quantitative* scores, but the *code* numbers referring to *categories*. For example, those for [V1] are coded as 1 = Sciences, 2 = Medicine, 3 = Engineering. In Table 14.2b, the elements of $\mathbf{Y}$ are presented as category names.

Table 14.2 Artificial example describing the faculties (FC) of students (ST), the subjects (SJ) at which they are best, and their orientation (OT), which is found in Adachi and Murakami (2011)

ST	(a) Data matrix $\mathbf{Y}$			(b) Data matrix $\mathbf{Y}$			(c) Indicator matrix $\mathbf{G} = [\mathbf{G}_1, \mathbf{G}_2, \mathbf{G}_3]$								
	Code number			Category*			$\mathbf{G}_1$ (FC)			$\mathbf{G}_2$ (SJ)				$\mathbf{G}_3$ (OT)	
	FC	SJ	OT	FC	SJ	OT	Sci	Med	Eng	Math	Bio	Phy	Chemo	Bs	Ap
1	3	4	2	Eng	Che	Ap	0	0	1	0	0	0	1	0	1
2	1	2	1	Sci	Bio	Bs	1	0	0	0	1	0	0	1	0
3	2	3	2	Med	Phy	Ap	0	1	0	0	0	1	0	0	1
4	1	1	1	Sci	Mat	Bs	1	0	0	1	0	0	0	1	0
5	2	2	1	Med	Bio	Bs	0	1	0	0	1	0	0	1	0
6	3	3	2	Eng	Phy	Ap	0	0	1	0	0	1	0	0	1
7	2	2	2	Med	Bio	Ap	0	1	0	0	1	0	0	0	1
8	1	3	1	Sci	Phy	Bs	1	0	0	0	0	1	0	1	0
9	2	4	2	Med	Che	Ap	0	1	0	0	0	0	1	0	1
10	3	1	1	Eng	Mat	Bs	0	0	1	1	0	0	0	1	0

*The names of categories are abbreviated as follows: *Eng* engineering, *Sci* sciences, *Med* medicine; *Che* chemistry, *Bio* biology, *Phy* physics, *Mat* mathematics; *Ap* applications, *Bs* basis

Each column of the data matrix in (a) or (b) can also be expressed as the *n-individuals* $\times$ K_j*-categories indicator* matrices

$$\mathbf{G}_j = \begin{bmatrix} \mathbf{g}_{1j}' \\ \vdots \\ \mathbf{g}_{ij}' \\ \vdots \\ \mathbf{g}_{nj}' \end{bmatrix} \quad (j = 1, 2, 3), \tag{14.33}$$

as in Table 14.2c. Here, the jth variable in (a) or (b) corresponds to $\mathbf{G}_j$, and the kth element g_{ijk} in the ith row $\mathbf{g}_{ij}'$ is defined as

$$g_{ijk} = \begin{cases} 1 & \text{if } k = y_{ij} \\ 0 & \text{otherwise} \end{cases}. \tag{14.34}$$

For example, $\mathbf{g}_{82}' = [0, 0, 1, 0]$, since $y_{82} = 3$: individual 8 shows 3 (=Physics) for variable 2. The indicator matrix $\mathbf{G}_j$ in (14.33) can also be called a *membership* matrix, as described in Sect. 7.1, as $\mathbf{G}_j$ stands for the membership of individuals to categories.

Let the number of columns of $\mathbf{G}_j$ be K_j, $j = 1, \ldots, J$, and $K = \sum_{j=1}^{J} K_j$. We further define an $n \times K$ block matrix as

$$\mathbf{G} = \left[\mathbf{G}_1, \ldots, \mathbf{G}_j, \ldots, \mathbf{G}_J\right]. \tag{14.35}$$

In the next sections, we refer to $\mathbf{G}$ rather than $\mathbf{G}_j$ as an *indicator* matrix.

14.5 Multiple Correspondence Analysis

The loss function for *multiple correspondence analysis* (*MCA*) is given by replacing $\mathbf{X}_j$ by $\mathbf{G}_j$ in the GCCA function (14.16). That is, MCA is formulated as minimizing

$$\eta(\mathbf{F}, \mathbf{C}) = \sum_{j=1}^{J} \left\|\mathbf{F} - \mathbf{G}_j\mathbf{C}_j\right\|^2 \tag{14.36}$$

subject to (14.36) and an additional constraint,

$$\mathbf{1}_n'\mathbf{F} = \mathbf{0}_m', \text{ or equivalently, } \mathbf{F} = \mathbf{JF}, \tag{14.37}$$

with $m \leq \text{rank}(\mathbf{JG})$. The equivalence in (14.37) has been proved in Note 3.1 (Chap. 3). The *K_j-categories* $\times$ *m-dimensions* matrix $\mathbf{C}_j$ to be obtained is called a *category score* matrix, as its kth row stands for the vector of scores which is suitable to be given to category k, as explained in the next section. There, we also explain why we refer to the columns of $\mathbf{C}_j$ as *dimensions*. For the same reason, an *n*-individuals $\times$ *m*-dimensions matrix $\mathbf{F}$ is called an *individual score* matrix. Why constraint (14.37) is added is explained next:

Note 14.6. Avoiding Trivial Solutions
Let $m = 1$ for the sake of simplicity. Then, $\mathbf{F}$ and $\mathbf{C}_j$ in (14.36) are the column vectors. Without (14.37), the MCA loss function (14.36) would attain the lower limit zero for

$$\mathbf{F} = \mathbf{1}_n \text{ and } \mathbf{C}_j = \mathbf{1}_{K_j}, \tag{14.38}$$

because (14.34) implies $\mathbf{G}_j\mathbf{1}_{K_j} = \mathbf{1}_n$. The solution in (14.38) is *trivial*, since it implies that the same score of "one" is given to all individuals and categories. This trivial solution does not satisfy (14.37); by adding it, the trivial one can be excluded from the solution.

As the minimization of (14.16) is equivalent to that of (14.29) in GCCA, the MCA problem, i.e., the minimization of (14.36) subject to (14.17) and (14.37), is equivalent to minimizing

$$h(\mathbf{F}, \mathbf{C}) = \left\| \mathbf{JGD}_{\mathrm{G}}^{-1/2} - \frac{1}{n}\mathbf{FC}'\mathbf{D}_{\mathrm{G}}^{1/2} \right\|^2 \tag{14.39}$$

over $\mathbf{F}$ and $\mathbf{C}$ under the same constraints, which is detailed in Theorem A.4.7 (Appendix A.4.4). Further, the theorem shows that the MCA solution is given by

$$\mathbf{F} = \sqrt{n}\mathbf{S}_m\mathbf{T}, \tag{14.40}$$

$$\mathbf{C} = \sqrt{n}\mathbf{D}_{\mathrm{G}}^{-1/2}\mathbf{P}_m\mathbf{\Theta}_m\mathbf{T}. \tag{14.41}$$

Here,

$$\mathbf{D}_{\mathrm{G}} = \begin{bmatrix} \mathbf{G}_1'\mathbf{G}_1 & & & \\ & \mathbf{G}_2'\mathbf{G}_2 & & \\ & & \ddots & \\ & & & \mathbf{G}_J'\mathbf{G}_J \end{bmatrix} \tag{14.42}$$

is the matrix in (14.18) with $\mathbf{X}_j$ replaced by $\mathbf{G}_j$, $\mathbf{T}$ is an $m \times m$ orthonormal matrix, and $\mathbf{S}_m$, $\mathbf{P}_m$, and $\mathbf{\Theta}_m$ are obtained through the SVD of $\mathbf{JGD}_{\mathrm{G}}^{-1/2}$ defined as

$$\mathbf{JGD}_{\mathrm{G}}^{-1/2} = \mathbf{S\Theta P}'. \tag{14.43}$$

Here, $\mathbf{S}'\mathbf{S} = \mathbf{P}'\mathbf{P} = \mathbf{I}_q$ with q = rank($\mathbf{JG}$) and $\mathbf{\Theta}$ is a diagonal matrix whose diagonal elements are arranged in descending order. That is, $\mathbf{S}_m$ and $\mathbf{P}_m$ contain the first m columns of $\mathbf{S}$ and $\mathbf{P}$, respectively, with $\mathbf{\Theta}_m$ the first $m \times m$ diagonal block of $\mathbf{\Theta}$. In this chapter, we do not use a rotation technique by setting $\mathbf{T}$ in (14.40) and (14.41) at $\mathbf{I}_m$, as explained with (A.4.33) in Appendix A.4.4.

We must mention that the block diagonal matrix $\mathbf{D}_{\mathrm{G}}^{-1/2}$ in (14.43) is simply a *diagonal* one. This can be verified by the fact that the $\mathbf{G}_1$ in Table 14.2c implies $\mathbf{G}_1'\mathbf{G}_1 = \begin{bmatrix} 3 & & \\ & 4 & \\ & & 3 \end{bmatrix}$. Thus, $(\mathbf{G}_1'\mathbf{G}_1)^{-1/2} = \begin{bmatrix} 1/\sqrt{3} & & \\ & 1/\sqrt{4} & \\ & & 1/\sqrt{3} \end{bmatrix}$. In general, $\mathbf{G}_j'\mathbf{G}_j$ and $(\mathbf{G}_j'\mathbf{G}_j)^{-1/2} (j = 1, \ldots, J)$ are diagonal matrices, which implies that $\mathbf{D}_{\mathrm{G}}$ and $\mathbf{D}_{\mathrm{G}}^{-1/2}$ are also diagonal.

14.6 Homogeneity Assumption

Table 14.3 presents the MCA solution of $\mathbf{F}$ and $\mathbf{C} = \left[\mathbf{C}_1', \ldots, \mathbf{C}_J'\right]'$ for the data set in Table 14.2 with m = 2. In Table 14.3, $\mathbf{f}_i'$ denotes the ith row of $\mathbf{F}$, which corresponds to the ith individual in Table 14.2, and $\mathbf{c}_{jk}'$ denotes the kth row of $\mathbf{C}_j$, which is associated with category k in variable j; for example, $\mathbf{c}_{23}'$ contains the scores for Phy (physics). The solution in Table 14.3 can be graphically represented

Table 14.3 MCA solution for the data in Table 14.2

F			C				
$\mathbf{f}_1'$	1.20	1.20	$\mathbf{C}_1$	$\mathbf{c}_{11}'$	Sci	−1.19	−0.10
$\mathbf{f}_2'$	−1.12	−0.94		$\mathbf{c}_{12}'$	Med	0.63	−0.84
$\mathbf{f}_3'$	0.83	−0.38		$\mathbf{c}_{13}'$	Eng	0.36	1.23
$\mathbf{f}_4'$	−1.56	0.59	$\mathbf{C}_2$	$\mathbf{c}_{21}'$	Math	−1.19	1.03
$\mathbf{f}_5'$	−0.27	−1.44		$\mathbf{c}_{22}'$	Bio	−0.26	−1.25
$\mathbf{f}_6'$	0.71	1.01		$\mathbf{c}_{23}'$	Phy	0.21	0.23
$\mathbf{f}_7'$	0.61	−1.37		$\mathbf{c}_{24}'$	Che	1.26	0.50
$\mathbf{f}_8'$	−0.90	0.05	$\mathbf{C}_3$	$\mathbf{c}_{31}'$	Bs	−0.93	−0.05
$\mathbf{f}_9'$	1.32	−0.19		$\mathbf{c}_{32}'$	Ap	0.93	0.05
$\mathbf{f}_{10}'$	−0.83	1.48					

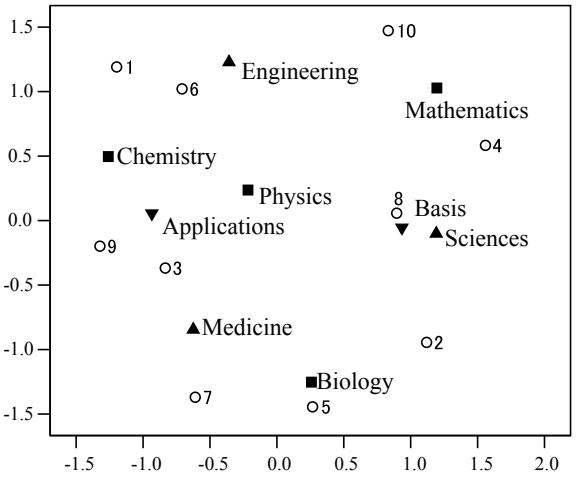


Fig. 14.2 Scatterplot of categories and individuals

as in Fig. 14.2, where individual i (= 1, …, n) is plotted as the point with its coordinate $\mathbf{f}_i$, and category k in variable j is plotted with its coordinate $\mathbf{c}_{jk}$. We can interpret the plot by noting inter-point *distances*. The rationale for this distance-based interpretation of MCA solutions is explained in the following paragraph.

MCA can be reformulated with the *homogeneity assumption*:

$$\begin{aligned}&\text{the scores for an } \textit{individual} \text{ should be } \textit{homogeneous to}\\ &\underline{\text{the scores for the } \textit{categories} \text{ to which the } \textit{individual belongs}.}\end{aligned} \tag{14.44}$$

Here, the underlined scores are expressed as the vector $\mathbf{c}_{jy_{ij}}$, which is the *category score* vector $\mathbf{c}'_{jk}$ with k set to the category y_{ij} (the *category number* that individual i shows for variable j). Assumption (14.44) requires $\left\|\mathbf{f}'_i - \mathbf{c}'_{jy_{ij}}\right\|^2$ to be small, and its sum over i and j can be expressed as

$$\sum_{j=1}^{J}\sum_{i=1}^{n}\left\|\mathbf{f}'_i - \mathbf{c}'_{jy_{ij}}\right\|^2 = \sum_{j=1}^{J}\sum_{i=1}^{n}\left\|\mathbf{f}'_i - \mathbf{g}'_{ij}\mathbf{C}_j\right\|^2 = \sum_{j=1}^{J}\left\|\begin{bmatrix}\mathbf{f}'_1\\ \vdots\\ \mathbf{f}'_n\end{bmatrix} - \begin{bmatrix}\mathbf{g}'_{1j}\\ \vdots\\ \mathbf{g}'_{nj}\end{bmatrix}\mathbf{C}_j\right\|^2. \tag{14.45}$$

Here, we have used

$$\mathbf{g}_{ij}'\mathbf{C}_j = \mathbf{g}_{ij}' \begin{bmatrix} \mathbf{c}_{j1}' \\ \vdots \\ \mathbf{c}_{jK_j}' \end{bmatrix} = \sum_{k=1}^{K_j} g_{ijk}\mathbf{c}_{jk}' = \mathbf{c}_{jy_{ij}}', \tag{14.46}$$

because of (14.34). We can find the equivalence of (14.45) to (14.36) by noting (14.33) and $\mathbf{F} = [\mathbf{f}_1, \ldots, \mathbf{f}_n]'$.

The inter-point distances in Fig. 14.2 allow us to capture the relationships among categories, among individuals, and between categories and individuals; we can consider the entities near one another to share similar features. For example, (1) the point for "Sciences" is found to be close to that for "Basis", which shows that the students in the department of "Sciences" tend to regard "Basic" sciences as important; (2) individuals 1 and 6 are similar students; (3) individual 3 is involved with "Medicine" and "Applications" (or applied sciences).

The spatial representation of the MCA solution as in Fig. 14.2 and its spatial interpretation show the reason why we refer to the columns of $\mathbf{F}$ and $\mathbf{C}_j$ as dimensions.

14.7 Bibliographical Notes

CCA is intelligibly introduced in Lattin, Carroll, and Green (2003) with real data examples, and detailed in Izenman (2008) and Kock (2014). The formulations of GCCA and MCA in this chapter are detailed in Gifi (1990). MCA is also intelligibly treated in Greenacre (2007). The analysis procedure called *correspondence analysis*, with the "multiple" deleted from MCA, is treated only in the next exercises. The relationships between correspondence analysis and MCA are detailed in Greenacre (1984, 2007).

We must mention that various terms have been used for referring to MCA and related procedures. For example, the term *homogeneity analysis* has been used in Gifi (1990). Other terms can be found in Hayashi (1952), Nishisato (1980), and Young (1981).

Recently, Shimodaira (2016) has proposed a procedure which can be viewed as a generalization of GCCA.

Exercises

14.1 Show that (14.16) can be rewritten as $||\mathbf{I}_J \otimes \mathbf{F} - \mathbf{X}^{\#}\mathbf{C}||^2$, where $\mathbf{X}^{\#} = \begin{bmatrix} \mathbf{X}_1 & & \\ & \ddots & \\ & & \mathbf{X}_J \end{bmatrix}$ is the $nJ \times p$ block diagonal matrix whose jth diagonal

block is $\mathbf{X}_j$, and $\mathbf{I}_J \otimes \mathbf{F} = \begin{bmatrix} \mathbf{F} \\ \vdots \\ \mathbf{F} \end{bmatrix}$ is the $nJ \times m$ block matrix whose all blocks are $\mathbf{F}$. The operator $\otimes$ is called *Kronecker product* and is detailed in Chap. 17.

14.2 Discuss how the generalized orthogonal rotation in Exercise 13.9 and GCCA are similar/different.

14.3 Let $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_p]$ contain standard scores with $\mathbf{1}_n'\mathbf{z}_j = 0$ and $n^{-1}\mathbf{z}_j'\mathbf{z}_j = 1 (j = 1, \ldots, p)$. We can substitute $\mathbf{z}_j$ for $\mathbf{X}_j$ in (14.16) to rewrite it as $\eta(\mathbf{F}, \mathbf{A}) = \sum_{j=1}^{p} \left\|\mathbf{F} - \mathbf{z}_j\mathbf{a}'_j\right\|^2$, with $\mathbf{C}$ in (14.16) replaced by $\mathbf{A} = \left[\mathbf{a}_1, \ldots, \mathbf{a}_p\right]'$. By noting the equivalence between (14.16) and (14.29), show that the minimization of $\eta(\mathbf{F}, \mathbf{A})$ subject to (14.17) is equivalent to the *principal component analysis* (*PCA*) for $\mathbf{Z}$, i.e., minimizing $||\mathbf{Z} - \mathbf{F}\mathbf{A}'||^2$ under (14.17).

14.4 Show that the function (14.29) multiplied by n can be rewritten as:

$$nf(\mathbf{F}, \mathbf{C}) = ||\mathbf{X}\mathbf{D}_{\mathrm{V}}^{-1/2} - \mathbf{F}\mathbf{C}'\mathbf{D}_{\mathrm{V}}^{1/2}||^2$$

with $\mathbf{D}_{\mathrm{V}} = \begin{bmatrix} \mathbf{V}_1 & & \\ & \ddots & \\ & & \mathbf{V}_J \end{bmatrix}$ the block diagonal matrix, whose jth block $\mathbf{V}_j$ is defined as $\mathbf{V}_j = n^{-1}\mathbf{X}_j'\mathbf{X}_j$ and is the covariance matrix for $\mathbf{X}_j$ if it is centered.

14.5 Let us constrain $\mathbf{C}_j$ in (14.36) to be $\mathbf{C}_j = \mathbf{q}_j\mathbf{a}_j'$, with $\mathbf{q}_j$ and $\mathbf{a}_j$ being $K_j \times 1$ and $m \times 1$ vectors, respectively. Then, (14.36) is rewritten as

$$\eta\left(\mathbf{F}, \mathbf{q}_j, \mathbf{a}_j\right) = \sum_{j=1}^{J} \left\|\mathbf{F} - \mathbf{G}_j\mathbf{q}_j\mathbf{a}'_j\right\|^2.$$

Show that its minimization over $\mathbf{F}$, $\mathbf{q}_1$, …, $\mathbf{q}_J$, and $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_J]$ subject to (14.17), (14.37), $\mathbf{1}_n'\mathbf{G}_j\mathbf{q}_j = 0$, and $n^{-1}\left(\mathbf{G}_j\mathbf{q}_i\right)'\mathbf{G}_j\mathbf{q}_i = 1$ is equivalent to minimizing $||\mathbf{G}_{\mathrm{Q}} - \mathbf{F}\mathbf{A}'||^2$ under the same constraints with $\mathbf{G}_{\mathrm{Q}} = [\mathbf{G}_1\mathbf{q}_1, \mathbf{G}_2\mathbf{q}_2, \ldots, \mathbf{G}_J\mathbf{q}_J]$ an $n \times J$ matrix (Gifi, 1990).

14.6 Discuss how the assignment of quantitative scores to categories and PCA are simultaneously performed in the procedure considered in Exercise 14.5.

14.7 Show that $\mathbf{N} = (n_{kl}) = \mathbf{G}_1'\mathbf{G}_2$ represents the $K_1 \times K_2$ *contingency table*, whose element n_{kl} expresses the number of individuals classified into category k for variable 1 and category l for variable 2.

14.8 Show that $\mathbf{G'}_1\mathbf{J}\mathbf{G}_2 = \mathbf{N} - n^{-1}\mathbf{D}_1\mathbf{1}_{K_1}\mathbf{1}'_{K_2}\mathbf{D}_2$, with $\mathbf{N}$ defined in Exercise 14.7, $\mathbf{J} = \mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}'_n$ the centering matrix, and $\mathbf{D}_j$ the $K_j \times K_j$ diagonal matrix whose kth diagonal element is the number of the individuals classified into category k for variable j (=1, 2).

14.9 The procedure called *correspondence analysis* with removing "multiple" from "multiple correspondence analysis" is performed for the contingency table $\mathbf{N}$ defined in Exercise 14.7 (Benzécri, 1992; Greenacre 1984). The loss function of correspondence analysis is expressed as:

$$f(\mathbf{C}_1, \mathbf{C}_2) = ||\widetilde{\mathbf{N}} - \frac{1}{n}\mathbf{D}_1^{1/2}\mathbf{C}_1\mathbf{C}'_2\mathbf{D}_2^{1/2}||^2, \tag{14.48}$$

which is minimized over $\mathbf{C}_1$ and $\mathbf{C}_2$, with

$$\widetilde{\mathbf{N}} = \mathbf{D}_1^{-1/2}(\mathbf{N} - n^{-1}\mathbf{D}_1\mathbf{1}_{K_1}\mathbf{1}'_{K_2}\mathbf{D}_2)\mathbf{D}_2^{-1/2} = \mathbf{D}_1^{-1/2}\mathbf{G'}_1\mathbf{J}\mathbf{G}_2\mathbf{D}_2^{-1/2}.$$

Show that (14.48) is minimized for

$$\mathbf{C}_1 = \sqrt{n}\mathbf{D}_1^{-1/2}\mathbf{U}_m\mathbf{\Delta}_m^{1/2} \quad \text{and} \quad \mathbf{C}_2 = \sqrt{n}\mathbf{D}_2^{-1/2}\mathbf{V}_m\mathbf{\Delta}_m^{1/2} \tag{14.49}$$

subject to $\mathbf{C}'_1\mathbf{D}_1\mathbf{C}_1 = \mathbf{C}'_2\mathbf{D}_2\mathbf{C}_2$ being a diagonal matrix. Here, $\mathbf{U}_m$ and $\mathbf{V}_m$ contain the first m columns of $\mathbf{U}$ and $\mathbf{V}$, respectively, while $\mathbf{\Delta}_m$ is the first $m \times m$ diagonal block of $\mathbf{\Delta}$, with $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{\Delta}$ obtained from the SVD $\widetilde{\mathbf{N}} = \mathbf{U}\Delta\mathbf{V}'$.

14.10 The solution of MCA for $\mathbf{G}$ with $K = 2$, i.e., $\mathbf{G} = [\mathbf{G}_1, \mathbf{G}_2]$, is given through the SVD (14.43) with $K = 2$, which is rewritten as:

$$\mathbf{J}[\mathbf{G}_1, \mathbf{G}_2]\begin{bmatrix}\mathbf{D}_1^{-1/2} & \\ & \mathbf{D}_2^{-1/2}\end{bmatrix} = \mathbf{S}\Theta[\mathbf{P'}_1, \mathbf{P'}_2], \tag{14.50}$$

with $\mathbf{P}_1$ ($K_1 \times r$) and $\mathbf{P}_2$ ($K_2 \times r$) the blocks of $\mathbf{P} = \begin{bmatrix}\mathbf{P}_1 \\ \mathbf{P}_2\end{bmatrix}$. Show that (14.50) leads to

$$\begin{bmatrix}\mathbf{D}_1^{-1/2} & \\ & \mathbf{D}_2^{-1/2}\end{bmatrix}\begin{bmatrix}\mathbf{G'}_1 \\ \mathbf{G'}_2\end{bmatrix}\mathbf{J}[\mathbf{G}_1, \mathbf{G}_2]\begin{bmatrix}\mathbf{D}_1^{-1/2} & \\ & \mathbf{D}_2^{-1/2}\end{bmatrix} = \begin{bmatrix}\mathbf{P}_1 \\ \mathbf{P}_2\end{bmatrix}\Theta[\mathbf{P'}_1, \mathbf{P'}_2]$$

and that its left-hand side can be rewritten as $\begin{bmatrix}\mathbf{M}_1 & \widetilde{\mathbf{N}} \\ \widetilde{\mathbf{N}}' & \mathbf{M}_2\end{bmatrix}$, which imply the equivalence of the correspondence analysis to MCA for $[\mathbf{G}_1, \mathbf{G}_2]$ with the constraint of $\mathbf{C}'_1\mathbf{D}_1\mathbf{C}_1 = \mathbf{C}'_2\mathbf{D}_2\mathbf{C}_2$ being a diagonal matrix. Here, the symbols have been the ones defined in Exercises 14.8 and 14.9, with $\mathbf{M}_j = \mathbf{I}_{K_j} - n^{-1}\mathbf{D}_j^{1/2}\mathbf{1}_{K_j}\mathbf{1}'_{K_j}\mathbf{D}_j^{1/2}$.

Chapter 15
Discriminant Analysis

Discriminant analysis refers to a group of statistical procedures for analyzing a data set with individuals *classified* into certain *groups*, where the results of the analysis are used for *finding* the *group* of a *new individual* that is not included in the above data set. The sections in this chapter can be classified into two parts: (1) Sects. 15.1–15.3 concern an approach *without using probabilities,* and (2) the remaining sections concern *probabilistic approaches*. In (1), a *canonical discriminant analysis* (*CDA*) procedure is introduced by modifying the multiple correspondence analysis in the last chapter. In (2), we introduce two probabilistic procedures using multivariate normal distributions. One of them is *linear discriminant analysis* (*LDA*), which is rooted in Fisher (1936) and found to be equivalent to CDA. The other is a *generalization* of LDA.

15.1 Modification of Multiple Correspondence Analysis

The *multiple correspondence analysis* (*MCA*) in the last chapter is performed for the n individuals $\times$ K-categories membership matrix (14.35). Here, let us consider a case where $J = 1$, i.e., $\mathbf{G} = \mathbf{G}_1$, and an n individuals $\times$ p-variables quantitative data matrix $\mathbf{X}$ corresponding to $\mathbf{G}$ is also given, with $\mathbf{1}_n'\mathbf{X} = \mathbf{0}_p'$. That is, the data set is expressed as an $n \times (K + p)$ block matrix $[\mathbf{G}, \mathbf{X}]$. An example of $[\mathbf{G}, \mathbf{X}]$ is shown in Table 15.1 (Fisher, 1936), in which individuals are *irises* whose categories are indicated by $\mathbf{G}$ and the individuals' features are described by $\mathbf{X}$. In this chapter, the column entities of $\mathbf{G}$ are called *groups* rather than categories.

K. Adachi, *Matrix-Based Introduction to Multivariate Data Analysis*,
https://doi.org/10.1007/978-981-15-4103-2_15

Table 15.1 Membership of irises for groups 1, 2, and 3 (**G**) and standardized scores for features of the irises (**X**). The original data are available at http://astro.temple.edu/∼alan/MMST/datasets.htm (Izenman, 2008)

Iris	G			X			
	1	2	3	SL*	SW*	PL*	PW*
1	1	0	0	−0.90	1.02	−1.34	−1.31
2	1	0	0	−1.14	−0.13	−1.34	−1.31
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
50	1	0	0	−1.02	0.56	−1.34	−1.31
51	0	1	0	1.40	0.33	0.53	0.26
52	0	1	0	0.67	0.33	0.42	0.39
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
100	0	1	0	−0.17	−0.59	0.19	0.13
101	0	0	1	0.55	0.56	1.27	1.71
102	0	0	1	−0.05	−0.82	0.76	0.92
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
150	0	0	1	0.07	−0.13	0.76	0.79

**SL* sepal length, *SW* sepal width, *PL* petal length, *PW* petal width

For the above **G**, the MCA loss function (14.36) is simplified into $\|\mathbf{F}-\mathbf{GC}\|^2$ without the symbol of summation and the subscript for **C**. Here, let the individual score matrix **F** be constrained as

$$\mathbf{F}=\mathbf{XB}, \tag{15.1}$$

with **B** a $p \times m$ coefficient matrix. Using (15.1) in $\|\mathbf{F}-\mathbf{GC}\|^2$, it is rewritten as

$$\eta(\mathbf{B},\mathbf{C})=\|\mathbf{XB}-\mathbf{GC}\|^2. \tag{15.2}$$

Further, the substitution of (15.1) into constraint (14.17) leads to

$$\frac{1}{n}\mathbf{B}'\mathbf{X}'\mathbf{XB}=\mathbf{I}_m. \tag{15.3}$$

Minimizing (15.2) over **B** and **C** subject to (15.3) is called *canonical discriminant analysis* (*CDA*), whose solution is detailed in Sect. 15.2. Before it, discriminant analysis is compared with clustering in the following note.

Note 15.1. Comparison to Cluster Analysis
Let us compare (15.2) with the loss function (7.4) in *k-means clustering* (Chap. 7). Deleting $\mathbf{B}$ from (15.2) leads to (7.4). Further, the matrix $\mathbf{G}$, which indicates the memberships of individuals to groups, is *known* in (15.2) (discriminant analysis), while $\mathbf{G}$ is *unknown* and to be obtained in (7.4) (cluster analysis). For this difference, discriminant analysis is referred to as *supervised classification*, while cluster analysis is called *unsupervised classification*, as the former is concerned with the classification when the data set exists that serves as the supervisor indicating the memberships, while such a data set or supervisor does not exist in the latter.

15.2 Canonical Discriminant Analysis

As shown in Appendix A.2.2, (15.2) is minimized for

$$\mathbf{C} = (\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}'\mathbf{F} = \mathbf{D}_{\mathrm{G}}^{-1}\mathbf{G}'\mathbf{X}\mathbf{B}, \tag{15.4}$$

given $\mathbf{G}$, with $\mathbf{D}_{\mathrm{G}} = \mathbf{G}'\mathbf{G}$, a $K \times K$ diagonal matrix. We can substitute (15.4) in (15.2) to rewrite it as

$$\begin{aligned}\eta(\mathbf{B}) &= \left\|\mathbf{X}\mathbf{B} - \mathbf{G}\mathbf{D}_{\mathrm{G}}^{-1}\mathbf{G}'\mathbf{X}\mathbf{B}\right\|^2 \\ &= \mathrm{tr}\,\mathbf{B}'\mathbf{X}'\mathbf{X}\mathbf{B} - 2\mathrm{tr}\mathbf{B}'\mathbf{X}'\mathbf{G}\mathbf{D}_{\mathrm{G}}^{-1}\mathbf{G}'\mathbf{X}\mathbf{B} + \mathrm{tr}\mathbf{B}'\mathbf{X}'\mathbf{G}\mathbf{D}_{\mathrm{G}}^{-1}(\mathbf{G}'\mathbf{G})\mathbf{D}_{\mathrm{G}}^{-1}\mathbf{G}'\mathbf{X}\mathbf{B} \\ &= nm - \mathrm{tr}\mathbf{B}'\mathbf{X}'\mathbf{G}\mathbf{D}_{\mathrm{G}}^{-1}\mathbf{G}'\mathbf{X}\mathbf{B},\end{aligned} \tag{15.5}$$

where we have used (15.3) and $\mathbf{G}'\mathbf{G} = \mathbf{D}_{\mathrm{G}}$. The minimization of (15.5) under (15.3) is equivalent to maximizing $\mathrm{tr}\mathbf{B}'\mathbf{X}'\mathbf{G}\mathbf{D}_{\mathrm{G}}^{-1}\mathbf{G}'\mathbf{X}\mathbf{B}$ subject to (15.3), whose solution is given as in Theorem A.4.9 (Appendix A.4.5). There, by setting $\mathbf{M}$ and $\mathbf{V}$ in (A.4.41) to $\mathbf{X}'\mathbf{G}\mathbf{D}_{\mathrm{G}}^{-1}\mathbf{G}'\mathbf{X}$ and $\mathbf{V} = n^{-1}\mathbf{X}'\mathbf{X}$, respectively, we have the solution for $\mathbf{B}$, as in (A.4.43).

Note 15.2. Another Formulation of CDA
As found above, CDA can be formulated as maximizing $\rho(\mathbf{B}) = \mathrm{tr}\mathbf{B}'\mathbf{S}\mathbf{B}$ over $\mathbf{B}$ subject to (15.3) with $\mathbf{S} = \mathbf{X}'\mathbf{G}\mathbf{D}_{\mathrm{G}}^{-1}\mathbf{G}'\mathbf{X}$. In a more popular introduction of CDA, (15.3) is replaced by $\mathbf{B}'\mathbf{W}\mathbf{B} = \mathbf{I}_m$ with $\mathbf{W} = n^{-1}(\mathbf{X}'\mathbf{X} - \mathbf{S})$: CDA is also formulated as maximizing $\rho(\mathbf{B})$ under $\mathbf{B}'\mathbf{W}\mathbf{B} = \mathbf{I}_m$. A reason for using (15.3) in this book is relating CDA to MCA.

Let us express the ith row of $\mathbf{X}$ $(n \times p)$ as $\mathbf{x}_i' = [x_{i1}, \ldots, x_{ip}]$ and the lth column of $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_m](p \times m)$ as $\mathbf{b}_l = [b_{1l}, \ldots, b_{pl}]'$. Then, the (i, l) element of (15.1) is expressed as

$$f_{il} = \mathbf{x}_i' \mathbf{b}_l = b_{1l} x_{i1} + \cdots + b_{pl} x_{ip}, \tag{15.6}$$

i.e., the weighted sum of the p variables in $\mathbf{x}_i$. Sum (15.6) is called the lth *discriminant score* for individual i, and the weights b_{1l}, …, b_{pl} are called the lth *discriminant coefficients*, with l = 1, …, m. The other parameter matrix in CDA is $\mathbf{C}$. Its rows of $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_K]'$ are associated with groups, and the kth row $\mathbf{c}_k'$ $(1 \times m)$ can be called the kth *group score vector*, as it stands for the features of the group.

Let us consider performing CDA for the iris data in Table 15.1, setting m = 2. This gives us $\mathbf{F}$, whose ith row is expressed as $\mathbf{f}_i' = [f_{i1}, f_{i2}] = [\mathbf{x}_i' \mathbf{b}_1, \mathbf{x}_i' \mathbf{b}_2] = \mathbf{x}_i' \mathbf{B}$, i.e., two discriminant scores for each individual. The resulting scores for the data set are expressed as

$$\mathbf{x}_i' \mathbf{b}_1 = 0.12 \times \mathrm{SL} + 0.12 \times \mathrm{SW} - 0.68 \times \mathrm{PL} - 0.38 \times \mathrm{PW}, \tag{15.7}$$

$$\mathbf{x}_i' \mathbf{b}_2 = -0.02 \times \mathrm{SL} - 0.84 \times \mathrm{SW} + 1.47 \times \mathrm{PL} - 1.94 \times \mathrm{PW}, \tag{15.8}$$

where the names of the variables in Table 15.1 and the solutions of the coefficients are substituted into x_{i1}, …, x_{ip} and b_{1l}, …, b_{pl} in (15.6), respectively. For example, the elements of the data vector $\mathbf{x}_1'(=[\mathrm{SL}, \mathrm{SW}, \mathrm{PL}, \mathrm{PW}]) = [-0.90, 1.02, -1.34, -1.31]$ for individual 1 can be substituted into the variables in (15.7) and (15.8) so that the discriminant score vector for individual 1 is given as $\mathbf{f}_1' = \mathbf{x}_1' \mathbf{B} = [\mathbf{x}_1' \mathbf{b}_1, \mathbf{x}_1' \mathbf{b}_2] = [1.42, -0.28]$. In Fig. 15.1, the vectors for all individuals, $\mathbf{f}_i' = \mathbf{x}_i' \mathbf{B} (i = 1, \ldots, 150)$, are plotted, with squares, circles, and triangles used for the individuals in Group 1, 2, and 3, respectively.

The CDA for the data in Table 15.1 also gives the solution of the group scores as

$$\mathbf{C} = \begin{bmatrix} \mathbf{c}_1' \\ \mathbf{c}_2' \\ \mathbf{c}_3' \end{bmatrix} = \begin{bmatrix} 1.33 & -0.19 \\ -0.31 & 0.65 \\ -1.01 & -0.46 \end{bmatrix}. \tag{15.9}$$

In Fig. 15.1, $\mathbf{c}_1$, $\mathbf{c}_2$, and $\mathbf{c}_3$ are represented as a filled square, circle, and triangle, respectively. There, we can find that the discriminant scores for the individuals in the same group are distributed mutually close, with their *center* being the *group score* vector. This can be mathematically shown in the next section.

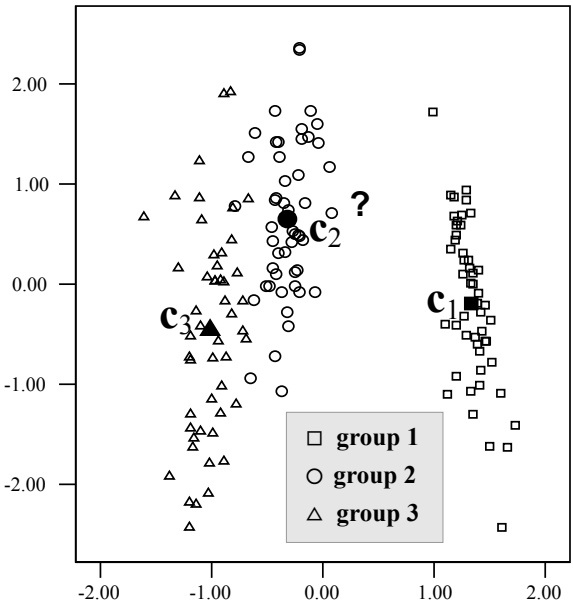


Fig. 15.1 Plots of individuals' discriminant scores and group scores

15.3 Minimum Distance Classification

Equation (15.4) for $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_K]'$ implies that its kth row $\mathbf{c}_k'$ is the *centroid*, i.e., the averaged vector of the discriminant score vectors $\mathbf{f}_i'$ for the individuals belonging to group k:

$$\mathbf{c}_k' = \frac{1}{n_k} \sum_{i \in g_k} \mathbf{f}_i' = \frac{1}{n_k} \sum_{i \in g_k} \mathbf{x}_i' \mathbf{B}. \tag{15.10}$$

Here, g_k expresses the set of the individuals in group k with their number denoted by n_k, and $\sum_{i \in g_k} \mathbf{f}_i'$ stands for the summation of $\mathbf{f}_i'$ over the individuals belonging to group k. The rows of $\mathbf{C}$ being averages can be verified by the following example: (15.4) is expressed as

$$\mathbf{C} = \mathbf{D}_{\mathrm{G}}^{-1} \mathbf{G}' \begin{bmatrix} \mathbf{f}_1' \\ \mathbf{f}_2' \\ \mathbf{f}_3' \\ \mathbf{f}_4' \\ \mathbf{f}_5' \end{bmatrix} = \begin{bmatrix} \frac{1}{3}(\mathbf{f}_2' + \mathbf{f}_3' + \mathbf{f}_5') \\ \frac{1}{2}(\mathbf{f}_1' + \mathbf{f}_4') \end{bmatrix}, \text{ when } \mathbf{G} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \text{ with } \mathbf{D}_{\mathrm{G}} = \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}.$$

Further, we can ascertain the closeness of the vectors $\mathbf{f}_i' = \mathbf{x}_i'\mathbf{B}$ in group k to $\mathbf{c}_k'$ from the fact that (15.2) is rewritten as

$$\eta(\mathbf{B}, \mathbf{C}) = \sum_{i=1}^{n} \left\| \mathbf{x}_i'\mathbf{B} - \mathbf{g}_i'\mathbf{C} \right\|^2 = \sum_{i=1}^{n} \left\| \mathbf{x}_i'\mathbf{B} - \mathbf{c}_{y_i}' \right\|^2. \tag{15.11}$$

Here, $\mathbf{g}_i'$ is the ith row of $\mathbf{G}$, y_i is the index number of the group to which individual i belongs, and we have used $\mathbf{g}_i'\mathbf{C} = \mathbf{c}_{y_i}'$. This implies that CDA is also based on the *homogeneity assumption*:

$$\underline{\text{the scores for an } \textit{individual} \text{ should be } \textit{homogeneous} \text{ to}} \\ \underline{\text{the scores for the } \textit{group} \text{ to which the } \textit{individual belongs}}, \tag{15.12}$$

which is the same as (14.44) except the term "categories" has been replaced by "group". Minimizing (15.11) allows $\mathbf{f}_i' = \mathbf{x}_i'\mathbf{B}$ to be close to $\mathbf{c}_{y_i}'$, with $\mathbf{c}_{y_i}'$ being the score of the group including individual i, which is also the centroid of the individual scores in that group, as shown in (15.10).

Let $\mathbf{x}_?'$ be a $1 \times p$ vector which is *not included* in $\mathbf{X}$ so that it is *unknown* to what *group* $\mathbf{x}_?'$ belongs. That is, our task is to *classify* $\mathbf{x}_?'$ into one of the groups $k = 1, \ldots, K$, in other words, to find the group to which $\mathbf{x}_?'$ should belong. Assumption (15.12) leads to the following *minimum distance classification*:

$$\mathbf{x} \text{ is classified into group } k^* \text{ with } \left\| \mathbf{x}'\mathbf{B} - \mathbf{c}_{k^*}' \right\| = \min_{1 \le k \le K} \left\| \mathbf{x}'\mathbf{B} - \mathbf{c}_k' \right\|. \tag{15.13}$$

Here, $\mathbf{x}'$ generally expresses a $1 \times p$ vector whose elements are associated with the p variables in $\mathbf{X}$. We illustrate the classification rule (15.13) with $\mathbf{x}$ equaling $\mathbf{x}_? = [1.8, \ 0.4, \ 0.1, \ -0.6]$. This is substituted into (15.6) to provide $\mathbf{x}_?'\mathbf{B} = [0.42, 0.94]$, with the elements of $\mathbf{B}$ given as in (15.7) and (15.8). The location of $\mathbf{x}_?'\mathbf{B}$ is shown by "**?**" in Fig. 15.1. By comparing its *distances* to $\mathbf{c}_1$, $\mathbf{c}_2$, and $\mathbf{c}_3$, we can find that $\mathbf{x}_?'\mathbf{B}$ is closest to $\mathbf{c}_2$, and thus, $\mathbf{x}_?$ is reasonably classified into Group 2.

15.4 Maximum Probability Classification

Beginning with this section, discriminant analysis will be formulated in a different manner: We start with a classification rule, in which the *distances* and "min" in (15.13) are replaced by *probabilities* and "max", respectively. The rule is stated as follows:

$$\mathbf{x} \text{ is classified into group } k^* \text{ with } P(g_{k^*}|\mathbf{x}) = \max_{1 \le k \le K} P(g_k|\mathbf{x}). \tag{15.14}$$

Here, $P(\mathrm{g}_k|\mathbf{x})$ stands for the probability that the individual showing $\mathbf{x}$ belongs to group k. This particular probability is called a *posterior probability* as it is related to considering the group from which $\mathbf{x}$ arises a posteriori, after $\mathbf{x}$ was observed. Interchanging g_k and $\mathbf{x}$ in $P(\mathrm{g}_k|\mathbf{x})$ gives the symbol $P(\mathbf{x}|\mathrm{g}_k)$, which is called a *group-conditional density*, and stands for the probability density of an individual in group k showing $\mathbf{x}$. Between $P(\mathrm{g}_k|\mathbf{x})$ and $P(\mathbf{x}|\mathrm{g}_k)$, the following equation is known to hold:

$$P(g_k|\mathbf{x}) = \frac{P(g_k)P(\mathbf{x}|g_k)}{\sum_{l=1}^{K} P(g_l)P(\mathbf{x}|g_l)}. \tag{15.15}$$

Here, $P(g_k)$ is a probability of a randomly chosen individual belonging to group k and called a *prior probability*, as it is given a priori, before $\mathbf{x}$ is observed. Equation (15.15) is known as the *Bayes' theorem*, as it was found by English pastor, Thomas Bayes (1701–1761). Thus, (15.14) is called the *Bayes' classification rule*.

As found in (15.15), we can obtain the posterior probability $P(g_k|\mathbf{x})$ necessary for classifying $\mathbf{x}$ with (15.14) if group-conditional densities $P(\mathbf{x}|g_k)$ and prior probabilities $P(g_k)$ ($k = 1, \ldots, K$) are estimated. This estimation is made using the data set $[\mathbf{G}, \mathbf{X}]$. The facts described in $[\mathbf{G}, \mathbf{X}]$ can also be expressed *without* using $\mathbf{G}$, by means of *rearranging* the individuals in $\mathbf{X}$ so that the ones belonging to the same group are collected in the same block. The rearrangement gives an n individuals $\times$ p-variables block matrix

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_k \\ \vdots \\ \mathbf{X}_K \end{bmatrix} \quad \text{with} \quad \mathbf{X}_k = \begin{bmatrix} \mathbf{x}'_{k1} \\ \vdots \\ \mathbf{x}'_{ki} \\ \vdots \\ \mathbf{x}'_{kn_k} \end{bmatrix}. \tag{15.16}$$

Here, $n = n_1 + \cdots + n_K$, and $\mathbf{x}_{ki}$ is the $p \times 1$ data vector for the ith one of the individuals belonging to group k. In the remaining sections, (15.16) is used for a data matrix with the memberships of individuals to groups known. Further, $P(\mathbf{x}|\mathrm{g}_k)$ is supposed to be the probability density of a *multivariate normal* (*MVN*) *distribution*:

$$\mathbf{x} \sim N_p(\boldsymbol{\mu}_k, \Sigma_k) \quad \text{for} \quad \mathbf{x} \in g_k, \tag{15.17}$$

with $\mathbf{x} \in g_k$ representing the fact that the individual showing $\mathbf{x}$ belongs to group k. That is, the group-conditional density for group k is given as

$$P(\mathbf{x}|g_k) = P(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_k|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right\} \tag{15.18}$$

by adding the subscript k to (8.9).

15.5 Normal Discrimination for Two Groups

In this section, the number of *groups* is restricted to *two* ($K = 2$), and the *covariance* matrix in (15.18) is supposed to be *homogeneous* between two groups:

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma} \tag{15.19}$$

Then, using (15.15), the rule (15.14) is rewritten as follows: $\mathbf{x}$ is classified into g_1 if $P(g_1)P(\mathbf{x}|g_1) \geq P(g_2)P(\mathbf{x}|g_2)$ or, equivalently,

$$\frac{p(\mathbf{x}|g_1)}{p(\mathbf{x}|g_2)} \geq \frac{p(g_2)}{p(g_1)}; \tag{15.20}$$

otherwise, $\mathbf{x}$ is classified into g_2. By changing both sides of (15.20) into their logarithm, we can rewrite it as $\log P(\mathbf{x}|\mathrm{g}_1) - \log P(\mathbf{x}|\mathrm{g}_2) \geq \log P(\mathrm{g}_2) - \log P(\mathrm{g}_1)$, or equivalently,

$$f(\mathbf{x}) = \log P(\mathbf{x}|\mathrm{g}_1) - \log P(\mathbf{x}|\mathrm{g}_2) + \log\{P(\mathrm{g}_1)/P(\mathrm{g}_2)\} \geq 0. \tag{15.21}$$

Further, by substituting (15.18) into (15.21) with the use of (15.19), we can rewrite the function in (15.21) as

$$\begin{aligned} f(\mathbf{x}) &= -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) + \log\{P(\mathrm{g}_1)/P(\mathrm{g}_2)\} \\ &= \mathbf{x}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \frac{1}{2}\left(\boldsymbol{\mu}_2'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1\right) + \log\{P(\mathrm{g}_1)/P(\mathrm{g}_2)\} \\ &= \mathbf{b}'\mathbf{x} + c = b_1 x_1 + \cdots + b_p x_p + c, \end{aligned} \tag{15.22}$$

with

$$\mathbf{b} = \left[b_1, \ldots, b_p\right]' = \boldsymbol{\Sigma}^{-1}(\mu_1 - \mu_2), \tag{15.23}$$

$$c = \frac{1}{2}\left(\boldsymbol{\mu}_2'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1\right) + \log\{P(\mathrm{g}_1)/P(\mathrm{g}_2)\}. \tag{15.24}$$

Rule (15.14) is thus simplified as

$$\mathbf{x} \text{ is classified into Group 1 if } f(\mathbf{x}) > 0; \text{ otherwise, into Group 2.} \tag{15.25}$$

As (15.22) is a linear function of $\mathbf{x}$, which is the weighted composite of variables, (15.22) is called a *linear discriminant function* (*LDF*), and the procedure for obtaining (15.22) is called *linear discriminant analysis* (*LDA*). As described in Appendix A.5.2, the maximum likelihood estimates of $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, and $\boldsymbol{\Sigma}$ needed for obtaining (15.22) are given by

$$\hat{\mu}_k' = \bar{\mathbf{x}}_k' = \frac{1}{n}\mathbf{1}_{n_k}'\mathbf{X}_k \quad (k = 1, 2), \tag{15.26}$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\left\{\sum_{i=1}^{n_1}(\mathbf{x}_{1i} - \bar{\mathbf{x}}_1)(\mathbf{x}_{1i} - \bar{\mathbf{x}}_1)' + \sum_{i=1}^{n_2}(\mathbf{x}_{2i} - \bar{\mathbf{x}}_2)(\mathbf{x}_{2i} - \bar{\mathbf{x}}_2)'\right\}. \tag{15.27}$$

These are substituted into $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, and $\boldsymbol{\Sigma}$ in (15.23) and (15.24) for providing $\mathbf{b}$ and c, though $P(g_1)/P(g_2)$ must also be estimated for obtaining c.

For example, we consider a case of $p = 2$, where

$$P(g_1) = P(g_2) \tag{15.28}$$

is supposed, and $\hat{\mu}_1 = [76.20, 61.42]'$, $\hat{\mu}_2 = [66.93, 72.16]'$, and $\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} 120.77 & 60.05 \\ 60.05 & 146.98 \end{bmatrix}$. By substituting these into (15.23) and (15.24), we have $\mathbf{b} = [0.14, -0.13]$ and $c = -1.40$. They lead to the LDF

$$f(\mathbf{x}) = 0.14x_1 - 0.13x_2 - 1.40. \tag{15.29}$$

The classification in which (15.29) is used for (15.25) can be graphically illustrated as in Fig. 15.2, where the bird's-eye view of the group-conditional densities for the two groups is depicted. As found there, the LDF (15.29) value of $\mathbf{x} = [x_1, x_2]$ is the coordinate on the axis called a *discriminant axis*. For example, let the point "**?**" in the figure indicate $\mathbf{x}_? = [58, 62]'$, i.e., a new observation to be classified. This leads to

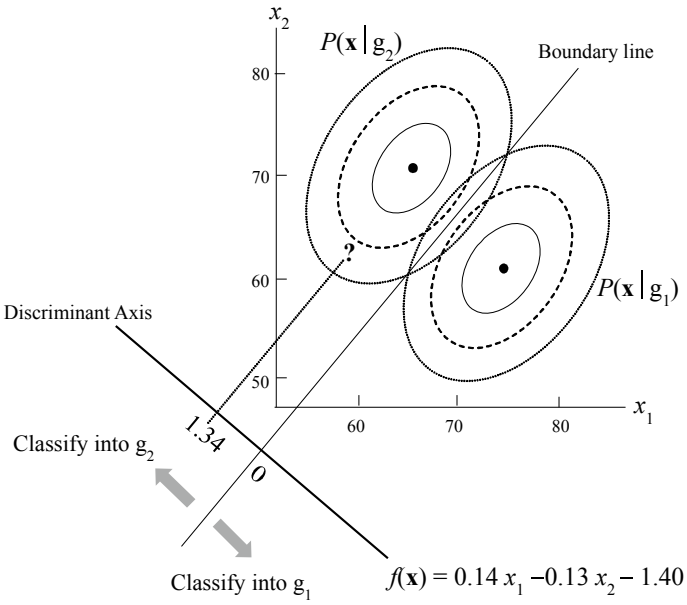


Fig. 15.2 Illustration of linear discriminant analysis

$$f\left([58, 62]'\right) = 0.14 \times 58 - 0.13 \times 62 - 1.40 = -1.34, \tag{15.30}$$

which is a coordinate on the discriminant axis. The LDF value in (15.30) is called a *discriminant score*. Since (15.30) is negative, (15.25) shows that $\mathbf{x}_? = [58, 62]'$ is to be classified into Group 2. Let us note the *boundary line* in Fig. 15.2. It defines the regions for two groups: The observations $\mathbf{x}$ located to the right of/below the line are classified into g_1 and those on the other side are classified into g_2.

15.6 Interpreting Solutions

For illustrating the interpretation of LDA solutions, we consider performing LDA for the 27 (employees) $\times$ 4 (personality traits) data matrix $\mathbf{X} = [\mathbf{X}_1', \mathbf{X}_2']'$ in Table 15.2a. Here, it is supposed that the personality traits of the employees are fit to their groups (i.e., departments). Substituting the solution of (15.23) and (15.24) in (15.22) leads to the LDF as

Table 15.2 Artificial example of the data for LDF (Adachi, 2006) and the resulting classification based on the discriminant scores

Department	Employ.	(a) Data				(b) Result	
		Social	Cooperative	Diligent	Creative	Score	Classi.
g_1	1	15	14	15	14	2.56	g_1
	2	11	13	17	17	−1.64	g_2^*
	3	16	14	17	26	1.34	g_1
	4	19	21	18	15	4.94	g_1
	5	18	26	21	15	3.93	g_1
	6	15	28	18	12	3.28	g_1
	7	17	19	12	10	6.41	g_1
	8	12	15	18	12	−0.68	g_2^*
	9	13	22	16	10	2.10	g_1
	10	14	26	18	6	2.79	g_1
	11	16	20	18	18	2.39	g_1
	12	11	15	20	15	−2.58	g_2^*
	13	20	21	17	20	5.70	g_1
	14	15	20	19	12	1.71	g_1
	15	13	13	17	16	−0.11	g_1
g_2	16	11	15	18	17	−1.82	g_2
	17	10	13	16	9	−1.22	g_2
	18	11	14	24	16	−4.65	g_2
	19	10	10	13	12	−0.50	g_2
	20	10	14	22	18	−4.61	g_2
	21	13	19	23	24	−2.72	g_2
	22	11	10	20	28	−4.36	g_2
	23	15	20	20	16	0.91	g_1^*
	24	12	22	23	13	−2.10	g_2
	25	10	11	18	10	−2.51	g_2
	26	12	10	19	27	−3.10	g_2
	27	10	14	21	19	−4.23	g_2

*Misclassification

$$f(\mathbf{x}) = 0.719x_1 + 0.139x_2 - 0.462x_3 - 0.084x_4 - 2.069, \tag{15.31}$$

where we have set the prior probabilities in (15.24) as $P(g_1) = 15/27$ and $P(g_2) = 12/27$, i.e., the proportions of the members in groups 1 and 2 in Table 15.2.

Let us consider assessing how correctly/incorrectly individuals are classified by the LDF in (15.31). An easy way to do so is to substitute each row vector of $\mathbf{X}$ into (15.31) and examine whether the resulting discriminant score shows the *correct*

classification or not. For example, the substitution of the first and second row vectors in Table 15.2a yields

$$f\left(\mathbf{x}_{11} = [15, 14, 15, 14]'\right) = 0.719 \times 15 + 0.139 \times 14 - 0.462 \times 15 - 0.084 \times 14 - 2.069 = 2.56, \tag{15.32}$$

$$f\left(\mathbf{x}_{12} = [11, 13, 17, 17]'\right) = 0.719 \times 11 + 0.139 \times 13 - 0.462 \times 17 - 0.084 \times 17 - 2.069 = -1.64, \tag{15.33}$$

respectively. Here, (15.32) implies correct classification since it gives a positive value, showing that $\mathbf{x}_{11}$ is to be classified into Group 1, and in reality, $\mathbf{x}_{11}$ belongs to Group 1. On the other hand, (15.33) implies *misclassification*, since (15.33) is negative and shows that $\mathbf{x}_{12}$ is to be classified into Group 2, but the examinee 2 belongs to Group 1 in fact. The scores obtained as above are shown in Table 15.2b, with the asterisks indicating misclassification.

By counting those asterisks, we can assess *misclassification rates*; the rate is 4/15 in Group 1, while it is 1/12 for Group 2, and the total rate is (4 + 1)/27 = 0.185. This assessment is known to *underestimate* the *misclassification rate* since the classification is made for the data vectors from which LDFs are obtained. This differs from a usual setting, in which a *new* data vector $\mathbf{x}_?$ to be classified is *not included* in the data set $\mathbf{X}$. However, procedures for more accurately assessing the rate are out of the scope of this book.

LDA is used not only for classification but also for finding the variables that characterize groups. For this purpose, the *standardized discriminant coefficients* are to be used that are obtained with LDA for standardized data. The coefficients for the standard scores transformed from the data in Table 15.2a are presented in Table 15.3. There, we can find the following:

(1) The persons to be classified into Group 1 are social and cooperative, but not diligent and creative, with particularly important characteristics being social and less diligent.
(2) The persons to be classified into Group 2 are diligent and creative, but not social and cooperative, with important characteristics being diligent and less social.

Let us consider performing the *CDA* in earlier sections for the data set in Table 15.2a with $m = 1$. CDA provides $\mathbf{B} = \mathbf{b} = [0.226, 0.044, -0.145, -0.26]'$,

Table 15.3 Standardized discriminant coefficients for the data in Table 15.2a

Social	Cooperative	Diligent	Creative
2.079	0.704	−1.289	−0.459

every element of which equals the corresponding coefficient in (15.31) divided by 3.1. Indeed, it is known that the coefficients for CDA are proportional to those of LDA, and the classifications made by CDA with $m = 1$ are *equivalent* to those by LDA when $P(g_1) = P(g_2)$, though its proof is omitted here. The discriminant analysis procedure differing from LDA and CDA is described in the following section.

15.7 Generalized Normal Discrimination

In this section, the classification by (15.14) is illustrated for the cases where $\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_K$ are supposed to be *heterogeneous* among groups. We consider the data matrix (15.16) with $n = 150$, $p = 2$, $K = 3$, and the 150 individuals randomly sampled. Let the statistics obtained from $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$ be summarized as in Fig. 15.3a; for example, the average vector $40^{-1}\mathbf{1}'_{40}\mathbf{X}_2$ for Group 2 is [25.9, 74.8], and the covariance matrix $44^{-1}\mathbf{X}'_3\mathbf{J}\mathbf{X}_3$ for Group 3 is $\begin{bmatrix} 435.1 & 212.6 \\ 212.6 & 168.4 \end{bmatrix}$ with $\mathbf{J}$ the centering matrix defined as (2.10).

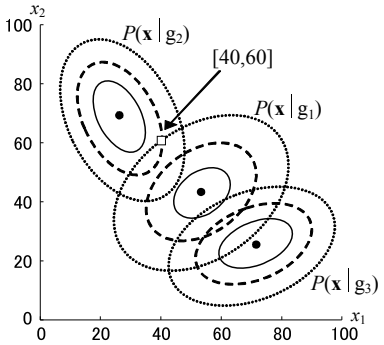
Prior probabilities can be estimated as $P(g_k) = n_k/n$:

$$P(g_1) = \frac{66}{150}, \quad P(g_2) = \frac{40}{150}, \text{ and } P(g_3) = \frac{44}{150} \tag{15.34}$$

for the data set in Fig. 15.3a. The *group-conditional density* is given as (15.18), whose parameters $\boldsymbol{\mu}_k$ and $\mathbf{\Sigma}_k$ can be estimated by the maximum likelihood method as described in Sect. 8.6 and illustrated in Sect. 8.8. The MLE of $\boldsymbol{\mu}_k$ and $\mathbf{\Sigma}_k$ is given by Eqs. (8.21) and (8.22) with the subscript k added as

Group	g_1		g_2		g_3	
n_k	66		40		44	
Variable	x_1	x_2	x_1	x_2	x_1	x_2
Average	52.1	43.3	25.9	74.8	71.0	23.4
Covariances	355.7	203.8	180.4	-198.8	435.1	212.6
	203.8	252.5	198.8	369.4	212.6	168.4

(a) Statistics for each group



(b) Group-conditional densities

Fig. 15.3 Statistics and probability densities for generalized normal discrimination

$$\hat{\boldsymbol{\mu}}_k = \bar{\mathbf{x}}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} \mathbf{x}_{ki}, \tag{15.35}$$

$$\hat{\boldsymbol{\Sigma}}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} (\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)(\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)'. \tag{15.36}$$

Let us substitute the statistics in Fig. 15.3a into the corresponding parts of (15.35) and (15.36). Using these results in (15.18), we have the group-conditional densities

$$P(\mathbf{x}|\mathrm{g}_1) = (2\pi)^{-p/2}\left|\begin{bmatrix} 355.7 & 203.8 \\ 203.8 & 252.5 \end{bmatrix}\right|^{-1/2} \times \exp\left\{-\frac{1}{2}\left(\mathbf{x} - \begin{bmatrix} 52.1 \\ 43.3 \end{bmatrix}\right)' \begin{bmatrix} 355.7 & 203.8 \\ 203.8 & 252.5 \end{bmatrix}^{-1} \left(\mathbf{x} - \begin{bmatrix} 52.1 \\ 43.3 \end{bmatrix}\right)\right\}, \tag{15.37}$$

$$P(\mathbf{x}|\mathrm{g}_2) = (2\pi)^{-p/2}\left|\begin{bmatrix} 180.4 & -198.8 \\ -198.8 & 369.4 \end{bmatrix}\right|^{-1/2} \times \exp\left\{-\frac{1}{2}\left(\mathbf{x} - \begin{bmatrix} 25.9 \\ 74.8 \end{bmatrix}\right)' \begin{bmatrix} 180.4 & -198.8 \\ -198.8 & 369.4 \end{bmatrix}^{-1} \left(\mathbf{x} - \begin{bmatrix} 25.9 \\ 74.8 \end{bmatrix}\right)\right\}, \tag{15.38}$$

$$P(\mathbf{x}|\mathrm{g}_3) = (2\pi)^{-p/2}\left|\begin{bmatrix} 435.1 & 212.6 \\ 212.6 & 166.4 \end{bmatrix}\right|^{-1/2} \times \exp\left\{-\frac{1}{2}\left(\mathbf{x} - \begin{bmatrix} 71.0 \\ 23.4 \end{bmatrix}\right)' \begin{bmatrix} 435.1 & 212.6 \\ 212.6 & 166.4 \end{bmatrix}^{-1} \left(\mathbf{x} - \begin{bmatrix} 71.0 \\ 23.4 \end{bmatrix}\right)\right\}, \tag{15.39}$$

with $\pi = 3.1416\ \ldots$ the circle ratio. In Fig. 15.3b, a bird's-eye view of (15.37)–(15.39) is drawn as in Fig. 8.4b. We may consider the figure as a map depicting three mountains whose tops are indicated by filled circles and counter lines are expressed by ellipses.

Let $\mathbf{x}_? = [40, 60]'$ indicated by a blank square in Fig. 15.3b be a new data vector for the individual whose membership to a group is *unknown*; our task is to classify $\mathbf{x}_?$ into one of groups 1, 2, and 3. This can be achieved by performing the calculus in the *Bayes' theorem* (15.15) and by using the *classification rule* (15.14).

By substituting $\mathbf{x}_? = [40, 60]'$ into (15.37), (15.38), and (15.39), we have the values of the group-conditional densities as $P\big([40, 60]'|\mathrm{g}_1\big) = 7.534 \times 10^{-5}$,

$P\big([40,60]'|\mathrm{g}_2\big) = 5.560 \times 10^{-4}$, and $P\big([40,60]'|\mathrm{g}_3\big) = 1.600 \times 10^{-13}$, respectively. Using these with (15.34), the numerator in the right-hand side of (15.15) is obtained as

$$P(\mathrm{g}_1)P\big([40,60]'|\mathrm{g}_1\big) = \frac{66}{150} \times 7.534 \times 10^{-5} = 3.315 \times 10^{-5}, \tag{15.40}$$

$$P(\mathrm{g}_2)P\big([40,60]'|\mathrm{g}_2\big) = \frac{40}{150} \times 5.560 \times 10^{-4} = 1.483 \times 10^{-4}, \tag{15.41}$$

$$P(\mathrm{g}_3)P\big([40,60]'|\mathrm{g}_3\big) = \frac{44}{150} \times 1.600 \times 10^{-13} = 4.693 \times 10^{-14} \tag{15.42}$$

for each group. Here, it should be noted that the denominator in the right-hand side of (15.15) is equivalent among different groups; we may *only compare* its *numerator* between groups for classification. This implies that (15.14) may be simplified to

$$\mathbf{x} \text{ is classified into group } k^* \text{ with } P(g_{k^*})P(\mathbf{x}|g_{k^*}) = \max_{1 \le k \le m} P(g_k)P(\mathbf{x}|g_k). \tag{15.43}$$

By this rule, we can compare (15.40), (15.41), and (15.42) to classify $\mathbf{x}_? = [40,60]'$ into Group 2 since (15.41) is the *highest* of the three values.

If we wish to perform not only the classification but also obtain the *posterior probability* of $\mathbf{x}_?$ belonging to the group, the denominator in the right-side hand of (15.15) must be obtained, which is the sum of $P(g_k)P(g_k|\mathbf{x})$ over k. The sum of (15.40)–(15.42) is given by

$$\begin{aligned}\sum_{l=1}^{m} P(g_l)P(\mathbf{x}|g_l) &= 3.315 \times 10^{-5} + 1.483 \times 10^{-4} + 4.693 \times 10^{-14} \\ &= 1.815 \times 10^{-4}.\end{aligned} \tag{15.44}$$

The use of this value and (15.41) in (15.15) leads to the posterior probability

$$P\big(g_2|[40,60]'\big) = \frac{1.483 \times 10^{-4}}{1.815 \times 10^{-4}} = 0.82. \tag{15.45}$$

Thus, the probability of $\mathbf{x}_?$ belonging to Group 2 is 0.82. This value can be regarded as expressing the *confidence* with which we classify $\mathbf{x}_?$ into g_2. In a parallel manner, the probability of $\mathbf{x}_?$ belonging to Group 1 can be obtained as $P\big(g_1|[40,60]'\big) = \frac{3.315 \times 10^{-5}}{1.815 \times 10^{-4}} = 0.18$, and $P\big(\mathrm{g}_3|[40,60]'\big) = 1 - P\big(g_1|[40,60]'\big) - P\big(g_2|[40,60]'\big)$ can be found to be almost zero.

15.8 Bibliographical Notes

A variety of discriminant analysis procedures are described in McLachlan (1992) and Hand (1997). Some new procedures in discriminant analysis are detailed in Hastie, Tibshirani, and Friedman (2009). An introduction to CDA as a modification of MCA is found in Adachi (2004).

Exercises

15.1 Matrices $\mathbf{V}_{\mathrm{B}} = \frac{1}{n}\left\{\sum_{k=1}^{K}\sum_{i=1}^{n_k}(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})'\right\} = \frac{1}{n}\left\{\sum_{k=1}^{K} n_k(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})'\right\}$, $\mathbf{V}_{\mathrm{W}} = \frac{1}{n}\left\{\sum_{k=1}^{K}\sum_{i=1}^{n_k}(\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)(\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)'\right\}$, and $\mathbf{V}_{\mathrm{T}} = \frac{1}{n}\left\{\sum_{k=1}^{K}\sum_{i=1}^{n_k}(\mathbf{x}_{ki} - \bar{\mathbf{x}})(\mathbf{x}_{ki} - \bar{\mathbf{x}})'\right\}$ are called *between-group*, *within-group*, and *total* covariance matrices, respectively, with $\mathbf{x}_{ki}$ the $p \times 1$ data vector for the ith individual in group k, $n = \sum_{i=1}^{K} n_k$, $\bar{\mathbf{x}}_k = n_k^{-1}\sum_{i=1}^{n_k}\mathbf{x}_{ki}$, and $\bar{\mathbf{x}} = n^{-1}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\mathbf{x}_{ki}$. Show $\mathbf{V}_{\mathrm{T}} = \mathbf{V}_{\mathrm{B}} + \mathbf{V}_{\mathrm{W}}$.

15.2 Let $\mathbf{x}_l'$ be the lth row vector of n individuals $\times$ p-variables data matrix $\mathbf{X} = \left[\mathbf{X}_1', \mathbf{X}_2'\right]' = [\mathbf{x}_1, \ldots, \mathbf{x}_n]'$ in (15.16) and $\mathbf{X}_{[l]}$ be the $(n - 1) \times p$ matrix obtained by removing $\mathbf{x}_l'$ from $\mathbf{X}$. In a *leaving-one-out* procedure, the following assessment is replicated over $l = 1, \ldots, n$: (15.23) and (15.24) are estimated with $\mathbf{X}_{[l]}$ and classification (15.25) with $\mathbf{x} = \mathbf{x}_l$ performed in order to assess whether the resulting classification is correct or not. It is known that misclassification rates are estimated better in the leaving-one-out procedure than in that illustrated in Sect. 15.6. Discuss why the rates are estimated better in the former procedure.

15.3 In *logistic discriminant analysis* for two groups, the posterior probability for Group 1 is expressed as $P(\mathrm{g}_1|\mathbf{x}) = \frac{1}{1+\exp(-\mathbf{b}'\mathbf{x}-c)}$, with $\mathbf{x}$ the vector containing observed variables, $\mathbf{b}$ the vector of coefficients, and $P(\mathrm{g}_2|\mathbf{x}) = 1 - P(\mathrm{g}_1|\mathbf{x})$. Discuss how the logistic and linear discriminant analyses are similar/different.

15.4 The *Mahalanobis distance* of $\mathbf{x}$ to group k is defined as $(\mathbf{x} - \boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k)$, with $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ the mean vector and covariance matrix for group k, respectively. Show that the classification rule (15.13), with its distances replaced by the Mahalanobis distances, is equivalent to the classification procedure in Sect. 15.7 with $P(g_k)$ and $|\boldsymbol{\Sigma}_k|$ constrained to be homogeneous among the groups.

15.5 Let us consider a case in which each element of the vector $\mathbf{x}$ in (15.15) takes either one or zero, and the jth element x_j takes one with probability θ_{jk} for $\mathbf{x}$ being included in group k. Show that if the elements of $\mathbf{x}$ are observed mutually independently, classification (15.14) is feasible using $P(\mathbf{x}|g_k) = \prod_{j=1}^{p}\theta_{jk}^{x_j}(1 - \theta_{jk})^{1-x_j}$ in (15.15).

15.6 Let us consider a variant of CDA with $\mathbf{G}$ unknown. This procedure is formulated as minimizing $\|\mathbf{XB} - \mathbf{GC}\|^2$ over $\mathbf{G} = (g_{ik})$, $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_K]'$ and

$\mathbf{B}$ subject to (7.1), (7.2), and (15.3). Show that the minimization can be attained by the following algorithm:

Step 1. Initialize $\mathbf{G}$ and obtain $\mathbf{V} = \mathbf{V}^{1/2}\mathbf{V}^{1/2}$ with $\mathbf{V} = n^{-1}\mathbf{X}'\mathbf{X}$.
Step 2. Obtain the EVD $\mathbf{X}'\mathbf{G}\mathbf{D}_{\mathrm{G}}^{-1}\mathbf{G}'\mathbf{X} = \mathbf{Q}\Theta^2\mathbf{Q}'$ to set $\mathbf{B} = \mathbf{V}^{-1/2}\mathbf{Q}_m$ with $\mathbf{D}_{\mathrm{G}} = \mathbf{G}'\mathbf{G}$.
Step 3. Obtain $\mathbf{C}$ by (15.4).
Step 4. Set $g_{ik} = 1$ if $\left\|\mathbf{x}_i'\mathbf{B} - \mathbf{c}_k'\right\|^2 = \min_{1 \le l \le K}\left\|\mathbf{x}_i'\mathbf{B} - \mathbf{c}_l'\right\|^2$ and $g_{ik} = 0$ otherwise, for $i = 1, \ldots, n$; $k = 1, \ldots, K$.
Step 5. Finish if convergence is reached; otherwise, go back to Step 2.

In Vichi and Kiers' (2001) *factorial K-means analysis* (*FKM*), (15.3) is replaced by $\mathbf{B}'\mathbf{B} = \mathbf{I}_m$.

15.7 There exists a *Bayesian method* for estimating parameters besides the least squares and maximum likelihood methods. In the Bayesian method, the fact is used that Bayes' theorem (15.15) can be generalized as

$$P(\boldsymbol{\theta}|\mathbf{X}) = \frac{P(\boldsymbol{\theta})P(\mathbf{X}|\boldsymbol{\theta})}{P(\mathbf{X})}. \tag{15.46}$$

Here, $\boldsymbol{\theta}$ is the vector containing parameters, while $\mathbf{X}$ is a data matrix, with $P(\boldsymbol{\theta})$ denoting the probability density function (PDF) of $\boldsymbol{\theta}$, $P(\mathbf{X})$ the PDF of $\mathbf{X}$ observed, $P(\mathbf{X}|\boldsymbol{\theta})$ the PDF of $\mathbf{X}$ for given $\boldsymbol{\theta}$, and $P(\boldsymbol{\theta}|\mathbf{X})$ the PDF of $\boldsymbol{\theta}$ for given $\mathbf{X}$. As found in (15.46), the parameters are also viewed as being randomly distributed in the Bayesian method. This method is formulated as the maximization of (15.46) over $\boldsymbol{\theta}$, or equivalently, maximizing $P(\boldsymbol{\theta})P(\mathbf{X}|\boldsymbol{\theta})$. Argue that $P(\boldsymbol{\theta})P(\mathbf{X}|\boldsymbol{\theta})$ is the product of the prior information for parameters and their likelihood.

15.8 A *penalized least squares method* for $n \times p$ data matrix $\mathbf{X}$ can be formulated as minimizing $\|\mathbf{X} - \mathbf{H}(\boldsymbol{\theta})\|^2 + \tau g(\boldsymbol{\theta})$ over parameter vector $\boldsymbol{\theta}$, with $\mathbf{H}(\boldsymbol{\theta})$ a function of $\boldsymbol{\theta}$ providing an $n \times p$ matrix, $g(\boldsymbol{\theta})$ a function of $\boldsymbol{\theta}$ giving a nonnegative scalar value, and τ a specified nonnegative scalar value. An example of the method is found in Exercise 4.11. Show that the Bayesian estimation method in Exercise 15.7 is equivalent to the penalized least squares one if $P(\mathbf{X}|\boldsymbol{\theta})$ in (15.46) takes the form of $P(\mathbf{X}|\boldsymbol{\theta}) = a \times \exp\left\{-b\|\mathbf{X} - \mathbf{H}(\boldsymbol{\theta})\|^2\right\}$ and τ is set to a certain value.

Chapter 16
Multidimensional Scaling

The keywords for describing *multidimensional scaling* (*MDS*) are the *coordinates* of objects, the *distances* between objects, and the corresponding *quasi-distances* observed as data. For example, let us suppose that the objects are cities such as London, Paris, and Amsterdam. Then, their *coordinates* are the locations of those cities on a map, which *define* the inter-city *distances*. We further suppose that the flight-times between those cities are observed as data, which are regarded as *quasi-distance data*, since they are approximately proportional to distances, but are *not equivalent* to them. The purpose of MDS is to *estimate the coordinates of objects*, i.e., their locations, *from quasi-distance data*; the coordinates are obtained so that their defined *distances* approximate *quasi-distance data*.

The origin of MDS can be found in Torgerson (1952). His approach is called *classical scaling*, which is equivalent to Gower's (1966) *principal coordinate analysis*. Those procedures are formulated with inter-object inner products rather than distances. Also, though they are not treated, their squares are considered in Takane, Young, and de Leeuw's (1977) procedure known as *alternate least squares scaling* (ALSCAL). In this chapter, only an MDS procedure is introduced in which *distances themselves* are considered and a computational technique called a *majorization algorithm* is used. This technique for MDS is rooted in de Leeuw (1977) and has been developed by Groenen (1993), Heiser (1991), and others.

16.1 Linking Coordinates to Quasi-distances

Let us use q_{ij} for the *observed quasi-distance* between objects i and j. Then, the data set of quasi-distances among n objects can be expressed as an $n \times n$ matrix

K. Adachi, *Matrix-Based Introduction to Multivariate Data Analysis*,
https://doi.org/10.1007/978-981-15-4103-2_16

$$\mathbf{Q} = \begin{bmatrix} q_{12} & q_{13} & \cdots & q_{1n} \\ & q_{23} & \cdots & q_{2n} \\ & & \ddots & \vdots \\ & & & q_{n-1,n} \end{bmatrix}. \tag{16.1}$$

Here, the lower left elements, i.e., the parts for q_{ij} with $i \geq j$, are blank, as it is supposed that $\mathbf{Q}$ is *symmetric* with $q_{ij} = q_{ji}$, and q_{ii} (the quasi-distance between an object and itself) is *not observed*. We also suppose $q_{ij} \geq 0$ in this chapter. One feature of data matrix $\mathbf{Q}$ is that the *same set of entities* occupies the *rows* and *columns*, which differs from the n-observations $\times$ p variables data matrices that have been treated in other chapters. Table 16.1 presents an example of $\mathbf{Q}$, which describes the perceived dissimilarities between objects (sports). They are quasi-distance data in that a pair of objects perceived *similar/dissimilar* corresponds to their being *close/distant*; however, the dissimilarities *differ* from the genuine *distances* defined mathematically as found in the next paragraph.

The purpose of MDS is to obtain an n-objects $\times$ m-dimensions matrix of the objects' *coordinates*,

$$\mathbf{A} = (a_{ik}) = \begin{bmatrix} \mathbf{a}_1' \\ \vdots \\ \mathbf{a}_n' \end{bmatrix}, \tag{16.2}$$

from (16.1). Here, $\mathbf{a}_i$ ($m \times 1$) is the *coordinate vector* indicating the location of object i, with the kth element a_{ik} the coordinate of i on dimension k. The *distance* between $\mathbf{a}_i$ and $\mathbf{a}_j$ is expressed as

$$\left\|\mathbf{a}_i - \mathbf{a}_j\right\| = \sqrt{\sum_{k=1}^{m} (a_{ik} - a_{jk})^2}. \tag{16.3}$$

Table 16.1 An example of $\mathbf{Q}$: rated dissimilarities between sports

	Baseball	Volleyball	Football	Tennis	Ping-pong	Basketball	Rugby	Softball
Baseball		5.6	5.0	4.6	4.4	5.4	6.0	1.2
Volleyball			5.4	4.4	4.2	3.0	5.4	5.4
Football				5.6	6.2	4.0	2.8	4.8
Tennis					2.0	5.8	6.4	4.2
Ping-pong						5.2	6.4	4.8
Basketball							4.6	5.2
Rugby								5.6
Softball								

This particular distance is called *Euclidean distance*, from the ancient Greek mathematician Euclid (or Eukleídēs in Greek), for distinguishing it from the other special definitions of distances.

Distance (16.3) can be linked with its *quasi*-version q_{ij} in (16.1) as

$$q_{ij} = \left\|\mathbf{a}_i - \mathbf{a}_j\right\| + e_{ij}, \tag{16.4}$$

with e_{ij} an error. Thus, MDS is formulated as minimizing the sum of squared errors, i.e.,

$$h(\mathbf{A}) = \sum_{i<j} \left(q_{ij} - \left\|\mathbf{a}_i - \mathbf{a}_j\right\|\right)^2, \tag{16.5}$$

is minimized over $\mathbf{A}$.

Note 16.1. Summation for $i < j$
The symbol $\sum_{i<j} x_{ij}$ stands for the summation of a set of x_{ij} that satisfies $i < j$. For example, let $\mathbf{X} = (x_{ij})$ be a 4×4 matrix, then $\sum_{i<j} x_{ij} = x_{12} + x_{13} + x_{14} + x_{23} + x_{24} + x_{34}$.

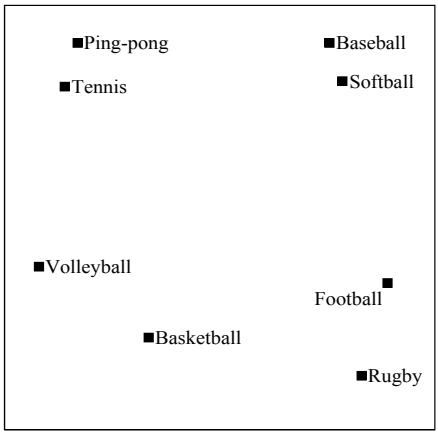
As found in (16.3), the distance is the *squared root* of $\left\|\mathbf{a}_i - \mathbf{a}_j\right\|^2$, which is far more *difficult to handle* than $\left\|\mathbf{a}_i - \mathbf{a}_j\right\|^2$. For dealing with that difficulty, some MDS procedures are formulated as fitting $\left\|\mathbf{a}_i - \mathbf{a}_j\right\|^2 = \|\mathbf{a}_i\|^2 + \left\|\mathbf{a}_j\right\|^2 - 2\mathbf{a}_i'\mathbf{a}_j$ to *squared* q_{ij} (Takane et al., 1977) or fitting *inner product* $\mathbf{a}_i'\mathbf{a}_j$ to the corresponding counterpart transformed from q_{ij} (Togerson, 1952; Gower, 1966), rather than minimizing (16.5). But, we will directly treat it in this chapter.

16.2 Illustration of an MDS Solution

For matrix $\mathbf{Q}$ in Table 16.1, MDS loss function (16.5) is minimized for the coordinate matrix $\mathbf{A}$ in Fig. 16.1A. This solution is graphically represented as in Fig. 16.1B, where the objects (sports) are plotted according to their *coordinates* in (A). We can see the plot as a usual *map*; the close/distant objects in the plot are similar/dissimilar in their features. For example, baseball and softball are closely located, which implies both are perceived to be similar, while rugby and ping-pong are distant, implying that they are dissimilar. This illustrates that we can *visually capture inter-objects relationships* in MDS solutions.

Dimension	1	2
Baseball	1.7	2.7
Volleyball	-2.8	-1.4
Football	2.6	-1.7
Tennis	-2.4	1.9
Ping-pong	-2.2	2.7
Basketball	-1.1	-2.7
Rugby	2.2	-3.4
Softball	1.9	2.0

(a) Coordinate Matrix



(b) Configuration

Fig. 16.1 MDS solution for the data in Table 16.1

The solution in Fig. 16.1 cannot explicitly be given. The iterative algorithm that provides the solution is described in the remaining sections.

16.3 Iterative Algorithm

Loss function (16.5) is expanded as $h(\mathbf{A}) = \sum_{i<j} q_{ij}^2 + \sum_{i<j} \left\|\mathbf{a}_i - \mathbf{a}_j\right\|^2 - 2\sum_{i<j} q_{ij}\left\|\mathbf{a}_i - \mathbf{a}_j\right\|$. Here, $\sum_{i<j} q_{ij}^2$ is a constant irrelevant to $\mathbf{A}$. Thus, the minimization of (16.5) is equivalent to minimizing

$$f(\mathbf{A}) = \sum_{i<j} \left\|\mathbf{a}_i - \mathbf{a}_j\right\|^2 - 2\sum_{i<j} q_{ij}\left\|\mathbf{a}_i - \mathbf{a}_j\right\|. \tag{16.6}$$

We will consider the latter.

Using $\mathbf{A}_{[t]}$ $(n \times m)$ for the coordinate matrix $\mathbf{A}$ obtained at the tth iteration, the outline of the iterative algorithm for minimizing (16.6) can be listed as follows:

Step 1. Initialize $\mathbf{A}_{[t]}$ with $t = 0$.
Step 2. Update $\mathbf{A}_{[t]}$ to $\mathbf{A}_{[t+1]}$ so that $f\left(\mathbf{A}_{[t]}\right) \geq f\left(\mathbf{A}_{[t+1]}\right)$.
Step 3. Finish if convergence is reached; otherwise, increase t by one and return to Step 2.

In Step 3, the convergence can be defined as $f\left(\mathbf{A}_{[t]}\right) - f\left(\mathbf{A}_{[t+1]}\right)$ is small enough to be ignored.

The update formula in Step 2 is given by

$$\mathbf{A}_{[t+1]} = \frac{1}{n}\mathbf{Q}_{(\mathbf{A}_{[t]})}\mathbf{A}_{[t]}. \tag{16.7}$$

Here, $\mathbf{Q}_{(\mathbf{A}_{[t]})}$ is the $n \times n$ matrix which is a function of $\mathbf{A}_{[t]}$ and is expressed as

$$\mathbf{Q}_{(\mathbf{A}_{[t]})} = \begin{bmatrix} \sum_{i=1}^{n} q_{i1}^{(\mathbf{A}_{[t]})} & & \\ & \ddots & \\ & & \sum_{i=1}^{n} q_{in}^{(\mathbf{A}_{[t]})} \end{bmatrix} - \begin{bmatrix} q_{11}^{(\mathbf{A}_{[t]})} & \cdots & q_{1n}^{(\mathbf{A}_{[t]})} \\ \vdots & \cdots & \vdots \\ q_{n1}^{(\mathbf{A}_{[t]})} & \cdots & q_{nn}^{(\mathbf{A}_{[t]})} \end{bmatrix}, \tag{16.8}$$

with the blanks standing for zero elements and $q_{ij}^{(\mathbf{A}_{[t]})}$ defined, using $\mathbf{a}_i^{[t]\prime}(1 \times m)$ for the ith row of $\mathbf{A}_{[t]}$, as

$$q_{ij}^{(\mathbf{A}_{[t]})} = \begin{cases} 0 & \text{if } \mathbf{a}_i^{[t]} = \mathbf{a}_j^{[t]} \\ \frac{q_{ij}}{\left\|\mathbf{a}_i^{[t]} - \mathbf{a}_j^{[t]}\right\|} & \text{otherwise} \end{cases}. \tag{16.9}$$

Why does (16.7) guarantee $f\left(\mathbf{A}_{[t]}\right) \geq f\left(\mathbf{A}_{[t+1]}\right)$? In order to explain this, we need a *long story* continuing over the next three sections. There, the following tasks are attained in turn:

(1) $\sum_{i<j}\left\|\mathbf{a}_i - \mathbf{a}_j\right\|^2$ in (16.6) is expressed in matrix form (Sect. 16.4).
(2) An inequality for $\sum_{i<j} q_{ij}\left\|\mathbf{a}_i - \mathbf{a}_j\right\|$ in (16.6) is derived (Sect. 16.5).
(3) We use the results of (1) and (2) to derive (16.7) (Sect. 16.6).

16.4 Matrix Expression for Squared Distances

In order to express *squared distance* $\left\|\mathbf{a}_i - \mathbf{a}_j\right\|^2$ in matrix form using $\mathbf{A}$, we introduce the elementary vectors in the following note:

Note 16.2. Elementary Vectors
Let $\mathbf{e}_i$ denote the $n \times 1$ vector filled with zeros, except only the ith element taking one. Such a vector is called an *elementary vector*. For example, $\mathbf{e}_2 = [0, 1, 0]'$ for $n = 3$. We can easily find that $\mathbf{e}_i'\mathbf{A} = \mathbf{a}_i'$ with $\mathbf{A}$ defined as (16.2); $\mathbf{e}_i'$ serves for selecting the ith row of a matrix.

Let $\mathbf{B} = \begin{bmatrix} \mathbf{b}_1' \\ \vdots \\ \mathbf{b}_n' \end{bmatrix}$ be an $n \times m$ matrix, like $\mathbf{A}$. Then, we have

$$\begin{aligned}(\mathbf{a}_i - \mathbf{a}_j)'(\mathbf{b}_i - \mathbf{b}_j) &= (\mathbf{e}_i'\mathbf{A} - \mathbf{e}_j'\mathbf{A})(\mathbf{e}_i'\mathbf{B} - \mathbf{e}_j'\mathbf{B})' = (\mathbf{e}_i - \mathbf{e}_j)'\mathbf{A}\mathbf{B}'(\mathbf{e}_i - \mathbf{e}_j) \\ &= \mathrm{tr}(\mathbf{e}_i - \mathbf{e}_j)'\mathbf{A}\mathbf{B}'(\mathbf{e}_i - \mathbf{e}_j) = \mathrm{tr}\mathbf{B}'(\mathbf{e}_i - \mathbf{e}_j)(\mathbf{e}_i - \mathbf{e}_j)'\mathbf{A} \\ &= \mathrm{tr}\mathbf{A}'\mathbf{H}_{ij}\mathbf{B},\end{aligned} \tag{16.10}$$

with

$$\mathbf{H}_{ij} = (\mathbf{e}_i - \mathbf{e}_j)(\mathbf{e}_i - \mathbf{e}_j)'. \tag{16.11}$$

For example, when $n = 3$,

$$\mathbf{H}_{12} = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \mathbf{H}_{13} = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{bmatrix}, \mathbf{H}_{23} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & -1 & 1 \end{bmatrix}. \tag{16.12}$$

If $\mathbf{B}$ is set to $\mathbf{A}$ in (16.10), we have the *squared distance*

$$\left\|\mathbf{a}_i - \mathbf{a}_j\right\|^2 = \mathrm{tr}\mathbf{A}'\mathbf{H}_{ij}\mathbf{A}. \tag{16.13}$$

This summation over $i < j$ is expressed in a simple form, using the following result:

Note 16.3. Use of Centering Matrix
It can be found that

$$\sum_{i<j} \mathbf{H}_{ij} = n\mathbf{I}_n - \mathbf{1}_n\mathbf{1}_n' = n\mathbf{J}, \tag{16.14}$$

with $\mathbf{J} = \mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n'$ the *centering matrix* defined in (2.10). Using (16.12), we can verify (16.14) as

$$\sum_{i<j}\mathbf{H}_{ij}=\mathbf{H}_{12}+\mathbf{H}_{13}+\mathbf{H}_{23}=\begin{bmatrix}2&-1&-1\\-1&2&-1\\-1&-1&2\end{bmatrix}$$
$$=3\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}-\begin{bmatrix}1&1&1\\1&1&1\\1&1&1\end{bmatrix}. \quad (16.15)$$

Using (16.13) and (16.14), we can rewrite $\sum_{i<j}\left\|\mathbf{a}_i-\mathbf{a}_j\right\|^2$ in (16.6) as

$$\sum_{i<j}\left\|\mathbf{a}_i-\mathbf{a}_j\right\|^2=\sum_{i<j}\mathrm{tr}\mathbf{A}'\mathbf{H}_{ij}\mathbf{A}=\mathrm{tr}\mathbf{A}'\sum_{i<j}\mathbf{H}_{ij}\mathbf{A}=n\mathrm{tr}\mathbf{A}'\mathbf{J}\mathbf{A}. \quad (16.16)$$

16.5 Inequality for Distances

This section concerns the term $\sum_{i<j}q_{ij}\left\|\mathbf{a}_i-\mathbf{a}_j\right\|$ in (16.6). The *distance* $\left\|\mathbf{a}_i-\mathbf{a}_j\right\|$ in that term is more difficult to handle than $\left\|\mathbf{a}_i-\mathbf{a}_j\right\|^2$. This difficulty can be dealt with by finding an *inequality* for $\sum_{i<j}q_{ij}\left\|\mathbf{a}_i-\mathbf{a}_j\right\|$ and $\sum_{i<j}q_{ij}(\mathbf{a}_i-\mathbf{a}_j)'(\mathbf{b}_i-\mathbf{b}_j)$, with $\mathbf{b}_i'$ being a row vector of $\mathbf{B}$ defined in Note 16.2. The first step for that task is using the following famous theorem:

Note 16.4. The Cauchy-Schwarz Inequality

$$\|\mathbf{x}\|\times\|\mathbf{y}\|\geq\mathbf{x}'\mathbf{y}. \quad (16.17)$$

Setting $\mathbf{x}=\mathbf{a}_i-\mathbf{a}_j$ and $\mathbf{y}=\mathbf{b}_i-\mathbf{b}_j$ in (16.17) and using $q_{ij}\geq 0$, we have

$$q_{ij}\left\|\mathbf{a}_i-\mathbf{a}_j\right\|\times\left\|\mathbf{b}_i-\mathbf{b}_j\right\|\geq q_{ij}(\mathbf{a}_i-\mathbf{a}_j)'(\mathbf{b}_i-\mathbf{b}_j), \quad (16.18)$$

which leads to

$$q_{ij}\left\|\mathbf{a}_i-\mathbf{a}_j\right\|\geq q_{ij}^{(\mathbf{B})}(\mathbf{a}_i-\mathbf{a}_j)'(\mathbf{b}_i-\mathbf{b}_j) \quad (16.19)$$

with

$$q_{ij}^{(\mathbf{B})} = \begin{cases} 0 & \text{if } \mathbf{b}_i = \mathbf{b}_j \\ \frac{q_{ij}}{\|\mathbf{b}_i - \mathbf{b}_j\|} & \text{otherwise} \end{cases}. \tag{16.20}$$

Here, it has been taken into consideration that the division by $\|\mathbf{b}_i - \mathbf{b}_j\| = 0$ cannot be defined, and $q_{ij}^{(\mathbf{B})}$ has the superscript "(**B**)" because $q_{ij}^{(\mathbf{B})}$ is a function of the rows of **B**.

We can use (16.10) to rewrite the right-hand side of (16.19) as

$$q_{ij}^{(\mathbf{B})}(\mathbf{a}_i - \mathbf{a}_j)'(\mathbf{b}_i - \mathbf{b}_j) = \mathrm{tr}\mathbf{A}'(q_{ij}^{(\mathbf{B})}\mathbf{H}_{ij})\mathbf{B}. \tag{16.21}$$

The left-hand side of (16.19) can be rewritten as

$$q_{ij}\|\mathbf{a}_i - \mathbf{a}_j\| = \begin{cases} 0 & \text{if } \mathbf{a}_i = \mathbf{a}_j \\ \frac{q_{ij}}{\|\mathbf{a}_i - \mathbf{a}_j\|}(\mathbf{a}_i - \mathbf{a}_j)'(\mathbf{a}_i - \mathbf{a}_j) & \text{otherwise} \end{cases}. \tag{16.22}$$

Its comparison with (16.20) allows us to find that (16.22) is further rewritten as

$$q_{ij}\|\mathbf{a}_i - \mathbf{a}_j\| = q_{ij}^{(\mathbf{A})}(\mathbf{a}_i - \mathbf{a}_j)'(\mathbf{a}_i - \mathbf{a}_j) = \mathrm{tr}\mathbf{A}'(q_{ij}^{(\mathbf{A})}\mathbf{H}_{ij})\mathbf{A}, \tag{16.23}$$

where $q_{ij}^{(\mathbf{A})}$ is defined by substituting $\mathbf{a}_i$ for $\mathbf{b}_i$ in (16.20), and we have also used (16.10).

The summation of both sides of (16.19) leads to

$$\sum_{i<j} q_{ij}\|\mathbf{a}_i - \mathbf{a}_j\| \geq \sum_{i<j} q_{ij}^{(\mathbf{B})}(\mathbf{a}_i - \mathbf{a}_i)'(\mathbf{b}_i - \mathbf{b}_j). \tag{16.24}$$

Here, we can use (16.21) and (16.23) to rewrite the left- and right-hand sides of (16.24) as

$$\sum_{i<j} q_{ij}\|\mathbf{a}_i - \mathbf{a}_j\| = \mathrm{tr}\mathbf{A}'\mathbf{Q}_{(\mathbf{A})}\mathbf{A}, \tag{16.25}$$

$$\sum_{i<j} q_{ij}^{(\mathbf{B})}(\mathbf{a}_i - \mathbf{a}_i)'(\mathbf{b}_i - \mathbf{b}_j) = \mathrm{tr}\mathbf{A}'\mathbf{Q}_{(\mathbf{B})}\mathbf{B}, \tag{16.26}$$

respectively, with

$$\mathbf{Q}_{(\mathbf{A})} = \sum_{i<j} q_{ij}^{(\mathbf{A})} \mathbf{H}_{ij} \text{ and } \mathbf{Q}_{(\mathbf{B})} = \sum_{i<j} q_{ij}^{(\mathbf{B})} \mathbf{H}_{ij}. \tag{16.27}$$

Thus, (16.24) is rewritten as

$$\mathrm{tr}\mathbf{A}'\mathbf{Q}_{(\mathbf{A})}\mathbf{A} \geq \mathrm{tr}\mathbf{A}'\mathbf{Q}_{(\mathbf{B})}\mathbf{B}, \tag{16.28}$$

which allows us to form the MDS algorithm described in the following section.

16.6 Majorization Algorithm

Using (16.16) and (16.25), MDS loss function (16.6) is rewritten as

$$f(\mathbf{A}) = n\mathrm{tr}\mathbf{A}'\mathbf{J}\mathbf{A} - 2\mathrm{tr}\mathbf{A}'\mathbf{Q}_{(\mathbf{A})}\mathbf{A}. \tag{16.29}$$

We also consider another function in which $\mathrm{tr}\mathbf{A}'\mathbf{Q}_{(\mathbf{A})}\mathbf{A}$ in (16.29) is replaced by (16.26):

$$g(\mathbf{A}, \mathbf{B}) = n\mathrm{tr}\mathbf{A}'\mathbf{J}\mathbf{A} - 2\mathrm{tr}\mathbf{A}'\mathbf{Q}_{(\mathbf{B})}\mathbf{B}. \tag{16.30}$$

By comparing (16.29) and (16.30) with (16.28), we can find

$$g(\mathbf{A}, \mathbf{B}) \geq f(\mathbf{A}). \tag{16.31}$$

Also, it should be noted that the substitution of $\mathbf{B}$ for $\mathbf{A}$ in (16.29) and (16.30) gives

$$g(\mathbf{B}, \mathbf{B}) = f(\mathbf{B}). \tag{16.32}$$

This equality and the inequality (16.31) lead to:

$$f(\mathbf{B}) = g(\mathbf{B}, \mathbf{B}) \geq g(\mathbf{A}^*, \mathbf{B}) \geq f(\mathbf{A}^*), \tag{16.33}$$

where $\mathbf{A}^*$ is the matrix $\mathbf{A}$ that minimizes $g(\mathbf{A}, \mathbf{B})$ for a given $\mathbf{B}$.

For finding $\mathbf{A}^*$, we use the fact that $\mathbf{Q}_{(\mathbf{B})}$ in (16.27) satisfies

$$\mathbf{J}\mathbf{Q}_{(\mathbf{B})} = \mathbf{Q}_{(\mathbf{B})} \text{ or } \sum_{i<j} q_{ij}^{(\mathbf{B})} \mathbf{J}\mathbf{H}_{ij} = \sum_{i<j} q_{ij}^{(\mathbf{B})} \mathbf{H}_{ij}. \tag{16.34}$$

This follows from the fact that (16.11) implies $\mathbf{1}_n'\mathbf{H}_{ij} = \mathbf{0}_n'$ and this is equivalent to $\mathbf{J}\mathbf{H}_{ij} = \mathbf{H}_{ij}$ since of (3.21). Using (16.34), we can rewrite (16.30) as

$$\begin{aligned} g(\mathbf{A}, \mathbf{B}) &= n\mathrm{tr}\mathbf{A}'\mathbf{J}\mathbf{A} - 2\mathrm{tr}\mathbf{A}'\mathbf{J}\mathbf{Q}_{(\mathbf{B})}\mathbf{B} \\ &= \left\| \sqrt{n}\,\mathbf{J}\mathbf{A} - \frac{1}{\sqrt{n}}\mathbf{Q}_{(\mathbf{B})}\mathbf{B} \right\|^2 - \frac{1}{n}\mathrm{tr}\mathbf{B}'\mathbf{Q}_{(\mathbf{B})}'\mathbf{Q}_{(\mathbf{B})}\mathbf{B}, \end{aligned} \tag{16.35}$$

because of (2.11) and (2.12). Given $\mathbf{B}$, (16.35) is minimized over $\mathbf{A}$ for

$$\sqrt{n}\mathbf{J}\mathbf{A} = \frac{1}{\sqrt{n}}\mathbf{Q}_{(\mathbf{B})}\mathbf{B}. \tag{16.36}$$

Here, we can suppose $\mathbf{A} = \mathbf{J}\mathbf{A}$; equivalently, $n^{-1}\mathbf{1}_n'\mathbf{A} = \mathbf{0}_n'$, as the *center of coordinates* $n^{-1}\mathbf{1}_n'\mathbf{A}$ *may be anywhere*; thus, we can set it to the *origin*. This allows (16.36) to be rewritten as $\mathbf{A} = n^{-1}\mathbf{Q}_{(\mathbf{B})}\mathbf{B}$. That is, when

$$\mathbf{A}^* = \mathbf{J}\mathbf{A}^* = \frac{1}{n}\mathbf{Q}_{(\mathbf{B})}\mathbf{B}, \tag{16.37}$$

(16.33) holds true. By setting $\mathbf{A}^* = \mathbf{A}_{[t+1]}$ and $\mathbf{B} = \mathbf{A}_{[t]}$ in (16.33) and (16.37), respectively, we have $f\left(\mathbf{A}_{[t]}\right) = g(\mathbf{A}_{[t]}, \mathbf{A}_{[t]}) \geq g\left(\mathbf{A}_{[t+1]}, \mathbf{A}_{[t]}\right) \geq f\left(\mathbf{A}_{[t+1]}\right)$, i.e., $f\left(\mathbf{A}_{[t]}\right) \geq f\left(\mathbf{A}_{[t+1]}\right)$, and the update Formula (16.7) for the coordinate matrix $\mathbf{A}$ to be obtained in MDS.

One feature of the derived algorithm is using an *auxiliary* function $g(\mathbf{A}, \mathbf{B})$ beside $f(\mathbf{A})$. The auxiliary function $g(\mathbf{A}, \mathbf{B})$ is called a *majorizing function*, as it majorizes $f(\mathbf{A})$ with (16.31). Algorithms with such majorizing functions are called *majorization algorithms*, and they are included in auxiliary function algorithms, as described in Appendix A.6.1.

16.7 Bibliographical Notes

Multidimensional scaling is detailed in Borg and Groenen (2005) and Cox and Cox (2000). A book-length description of majorization algorithms is found in Groenen (1993). Applications of MDS are intelligibly illustrated in Borg, Groenen, and Mair (2013).

Though quasi-distance q_{ij} is restricted to a nonnegative value in this chapter, Heiser (1991) generalized the algorithm so that it is feasible for q_{ij} being negative.

Exercises

16.1. Let $\mathbf{D}^{(2)}$ be the $n \times n$ matrix whose (i, j) element is the squared distance $\left\|\mathbf{a}_i - \mathbf{a}_j\right\|^2$ between the ith and jth rows of $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_n]'$. Show that $\mathbf{D}^{(2)} = \mathbf{1}_n\mathbf{1}_n'\text{diag}(\mathbf{AA}') - 2\mathbf{AA}' + \text{diag}(\mathbf{AA}')\mathbf{1}_n\mathbf{1}_n'$, where $\text{diag}(\mathbf{AA}')$ denotes the $n \times n$ diagonal matrix whose diagonal elements are those of $\mathbf{AA}'$, as defined in Note 12.1.

16.2. Show that $-2^{-1}\mathbf{JD}^{(2)}\mathbf{J} = \mathbf{AA}'$, subject to $\mathbf{A} = \mathbf{JA}$, with $\mathbf{J} = \mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n'$ and $\mathbf{D}^{(2)}$ defined in Exercise 16.1, and discuss the rationale of minimizing $\left\|-2^{-1}\mathbf{JQJ} - \mathbf{AA}'\right\|^2$ over $\mathbf{A}$ (Gower, 1966; Torgerson, 1952).

16.3. It is known that the differentiation of (16.3) with respect to a_{ij} is proportional to $\left\|\mathbf{a}_i - \mathbf{a}_j\right\|^{-1}$, which implies that an algorithm using the differentiation for MDS fails when $\mathbf{a}_i = \mathbf{a}_j$ arises. Show that the majorization algorithm in this chapter does not fail for $\mathbf{a}_i = \mathbf{a}_j$.

16.4. Show that the MDS solution minimizing (16.6) can be rotated.

16.5. Let $\mathbf{Q}_s = (q_{sij})$ be an $n \times n$ quasi-distance data matrix obtained from source $s = 1, \ldots, S$, with q_{sij} the (i, j) element of $\mathbf{Q}_s$. In an extended version of MDS for $\mathbf{Q}_1, \ldots, \mathbf{Q}_S$, the loss function is defined as $\sum_{s=1}^{S}\sum_{i<j}(q_{sij} - d_{sij})^2$, with d_{sij} the *weighted Euclidean distance* defined as

$$d_{sij} = \sqrt{\sum_{k=1}^{m} w_{sk}^2(a_{ik} - a_{jk})^2}. \tag{16.38}$$

The above loss function is minimized over $\mathbf{A}$ and w_{sk} $(s = 1, \ldots, S;$ $k = 1, \ldots, m)$ subject to a certain constraint on $\mathbf{A}$. Here, $\mathbf{A} = (a_{ik})$ does not have subscript s, while w_{sk} does, implying that w_{sk} serves to explain the differences of $\mathbf{Q}_s$ across sources $s = 1, \ldots, S$. Discuss how w_{sk} explains those differences.

16.6. Show that (16.38) is rewritten as $\{(\mathbf{a}_i - \mathbf{a}_j)'\mathbf{W}_s^2(\mathbf{a}_i - \mathbf{a}_j)\}^{1/2} = \left\|\mathbf{W}_s\mathbf{a}_i - \mathbf{W}_s\mathbf{a}_j\right\| = \left\|\mathbf{W}_s(\mathbf{a}_i - \mathbf{a}_j)\right\|$ with $\mathbf{W}_s = \begin{bmatrix} w_{s1} & & \\ & \ddots & \\ & & w_{sm} \end{bmatrix}$ an $m \times m$ diagonal matrix.

16.7. Show that $\mathbf{A}$ cannot be rotated in the extended MDS considered in Exercise 16.5, except for special cases.

16.8. Distance (16.38) can be rewritten as $d_{sij} = \sqrt{\sum_{k=1}^{m}\frac{w_{sk}^2}{c_k^2}(c_k a_{ik} - c_k a_{jk})^2}$. Show that the solution is not unique without a constraint on $\mathbf{A}$ and the solution can be determined uniquely by constraining each column of $\mathbf{A}$ to be standardized.

16.9. Show that

$$\mathbf{D}^{(2)}_{(\mathbf{F},\mathbf{C})} = \mathrm{diag}(\mathbf{F}\mathbf{F}')\mathbf{1}_n\mathbf{1}_p' - 2\mathbf{F}\mathbf{C}' + \mathbf{1}_n\mathbf{1}_p'\mathrm{diag}(\mathbf{C}\mathbf{C}') \tag{16.39}$$

expresses the $n \times p$ matrix whose (i, j) element is the squared distance between the ith row of $\mathbf{F}$ ($n \times m$) and the jth row of $\mathbf{C}$ ($p \times m$). Here, $\mathrm{diag}(\mathbf{F}\mathbf{F}')$ is defined as in Note 12.1.

16.10. Let us consider approximating $n \times p$ data matrix $\mathbf{X} = (x_{ij})$ by $\mathbf{D}_{(\mathbf{F},\,\mathbf{C})}$, whose elements are the square roots of the corresponding ones in (16.39), i.e., minimizing $\left\|\mathbf{X} - \mathbf{D}_{(\mathbf{F},\mathbf{C})}\right\|^2$ over $\mathbf{F}$ and $\mathbf{C}$, with $m \leq \min(n, p)$. Discuss for what types of $\mathbf{X}$ the above minimization is useful.

Part V
Advanced Procedures

In this part, we start with advanced matrix operations (Chap. 17) as a preparation for the chapters that follow. The matrix decomposition (MD) formulation of exploratory factor analysis (EFA) is introduced in Chap. 18. This formulation allows us to directly contrast the solution of EFA with that of principal component analysis (PCA), as described in Chap. 19. Three-way PCA, which is specially designed for three-way data, is treated in Chap. 20. Finally, in Chaps. 21 and 22, sparse multivariate analysis procedures are introduced, in which sparse solutions are estimated. These refer to solutions including a number of zeros. Such sparse approaches originate in regression analysis as discussed in Chap. 21. Furthermore, the factor analysis (FA) version can deal with the difficulties present in confirmatory FA (Chap. 10), as explained in Chap. 22.

Chapter 17
Advanced Matrix Operations

In this chapter we introduce matrix operations that are more advanced than those treated so far. We start by describing systems of linear equations, and then introduce the *Moore–Penrose* (*MP*) *inverse*, considered as one of the most important operations for statistics, as well as *singular value decomposition* (*SVD*). The MP inverse is closely related to SVD and more useful than the ordinary inverse matrix, which is regarded as a special case of the MP inverse. The MP inverse allows the least squares problems to be generally formulated and is a bridge to *orthogonal complement matrices*. Finally, we introduce other classes of matrix operations; the *Kronecker product*, *Khatri–Rao product*, *vec operator*, and *Hadamard product*.

17.1 Introductory Systems of Linear Equations

A set of equations such as

$$\begin{cases} 3a - 5b + 9c = 7 \\ -a + 6b - 7c = 1 \\ 4a + 7b + c = -5 \end{cases} \tag{17.1}$$

is called a *system of linear equations*. The solving of (17.1) equates to obtaining values of a, b, and c that satisfy all equations. This problem can be easily solved using a matrix and vectors: by defining $\mathbf{X} = \begin{bmatrix} 3 & -5 & 9 \\ -1 & 6 & -7 \\ 4 & 7 & 1 \end{bmatrix}$, $\mathbf{b} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$, $\mathbf{y} = \begin{bmatrix} 7 \\ 1 \\ -5 \end{bmatrix}$, (17.1) is rewritten as

K. Adachi, *Matrix-Based Introduction to Multivariate Data Analysis*,
https://doi.org/10.1007/978-981-15-4103-2_17

$$\mathbf{Xb} = \mathbf{y} \tag{17.2}$$

and the solution of $\mathbf{b}$ is given by $\mathbf{b} = \mathbf{X}^{-1}\mathbf{Xb} = \mathbf{X}^{-1}\mathbf{y}$, i.e.,

$$\mathbf{b} = \mathbf{X}^{-1}\mathbf{y}. \tag{17.3}$$

Since the inverse of $\mathbf{X}$ can be found to be $\mathbf{X}^{-1} = \begin{bmatrix} 2.62 & 3.24 & -0.90 \\ -1.29 & -1.57 & 0.57 \\ -1.48 & -1.95 & 0.62 \end{bmatrix}$, the solution of $\mathbf{b}$ is given by $\mathbf{b} = \mathbf{X}^{-1}\mathbf{y} = \begin{bmatrix} 2.62 & 3.24 & -0.90 \\ -1.29 & -1.57 & 0.57 \\ -1.48 & -1.95 & 0.62 \end{bmatrix} \begin{bmatrix} 7 \\ 1 \\ -5 \end{bmatrix} = \begin{bmatrix} 26.10 \\ -13.43 \\ -15.38 \end{bmatrix}$.

Any system of linear equations can be expressed in the form of (17.2). Thus, we define $\mathbf{X}$, $\mathbf{b}$, and $\mathbf{y}$ as an $n \times p$ matrix, $p \times 1$ vector, and $n \times 1$ vector, respectively. In the last example, $n = p$ and the existence of $\mathbf{X}^{-1}$ have been supposed. In the next section, however, they are not.

17.2 Moore–Penrose Inverse and System of Linear Equations

This example of a system of linear equations: $\begin{cases} 3a - 5b + 9c = 7 \\ -a + 6b - 7c = 1 \\ 4a + 7b + c = -5 \\ 2a - 8b + 3c = 6 \end{cases}$, with an equation added to (17.1), does *not have a solution*, i.e., no vector $\mathbf{b} = [a, b, c]'$ exists that satisfies those four equations. On the other hand, the system

$$\begin{cases} 3a - 5b + 9c = 7 \\ -a + 6b - 7c = 1 \end{cases}, \tag{17.4}$$

with one equation deleted from (17.1) has *multiple solutions*, i.e., the vector $\mathbf{b} = [a, b, c]'$ satisfying (17.4) is *not unique*, in contrast to the unique solution for (17.1). These examples show that we must consider whether a system has a solution or not, and that we must consider how a solution is expressed if it exists. To consider them, the *Moore–Penrose* (*MP*) *inverse* matrix can be useful.

Note 17.1. The Moore–Penrose (MP) Inverse
For any $n \times p$ matrix $\mathbf{X}$, the $p \times n$ matrix $\mathbf{X}^+$ satisfying

$$\mathbf{X}\mathbf{X}^+\mathbf{X} = \mathbf{X},\ \mathbf{X}^+\mathbf{X}\mathbf{X}^+ = \mathbf{X}^+,\ (\mathbf{X}\mathbf{X}^+)' = \mathbf{X}\mathbf{X}^+,\ \text{and}\ (\mathbf{X}^+\mathbf{X})' = \mathbf{X}^+\mathbf{X} \tag{17.5}$$

can be uniquely determined. The matrix $\mathbf{X}^+$ is called the *Moore–Penrose* (*MP*) *inverse* of $\mathbf{X}$.

If $\mathbf{X}$ is nonsingular with $n = p$ and $\text{rank}(\mathbf{X}) = n$, $\mathbf{X}^+$ equals the inverse matrix $\mathbf{X}^{-1}$ for $\mathbf{X}$ introduced in Note 4.2:

$$\mathbf{X}^{-1} = \mathbf{X}^+ \quad \text{if } \mathbf{X} \text{ is nonsingular.} \tag{17.6}$$

This implies that the *inverse matrix* is a *special case* of the MP inverse.

The MP inverse of $\mathbf{X}$ has the following properties:

$$(\mathbf{X}^+)^+ = \mathbf{X} \ \text{ and } \ (\mathbf{X}^+)' = (\mathbf{X}')^+. \tag{17.7}$$

Further, it holds

$$\mathbf{X}^+ = \mathbf{X}' \quad \text{if} \quad \mathbf{X}'\mathbf{X} = \mathbf{I}_p. \tag{17.8}$$

Let us consider the system of equations, $\mathbf{X}\mathbf{b} = \mathbf{y}$, with $\mathbf{b}$ $(p \times 1)$ unknown but $\mathbf{X}$ $(n \times p)$ and $\mathbf{y}$ $(n \times 1)$ given. Using the MP inverse, the *existence of the solutions* can be shown as follows:

Note 17.2. On the Existence of Solutions
The following three statements are known to be equivalent (e.g., Schott, 2015, Sect. 6.2; Seber, 2008, Sect. 13.1.1):

$$\text{The equation } \mathbf{X}\mathbf{b} = \mathbf{y} \text{ for given } \mathbf{X} \text{ and } \mathbf{y} \text{ has a solution of } \mathbf{b}. \tag{17.9}$$

$$\mathbf{X}\mathbf{X}^+\mathbf{y} = \mathbf{y}. \tag{17.10}$$

$$\text{rank}([\mathbf{X}, \mathbf{y}]) = \text{rank}(\mathbf{X}). \tag{17.11}$$

If (17.9), (17.10), or (17.11) holds, all solutions of $\mathbf{X}\mathbf{b} = \mathbf{y}$ can be expressed as

$$\mathbf{b} = \mathbf{X}^+\mathbf{y} + (\mathbf{I}_p - \mathbf{X}^+\mathbf{X})\mathbf{q}, \tag{17.12}$$

with $\mathbf{q}$ an *arbitrary* $p \times 1$ vector. This fact is proved as follows. We can find (17.12) $\Rightarrow \mathbf{Xb} = \mathbf{y}$ by substituting (17.12) in $\mathbf{Xb}$; this leads to $\mathbf{X}\{\mathbf{X}^+\mathbf{y} + (\mathbf{I}_p - \mathbf{X}^+\mathbf{X})\mathbf{q}\} = \mathbf{X}\mathbf{X}^+\mathbf{y} - (\mathbf{X} - \mathbf{X}\mathbf{X}^+\mathbf{X})\mathbf{q} = \mathbf{y} + (\mathbf{X} - \mathbf{X})\mathbf{q} = \mathbf{y}$, because of (17.5) and (17.10). Conversely, $\mathbf{Xb} = \mathbf{y}$ implies $\mathbf{X}^+\mathbf{Xb} = \mathbf{X}^+\mathbf{y}$ and thus $\mathbf{b} = \mathbf{b} + \mathbf{X}^+\mathbf{y} - \mathbf{X}^+\mathbf{Xb} = \mathbf{X}^+\mathbf{y} + (\mathbf{I}_p - \mathbf{X}^+\mathbf{X})\mathbf{b}$: (17.12) holds for $\mathbf{q} = \mathbf{b}$.

Equation (17.12) implies

$$\mathbf{b} \text{ is} \begin{cases} \text{unique, if } \mathbf{X}^+\mathbf{X} = \mathbf{I}_p \\ \text{not unique, otherwise} \end{cases}. \tag{17.13}$$

To illustrate Note 17.2, we rewrite (17.4) in the form of $\mathbf{Xb} = \mathbf{y}$, with $\mathbf{X} = \begin{bmatrix} 3 & -5 & 9 \\ -1 & 6 & -7 \end{bmatrix}$, $\mathbf{b} = [a, \ b, \ c]'$, $\mathbf{y} = [7, \ 1]'$. Then, we have $\mathbf{X}^+ = \begin{bmatrix} 0.24 & 0.26 \\ 0.22 & 0.31 \\ 0.15 & 0.09 \end{bmatrix}$ which can be found to satisfy (17.10), and (17.12) shows that the solution of (17.9) is given by $\mathbf{b} = \mathbf{X}^+\mathbf{y} + (\mathbf{I}_3 - \mathbf{X}^+\mathbf{X})\mathbf{q} = \begin{bmatrix} 1.94 \\ 1.83 \\ 1.15 \end{bmatrix} + \begin{bmatrix} 0.54 & -0.34 & -0.37 \\ -0.34 & 0.21 & 0.23 \\ -0.37 & 0.23 & 0.25 \end{bmatrix}\mathbf{q}$.

17.3 Singular Value Decomposition and the Moore–Penrose Inverse

Let $\mathbf{X}$ be an $n \times p$ matrix with $\text{rank}(\mathbf{X}) = r \leq \min(n, p)$ and its *singular value decomposition* (*SVD*) defined as in Theorem A.3.2, i.e.,

$$\mathbf{X} = \mathbf{K}\boldsymbol{\Lambda}\mathbf{L}', \tag{17.14}$$

with $\mathbf{K}'\mathbf{K} = \mathbf{L}'\mathbf{L} = \mathbf{I}_r$ and $\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_r \end{bmatrix}$ is an $r \times r$ diagonal matrix whose diagonal elements are all positive. Then, its *MP inverse* is expressed as

$$\mathbf{X}^{+} = \mathbf{L}\,\boldsymbol{\Lambda}^{-1}\mathbf{K}' \tag{17.15}$$

with $\boldsymbol{\Lambda}^{-1} = \begin{bmatrix} 1/\lambda_1 & & \\ & \ddots & \\ & & 1/\lambda_r \end{bmatrix}$. We can easily ascertain that (17.15) satisfies (17.5). The MP inverse of $\mathbf{X}$ may be *defined by* (17.15) rather than (17.5).

The SVD expression (17.15) for the MP inverse allows us to easily derive the properties listed next.

Note 17.3. Properties of the MP Inverse

$$\mathbf{X}\mathbf{X}^{+} \text{ and } \mathbf{X}^{+}\mathbf{X} \text{ are symmetric and idempotent} \tag{17.16}$$

$$\mathbf{X}'\mathbf{X}\mathbf{X}^{+} = \mathbf{X}' = \mathbf{X}^{+}\mathbf{X}\mathbf{X}' \tag{17.17}$$

$$\mathbf{X}'\mathbf{X}^{+\prime}\mathbf{X}^{+} = \mathbf{X}^{+} = \mathbf{X}^{+}\mathbf{X}^{+\prime}\mathbf{X}' \tag{17.18}$$

$$(\mathbf{X}'\mathbf{X})^{+} = \mathbf{X}^{+}\mathbf{X}^{+\prime}, (\mathbf{X}\mathbf{X}')^{+} = \mathbf{X}^{+\prime}\mathbf{X}^{+} \tag{17.19}$$

$$(\mathbf{X}'\mathbf{X})^{+}\mathbf{X}' = \mathbf{X}^{+} = \mathbf{X}'(\mathbf{X}\mathbf{X}')^{+} \tag{17.20}$$

$$\mathbf{X}(\mathbf{X}'\mathbf{X})^{+}\mathbf{X}'\mathbf{X} = \mathbf{X} = \mathbf{X}\mathbf{X}'(\mathbf{X}\mathbf{X}')^{+}\mathbf{X} \tag{17.21}$$

$$\operatorname{rank}(\mathbf{X}) = \operatorname{rank}(\mathbf{X}^{+}) = \operatorname{rank}(\mathbf{X}\mathbf{X}^{+}) = \operatorname{rank}(\mathbf{X}^{+}\mathbf{X}) \tag{17.22}$$

$$\mathbf{X}\mathbf{X}^{+} = \mathbf{I}_n \text{ if } \operatorname{rank}(\mathbf{X}) \text{ equals } n(\text{the number of the rows of } \mathbf{X}) \tag{17.23}$$

$$\mathbf{X}^{+}\mathbf{X} = \mathbf{I}_p \text{ if } \operatorname{rank}(\mathbf{X}) \text{ equals } p(\text{the number of the columns of } \mathbf{X}) \tag{17.24}$$

For example, (17.17) can be found using the fact that (17.14) and (17.15) imply

$$\mathbf{X}\mathbf{X}^{+} = \mathbf{K}\mathbf{K}' \quad \text{and} \quad \mathbf{X}^{+}\mathbf{X} = \mathbf{L}\mathbf{L}'. \tag{17.25}$$

They lead to $\mathbf{X}'\mathbf{X}\mathbf{X}^{+} = \mathbf{L}\boldsymbol{\Lambda}\mathbf{K}'(\mathbf{K}\mathbf{K}') = \mathbf{L}\boldsymbol{\Lambda}\mathbf{K}' = \mathbf{X}'$ and $\mathbf{X}^{+}\mathbf{X}\mathbf{X}' = (\mathbf{L}\mathbf{L}')\mathbf{L}\boldsymbol{\Lambda}\mathbf{K}' = \mathbf{L}\boldsymbol{\Lambda}\mathbf{K}' = \mathbf{X}'$. Further, (17.24) follows from (17.25) and (A.3.2), and these two equations lead to (17.23) in a parallel manner.

We can also find (17.20) as follows. (17.14) leads to $\mathbf{X}'\mathbf{X} = \mathbf{L}\boldsymbol{\Lambda}^2\mathbf{L}'$ and $\mathbf{X}\mathbf{X}' = \mathbf{K}\boldsymbol{\Lambda}^2\mathbf{K}$. From (17.15), their MP inverses are expressed as

$$(\mathbf{X}'\mathbf{X})^+ = \mathbf{L}\boldsymbol{\Lambda}^{-2}\mathbf{L}' \quad \text{and} \quad (\mathbf{X}\mathbf{X}')^+ = \mathbf{K}\boldsymbol{\Lambda}^{-2}\mathbf{K}'. \tag{17.26}$$

Multiplication of the matrices in (17.26) and the transposition of (17.14) lead to $(\mathbf{X}'\mathbf{X})^+\mathbf{X}' = \mathbf{L}\boldsymbol{\Lambda}^{-2}\mathbf{L}'\mathbf{L}\boldsymbol{\Lambda}\mathbf{K}' = \mathbf{L}\boldsymbol{\Lambda}^{-1}\mathbf{K}' = \mathbf{X}^+$ and $\mathbf{X}'(\mathbf{X}\mathbf{X}')^+ = \mathbf{L}\boldsymbol{\Lambda}\mathbf{K}'\mathbf{K}\boldsymbol{\Lambda}^{-2}\mathbf{K}' = \mathbf{L}\boldsymbol{\Lambda}^{-1}\mathbf{K}' = \mathbf{X}^+$.

The properties in Note 17.3 and others are proved in Magnus and Neudecker (2019, pp. 38–39) without using SVD.

17.4 Least Squares Problem Solved with Moore–Penrose Inverse

As explained in Appendix A.2.2, the *least square function*

$$f(\mathbf{B}) = \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|^2 \tag{17.27}$$

is minimized for $\mathbf{B} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}$ if $\mathbf{X}'\mathbf{X}$ is nonsingular. However, this condition of nonsingularity is not indispensable, as shown using the MP inverse in the next paragraph.

Function (17.27) is minimized for

$$\mathbf{B} = \mathbf{X}^+\mathbf{Y}, \quad \text{i.e.,} \quad \mathbf{X}\mathbf{B} = \mathbf{P}_\mathbf{X}\mathbf{Y}, \tag{17.28}$$

with

$$\mathbf{P}_\mathbf{X} = \mathbf{X}\mathbf{X}^+ = \mathbf{K}\mathbf{K}' \tag{17.29}$$

using the fact that (17.27) is decomposed as

$$\|\mathbf{Y} - \mathbf{X}\mathbf{B}\|^2 = \|\mathbf{Y} - \mathbf{P}_\mathbf{X}\mathbf{Y}\|^2 + \|\mathbf{P}_\mathbf{X}\mathbf{Y} - \mathbf{X}\mathbf{B}\|^2. \tag{17.30}$$

On the right side, only the term $\|\mathbf{P}_\mathbf{X}\mathbf{Y} - \mathbf{X}\mathbf{B}\|^2$ is dependent on $\mathbf{B}$, and this term becomes zero for (17.28). The decomposition (17.30) is derived as follows: (17.27) can be rewritten as

$$\begin{aligned} \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|^2 &= \|\mathbf{Y} - \mathbf{P}_\mathbf{X}\mathbf{Y} + \mathbf{P}_\mathbf{X}\mathbf{Y} - \mathbf{X}\mathbf{B}\|^2 \\ &= \|\mathbf{Y} - \mathbf{P}_\mathbf{X}\mathbf{Y}\|^2 + \|\mathbf{P}_\mathbf{X}\mathbf{Y} - \mathbf{X}\mathbf{B}\|^2 + 2\mathrm{tr}\mathbf{C}, \end{aligned} \tag{17.31}$$

with

$$\mathbf{C} = (\mathbf{Y} - \mathbf{P}_X\mathbf{Y})'(\mathbf{P}_X\mathbf{Y} - \mathbf{XB}) = \mathbf{Y}'\mathbf{P}_X\mathbf{Y} - \mathbf{Y}'\mathbf{XB} - \mathbf{Y}'\mathbf{P}'_X\mathbf{P}_X\mathbf{Y} + \mathbf{Y}'\mathbf{P}'_X\mathbf{XB}. \tag{17.32}$$

This $\mathbf{C}$ is found to be a zero matrix, since (17.29) leads to $\mathbf{P}_X = \mathbf{P}'_X$, $\mathbf{P}'_X\mathbf{P}_X = \mathbf{P}_X$, and $\mathbf{P}'_X\mathbf{X} = \mathbf{P}_X\mathbf{X} = \mathbf{X}\mathbf{X}^+\mathbf{X} = \mathbf{X}$.

The above facts suggest that the regression analysis formulated as (4.8) is feasible even if p (the number of the columns of $\mathbf{X}$ containing explanatory variables) is greater than n (the number of the rows of $\mathbf{X}$), although the *uselessness* of such analysis is discussed in Sect. 21.6.

In A.2.2, a *projection matrix* in (A.2.10) is introduced on the supposition that $\mathbf{X}'\mathbf{X}$ is nonsingular. However, this is not requisite: A necessary and sufficient condition for a matrix $\mathbf{M}$ to be a projection matrix is $\mathbf{M}' = \mathbf{M}$ and $\mathbf{MM} = \mathbf{M}$ (Yanai, Takeuchi, & Takane, 2011). Thus, (17.29) is a projection matrix. In more detail, the projection in this book refers to one in a narrow sense. It can be defined in a wider sense (Yanai et al., 2011).

Now, let us consider a generalized least squares problem of minimizing

$$f(\mathbf{G}) = \|\mathbf{Y} - \mathbf{XGZ}'\|^2 \tag{17.33}$$

over $\mathbf{G}$ for given $\mathbf{Y}$, $\mathbf{X}$, and $\mathbf{Z}$. This is called the *Penrose regression* problem (Penrose, 1956). Function (17.33) is minimized for

$$\mathbf{G} = \mathbf{X}^+\mathbf{YZ}^{+\prime}, \text{ i.e., } \mathbf{XGZ}' = \mathbf{P}_X\mathbf{YP}'_Z \tag{17.34}$$

with

$$\mathbf{P}_X = \mathbf{X}\mathbf{X}^+ \text{ and } \mathbf{P}_Z = \mathbf{Z}\mathbf{Z}^+. \tag{17.35}$$

This result follows from reexpressing (17.33) as

$$\begin{aligned} \|\mathbf{Y} - \mathbf{XGZ}\|^2 &= \left\|\mathbf{Y} - \mathbf{P}_X\mathbf{YP}'_Z + \mathbf{P}_X\mathbf{YP}'_Z - \mathbf{XGZ}'\right\|^2 \\ &= \left\|\mathbf{Y} - \mathbf{P}_X\mathbf{YP}'_Z\right\|^2 + \left\|\mathbf{P}_X\mathbf{YP}'_Z - \mathbf{XGZ}'\right\|^2 + 2\mathrm{tr}\mathbf{N}, \end{aligned} \tag{17.36}$$

with

$$\begin{aligned} \mathrm{tr}\mathbf{N} &= \mathrm{tr}\left(\mathbf{Y} - \mathbf{P}_X\mathbf{YP}'_Z\right)'\left(\mathbf{P}_X\mathbf{YP}'_Z - \mathbf{XGZ}'\right) \\ &= \mathrm{tr}\mathbf{Y}'\mathbf{P}_X\mathbf{YP}'_Z - \mathrm{tr}\mathbf{Y}'\mathbf{XGZ}' - \mathrm{tr}\mathbf{P}_Z\mathbf{Y}'\mathbf{P}'_X\mathbf{P}_X\mathbf{YP}'_Z + \mathrm{tr}\mathbf{P}_Z\mathbf{Y}'\mathbf{P}'_X\mathbf{XGZ}' \\ &= \mathrm{tr}\mathbf{Y}'\mathbf{P}_X\mathbf{YP}'_Z - \mathrm{tr}\mathbf{Y}'\mathbf{XGZ}' - \mathrm{tr}\mathbf{Y}'\mathbf{P}_X\mathbf{YP}'_Z + \mathrm{tr}\mathbf{Y}'\mathbf{XGZ}' = 0. \end{aligned} \tag{17.37}$$

Here, we have used the fact that (17.29) and (17.35) lead to $\mathbf{P}'_{\mathrm{X}} = \mathbf{P}_{\mathrm{X}}$, $\mathbf{P}'_{\mathrm{X}}\mathbf{P}_{\mathrm{X}} = \mathbf{P}_{\mathrm{X}}$, and $\mathbf{P}'_{\mathrm{X}}\mathbf{X} = \mathbf{P}_{\mathrm{X}}\mathbf{X} = \mathbf{X}\mathbf{X}^{+}\mathbf{X} = \mathbf{X}$, which also implies $\mathbf{P}'_{\mathrm{Z}} = \mathbf{P}_{\mathrm{Z}}$, $\mathbf{P}'_{\mathrm{Z}}\mathbf{P}_{\mathrm{Z}} = \mathbf{P}_{\mathrm{Z}}$, and $\mathbf{Z}'\mathbf{P}_{\mathrm{Z}} = \left(\mathbf{P}'_{\mathrm{Z}}\mathbf{Z}\right)' = (\mathbf{P}_{\mathrm{Z}}\mathbf{Z})' = \mathbf{Z}'$. Thus, on the right side of (17.36), $\mathbf{G}$ appears only in $\left\|\mathbf{P}_{\mathrm{X}}\mathbf{Y}\mathbf{P}'_{\mathrm{Z}} - \mathbf{X}\mathbf{G}\mathbf{Z}'\right\|^2 \geq 0$, which is zero in the case of (17.34).

17.5 Orthogonal Complement Matrix

Let us next consider a matrix whose columns are orthogonal to those of $\mathbf{X}$ $(n \times p)$ with its SVD defined as (17.14) and $r = \mathrm{rank}(\mathbf{X})$:

Note 17.4. Orthogonal Complement Matrix
The $n \times q$ matrix $\mathbf{X}_{\perp}$ satisfying

$$\mathbf{X}'\mathbf{X}_{\perp} = {}_{p}\mathbf{O}_{q} \tag{17.38}$$

is called the *orthogonal complement* (*OC*) *matrix* of $\mathbf{X}$. A matrix of $\mathbf{X}_{\perp}$ in (17.38) is generally expressed as

$$\mathbf{X}_{\perp} = (\mathbf{I}_n - \mathbf{X}'^{+}\mathbf{X}')\mathbf{M}. \tag{17.39}$$

Here, $\mathbf{M}$ is an arbitrary $n \times q$ matrix. This fact is proved in the next paragraph.

We can find (17.39) $\Rightarrow$ (17.38), by substituting (17.39) on the left side of (17.38): $\mathbf{X}'\left(\mathbf{I}_n - \mathbf{X}'^{+}\mathbf{X}'\right)\mathbf{M} = (\mathbf{X}' - \mathbf{X}'\mathbf{X}'^{+}\mathbf{X}')\mathbf{M} = (\mathbf{X}' - \mathbf{X}')\mathbf{M} = {}_{p}\mathbf{O}_{q}$. Conversely, (17.38) implies $\mathbf{X}'^{+}\mathbf{X}'\mathbf{X}_{\perp} = {}_{n}\mathbf{O}_{q}$ and thus $\mathbf{X}_{\perp} = \mathbf{X}_{\perp} - \mathbf{X}'^{+}\mathbf{X}'\mathbf{X}_{\perp} = (\mathbf{I}_n - \mathbf{X}'^{+}\mathbf{X}')\mathbf{X}_{\perp}$: (17.39) holds for $\mathbf{M} = \mathbf{X}_{\perp}$.

We can use (17.7) and (17.25) to rewrite (17.39) as

$$\mathbf{X}_{\perp} = \left[\mathbf{I}_n - (\mathbf{X}\mathbf{X}^{+})'\right]\mathbf{M} = (\mathbf{I}_n - \mathbf{K}\mathbf{K}')\mathbf{M}. \tag{17.40}$$

Now, we suppose that $\mathbf{X}$ is *centered* with

$$\mathbf{1}'_n\mathbf{X} = \mathbf{0}'_p. \tag{17.41}$$

Then, the *centered* OC matrix of $\mathbf{X}$, i.e., the $n \times q$ matrix $\bar{\mathbf{X}}_{\perp}$ that satisfies both

$$\mathbf{X}'\bar{\mathbf{X}}_{\perp} = {}_{p}\mathbf{O}_{q} \quad \text{and} \quad \mathbf{1}'_n\bar{\mathbf{X}}_{\perp} = \mathbf{0}'_q, \tag{17.42}$$

is generally expressed as

$$\bar{\mathbf{X}}_{\perp} = (\mathbf{I}_n - \mathbf{X}'^{+}\mathbf{X}')\bar{\mathbf{M}} \tag{17.43}$$

with $\bar{\mathbf{M}}$ an arbitrary $n \times q$ matrix satisfying

$$\mathbf{1}_n' = \bar{\mathbf{M}} = \mathbf{0}_q'. \tag{17.44}$$

This fact is shown in the next paragraph.

We can find (17.43) $\Rightarrow$ (17.42) as follows: (17.43) $\Rightarrow \mathbf{X}'\bar{\mathbf{X}}_{\perp} =_p \mathbf{O}_q$ is derived as (17.39) $\Rightarrow$ (17.38) was derived in Note 17.4, while (17.43) $\Rightarrow \mathbf{1}_n'\bar{\mathbf{X}}_{\perp} = \mathbf{0}_q'$ follows from the fact that (17.43) can be rewritten as $\bar{\mathbf{X}}_{\perp} = (\mathbf{I}_n - \mathbf{K}\mathbf{K}')\bar{\mathbf{M}}$ using (17.40). This leads to

$$\mathbf{1}_n'\bar{\mathbf{X}}_{\perp} = (\mathbf{1}_n' - \mathbf{1}_n'\mathbf{K}\mathbf{K}')\bar{\mathbf{M}} = \mathbf{1}_n'\bar{\mathbf{M}} = \mathbf{0}_q'. \tag{17.45}$$

Here, we have used (17.44) and the equality $\mathbf{1}_n'\mathbf{K} = \mathbf{0}_r'$ that follows from (17.14) implying $\mathbf{K} = \mathbf{X}\mathbf{L}\boldsymbol{\Lambda}^{-1}$ with (17.41). Conversely, (17.42) $\Rightarrow$ (17.43) follows from the fact that (17.42) leads to $\mathbf{X}'^{+}\mathbf{X}'\bar{\mathbf{X}}_{\perp} =_n \mathbf{O}_q$ and thus $\bar{\mathbf{X}}_{\perp} = \bar{\mathbf{X}}_{\perp} - \mathbf{X}'^{+}\mathbf{X}'\bar{\mathbf{X}}_{\perp} = (\mathbf{I}_n - \mathbf{X}'^{+}\mathbf{X}')\bar{\mathbf{X}}_{\perp}$: (17.43) holds for $\bar{\mathbf{M}} = \bar{\mathbf{X}}_{\perp}$ with this satisfying (17.44).

Now, let us suppose that (17.41) does not necessarily hold and $s = \mathrm{rank}(\mathbf{X}_{\perp}) \geq m$, with $\mathbf{X}_{\perp}$ given by (17.39). On these suppositions, we consider an $n \times m$ matrix $\mathbf{X}_{\perp}^{*}$, which can be called the *column-orthonormal OC matrix* of $\mathbf{X}$, i.e., it satisfies both

$$\mathbf{X}'\mathbf{X}_{\perp}^{*} =_p \mathbf{O}_m \quad \text{and} \quad \mathbf{X}_{\perp}^{*\prime}\mathbf{X}_{\perp}^{*} = \mathbf{I}_m. \tag{17.46}$$

Such a matrix $\mathbf{X}_{\perp}^{*}$ can be obtained through the SVD of (17.39):

$$(\mathbf{I}_n - \mathbf{X}'^{+}\mathbf{X}')\mathbf{M} = \mathbf{V}\boldsymbol{\Theta}\mathbf{W}'. \tag{17.47}$$

Here, $\mathbf{V}'\mathbf{V} = \mathbf{W}'\mathbf{W} = \mathbf{I}_s$, and $\boldsymbol{\Theta}$ is an $s \times s$ diagonal matrix with its diagonal elements all positive. The matrix $\mathbf{X}_{\perp}^{*}$ defined as

$$\mathbf{X}_{\perp}^{*} = \mathbf{V}\mathbf{S} = (\mathbf{I}_n - \mathbf{X}'^{+}\mathbf{X}')\mathbf{M}\mathbf{W}\boldsymbol{\Theta}^{-1}\mathbf{S}. \tag{17.48}$$

satisfies (17.46), with $\mathbf{S}$ an arbitrary $s \times m$ matrix meeting $\mathbf{S}'\mathbf{S} = \mathbf{I}_m$ and the last identity in (17.48) following from (17.47). That identity and $\mathbf{S}'\mathbf{V}'\mathbf{V}\mathbf{S} = \mathbf{S}'\mathbf{S} = \mathbf{I}_m$ allow us to find that (17.48) meets (17.46).

The *centered* OC matrix and the *column-orthonormal* OC matrix were treated in the last paragraphs. Now, we consider a *centered* and *column-orthonormal* OC matrix of the centered $\mathbf{X}$, on the supposition of $t = \mathrm{rank}(\bar{\mathbf{X}}_\perp) \geq m$ with $\bar{\mathbf{X}}_\perp$ given by (17.43). That is, to be considered is the $n \times m$ matrix $\bar{\mathbf{X}}_\perp^*$ that satisfies the three conditions

$$\mathbf{X}'\bar{\mathbf{X}}_\perp^* = {}_p\mathbf{O}_m, \quad \bar{\mathbf{X}}_\perp^{*\prime}\bar{\mathbf{X}}_\perp^* = \mathbf{I}_m, \quad \text{and} \quad \mathbf{1}_n'\bar{\mathbf{X}}_\perp^* = \mathbf{0}_q'. \tag{17.49}$$

for (17.41). This matrix $\bar{\mathbf{X}}_\perp^*$ is given through the SVD of (17.43) defined as

$$(\mathbf{I}_n - \mathbf{X}'^{+}\mathbf{X}')\bar{\mathbf{M}} = \mathbf{H}\boldsymbol{\Omega}\mathbf{Q}'. \tag{17.50}$$

Here, $\bar{\mathbf{M}}$ meets (17.44), $\mathbf{H}'\mathbf{H} = \mathbf{Q}'\mathbf{Q} = \mathbf{I}_t$, and $\boldsymbol{\Omega}$ is a $t \times t$ diagonal matrix whose diagonal elements are all positive. The matrix defined as

$$\bar{\mathbf{X}}_\perp^* = \mathbf{H}\mathbf{T} = (\mathbf{I}_n - \mathbf{X}'^{+}\mathbf{X}')\bar{\mathbf{M}}\mathbf{Q}\boldsymbol{\Omega}^{-1}\mathbf{T}. \tag{17.51}$$

satisfies (17.49), with $\mathbf{T}'\mathbf{T} = \mathbf{I}_m$ and the last identity following from (17.50). As (17.45) is found, we can find (17.51) to satisfy $\mathbf{1}_n'\bar{\mathbf{X}}_\perp^* = \mathbf{0}_q'$. This fact, the last identity in (17.51), and $\mathbf{T}'\mathbf{H}'\mathbf{H}\mathbf{T} = \mathbf{I}_m$ allow us to find that (17.51) meets (17.49).

For example, for $\mathbf{X} = \begin{bmatrix} 6 & -2 & 3 \\ 3 & 1 & -5 \\ 2 & 0 & 2 \\ -2 & -1 & 5 \\ -3 & 2 & -2 \\ -6 & 0 & -3 \end{bmatrix}$ with (17.41), $\bar{\mathbf{X}}_\perp^* = \begin{bmatrix} -0.154 & -0.500 \\ -0.239 & 0.489 \\ 0.795 & 0.163 \\ -0.411 & 0.460 \\ -0.240 & -0.514 \\ 0.249 & -0.097 \end{bmatrix}$ obtained through (17.51) is one of the 6×2 matrices satisfying (17.49).

17.6 Kronecker Product

The operations in this and the next two sections are only used in Chap. 20. Hence, in these three sections, symbols are used in the same convention as Chap. 20: we express the series of integers as $k = 1, \ldots, K$ and $p = 1, \ldots, P$, for example, rather than, $i = 1, \ldots, n$ and $j = 1, \ldots, p$ often used so far. Thus, the case of the characters should be carefully noted (e.g., "K" or "k").

The *Kronecker product*, which is denoted by $\otimes$, is defined as follows.

Note 17.5. The Kronecker Product
From two matrices $\mathbf{C} = (c_{kr})(K \times R)$ and $\mathbf{B} = \left(b_{jq}\right)(J \times Q)$, the *Kronecker product* gives the $KJ \times RQ$ matrices

$$\mathbf{C} \otimes \mathbf{B} = \begin{bmatrix} c_{11}\mathbf{B} & \cdots & c_{1R}\mathbf{B} \\ & \vdots & \\ c_{K1}\mathbf{B} & \cdots & c_{KR}\mathbf{B} \end{bmatrix} \quad \text{and} \quad \mathbf{B} \otimes \mathbf{C} = \begin{bmatrix} b_{11}\mathbf{C} & \cdots & b_{1Q}\mathbf{C} \\ & \vdots & \\ b_{J1}\mathbf{C} & \cdots & b_{JQ}\mathbf{C} \end{bmatrix}. \tag{17.52}$$

For example, if $\mathbf{C} = \begin{bmatrix} 1 & 0 \\ -3 & 2 \end{bmatrix}$ and $\mathbf{B} = \begin{bmatrix} -6 & 4 \\ 5 & -7 \end{bmatrix}$, then

$$\mathbf{C} \otimes \mathbf{B} = \begin{bmatrix} -6 & 4 & 0 & 0 \\ 5 & -7 & 0 & 0 \\ 18 & -12 & -12 & 8 \\ -15 & 21 & 10 & -14 \end{bmatrix} \quad \text{and}$$

$$\mathbf{B} \otimes \mathbf{C} = \begin{bmatrix} -6 & 0 & 4 & 0 \\ 18 & -12 & -12 & 8 \\ 5 & 0 & -7 & 0 \\ -15 & 10 & 21 & -14 \end{bmatrix}.$$

Let two vectors be defined as $\mathbf{c} = [c_1, \ldots, c_K]'$ and $\mathbf{b} = [b_1, \ldots b_J]$. Their Kronecker product $\mathbf{c} \otimes \mathbf{b}$ gives the $KJ \times 1$ vector:

$$\mathbf{c} \otimes \mathbf{b} = \begin{bmatrix} c_1\mathbf{b} \\ \vdots \\ c_K\mathbf{b} \end{bmatrix}. \tag{17.53}$$

We also find

$$\mathbf{c} \otimes \mathbf{b}' = \begin{bmatrix} c_1\mathbf{b}' \\ \vdots \\ c_K\mathbf{b}' \end{bmatrix} = \begin{bmatrix} c_1 b_1 & \cdots & c_1 b_J \\ \vdots & \vdots & \vdots \\ c_K b_1 & \cdots & c_K b_J \end{bmatrix} = \mathbf{c}\mathbf{b}' = [b_1\mathbf{c}, \ldots, b_J\mathbf{c}] = \mathbf{b}' \otimes \mathbf{c}. \tag{17.54}$$

The Kronecker product has the following properties:

$$(\mathbf{C} \otimes \mathbf{B})' = \mathbf{C}' \otimes \mathbf{B}', \tag{17.55}$$

$$(\mathbf{C} \otimes \mathbf{B})(\mathbf{X} \otimes \mathbf{Y}) = (\mathbf{C}\mathbf{X}) \otimes (\mathbf{B}\mathbf{Y}), \tag{17.56}$$

$$\mathrm{tr}(\mathbf{C} \otimes \mathbf{B}) = (\mathrm{tr}\mathbf{C})(\mathrm{tr}\mathbf{B}), \tag{17.57}$$

$$\mathrm{rank}(\mathbf{C} \otimes \mathbf{B}) = \mathrm{rank}(\mathbf{C})\mathrm{rank}(\mathbf{B}), \tag{17.58}$$

$$(\mathbf{C} \otimes \mathbf{B})^{+} = \mathbf{C}^{+} \otimes \mathbf{B}^{+}. \tag{17.59}$$

This and (17.58) imply that $(\mathbf{C} \otimes \mathbf{B})^{-1} = \mathbf{C}^{-1} \otimes \mathbf{B}^{-1}$ if $\mathbf{C}$ and $\mathbf{B}$ are nonsingular.

17.7 Khatri–Rao Product

While the symbol • is used to represent various operations in the literature, we use it to denote the *Khatri–Rao Product* (Rao & Mitra, 1971; Rao & Rao, 1998) in this book. The definition of the product is as follows.

Note 17.6. The Khatri–Rao Product
From $\mathbf{C} = (c_{kp}) = [\mathbf{c}_1, \ldots, \mathbf{c}_P](K \times P)$ and $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_P](J \times P)$, the *Khatri–Rao product* gives the $KJ \times P$ matrix

$$\begin{aligned} \mathbf{C} \bullet \mathbf{B} &= [\mathbf{c}_1, \ldots, \mathbf{c}_P] \bullet [\mathbf{b}_1, \ldots, \mathbf{b}_P] = [\mathbf{c}_1 \otimes \mathbf{b}_1, \ldots, \mathbf{c}_P \otimes \mathbf{b}_P] \\ &= \begin{bmatrix} c_{11}\mathbf{b}_1 & & c_{1P}\mathbf{b}_P \\ \vdots & \cdots & \vdots \\ c_{K1}\mathbf{b}_1 & & c_{KP}\mathbf{b}_P \end{bmatrix}. \end{aligned} \tag{17.60}$$

For example, if $\mathbf{C} = \begin{bmatrix} 1 & 0 \\ -3 & 2 \end{bmatrix}$ and $\mathbf{B} = \begin{bmatrix} -6 & 4 \\ 5 & -7 \end{bmatrix}$, then

$$\mathbf{C} \bullet \mathbf{B} = \begin{bmatrix} -6 & 0 \\ 5 & 0 \\ 18 & 8 \\ -15 & -14 \end{bmatrix} \quad \text{and} \quad \mathbf{B} \bullet \mathbf{C} = \begin{bmatrix} -6 & 0 \\ 18 & 8 \\ 5 & 0 \\ -15 & -14 \end{bmatrix}.$$

For $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_P](I \times P)$ and $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_P](J \times P)$, we have

$$(\mathbf{B} \bullet \mathbf{A})'(\mathbf{B} \bullet \mathbf{A}) = \mathbf{I}_P \ \text{ if } \ \mathbf{A}'\mathbf{A} = \mathbf{B}'\mathbf{B} = \mathbf{I}_P. \tag{17.61}$$

This can be proved as follows: $\mathbf{B} \bullet \mathbf{A} = [\mathbf{b}_1 \otimes \mathbf{a}_1, \ldots, \mathbf{b}_p \otimes \mathbf{a}_p]$ leads to

$$\begin{aligned}
(\mathbf{B} \bullet \mathbf{A})'(\mathbf{B} \bullet \mathbf{A}) &= \begin{bmatrix} (\mathbf{b}_1 \otimes \mathbf{a}_1)' \\ \vdots \\ (\mathbf{b}_P \otimes \mathbf{a}_P)' \end{bmatrix} [\mathbf{b}_1 \otimes \mathbf{a}_1, \ldots, \mathbf{b}_P \otimes \mathbf{a}_P] \\
&= \begin{bmatrix} (\mathbf{b}_1' \otimes \mathbf{a}_1')(\mathbf{b}_1 \otimes \mathbf{a}_1) & \cdots & (\mathbf{b}_1' \otimes \mathbf{a}_1')(\mathbf{b}_P \otimes \mathbf{a}_P) \\ \vdots & \vdots & \vdots \\ (\mathbf{b}_P' \otimes \mathbf{a}_P')(\mathbf{b}_1 \otimes \mathbf{a}_1) & \cdots & (\mathbf{b}_P' \otimes \mathbf{a}_P')(\mathbf{b}_P \otimes \mathbf{a}_P) \end{bmatrix} \\
&= \begin{bmatrix} (\mathbf{b}_1'\mathbf{b}_1) \otimes (\mathbf{a}_1'\mathbf{a}_1) & \cdots & (\mathbf{b}_1'\mathbf{b}_P) \otimes (\mathbf{a}_1'\mathbf{a}_P) \\ \vdots & \vdots & \vdots \\ (\mathbf{b}_P'\mathbf{b}_1) \otimes (\mathbf{a}_P'\mathbf{a}_1) & \cdots & (\mathbf{b}_P'\mathbf{b}_P) \otimes (\mathbf{a}_P'\mathbf{a}_P) \end{bmatrix} \\
&= \begin{bmatrix} (\mathbf{b}_1'\mathbf{b}_1)(\mathbf{a}_1'\mathbf{a}_1) & \cdots & (\mathbf{b}_1'\mathbf{b}_P)(\mathbf{a}_1'\mathbf{a}_P) \\ \vdots & \vdots & \vdots \\ (\mathbf{b}_P'\mathbf{b}_1)(\mathbf{a}_P'\mathbf{a}_1) & \cdots & (\mathbf{b}_P'\mathbf{b}_P)(\mathbf{a}_P'\mathbf{a}_P) \end{bmatrix},
\end{aligned} \tag{17.62}$$

where we have used (17.55), (17.56), and the inner product of vectors providing a scalar. On the rightest side of (17.62), it can be found that $(\mathbf{b}_p'\mathbf{b}_q)(\mathbf{a}_p'\mathbf{a}_q)$ is 1 for $p = q$, but 0 for $p \neq q$, since $\mathbf{A}'\mathbf{A} = \mathbf{B}'\mathbf{B} = \mathbf{I}_P$ in (17.61) implies that $\mathbf{b}_p'\mathbf{b}_q$ and $\mathbf{a}_p'\mathbf{a}_q$ are both equal to one if $p = q$, and otherwise zero.

17.8 Vec Operator

The *vec operator*, which is denoted by vec(), stacks the columns of a matrix vertically to produce a vector, as shown next.

Note 19.3. The Vec Operator
For $I \times J$ matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_J]$, the operator gives $JI \times 1$ vector:

$$\text{vec}(\mathbf{X}) = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_J \end{bmatrix} \tag{17.63}$$

For example, if $\mathbf{X} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$, then $\text{vec}(\mathbf{X}) = \begin{bmatrix} 1 \\ 3 \\ 2 \\ 4 \end{bmatrix} = [1\,3\,2\,4]'$.

For vectors $\mathbf{x} = [x_1, \ldots, x_I]'$ and $\mathbf{y} = [y_1, \ldots, y_J]'$

$$\text{vec}(\mathbf{x}) = \text{vec}(\mathbf{x}') = \mathbf{x} \tag{17.64}$$

$$\text{vec}(\mathbf{x}\mathbf{y}') = \text{vec}([y_1\mathbf{x}, \ldots, y_J\mathbf{x}]) = \begin{bmatrix} y_1\mathbf{x} \\ \vdots \\ y_J\mathbf{x} \end{bmatrix} = \mathbf{y} \otimes \mathbf{x} \tag{17.65}$$

Let $\mathbf{X}_1, \ldots, \mathbf{X}_K$ be matrices of the same size with $\alpha_1, \ldots, \alpha_K$ scalars. It can then be found that $\text{vec}(\alpha_1\mathbf{X}_1) = \alpha_1\text{vec}(\mathbf{X}_1)$ and $\text{vec}(\mathbf{X}_1 + \mathbf{X}_2) = \text{vec}(\mathbf{X}_1) + \text{vec}(\mathbf{X}_2)$. This implies $\text{vec}(\alpha_1\mathbf{X}_1 + \alpha_2\mathbf{X}_2) = \alpha_1\text{vec}(\mathbf{X}_1) + \alpha_2\text{vec}(\mathbf{X}_2)$. This can be generalized as

$$\text{vec}\left(\sum_{k=1}^{K} \alpha_k \mathbf{X}_k\right) = \sum_{k=1}^{K} \alpha_k \text{vec}(\mathbf{X}_k). \tag{17.66}$$

For any $I \times J$ matrix $\mathbf{X}$, $P \times I$ matrix $\mathbf{Y}$, and $J \times Q$ matrix $\mathbf{Z}$,

$$\text{vec}(\mathbf{YXZ}) = (\mathbf{Z}' \otimes \mathbf{Y})\text{vec}(\mathbf{X}) \;\text{ and }\; \text{vec}(\mathbf{YXZ})' = \text{vec}(\mathbf{X})'(\mathbf{Z} \otimes \mathbf{Y}') \tag{17.67}$$

hold true. Here, the left equality is derived as follows: Let $\mathbf{e}_j'$ denote a $1 \times J$ elementary vector which is filled with zeros except for the jth element which is one. Then, we have

$$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_J] = \sum_{j=1}^{J} \mathbf{x}_j \mathbf{e}'_j. \tag{17.68}$$

Using this, (17.54), (17.56), (17.65), and (17.66), we can show

$$\begin{aligned}
\mathrm{vec}(\mathbf{YXZ}) &= \mathrm{vec}\left(\mathbf{Y}\sum_{j=1}^{J}\mathbf{x}_j\mathbf{e}'_j\mathbf{Z}\right) = \mathrm{vec}\left(\sum_{j=1}^{J}\mathbf{Y}\mathbf{x}_j\mathbf{e}'_j\mathbf{Z}\right) = \sum_{j=1}^{J}\mathrm{vec}(\mathbf{Y}\mathbf{x}_j\mathbf{e}'_j\mathbf{Z}) \\
&= \sum_{j=1}^{J}\mathrm{vec}[\mathbf{Y}\mathbf{x}_j(\mathbf{Z}'\mathbf{e}_j)'] = \sum_{j=1}^{J}[(\mathbf{Z}'\mathbf{e}_j)\otimes(\mathbf{Y}\mathbf{x}_j)] = \sum_{j=1}^{J}[(\mathbf{Z}'\otimes\mathbf{Y})(\mathbf{e}_j\otimes\mathbf{x}_j)] \\
&= (\mathbf{Z}'\otimes\mathbf{Y})\sum_{j=1}^{J}(\mathbf{e}_j\otimes\mathbf{x}_j) = (\mathbf{Z}'\otimes\mathbf{Y})\sum_{j=1}^{J}\mathrm{vec}(\mathbf{x}_j\mathbf{e}'_j) = (\mathbf{Z}'\otimes\mathbf{Y})\mathrm{vec}\left(\sum_{j=1}^{J}\mathbf{x}_j\mathbf{e}'_j\right) \\
&= (\mathbf{Z}'\otimes\mathbf{Y})\mathrm{vec}(\mathbf{X}).
\end{aligned}$$

The right equality in (17.67) is derived from the left one using (17.55).

17.9 Hadamard Product

The *Hadamard product* of two matrices $\mathbf{A} = (a_{ij})$ and $\mathbf{B} = (b_{ij})$ of the same size $n \times p$ is defined as the element-wise product:

$$\mathbf{A}\odot\mathbf{B} = (a_{ij}b_{ij}) = \begin{bmatrix} a_{11}b_{11} & \cdots & a_{1p}b_{1p} \\ \vdots & \vdots & \vdots \\ a_{n1}b_{n1} & \cdots & a_{np}b_{np} \end{bmatrix}. \tag{17.69}$$

For example, if $\mathbf{A} = \begin{bmatrix} 1 & 0 \\ -3 & 2 \end{bmatrix}$ and $\mathbf{B} = \begin{bmatrix} -6 & 4 \\ -5 & -7 \end{bmatrix}$, then $\mathbf{A}\odot\mathbf{B} = \mathbf{B}\odot\mathbf{A} = \begin{bmatrix} -6 & 0 \\ 15 & -14 \end{bmatrix}$.

17.10 Bibliographical Notes

Banerjee and Roy (2014) and Schott (2005) are among the textbooks recommended for learning the advanced matrix operation. Exhaustive descriptions for the matrix operation are also found in Harville (1997) and Seber (2008). Although differential calculus for matrices is the main subject in Magnus and Neudecker (2019), its Chaps. 1–3 and 11 are useful for understanding matrix algebra. The computational aspects in matrices are specially emphasized in Eldén (2007) and Gentle (2017). Formulas for matrix operations are exhaustively listed in Lütkepohl (1996).

Horn and Johnson (2019) is among the most advanced books of matrix algebra, as is Golub and van Loan (2013) for matrix computations.

Exercises

17.1. Let us replace the vectors $\mathbf{y}$ and $\mathbf{b}$ in the equation $\mathbf{Xb} = \mathbf{y}$ by matrices as $\mathbf{XB} = \mathbf{Y}$: Consider the equation $\mathbf{XB} = \mathbf{Y}$ for given $\mathbf{X}$ $(n \times p)$ and $\mathbf{Y}$ $(n \times q)$. Show that if $\mathbf{XX}^+\mathbf{Y} = \mathbf{Y}$, the solution of $\mathbf{B}$ $(p \times q)$ for the equation is given by $\mathbf{B} = \mathbf{X}^+\mathbf{Y} + \left(\mathbf{I}_p - \mathbf{X}^+\mathbf{X}\right)\mathbf{Q}$ with $\mathbf{Q}$ an arbitrary $p \times q$ matrix.

17.2. Let $\mathbf{S}$ be symmetric and its *eigenvalue decomposition* (*EVD*) be defined as $\mathbf{S} = \mathbf{E}\Theta\mathbf{E}'$, with $\mathrm{rank}(\mathbf{S}) = r, \mathbf{E}'\mathbf{E} = \mathbf{I}_r$, and $\boldsymbol{\Theta}$ being the diagonal matrix whose diagonal elements are not zero. Show $\mathbf{S}^+ = \mathbf{E}\Theta^{-1}\mathbf{E}'$. See Appendix A.3.4 for the EVD.

17.3. Use the SVD of $\mathbf{X}$ to show (17.21) and (17.22).

17.4. Argue how the MP inverse is more useful than the inverse matrix.

17.5. Let $\mathbf{N}$ be $n \times n$. Show that if $\mathrm{rank}(\mathbf{N}) = n$, the $n \times q$ orthogonal complement (OC) matrix of $\mathbf{N}$ is ${}_n\mathbf{O}_q$.

17.6. Let $\mathbf{Y}$ $(m \times p)$ be the *row-orthogonal complement matrix* of $\mathbf{X}$ $(n \times p)$ satisfying $\mathbf{YX}' =_m \mathbf{O}_n$. Show $\mathbf{Y} = \mathbf{R}\left(\mathbf{I}_p - \mathbf{X}'\mathbf{X}'^+\right)$ with $\mathbf{R}$ an arbitrary $m \times p$ matrix.

17.7. Discuss the equivalence between minimizing $\|\mathbf{X} - (\mathbf{FA}' + \mathbf{F}_\perp\mathbf{B}')\|^2$ over $\mathbf{F}$ $(n \times p)$ and minimizing $\|\mathbf{X} - (\mathbf{FA}' + \mathbf{UB}')\|^2$ over $\mathbf{F}$ and $\mathbf{U}$ $(n \times q)$ subject to $\mathbf{F}'\mathbf{U} =_p \mathbf{O}_q$, where $\mathbf{F}_\perp$ is the $n \times q$ OC matrix of $\mathbf{F}$.

17.8. Let $\mathbf{G} = [\mathbf{F}, \mathbf{U}]$ be the $n \times (p + q)$ block matrix with $\mathbf{F}$ and $\mathbf{U}$ being $n \times p$ and $n \times q$, respectively. Show that $\mathbf{G}'\mathbf{G} = \mathbf{I}_{p+q}$ implies $\mathbf{F}$ being the column-orthonormal OC matrix of $\mathbf{U}$ and this being the column-orthonormal OC matrix of $\mathbf{F}$.

17.9. Let each of $I \times 1$ random vectors $\mathbf{x}_1, \ldots, \mathbf{x}_K$ follows the I-variate normal distribution whose mean vector is $\boldsymbol{\mu}$ $(I \times 1)$ and covariance matrix is $\boldsymbol{\Sigma}$ $(I \times I)$. Show that we can express

$$\begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_K \end{bmatrix} \sim N_{KI}(\mathbf{1}_K \otimes \boldsymbol{\mu}, \mathbf{I}_K \otimes \boldsymbol{\Sigma}),$$

if $\mathbf{x}_K$ and $\mathbf{x}_l$ are mutually uncorrelated for $k \neq l$ $(k = 1, \ldots, K; l = 1, \ldots, K)$.

17.10. Let $\mathbf{X}_k = \big(x_{ijk}\big)(k = 1, \ldots, K)$ an $I \times J$ data matrix, whose (i, j) element is x_{ijk} and modeled as $x_{ijk} = \alpha_i + \beta_j + e_{ijk}$ with e_{ijk} an error. Show that the model can be rewritten as $[\text{vec}(\mathbf{X}_1), \ldots, \text{vec}(\mathbf{X}_K)] = (\mathbf{1}_J \otimes \boldsymbol{\alpha})\mathbf{1}'_K + (\boldsymbol{\beta} \otimes \mathbf{1}_I)\mathbf{1}'_K + [\text{vec}(\mathbf{E}_1), \ldots, \text{vec}(\mathbf{E}_K)]$ with $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_I]'$, $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_J]'$, and $\mathbf{E}_k(k = 1, \ldots, K)$ being the $I \times J$ matrix whose (i, j) element is e_{ijk}.

17.11. Let $\mathbf{X}$ and $\mathbf{Y}$ be $n \times p$ matrices. Show $\mathbf{1}'_n(\mathbf{X} \odot \mathbf{Y})\mathbf{1}_p = \text{tr}\mathbf{X}\mathbf{Y}'$.

17.12. Let $\mathbf{R} = \mathbf{V}\mathbf{D}\mathbf{V}'$ be a $p \times p$ matrix with $\mathbf{D}$ being an $r \times r$ diagonal matrix. Show that $(\mathbf{V} \odot \mathbf{V})\mathbf{D}\mathbf{1}_r = \mathbf{1}_p$ stands for all diagonal elements of $\mathbf{R}$ being ones.

17.13. Discuss the implication of the following problem: for a $n \times p$ data matrix $\mathbf{X} = (x_{ij})$ including unobserved elements (i.e., missing ones), minimize $f(\mathbf{Z}, \mathbf{W}) = \|(\mathbf{W} \odot \mathbf{X}) - \mathbf{Z}\|^2$ over $\mathbf{W} = (w_{ij})$ and $\mathbf{Z} = (z_{ij})$, subject to the constraints: rank($\mathbf{Z}$) < rank($\mathbf{X}$), and $w_{ij} = 1$ if x_{ij} is observed, but w_{ij} being unknown otherwise.

Chapter 18
Exploratory Factor Analysis (Part 2)

In Chap. 12, exploratory factor analysis (EFA) was formulated as a probabilistic model. However, EFA can also be formulated as a kind of *matrix decomposition* problem, without using the notion of probabilities. This formulation of EFA was proposed in 2001 by Professor Henk A. L. Kiers at the University of Groningen, as found in Sŏcan's (2003, p. 17) Ph.D. thesis from the same university. In this formulation, common and unique factor scores, loadings, and unique variances are all treated as *fixed unknown parameters in matrices*. As it leads to a procedure which is fully based on matrix algebra, the EFA procedure can be referred to as *matrix decomposition factor analysis* (*MDFA*). In contrast, the procedure in Chapter 12 can be called *latent variable factor analysis* (*LVFA*), as factor scores are treated as random latent variables. MDFA and LVFA are found to provide almost equivalent solutions of factor loadings and unique variances. However, the *strengths* of MDFA are that *essential properties of FA* are elucidated with the frameworks of matrix algebra, the *model part* approximating a data set can be *identified*, and the *optimal factor scores* can be expressed by a formula, which cannot be done in LVFA.

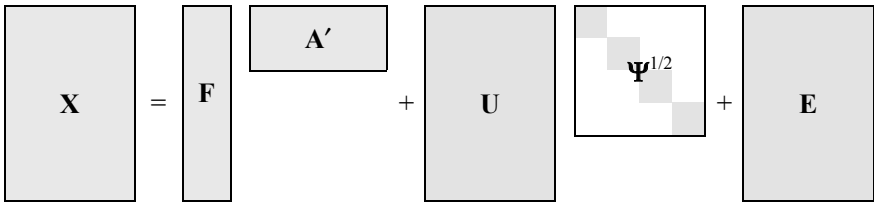
18.1 Matrix Decomposition Formulation

Professor Henk A. L. Kiers at the University of Groningen, Netherlands, proposed the *matrix decomposition formulation* of exploratory factor analysis (EFA) in 2001, as mentioned in this chapter's introduction. This formulation is introduced simply by adding the matrix product $\mathbf{U}\boldsymbol{\Psi}^{1/2}$ into the PCA model (5.1) or (5.3) (Chap. 5). EFA can be modeled as

$$\mathbf{X} = \mathbf{F}\mathbf{A}' + \mathbf{U}\boldsymbol{\Psi}^{1/2} + \mathbf{E} \tag{18.1}$$

K. Adachi, *Matrix-Based Introduction to Multivariate Data Analysis*,
https://doi.org/10.1007/978-981-15-4103-2_18

for an n-individuals $\times$ p-variables centered data matrix $\mathbf{X} = (x_{ij})$ with $\mathbf{1}_n'\mathbf{X} = \mathbf{0}_p'$ and $n > p$. Here, the superscript 1/2 is attached to $\boldsymbol{\Psi}$ as $\boldsymbol{\Psi}^{1/2}$, so that its square corresponds to (12.5) in Chap. 12 as explained later. The model (18.1) can also be expressed visually as



(18.2)

The matrices in (18.1) or (18.2) are listed as follows:

$\mathbf{F} = (f_{ik})$: n (individuals) $\times$ m (common factors) matrix of common factor scores;
$\mathbf{U} = (u_{ij})$: n (individuals) $\times$ p (unique factors) matrix of unique factor scores;
$\mathbf{E} = (e_{ij})$: n (individuals) $\times$ p (variables) matrix of errors;
$\mathbf{A} = (a_{jk})$: p (variables) $\times$ m (common factors) matrix of factors loadings;
$\boldsymbol{\Psi}^{1/2}$: p (unique factors) $\times$ p (variables) diagonal matrix, whose jth diagonal element is expressed as $\psi_j^{1/2}$.

A key point is that $m < p$ ($< n$), i.e., the number of common factors is less than the number of variables, as shown in (18.2), where $\mathbf{F}$ is narrower than $\mathbf{X}$. The relationships of p variables to m common factors are described by the loading matrix $\mathbf{A}$. On the other hand, the number of unique factors is equal to the number of variables: $\mathbf{X}$ and $\mathbf{U}$ are of the same size as in (18.2). Further, the matrix $\boldsymbol{\Psi}^{1/2}$, which relates $\mathbf{U}$ to $\mathbf{X}$, is diagonal. This implies that each column of $\mathbf{U}$ and $\mathbf{X}$ has a one-to-one correspondence.

The differences between the common factors in $\mathbf{F}$ and the unique factors in $\mathbf{U}$ can be seen in the diagram in Fig. 18.1, which illustrates the EFA model (18.1) or

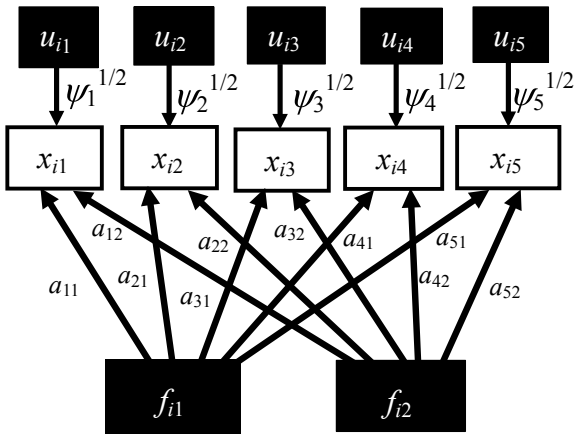


Fig. 18.1 Graphical representation of matrix decomposition FA (MDFA) with $p = 5$ and $m = 2$

(18.2) with $p = 5$ and $m = 2$. As seen here, a few common factors account for all variables with the loadings in $\mathbf{A} = (a_{jk})$ being coefficients, while the jth unique factor u_{ij} specifically affects the jth variable x_{ij} with $\psi_j^{1/2}$ a coefficient. The diagram has the same form as Fig. 12.1, with factors depicted as squares rather than ellipses and circles in Fig. 12.1. This is in order to emphasize the difference of *factor scores* being treated as elements in *fixed matrices* in (18.1) as opposed to latent variables in Chap. 12, as detailed later.

The common and unique factor score matrices $\mathbf{F}$ and $\mathbf{U}$ in (18.1) or (18.2) are constrained through

$$\mathbf{1}_n'\mathbf{F} = \mathbf{0}_m \quad \text{and} \quad \mathbf{1}_n'\mathbf{U} = \mathbf{0}_p, \tag{18.3}$$

$$\frac{1}{n}\mathbf{F}'\mathbf{F} = \mathbf{I}_m, \quad \frac{1}{n}\mathbf{U}'\mathbf{U} = \mathbf{I}_p, \quad \text{and} \quad \mathbf{F}'\mathbf{U} =_m \mathbf{O}_p. \tag{18.4}$$

The constraints in (18.3) and (18.4) imply that the factor scores are centered and standardized, with the scores in a column of $[\mathbf{F}, \mathbf{U}]$ uncorrelated to those in the other columns. This uncorrelatedness can also be found in Fig. 18.1, where no factors are linked with each other.

The matrices $\mathbf{F}$, $\mathbf{A}$, $\mathbf{U}$, and $\mathbf{\Psi}^{1/2}$ in (18.1) or (18.2) are treated as *unknown parameter matrices* to be estimated. For the estimation, a *least square function* is defined as the sum of the squared elements of $\mathbf{E}$ in (18.1) or (18.2):

$$\begin{aligned} f\left(\mathbf{F}, \mathbf{U}, \mathbf{A}, \mathbf{\Psi}^{1/2}\right) &= f(\mathbf{Z}, \mathbf{B}) = \|\mathbf{E}\|^2 = \left\|\mathbf{X} - \left(\mathbf{F}\mathbf{A}' + \mathbf{U}\mathbf{\Psi}^{1/2}\right)\right\|^2 \\ &= \left\|\mathbf{X} - [\mathbf{F}, \mathbf{U}]\left[\mathbf{A}, \mathbf{\Psi}^{1/2}\right]'\right\|^2 = \|\mathbf{X} - \mathbf{Z}\mathbf{B}'\|^2, \end{aligned} \tag{18.5}$$

where $\mathbf{Z} = [\mathbf{F}, \mathbf{U}]$ and $\mathbf{B} = [\mathbf{A}, \mathbf{\Psi}^{1/2}]$ are $n \times (m + p)$ and $p \times (m + p)$ block matrices. Using $\mathbf{Z}$, the constraints in (18.3) and (18.4) are summarized as

$$\mathbf{1}_n'\mathbf{Z} = \mathbf{0}_{m+p}, \tag{18.6}$$

$$\frac{1}{n}\mathbf{Z}'\mathbf{Z} = \mathbf{I}_{m+p}. \tag{18.7}$$

That is, EFA can be formulated by minimizing (18.5) over $\mathbf{Z}$ and $\mathbf{B}$ subject to (18.6) and (18.7) in the matrix decomposition formulation. We call this procedure as *matrix decomposition factor analysis* (*MDFA*). Here, it should be noted that constraint (18.7) or (18.4) implies $n \geq m + p$, since of the following facts: (3.34) and (18.7) lead to $\text{rank}(\mathbf{Z}) = m + p$, and the comparison of this result with (3.32) allows us to find that $m + p \leq \min(n, m + p)$, i.e., $m + p \leq n$ is required.

18.2 Comparisons to Latent Variable Formulation

The model for MDFA expressed in (18.1–18.4) can be compared to the corresponding model in Chap. 12 by replacing $\mathbf{e}$ in the latter with $\mathbf{\Psi}^{1/2}\mathbf{u}$. The replacement allows (12.3) and (12.4) to be rewritten as

$$\mathbf{x} = \mathbf{A}\mathbf{f} + \mathbf{\Psi}^{1/2}\mathbf{u}, \tag{18.8}$$

$$\mathbf{\Psi}^{1/2}\mathbf{u} \sim N_p(\mathbf{0}_p, \mathbf{\Psi}), \quad \text{or equivalently,} \quad \mathbf{u} \sim N_p(\mathbf{0}_p, \mathbf{I}_p). \tag{18.9}$$

with $\mathbf{f}$ ($m \times 1$) and $\mathbf{u}$ ($p \times 1$) common and unique factor score vectors, respectively, Here, we note (12.6) again:

$$\mathbf{f} \sim N_m(\mathbf{0}_m, \mathbf{I}_m). \tag{18.10}$$

The assumption of mutual independence of $\mathbf{f}$ and $\mathbf{e}$ described in Sect. 12.3 is equivalent to assuming that $\mathbf{f}$ and $\mathbf{u}$ are distributed mutually independently. In (18.8–18.10), the scores in $\mathbf{f}$ and $\mathbf{u}$ are treated as *random latent variables* which can take various values. In this sense, we can refer to the EFA procedure in Chap. 12 as *latent variable factor analysis* (*LVFA*).

LVFA can be related to MDFA in the last section as follows: The transposes of the vectors $\mathbf{f}$ and $\mathbf{u}$ in the LVFA model (18.8) correspond to each row of $\mathbf{F}$ and $\mathbf{U}$ in the MDFA model in (18.1) or (18.2). The LVFA assumption $\mathbf{u} \sim N_p(\mathbf{0}_p, \mathbf{I}_p)$ in (18.9) corresponds to the MDFA constraints $\mathbf{1}_n'\mathbf{U} = \mathbf{0}_p$ in (18.3) and $n^{-1}\mathbf{U}'\mathbf{U} = \mathbf{I}_p$ in (18.4). Analogously, the LVFA assumption (18.10) corresponds to the MDFA constraints $\mathbf{1}_n'\mathbf{F} = \mathbf{0}_p$ in (18.3) and $n^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_m$ in (18.4). Finally, the LVFA assumption of mutual independence of $\mathbf{f}$ and $\mathbf{u}$ is associated with the MDFA constraint $\mathbf{F}'\mathbf{U} = {}_m\mathbf{O}_p$ in (18.4).

The covariance matrix among the columns of the unique factor part $\mathbf{U}\mathbf{\Psi}^{1/2}$ in the MDFA model (18.1) is given by $n^{-1}\left(\mathbf{U}\mathbf{\Psi}^{1/2}\right)'\mathbf{J}\mathbf{U}\mathbf{\Psi}^{1/2} = \mathbf{\Psi}$ with $\mathbf{J} = \mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n'$, since of (18.4) and $\mathbf{J}\mathbf{U}\mathbf{\Psi}^{1/2} = \mathbf{U}\mathbf{\Psi}^{1/2}$ following from (3.20) and (18.3). Thus, the diagonal elements of $\mathbf{\Psi}$, i.e., $\psi_1, \ldots, \psi_p$, can be called *unique variances*, like those in LVFA (Sect. 12.3).

One difference between LVFA and MDFA is that the unique factor part $\mathbf{e} = \mathbf{\Psi}^{1/2}\mathbf{u}$ can be viewed as an error variable in LVFA (18.8), while the error matrix $\mathbf{E}$ is a necessary addition to the unique factor part $\mathbf{U}\mathbf{\Psi}^{1/2}$ in MDFA (18.1). This is due to the fact that the part $\mathbf{F}\mathbf{A}' + \mathbf{U}\mathbf{\Psi}^{1/2}$ in (18.1) is a fixed matrix and cannot be equalized to $\mathbf{X}$. In spite of this difference, the LVFA and MDFA *solutions* of $\mathbf{A}$ and $\mathbf{\Psi}$ have been found to be *almost equivalent*, as illustrated in the next section.

We must note that the *factor scores* in $\mathbf{F}$ and $\mathbf{U}$ are the parameters to be *optimally estimated* in MDFA as in (18.5). On the other hand, the scores are treated as random vectors $\mathbf{f}$ and $\mathbf{u}$ in LVFA and their optimal estimation is out of scope, in line with $\mathbf{f}$ and $\mathbf{u}$ being absent from the LVFA objective function (12.9); its optimization only

aims to obtain the optimal $\mathbf{A}$ and $\mathbf{\Psi}$, though the resulting $\mathbf{A}$ and $\mathbf{\Psi}$ allows the $\mathbf{f}$ value for a particular individual to be obtained through optimizing an objective function other than (12.9) (e.g., Yanai & Ichikawa, 2007; Mulaik, 2010).

18.3 Solution of Loadings and Unique Variances

Before explaining how (18.5) is minimized in MDFA, we illustrate its solution for $\mathbf{B} = [\mathbf{A}, \mathbf{\Psi}^{1/2}]$. Here, it must be noted that the solution can also be *rotated* by the procedures in Chap. 13, as the MDFA solution has the same rotational indeterminacy as explained in Sect. 12.5. That is, we can substitute $\mathbf{AT}$ for $\mathbf{A}$ and $\mathbf{FT}$ for $\mathbf{F}$ in (18.1–18.4), without changing the equations. Here, $\mathbf{T}$ is an $m \times m$ orthonormal matrix with $\mathbf{T}'\mathbf{T} = \mathbf{T}\mathbf{T}' = \mathbf{I}_m$. Furthermore, this condition may be relaxed as $\mathbf{T}$ being a nonsingular matrix satisfying $\mathrm{diag}(\mathbf{T}'\mathbf{T}) = \mathbf{I}_m$. Here, $\mathrm{diag}(\mathbf{T}'\mathbf{T})$ stands for the $m \times m$ diagonal matrix whose diagonal elements are those of $\mathbf{T}'\mathbf{T}$, as defined in Note 12.1. Hence, we can substitute $\mathbf{AT}'^{-1}$ for $\mathbf{A}$ and $\mathbf{FT}$ for $\mathbf{F}$ in (18.1–18.4), without changing the equations, except $n^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_m$. This exception is not a problematic one as explained in Sect. 12.5. When the resulting $\mathbf{F}$ in MDFA with constraint $n^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_m$ is rotated to $\mathbf{FT}$ with $\mathrm{diag}(\mathbf{T}'\mathbf{T}) = \mathbf{I}_m$, the substitution of the resulting $\mathbf{FT}$ into $\mathbf{F}$ in $n^{-1}\mathbf{F}'\mathbf{F}$ leads to $n^{-1}\mathbf{T}'\mathbf{F}'\mathbf{F}\mathbf{T} = \mathbf{T}'(n^{-1}\mathbf{F}'\mathbf{F})\mathbf{T} = \mathbf{T}'\mathbf{T}$. This matrix can be regarded as containing factor correlations, since $\mathrm{diag}(\mathbf{T}'\mathbf{T}) = \mathbf{I}_m$.

For standard scores of the data set in Table 10.1, MDFA followed by the oblique geomin rotation (Sect. 13.5) provides the loadings, unique variances, and factor correlation presented in Table 18.1. Let us compare the results with the corresponding LVFA ones in Table 12.1(B). One difference is that the former table has the column "Residual", in which $r_j = n^{-1}\sum_{i=1}^{n} e_{ij}^2$ for variable j is presented with e_{ij} the (i, j) element of the resulting $\mathbf{E}$. This r_j is explained later in Sect. 18.7. The index r_j is not presented in the LV approach, whose model does not have $\mathbf{E}$ as described in Sect. 18.2.

Table 18.1 MDFA solution for the standard scores of the data in Table 10.1(A) with $\mathbf{AT}$ obtained by the geomin rotation

j	$\mathbf{AT}$		ψ_j	Residual
A	0.82	0.07	0.25	0.0026
C	−0.14	0.84	0.38	0.0007
I	0.74	0.04	0.42	0.0015
B	−0.04	−0.81	0.30	0.0014
T	0.03	0.89	0.18	0.0013
V	0.88	0.01	0.22	0.0019
H	−0.82	0.09	0.39	0.0028
P	0.23	0.59	0.47	0.0016
Correlation	0.48			

Except for the "Residual" column, the solutions from MDFA and LVFA [i.e., Tables 18.1 and 12.1(B)] are *almost equivalent*. Thus, the interpretation of the loadings, unique variances, and factor correlation is also the same for the MDFA and LVFA solutions. Such broad equivalence of MDFA and LVFA solutions can be found for other data sets (Adachi & Trendafilov, 2019).

18.4 Iterative Algorithm

In this section, we present the MDFA algorithm for obtaining $\mathbf{B} = [\mathbf{A}, \mathbf{\Psi}^{1/2}]$, leaving its derivation to be explained later. Here, we suppose

$$\text{rank}(\mathbf{XB}) = p. \tag{18.11}$$

This implies rank $(\mathbf{X}) = p$ and rank$(\mathbf{B}) = p$ (the number of rows in $\mathbf{B}$), with the latter leading to

$$\mathbf{BB}^{+} = \mathbf{I}_p \tag{18.12}$$

because of (17.23).

Let $\mathbf{V} = n^{-1}\mathbf{X}'\mathbf{X}$ be the inter-variable covariance matrix and

$$\mathbf{S}_{\text{XZ}} = \frac{1}{n}\mathbf{X}'\mathbf{Z} = \left[\frac{1}{n}\mathbf{X}'\mathbf{F}, \frac{1}{n}\mathbf{X}'\mathbf{U}\right] = [\mathbf{S}_{\text{XE}}, \mathbf{S}_{\text{XU}}] \tag{18.13}$$

denote the p-variables $\times$ $(m + p)$-factors covariance matrix, where $\mathbf{S}_{\text{XF}} = n^{-1}\mathbf{X}'\mathbf{F}$ contains the covariances of p variables to m common factors, and $\mathbf{S}_{\text{XU}} = n^{-1}\mathbf{X}'\mathbf{U}$ consists of the covariances of p variables to p unique factors. Then, the MDFA algorithm for obtaining $\mathbf{B} = [\mathbf{A}, \mathbf{\Psi}^{1/2}]$ can be listed as follows:

Note 18.1. Algorithm for Obtaining B

Step 1. Initialize $\mathbf{B} = [\mathbf{A}, \mathbf{\Psi}^{1/2}]$

Step 2. Perform the *eigenvalue decomposition* (*EVD*) of $\mathbf{B}'\mathbf{VB}$ defined as

$$\mathbf{B}'\mathbf{VB} = \mathbf{Q}\mathbf{\Theta}^2\mathbf{Q}', \tag{18.14}$$

with $\mathbf{Q}'\mathbf{Q} = \mathbf{I}_p$ and $\mathbf{\Theta}^2$ a $p \times p$ diagonal matrix.

Step 3. Update $\mathbf{S}_{\text{XZ}}$ as

$$\mathbf{S}_{\text{XZ}} = \mathbf{B}^{+\prime}\mathbf{Q}\mathbf{\Theta}\mathbf{Q}'. \tag{18.15}$$

Step 4. Update $\mathbf{B} = [\mathbf{A}, \mathbf{\Psi}^{1/2}]$ as

$$\mathbf{B} = [\mathbf{S}_{\mathrm{XF}}, \mathrm{diag}(\mathbf{S}_{\mathrm{XU}})], \text{i.e.}, \mathbf{A} = \mathbf{S}_{\mathrm{XF}} \text{ and } \boldsymbol{\Psi}^{1/2} = \mathrm{diag}(\mathbf{S}_{\mathrm{XU}}). \tag{18.16}$$

Step 5. Finish if the decrease in the standardized loss function value

$$f_{\mathrm{s}} = 1 - \frac{\mathrm{tr}\mathbf{B}\mathbf{B}'}{\mathrm{tr}\mathbf{V}} \tag{18.17}$$

from the previous round is small enough; otherwise, go back to Step 2.

Here, the EVD used in Step 2 is detailed in Note 6.1 and A.3.4.

In Note 18.1, we find that obtaining $\mathbf{Z} = [\mathbf{F}, \mathbf{U}]$ is *unnecessary* for finding the solution of $\mathbf{B} = [\mathbf{A}, \boldsymbol{\Psi}^{1/2}]$: The algorithm is *only involved in* $\mathbf{S}_{\mathrm{XZ}}$ and $\mathbf{B}$. This is because (18.5) can be expanded and rewritten using (18.5) as

$$\begin{aligned} f(\mathbf{Z}, \mathbf{B}) = \mathrm{tr}\mathbf{X}'\mathbf{X} - 2\mathrm{tr}\mathbf{X}'\mathbf{Z}\mathbf{B}' + \mathrm{tr}\mathbf{B}\mathbf{Z}'\mathbf{Z}\mathbf{B}' &= n\mathrm{tr}\mathbf{V} - 2\mathrm{tr}\mathbf{X}'\mathbf{Z}\mathbf{B}' + n\mathrm{tr}\mathbf{B}\mathbf{B}' \\ &= n(\mathrm{tr}\mathbf{V} - 2\mathrm{tr}\mathbf{S}_{\mathrm{XZ}}\mathbf{B}' + \mathrm{tr}\mathbf{B}\mathbf{B}') = f(\mathbf{S}_{\mathrm{XZ}}, \mathbf{B}). \end{aligned} \tag{18.18}$$

That is, (18.5) can be expressed as a function of $\mathbf{S}_{\mathrm{XZ}}$ and $\mathbf{B}$, but not $\mathbf{Z}$. Furthermore, only if the covariance matrix $\mathbf{V}$ is available, the original data $\mathbf{X}$ is indispensable for obtaining the solution of $\mathbf{B} = [\mathbf{A}, \boldsymbol{\Psi}^{1/2}]$. How the formulas (18.14–18.17) are derived is explained in the next two sections.

18.5 Estimation of Covariances Between Variables and Factor Scores

The goal of this section is to finally show that the covariance matrix (18.13) for the optimal $\mathbf{Z}$ is expressed as (18.15). In order to achieve this, we need three paragraphs and a note, in which it is explained how the optimal $\mathbf{Z}$ minimizing (18.5) under (18.6) and (18.7) are expressed for $\mathbf{B}$ given.

We should note $f(\mathbf{Z}, \mathbf{B}) = n\mathrm{tr}\mathbf{V} - 2\mathrm{tr}\mathbf{X}'\mathbf{Z}\mathbf{B}' + n\mathrm{tr}\mathbf{B}\mathbf{B}'$ in (18.18), where only $-2\mathrm{tr}\mathbf{X}'\mathbf{Z}\mathbf{B}'$ is a function of $\mathbf{Z}$. Thus, the minimization of (18.18) over $\mathbf{Z}$ amounts to maximizing

$$g(\mathbf{Z}) = \frac{1}{n}\mathrm{tr}\mathbf{X}'\mathbf{Z}\mathbf{B}' = \frac{1}{n}\mathrm{tr}(\mathbf{X}\mathbf{B})'\mathbf{Z} = \mathrm{tr}\left(\frac{1}{\sqrt{n}}\mathbf{X}\mathbf{B}\right)'\left(\frac{1}{\sqrt{n}}\mathbf{Z}\right). \tag{18.19}$$

Now, let us only consider the constraint (18.7), ignoring (18.6). As explained later in Note 18.2, the maximization of (18.19) subject to (18.7) is attained for

$$\mathbf{Z} = \sqrt{n}\widetilde{\mathbf{P}}\widetilde{\mathbf{Q}}' = \sqrt{n}\mathbf{P}\mathbf{Q}' + \sqrt{n}\mathbf{P}_\perp\mathbf{Q}_\perp' = \mathbf{X}\mathbf{B}\mathbf{Q}\boldsymbol{\Theta}^{-1}\mathbf{Q}' + \sqrt{n}\mathbf{P}_\perp\mathbf{Q}_\perp'. \tag{18.20}$$

Here, $\widetilde{\mathbf{P}} = [\mathbf{P}, \mathbf{P}_\perp]$ and $\widetilde{\mathbf{Q}} = [\mathbf{Q}, \mathbf{Q}_\perp]$ are the $n \times (p + m)$ and $(p + m) \times (p + m)$ block matrices, respectively, and $\boldsymbol{\Theta}$ is an $p \times p$ diagonal matrix whose diagonal elements are positive, with the last identity in (18.20) and the blocks of $\tilde{\mathbf{P}}$ and $\tilde{\mathbf{Q}}$ detailed in the next paragraph.

The matrices $\mathbf{P}$ $(n \times p)$, $\mathbf{Q}$ $((p + m) \times p)$, and, $\boldsymbol{\Theta}$ are obtained through the *singular value decomposition* (*SVD*) of $n^{-1/2}\mathbf{X}\mathbf{B}$, which is defined as

$$\frac{1}{\sqrt{n}}\mathbf{X}\mathbf{B} = \mathbf{P}\boldsymbol{\Theta}\mathbf{Q}' \tag{18.21}$$

with $\mathbf{P}'\mathbf{P} = \mathbf{Q}'\mathbf{Q} = \mathbf{I}_p$. The remaining $\mathbf{P}_\perp$ $(n \times m)$ and $\mathbf{Q}_\perp$ $((p + m) \times m)$ are the *column-orthonormal orthogonal complement* (*OC*) *matrices* of $\mathbf{P}$ and $\mathbf{Q}$, respectively. Recall the explanation in Sect. 17.5 of OC matrices. That is, $\mathbf{P}_\perp$ and $\mathbf{Q}_\perp$ allow $\widetilde{\mathbf{P}}$ and $\widetilde{\mathbf{Q}}$ to satisfy

$$\widetilde{\mathbf{P}}'\widetilde{\mathbf{P}} = \widetilde{\mathbf{Q}}'\widetilde{\mathbf{Q}} = \mathbf{I}_{p+m}, \text{ i.e., } \begin{bmatrix} \mathbf{P}'\mathbf{P} & \mathbf{P}'\mathbf{P}_\perp \\ \mathbf{P}_\perp'\mathbf{P} & \mathbf{P}_\perp'\mathbf{P}_\perp \end{bmatrix} = \begin{bmatrix} \mathbf{Q}'\mathbf{Q} & \mathbf{Q}'\mathbf{Q}_\perp \\ \mathbf{Q}_\perp'\mathbf{Q} & \mathbf{Q}_\perp'\mathbf{Q}_\perp \end{bmatrix} = \begin{bmatrix} \mathbf{I}_p & \\ & \mathbf{I}_m \end{bmatrix}, \tag{18.22}$$

with the blank blocks in the right matrix filled with zeros. The equality $n^{1/2}\mathbf{P}\mathbf{Q}' = \mathbf{X}\mathbf{B}\mathbf{Q}\boldsymbol{\Theta}^{-1}\mathbf{Q}'$ in (18.20) is derived by post-multiplying both sides of (18.21) by $n^{1/2}\mathbf{Q}\boldsymbol{\Theta}^{-1}\mathbf{Q}'$.

Note 18.2. ten Berge's (1993) Theorem for Obtaining a Higher Rank Matrix

The factor score matrix $\mathbf{Z}$ is considered the *higher rank* matrix, in that (18.7) implies rank($\mathbf{Z}$) = $m + p$, which is greater than (18.11), with $\mathbf{X}\mathbf{B}$ in (18.11) corresponding to $\mathbf{Z}$ as in (18.19).

We can use Definition A.4.1 and Theorem A.4.1 (in Appendix A.4.1) to prove that (18.19) is maximized for (18.20) subject to (18.7), as follows: Substituting (18.21) in (18.19), this is rewritten as $g(\mathbf{Z}) = \text{tr}\mathbf{Q}\boldsymbol{\Theta}\mathbf{P}'\left(n^{-1/2}\mathbf{Z}\right) = \text{tr}\mathbf{P}'\left(n^{-1/2}\mathbf{Z}\right)\mathbf{Q}\boldsymbol{\Theta}$. From (18.7), $\mathbf{P}'\mathbf{P} = \mathbf{Q}'\mathbf{Q} = \mathbf{I}_p$, and (A.4.2), $\mathbf{P}'(n^{-1/2}\mathbf{Z})\mathbf{Q}$ is a $p \times p$ suborthonormal matrix with its rank $\leq p$. Further, $\boldsymbol{\Theta}$ is a diagonal matrix whose diagonal elements are positive. These properties lead to $g(\mathbf{Z}) = \text{tr}\mathbf{P}'\left(n^{-1/2}\mathbf{Z}\right)\mathbf{Q}\boldsymbol{\Theta} \leq \text{tr}\boldsymbol{\Theta}$. The upper limit tr$\boldsymbol{\Theta}$ is attained for (18.20) with

$$g\left(\sqrt{n}\widetilde{\mathbf{P}}\widetilde{\mathbf{Q}}'\right) = \mathrm{tr}\mathbf{P}'\left(\widetilde{\mathbf{P}}\widetilde{\mathbf{Q}}'\right)\mathbf{Q}\boldsymbol{\Theta} = \mathrm{tr}\mathbf{P}'\left(\mathbf{P}\mathbf{Q}' + \mathbf{P}_\perp\mathbf{Q}_\perp'\right)\mathbf{Q}\boldsymbol{\Theta} = \mathrm{tr}\mathbf{Q}'\mathbf{Q}\boldsymbol{\Theta} = \mathrm{tr}\boldsymbol{\Theta}$$

which is derived using the equation $\mathbf{P}'\mathbf{P}_\perp = {}_p\mathbf{O}_m$or $\mathbf{Q}'\mathbf{Q}_\perp = {}_p\mathbf{O}_m$ following from (18.22). Further, this equation shows that (18.20) satisfies (18.7) as

$$\frac{1}{n}\left(\sqrt{n}\widetilde{\mathbf{P}}\widetilde{\mathbf{Q}}'\right)'\sqrt{n}\widetilde{\mathbf{P}}\widetilde{\mathbf{Q}}' = \widetilde{\mathbf{Q}}\widetilde{\mathbf{P}}'\widetilde{\mathbf{P}}\widetilde{\mathbf{Q}}' = \widetilde{\mathbf{Q}}\widetilde{\mathbf{Q}}' = \mathbf{I}_{p+m},$$

where the fact has been used that column-orthonormal and square $\widetilde{\mathbf{Q}}$ also satisfies $\widetilde{\mathbf{Q}}\widetilde{\mathbf{Q}}' = \mathbf{I}_{p+m}$ as explained in Appendix A.1.2.

We should note that $\mathbf{P}_\perp$ and $\mathbf{Q}_\perp$, i.e., the *column-orthonormal OC matrices* of $\mathbf{P}$ and $\mathbf{Q}$, are not unique, as shown in Sect. 17.5. That is, infinitely many matirces $\mathbf{P}_\perp$ and $\mathbf{Q}_\perp$ exist that satisfy (18.22). Any of them can be substituted into $\mathbf{P}_\perp$ and $\mathbf{Q}_\perp$ in the block matrices $\widetilde{\mathbf{P}} = [\mathbf{P}, \mathbf{P}_\perp]$ and $\widetilde{\mathbf{Q}} = [\mathbf{Q}, \mathbf{Q}_\perp]$ used in above equations.

We should remember that (18.20) has been derived without considering the constraint (18.6). However, we can show that (18.6) is also satisfied by (18.20), using $n^{1/2}\mathbf{P}\mathbf{Q}' = \mathbf{X}\mathbf{B}\mathbf{Q}\boldsymbol{\Theta}^{-1}\mathbf{Q}'$ in (18.20) and $\mathbf{1}_n'\mathbf{X} = \mathbf{0}_p'$. They lead to $\mathbf{1}_n'\mathbf{P}\mathbf{Q}' = \mathbf{0}_{p+m}'$, and both sides of this equation are post-multiplied by $\mathbf{Q}$ to give $\mathbf{1}_n'\mathbf{P} = \mathbf{0}_p'$. This guarantees the existence of $\mathbf{P}_\perp$ satisfying $\mathbf{1}_n\mathbf{P}_\perp = \mathbf{0}_m'$ and (18.22), as found from the fact that (17.51) satisfies (17.49) for (17.41). The above equalities $\mathbf{1}_n'\mathbf{P} = \mathbf{0}_p'$ and $\mathbf{1}_n\mathbf{P}_\perp = \mathbf{0}_m'$ show that the pre-multiplication of the left side in (18.20) by $\mathbf{1}_n'$ leads to $\mathbf{1}_n'\mathbf{Z} = n^{1/2}\left(\mathbf{1}_n'\mathbf{P}\mathbf{Q}' + \mathbf{1}_n'\mathbf{P}_\perp\mathbf{Q}_\perp'\right) = \mathbf{0}_{p+m}'$, i.e., (18.20) satisfies (18.6). Thus, we can conclude that (18.5) is minimized for (18.20) under (18.6) and (18.7), for a given $\mathbf{B}$.

The covariance matrix $\mathbf{S}_{XZ} = n^{-1}\mathbf{X}'\mathbf{Z}$ for the optimal $\mathbf{Z}$ (18.20) is given by (18.15) as follows. Post-multiplying the both side of (18.21) by $\mathbf{B}^+$ leads to

$$\frac{1}{\sqrt{n}}\mathbf{X} = \mathbf{P}\boldsymbol{\Theta}\mathbf{Q}'\mathbf{B}^+, \quad \text{or equivalently}, \quad \frac{1}{\sqrt{n}}\mathbf{X}' = \mathbf{B}^{+\prime}\mathbf{Q}\boldsymbol{\Theta}\mathbf{P}', \tag{18.23}$$

from (18.12). The right equation in (18.23) is multiplied by $n^{-1/2}$ and post-multiplied by (18.20) to give $n^{-1}\mathbf{X}'\mathbf{Z} = n^{-1/2}\mathbf{B}^{+\prime}\mathbf{Q}\boldsymbol{\Theta}\mathbf{P}'\left(n^{1/2}\mathbf{P}\mathbf{Q}' + n^{1/2}\mathbf{P}_\perp\mathbf{Q}_\perp'\right)$, and this equation is rewritten as (18.15), using (18.13) and (18.22). Here, it should be noted that the update by (18.15) requires $\mathbf{B}$, $\mathbf{Q}$, and $\boldsymbol{\Theta}$. Among the three matrices, $\mathbf{B}$ is given in the next section, while $\mathbf{Q}$ and $\boldsymbol{\Theta}$ are given through the *EVD* in (18.14) following from the pre-multiplication of (18.21) by its transpose: $\left(n^{-1/2}\mathbf{X}\mathbf{B}\right)'\left(n^{-1/2}\mathbf{X}\mathbf{B}\right) = \left(\mathbf{P}\boldsymbol{\Theta}\mathbf{Q}'\right)'\mathbf{P}\boldsymbol{\Theta}\mathbf{Q}'$, i.e., $\mathbf{B}'\left(n^{-1}\mathbf{X}'\mathbf{X}\right)\mathbf{B} = \mathbf{Q}\boldsymbol{\Theta}\mathbf{P}'\mathbf{P}\boldsymbol{\Theta}\mathbf{Q}'$.

18.6 Estimation of Loadings and Unique Variances

For a given $\mathbf{S}_{XZ}$, the optimal $\mathbf{B}$ is given by (18.16). This follows from decomposing that loss function (18.5) as

$$f(\mathbf{Z},\mathbf{B}) = \|\mathbf{X}-\mathbf{Z}\mathbf{B}'\|^2 = \left\|\mathbf{X}-\mathbf{Z}\mathbf{S}'_{XZ}\right\|^2 + n\|\mathbf{S}_{XZ}-\mathbf{B}\|^2. \tag{18.24}$$

This decomposition is derived from the fact that (18.5) can be rewritten as $\left\|\mathbf{X}-\mathbf{Z}\mathbf{S}'_{XZ}+\mathbf{Z}\mathbf{S}'_{XZ}-\mathbf{Z}\mathbf{B}'\right\|^2 = \left\|\mathbf{X}-\mathbf{Z}\mathbf{S}'_{XZ}\right\|^2 + w - 2h$. Here, $w = \left\|\mathbf{Z}\mathbf{S}'_{XZ}-\mathbf{Z}\mathbf{B}'\right\|^2 = n\|\mathbf{S}_{XZ}-\mathbf{B}\|^2$ since of (18.7), and $h = \mathrm{tr}\left(\mathbf{X}-\mathbf{Z}\mathbf{S}'_{XZ}\right)'\left(\mathbf{Z}\mathbf{S}'_{XZ}-\mathbf{Z}\mathbf{B}'\right) = n\mathrm{tr}\mathbf{S}_{XZ}\mathbf{S}_{XZ'} - n\mathrm{tr}\mathbf{S}_{XZ}\mathbf{B}' - n\mathrm{tr}\mathbf{S}_{XZ}\mathbf{S}'_{XZ} + n\mathrm{tr}\mathbf{S}_{XZ}\mathbf{B}' = 0$ since of (18.7) and (18.13). Thus, we have (18.24).

On the right side of (18.24), $\mathbf{B} = [\mathbf{A}, \boldsymbol{\Psi}^{1/2}]$ appears only in $\|\mathbf{S}_{XZ}-\mathbf{B}\|^2 = \left\|[\mathbf{S}_{XF},\mathbf{S}_{XU}] - \left[\mathbf{A},\boldsymbol{\Psi}^{1/2}\right]\right\|^2$. This can further be decomposed as

$$\|\mathbf{S}_{XZ}-\mathbf{B}\|^2 = \|\mathbf{S}_{XF}-\mathbf{A}\|^2 + \left\|\mathrm{diag}(\mathbf{S}_{XU}) - \boldsymbol{\Psi}^{1/2}\right\|^2 + \|\mathbf{S}_{XU} - \mathrm{diag}(\mathbf{S}_{XU})\|^2, \tag{18.25}$$

using the fact that $\boldsymbol{\Psi}^{1/2}$ is diagonal. On the right side, the part relevant to $\mathbf{B} = [\mathbf{A}, \boldsymbol{\Psi}]$ is $\|\mathbf{S}_{XF}-\mathbf{A}\|^2 + \left\|\mathrm{diag}(\mathbf{S}_{XU}) - \boldsymbol{\Psi}^{1/2}\right\|^2$, whose lower limit zero is attained for (18.16).

Finally, we show why (18.17) is the *standardized loss function value*. Let us inspect $\mathrm{tr}\mathbf{S}_{XZ}\mathbf{B}'$ in loss function (18.18). This can be rewritten using $\mathbf{B}' = [\mathbf{A}, \boldsymbol{\Psi}^{1/2}]'$ and (18.13) as $\mathrm{tr}\mathbf{S}_{XZ}\mathbf{B}' = \mathrm{tr}\mathbf{S}_{XF}\mathbf{A}' + \mathrm{tr}\mathbf{S}_{XU}\boldsymbol{\Psi}^{1/2}$. Substituting (18.16) into $\mathbf{B}$ in $\mathrm{tr}\mathbf{S}_{XZ}\mathbf{B}'$, we have

$$\mathrm{tr}\mathbf{S}_{XZ}\mathbf{B}' = \mathrm{tr}\mathbf{A}\mathbf{A}' + \mathrm{tr}\mathbf{S}_{XU}\boldsymbol{\Psi}^{1/2} = \mathrm{tr}\mathbf{A}\mathbf{A}' + \mathrm{tr}\boldsymbol{\Psi}^{1/2}\boldsymbol{\Psi}^{1/2} = \mathrm{tr}\mathbf{A}\mathbf{A}' + \mathrm{tr}\boldsymbol{\Psi} = \mathrm{tr}\mathbf{B}\mathbf{B}' \tag{18.26}$$

where we have used $\mathrm{tr}\mathbf{S}_{XU}\boldsymbol{\Psi}^{1/2} = \mathrm{tr}\left\{\mathrm{diag}(\mathbf{S}_{XU})\boldsymbol{\Psi}^{1/2}\right\}$, since $\boldsymbol{\Psi}^{1/2}$ is diagonal. Using (18.26) in (18.18), its attained value is found to be

$$f(\mathbf{S}_{XZ},\mathbf{B}) = n(\mathrm{tr}\mathbf{V} - 2\mathrm{tr}\mathbf{B}\mathbf{B}' + \mathbf{tr}\mathbf{B}\mathbf{B}') = n(\mathrm{tr}\mathbf{V} - \mathrm{tr}\mathbf{B}\mathbf{B}'). \tag{18.27}$$

By dividing this by $n\mathrm{tr}\mathbf{V}$, we have (18.17), which is a standardized index in that it takes a value within the range [0, 1]. The property of (18.17) ≥ 0 follows from (18.27) ≥ 0 which is derived from the fact (18.5) ≥ 0 and thus (18.18) ≥ 0. On the other hand, (18.17) ≤ 1 follows from the fact that (18.27) ≥ 0 allows us to find $\mathrm{tr}\mathbf{V} \geq \mathrm{tr}\mathbf{B}\mathbf{B}' = \|\mathbf{B}\|^2 \geq 0$.

18.7 Identifiability of the Model Part and Residuals

In this and next sections, the *solutions* for $\mathbf{Z} = [\mathbf{F}, \mathbf{U}]$ and $\mathbf{B} = [\mathbf{A}, \boldsymbol{\Psi}]$ are expressed as $\widehat{\mathbf{Z}} = [\widehat{\mathbf{F}}, \widehat{\mathbf{U}}]$, and $\widehat{\mathbf{B}} = \left[\widehat{\mathbf{A}}, \widehat{\boldsymbol{\Psi}}^{1/2}\right]$, with $\widehat{\mathbf{F}}$, $\widehat{\mathbf{U}}$, $\widehat{\mathbf{A}}$, and $\widehat{\boldsymbol{\Psi}}^{1/2}$ denoting the solutions of $\mathbf{F}$, $\mathbf{U}$, $\mathbf{A}$, and $\boldsymbol{\Psi}$, respectively. Further, we use $\mathbf{S}_{\mathrm{X}\widehat{\mathrm{Z}}} = n^{-1}\mathbf{X}'\widehat{\mathbf{Z}}$ for the p-variables $\times$ $(m + p)$-factors covariance matrix based on the solution of $\mathbf{Z}$.

One *advantage* of MDFA over LVFA in Chap. 12 is that the error matrix $\mathbf{E}$ based on the solution, i.e., the *residual matrix*

$$\widehat{\mathbf{E}} = \mathbf{X} - \left(\widehat{\mathbf{F}}\widehat{\mathbf{A}}' + \widehat{\mathbf{U}}\widehat{\boldsymbol{\Psi}}^{1/2}\right) = \mathbf{X} - \widehat{\mathbf{Z}}\widehat{\mathbf{B}}', \tag{18.28}$$

can be *obtained* in MDFA. This follows from the fact that the *model part* $\widehat{\mathbf{F}}\widehat{\mathbf{A}}' + \widehat{\mathbf{U}}\widehat{\boldsymbol{\Psi}}^{1/2} = \widehat{\mathbf{Z}}\widehat{\mathbf{B}}'$ can be *computed* as shown next:

Note 18.3. Identifiability of the FA Model Part (Adachi & Trendafilov's (2018a) Theorem 3.1)

$$\widehat{\mathbf{Z}}\widehat{\mathbf{B}}' = \sqrt{n}\mathbf{P}\mathbf{Q}'\widehat{\mathbf{B}}' = \mathbf{X}\widehat{\mathbf{B}}\mathbf{Q}\boldsymbol{\Theta}^{-1}\mathbf{Q}'\widehat{\mathbf{B}}' = \mathbf{X}\mathbf{V}^{-1}\mathbf{S}_{\mathrm{X}\widehat{\mathrm{Z}}}\widehat{\mathbf{B}}' = \mathbf{X}\widehat{\mathbf{B}}\mathbf{S}'_{\mathrm{X}\widehat{\mathrm{Z}}}\mathbf{V}^{-1}. \tag{18.29}$$

This is proved in the following paragraphs.

The first equality $\widehat{\mathbf{Z}}\widehat{\mathbf{B}}' = n^{1/2}\mathbf{P}\mathbf{Q}'\widehat{\mathbf{B}}'$ in (18.29) is derived as follows. Pre-multiplying both sides of (18.21) by $n^{1/2}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ leads to $\widehat{\mathbf{B}} = n^{1/2}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}\boldsymbol{\Theta}\mathbf{Q}'$, which implies $\widehat{\mathbf{B}}\mathbf{Q}_\perp = {}_p\mathbf{O}_m$ since $\mathbf{Q}'\mathbf{Q}_\perp = {}_p\mathbf{O}_m$ follows from (18.22). The equalities $\widehat{\mathbf{B}}\mathbf{Q}_\perp = {}_p\mathbf{O}_m$ and (18.20) lead to $\widehat{\mathbf{Z}}\widehat{\mathbf{B}}' = \left(n^{1/2}\mathbf{P}\mathbf{Q}' + n^{1/2}\mathbf{P}_\perp\mathbf{Q}'_\perp\right)\widehat{\mathbf{B}}' = n^{1/2}\mathbf{P}\mathbf{Q}'\widehat{\mathbf{B}}'$.

The second identity $n^{1/2}\mathbf{P}\mathbf{Q}'\widehat{\mathbf{B}}^{'} = \mathbf{X}\widehat{\mathbf{B}}\mathbf{Q}\boldsymbol{\Theta}^{-1}\mathbf{Q}'\widehat{\mathbf{B}}^{'}$ in (18.29) follows from the equation $n^{1/2}\mathbf{P}\mathbf{Q}' = \mathbf{X}\widehat{\mathbf{B}}\mathbf{Q}\boldsymbol{\Theta}^{-1}\mathbf{Q}'$ in (18.20) with $\mathbf{B}$ rewritten as $\widehat{\mathbf{B}}$: Both sides of this equation, post-multiplied by $\widehat{\mathbf{B}}^{'}$, provide the second identity.

Before deriving the third equality $\mathbf{X}\widehat{\mathbf{B}}\mathbf{Q}\boldsymbol{\Theta}^{-1}\mathbf{Q}'\widehat{\mathbf{B}}^{'} = \mathbf{X}\mathbf{V}^{-1}\mathbf{S}_{\mathrm{X}\widehat{\mathrm{Z}}}\widehat{\mathbf{B}}^{'}$, we must prove

$$\mathbf{S}_{\mathrm{X}\widehat{\mathrm{Z}}} = \mathbf{V}\widehat{\mathbf{B}}\mathbf{Q}\boldsymbol{\Theta}^{-1}\mathbf{Q}'. \tag{18.30}$$

This can be shown by pre-multiplying both sides of (18.20) by $n^{-1}\mathbf{X}'$, which leads to $\mathbf{S}_{\mathbf{X}\widehat{\mathbf{Z}}} = \mathbf{V}\widehat{\mathbf{B}}\mathbf{Q}\boldsymbol{\Theta}^{-1}\mathbf{Q}' + n^{1/2}\mathbf{X}'\mathbf{P}_{\perp}\mathbf{Q}'_{\perp}$. Here, $\mathbf{X}'\mathbf{P}_{\perp}\mathbf{Q}'_{\perp} =_p \mathbf{O}_p$, from (18.23) and $\mathbf{P}'\mathbf{P}_{\perp} =_p \mathbf{O}_p$. Thus, we have (18.30), which implies $\mathbf{V}^{-1}\mathbf{S}_{\mathbf{X}\widehat{\mathbf{Z}}}\widehat{\mathbf{B}}' = \widehat{\mathbf{B}}\mathbf{Q}\boldsymbol{\Theta}^{-1}\mathbf{Q}'\widehat{\mathbf{B}}'$. Pre-multiplying both sides by $\mathbf{X}$ leads to the third equality in (18.29).

The last identity $\mathbf{X}\mathbf{V}^{-1}\mathbf{S}_{\mathbf{X}\widehat{\mathbf{Z}}}\widehat{\mathbf{B}}' = \mathbf{X}\widehat{\mathbf{B}}\mathbf{S}'_{\mathbf{X}\widehat{\mathbf{Z}}}\mathbf{V}^{-1}$ follows from $\widehat{\mathbf{B}}\mathbf{Q}\boldsymbol{\Theta}^{-1}\mathbf{Q}'\widehat{\mathbf{B}}'$ in (18.29) being symmetric. Thus, $\mathbf{V}^{-1}\mathbf{S}_{\mathbf{X}\widehat{\mathbf{Z}}}\widehat{\mathbf{B}}' = \widehat{\mathbf{B}}\mathbf{Q}\boldsymbol{\Theta}^{-1}\mathbf{Q}'\widehat{\mathbf{B}}'$ equals its transpose: $\mathbf{V}^{-1}\mathbf{S}_{\mathbf{X}\widehat{\mathbf{Z}}}\widehat{\mathbf{B}}' = \widehat{\mathbf{B}}\mathbf{S}'_{\mathbf{X}\widehat{\mathbf{Z}}}\mathbf{V}^{-1}$. Pre-multiplying both sides by $\mathbf{X}$ leads to the last identity in (18.29). This completes the proof for (18.29).

Here, we should note that (18.30) may be substituted for (18.15) in Note 18.1.

Table 18.2(A) and (B) show (18.29) and (18.28) values, respectively, which were obtained with MDFA for the standard scores of the data set in Table 10.1. Here, we can find that the absolute values of the residuals in (B) are much smaller than those of the model part in (A). This allows us to consider that the FA model $\mathbf{FA}' + \mathbf{U}\boldsymbol{\Psi}^{1/2}$ fit well to the data set $\mathbf{X}$. However, we can also find a few residuals of large absolute values in $\widehat{\mathbf{E}} = \left(\hat{e}_{ij}\right)$. Such residuals suggest that the corresponding observations deviate from the FA model. For example, $\hat{e}_{71} = 0.108$ is relatively large in Table 18.2(B). This suggests that the score of the seventh participant for A (aggressiveness) may be substantially larger than the score predicted by the FA model.

The "Residual" in Table 18.1 shows $r_j = n^{-1}\left\|\widehat{\mathbf{e}}_j\right\|^2$ with $\hat{\mathbf{e}}_j$ the jth column of (18.28). That is, r_j is the average of the squared residuals. This can be interpreted as the size of residuals for variable j, which remain unaccounted by the FA model part (18.29). We can also call $r_j = n^{-1}\left\|\widehat{\mathbf{e}}_j\right\|^2$ the residual variance for variable j since

$$\mathbf{1}'_n\widehat{\mathbf{e}}_j = 0, \quad \text{or equivalently,} \quad \mathbf{1}'_n\widehat{\mathbf{E}} = \mathbf{0}_{p'} \tag{18.31}$$

which follows from $\mathbf{X}$ and $\widehat{\mathbf{Z}}\widehat{\mathbf{B}}'$ on the right side of (18.28) being centered. The residual matrix $\widehat{\mathbf{E}}$ and the size of residuals $r_j = n^{-1}\left\|\widehat{\mathbf{e}}_j\right\|^2$ cannot be estimated in the LVFA approach, as the term associated with $\mathbf{E}$ does not appear in (18.8).

Table 18.2 Parts of the resulting model values (A) and residuals (B)

Ind.	A	C	I	B	T	V	H	P
(A) Model part: $\widehat{\mathbf{F}}\widehat{\mathbf{A}}' + \widehat{\mathbf{U}}\widehat{\mathbf{\Psi}}^{1/2} = \widehat{\mathbf{Z}}\widehat{\mathbf{B}}'$								
1	2.835	1.359	1.974	−2.393	1.940	1.986	−1.205	1.639
2	−2.044	−2.356	−1.203	1.690	−2.810	−1.189	0.921	−1.841
3	−0.047	0.414	0.493	0.363	1.408	−0.613	0.329	0.270
4	−0.610	0.473	−0.418	−1.679	1.313	0.037	0.920	0.976
5	0.708	−0.517	0.413	0.303	0.186	0.067	0.368	0.227
6	−0.693	−0.528	−1.133	−0.320	0.163	−1.195	−0.133	−0.378
7	0.594	1.301	−0.336	−0.367	1.409	−1.155	0.340	1.680
•	•	•	•	•	•	•	•	•
•	•	•	•	•	•	•	•	•
•	•	•	•	•	•	•	•	•
98	−0.699	1.359	−1.121	0.315	0.747	−1.135	1.006	−1.160
99	−0.071	−1.459	−1.167	2.284	−1.027	0.055	0.296	0.972
100	−0.692	−0.493	0.443	1.047	−1.605	0.084	−0.674	−1.830
(B) Residuals: $\hat{\mathbf{E}} = \mathbf{X} - \left(\widehat{\mathbf{F}}\widehat{\mathbf{A}}' + \widehat{\mathbf{U}}\widehat{\mathbf{\Psi}}^{1/2}\right)$								
Ind.	A	C	I	B	T	V	H	P
1	−0.048	0.004	0.046	0.042	0.006	−0.045	−0.043	0.054
2	−0.035	−0.014	0.030	−0.022	0.010	0.019	0.011	−0.023
3	0.054	0.016	−0.070	−0.035	−0.056	0.065	0.058	0.001
4	−0.078	−0.044	0.043	−0.002	0.039	0.038	0.012	0.006
5	−0.006	0.013	0.010	0.026	−0.020	0.008	0.019	0.044
6	0.005	0.024	−0.041	−0.022	0.003	0.025	−0.025	−0.063
7	0.108	0.061	−0.039	0.026	−0.057	−0.014	0.047	0.014
•	•	•	•	•	•	•	•	•
•	•	•	•	•	•	•	•	•
•	•	•	•	•	•	•	•	•
98	0.011	0.003	−0.052	0.014	0.012	−0.034	−0.074	0.008
99	0.078	0.022	−0.007	0.053	0.006	0.020	0.091	0.010
100	0.004	−0.011	−0.020	−0.050	−0.008	−0.009	−0.029	−0.034

18.8 Factor Scores as Higher Rank Approximations

The estimation of the *factor score* matrix $\mathbf{Z}$ can be viewed as a *higher rank approximation* problem. This is because, we can use (18.7) to rewrite the MDFA loss function (18.5) as

$$f(\mathbf{Z}|\mathbf{B}) = \|\mathbf{Z} - \mathbf{X}\mathbf{B}\|^2 + \|\mathbf{X}\|^2 + n\|\mathbf{B}\|^2 - \|\mathbf{X}\mathbf{B}\|^2 - n(p+m). \tag{18.32}$$

Here, we find that only the term $\|\mathbf{Z} - \mathbf{XB}\|^2$ is a function of $\mathbf{Z}$. Thus, the minimization of (18.5) over $\mathbf{Z}$ for a given $\mathbf{B}$ amounts to minimizing $\|\mathbf{Z} - \mathbf{XB}\|^2$ over $\mathbf{Z}$. This is the problem of approximating $\mathbf{XB}$ by the *higher rank* matrix $\mathbf{Z}$, as $\text{rank}(\mathbf{Z}) = m + p > \text{rank}(\mathbf{XB}) = p$, as described in Note 18.2.

The solution of $\mathbf{Z}$ for the higher-rank approximation is given by (18.20), which can be rewritten as

$$\widehat{\mathbf{Z}} = \widehat{\mathbf{Z}}_1 + \widehat{\mathbf{Z}}_2 \tag{18.33}$$

with $\widehat{\mathbf{Z}}_1 = n^{1/2}\mathbf{PQ}' = \mathbf{XBQ}\Theta^{-1}\mathbf{Q}'$ and $\widehat{\mathbf{Z}}_2 = n^{1/2}\mathbf{P}_\perp\mathbf{Q}'_\perp$. Here, $\widehat{\mathbf{Z}}_1$ can be *uniquely determined*, while $\widehat{\mathbf{Z}}_2 = n^{1/2}\mathbf{P}_\perp\mathbf{Q}'_\perp$ *cannot*, since infinitely many matrices $\mathbf{P}_\perp(n \times m)$ and $\mathbf{Q}_\perp(p \times m)$ exist which allow to $\widetilde{\mathbf{P}}$ and $\widetilde{\mathbf{Q}}$ to satisfy (18.22), as described in Note 18.2. Hence, the optimal factor scores $\widehat{\mathbf{Z}}$ are *not unique*, despite the fact that its pre-multiplication by $\widehat{\mathbf{B}}'$ yields the model part $\widehat{\mathbf{Z}}\widehat{\mathbf{B}}'$ *determined uniquely* as in (18.29).

The two matrices on the right side of (18.33) satisfy

$$\widehat{\mathbf{Z}}_1'\widehat{\mathbf{Z}}_2 =_p \mathbf{O}_m, \tag{18.34}$$

from (18.22), with

$$\left\|\widehat{\mathbf{Z}}_2\right\| = \sqrt{nm} \tag{18.35}$$

following from $\left\|\widehat{\mathbf{Z}}_2\right\|^2 = n\text{tr}\mathbf{Q}_\perp\mathbf{P}'_\perp\mathbf{P}_\perp\mathbf{Q}'_\perp = n\text{tr}\mathbf{Q}'_\perp\mathbf{Q}_\perp = n\text{tr}\mathbf{I}_m$. On the basis of these properties, Adachi and Trendafilov (2018a) have presented the diagram in Fig. 18.2, where the matrices $\widehat{\mathbf{Z}}$ and $\widehat{\mathbf{Z}}_1$ in (18.33) are depicted as arrows (or vectors). Here, the arrows for $\widehat{\mathbf{Z}}$ and $\widehat{\mathbf{Z}}_1$ are illustrated so that $\widehat{\mathbf{Z}}_1$ and $\widehat{\mathbf{Z}} - \widehat{\mathbf{Z}}_1 = \widehat{\mathbf{Z}}_2$ intersect orthogonally, from (18.34). Further, the endpoint of the arrow for $\widehat{\mathbf{Z}}$ is depicted to form a *circle* whose center is $\widehat{\mathbf{Z}}_1$ and radius is $(nm)^{1/2}$, from (18.35). In other words, in Fig. 18.2, $\widehat{\mathbf{Z}}$ forms a *cone*, which traces a circle around its center at

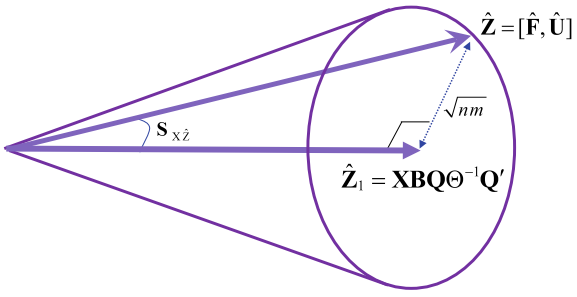


Fig. 18.2 Adachi and Trendafilov's cone of common-unique factor scores

$\widehat{\mathbf{Z}}_1$. Here, any $\widehat{\mathbf{Z}}$ whose endpoint is on the *circle* is *optimal*. This suggests that imposing an additional constraint on $\mathbf{Z}$ allows for a useful $\widehat{\mathbf{Z}}$ to be found, as shown in Uno, Adachi, and Trendafilov (2019).

18.9 Bibliographical Notes

As already mentioned, the matrix decomposition formulation of FA (MDFA) was proposed in 2001 by Professor Henk A. L. Kiers at the University of Groningen, as found in Sŏcan's thesis (2013). Independently, de Leeuw (2004) also presented a description of MDFA. Later, Unkel and Trendafilov (2010) reviewed some formulations of FA detailing MDFA. In the above literature, the MDFA algorithms described needed the *original data matrix* $\mathbf{X}$. In contrast, the algorithm in this chapter which *only needs the covariance matrix* $\mathbf{V}$ was proposed by Adachi (2012).

Some properties of the MDFA solution described in this chapter have been detailed in Adachi and Trendafilov (2018a) along with other properties. Also, Stegeman (2016) has discussed properties of the MDFA solution. Further, Stegeman (2016) has proposed a *constrained version* of MDFA which is not treated in this book. It is argued that FA can be classified into the *three* types, latent variable FA (LVFA), MDFA, and Stegeman's (2016) constrained MDFA, in Adachi's (2019) comprehensive review of FA formulations.

Exercises

18.1. Let us define the rows and columns of the matrices in (18.5) as $\mathbf{E} = [\mathbf{e}_1, \ldots, \mathbf{e}_p], \mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_p], \mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_p]$, $\mathbf{B} = \begin{bmatrix} \mathbf{b}_1' \\ \vdots \\ \mathbf{b}_p' \end{bmatrix}$, and $\mathbf{A} = \begin{bmatrix} \mathbf{a}_1' \\ \vdots \\ \mathbf{a}_p' \end{bmatrix}$. Show that (18.5) can be rewritten as $f(\mathbf{Z}, \mathbf{B}) = \sum_{j=1}^{p} f_j(\mathbf{Z}, \mathbf{b}_j)$ with

$$f_j(\mathbf{Z}, \mathbf{b}_j) = \|\mathbf{e}_j\|^2 = \left\|\mathbf{x}_j - \mathbf{F}\mathbf{a}_j - \psi_j^{1/2}\mathbf{u}_j\right\|^2 = \|\mathbf{x}_j - \mathbf{Z}\mathbf{b}_j\|^2. \tag{18.36}$$

18.2. Let $1 \times (p + m)$ vector $\mathbf{s}_j^{\mathrm{XZ}'}$ denote the jth row of (18.13) with $\mathbf{S}_{\mathrm{XZ}} = [\mathbf{s}_1^{\mathrm{XZ}}, \ldots, \mathbf{s}_p^{\mathrm{XZ}}]'$ $(p \times (p + m))$. Show that (18.36) can be decomposed as

$$f_j(\mathbf{Z}, \mathbf{b}_j) = \left\|\mathbf{x}_j - \mathbf{Z}_{\mathbf{s}_j^{\mathrm{XZ}}}\right\|^2 + n\left\|\mathbf{s}_j^{\mathrm{xz}} - \mathbf{b}_j\right\|^2 \tag{18.37}$$

under (18.7), using $f_j(\mathbf{Z}, \mathbf{b}_j) = \left\| \left(\mathbf{x}_i - \mathbf{Z}\mathbf{s}_j^{\mathrm{XZ}} \right) + \left(\mathbf{Z}\mathbf{s}_j^{\mathrm{XZ}} - \mathbf{Z}\mathbf{b}_j \right) \right\|^2$ (Adachi & Trendafilov, 2018a, Theorem 2.1).

18.3. Use (18.7) and (18.13) to show $n^{-1}\,\widehat{\mathbf{E}}'\widehat{\mathbf{Z}} = n^{-1}\left(\mathbf{X} - \widehat{\mathbf{Z}}\widehat{\mathbf{B}}'\right)'\widehat{\mathbf{Z}} = \mathbf{S}_{\mathrm{X}\widehat{\mathrm{Z}}} - \widehat{\mathbf{B}}$.

18.4. Let the covariance matrices of the residuals to the optimal common factor scores and unique factor scores be denoted as $\mathbf{S}_{\widehat{\mathrm{E}}\widehat{\mathrm{F}}} = n^{-1}\widehat{\mathbf{E}}'\widehat{\mathbf{F}}$ and $\mathbf{S}_{\widehat{\mathrm{E}}\widehat{\mathrm{U}}} = n^{-1}\widehat{\mathbf{E}}'\widehat{\mathbf{U}}$. Show that the equation shown in Exercise 18.3 implies

$$\mathbf{S}_{\widehat{\mathrm{E}}\widehat{\mathrm{F}}} =_p \mathbf{O}_m \quad \text{and} \quad \mathbf{S}_{\widehat{\mathrm{E}}\widehat{\mathrm{U}}} = \mathrm{Offd}\left(\mathbf{S}_{\mathrm{X}\widehat{\mathrm{U}}}\right), \tag{18.38}$$

where $\mathbf{S}_{\mathrm{X}\widehat{\mathrm{U}}} = n^{-1}\mathbf{X}'\widehat{\mathbf{U}}$ contains the covariances of the variables to the optimal unique factor scores, and $\mathrm{Offd}(\mathbf{S}_{\mathrm{X}\widehat{\mathrm{U}}}) = \mathbf{S}_{\mathrm{X}\widehat{\mathrm{U}}} - \mathrm{diag}(\mathbf{S}_{\mathrm{X}\widehat{\mathrm{U}}})$: The diagonal elements of $\mathrm{Offd}(\mathbf{S}_{\mathrm{X}\widehat{\mathrm{U}}})$ are zeros and its off-diagonal elements are those of $\mathbf{S}_{\mathrm{X}\widehat{\mathrm{U}}}$ (Adachi & Trendafilov, 2018a, Theorem 4.1).

18.5. Use (18.12), (18.14), and (18.15) to show that

$$\mathbf{S}_{\mathrm{X}\widehat{\mathrm{Z}}}\mathbf{S}'_{\mathrm{X}\widehat{\mathrm{Z}}} = \mathbf{V}\,: \tag{18.39}$$

the product of $\mathbf{S}_{\mathrm{X}\widehat{\mathrm{Z}}}$ (the covariance matrix of p variables to the $m + p$ optimal factor scores) and its transpose equals the inter-variable covariance matrix (Adachi & Trendafilov, 2018a, Lemma 4.1).

18.6. Note that $\mathbf{Z}$ in (18.20) can be rewritten as $\widehat{\mathbf{Z}}$, since (18.20) shows the solution of $\mathbf{Z}$. Use (18.7) and (18.39) to show that when (18.20) is substituted into $\mathbf{Z}$ in (18.37), the first term of in the right-hand side of (18.37) vanishes:

$$\left\| \mathbf{x}_j - \widehat{\mathbf{Z}}\mathbf{s}_j^{\mathrm{X}\widehat{\mathrm{Z}}} \right\|^2 = 0, \quad \text{i.e.,} \quad f_j\left(\widehat{\mathrm{Z}}, \mathbf{b}_j\right) = n\left\| \mathbf{s}_j^{\mathrm{X}\widehat{\mathrm{Z}}} - \mathbf{b}_j \right\|^2 \tag{18.40}$$

with $\mathbf{s}_j^{\mathrm{X}\widehat{\mathrm{Z}}\prime}$, the jth row of $\mathbf{S}_{\mathrm{X}\widehat{\mathrm{Z}}}$ (Adachi & Trendafilov, 2018a, Lemma 4.2).

18.7. Show that the loss function (18.5) or (18.24), in which the solution is substituted, can be expressed as

$$f(\widehat{\mathbf{Z}}, \widehat{\mathbf{B}}) = n\left\| \mathrm{Offd}\left(\mathbf{S}_{\mathrm{X}\widehat{\mathrm{U}}}\right) \right\|^2, \tag{18.41}$$

using (18.16), (18.38), and (18.40), with $\mathrm{Offd}(\mathbf{S}_{\mathrm{X}\widehat{\mathrm{U}}})$ defined in Exercise 18.4.

18.8. Consider what (18.41) implies. Hints are found in Adachi and Trendafilov (2018a, Sect. 4).

18.9. Discuss how the *confirmatory* FA based on MDFA can be formulated, by making note of (18.24) and (18.25).

18.10. Let the blocks of the matrices $\widehat{\mathbf{Z}}_1$ and $\widehat{\mathbf{Z}}_2$ in (18.33) be defined as $\widehat{\mathbf{Z}}_1 = [\mathbf{F}_1, \mathbf{U}_1]$ and $\widehat{\mathbf{Z}}_2 = [\mathbf{F}_2, \mathbf{U}_2]$, where $\mathbf{F}_1$ and $\mathbf{F}_2$ are $n \times m$, while $\mathbf{U}_1$ and $\mathbf{U}_2$ are $n \times p$. Uno et al., (2019) has proposed a *factor score identification* procedure, in which $\|[\mathbf{GC}, \mathbf{U}_1] - ([\mathbf{F}_1, \mathbf{U}_1] + [\mathbf{F}_2, \mathbf{U}_2])\|^2$ is minimized over $\mathbf{G}$, $\mathbf{C}$, and $[\mathbf{F}_2, \mathbf{U}_2]$ for the unique $\widehat{\mathbf{Z}}_1 = [\mathbf{F}_1, \mathbf{U}_1]$ given by MDFA, where $\mathbf{G} = (g_{ik})$ is an n (individuals) $\times$ K (clusters) membership matrix defined in (7.1) and (7.2), $\mathbf{C}$ is an unconstrained K (clusters) $\times$ m (common factors) matrix, and $[\mathbf{F}_2, \mathbf{U}_2]$ is constrained so as to satisfy (18.6) and (18.7). Discuss the purpose of the factor score identification procedure.

Chapter 19
Principal Component Analysis Versus Factor Analysis

In this chapter, we refer to exploratory factor analysis simply as factor analysis and consider the principal component analysis formulated as reduced rank approximation as in Chap. 5. *Principal component analysis* (*PCA*) and *factor analysis* (*FA*) can be performed for identical data sets, with the purpose of dimension reduction. This reduction means that p observed variables, i.e., the p-dimensional scores, are reduced to lower-dimensional scores. The lower dimensions correspond to the m *principal components* in PCA and the m *common factors* in FA, with $m < p$. A major purpose of this chapter is to introduce mathematical facts that *contrast PCA* and *FA solutions* for an identical data set. The facts elucidate *crucial differences* between *PCA* and *FA*, which can suggest *whether* PCA or FA should be used for a particular data set.

19.1 Motivational Examples

An identical data set can be analyzed by both principal component analysis (PCA) and factor analysis (FA) for the purpose of dimension reduction. In doing so, one is led to ask, "*How similar/different* are the resulting *PCA* and *FA solutions*?" To answer this question, we performed PCA and FA for the correlation matrices in Tables 19.1 and 19.2, where (5.25) was considered as a constraint in PCA. The resulting solutions are shown in Tables 19.3 and 19.4, where the loading matrices have been rotated by (orthogonal) varimax rotation (Chap. 13), with UV and Res the abbreviations for unique and residual variances, respectively. In the tables, the FA solutions are those for *matrix decomposition FA* (*MDFA*) and *latent variable FA* (*LVFA*), which were treated in Chaps. 18 and 12, respectively. As the residual variances for variables cannot be obtained in LVFA, the corresponding column is not presented in Tables 19.3 and 19.4. Apart from this point, the LVFA and MDFA solutions are found to be *almost equivalent*.

K. Adachi, *Matrix-Based Introduction to Multivariate Data Analysis*,
https://doi.org/10.1007/978-981-15-4103-2_19

Table 19.1 Correlation matrix for Adachi and Trendafiov's (2018a, p. 409) data set which is a part of Tanaka and Tarumi's (1995) test score data, with the upper triangular elements omitted

Variable	1	2	3	4	5
1. Japanese	1				
2. English	0.553	1			
3. Social studies	0.363	0.503	1		
4. Mathematics	0.447	0.330	0.287	1	
5. Sciences	0.388	0.279	0.076	0.563	1

We should note the following observations for the solutions in Tables 19.3 and 19.4:

[O1] We can find that a number of PCA loadings are boldfaced, where the absolute values of the boldfaced loadings are greater than their FA counterparts: the *magnitudes of PCA loadings* tend to be *greater* than the *FA ones*.

[O2] The *residual variances for PCA* are *greater* than *those for MDFA*.

[O3] The *residual variances for PCA* are *smaller* than the *unique variances in FA*.

[O1]–[O3] are *empirical* findings, which will not always be the case. However, some *mathematical* facts, which always hold and further *suggest* the tendencies [O1]–[O3], can be deduced by comparing *PCA* and *MDFA* solutions, as described in Sects. 19.4–19.6. These facts can also be *empirically generalized to LVFA* as described in Sect. 19.7.

19.2 Comparisons of Models

Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_p]$ be an n-individuals $\times$ p-variables centered data matrix, with $\mathbf{1}_n'\mathbf{X} = \mathbf{0}_p'$. From here to Sect. 19.6, we refer to MDFA simply as FA. As described in Sect. 18.1, FA can be modeled as (18.1) or (18.2):

$$\mathbf{X} = \mathbf{F}\mathbf{A}' + \mathbf{U}\boldsymbol{\Psi}^{1/2} + \mathbf{E}_{\mathrm{FA}}. \tag{19.1}$$

Here, the subscript FA has been attached to the error matrix $\mathbf{E}$ to distinguish it from that appearing in the next paragraph. As listed after (18.2), $\mathbf{F}$ ($n \times m$) contains common factor scores, $\mathbf{U}$ ($n \times p$) contains unique factor scores, $\mathbf{A}$ ($p \times m$) consists of factor loadings, and $\boldsymbol{\Psi}^{1/2}$ ($p \times p$) is the diagonal matrix whose j th diagonal element $\psi_j^{1/2}$ is the square root of the unique variance for variable j, with $p > m$.

By simply removing the *unique factor* part $\mathbf{U}\boldsymbol{\Psi}^{1/2}$ from (19.1), we have the PCA model (5.1) or (5.3) (Chap. 5). For distinguishing $\mathbf{F}\mathbf{A}'$ in (5.1) from that in (19.1), we substitute $\mathbf{P}\mathbf{C}'$ for $\mathbf{F}\mathbf{A}'$ in (5.1): PCA is modeled as

Table 19.2 Correlations among 12 personality traits (Yanai and Ichikawa, 2007, Table 19.2)

Variable		1	2	3	4	5	6	7	8	9	10	11	12
1	Extraversion	1											
2	Activity	0.525	1										
3	Empathy	0.235	0.311	1									
4	Novelty	0.280	0.418	0.091	1								
5	Durability	0.136	0.316	0.336	0.070	1							
6	Regularity	0.208	0.377	0.374	0.105	0.573	1						
7	Self-revelation	0.323	0.346	0.041	0.423	0.030	0.118	1					
8	Aggressiveness	0.049	0.189	–0.188	0.322	–0.110	0.099	0.374	1				
9	Uncooperativeness	–0.198	–0.038	–0.262	0.005	–0.052	–0.002	0.142	0.371	1			
10	Inferiority feeling	–0.342	–0.453	–0.052	–0.232	–0.220	–0.121	–0.064	0.083	0.299	1		
11	Nervousness	–0.189	–0.070	0.165	–0.053	0.046	0.197	0.153	0.248	0.175	0.386	1	
12	Depression	–0.245	–0.207	–0.025	0.003	–0.092	–0.047	0.130	0.314	0.403	0.518	0.592	1

Table 19.3 Solutions for the correlations in Table 19.1

Variable	PCA			MDFA				LVFA		
	Loadings		Res	Loadings		UV	Res	Loadings		UV
Japanese	**0.51**	**0.62**	0.13	0.38	0.60	0.50	0.001	0.37	0.61	0.50
English	**0.25**	**0.81**	0.08	0.21	0.76	0.37	0.002	0.21	0.76	0.38
Social Studies	–0.02	**0.86**	0.07	0.03	0.65	0.58	0.002	0.02	0.65	0.58
Mathematics	**0.80**	0.26	0.08	0.59	0.34	0.53	0.003	0.58	0.34	0.55
Sciences	**0.90**	0.02	0.03	0.89	0.10	0.19	0.001	0.90	0.11	0.17
Sum of squares	3.62		1.38	2.81		2.18	0.008	2.82		2.18

$$\mathbf{X} = \mathbf{PC}' + \mathbf{E}_{\mathrm{PC}}. \tag{19.2}$$

Here, the n (individuals) $\times$ m (components) matrix $\mathbf{P}$ contains PC scores, $\mathbf{C}$ ($p \times m$) consists of component loadings, and $\mathbf{E}_{\mathrm{PC}}$ ($n \times p$) contains errors.

The implication of $\mathbf{PC}'$ in the PCA model (19.2) can be illustrated through Fig. 19.1a: the variables $\mathbf{x}_1, \ldots, \mathbf{x}_p$ are *commonly* explained by the PC score vectors in $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_m]$ weighted by the loadings in $\mathbf{C} = (c_{jk})$, while the errors in $\mathbf{E}_{\mathrm{PC}}$ remain unexplained. At this point, we can call $\mathbf{PC}'$ the *common part*. On the other hand, the FA model (19.1) can be illustrated through Fig. 19.1b: the common factor vectors $\mathbf{F} = [\mathbf{f}_1, \ldots, \mathbf{f}_m]$ account for all variables with the loadings in $\mathbf{A} = (a_{jk})$ coefficients, while each unique factor vector in $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_p]$ uniquely affects the corresponding variable with the diagonal element $\psi_j^{1/2}$ in $\mathbf{\Psi}^{1/2}$ being a coefficient. Thus, $\mathbf{FA}'$ serves as the *common part* the same way $\mathbf{PC}'$ does in PCA, while $\mathbf{U\Psi}^{1/2}$ can be called the *unique part*, which is absent in PCA.

At this stage, we can clearly answer

Choose FA if the unique part should be extracted (19.3)

in regards to the question of whether FA or PCA should be used for a data set. Cases in which different answers should be given are presented in Sects. 19.4 and 19.5.

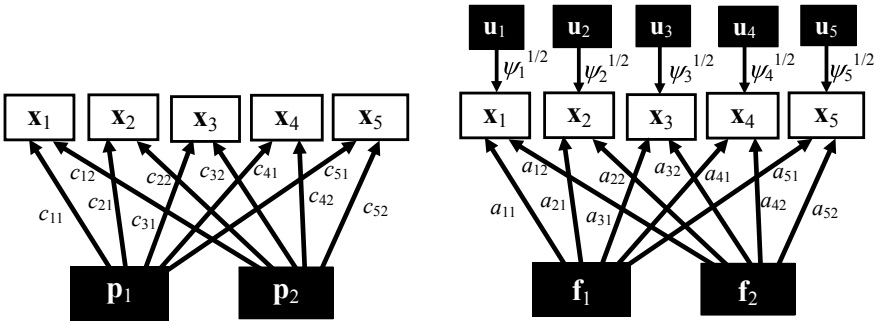
19.3 Solutions and Decomposition of the Sum of Squares

As described in Chap. 5, PCA can be formulated as minimizing the squared norm of the error matrix $\mathbf{E}_{\mathrm{PC}}$ in (19.2), i.e.,

$$f_{\mathrm{PC}}(\mathbf{P}, \mathbf{C}) = \|\mathbf{E}_{\mathrm{PC}}\|^2 = \|\mathbf{X} - \mathbf{PC}'\|^2 \tag{19.4}$$

Table 19.4 Solutions for the correlations in Table 19.2

	PCA				MDFA					LVFA			
	Loadings			Res	Loadings			UV	Res	Loadings			UV
Extraversion	**0.28**	**–0.43**	**0.50**	0.48	0.25	–0.34	0.46	0.60	0.01	0.24	−0.33	0.47	0.61
Activity	**0.44**	**–0.31**	0.64	0.30	0.43	–0.26	0.64	0.33	0.01	0.42	−0.25	0.65	0.34
Empathy	**0.75**	–0.04	**–0.04**	0.43	0.61	–0.04	0.01	0.61	0.02	0.60	−0.02	0.02	0.64
Novelty	0.05	**–0.12**	**0.72**	0.46	0.05	–0.07	0.61	0.61	0.01	0.04	−0.05	0.62	0.62
Durability	**0.75**	–0.07	0.02	0.43	0.65	–0.07	0.05	0.55	0.02	0.66	−0.08	0.05	0.55
Regularity	**0.78**	**0.08**	0.18	0.35	0.71	0.06	0.18	0.45	0.02	0.71	0.05	0.17	0.47
Self-revelation	**0.0**5	0.13	**0.75**	0.42	0.04	0.14	0.63	0.58	0.01	0.03	0.16	0.63	0.58
Aggressiveness	**–0.15**	**0.40**	**0.67**	0.36	–0.14	0.38	0.58	0.50	0.01	−0.13	0.37	0.55	0.54
Uncooperativeness	**–0.28**	**0.56**	**0.26**	0.54	–0.24	0.46	0.18	0.67	0.02	−0.23	0.45	0.17	0.72
Inferiority feeling	–0.16	**0.72**	–0.28	0.38	–0.18	0.63	-0.29	0.49	0.01	−0.18	0.62	−0.30	0.49
Nervousness	**0.32**	**0.76**	**0.04**	0.31	0.27	0.71	0.02	0.41	0.01	0.26	0.72	0.02	0.41
Depression	–0.06	**0.85**	**0.07**	0.27	–0.09	0.83	0.03	0.30	0.01	−0.10	0.83	0.02	0.30
Sum of squares		7.24		1.94		5.78		3.25	0.00		5.73		3.45



(a) Principal Component Analysis (b) Factor Analysis

Fig. 19.1 Graphical representation of PCA and FA with $p = 5$ and $m = 2$

over $\mathbf{P}$ and $\mathbf{C}$ subject to

$$\frac{1}{n}\mathbf{P}'\mathbf{P} = \mathbf{I}_m. \tag{19.5}$$

Let $\widehat{\mathbf{P}}$ and $\widehat{\mathbf{C}}$ denote the solutions of $\mathbf{P}$ and $\mathbf{C}$, respectively, with $\widehat{\mathbf{E}}_{\mathrm{PC}} = \mathbf{X} - \widehat{\mathbf{P}}\widehat{\mathbf{C}}'$ containing the resulting values of errors, i.e., residuals.

As described in Chap. 18, FA can be formulated as minimizing the squared norm of the matrix $\mathbf{E}_{\mathrm{FA}}$ in (19.1), i.e.,

$$f_{\mathrm{FA}}(\mathbf{F}, \mathbf{A}, \mathbf{U}, \boldsymbol{\Psi}) = \|\mathbf{E}_{\mathrm{FA}}\|^2 = \left\|\mathbf{X} - \left(\mathbf{F}\mathbf{A}' + \mathbf{U}\boldsymbol{\Psi}^{1/2}\right)\right\|^2 \tag{19.6}$$

over $\mathbf{F}$, $\mathbf{A}$, $\mathbf{U}$, and $\boldsymbol{\Psi}$ under the constraints in (18.4):

$$\frac{1}{n}\mathbf{F}'\mathbf{F} = \mathbf{I}_m, \quad \frac{1}{n}\mathbf{U}'\mathbf{U} = \mathbf{I}_p, \quad \text{and} \quad \mathbf{F}'\mathbf{U} =_m \mathbf{O}_p. \tag{19.7}$$

Here, the other constraint (18.3) need not be considered, since it is satisfied by the solution minimizing (19.6) under (19.7) when $\mathbf{X}$ is centered, as explained in Sect. 18.5. Let the solutions of $\mathbf{F}$, $\mathbf{A}$, $\mathbf{U}$, and $\boldsymbol{\Psi}$ be $\widehat{\mathbf{F}}$, $\widehat{\mathbf{A}}$, $\widehat{\mathbf{U}}$ and $\widehat{\boldsymbol{\Psi}}$, respectively, with $\widehat{\mathbf{E}}_{\mathrm{FA}} = \mathbf{X} - \left(\widehat{\mathbf{F}}\widehat{\mathbf{A}}' + \widehat{\mathbf{U}}\widehat{\boldsymbol{\Psi}}^{1/2}\right)$ containing residuals.

The following *decompositions* play a key role for comparing the PCA and FA solutions:

Note 19.1. Decompositions of Sum of Squares in PCA and FA

The sum of squares for the centered data matrix, i.e., $\|\mathbf{X}\|^2$, can be decomposed into sums of squares in the PCA and FA solutions:

PCA:

$$\|\mathbf{X}\|^2 = \left\|\widehat{\mathbf{P}}\widehat{\mathbf{C}}'\right\|^2 + \left\|\widehat{\mathbf{E}}_{\mathrm{PC}}\right\|^2 = n\left\|\widehat{\mathbf{C}}\right\|^2 + \left\|\widehat{\mathbf{E}}_{\mathrm{PC}}\right\|^2, \tag{19.8}$$

$$\mathrm{FA}: \|\mathbf{X}\|^2 = \left\|\widehat{\mathbf{F}}\widehat{\mathbf{A}}'\right\|^2 + \left\|\widehat{\mathbf{U}}\widehat{\mathbf{\Psi}}^{1/2}\right\|^2 + \left\|\widehat{\mathbf{E}}_{\mathrm{FA}}\right\|^2 = n\left\|\widehat{\mathbf{A}}\right\|^2 + n\mathrm{tr}\widehat{\mathbf{\Psi}} + \left\|\widehat{\mathbf{E}}_{\mathrm{FA}}\right\|^2. \tag{19.9}$$

These equations are derived as in the next paragraphs.

The decomposition (19.8) for PCA follows from Note 5.3: through the singular value decomposition $\mathbf{X} = \mathbf{K}\mathbf{\Lambda}\mathbf{L}'$ in Note 5.1, the PCA solution $\widehat{\mathbf{P}}\widehat{\mathbf{C}}'$ minimizing (19.4) is given by $\mathbf{K}_m\mathbf{\Lambda}_m\mathbf{L}_m'$ as shown in Notes 5.2 and 5.3, with $\widehat{\mathbf{E}}_{\mathrm{PC}} = \mathbf{X} - \widehat{\mathbf{P}}\widehat{\mathbf{C}}' = \mathbf{K}_{[m]}\mathbf{\Lambda}_{[m]}\mathbf{L}_{[m]'}$. The orthogonality $\mathbf{K}_m'\mathbf{K}_{[m]} = {}_m\mathbf{O}_{p-m}$ leads to $\left(\widehat{\mathbf{P}}\widehat{\mathbf{C}}'\right)'\widehat{\mathbf{E}}_{\mathrm{PC}} = {}_p\mathbf{O}_p$. This property and (19.2) allow us to find the first identity in (19.8). Its last identity follows from (19.5) which implies $\left\|\widehat{\mathbf{P}}\widehat{\mathbf{C}}'\right\|^2 = n\mathrm{tr}\widehat{\mathbf{C}}'\widehat{\mathbf{C}} = n\left\|\widehat{\mathbf{C}}\right\|^2$.

The decomposition (19.9) for FA follows from (18.27) and (19.7): using $\mathbf{B} = [\mathbf{A}, \mathbf{\Psi}^{1/2}]$ and $\mathbf{V} = n^{-1}\mathbf{X}'\mathbf{X}$, (18.27) can be rewritten as $\left\|\widehat{\mathbf{E}}_{\mathrm{FA}}\right\|^2 = \|\mathbf{X}\|^2 - \left(n\left\|\widehat{\mathbf{A}}\right\|^2 + n\mathrm{tr}\widehat{\mathbf{\Psi}}\right)$, which shows the equality of the left and right sides in (19.9). The last identity in (19.9) is derived from that (19.7) implies $\left\|\widehat{\mathbf{F}}\widehat{\mathbf{A}}'\right\|^2 = n\left\|\widehat{\mathbf{A}}\right\|^2$ and $\left\|\widehat{\mathbf{U}}\widehat{\mathbf{\Psi}}^{1/2}\right\|^2 = n\mathrm{tr}\widehat{\mathbf{\Psi}}$.

In *decompositions* (19.8) and (19.9), $\left\|\widehat{\mathbf{P}}\widehat{\mathbf{C}}'\right\|^2 = n\left\|\widehat{\mathbf{C}}\right\|^2$ and $\left\|\widehat{\mathbf{F}}\widehat{\mathbf{A}}'\right\|^2 = n\left\|\widehat{\mathbf{A}}\right\|^2$ stand for the *sizes of the common part* in the PCA solution and the FA counterpart, respectively. On the other hand, $\left\|\widehat{\mathbf{U}}\widehat{\mathbf{\Psi}}^{1/2}\right\|^2 = n\mathrm{tr}\widehat{\mathbf{\Psi}}$ in (19.9) stands for the *size of the unique part*, which is specific to FA. The remaining $\left\|\widehat{\mathbf{E}}_{\mathrm{PC}}\right\|^2$ and $\left\|\widehat{\mathbf{E}}_{\mathrm{FA}}\right\|^2$ in (19.8) and (19.9) stand for the *sizes of the residuals* that remain unaccounted for by components/factors. The decomposition of $\|\mathbf{X}\|^2$ into the above sums of squares can be seen in Fig. 19.2. Here, the *areas* of the sums of squares *differ* between PCA and FA. These differences are explained by the inequalities presented in the next three sections. The inequalities also suggest that the observations [O1]–[O3] in Sect. 19.1 are commonly found.

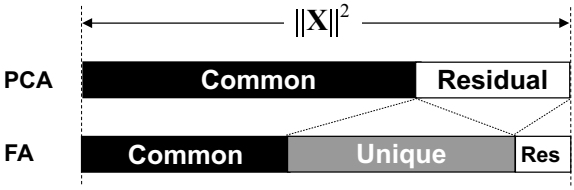


Fig. 19.2 Relative sizes of common parts, unique parts, and residuals in PCA and FA solutions

19.4 Larger Common Part of Principal Component Analysis

Let us consider the left part in Fig. 19.2. Here, the *size of the common part* $n\left\|\widehat{\mathbf{C}}\right\|^2 = \left\|\widehat{\mathbf{P}}\widehat{\mathbf{C}}'\right\|^2$ for PCA is depicted so as to be greater than the FA counterpart $n\left\|\widehat{\mathbf{A}}\right\|^2 = \left\|\widehat{\mathbf{F}}\widehat{\mathbf{A}}'\right\|^2$. This follows from the next theorem:

Note 19.2. Larger Sum of Squared PCA Loadings (Adachi and Trendafilov, 2019, Theorem 2)
For a given $\mathbf{X}$, the *sum of squared PCA loadings* is *always larger* than or equal to the *FA* counterpart under constraints (19.5) and (19.7):

$$\left\|\widehat{\mathbf{C}}\right\|^2 = \left\|\widehat{\mathbf{C}}\mathbf{T}_P\right\|^2 \geq \left\|\widehat{\mathbf{A}}\right\|^2 = \left\|\widehat{\mathbf{A}}\mathbf{T}_F\right\|^2. \tag{19.10}$$

Here, $\mathbf{T}_P$ and $\mathbf{T}_F$ are arbitrary $m \times m$ orthonormal matrices with $\mathbf{T}_P'\mathbf{T}_P = \mathbf{I}_m$ and $\mathbf{T}_F'\mathbf{T}_F = \mathbf{I}_m$. This implies that the common part in PCA is always larger than or equal to the FA counterpart, even *after orthogonal rotation*. The proof is provided in Sect. 19.9.

The inequality (19.10) provides the guideline:

Choose PCA rather than FA for the purpose of extracting a larger common part from data. (19.11)

We can rewrite (19.10) as

$$\sum_{j=1}^{p}\sum_{k=1}^{m}\left|c_{jk}\right|^2 \geq \sum_{j=1}^{p}\sum_{k=1}^{m}\left|a_{jk}\right|^2. \tag{19.12}$$

Here, c_{jk} and a_{jk} are the (j, k) elements of $\widehat{\mathbf{C}}\mathbf{T}_P$ and $\widehat{\mathbf{A}}\mathbf{T}_F$, respectively, with $\mathbf{I}_m \in \mathbf{T}_P$ and $\mathbf{I}_m \in \mathbf{T}_F$. Inequality (19.12) suggests the observation [O1] in Sect. 19.1, i.e., that the absolute loading $|c_{jk}|$ for PCA tends to be greater than the FA counterpart $|a_{jk}|$ before and after orthogonal rotation, although $|c_{jk}| \leq |a_{jk}|$ can also be observed.

19.5 Better Fit of Factor Analysis

Let us consider the right part in Fig. 19.2. Here, the *size of residuals* $\left\|\widehat{\mathbf{E}}_{\mathrm{PC}}\right\|^2$ for *PCA* is depicted so as to be *greater* than the *FA* counterpart $\left\|\widehat{\mathbf{E}}_{\mathrm{FA}}\right\|^2$. This follows from the next theorem:

Note 19.3. Better fit of FA (Adachi and Trendafilov, 2019, Theorem 1) For a given $\mathbf{X}$, the *FA* solution always shows a *better* or equivalent *fit* compared to the *PCA* solution. In other words, the *sum of squared residuals* in FA does not exceed the PCA counterpart:

$$\left\|\widehat{\mathbf{E}}_{\mathrm{FA}}\right\|^2 \leq \left\|\widehat{\mathbf{E}}_{\mathrm{PC}}\right\|^2 \tag{19.13}$$

This proof is given in Sect. 19.9.

Inequality (19.13) provides the guideline:

Choose FA for the purpose of accounting for data better. (19.14)

The index "Res" for variable j in Tables 19.3 and 19.4 are defined as

$$\mathrm{RES}_j^{\mathrm{FA}} = \frac{1}{n}\left\|\hat{\mathbf{e}}_j^{[\mathrm{FA}]}\right\|^2 \text{ for FA and } \mathrm{RES}_j^{\mathrm{PC}} = \frac{1}{n}\left\|\hat{\mathbf{e}}_j^{[\mathrm{PC}]}\right\|^2 \text{for PCA}, \tag{19.15}$$

with $\hat{\mathbf{e}}_j^{[\mathrm{PC}]}$ and $\hat{\mathbf{e}}_j^{[\mathrm{FA}]}$ being the jth columns of $\widehat{\mathbf{E}}_{\mathrm{PC}}$ and $\widehat{\mathbf{E}}_{\mathrm{FA}}$, respectively. These two matrices are centered as shown in Chaps. 5 and 18, which implies that the two indices in (19.15) stand for residual variances. From (19.15), we can find that the sum of residual variances over variables equal to the average of squared residuals:

$$\sum_{j=1}^{p} \mathrm{RES}_j^{\mathrm{FA}} = \frac{1}{n}\left\|\widehat{\mathbf{E}}_{\mathrm{FA}}\right\|^2 \quad \text{and} \quad \sum_{j=1}^{p} \mathrm{RES}_j^{\mathrm{PC}} = \frac{1}{n}\left\|\widehat{\mathbf{E}}_{\mathrm{PC}}\right\|^2. \tag{19.16}$$

By comparing these equalities with (19.13), we can find that it is rewritten as

$$\sum_{j=1}^{q} \mathrm{RES}_j^{\mathrm{FA}} \leq \sum_{j=1}^{q} \mathrm{RES}_j^{\mathrm{PC}}. \tag{19.17}$$

This inequality suggests that the residual variance $\mathrm{RES}_j^{\mathrm{PC}}$ for each variable in PCA tends to be greater than the FA counterpart $\mathrm{RES}_j^{\mathrm{FA}}$, as written in [O2] in Sect. 19.1, though exceptions can exist.

We should note that (19.8) and (19.9) are rewritten as $\left\|\widehat{\mathbf{E}}_{\mathrm{PC}}\right\|^2 = \|\mathbf{X}\|^2 - n\left\|\widehat{\mathbf{C}}\right\|^2$ and $\left\|\widehat{\mathbf{E}}_{\mathrm{FA}}\right\|^2 = \|\mathbf{X}\|^2 - n\left\|\widehat{\mathbf{A}}\right\|^2 - n\mathrm{tr}\widehat{\mathbf{\Psi}}$, respectively. Using these equations in (19.13), we have $\|\mathbf{X}\|^2 - n\left\|\widehat{\mathbf{A}}\right\|^2 - n\mathrm{tr}\widehat{\mathbf{\Psi}} \leq \|\mathbf{X}\|^2 - n\left\|\widehat{\mathbf{C}}\right\|^2$. Detracting $\|\mathbf{X}\|^2$ from both sides of this inequality leads to $-n\left\|\widehat{\mathbf{A}}\right\|^2 - n\mathrm{tr}\widehat{\mathbf{\Psi}} \leq -n\left\|\widehat{\mathbf{C}}\right\|^2$, which implies the following:

Note 19.4. Upper Limit of the Sum of Squared PCA Loadings (Adachi and Trendafilov, 2019, Theorem 3)
For a given $\mathbf{X}$, the sum of the squared PCA loadings cannot exceed the sum of the squared loadings and unique variances in the FA solution:

$$\left\|\widehat{\mathbf{C}}\right\|^2 = \left\|\widehat{\mathbf{C}}\mathbf{T}_P\right\|^2 \leq \left\|\widehat{\mathbf{A}}\right\|^2 + \mathrm{tr}\widehat{\mathbf{\Psi}} = \left\|\widehat{\mathbf{A}}\mathbf{T}_F\right\|^2 + \mathrm{tr}\widehat{\mathbf{\Psi}}, \tag{19.18}$$

with $\mathbf{T}_P$ and $\mathbf{T}_F$ arbitrary orthonormal matrices.

This inequality shows the *upper* limit of the PCA common part, in contrast to (19.10) which shows its *lower* limit.

19.6 Largeness of Unique Variances in Factor Analysis

Let us consider the right and middle parts in Fig. 19.2. Here, the area of the *FA unique part* is *greater* than that of the *PCA residual part*. This is suggested by the following theorem:

Note 19.5. Lower Limit of the Sum of Unique Variances. (Adachi and Trendafilov, 2019, Theorem 4)
For a given $\mathbf{X}$, the *sum of the unique variances* in FA is larger than or equal to the average of squared residuals for PCA minus the average for FA:

$$\mathrm{tr}\,\widehat{\boldsymbol{\Psi}} \geq \frac{1}{n}\left\|\widehat{\mathbf{E}}_{\mathrm{PC}}\right\|^2 - \frac{1}{n}\left\|\widehat{\mathbf{E}}_{\mathrm{FA}}\right\|^2. \tag{19.19}$$

This proof is given in Sect. 19.9.

By comparing (19.16) with (19.19), we find that the latter is rewritten as

$$\sum_{j=1}^{p}\psi_j \geq \sum_{j=1}^{p}\left(\mathrm{RES}_j^{\mathrm{PC}} - \mathrm{RES}_j^{\mathrm{FA}}\right). \tag{19.20}$$

This suggests that ψ_j tends to be greater than $\mathrm{RES}_j^{\mathrm{PC}} - \mathrm{RES}_j^{\mathrm{FA}}$ and further that if $\mathrm{RES}_j^{\mathrm{FA}}$ is small enough, ψ_j^2 tends to be greater than $\mathrm{RES}_j^{\mathrm{PC}}$, i.e., the *FA unique variance* for each variable tends to be *larger* than the *PCA residual variance* for that variable, as written in [O3] in Sect. 19.1.

19.7 Inequalities for Latent Variable Factor Analysis

The mathematical results in Notes 19.2–19.5 are based on *MDFA*, rather than LVFA. However, Adachi and Trendafilov (2019) found that *LVFA almost always provides the solutions shown by the inequalities* in Notes 19.2 and 12.4. This is suggested by a broad equivalence of the MDFA and LVFA solutions, which is shown in Tables 19.3 and 19.4, and other data sets (Adachi and Trendafilov, 2019).

The inequalities in Notes 12.3 and 12.5 do not make sense in LVFA whose model does not include $\mathbf{E}_{\mathrm{FA}}$. However, the relationship in (19.19) with $n^{-1}\left\|\widehat{\mathbf{E}}_{\mathrm{FA}}\right\|^2$ removed, i.e., $\mathrm{tr}\,\widehat{\boldsymbol{\Psi}}$ being greater than $n^{-1}\left\|\widehat{\mathbf{E}}_{\mathrm{PC}}\right\|^2$, is almost always found in the LVFA solutions (Adachi and Trendafilov, 2019).

19.8 Inequalities After Nonsingular Transformation

The FA (19.1) and PCA (19.2) models can be rewritten as

$$\mathbf{X} = \mathbf{F}\mathbf{A}' + \mathbf{U}\boldsymbol{\Psi}^{1/2} + \mathbf{E}_{\mathrm{FA}} = \mathbf{F}\mathbf{N}_F\mathbf{N}_F^{-1}\mathbf{A}' + \mathbf{U}\boldsymbol{\Psi}^{1/2} + \mathbf{E}_{\mathrm{FA}}. \tag{19.21}$$

$$\mathbf{X} = \mathbf{P}\mathbf{C}' + \mathbf{E}_{\mathrm{PC}} = \mathbf{P}\mathbf{N}_P\mathbf{N}_P^{-1}\mathbf{C}' + \mathbf{E}_{\mathrm{PC}} \tag{19.22}$$

with $\mathbf{N}_P$ and $\mathbf{N}_F$ arbitrary nonsingular $m \times m$ matrices. Without the conditions $n^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_m$ and (19.5), (19.21), and (19.22) show that we can regard $\mathbf{A}\mathbf{N}_F{}'^{-1}$ and $\mathbf{C}\mathbf{N}_P{}'^{-1}$ as loading matrices, $\mathbf{F}\mathbf{N}_F$ as a common factor score matrix, and $\mathbf{P}\mathbf{N}_P$ as a PC score matrix. It implies that the *oblique* rotation treated in Chap. 13 can be performed for FA and PCA solutions.

Even after the *nonsingular* transformations in (19.21) and (19.22) are performed, the inequalities in (19.13) and (19.19) hold true, since the matrices $\mathbf{E}_{\mathrm{FA}}$, $\mathbf{E}_{\mathrm{PC}}$, and $\boldsymbol{\Psi}$ appearing in those inequalities are irrelevant to $\mathbf{N}_P$ and $\mathbf{N}_F$ as found in (19.21) and (19.22).

However, (19.10) and (19.18) are involved in the loading matrices $\mathbf{A}$ and $\mathbf{C}$, which are transformed by $\mathbf{N}_F^{-1}$ and $\mathbf{N}_P^{-1}$ in (19.21) and (19.22). Since the orthonormal matrices $\mathbf{T}_P$ and $\mathbf{T}_F$ in (19.10) and (19.18) cannot be replaced by $\mathbf{N}_P$ and $\mathbf{N}_F$, neither $\left\|\widehat{\mathbf{C}}\mathbf{N}_P\right\|^2 \le \left\|\widehat{\mathbf{A}}\mathbf{N}_P\right\|^2$ nor $\left\|\widehat{\mathbf{C}}\mathbf{N}_P\right\|^2 \le \left\|\widehat{\mathbf{A}}\mathbf{N}_P\right\|^2 + \mathrm{tr}\widehat{\boldsymbol{\Psi}}$ hold in general. The following inequalities described next, however, do hold.

Note 19.6. Larger PCA Common Part and its Upper Limit (Adachi and Trendafilov, 2019, Theorems 2 and 3)
For a given $\mathbf{X}$, we have

$$\left\|\widehat{\mathbf{P}}\widehat{\mathbf{C}}'\right\|^2 = \left\|\widehat{\mathbf{P}}\mathbf{N}_p\mathbf{N}_p^{-1}\widehat{\mathbf{C}}'\right\|^2 \ge \left\|\widehat{\mathbf{F}}\widehat{\mathbf{A}}'\right\|^2 = \left\|\widehat{\mathbf{F}}\mathbf{N}_F\mathbf{N}_F^{-1}\widehat{\mathbf{A}}'\right\|^2. \tag{19.23}$$

with $\mathbf{N}_P$ and $\mathbf{N}_F$ arbitrary $m \times m$ nonsingular matrices. The upper limit of $\left\|\widehat{\mathbf{P}}\widehat{\mathbf{C}}'\right\|^2$ is given by

$$\left\|\widehat{\mathbf{P}}\widehat{\mathbf{C}}'\right\|^2 = \left\|\widehat{\mathbf{P}}\mathbf{N}_P\mathbf{N}_p^{-1}\widehat{\mathbf{C}}'\right\|^2 \le \left\|\widehat{\mathbf{F}}\widehat{\mathbf{A}}'\right\|^2 + n\mathrm{tr}\widehat{\boldsymbol{\Psi}} = \left\|\widehat{\mathbf{F}}\mathbf{N}_F\mathbf{N}_F^{-1}\widehat{\mathbf{A}}'\right\|^2 + n\mathrm{tr}\widehat{\boldsymbol{\Psi}} \tag{19.24}$$

These result can be proved as follows: (19.5) and (19.7) imply $\left\|\widehat{\mathbf{C}}\right\|^2 = n^{-1}\left\|\widehat{\mathbf{P}}\widehat{\mathbf{C}}'\right\|^2$ and $\left\|\widehat{\mathbf{A}}\right\|^2 = n^{-1}\left\|\widehat{\mathbf{F}}\widehat{\mathbf{A}}'\right\|^2$. We can use these equalities in (19.10) and (19.18) to obtain (19.23) and (19.24).

Inequality (19.23) shows that the suggestion in (19.11) is valid even after the *nonsingular* transformations in (19.21) and (19.22), if the *common parts* refer to $\left\|\widehat{\mathbf{F}}\widehat{\mathbf{A}}'\right\|^2$ and $\left\|\widehat{\mathbf{P}}\widehat{\mathbf{C}}'\right\|^2$ rather than $\left\|\widehat{\mathbf{A}}\right\|^2$ and $\left\|\widehat{\mathbf{C}}\right\|^2$.

19.9 Proofs for Inequalities

The inequalities in Notes 19.2, 19.3, and 19.5 are proved in this section.

We can prove (19.10) in Note 19.2 as follows: The PCA loss function (19.4) is expanded as $f_{\mathrm{PC}}(\mathbf{P},\mathbf{C}) = ||\mathbf{X}||^2 - 2\mathrm{tr}\,\mathbf{X}'\mathbf{P}\mathbf{C}' + \|\mathbf{P}\mathbf{C}'\|^2$. By substituting $\widehat{\mathbf{P}}\widehat{\mathbf{C}}'$ for $\mathbf{P}\mathbf{C}'$ in $f_{\mathrm{PC}}(\mathbf{P}, \mathbf{C})$ and using (19.5) and $\mathbf{V} = n^{-1}\mathbf{X}'\mathbf{X}$, we have

$$f_{\mathrm{PC}}(\widehat{\mathbf{P}}, \widehat{\mathbf{C}}) = n\mathrm{tr}\mathbf{V} - 2\mathrm{tr}\mathbf{X}'\widehat{\mathbf{P}}\widehat{\mathbf{C}}' + n\mathrm{tr}\widehat{\mathbf{C}}\widehat{\mathbf{C}}' = n\left(\mathrm{tr}\mathbf{V} - \mathrm{tr}\widehat{\mathbf{C}}\widehat{\mathbf{C}}'\right). \tag{19.25}$$

Here, we have also used the fact that (5.30) can be rewritten as $\widehat{\mathbf{C}} = n^{-1}\mathbf{X}'\widehat{\mathbf{P}}$ using the notation in this chapter. Now, let us substitute the *FA solution* $\widehat{\mathbf{F}}\widehat{\mathbf{A}}'$ into $\mathbf{P}\mathbf{C}'$ in the *PCA function* (19.4). Then, we have $f_{\mathrm{PC}}(\widehat{\mathbf{F}}, \widehat{\mathbf{A}}) = \left\|\mathbf{X} - \widehat{\mathbf{F}}\widehat{\mathbf{A}}'\right\|^2 = n\left(\mathrm{tr}\mathbf{V} - 2\mathrm{tr}\mathbf{S}_{\mathrm{XF}}\widehat{\mathbf{A}}' + \mathrm{tr}\widehat{\mathbf{A}}\widehat{\mathbf{A}}'\right) = n\left(\mathrm{tr}\mathbf{V} - \mathrm{tr}\widehat{\mathbf{A}}\widehat{\mathbf{A}}'\right)$ with $\mathbf{S}_{\mathrm{XF}} = n^{-1}\mathbf{X}'\widehat{\mathbf{F}}$, using (19.7) and (18.16). Clearly, $f_{\mathrm{PC}}(\widehat{\mathbf{F}}, \widehat{\mathbf{A}}) = n\left(\mathrm{tr}\mathbf{V} - \mathrm{tr}\widehat{\mathbf{A}}\widehat{\mathbf{A}}'\right)$ cannot be lower than (19.25), since the *PCA solution* is the *best reduced rank approximation* as shown in Note 5.3 and Theorem A.4.5 with (A.4.17). Thus, we have

$$\mathrm{tr}\mathbf{V} - \mathrm{tr}\widehat{\mathbf{C}}\widehat{\mathbf{C}}' \le \mathrm{tr}\mathbf{V} - \mathrm{tr}\widehat{\mathbf{A}}\widehat{\mathbf{A}}'. \tag{19.26}$$

This result and the orthonormality of $\mathbf{T}_P$ and $\mathbf{T}_F$ give (19.10).

The inequality (19.13) in Note 19.3 can be proved as follows: if $\boldsymbol{\Psi}$ is restricted to ${}_p\mathbf{O}_p$, the FA loss function (19.6) with (19.7) is equivalent to the PCA function (19.4) with (19.5): $\|\mathbf{X} - \mathbf{P}\mathbf{C}'\|^2 = \left\|\mathbf{X} - \mathbf{F}\mathbf{A}' - \mathbf{U}_p\mathbf{O}_p\right\|^2$, and its minimization is independent of the two constraints in (19.7) except $n^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_m$ which is equivalent to (19.5). Thus, we have $f_{\mathrm{PC}}(\widehat{\mathbf{P}}, \widehat{\mathbf{C}}) = f_{\mathrm{FA}}\left(\widehat{\mathbf{P}}, \widehat{\mathbf{C}}, \mathbf{U}, {}_p\mathbf{O}_p\right)$, where the right term is the *FA function* (19.6) with $\mathbf{F} = \widehat{\mathbf{P}}, \mathbf{A} = \widehat{\mathbf{C}}$, and $\boldsymbol{\Psi} = {}_p\mathbf{O}_p$. Obviously, $f_{\mathrm{FA}}\left(\widehat{\mathbf{P}}, \widehat{\mathbf{C}}, \mathbf{U}, {}_p\mathbf{O}_p\right)$ cannot be lower than $f_{\mathrm{FA}}(\widehat{\mathbf{F}}, \widehat{\mathbf{A}}, \widehat{\mathbf{U}}, \widehat{\boldsymbol{\Psi}})$, i.e., the minimum of the FA loss function, which leads to

$$f_{\mathrm{PC}}(\widehat{\mathbf{P}}, \widehat{\mathbf{C}}) = f_{\mathrm{FA}}\left(\widehat{\mathbf{P}}, \widehat{\mathbf{C}}, \mathbf{U}, {}_p\mathbf{O}_p\right) \ge f_{\mathrm{FA}}(\widehat{\mathbf{F}}, \widehat{\mathbf{A}}, \widehat{\mathbf{U}}, \widehat{\boldsymbol{\Psi}}), \tag{19.27}$$

i.e., $f_{\mathrm{PC}}(\widehat{\mathbf{P}}, \widehat{\mathbf{C}}) \ge f_{\mathrm{FA}}(\widehat{\mathbf{F}}, \widehat{\mathbf{A}}, \widehat{\mathbf{U}}, \widehat{\boldsymbol{\Psi}})$. This result, (19.4), and (19.6) show (19.13).

We can prove (19.19) in Note 19.5 as follows: (18.27) and $\mathbf{B} = [\mathbf{A}, \mathbf{\Psi}^{1/2}]$ imply $n^{-1}\left\|\widehat{\mathbf{E}}_{\mathrm{FA}}\right\|^2 = \mathrm{tr}(\mathbf{V} - \widehat{\mathbf{A}}\widehat{\mathbf{A}}' - \widehat{\mathbf{\Psi}})$, i.e., $\mathrm{tr}\mathbf{V} - \mathrm{tr}\widehat{\mathbf{A}}\widehat{\mathbf{A}}' = \mathrm{tr}\widehat{\mathbf{\Psi}} + n^{-1}\left\|\widehat{\mathbf{E}}_{\mathrm{FA}}\right\|^2$. We can also rewrite (19.25) as $\mathrm{tr}\mathbf{V} - \mathrm{tr}\widehat{\mathbf{C}}\widehat{\mathbf{C}}' = n^{-1}\left\|\widehat{\mathbf{E}}_{\mathrm{PC}}\right\|^2$. Using these equalities in (19.26), we have $n^{-1}\left\|\widehat{\mathbf{E}}_{\mathrm{PC}}\right\|^2 \leq \left\|\widehat{\mathbf{\Psi}}\right\| + n^{-1}\left\|\widehat{\mathbf{E}}_{\mathrm{FA}}\right\|^2$, which can be rewritten as (19.19).

19.10 Bibliographical Notes

Major parts of this chapter are based on the results of Adachi and Trendafilov (2019). Among the literature published prior to this, it is difficult to find any which clearly indicate the differences between PCA and FA solutions. However, useful comparisons between PCA and latent variable FA are found in Ogasawara (2000), Schneeweiss and Mathes (1995), and the series of the papers in Volume 25, Issue 1 of the journal *Multivariate Behavioral Research* which starts with Velicer and Jackson (1990).

Exercise

19.1. Summarize the respective cases where PCA and FA should be used.
19.2. Discuss how the PCA solution can be obtained explicitly, while FA cannot.
19.3. Discuss how singular value decomposition is used for obtaining PCA and MDFA solutions.
19.4. Argue that (19.27) shows PCA being equivalent to the constrained FA with unique variances restricted to zeros.
19.5. Show that the PCA loss function (19.4) can be rewritten as

$$f_{\mathrm{PC}}(\mathbf{P}|\mathbf{C}) = \|\mathbf{P} - \mathbf{X}\mathbf{C}\|^2 + \|\mathbf{X}\|^2 + n\|\mathbf{C}\|^2 - \|\mathbf{X}\mathbf{C}\|^2 - nm \tag{19.28}$$

subject to (19.5).
19.6. Comparing (19.28) with (18.32), discuss the following statement: in contrast to the fact that obtaining $\mathbf{P}$ in PCA can be regarded as a *lower* rank approximation problem, obtaining $\mathbf{Z} = [\mathbf{F}, \mathbf{U}]$ in FA can be viewed as a *higher* rank approximation problem.
19.7. *Factor indeterminacy* refers to the property that the *optimal factor score matrix cannot be uniquely determined*. Discuss the differences between this indeterminacy and the *rotational indeterminacy*.
19.8. Discuss that the optimal PC score matrix in PCA can be uniquely determined as a function of $\mathbf{X}$, while the optimal factor score matrix in MDFA cannot be uniquely determined.

Chapter 20
Three-Way Principal Component Analysis

In Chap. 5, principal component analysis (PCA) was introduced as the reduced rank approximation of a data matrix. This matrix should be noted to be a two-way array of rows × columns. We often encounter *three-way data* arrays, however, an example of which is a set of scores of examinees for multiple tests administered on different occasions. These scores form a three-way array of examinees × tests × occasions. Modified PCA procedures specified for similar three-way data are known as *three-way PCA* (*3WPCA*). Popular 3WPCA procedures are introduced in this chapter.

20.1 Tucker3 and Parafac Models

Let a *three-way data array* be denoted as

$$\dddot{\mathbf{X}} = \left\{x_{ijk}; i = 1, \ldots, I; j = 1, \ldots, J; k = 1, \ldots, K\right\}. \tag{20.1}$$

This can be depicted as the left cube in Fig. 20.1. For example, x_{ijk} could stand for the brightness of the (i, j) element or pixel in an image recorded at time k. Another example is a case where x_{ijk} is the score of examinee i for test j on occasion k.

One popular *three-way principal component analysis* (*3WPCA*) procedure is *Tucker3*, which is also called *Tucker decomposition*. Those names follow Tucker (1966) who proposed the procedure. Tucker3 is modeled as

$$x_{ijk} = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} a_{ip}b_{jq}c_{kr}g_{pqr} + e_{ijk} \tag{20.2}$$

K. Adachi, *Matrix-Based Introduction to Multivariate Data Analysis*,
https://doi.org/10.1007/978-981-15-4103-2_20

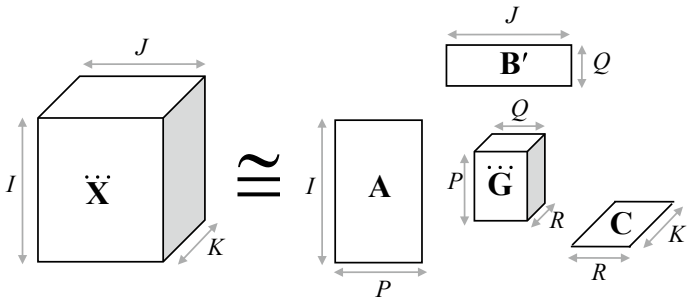


Fig. 20.1 Pictorial representation of the Tucker3 model

with e_{ijk} an error, $I \geq P$, $J \geq Q$, and $K \geq R$. Here, we have used the sets of symbols $\{a, b, c\}$, $\{I, J, K\}$, and $\{P, Q, R\}$, whose elements are simply in alphabetical order, for the sake of easily labeling their correspondence to each of the three ways. Let us define three matrices as $\mathbf{A} = \left(a_{ip}\right) = [\mathbf{a}_1, \ldots, \mathbf{a}_P]$ $(I \times P)$, $\mathbf{B} = \left(b_{jq}\right) = [\mathbf{b}_1, \ldots, \mathbf{b}_Q](J \times Q)$, and $\mathbf{C} = (c_{kr}) = [\mathbf{c}_1, \ldots, \mathbf{c}_R]$ $(K \times R)$, with a three-way array

$$\dddot{\mathbf{G}} = \left\{g_{pqr}; p = 1, \ldots, P; q = 1, \ldots, Q; r = 1, \ldots, R\right\}. \tag{20.3}$$

which is called a *core array*. The elements in $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\dddot{\mathbf{G}}$ are the unknown parameters to be estimated in Tucker3.

The implications of the Tucker3 model (20.2) can be viewed in the three-dimensional diagram in Fig. 20.1. This depicts the assumption that data cube $\dddot{\mathbf{X}}(I \times J \times K)$ is underlain by a smaller *core* cube $\dddot{\mathbf{G}}$ $(P \times Q \times R)$. This cube describes the relationships among the P, Q, and R *components* which correspond to the columns of $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$, respectively. These three matrices describe how the I, J, and K entities (surrounding $\dddot{\mathbf{X}}$) *load* the P, Q, and R components.

Another popular 3WPCA procedure is *Parafac*, which was proposed by Harshman (1970) and Carroll and Chang (1970), and whose root is also found in Hitchcock (1927). The name Parafac originates from the abbreviation of parallel factor analysis. This term is misleading as explained later, but we use the name Parafac as it is prevalent. Its model is a constrained variant of (20.2), in which g_{pqr} is restricted so that $g_{pqr} = 1$ for $p = q = r$ and $g_{pqr} = 0$ otherwise, with $P = Q = R$. Hence:

$$x_{ijk} = \sum_{p=1}^{P} a_{ip} b_{jp} c_{kp} + e_{ijk}. \tag{20.4}$$

Here, g_{pqr} in (20.2) disappears: Parafac can also be represented in Fig. 20.1 in which the cube $\dddot{\mathbf{G}}$ is removed, but $\mathbf{A}$ $(I \times P)$, $\mathbf{B}$ $(J \times P)$, and $\mathbf{C}$ $(K \times P)$ remain

with the numbers of their columns constrained identically to be P. The pth columns of $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are associated mutually to give the product $a_{ip}b_{jp}c_{jp}$ and its sum over p approximates x_{ijk}.

In Tucker3 and Parafac, the parameters are estimated with the least squares method: the sum of the squared errors e_{ijk}^2 is minimized over the parameters. The algorithms for the minimization are described later in Sects. 20.5–20.8.

3WPCA used to be called three-way factor analysis. This naming is misleading, as is the name parallel factor analysis, since 3WPCA including Tucker3 and Parafac is clearly different from the factor analysis characterized by unique factors (Chap. 19): these are not included in the 3WPCA models. Furthermore, in the next section, more straightforward evidence is given showing that 3WPCA belongs to the *family of PCA*.

20.2 Hierarchical Relationships Among PCA and 3WPCA

Let $\mathbf{X}_k$, $k = 1, \ldots, K$, denote the $I \times J$ matrices whose (i, j) element is x_{ijk} in (20.1):

$$\mathbf{X}_k = \left(x_{ijk}\right) \quad (k = 1, \ldots, K) \tag{20.5}$$

These are vertically stacked to form the $KI \times J$ block matrix $\widetilde{\mathbf{X}} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_K \end{bmatrix}$.

Table 20.1 presents an example of $\widetilde{\mathbf{X}}$, whose contents are explained in the next section. The purpose of this section is to show that 3WPCA procedures are *constrained versions of the PCA* in Chap. 5 performed for $\widetilde{\mathbf{X}}$:

$$\text{PCA} \succ \text{Tucker2} \succ \text{Tucker3} \succ \text{Parafac}. \tag{20.6}$$

Here, the symbol $\succ$ delineates the order of *constrainedness*, so that the procedure after $\succ$ is a constrained version of the one before it. Tucker2 is introduced later.

The *PCA* model for $\widetilde{\mathbf{X}}$ with the number of components $Q \leq \min(KI, J)$ can be expressed as $\widetilde{\mathbf{X}} = \widetilde{\mathbf{A}}\mathbf{B}' + \widetilde{\mathbf{E}}$, using $\widetilde{\mathbf{A}} = \begin{bmatrix} \mathbf{A}_1 \\ \vdots \\ \mathbf{A}_K \end{bmatrix}$ a PC score matrix of $KI \times Q$, $\widetilde{\mathbf{E}} = \begin{bmatrix} \mathbf{E}_1 \\ \vdots \\ \mathbf{E}_K \end{bmatrix}$ a $KI \times J$ error matrix, and $\mathbf{B}$ a $J \times Q$ loading matrix, where $\mathbf{A}_k$ and $\mathbf{E}_k$ ($k = 1, \ldots, K$) are $I \times Q$ and $I \times J$, respectively. Here, we have used the characters $\mathbf{A}$ and $\mathbf{B}$, so that they correspond to those used for 3WPCA. The above PCA model can be rewritten as

Table 20.1 Three-way data array of 15 concepts × 10 adjectives × 6 occasions (Osgood and Luria, 1954)

Occasion	Concept	Adjective									
		Hot	Worthless	Relaxed	Large	Slow	Clean	Strong	Distasteful	Shallow	Passive
$\mathbf{X}_1$ *Eve White 1*	Love	1.0	−3.0	0.0	3.0	3.0	3.0	3.0	−3.0	−3.0	−3.0
	Child	0.0	−3.0	0.0	−3.0	−1.5	3.0	−3.0	−3.0	−3.0	−3.0
	Doctor	2.0	−3.0	3.0	3.0	−1.0	3.0	3.0	−3.0	−3.0	−3.0
	Me	0.0	0.0	−3.0	−3.0	1.5	1.5	−3.0	−1.0	0.5	0.0
	Job	1.0	−3.0	−2.0	1.5	1.5	3.0	1.5	−3.0	−2.0	−1.5
	Mental sick	0.0	−3.0	−3.0	3.0	3.0	1.5	−3.0	3.0	−3.0	−3.0
	Mother	−2.5	−3.0	−2.5	1.5	−3.0	3.0	3.0	−3.0	−1.5	−3.0
	Peace of mind	1.5	−3.0	3.0	3.0	3.0	3.0	3.0	−3.0	−3.0	3.0
	Fraud	−3.0	3.0	−3.0	−3.0	−3.0	−3.0	−3.0	3.0	3.0	0.0
	Spouse	0.0	−1.0	0.0	1.0	0.0	1.0	0.0	0.0	−1.5	0.5
	Self-control	0.0	−3.0	1.0	2.5	2.0	1.0	3.0	−3.0	−3.0	−3.0
	Hatred	−3.0	3.0	−3.0	−3.0	3.0	−3.0	−3.0	3.0	0.0	−3.0
	Father	2.0	−3.0	−2.5	3.0	0.0	3.0	3.0	−3.0	−3.0	−1.5
	Confusion	−3.0	−1.0	−3.0	3.0	−3.0	−2.5	−3.0	3.0	−3.0	−3.0
	Sex	−0.5	−1.5	0.5	2.0	1.0	1.5	0.0	0.5	−2.5	0.0
$\mathbf{X}_2$ *Eve White 2*	Love	1.5	−3.0	−1.5	2.0	0.5	3.0	1.0	−2.0	−2.0	−1.5
	Child	2.0	−3.0	3.0	3.0	−2.0	3.0	−3.0	−3.0	−2.0	−3.0
	Doctor	1.0	−3.0	3.0	3.0	−2.0	3.0	3.0	−3.0	−2.5	−3.0
	Me	−0.5	2.0	−3.0	−2.0	2.0	2.0	−3.0	1.5	0.0	1.0
	Job	0.0	−3.0	−2.0	2.5	0.0	3.0	2.0	−2.5	−1.5	0.0
	Mental sick	1.0	0.0	−3.0	2.5	2.0	2.5	−1.0	1.0	−3.0	−3.0
	Mother	2.0	−3.0	−0.5	3.0	−2.0	3.0	2.5	−3.0	−2.5	−3.0
	Peace of mind	0.0	−3.0	2.5	3.0	2.0	3.0	3.0	−3.0	−3.0	−0.5

(continued)

Table 20.1 (continued)

Occasion	Concept	Adjective									
		Hot	Worthless	Relaxed	Large	Slow	Clean	Strong	Distasteful	Shallow	Passive
	Fraud	−3.0	3.0	−3.0	−3.0	−2.0	−3.0	−3.0	3.0	−3.0	−2.0
	Spouse	−1.5	−0.5	1.5	1.5	0.0	0.5	−0.5	1.5	−0.5	0.5
	Self-control	1.0	−3.0	−1.0	2.5	2.0	2.5	2.0	−2.5	−2.0	−2.0
	Hatred	−3.0	3.0	−2.5	−3.0	−1.0	−3.0	−3.0	3.0	−3.0	−2.0
	Father	1.0	−3.0	2.0	3.0	−2.0	3.0	3.0	−3.0	−2.0	−3.0
	Confusion	0.0	1.5	−3.0	1.0	−2.0	2.0	−2.5	3.0	−3.0	−2.5
	Sex	−2.0	0.5	−1.5	−1.5	0.0	−1.0	−1.5	1.5	−0.5	1.0
$\mathbf{X}_3$ *Eve Black 1*	Love	0.0	3.0	−3.0	−3.0	3.0	−2.0	−3.0	3.0	3.0	3.0
	Child	−3.0	3.0	−3.0	3.0	0.0	−3.0	3.0	3.0	3.0	3.0
	Doctor	3.0	−3.0	3.0	3.0	−3.0	3.0	3.0	−2.5	−3.0	−3.0
	Me	3.0	−3.0	3.0	0.0	−3.0	3.0	3.0	−3.0	−3.0	−3.0
	Job	−3.0	3.0	0.0	1.5	3.0	0.0	−3.0	3.0	3.0	3.0
	Mental sick	−3.0	3.0	−1.5	0.0	−3.0	−2.5	−3.0	3.0	0.0	3.0
	Mother	−3.0	0.5	0.0	0.0	−3.0	3.0	0.0	3.0	3.0	−3.0
	Peace of mind	3.0	−3.0	3.0	3.0	−3.0	3.0	3.0	−3.0	−3.0	−3.0
	Fraud	2.5	−2.0	2.0	2.5	−2.5	1.0	2.0	−2.0	−2.5	−2.5
	Spouse	−1.5	3.0	0.0	−1.5	3.0	0.0	0.0	1.5	0.0	3.0
	Self-control	−0.5	0.5	−2.0	2.0	−2.0	0.5	1.0	−0.5	−2.0	0.5
	Hatred	2.0	−2.5	2.5	2.5	−3.0	2.0	2.5	−2.0	−2.5	−2.5
	Father	3.0	−3.0	−3.0	−3.0	0.0	3.0	−3.0	−3.0	3.0	−3.0
	Confusion	0.0	3.0	−2.5	0.0	0.5	−2.5	−3.0	0.0	0.0	0.0
	Sex	−3.0	3.0	−3.0	−3.0	0.0	0.0	−3.0	3.0	3.0	3.0

(continued)

Table 20.1 (continued)

Occasion	Concept	Adjective									
		Hot	Worthless	Relaxed	Large	Slow	Clean	Strong	Distasteful	Shallow	Passive
$\mathbf{X}_4$ *Eve Black 2*	Love	−3.0	3.0	−3.0	−3.0	−3.0	−2.0	−3.0	3.0	3.0	3.0
	Child	−3.0	2.0	−2.5	−3.0	−0.5	−1.0	−2.5	2.0	−1.5	3.0
	Doctor	3.0	−3.0	3.0	3.0	−3.0	3.0	3.0	−3.0	−3.0	−3.0
	Me	3.0	−3.0	3.0	3.0	−2.0	3.0	3.0	−3.0	−3.0	−3.0
	Job	−3.0	3.0	−3.0	−3.0	3.0	−3.0	−3.0	3.0	3.0	3.0
	Mental sick	−3.0	3.0	−3.0	−3.0	−3.0	−3.0	−3.0	3.0	3.0	3.0
	Mother	−2.5	−1.5	−2.0	0.5	−3.0	0.0	3.0	1.5	−2.5	−3.0
	Peace of mind	1.5	−1.0	2.5	1.5	−3.0	2.0	2.0	−2.5	−2.0	−3.0
	Fraud	2.0	−2.0	2.0	2.5	−2.5	2.0	2.0	−2.0	−2.0	−2.5
	Spouse	−3.0	3.0	−2.0	−3.0	2.5	−3.0	−3.0	3.0	3.0	3.0
	Self-control	−2.0	0.5	−2.0	0.5	−2.5	0.5	1.0	0.5	−2.5	0.0
	Hatred	2.0	−2.0	1.0	2.0	−2.5	2.5	2.0	−2.0	−2.0	−2.5
	Father	2.5	−3.0	−2.0	0.0	−2.5	3.0	0.0	−3.0	−2.0	−3.0
	Confusion	−3.0	3.0	−3.0	−3.0	−2.5	−3.0	−3.0	3.0	0.0	3.0
	Sex	−3.0	3.0	−3.0	−3.0	1.5	−3.0	−3.0	3.0	3.0	3.0
$\mathbf{X}_5$ *Jane 1*	Love	0.0	−3.0	3.0	3.0	3.0	3.0	3.0	−3.0	−3.0	−1.5
	Child	0.0	−3.0	1.0	3.0	−1.0	3.0	0.0	−3.0	−3.0	−3.0
	Doctor	0.0	−3.0	3.0	3.0	−3.0	3.0	3.0	−3.0	−3.0	−3.0
	Me	0.0	0.0	−2.0	0.0	0.5	3.0	0.5	−0.5	−2.0	−2.0
	Job	0.0	−3.0	−2.5	3.0	3.0	3.0	2.5	0.5	−3.0	0.5
	Mental sick	0.0	−2.0	−3.0	3.0	3.0	3.0	0.0	1.0	−3.0	−3.0
	Mother	0.5	−3.0	−2.5	1.5	−0.5	3.0	−2.0	−2.0	−3.0	−3.0
	Peace of mind	0.0	−3.0	3.0	3.0	3.0	3.0	3.0	−3.0	−3.0	1.5

(continued)

Table 20.1 (continued)

Occasion	Concept	Adjective									
		Hot	Worthless	Relaxed	Large	Slow	Clean	Strong	Distasteful	Shallow	Passive
	Fraud	−3.0	3.0	−2.5	0.5	−2.5	−3.0	−3.0	3.0	−3.0	2.0
	Spouse	0.0	0.0	3.0	3.0	−1.0	3.0	3.0	−3.0	−2.0	0.0
	Self-control	0.0	−3.0	3.0	3.0	3.0	3.0	3.0	−3.0	−3.0	−1.5
	Hatred	−3.0	2.0	−3.0	0.5	3.0	−2.0	−3.0	3.0	−3.0	−1.5
	Father	2.0	−3.0	−2.5	3.0	−2.0	3.0	3.0	−3.0	−3.0	−3.0
	Confusion	0.0	−1.5	−3.0	2.0	−2.5	3.0	0.0	3.0	−3.0	−2.5
	Sex	0.0	−3.0	3.0	2.5	0.0	3.0	3.0	−3.0	0.0	0.0
$\mathbf{X}_6$ *Jane 2*	Love	2.0	−3.0	3.0	3.0	2.5	3.0	3.0	−3.0	−3.0	−2.5
	Child	2.5	−3.0	1.0	2.5	−2.5	3.0	2.0	−3.0	−2.0	−3.0
	Doctor	2.5	−3.0	3.0	3.0	−2.0	3.0	3.0	−3.0	−2.5	−3.0
	Me	1.5	−2.0	−3.0	0.5	−2.0	3.0	0.0	−2.0	−1.0	−2.5
	Job	2.0	−3.0	−2.0	2.5	3.0	3.0	2.5	−2.0	−2.5	−2.0
	Mental sick	0.0	−2.0	−3.0	2.5	2.0	3.0	1.0	0.0	−2.5	−3.0
	Mother	0.5	−3.0	−2.0	2.5	−2.0	3.0	2.0	−−3.0	−2.0	−3.0
	Peace of mind	2.5	−3.0	3.0	3.0	2.0	3.0	3.0	−3.0	−2.5	−1.5
	Fraud	−2.0	3.0	−2.0	−2.5	−2.0	−2.0	−2.0	3.0	−2.0	0.5
	Spouse	2.0	−3.0	3.0	3.0	−1.5	3.0	3.0	−3.0	−2.0	−3.0
	Self-control	2.0	−3.0	2.5	3.0	2.0	3.0	3.0	−3.0	−2.5	−2.0
	Hatred	−2.5	3.0	−2.5	−2.0	0.0	−2.0	−2.0	3.0	−2.0	−2.0
	Father	2.5	−3.0	−2.0	3.0	−2.0	3.0	3.0	−3.0	−2.5	−3.0
	Confusion	0.5	0.0	−2.5	0.0	−2.0	0.0	0.0	1.5	−2.5	−1.5
	Sex	2.5	−3.0	3.0	3.0	0.0	3.0	2.5	−3.0	−3.0	−2.5

$$\mathbf{X}_k = \mathbf{A}_k\mathbf{B}' + \mathbf{E}_k. \tag{20.7}$$

This PCA model is also called *Tucker1* (Tucker, 1966), as *constraining* it leads to Tucker2, and Tucker3, as explained in the following paragraphs.

The *Tucker2* model is derived by constraining $\mathbf{A}_k$ in (20.7) as $\mathbf{A}_k = \mathbf{A}\mathbf{H}_k$:

$$\mathbf{X}_k = \mathbf{A}\mathbf{H}_k\mathbf{B}' + \mathbf{E}_k \tag{20.8}$$

with $\mathbf{A}$ being $I \times P$, $\mathbf{H}_k$ being $P \times Q$, and $I \geq P$. Here, it should be noted that $\mathbf{H}_k$ has the subscript k which is not possessed by $\mathbf{A}$ and $\mathbf{B}$: they are *invariant* across different k, while $\mathbf{H}_k$ serves to *explain* the *differences* in $\mathbf{X}_k$ across k.

In order to show how the Tucker2 model is constrained to lead to Tucker3, we arrange the PQR elements in the core array (20.3) in $P \times Q$ matrices $\mathbf{G}_1, \ldots, \mathbf{G}_r, \ldots, \mathbf{G}_R$, with g_{pqr} the (p, q) element of $\mathbf{G}_r$:

$$\mathbf{G}_r = \left(g_{pqr}\right) \quad (r = 1, \ldots, R). \tag{20.9}$$

Tucker3 is a *constrained version of Tucker2* modeled as (20.8), in which $\mathbf{H}_k$ is restricted to the sum of (20.9) weighted by c_{kr} in (20.2):

$$\mathbf{H}_k = c_{k1}\mathbf{G}_1 + \ldots + c_{kR}\mathbf{G}_R = \sum_{r=1}^{R} c_{kr}\mathbf{G}_r \tag{20.10}$$

with $R \leq K$. Using (20.10) in (20.8) leads to the Tucker3 model

$$\mathbf{X}_k = \mathbf{A}\left(\sum_{r=1}^{R} c_{kr}\mathbf{G}_r\right)\mathbf{B}' + \mathbf{E}_k. \tag{20.11}$$

Its equivalence to (20.2) is shown next:

Note 20.1. Product of Three Matrices

In this chapter, we often encounter the products of three matrices. An example of them is the product $\mathbf{A}\left(\sum_{r=1}^{R} c_{kp}\mathbf{G}_r\right)\mathbf{B}'$ in (20.11), which is rewritten as $\sum_{r=1}^{R} \mathbf{A}\left(c_{kp}\mathbf{G}_r\right)\mathbf{B}'$.

Using $\tilde{\mathbf{a}}_i' = [a_{i1}, \ldots, a_{iP}]$ for the ith row of $\mathbf{A}$ and $\tilde{\mathbf{b}}_j' = \left[b_{j1}, \ldots, b_{jQ}\right]$ for the jth row of $\mathbf{B}$, we can find that the (i, j) element of $\mathbf{A}(c_{kr}\mathbf{G}_r)\mathbf{B}'$ is expressed as

$$\tilde{\mathbf{a}}_i'(c_{kr}\mathbf{G}_r)\tilde{\mathbf{b}}_j = \sum_{p=1}^{P}\sum_{q=1}^{Q} a_{ip}(c_{kr}g_{pqr})b_{jq}. \tag{20.12}$$

The (i, j) element of $\mathbf{A}\left(\sum_{r=1}^{R} c_{kp}\mathbf{G}_r\right)\mathbf{B}' = \sum_{r=1}^{R} \mathbf{A}\left(c_{kp}\mathbf{G}_r\right)\mathbf{B}'$ is given by the sum of (20.12) over $r = 1, \ldots, R$:

$$\sum_{r=1}^{R} \tilde{\mathbf{a}}_i'(c_{kr}\mathbf{G}_r)\tilde{\mathbf{b}}_j = \sum_{r=1}^{R}\sum_{p=1}^{P}\sum_{r=1}^{Q} a_{ip}(c_{kr}g_{pqr})b_{jq} = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} a_{ip}b_{jq}c_{kr}g_{pqr}. \tag{20.13}$$

Equalizing this plus e_{ijk} to x_{ijk} leads to (20.2), with e_{ijk} the (i, j) element of $\mathbf{E}_k$.

Finally, *Parafac* is a *constrained version of the Tucker3* modeled in (20.11), in which P, Q, and R are constrained through $P = Q = R$ and $\mathbf{G}_r$ is restricted to a matrix filled with zeros except for the rth diagonal element which equals one. Hence, the Parafac model can be expressed as

$$\mathbf{X}_k = \mathbf{A}\mathbf{D}_k\mathbf{B}' + \mathbf{E}_k, \tag{20.14}$$

with

$$\mathbf{D}_k = \begin{bmatrix} c_{k1} & & \\ & \ddots & \\ & & c_{kP} \end{bmatrix} \tag{20.15}$$

being the diagonal matrix whose diagonal elements are $c_{k1}, \ldots, c_{kP}$. The (i, j) element of $\mathbf{A}\mathbf{D}_k\mathbf{B}'$ in (20.14) is expressed as $\tilde{\mathbf{a}}_i'\mathbf{D}_k\tilde{\mathbf{b}}_j = \sum_{p=1}^{P} a_{ip}c_{kp}b_{jq}$, which allows us to find the equivalence of (20.4) and (20.14).

The above facts lead to the hierarchical relationship in (20.6). It shows that 3WPCA procedures are directly derived from the PCA in Chap. 5 and are not extensions of PCA but rather constrained versions.

Before describing how the parameters are estimated in 3WPCA, we illustrate the Parafac and Tucker3 solutions for the data set in Table 20.1 and explain how the solutions are interpreted on the basis of (20.11) and (20.14), in Sects 20.3 and 20.4.

20.3 Parafac Solution

The data in Table 20.1 consist of the I (=15 concepts) $\times$ J (=10 adjectives) matrices $\mathbf{X}_k$ obtained over K = 6 occasions. The data were originally observed by Osgood and Luria's (1954) for a female suffering from *triple personality disorder*: she had the three types of personalities named *Eve White*, *Eve Black*, and *Jane*. During episodes for each of her three personality types, data were observed on two occasions, as found in the left column in Table 20.1. Its element x_{ijk} indicates her rating on occasion k and has been transformed from Osgood and Luria's (1954) original score so that x_{ijk} ranges from −3 to 3. That is, positive x_{ijk} stands for what

extent the concept i is rated to be featured by adjective j, while negative x_{ijk} indicates how strongly the concept i is featured by the antonym of adjective j. For example, $x_{111} = 1$ stands for her thinking *love* to be *hot* at degree 1, while $x_{114} = -3$ indicates that she thinks *love* to be *hot* at degree −3, i.e., to be rather *cold* at degree 3.

For the data in Table 20.1, Parafac was performed with $P = 3$. The solution is presented in Table 20.2. Here, the columns (i.e., components) in **A** and **B** are indicated by the labels, *belongings*, *ill feeling*, … , *goodness*, referring to the interpretation of those components. For example, the second component in **A** has been named *ill feeling*, as this can be associated with the concepts of *mental sick* and *confusion* showing the large positive loadings for that component. On the other hand, the third component in **B** can be interpreted as representing *goodness*, as the adjectives *large* and *clean* associated with *goodness* present positive loadings, but

Table 20.2 Parafac solution for the data set in Table 20.1

Concept	A			B			
	Belongings	Ill Feeling	Routine	Adjective	Activity	Erethism	Goodness
Love	−0.15	0.00	**0.39**	Hot	**0.31**	−0.19	0.11
Child	−0.10	0.14	0.24	Worthless	−0.07	−0.01	**−0.44**
Doctor	**0.50**	−0.09	0.26	Relaxed	**0.30**	**−0.64**	0.04
Me	**0.34**	0.19	0.00	Large	−0.08	0.10	**0.44**
Job	**−0.33**	0.15	**0.3**1	Slow	**−0.73**	−0.12	**0.35**
Mental sick	−0.12	**0.44**	0.25	Clean	0.03	0.10	**0.44**
Mother	0.20	0.27	0.24	Strong	0.18	**−0.30**	0.27
Peace of mind	**0.41**	−0.25	**0.33**	Distasteful	−0.15	0.27	**−0.34**
Fraud	0.22	**0.34**	**−0.30**	Shallow	−0.17	**−0.42**	−0.28
Spouse	−0.22	−0.05	0.19	Passive	**−0.44**	**−0.43**	−0.13
Self-control	0.12	−0.02	**0.33**			**C**	
Hatred	0.21	**0.42**	−0.21	Occasion	*B-A*[a]	*I-E*[b]	*R-G*[c]
Father	0.25	0.15	0.28	Eve White 1	1.52	12.39	20.34
Confusion	−0.11	**0.52**	0.03	Eve White 2	3.74	11.41	17.73
Sex	−0.24	−0.01	0.21	Eve Black 1	17.47	2.59	−5.40
				Eve Black 2	17.96	4.01	−9.43
				Jane 1	2.38	12.69	21.35
				Jane 2	3.32	9.70	23.20

[a]Belongings–Activity
[b]Ill Feeling–Erethism
[c]Routine–Goodness

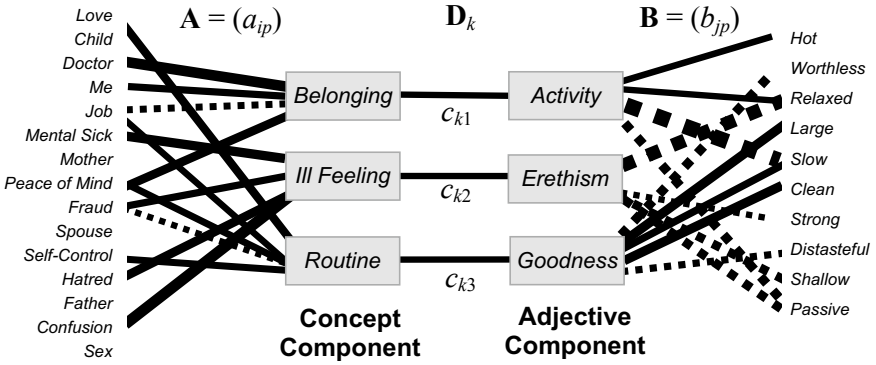


Fig. 20.2 Network representation of the Parafac solution in Table 20.2

worthless and *distasteful* associated with *badness* show negative loadings for the component. The other components have been interpreted in a similar manner.

The Parafac solution can be visually represented in the network diagram in Fig. 20.2. On the left side, the concept i and component p with $|a_{ip}| > 0.3$ are linked, while on the right side the adjective j and component p with $|b_{jp}| > 0.3$ are linked. Here, the widths of the paths expressing links are proportional to the absolute values of the corresponding loadings, with positive or negative value indicated by the paths being real or dotted.

As found in middle section of Fig. 20.2, the diagonal elements of $\mathbf{D}_k$ in (20.15), i.e., the elements of $\mathbf{C} = (c_{kp})$, indicate the relationships of the components in $\mathbf{A}$ to those in $\mathbf{B}$. This role of c_{kp} can be understood by noting that P routes exist from concept i to adjective j. For example, from *Fraud* to *Slow*, there are possible $P = 3$ routes: [1] *Fraud–Belongings–Activity–Slow*, [2] *Fraud–Ill Feeling–Erethism–Slow*, and [3] *Fraud–* Routine*–Goodness–Slow*, with the symbol "–" indicating a path. Here, each route consists of the three paths, which are associated with coefficients a_{ip}, c_{kp}, and b_{jp}. Their product $a_{ip}c_{kp}b_{jp} = a_{ip}b_{jp}c_{kp}$ is summed over $p = 1, \ldots, P$ to provide the model part $\sum_{p=1}^{P} a_{ip}b_{jp}c_{kp}$ in (20.4) approximating x_{ijk}. For this reason, Fig. 20.2 can be viewed as the *network representation of the Parafac solution* in Table 20.2.

We should note in (20.4) that c_{kp} has the subscript k, which is not possessed by a_{ip} and b_{jp}. This implies that the left and right links associated with $\mathbf{A}$ and $\mathbf{B}$ in Fig. 20.2 are *invariant across* $k = 1, \ldots, K$, but the *inter-component links* (middle of the figure) *differ* across K occasions. The inter-occasion differences can be seen by noting the $\mathbf{C} = (c_{kp})$ solution in Table 20.2. Here, we can find that the c_{kp} values for the four occasions concerning *Eve White* and *Jane* are mutually similar, but differ from the two occasions concerning *Eve Black*. That is, the occasions ($k = 1, \ldots, 6$) are classified into the *Eve White–Jane* (*EWJ*) and *Eve Black* (*EB*) groups. In the former, the c_{kp} values are all positive, which implies that the three pairs of components, [1] *Belongings–Activity*, [2] *Ill Feeling–Erethism*, and [3] *Routine–*

Goodness, have positive associations. On the other hand, the c_{k3} values in the *EB* group are negative, which implies that *Routine* is negatively associated with *Goodness*. The two groups also differ in the following ways: in the *EWJ* group, the associations in [2] and [3] are stronger than that in [1], as c_{k2} and c_{k3} are larger than c_{k1}. In contrast, the association in [1] is remarkably stronger in the *EB* group.

20.4 Tucker3 Solution

For the data in Table 20.1, we preformed Tucker3 with $P = Q = 3$ and $R = 2$. The resulting solution is presented in Table 20.3. Here, the names of components are the same as in Table 20.2, as the components in **A** and **B** can be interpreted in the same manner as in the Parafac solution.

As described in Sect. 20.2, the Tucker3 model (20.11) follows from (20.8) with $\mathbf{H}_k$ constrained as (20.10). On the basis of (20.8), the Tucker3 solution in Table 20.3 can be represented in the network diagram in Fig. 20.3, which is depicted in the same manner as in Fig. 20.2. How the diagram in Fig. 20.3 relates to (20.8) can be understood by noting the routes from concepts to adjectives. Let $R_{i\text{-}p\text{-}q\text{-}j}$ denote the route from concept i to adjective j by way of the pth component linked to i and the qth component linked to j. Then, all possible routes from i to j are expressed as $\{R_{i\text{-}p\text{-}q\text{-}j}\colon p = 1, \ldots, P; q = 1, \ldots, Q\}$. Each of these routes consists of three paths associated with coefficients a_{ip}, h_{pqk}, and b_{jq}, where h_{pqk} is the (p, q) element of $\mathbf{H}_k$. The product of those coefficients, $a_{ip}h_{pqk}b_{jq}$, are summed over $p = 1, \ldots, P$ and $q = 1, \ldots, Q$ to provide $\sum_{p=1}^{P}\sum_{q=1}^{Q} a_{ip}h_{pqk}b_{jq}$, which is the (i, j) element of the model part $\mathbf{AH}_k\mathbf{B}'$ in (20.8).

We should note that the *inter-component links* shown in the middle of Fig. 20.3 differ from the corresponding links in Fig. 20.2. In Fig. 20.2, the components in **A** are linked in parallel with those in **B**, since the matrix $\mathbf{D}_k$ specifying the links is diagonal with its off-diagonal elements zeros. In contrast, in Fig. 20.3 each component in **A** is linked to all of the ones in **B**, since the matrix $\mathbf{H}_k$ in (20.8) specifying the links is not diagonal.

Matrix $\mathbf{H}_k$ is constrained as in (20.10). It is expressed as $\mathbf{H}_k = c_{k1}\mathbf{G}_1 + c_{k2}\mathbf{G}_2$ in this example with $R = 2$. Here, $\mathbf{G}_1$ and $\mathbf{G}_2$ do not have the subscript k, which implies that the differences of $\mathbf{H}_k$ across occasions $k = 1, \ldots, K$ are specified by weights c_{kr}. In Table 20.3, the similarities/differences of the $[c_{k1}, c_{k2}]$ values among the six occasions show that they are classified into *Eve White–Jane* (*EWJ*) and *Eve Black* (*EB*) groups. Figure 20.4a, b illustrate $\mathbf{H}_1 = c_{11}\mathbf{G}_1 + c_{12}\mathbf{G}_2$ for *Eve White 1* in the former group and $\mathbf{H}_4 = c_{41}\mathbf{G}_1 + c_{42}\mathbf{G}_2$ for *Eve Black 2* in the latter group. Here, the striking difference between $\mathbf{H}_1$ and $\mathbf{H}_4$ is that the positive link between *Belongingness* and *Goodness* is far stronger in (b) than in (a), and that the link between *Routine* and *Goodness* in (a) shows *Routine* being rated *Good*, whereas the counterpart in (b) does not, implying it is to some extent bad.

Table 20.3 Tucker3 solution for the data set in Table 20.1

A				B				C		
Concept	Belongings	Ill feeling	Routine	Adjective	Activity	Erethism	Goodness	Occasion	G_1	G_2
Love	−0.18	0.00	**0.38**	Hot	0.15	−0.21	**0.30**	Eve White 1	0.17	0.46
Child	−0.09	0.12	0.28	Worthless	0.27	−0.02	**−0.36**	Eve White 2	0.23	0.41
Doctor	**0.50**	−0.04	0.28	Relaxed	0.17	**−0.66**	0.26	Eve Black 1	0.61	−0.20
Me	**0.32**	0.22	−0.03	Large	**−0.34**	0.12	0.26	Eve Black 2	0.72	−0.26
Job	−0.22	0.09	0.29	Slow	**−0.75**	−0.13	−0.25	Jane 1	0.14	0.49
Mental sick	−0.21	**0.44**	0.21	Clean	**−0.32**	0.15	**0.32**	Jane 2	0.12	0.53
Mother	0.16	**0.31**	0.25	Strong	−0.08	−0.27	**0.32**			
Peace of mind	**0.41**	−0.20	0.27	Distasteful	0.14	0.24	**−0.36**			
Fraud	0.16	**0.36**	**−0.30**	Shallow	0.04	**−0.42**	**−0.31**			
Spouse	−0.21	−0.08	0.22	Passive	−0.26	**−0.41**	**−0.39**			
Self-control	0.15	−0.01	**0.31**			**G_1**			**G_2**	
Hatred	0.18	**0.42**	−0.25		Activity	Erethism	Goodness	Activity	Erethism	Goodness
Father	0.28	0.17	0.29	Belonging	−0.05	−0.20	**35.34**	−0.07	−0.38	−3.43
Confusion	−0.19	**0.51**	0.05	Ill feeling	5.90	**12.06**	−0.29	−0.05	**20.04**	0.00
Sex	−0.28	−0.02	0.22	Routine	−3.36	4.82	0.00	**−18.64**	0.06	**41.50**

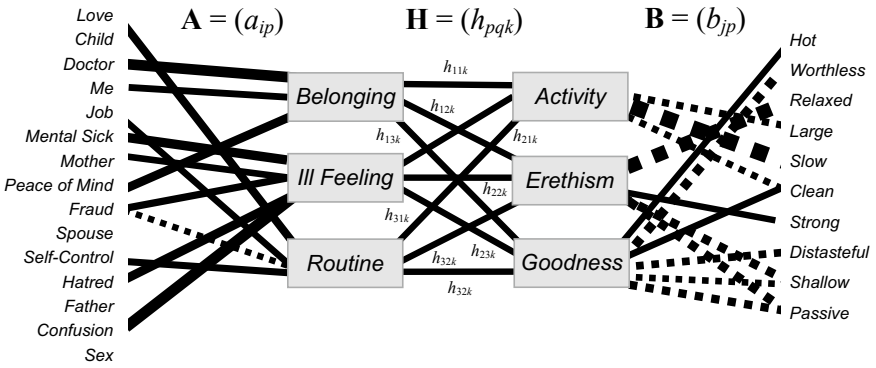


Fig. 20.3 Network representation of the Tucker3 solution in Table 20.3

(a) $\mathbf{H}_1$: Eve White 1 (b) $\mathbf{H}_4$: Eve Black 2

Fig. 20.4 Inter-component links across two occasions, with the links whose absolute values are less than 0.1 omitted, the widths of the paths proportional to the absolute values of the corresponding elements, and positive or negative value indicated by the paths being real or dotted

20.5 Unconstrained Parafac Algorithm

Algorithms for 3WPCA are described in custom by arranging (20.5) *horizontally* in the $I \times KJ$ block matrix

$$\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_k, \ldots, \mathbf{X}_K]. \tag{20.16}$$

For this data matrix, the Parafac model (20.4) or (20.14) can be rewritten as

$$\mathbf{X} = \mathbf{A}(\mathbf{C} \bullet \mathbf{B})' + \mathbf{E}. \tag{20.17}$$

using the *Khatri–Rao product* defined in (17.60). Here, $\mathbf{E} = [\mathbf{E}_1, \ldots, \mathbf{E}_k, \ldots, \mathbf{E}_K]$ is the $I \times KJ$ matrix with the (i, j) element of $\mathbf{E}_k$ being e_{ijk}. The equivalence of (20.17) to (20.14) is explained next:

Note 20.2. Parafac Model with the Khatri–Rao Product (Part 1)
The transpose of (17.60) pre-multiplied by $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_P]$ is expressed as

$$\begin{aligned}\mathbf{A}(\mathbf{C} \bullet \mathbf{B})' &= [\mathbf{a}_1, \ldots, \mathbf{a}_P] \begin{bmatrix} c_{11}\mathbf{b}_1' & \cdots & c_{K1}\mathbf{b}_1' \\ \vdots & \vdots & \vdots \\ c_{1P}\mathbf{b}_P' & \cdots & c_{KP}\mathbf{b}_P' \end{bmatrix} \\ &= \left[\sum_{p=1}^{P} c_{1p}\mathbf{a}_p\mathbf{b}_p', \ldots, \sum_{p=1}^{P} c_{Kp}\mathbf{a}_p\mathbf{b}_p' \right],\end{aligned}$$

whose kth block is $\sum_{p=1}^{P} c_{kp}\mathbf{a}_p\mathbf{b}_p'$, with $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_P]$. By taking account of this result and (20.16), we can find that (20.17) is rewritten as

$$\mathbf{X}_k = \sum_{p=1}^{P} c_{kp}\mathbf{a}_p\mathbf{b}_p' + \mathbf{E}_k = [\mathbf{a}_1, \ldots, \mathbf{a}_P] \begin{bmatrix} c_{k1} & & \\ & \ddots & \\ & & c_{kP} \end{bmatrix} \begin{bmatrix} \mathbf{b}_1' \\ \vdots \\ \mathbf{b}_P' \end{bmatrix} + \mathbf{E}_k,$$

i.e., (20.14) with the substitution (20.15).

The Parafac model (20.17) is equivalent to

$$\mathbf{X}^{\#} = \mathbf{B}(\mathbf{C} \bullet \mathbf{A})' + \mathbf{E}^{\#}, \tag{20.18}$$

$$\mathbf{X}^{*} = \mathbf{C}(\mathbf{B} \bullet \mathbf{A})' + \mathbf{E}^{*}, \tag{20.19}$$

as shown in the next note. Here, $\mathbf{X}^{\#}$ ($J \times KI$) and $\mathbf{X}^{*}$ ($K \times JI$) are obtained by arranging (20.5) as

$$\mathbf{X}^{\#} = \left[\mathbf{X}_1', \ldots, \mathbf{X}_k', \ldots, \mathbf{X}_K'\right], \tag{20.20}$$

$$\mathbf{X}^{*} = \begin{bmatrix} \mathrm{vec}(\mathbf{X}_1)' \\ \vdots \\ \mathrm{vec}(\mathbf{X}_K)' \end{bmatrix} = \left[\mathbf{X}_1^{*}, \ldots, \mathbf{X}_j^{*}, \ldots, \mathbf{X}_J^{*}\right], \tag{20.21}$$

and the error matrices $\mathbf{E}^{\#}$ and $\mathbf{E}^{*}$ are the variants of $\mathbf{E}$ whose elements are arranged so as to correspond to (20.20) and (20.21), respectively, with vec() defined as (17.63) and $\mathbf{X}_j^{*}$ the $K \times I$ matrix whose (k, i) element is x_{ijk}. The equivalence of (20.17) to (20.18) and (20.19) is shown next:

Note 20.3. Parafac Model with the Khatri–Rao Product (Part 2)
Swapping the matrices $\mathbf{A}$ and $\mathbf{B}$ in Note 20.2 leads to

$$\mathbf{B}(\mathbf{C} \bullet \mathbf{A})' = [\mathbf{b}_1, \ldots, \mathbf{b}_P] \begin{bmatrix} c_{11}\mathbf{a}_1' & \cdots & c_{K1}\mathbf{a}_1' \\ \vdots & \vdots & \vdots \\ c_{1P}\mathbf{a}_P' & \cdots & c_{KP}\mathbf{a}_P' \end{bmatrix} = \left[\sum_{p=1}^{P} c_{1p}\mathbf{b}_p\mathbf{a}_p', \ldots, \sum_{p=1}^{P} c_{Kp}\mathbf{b}_p\mathbf{a}_p' \right],$$

whose kth block is $\sum_{p=1}^{P} c_{kp}\mathbf{b}_p\mathbf{a}_p'$. Since this is the transpose of $\sum_{p=1}^{P} c_{kp}\mathbf{a}_p\mathbf{b}_p'$ in Note 20.2, we can find that $\sum_{p=1}^{P} c_{kp}\mathbf{b}_p\mathbf{a}_p'$ corresponds to $\mathbf{X}_k'$ so that (20.17) is rewritten as (20.18).

By using $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_P]$ and swapping $\mathbf{B}$ and $\mathbf{C}$ in the above $\mathbf{B}(\mathbf{C} \bullet \mathbf{A})'$, we have

$$\mathbf{C}(\mathbf{B} \bullet \mathbf{A})' = [\mathbf{c}_1, \ldots, \mathbf{c}_P] \begin{bmatrix} b_{11}\mathbf{a}_1' & \cdots & b_{J1}\mathbf{a}_1' \\ \vdots & \vdots & \vdots \\ b_{1P}\mathbf{a}_P' & \cdots & b_{JP}\mathbf{a}_P' \end{bmatrix} = \left[\sum_{p=1}^{P} b_{1p}\mathbf{c}_p\mathbf{a}_p', \ldots, \sum_{p=1}^{P} b_{Jp}\mathbf{c}_p\mathbf{a}_p' \right],$$

whose jth block is the $K \times I$ matrix $\sum_{p=1}^{P} b_{jp}\mathbf{c}_p\mathbf{a}_p'$ with its (k, i) element $\sum_{p=1}^{P} b_{jp}c_{kp}a_{ip} = \sum_{p=1}^{P} a_{ip}b_{jp}c_{kp}$. This corresponds to the (k, i) element of $\mathbf{X}_j^*$ in (20.21), i.e., x_{ijk}, which shows the equivalence of (20.19) to (20.4) or (20.17).

Using (20.17)–(20.19), the sum of the squared errors for Parafac is expressed in three forms as in

$$f_{\mathrm{P}}(\mathbf{A}, \mathbf{B}, \mathbf{C}) = \left\|\mathbf{X} - \mathbf{A}(\mathbf{C} \bullet \mathbf{B})'\right\|^2 = \left\|\mathbf{X}^{\#} - \mathbf{B}(\mathbf{C} \bullet \mathbf{A})'\right\|^2 = \left\|\mathbf{X}^{*} - \mathbf{C}(\mathbf{B} \bullet \mathbf{A})'\right\|^2, \tag{20.22}$$

This minimization over $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ can be attained by alternately solving the following three problems:

[P1] minimize $\|\mathbf{X} - \mathbf{A}(\mathbf{C} \bullet \mathbf{B})'\|^2$ over $\mathbf{A}$ with $\mathbf{B}$ and $\mathbf{C}$ kept fixed,
[P2] minimize $\|\mathbf{X}^{\#} - \mathbf{B}(\mathbf{C} \bullet \mathbf{A})'\|^2$ over $\mathbf{B}$ with $\mathbf{A}$ and $\mathbf{C}$ kept fixed,
[P3] minimize $\|\mathbf{X}^{*} - \mathbf{C}(\mathbf{B} \bullet \mathbf{A})'\|^2$ over $\mathbf{C}$ with $\mathbf{A}$ and $\mathbf{B}$ kept fixed.

Here, we can find every problem to be the *regression* of data onto the matrix defined by the Khatri–Rao product. For example, $\left\|\mathbf{X} - \mathbf{A}(\mathbf{C} \bullet \mathbf{B})'\right\|^2 = \|\mathbf{X}' - (\mathbf{C} \bullet \mathbf{B})\mathbf{A}'\|^2$ in [P1] is the loss function for the regression of $\mathbf{X}'$ onto $(\mathbf{C} \bullet \mathbf{B})$ with $\mathbf{A}'$ the coefficient matrix. Its solution is explicitly given by $\mathbf{A}' = (\mathbf{C} \bullet \mathbf{B})^+\mathbf{X}'$, or equivalently, $\mathbf{A} = \mathbf{X}(\mathbf{C} \bullet \mathbf{B})^{+\prime}$, as explained with (17.28) in Sect. 17.4. In a parallel manner, the solutions for [P2] and [P3] can be obtained. Thus, the Parafac algorithm for minimizing (20.22) can be summarized as follows:

Step 1. Initialize $\mathbf{B}$ and $\mathbf{C}$.
Step 2. Update $\mathbf{A} = \mathbf{X}(\mathbf{C} \bullet \mathbf{B})^{+\prime}$.
Step 3. Update $\mathbf{B} = \mathbf{X}^{\#}(\mathbf{C} \bullet \mathbf{A})^{+\prime}$.
Step 4. Update $\mathbf{C} = \mathbf{X}^{*}(\mathbf{B} \bullet \mathbf{A})^{+\prime}$.
Step 5. Finish if convergence is reached; otherwise, go back to Step 2.

20.6 Constrained Parafac Algorithm

A drawback of the procedure in the last section is that it sometimes provides the solutions in which $\mathbf{A}$, $\mathbf{B}$, or $\mathbf{C}$ is *nearly rank-deficient*. This term is explained next:

Note 20.4. Nearly Rank-Deficient Matrix and Condition Number
Let $\mathbf{N}$ be an $n \times p$ matrix to be determined. If rank($\mathbf{N}$) is *nearly* less than min(n, p), $\mathbf{N}$ can be said to be *nearly* rank-deficient. Here, we should note the adverb *nearly*. For example, let $\mathbf{R} = \begin{bmatrix} 1 & 3 \\ 2 & 6 \end{bmatrix}$ and $\mathbf{N} = \begin{bmatrix} 1 & 2.9 \\ 2 & 6.1 \end{bmatrix}$. We find rank($\mathbf{R}$) = 1 < rank($\mathbf{N}$) = 2, but $\mathbf{N} \cong \mathbf{R}$. Then, $\mathbf{N}$ can be said to be *nearly* rank-deficient.

How *nearly* rank($\mathbf{N}$) is less than min(n, p) can be indicated by the largeness of the *condition number*, which is defined as the ratio of the largest singular value of $\mathbf{N}$ to its smallest nonzero singular value.

A solution including a nearly rank-deficient matrix is not useful, as rows/columns are indistinctive and redundant.

A remedy for avoiding such solutions in Parafac is to impose the *column-orthonormality constraints* on two of $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ (Kroonenberg, 2008; Smilde, Bro, & Geladi, 2004). One example is

$$\mathbf{A}'\mathbf{A} = \mathbf{B}'\mathbf{B} = \mathbf{I}_P. \tag{20.23}$$

A *constrained Parafac* procedure can be formulated by minimizing (20.22) under (20.23). Indeed, this procedure has been used for the solution in Table 20.2, whose algorithm consists of alternately solving the problems [P1], [P2], and [P3] in Sect. 20.5 subject to (20.23).

Let us consider [P1] subject to (20.23), i.e., minimizing $f_{\mathrm{P}}(\mathbf{A}) = \|\mathbf{X} - \mathbf{A}(\mathbf{C} \bullet \mathbf{B})'\|^2$ over $\mathbf{A}$ under $\mathbf{A}'\mathbf{A} = \mathbf{I}_P$. Using this constraint, $f_{\mathrm{P}}(\mathbf{A})$ can be rewritten as $f_{\mathrm{P}}(\mathbf{A}) = \|\mathbf{X}\|^2 - 2\mathrm{tr}(\mathbf{C} \bullet \mathbf{B})'\mathbf{X}'\mathbf{A} + \|(\mathbf{C} \bullet \mathbf{B})\|^2$, which shows that the problem amounts to maximizing $\mathrm{tr}(\mathbf{C} \bullet \mathbf{B})'\mathbf{X}'\mathbf{A}$ over $\mathbf{A}$ subject to $\mathbf{A}'\mathbf{A} = \mathbf{I}_P$. This can be attained through the *singular value decomposition* (*SVD*) defined as $\mathbf{X}(\mathbf{C} \bullet \mathbf{B}) = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}'$, with $\mathbf{U}'\mathbf{U} = \mathbf{V}'\mathbf{V} = \mathbf{I}_P$ and $\boldsymbol{\Delta}$ a $P \times P$ diagonal matrix, as found in Theorem A.4.2: the optimal $\mathbf{A}$ is given by $\mathbf{A} = \mathbf{U}\mathbf{V}'$. Analogously, the solution of [P2] subject to $\mathbf{B}'\mathbf{B} = \mathbf{I}_P$ in (20.23) is given by $\mathbf{B} = \mathbf{U}^{\#}\mathbf{V}^{\#\prime}$, whose right side hand is obtained through the SVD $\mathbf{X}^{\#}(\mathbf{C} \bullet \mathbf{A}) = \mathbf{U}^{\#}\boldsymbol{\Delta}^{\#}\mathbf{V}^{\#\prime}$. Since the remaining parameter matrix $\mathbf{C}$ is unconstrained, its solution is obtained by Step 4 in Sect. 20.5: with $\mathbf{A}$ and $\mathbf{B}$ fixed, the optimal $\mathbf{C}$ is given by $\mathbf{C} = \mathbf{X}^{*}(\mathbf{B} \bullet \mathbf{A})^{+\prime}$. Here, $(\mathbf{B} \bullet \mathbf{A})^{+\prime}$ can be *simplified* into $\mathbf{B} \bullet \mathbf{A}$ under (20.23), which is derived using the identity (17.61) with (20.23):

$$(\mathbf{B} \bullet \mathbf{A})'(\mathbf{B} \bullet \mathbf{A}) = \mathbf{I}_P. \tag{20.24}$$

By comparing this with (17.8), we find $(\mathbf{B} \bullet \mathbf{A})^{+\prime} = (\mathbf{B} \bullet \mathbf{A})'' = \mathbf{B} \bullet \mathbf{A}$. Thus, the Parafac algorithm subject to (20.23) can be listed as follows:

Step 1. Initialize $\mathbf{B}$ and $\mathbf{C}$.
Step 2. Update $\mathbf{A}$ with $\mathbf{A} = \mathbf{U}\mathbf{V}'$.
Step 3. Update $\mathbf{B}$ with $\mathbf{B} = \mathbf{U}^{\#}\mathbf{V}^{\#\prime}$.
Step 4. Update $\mathbf{C} = \mathbf{X}^{*}(\mathbf{B} \bullet \mathbf{A})$.
Step 5. Finish if convergence is reached; otherwise, go back to Step 2.

By substituting $\mathbf{C} = \mathbf{X}^{*}(\mathbf{B} \bullet \mathbf{A})$ into the final term of (20.22), the attained value of the loss function can be expressed as

$$\begin{aligned}
&\left\|\mathbf{X}^{*} - \mathbf{X}^{*}(\mathbf{B} \bullet \mathbf{A})(\mathbf{B} \bullet \mathbf{A})'\right\|^2 \\
&\quad = \|\mathbf{X}^{*}\|^2 - 2\mathrm{tr}\mathbf{X}^{*\prime}\mathbf{X}^{*}(\mathbf{B} \bullet \mathbf{A})(\mathbf{B} \bullet \mathbf{A})' + \mathrm{tr}(\mathbf{B} \bullet \mathbf{A})(\mathbf{B} \bullet \mathbf{A})'\mathbf{X}^{*}\mathbf{X}^{*}(\mathbf{B} \bullet \mathbf{A})(\mathbf{B} \bullet \mathbf{A})' \\
&\quad = \|\mathbf{X}^{*}\|^2 - \mathrm{tr}(\mathbf{B} \bullet \mathbf{A})'\mathbf{X}^{*\prime}\mathbf{X}^{*}(\mathbf{B} \bullet \mathbf{A}) = \|\mathbf{X}\|^2(1 - GOF_{\mathrm{P}}).
\end{aligned} \tag{20.25}$$

Here, (20.24) has been used and

$$GOF_{\mathrm{P}} = \frac{\|\mathbf{X}^{*}(\mathbf{B} \bullet \mathbf{A})\|^2}{\|\mathbf{X}\|^2} = \frac{\|\mathbf{C}\|^2}{\|\mathbf{X}\|^2}. \tag{20.26}$$

is the *standardized goodness-of-fit index*, which takes a value within the range [0, 1]. The monotonic increase in (20.26) with the iteration of Steps 2 to 4 follows from

the monotonic decrease of (20.25). We can use (20.26) to check the convergence, definable as a change in value of (20.26) from the previous round being small enough to be ignored. The resulting (20.26) value for the solution in Table 20.2 was 0.56.

20.7 Tucker3 Algorithm: The Optimal Core Array

For the block data matrix (20.16), the Tucker3 model (20.2) or (20.11) is rewritten as

$$\mathbf{X} = \mathbf{AG}(\mathbf{C} \otimes \mathbf{B})' + \mathbf{E}, \tag{20.27}$$

using the *Kronecker product* defined in (17.52). Here, $\mathbf{G} = [\mathbf{G}_1, \ldots, \mathbf{G}_R]$ is the $P \times RQ$ block matrix whose rth block $\mathbf{G}_r$ is defined as (20.9), and $\mathbf{E} = [\mathbf{E}_1, \ldots, \mathbf{E}_K]$ is the $I \times KJ$ matrix whose kth block is $\mathbf{E}_k$ $(I \times J)$ with its (i, j) element e_{ijk} corresponding to x_{ijk}. The equivalence of (20.27) to (20.11) is explained next:

Note 20.5. Expression of the Tucker3 Model with the Kronecker Product
Using (17.52) and (17.55), we have

$$\begin{aligned}
(\mathbf{C} \otimes \mathbf{B})' = \mathbf{C}' \otimes \mathbf{B}' &= \begin{bmatrix} c_{11}\mathbf{B}' & \cdots & c_{K1}\mathbf{B}' \\ \vdots & \vdots & \vdots \\ c_{1R}\mathbf{B}' & \cdots & c_{KR}\mathbf{B}' \end{bmatrix} \text{ and} \\
\mathbf{AG}(\mathbf{C} \otimes \mathbf{B})' &= \mathbf{A}[\mathbf{G}_1, \ldots, \mathbf{G}_R] \begin{bmatrix} c_{11}\mathbf{B}' & \cdots & c_{K1}\mathbf{B}' \\ \vdots & \vdots & \vdots \\ c_{1R}\mathbf{B}' & \cdots & c_{KR}\mathbf{B}' \end{bmatrix} \\
&= \mathbf{A}\left[\sum_{r=1}^{R} c_{1r}\mathbf{G}_r\mathbf{B}', \ldots, \sum_{r=1}^{R} c_{K}\mathbf{G}_r\mathbf{B}'\right] \\
&= \left[\mathbf{A}\left(\sum_{r=1}^{R} c_{1r}\mathbf{G}_r\right)\mathbf{B}', \ldots, \mathbf{A}\left(\sum_{r=1}^{R} c_{Kr}\mathbf{G}_r\right)\mathbf{B}'\right],
\end{aligned} \tag{20.28}$$

whose kth block is $\mathbf{A}\left(\sum_{r=1}^{R} c_{kr}\mathbf{G}_r\right)\mathbf{B}'$ $(I \times J)$. By comparing (20.27) and (20.28) with (20.16), we find that (20.27) is equivalent to (20.11).

Tucker3 is thus formulated by minimizing the least squares function for (20.27):

$$f_{\mathrm{T3}}(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{G}) = \|\mathbf{E}\|^2 = \left\|\mathbf{X} - \mathbf{A}\mathbf{G}(\mathbf{C} \otimes \mathbf{B})'\right\|^2. \tag{20.29}$$

Here, we can constrain $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ as

$$\mathbf{A}'\mathbf{A} = \mathbf{I}_P, \quad \mathbf{B}'\mathbf{B} = \mathbf{I}_Q, \quad \text{and} \quad \mathbf{C}'\mathbf{C} = \mathbf{I}_R \tag{20.30}$$

without loss of generality, since the minimum value of (20.29) remains the same whether (20.30) is imposed or not, as explained in Sect. 20.9.

Let us consider the minimization of (20.29) over $\mathbf{G}$ for given $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$. We can find that (20.29) has the same form as (17.33): the minimization of (20.29) is the *Penrose regression* problem formulated by minimizing (17.33). Thus, (17.34) shows that (20.29) is minimized for

$$\mathbf{G} = \mathbf{A}^+\mathbf{X}(\mathbf{C} \otimes \mathbf{B})^{+\prime} = \mathbf{A}'\mathbf{X}(\mathbf{C} \otimes \mathbf{B}). \tag{20.31}$$

Here, the last identity is derived using (17.7), (17.8), (17.55), and (17.59): those and (20.30) lead to $(\mathbf{C} \otimes \mathbf{B})^{\prime +} = (\mathbf{C}' \otimes \mathbf{B}')^+ = \mathbf{C}'^+ \otimes \mathbf{B}'^+ = \mathbf{C}^{+\prime} \otimes \mathbf{B}^{+\prime} = \mathbf{C}'' \otimes \mathbf{B}'' = \mathbf{C} \otimes \mathbf{B}$, with $\mathbf{A}^+ = \mathbf{A}'$ following from (17.8) and (20.30).

We can expand the right term in (20.31) as

$$\begin{aligned}\mathbf{A}'\mathbf{X}(\mathbf{C} \otimes \mathbf{B}) &= \mathbf{A}'[\mathbf{X}_1, \ldots, \mathbf{X}_K]\begin{bmatrix} c_{11}\mathbf{B} & \cdots & c_{1R}\mathbf{B} \\ \vdots & \vdots & \vdots \\ c_{K1}\mathbf{B} & \cdots & c_{KR}\mathbf{B} \end{bmatrix} \\ &= \left[\sum_{k=1}^{K} c_{k1}\mathbf{A}'\mathbf{X}_k\mathbf{B}, \ldots, \sum_{k=1}^{K} c_{lR}\mathbf{A}'\mathbf{X}_k\mathbf{B}\right].\end{aligned} \tag{20.32}$$

Its rth block is the $P \times Q$ matrix $\sum_{k=1}^{K} c_{kr}\mathbf{A}'\mathbf{X}_k\mathbf{B}$, whose (p, q) element is $\sum_{k=1}^{K} c_{kr}\mathbf{a}_p'\mathbf{X}_k\mathbf{b}_q$, while the rth block of $\mathbf{G}$ in (20.31) is $\mathbf{G}_r$ ($P \times Q$) whose (p, q) element is g_{pqr}. These facts show that (20.31) is rewritten as

$$\begin{aligned}g_{pqr} = \sum_{k=1}^{K} c_{kr}\mathbf{a}_p'\mathbf{X}_k\mathbf{b}_q &= \sum_{k=1}^{K} c_{kr}\left[\sum_{i=1}^{I} a_{ip}x_{i1k}, \ldots, \sum_{i=1}^{I} a_{ip}x_{iJk}\right]\mathbf{b}_q \\ &= \sum_{k=1}^{K} c_{kr}\sum_{j=1}^{J}\sum_{i=1}^{I} a_{ip}x_{ijk}b_{jq} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} a_{ip}b_{jq}c_{kr}x_{ijk}.\end{aligned} \tag{20.33}$$

By comparing this with the Tucker3 model (20.2), we can find a kind of *parallel relationship*: the model part in (20.2) is the *sum* of the core elements g_{pqr} *weighted* by loadings a_{ip}, b_{jq}, and c_{kr} over p, q, and r, which approximates the observation

x_{ijk}. On the other hand, (20.33) shows that the solution of core element g_{pqr} is the *sum* of observations x_{ijk} *weighted* by loadings a_{ip}, b_{jq}, and c_{kr} over i, j, and k. This is similar to the fact that the *PC scores* in two-way PCA (Chaps. 5 and 6) are given as the weighted composite of observations.

By substituting (20.31) into the final term of (20.29), we have

$$\begin{aligned}\left\|\mathbf{X}-\mathbf{A}\mathbf{A}'\mathbf{X}(\mathbf{C}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{B})'\right\|^2 &= \|\mathbf{X}\|^2-2\mathrm{tr}\mathbf{X}'\mathbf{A}\mathbf{A}'\mathbf{X}(\mathbf{C}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{B})'\\ &\quad+\mathrm{tr}(\mathbf{C}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{B})'\mathbf{X}'\mathbf{A}\mathbf{A}'\mathbf{A}\mathbf{A}'\mathbf{X}(\mathbf{C}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{B})'\\ &= \|\mathbf{X}\|^2-2\mathrm{tr}\mathbf{A}'\mathbf{X}(\mathbf{C}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{B})'\mathbf{X}'\mathbf{A}\\ &\quad+\mathrm{tr}\mathbf{A}'\mathbf{X}(\mathbf{C}\otimes\mathbf{B})\{(\mathbf{C}\otimes\mathbf{B})'(\mathbf{C}\otimes\mathbf{B})\}(\mathbf{C}\otimes\mathbf{B})'\mathbf{X}'\mathbf{A}\\ &= \|\mathbf{X}\|^2-g(\mathbf{A},\mathbf{B},\mathbf{C}),\end{aligned} \tag{20.34}$$

with

$$g(\mathbf{A},\mathbf{B},\mathbf{C})=\mathrm{tr}\mathbf{A}'\mathbf{X}(\mathbf{C}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{B})'\mathbf{X}'\mathbf{A}=\|\mathbf{A}'\mathbf{X}(\mathbf{C}\otimes\mathbf{B})\|^2=\|\mathbf{G}\|^2. \tag{20.35}$$

Here, we have used the fact that (17.55), (17.56), and (20.30) imply $(\mathbf{C}\otimes\mathbf{B})'(\mathbf{C}\otimes\mathbf{B})=(\mathbf{C}'\otimes\mathbf{B}')(\mathbf{C}\otimes\mathbf{B})=(\mathbf{C}'\mathbf{C})\otimes(\mathbf{B}'\mathbf{B})=\mathbf{I}_{QR}$, and the last identity in (20.35) follows from (20.31). Equation (20.34) shows that its *minimization* over $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ under (20.30) amounts to *maximizing* (20.35) subject to (20.30). The resulting $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ can be substituted in (20.31) to provide the solution of $\mathbf{G}$.

We should note that the last identity in (20.35) is the *sum* of the *squared elements* in the core $\mathbf{G}$. Upon maximizing this, Tucker3 is formulated. This is similar to the fact that two-way PCA can be formulated as the *maximization* of the *variance of PC scores* as described in Chap. 6.

20.8 Tucker3 Algorithm: Iterative Solution

The maximization of (20.35) subject to (20.30) can be attained by alternately iterating steps, in each of which $\mathbf{A}$, $\mathbf{B}$, or $\mathbf{C}$ is optimally updated so that (20.35) is maximized. The steps are described in the next paragraphs.

First, let us consider maximizing (20.35) over $\mathbf{A}$ subject to (20.30), i.e., $\mathbf{A}'\mathbf{A}=\mathbf{I}_P$, with $\mathbf{B}$ and $\mathbf{C}$ kept fixed. It is attained for

$$\mathbf{A}=\mathrm{EV}_{I\times P}\left[\mathbf{X}(\mathbf{C}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{B})'\mathbf{X}'\right]. \tag{20.36}$$

Here, $\mathrm{EV}_{I\times P}[\mathbf{M}]$ is the function which provides the $I \times P$ matrix, whose columns are the *eigenvectors* corresponding to the P largest *eigenvalues* of an $I \times I$ matrix $\mathbf{M}$, with $I \geq P$. The fact that (20.35) is maximized under (20.30) for (20.36) follows from Theorem A.4.4 in Appendix A.4.2.

Next, let us consider maximizing (20.35) over $\mathbf{B}$ subject to $\mathbf{B}'\mathbf{B} = \mathbf{I}_Q$, with $\mathbf{A}$ and $\mathbf{C}$ fixed. For this problem, we use the fact that (20.35) can be rewritten as

$$g(\mathbf{A}, \mathbf{B}, \mathbf{C}) = \mathrm{tr}\mathbf{B}'\mathbf{X}^{\#}(\mathbf{C} \otimes \mathbf{A})(\mathbf{C} \otimes \mathbf{A})'\mathbf{X}^{\#\prime}\mathbf{B} = \left\|\mathbf{B}'\mathbf{X}^{\#}(\mathbf{C} \otimes \mathbf{A})\right\|^2. \tag{20.37}$$

with $\mathbf{X}^{\#}$ defined as (20.20). Function (20.35) or (20.37) is maximized under $\mathbf{B}'\mathbf{B} = \mathbf{I}_Q$ for

$$\mathbf{B} = \mathrm{EV}_{J\times Q}\left[\mathbf{X}^{\#}(\mathbf{C} \otimes \mathbf{A})(\mathbf{C} \otimes \mathbf{A})'\mathbf{X}^{\#\prime}\right], \tag{20.38}$$

as (20.36) is derived from (20.35). The equivalence of (20.35) and (20.37) follows from the fact that $\mathbf{B}'\mathbf{X}^{\#}(\mathbf{C} \otimes \mathbf{A})$ in (20.37) is expanded as

$$\begin{aligned}\mathbf{B}'\mathbf{X}^{\#}(\mathbf{C} \otimes \mathbf{A}) &= \mathbf{B}'\left[\mathbf{X}'_1, \ldots, \mathbf{X}'_K\right]\begin{bmatrix} c_{11}\mathbf{A} & \cdots & c_{1R}\mathbf{A} \\ \vdots & \vdots & \vdots \\ c_{K1}\mathbf{A} & \cdots & c_{KR}\mathbf{A} \end{bmatrix} \\ &= \left[\sum_{k=1}^{K} c_{k1}\mathbf{B}'\mathbf{X}'_k\mathbf{A}, \cdots, \sum_{k=1}^{K} c_{kR}\mathbf{B}'\mathbf{X}'_k\mathbf{A}\right] :\end{aligned} \tag{20.39}$$

each block of the right matrix in (20.39) is merely the transpose of the counterpart in (20.32).

Finally, let us consider maximizing (20.35) over $\mathbf{C}$ subject to $\mathbf{C}'\mathbf{C} = \mathbf{I}_R$, with $\mathbf{A}$ and $\mathbf{B}$ fixed. For this problem, we use the fact that (20.35) can be rewritten as

$$g(\mathbf{A}, \mathbf{B}, \mathbf{C}) = \mathrm{tr}\mathbf{C}'\mathbf{X}^{*}(\mathbf{B} \otimes \mathbf{A})(\mathbf{B} \otimes \mathbf{A})'\mathbf{X}^{*\prime}\mathbf{C} = \left\|\mathbf{C}'\mathbf{X}^{*}(\mathbf{B} \otimes \mathbf{A})\right\|^2. \tag{20.40}$$

with $\mathbf{X}^{*}$ defined as (20.21). Function (20.35) or (20.40) is maximized under $\mathbf{C}'\mathbf{C} = \mathbf{I}_R$ for

$$\mathbf{C} = \mathrm{EV}_{K\times R}\left[\mathbf{X}^{*}(\mathbf{B} \otimes \mathbf{A})(\mathbf{B} \otimes \mathbf{A})'\mathbf{X}^{*\prime}\right] \tag{20.41}$$

as (20.36) is derived from (20.35). The equivalence of (20.40) and (20.35) is shown as follows: by using (17.67) and expressing the $k$tk row of $\mathbf{C}$ as $\tilde{\mathbf{c}}'_k = [c_{k1}, \ldots, c_{kR}]$, we can rewrite $\mathbf{C}'\mathbf{X}^{*}(\mathbf{B} \otimes \mathbf{A})$ in (20.40) as

$$\mathbf{C}'\mathbf{X}^*(\mathbf{B}\otimes\mathbf{A}) = \mathbf{C}'\begin{bmatrix}\mathrm{vec}(\mathbf{X}_1)'\\ \vdots \\ \mathrm{vec}(\mathbf{X}_K)'\end{bmatrix}(\mathbf{B}\otimes\mathbf{A}) = \mathbf{C}'\begin{bmatrix}\mathrm{vec}(\mathbf{X}_1)'(\mathbf{B}\otimes\mathbf{A}'')\\ \vdots \\ \mathrm{vec}(\mathbf{X}_K)'(\mathbf{B}\otimes\mathbf{A}'')\end{bmatrix}$$
$$= \mathbf{C}'\begin{bmatrix}\mathrm{vec}(\mathbf{A}'\mathbf{X}_1\mathbf{B})'\\ \vdots \\ \mathrm{vec}(\mathbf{A}'\mathbf{X}_K\mathbf{B})'\end{bmatrix} = [\tilde{\mathbf{c}}_1\cdots\tilde{\mathbf{c}}_K]\begin{bmatrix}\mathrm{vec}(\mathbf{A}'\mathbf{X}_1\mathbf{B})'\\ \vdots \\ \mathrm{vec}(\mathbf{A}'\mathbf{X}_K\mathbf{B})'\end{bmatrix} = \sum_{k=1}^{K}\tilde{\mathbf{c}}_k\mathrm{vec}(\mathbf{A}'\mathbf{X}_k\mathbf{B})'. \tag{20.42}$$

Here, $\mathbf{A}'\mathbf{X}_k\mathbf{B} = \begin{bmatrix}\mathbf{a}_1'\mathbf{X}_k\mathbf{b}_1 & \cdots & \mathbf{a}_1'\mathbf{X}_k\mathbf{b}_Q\\ \vdots & \vdots & \vdots \\ \mathbf{a}_P'\mathbf{X}_k\mathbf{b}_1 & \cdots & \mathbf{a}_P'\mathbf{X}_k\mathbf{b}_Q\end{bmatrix}$ and $\mathrm{vec}(\mathbf{A}'\mathbf{X}_k\mathbf{B})' = \left[\mathbf{a}_1'\mathbf{X}_k\mathbf{b}_1,\ldots,\mathbf{a}_P'\mathbf{X}_k\mathbf{b}_1,\ldots,\mathbf{a}_1'\mathbf{X}_k\mathbf{b}_Q,\ldots,\mathbf{a}_P'\mathbf{X}_k\mathbf{b}_Q\right]$. Thus, we can find that (20.42) is the $R \times PQ$ matrix whose elements are expressed as (20.33). It implies that (20.40) equals (20.35).

The above facts are sufficient to now describe the Tucker3 algorithm:

Step 1. Initialize **B** and **C**.
Step 2. Update **A** with (20.36)
Step 3. Update **B** with (20.38)
Step 4. Update **C** with (20.41)
Step 5. Obtain (20.31) and finish if convergence is reached; otherwise, go back to Step 2.

Equations (20.34) and (20.35) show that the attained value of the Tucker3 loss function is expressed as $\|\mathbf{X}\|^2(1 - GOF_{\mathrm{T3}})$ with

$$GOF_{\mathrm{T3}} = \frac{g(\mathbf{A},\mathbf{B},\mathbf{C})}{\|\mathbf{X}\|^2} = \frac{\|\mathbf{G}\|^2}{\|\mathbf{X}\|^2}. \tag{20.43}$$

The value of (20.43) with its range [0, 1] expresses the *standardized goodness-of-fit* of the Tucker3 solution and is convenient for checking the convergence. The (20.43) value was 0.72 for the solution in Table 20.3.

20.9 Three-Way Rotation in Tucker3

The *Tucker3* solution is *not uniquely determined*, as shown in the following. Let the solutions of **A**, **B**, **C**, and **G** be transformed into

$$\widetilde{\mathbf{A}} = \mathbf{AS}, \widetilde{\mathbf{B}} = \mathbf{BT}, \widetilde{\mathbf{C}} = \mathbf{CU}, \text{and } \widetilde{\mathbf{G}} = \mathbf{S}^{-1}\mathbf{G}\left(\mathbf{U}^{-1} \otimes \mathbf{T}^{-1}\right)' \tag{20.44}$$

with $\mathbf{S}$ ($P \times P$), $\mathbf{T}$ ($Q \times Q$), and $\mathbf{U}$ ($R \times R$) any *nonsingular matrices*. Even if $\widetilde{\mathbf{A}}$, $\widetilde{\mathbf{B}}$, $\widetilde{\mathbf{C}}$, and $\widetilde{\mathbf{G}}$ transformed according to (20.44) are substituted into $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{G}$ in the loss function (20.29), respectively, it remains unchanged as

$$\begin{aligned}
\left\|\mathbf{X} - \widetilde{\mathbf{A}}\widetilde{\mathbf{G}}(\widetilde{\mathbf{C}} \otimes \widetilde{\mathbf{B}})'\right\|^2 &= \left\|\mathbf{X} - \mathbf{ASS}^{-1}\mathbf{G}\left(\mathbf{U}^{-1} \otimes \mathbf{T}^{-1}\right)'\left\{(\mathbf{CU})' \otimes (\mathbf{BT})'\right\}\right\|^2 \\
&= \left\|\mathbf{X} - \mathbf{AG}\left(\mathbf{U}^{-1\prime} \otimes \mathbf{T}^{-1\prime}\right)\left\{(\mathbf{U}'\mathbf{C}') \otimes (\mathbf{T}'\mathbf{B}')\right\}\right\|^2 \\
&= \left\|\mathbf{X} - \mathbf{AG}\left(\mathbf{U}'^{-1} \otimes \mathbf{T}'^{-1}\right)(\mathbf{U}' \otimes \mathbf{T}')(\mathbf{C}' \otimes \mathbf{B}')\right\|^2 \\
&= \left\|\mathbf{X} - \mathbf{AG}\left\{\left(\mathbf{U}'^{-1}\mathbf{U}'\right) \otimes \left(\mathbf{T}'^{-1}\mathbf{T}'\right)\right\}(\mathbf{C}' \otimes \mathbf{B}')\right\|^2 \\
&= \left\|\mathbf{X} - \mathbf{AG}(\mathbf{C} \otimes \mathbf{B})'\right\|^2,
\end{aligned} \tag{20.45}$$

where (17.55) and (17.56) have been used. Thus, if $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{G}$ give the Tucker3 solution minimizing (20.29) subject to (20.30), (20.44) is also the *solution* that minimizes (20.29) or (20.45), but is not restricted by (20.30). For this reason, we can impose the constraint (20.30) for $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$, *without loss of generality*.

The property that the transformation (20.44) is allowed can be exploited so as to produce interpretable $\widetilde{\mathbf{A}}$, $\widetilde{\mathbf{B}}$, $\widetilde{\mathbf{C}}$, and $\widetilde{\mathbf{G}}$ by choosing appropriate $\mathbf{S}$, $\mathbf{T}$, and $\mathbf{U}$: namely, we can perform a *rotation* procedure for $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{G}$ as described in Chap. 13. However, the rotation required in Tucker 3 differs from that in Chap. 13, in that three rotation matrices $\mathbf{S}$, $\mathbf{T}$, and $\mathbf{U}$ are to be obtained in this chapter.

From here, let $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{G}$ be the Tucker 3 solutions satisfying (20.30) and also $\widetilde{\mathbf{A}}$, $\widetilde{\mathbf{B}}$, and $\widetilde{\mathbf{C}}$ be *constrained* as $\widetilde{\mathbf{A}}'\widetilde{\mathbf{A}} = \mathbf{I}_P$, $\widetilde{\mathbf{B}}'\widetilde{\mathbf{B}} = \mathbf{I}_Q$, and $\widetilde{\mathbf{C}}'\widetilde{\mathbf{C}} = \mathbf{I}_R$. These constraints are equivalent to

$$\mathbf{S}'\mathbf{S} = \mathbf{SS}' = \mathbf{I}_P, \quad \mathbf{T}'\mathbf{T} = \mathbf{TT}' = \mathbf{I}_Q, \quad \mathbf{U}'\mathbf{U} = \mathbf{UU}' = \mathbf{I}_R, \tag{20.46}$$

since $\widetilde{\mathbf{A}}$, $\widetilde{\mathbf{B}}$, and $\widetilde{\mathbf{C}}$ in (20.44) with (20.30) and (20.46) lead to $\widetilde{\mathbf{A}}'\widetilde{\mathbf{A}} = \mathbf{I}_P$, $\widetilde{\mathbf{B}}'\widetilde{\mathbf{B}} = \mathbf{I}_Q$, and $\widetilde{\mathbf{C}}'\widetilde{\mathbf{C}} = \mathbf{I}_R$. The constraint (20.46) simplifies $\widetilde{\mathbf{G}}$ in (20.44) to $\widetilde{\mathbf{G}} = \mathbf{S}'\mathbf{G}(\mathbf{U} \otimes \mathbf{T})$. For obtaining suitable $\widetilde{\mathbf{A}}$, $\widetilde{\mathbf{B}}$, $\widetilde{\mathbf{C}}$ and $\widetilde{\mathbf{G}} = \mathbf{S}'\mathbf{G}(\mathbf{U} \otimes \mathbf{T})$ subject to (20.46), some *three-way rotation* procedures have been proposed (e.g., Kroonenberg, 2008, Chap. 10).

Out of these, we have used orthogonal *three-way simplimax rotation* (Kiers, 1998a) to obtain the solution in Sect. 20.4. In this rotation, the matrices $\mathbf{S}$, $\mathbf{T}$, and $\mathbf{U}$ are obtained that allow the *transformed core matrix* $\widetilde{\mathbf{G}} = \mathbf{S}'\mathbf{G}(\mathbf{U} \otimes \mathbf{T})$ to *approximate* a $P \times RQ$ *target matrix* $\mathbf{G}_{\mathrm{T}}$. Here, this target includes *a number of zero elements*. Thus, the resulting $\widetilde{\mathbf{G}}$ includes a number of the elements *close to zeros*,

which may not be noted. This facilitates interpretation of the core matrix. For this approximation, the simplimax rotation is formulated as minimizing

$$\phi(\mathbf{S}, \mathbf{T}, \mathbf{U}, \mathbf{G}_{\mathrm{T}}) = \left\| \widetilde{\mathbf{G}} - \mathbf{G}_{\mathrm{T}} \right\|^2 = \left\| \mathbf{S}'\mathbf{G}(\mathbf{U} \otimes \mathbf{T}) - \mathbf{G}_{\mathrm{T}} \right\|^2 \tag{20.47}$$

over $\mathbf{S}$, $\mathbf{T}$, $\mathbf{U}$, and $\mathbf{G}_{\mathrm{T}}$ subject to (20.46) and

$$\mathrm{N}_0(\mathbf{G}_{\mathrm{T}}) = U. \tag{20.48}$$

Here, $\mathrm{N}_0(\mathbf{G}_{\mathrm{T}})$ denotes the *number of zero elements* in $\mathbf{G}_{\mathrm{T}}$ and U is a specified integer. An interesting feature of the simplimax rotation is that the target $\mathbf{G}_{\mathrm{T}}$ is also to be estimated: only its number of zero elements $\mathrm{N}_0(\mathbf{G}_{\mathrm{T}})$ is known to be U, and thus the locations of those elements are to be estimated jointly with the values of nonzero parameters. The solution in Sect. 20.4 resulted with U being half of the elements in $\mathbf{G}_{\mathrm{T}}$, i.e., $U = PQR/2 = (3 \times 3 \times 2)/2 = 9$.

The solution of the simplimax rotation can be obtained through alternately iterating the steps, in each of which (20.47) is minimized over one of $\mathbf{S}$, $\mathbf{T}$, $\mathbf{U}$, and $\mathbf{G}_{\mathrm{T}}$ under (20.46) and (20.48) with the remaining three matrices fixed.

First, let us consider minimizing (20.47) over $\mathbf{S}$ subject to (20.46) with $\mathbf{T}$, $\mathbf{U}$, and $\mathbf{G}_{\mathrm{T}}$ fixed. We can find this minimization to be attained by the *orthogonal Procrustes rotation* (13.21) (Chap. 13), since (20.47) is rewritten as $\|\mathbf{G}_{\mathrm{T}}' - \{\mathbf{G}(\mathbf{U} \otimes \mathbf{T})\}'\mathbf{S}\|^2$ which has the same form as the function in (13.21), with the constraint for $\mathbf{S}$ in (20.46) having the same form as that in (13.21). In similar manners, the minimization of (20.47) over $\mathbf{T}$ and the minimization over $\mathbf{U}$ under (20.46) can also be attained by the orthogonal Procrustes rotation (13.21), through rewriting (20.47) suitably, though its details are omitted here.

Now, let us consider the step for minimizing (20.47) over $\mathbf{G}_{\mathrm{T}}$ under (20.48) with $\mathbf{S}$, $\mathbf{T}$, and $\mathbf{U}$ fixed. Using $\widetilde{\mathbf{G}} = (\tilde{g}_{ps})$ and $\mathbf{G}_{\mathrm{T}} = (g_{ps}^{[\mathrm{T}]})$, we can rewrite (20.47) as

$$\phi(\mathbf{G}_T) = \sum_{(p,s)\in\aleph_0} \tilde{g}_{ps}^2 + \sum_{(p,s)\in\aleph_\#} (g_{ps}^{[\mathrm{T}]} - \tilde{g}_{ps})^2 \geq \sum_{(p,s)\in\aleph_0} \tilde{g}_{ps}^2. \tag{20.49}$$

Here, $\aleph_0$ denotes the set of U pairs of (p, s) with elements $g_{ps}^{[\mathrm{T}]}$ to be zero, $\aleph_\#$ is the set of $PQR-U$ pairs of (p, s) with $g_{ps}^{[\mathrm{T}]}$ to be nonzero, $\sum_{(p,s)\in\aleph_0} \tilde{g}_{ps}^2$ stands for the summation of $\tilde{g}_{ps}^2$ over the (p, s) contained in $\aleph_0$, and we have used $\sum_{(p,s)\in\aleph_0} (g_{ps}^{[\mathrm{T}]} - \tilde{g}_{ps})^2 = \sum_{(p,s)\in\aleph_0} (0 - \tilde{g}_{ps})^2 = \sum_{(p,s)\in\aleph_0} \tilde{g}_{ps}^2$. The inequality in (20.49) shows that $\phi(\mathbf{G}_{\mathrm{T}})$ attains its *lower limit* $\sum_{(p,s)\in\aleph_0} \tilde{g}_{ps}^2$ when the element $g_{ps}^{[\mathrm{T}]}$ with $(p, s)\in\aleph_\#$ is set equal to $\tilde{g}_{ps}$ so that $\sum_{(p,s)\in\aleph_\#} (g_{ps}^{[\mathrm{T}]} - \tilde{g}_{ps})^2 = \sum_{(p,s)\in\aleph_\#} (\tilde{g}_{ps} - \tilde{g}_{ps})^2 = 0$. Further, the *limit* $\sum_{(p,s)\in\aleph_0} \tilde{g}_{ps}^2$ is *minimum*, when $\aleph_0$ contains the (p, s) for the U smallest elements among all ones of $\widetilde{\mathbf{G}} \odot \widetilde{\mathbf{G}} = (\tilde{g}_{ps}^2)$,

with $\odot$ standing for the *Hadamard product* defined in (17.69). The optimal $\mathbf{G}_{\mathrm{T}} = (g_{ps}^{[\mathrm{T}]})$ is thus given by

$$g_{ps}^{[\mathrm{T}]} = \begin{cases} 0 & \text{if } \tilde{g}_{ps}^2 \leq \tilde{g}_{[U]}^2 \\ \tilde{g}_{ps} & \text{otherwise} \end{cases} \tag{20.50}$$

with $\tilde{g}_{[U]}^2$ the Uth smallest value among all elements of $\widetilde{\mathbf{G}} \odot \widetilde{\mathbf{G}}$.

The oblique version of the three-way simplimax rotation with (20.46) relaxed is a major topic in Kiers (1998a), though we treat only the orthogonal version here. The target matrix $\mathbf{G}_{\mathrm{T}}$ can be said to be a *sparse* matrix, which will be a keyword in the next two chapters.

20.10 Bibliographical Notes

Chemometricians Smilde et al. (2004) and psychometrician Kroonenberg (2008) have published books in which *multi-way* PCA procedures are reviewed comprehensively. In their description, multi-way PCA includes three-, four-, and five-way PCA as special cases: 3WPCA procedures can be extended straightforwardly to accommodate such cases, as found in Kroonenberg (2008) and Smilde et al. (2004). Adachi (2016) also reviews 3WPCA compactly within one chapter.

Exercises

20.1. For $\mathbf{a}_p = [a_{1p}, \ldots, a_{Ip}]'$, $\mathbf{b}_q = [b_{1q}, \ldots, b_{Jq}]'$, and $\mathbf{c}_r = [c_{1r}, \ldots, c_{Kr}]'$, the *three-way tensor product* is defined as

$$\mathbf{a}_p \circ \mathbf{b}_q \circ \mathbf{c}_r = \{a_{ip} b_{jq} c_{kr}; i = 1, \ldots, I; j = 1, \ldots, J; k = 1, \ldots, K\}: \tag{20.51}$$

the tensor product $\mathbf{a}_p \circ \mathbf{b}_q \circ \mathbf{c}_r$ provides the three-way $I \times J \times K$ array on the right side. Show that the *Tucker3* model (20.2) can be rewritten as

$$\dddot{\mathbf{X}} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} (\mathbf{a}_p \circ \mathbf{b}_q \circ \mathbf{c}_r) g_{pqr} + \dddot{\mathbf{E}} \tag{20.52}$$

with $\dddot{\mathbf{X}}$ the three-way data array defined as (20.1) and $\dddot{\mathbf{E}} = \{e_{ijk}; i = 1, \ldots, I; j = 1, \ldots, J; k = 1, \ldots, K\}$ the three-way array of errors.

20.2. Show that the *Parafac* model (20.4) can be rewritten as

$$\dddot{\mathbf{X}} = \sum_{p=1}^{P} (\mathbf{a}_p \circ \mathbf{b}_p \circ \mathbf{c}_p) + \dddot{\mathbf{E}} \tag{20.53}$$

with $\mathbf{a}_p = \left[a_{1p}, \ldots, a_{Ip}\right]', \mathbf{b}_p = \left[b_{1p}, \ldots, b_{Jp}\right]', \mathbf{c}_p = \left[c_{1p}, \ldots, c_{Kp}\right]'$, and $\ddot{\mathbf{E}}$ defined in Exercise 20.1.

20.3. Show the equivalence of $\mathbf{a}_p \circ \mathbf{b}_q$ and $\mathbf{a}_p \mathbf{b}_q'$: the *vector* product $\mathbf{a}_p \mathbf{b}_q'$ can be regarded as a two-way version of the *tensor* product.

20.4. For $\mathbf{a}_p = [a_{1p}, \ldots, a_{Ip}]'$, $\mathbf{b}_q = [b_{1q}, \ldots, b_{Jq}]'$, $\mathbf{c}_r = [c_{1r}, \ldots, c_{Kr}]'$, and $\mathbf{d}_s = [d_{1s}, \ldots, d_{Ls}]'$, the *four-way tensor product* is defined as

$$\begin{aligned}&\mathbf{a}_p \circ \mathbf{b}_q \circ \mathbf{c}_r \circ \mathbf{d}_s \\ &\quad = \left\{a_{ip} b_{jq} c_{kr} d_{ls}; i = 1, \ldots, I; j = 1, \ldots, J; k = 1, \ldots, K; l = 1, \ldots, L\right\}.\end{aligned} \tag{20.54}$$

Discuss how a *four-way extension of Tucker3* is modeled as

$$\ddddot{\mathbf{X}} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} \sum_{s=1}^{S} \left(\mathbf{a}_p \circ \mathbf{b}_q \circ \mathbf{c}_r \circ \mathbf{d}_s\right) g_{pqrs} + \ddddot{\mathbf{E}} \tag{20.55}$$

for the four-way $I \times J \times K \times L$ data array expressed as $\ddddot{X} = \left\{x_{ijkl}; i = 1, \ldots, I; j = 1, \ldots, J; k = 1, \ldots, K; l = 1, \ldots, L\right\}$, with $\ddddot{E}$ a four-way array of errors.

20.5. By extending (20.54) and (20.55), discuss how Tucker3 can be generalized for a N-way data array.

20.6. Show that the *Parafac* loss function (20.22) can be rewritten as

$$\left\|\mathbf{X_A} - \mathbf{F_{BC}}\mathbf{A}'\right\|^2 = \left\|\mathbf{X_B} - \mathbf{F_{CA}}\mathbf{B}'\right\|^2 = \left\|\mathbf{X_C} - \mathbf{F_{AB}}\mathbf{C}'\right\|^2 \tag{20.56}$$

Here, $\{(j-1)K + k\}$-th rows of $\mathbf{X_A}$ $(JK \times I)$ and $\mathbf{F_{BC}}$ $(JK \times P)$ are $[x_{1jk}, \ldots, x_{Ijk}]$ and $[b_{j1}c_{k1}, \ldots, b_{jP}c_{kP}]$, respectively; the $\{(k-1)I + i\}$-th rows of $\mathbf{X_B}$ $(KI \times J)$ and $\mathbf{F_{CA}}$ $(KI \times P)$ are $[x_{i1k}, \ldots, x_{iJk}]$ and $[c_{k1}a_{i1}, \ldots, c_{kP}a_{iP}]$, respectively; the $\{(i-1)J + j\}$-th rows of $\mathbf{X_C}$ $(IJ \times K)$ and $\mathbf{F_{AB}}$ $(IJ \times P)$ are $[x_{ij1}, \ldots, x_{ijK}]$ and $[a_{i1}b_{j1}, \ldots, a_{iP}b_{jP}]$, respectively (Adachi, 2013b).

20.7. Show a *Parafac* algorithm for minimizing (20.56).

20.8. Show that the *Parafac* model (20.14) with (20.15) can be rewritten as

$$\mathrm{vec}(\mathbf{X}_k) = \mathbf{U}\mathbf{d}_k + \mathrm{vec}(\mathbf{E}_k) \tag{20.57}$$

using $\mathbf{U} = [\mathbf{b}_1 \otimes \mathbf{a}_1, \ldots, \mathbf{b}_P \otimes \mathbf{a}_P]\,(JI \times P)$ and $\mathbf{d}_k = \mathbf{D}_k \mathbf{1}_P$ $(P \times 1)$, with $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_P]$ and $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_P]$ (ten Berge, 1993). Hints can be found in (17.65), (17.66), and the fact that (20.14) is rewritten as $\mathbf{X}_k = \sum_{p=1}^{P} c_{kp}\mathbf{a}_p\mathbf{b}_p' + \mathbf{E}_k$, or equivalently, $\mathrm{vec}(\mathbf{X}_k) = \mathrm{vec}\left(\sum_{p=1}^{P} c_{kp}\mathbf{a}_p\mathbf{b}_p'\right) + \mathrm{vec}(\mathbf{E}_k)$.

20.9. Show that the *Tucker3* loss function (20.29) can be rewritten as

$$\left\|\mathbf{X}-\mathbf{A}\mathbf{G}(\mathbf{C}\otimes\mathbf{B})'\right\|^2=\left\|\mathbf{X}^{\#}-\mathbf{B}\mathbf{G}^{\#}(\mathbf{C}\otimes\mathbf{A})'\right\|^2=\left\|\mathbf{X}^{*}-\mathbf{C}\mathbf{G}^{*}(\mathbf{B}\otimes\mathbf{A})'\right\|^2 \tag{20.58}$$

with (20.20) and (20.21). Here, $\mathbf{G}^{\#}=[\mathbf{G}_1',\ldots,\mathbf{G}_R']$ $(Q\times RP)$, and $\mathbf{G}^{*}=$

$$\begin{bmatrix}\mathrm{vec}(\mathbf{G}_1)'\\ \vdots\\ \mathrm{vec}(\mathbf{G}_R)'\end{bmatrix}=\left[\mathbf{G}_1^{*},\ldots,\mathbf{G}_q^{*},\ldots,\mathbf{G}_Q^{*}\right](R\times QP)$$ with $\mathbf{G}_q^{*}$ the $R\times P$ matrix whose (r, p) element is g_{pqr}.

20.10. The *indeterminacy of the Tucker3 solution* is shown by (20.45) with the transformation (20.44). Show that the indeterminacy can be also shown by

$$\left\|\mathbf{X}^{\#}-\mathbf{B}\mathbf{G}^{\#}(\mathbf{C}\otimes\mathbf{A})'\right\|^2=\left\|\mathbf{X}^{\#}-\widetilde{\mathbf{B}}\widetilde{\mathbf{G}}^{\#}(\widetilde{\mathbf{C}}\otimes\widetilde{\mathbf{A}})'\right\|^2, \tag{20.59}$$

$$\left\|\mathbf{X}^{*}-\mathbf{C}\mathbf{G}^{*}(\mathbf{B}\otimes\mathbf{A})'\right\|^2=\left\|\mathbf{X}^{*}-\widetilde{\mathbf{C}}\widetilde{\mathbf{G}}^{*}(\widetilde{\mathbf{B}}\otimes\widetilde{\mathbf{A}})'\right\|^2, \tag{20.60}$$

on the basis of (20.58). Here, $\widetilde{\mathbf{A}}=\mathbf{A}\mathbf{S},\widetilde{\mathbf{B}}=\mathbf{B}\mathbf{T},\widetilde{\mathbf{C}}=\mathbf{C}\mathbf{U},\widetilde{\mathbf{G}}^{\#}=\mathbf{T}^{-1}\,G^{\#}(\mathbf{U}^{-1}\otimes\mathbf{S}^{-1})'$ and $\widetilde{\mathbf{G}}^{*}=\mathbf{U}^{-1}\mathbf{G}^{*}(\mathbf{T}^{-1}\otimes\mathbf{S}^{-1})'$.

20.11. Kiers (1998b) has proposed a *three-way rotation* technique for Tucker 3 alternative to the procedure in Sect. 20.9. In this method, the function

$$\begin{aligned}\eta(\mathbf{S},\mathbf{T},\mathbf{U})=w_1\mathrm{Simp}\left(\widetilde{\mathbf{G}}'\right)+w_2\mathrm{Simp}\left(\widetilde{\mathbf{G}}^{\#\prime}\right)+w_3\mathrm{Simp}\left(\widetilde{\mathbf{G}}^{*\prime}\right)\\ +w_4\mathrm{Simp}(\widetilde{\mathbf{A}})+w_5\mathrm{Simp}(\widetilde{\mathbf{B}})+w_6\mathrm{Simp}(\widetilde{\mathbf{C}})\end{aligned} \tag{20.61}$$

is maximized over $\mathbf{S}$, $\mathbf{T}$, and $\mathbf{U}$ under (20.46), for the Tucker3 solution subject to (20.30). Here, $\widetilde{\mathbf{A}}$, $\widetilde{\mathbf{B}}$, $\widetilde{\mathbf{C}}$, $\widetilde{\mathbf{G}}$, $\widetilde{\mathbf{G}}^{\#}$, and $\widetilde{\mathbf{G}}^{*}$ are those in (20.44), (20.59), and (20.60), with $\widetilde{\mathbf{G}}$, $\widetilde{\mathbf{G}}^{\#}$, and $\widetilde{\mathbf{G}}^{*}$ simplified to $\widetilde{\mathbf{G}}=\mathbf{S}'\mathbf{G}(\mathbf{U}\otimes\mathbf{T})$, $\widetilde{\mathbf{G}}^{\#}=\mathbf{T}'\mathbf{G}^{\#}(\mathbf{U}\otimes\mathbf{S})$, and $\widetilde{\mathbf{G}}^{*}=\mathbf{U}'\mathbf{G}^{*}(\mathbf{T}\otimes\mathbf{S})$ since of (20.46). The scalars $w_1,\ldots,w_6$ in (20.61) are nonnegative weights to be prespecified, and the function Simp(•) in (20.61) is the *varimax rotation* function in (13.12), defined as

$$\mathrm{Simp}(\mathbf{V})=\frac{1}{L}\sum_{m=1}^{M}\sum_{l=1}^{L}\left(v_{lm}^2-\frac{1}{L}\sum_{n=1}^{L}v_{nm}^2\right)^2$$

for an $L \times M$ matrix $\mathbf{V} = (v_{lm})$. The maximization of (20.61) over $\mathbf{S}$, $\mathbf{T}$, and $\mathbf{U}$ subject to (20.46) can be called *three-way varimax rotation*. Discuss the differences of this rotation to the *three-way simplimax rotation* in Sect. 20.9.

20.12. Show how the *Tucker2* modeled in (20.8) can be formulated as the minimization of

$$f(\mathbf{A}, \mathbf{B}, \mathbf{H}) = \left\|\mathbf{X} - \mathbf{AH}(\mathbf{I}_K \otimes \mathbf{B})'\right\|^2 \tag{20.62}$$

over $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{H} = [\mathbf{H}_1, \ldots, \mathbf{H}_K]$ $(P \times KQ)$.

Chapter 21
Sparse Regression Analysis

A matrix or vector is said to be sparse when it includes a number of zero elements. Hence, the term *sparse estimation* refers to estimating a number of parameters as zeros. The developments in multivariate analysis procedures with sparse estimation started from modifications to the *multiple regression analysis* introduced in Chap. 4. A number of modified regression procedures have been developed so that the resulting regression coefficient vector is sparse, and can be generally referred to as *sparse regression analysis*. Among those procedures, the first was proposed by Tibshirani (1996) and called *lasso*. One of the main purposes of sparse regression analysis can be regarded as *removing useless variables computationally* in order to select useful ones: The explanatory variables, whose coefficients are estimated as zeros, are removed from a set of variables to determine a dependent variable.

21.1 Illustration of Sparse Solution

Let us recall the *regression analysis* that was presented in Chap. 4. It is modeled as

$$\underset{\mathbf{y}}{\begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix}} = \underset{\mathbf{X}}{\begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix}} \underset{\mathbf{b}}{\begin{bmatrix} b_1 \\ \vdots \\ b_j \\ \vdots \\ b_p \end{bmatrix}} + c \underset{\mathbf{1}_n}{\begin{bmatrix} 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{bmatrix}} + \underset{\mathbf{e}}{\begin{bmatrix} e_1 \\ \vdots \\ e_i \\ \vdots \\ e_n \end{bmatrix}}, \tag{21.1}$$

K. Adachi, *Matrix-Based Introduction to Multivariate Data Analysis*,
https://doi.org/10.1007/978-981-15-4103-2_21

where the squared norm of the error vector **e**, i.e.,

$$f(\mathbf{b}, c) = \|\mathbf{e}\|^2 = \|\mathbf{y} - \mathbf{X}\mathbf{b} - c\mathbf{1}_n\|^2, \tag{21.2}$$

is minimized over the coefficient vector **b** and intercept c. The solution of c must satisfy (4.9). Substituting this into c in (21.2), this is simplified to

$$f(\mathbf{b}) = \|\mathbf{e}\|^2 = \|\mathbf{J}\mathbf{y} - \mathbf{J}\mathbf{X}\mathbf{b}\|^2, \tag{21.3}$$

with $\mathbf{J} = \mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n'$ the centering matrix. The solution of **b** minimizing (21.3) is given by (4.12), and using this in (4.9) leads to the solution of c. In this section, we call this procedure *standard regression*, to distinguish it from *sparse regression* introduced in this chapter. Furthermore, we refer to the function (21.3) without c, rather than (21.2) including c, as the *standard regression function*, as the sparse regression procedures differ from standard regression solely in that another function (shown in the later sections) is added to (21.3). The solution c is also given by using the resulting **b** in (4.9) in sparse regression.

To illustrate what solutions are provided by sparse regression, we use an n (252 persons) $\times$ p (14 variables) matrix of the standard scores for body fat data, cited from the website https://astro.temple.edu/∼alan/MMST/datasets.html in Izenman's (2008) textbook. Here, the variables consist of the 13 physical features shown in the left column of Table 21.1 and a fat index. For the data set, we performed standard and sparse regression, with the 13 physical features treated as explanatory variables (**X**) for predicting the fat index (**y**). In the table, the sparse regression procedures have been labeled lasso and L_0, which are explained later. The solutions are shown in Table 21.1. Here, blank cells indicate estimates of zero, demonstrating the *sparse coefficient vectors* resulting in the sparse regression. For example, the seven coefficients b_2, b_5, b_7, …, b_{11} are estimated as zeros in lasso (Table 21.1B): Its solution leads to the equation explaining fatness:

$$\begin{aligned}\text{fat} = {} & 0.091\,\text{age} - 0.088\,\text{height} - 0.109\,\text{neck} \\ & + 0.942\,\text{abdomen} + 0.073\,\text{forearm} - 0.168\,\text{wrist} + \textit{error}.\end{aligned}$$

Here, the seven variables whose coefficients were estimated as zero have vanished.

The bottom row in Table 21.1 shows the values of the BIC (8.25). How BIC is incorporated in regression analysis is explained in Sect. 21.4. As described in Sect. 8.7, a model with a smaller BIC value is considered as better: Table 21.1 shows that the lasso solution is the best among the three ones.

In regression analysis, it is often inevitable to select a subset of the explanatory variables useful for predicting a dependent variable among the whole set. The selection can be restated as removing the explanatory variables that are useless for the prediction. This selection or removal is referred to as *variable selection*. Sparse

Table 21.1 Solutions of the standard and sparse regression for bodyfat data

Procedure		(A) Standard	(B) Lasso	(C) L_0
Age	b_1	0.097	0.091	0.093
Weight	b_2	−0.342		−0.339
Height	b_3	−0.013	−0.088	
Neck	b_4	−0.138	−0.109	−0.144
Chest	b_5	−0.017		
Abdomen	b_6	1.231	0.942	1.226
Hip	b_7	−0.161		−0.160
Thigh	b_8	0.156		0.159
Knee	b_9	0.004		
Ankle	b_{10}	0.036		
Biceps	b_{11}	0.066		0.065
Forearm	b_{12}	0.110	0.073	0.109
Wrist	b_{13}	−0.184	−0.168	−0.172
BIC		1128.5	1104.4	1107.1
Weight			$w = 7.9$	$w = 0.56$

regression procedures jointly perform both variable selection and parameter estimation. They can be expressed as *estimating*

[1] what variables are to be *excluded* with their *coefficients* as *zero*
[2] the *values* of *nonzero coefficients*.

simultaneously and *optimally*.

21.2 Penalized Least Squares Method and Lasso

In sparse regression, the simultaneous estimation of [1] and [2] above is attained by a *penalized* least squares method. This term, which is also called a *regularized* least squares method, generally refers to minimizing the composite of a least squares function and an *additional function*. In sparse regression, the least square function is the *standard regression function* (21.3), while the latter is a *penalty function* which penalizes nonzero elements in $\mathbf{b}$. Using Pen($\mathbf{b}$) for the penalty function, sparse regression can be formulated as minimizing

$$f_{\mathrm{PEN}}(\mathbf{b}) = \|\mathbf{Jy} - \mathbf{JXb}\|^2 + nw\mathrm{Pen}(\mathbf{b}) \tag{21.4}$$

over $\mathbf{b}$ for a given $w \geq 0$. Here, wPen($\mathbf{b}$) has been multiplied by n merely for the sake of the convenience during the subsequent derivations of equations. The role of w is to tune the two functions $\|\mathbf{Jy} - \mathbf{JXb}\|^2$ and Pen($\mathbf{b}$). A parameter w, which tunes multiple functions, is referred to as a *tuning parameter*. It can also be called a *penalty weight*, since it determines the importance (or weight) of Pen($\mathbf{b}$) relative to $\|\mathbf{Jy} - \mathbf{JXb}\|^2$.

The procedures used in sparse regression can be distinguished according to what functions are used for Pen(**b**). Among them, Tibshirani's (1996) method called *lasso* was the first proposed and best known. In lasso,

$$\text{Pen}(\mathbf{b}) = \|\mathbf{b}\|_1 = \sum_{j=1}^{p} |b_j| : \tag{21.5}$$

Pen(**b**) is called the L_1 norm as explained next:

Note 21.1. L_q Norm
As found in (21.5), $\|\mathbf{b}\|_1$ stands for the *sum of the absolute values* of the elements. This sum is called the L_1 norm of **b**. More generally, the L_q norm is defined as

$$\|\mathbf{A}\|_q = \left(\sum_{i=1}^{n} \sum_{j=1}^{p} |a_{ij}|^q \right)^{1/q} \quad (q > 0) \tag{21.6}$$

for $n \times p$ $\mathbf{A} = (a_{ij})$. According to this terminology and notation, the norm $\|\mathbf{A}\|$ used so far is called the L_2 *norm* and must be replaced by $\|\mathbf{A}\|_2$. However, we continue to use $\|\mathbf{A}\|$ for the L_2 norm and $\|\mathbf{A}\|^2$ for the squared L_2 norm.

Using (21.5) in (21.4), we have the *lasso loss function*

$$f_{L1}(\mathbf{b}) = \|\mathbf{Jy} - \mathbf{JXb}\|^2 + nw\|\mathbf{b}\|_1 \tag{21.7}$$

Why this minimization leads to the sparse **b** in Table 21.1B is described in the next section.

The name *lasso* originates from the abbreviation of *least absolute selection and shrinkage operator*. Here, "*least absolute*" and "*selection*" stand for (21.5) being based on absolute values and usable for variable selection, while the reference to "*shrinkage*" will be mentioned in Sect. 21.5.

21.3 Coordinate Descent Algorithm for Lasso

Of the algorithms for minimizing (21.7), we introduce the *alternate least squares* (*ALS*) approach, which is also called a *coordinate descent algorithm* in some sparse regression literature (Hastie, Tibshirani, & Wainwright, 2015). In the algorithm, a

procedure is alternately iterated to optimally update *each of the coefficients* b_j ($j = 1, \ldots, p$). How the optimal b_j is obtained is explained in the next paragraphs.

We can rewrite (21.3) as

$$f(\mathbf{b}) = \left\|\mathbf{J}\mathbf{y} - \mathbf{J}\left(b_1\mathbf{x}_1 + \ldots + b_p\mathbf{x}_p\right)\right\|^2 = \left\|\mathbf{r}_{(j)} - b_j\mathbf{J}\mathbf{x}_j\right\|^2. \tag{21.8}$$

Here, $\mathbf{X} = \left[\mathbf{x}_1, \ldots, \mathbf{x}_p\right]$, $\mathbf{b} = \left[b_1, \ldots, b_p\right]'$, and

$$\mathbf{r}_{(j)} = \mathbf{J}\mathbf{y} - \sum_{k \neq j} b_k \mathbf{J}\mathbf{x}_k. \tag{21.9}$$

with $\Sigma_{k \neq j} b_k \mathbf{J}\mathbf{x}_k$ standing for the sum of $b_k\mathbf{J}\mathbf{x}_k$ over all $k = 1, \ldots, p$ except j. Using (21.5) and (21.8), the lasso loss function (21.7) is rewritten as

$$\begin{aligned} f_{L1}(\mathbf{b}) &= \left\|\mathbf{r}_{(j)} - b_j\mathbf{J}\mathbf{x}_j\right\|^2 + nw\left|b_j\right| + nw\sum_{k \neq j}\left|b_k\right| \\ &= \left\|\mathbf{r}(j)\right\|^2 + \left(\mathbf{x}_j'\mathbf{J}\mathbf{x}_j\right)b_j^2 - 2\left(\mathbf{r}_{(j)}'\mathbf{J}\mathbf{x}_j\right)b_j + nw\left|b_j\right| + nw\sum_{k \neq j}\left|b_k\right| \\ &= \left\|\mathbf{r}_{(j)}\right\|^2 + \left(\mathbf{x}_j'\mathbf{J}\mathbf{x}_j\right)g_j(b_j) + nw\sum_{k \neq j}\left|b_k\right|, \end{aligned} \tag{21.10}$$

On the right side of (21.10), only $g_j(b_j)$ is a function of b_j, which is expressed as

$$\begin{aligned} g_j(b_j) &= b_j^2 - 2\frac{\mathbf{r}_{(j)}'\mathbf{J}\mathbf{x}_j}{\mathbf{x}_j'\mathbf{J}\mathbf{x}_j}b_j + \frac{nw}{\mathbf{x}_j'\mathbf{J}\mathbf{x}_j}\left|b_j\right| \\ &= b_j^2 - 2\frac{\mathbf{r}_{(j)}'\mathbf{J}\mathbf{x}_j}{nv_{jj}}b_j + \frac{w}{v_{jj}}\left|b_j\right| = b_j^2 - 2r_j\left(\mathbf{b}_{[j]}\right)b_j + 2d_j(w)\left|b_j\right|. \end{aligned} \tag{21.11}$$

Here, $v_{jj} = n^{-1}\mathbf{x}_j'\mathbf{J}\mathbf{x}_j$ is the variance of the jth explanatory variable in $\mathbf{x}_j$, $\mathbf{b}_{[j]}$ is the $(p - 1) \times 1$ vector consisting of $b_1, \ldots, b_p$ except b_j,

$$r_j\left(\mathbf{b}_{[j]}\right) = \frac{\mathbf{r}_{(j)}'\mathbf{J}\mathbf{x}_j}{nv_{jj}} = \frac{\mathbf{y}'\mathbf{J}\mathbf{x}_j}{nv_{jj}} - \frac{\sum_{k \neq j} b_k\mathbf{x}_k'\mathbf{J}\mathbf{x}_j}{nv_{jj}} = \frac{v_j^{[y]}}{v_{jj}} - \frac{\sum_{k \neq j} b_k v_{jk}}{v_{jj}}, \tag{21.12}$$

$$d_j(w) = \frac{w}{2v_{jj}} \geq 0, \tag{21.13}$$

with $v_{jk} = n^{-1}\mathbf{x}_k'\mathbf{J}\mathbf{x}_j = n^{-1}\mathbf{x}_j'\mathbf{J}\mathbf{x}_k$ and $v_j^{[y]} = n^{-1}\mathbf{y}'\mathbf{J}\mathbf{x}_j$ being the covariances of the jth explanatory variable to the kth one and the dependent variable, respectively. The inequality in (21.13) follows from $w \geq 0$ and $v_{jj} > 0$. The notations $r_j(\mathbf{b}_{[j]})$ and $d_j(w)$ are used in (21.11)–(21.13) in order to indicate that they are functions of $\mathbf{b}_{[j]}$ and w.

The coefficient b_j minimizing (21.11) is the optimal one that minimizes the lasso loss function (21.7) with the other coefficients b_k ($k \neq j$) fixed. The minimizer of (21.11) is given by

$$b_j = \begin{cases} 0 & \text{if } \left|r_j(\mathbf{b}_{[j]})\right| \leq d_j(w) \\ \text{sign}[r_j(\mathbf{b}_{[j]})]\left\{\left|r_j(\mathbf{b}_{[j]})\right| - d_j(w)\right\} & \text{otherwise} \end{cases}, \tag{21.14}$$

as explained in Note 21.2 below. Here, sign[t] expresses the sign of a scalar t: sign[t] = 1 for $t > 0$, sign[t] = 0 for $t = 0$, and sign[t] = −1 for $t < 0$. In (21.14), we find that the solution of b_j is exactly zero for $\left|r_j(\mathbf{b}_{[j]})\right| \leq d_j(w)$.

We can obtain the optimal $\mathbf{b}$ through alternate iteration of updating b_j with (21.14) for $j = 1, \ldots, p$. Thus, the *algorithm for lasso* can be summarized as follows:

Step 1. Initialize $\mathbf{b}$
Step 2. For each of $j = 1, \ldots, p$, perform the following: Obtain $r_j(\mathbf{b}_{[j]})$ using the current b_k values in (21.12) to update b_j with (21.14).
Step 3. Finish if convergence is reached; otherwise, go back to Step 2.

Here, we should note that the current $r_j(\mathbf{b}_{[j]})$ value must be obtained before the update of b_j in Step 2, since $r_j(\mathbf{b}_{[j]})$ in (21.14) is a function of the coefficients $b_1, \ldots, b_p$ except b_j. The initialization in Step 1 is made by setting $\mathbf{b}$ to the standard regression solution for the computations in this chapter.

Note 21.2. Minimizing a Quadratic Function plus L_1 Norm
For the sake of simplicity, we omit $(\mathbf{b}_{[j]})$, (w), and the subscript j from the symbols in (21.11). Hence, (21.11) is simplified to $g(b) = b^2 - 2rb + 2d|b|$, which can be rewritten as

$$g(b) = b^2 - 2rb + 2db \quad \text{for } b \geq 0, \tag{21.15}$$

$$g(b) = b^2 - 2rb - 2db \quad \text{for } b < 0. \tag{21.16}$$

Here, the inequality in (21.13), i.e., $d \geq 0$, should be kept in mind. The *shape* of function $g(b)$ and the *solution* of b *depend* on *which inequality holds*, $r > d$ (≥ 0), $r < -d$ (≤ 0), or $-d \leq r \leq d$. This is illustrated in Fig. 21.1, where we can see that the solution of b is zero if $-d \leq r \leq d$, but not be zero, otherwise. This is shown by formulas in the next paragraphs.

First, let us consider the cases with $-d \leq r \leq d$. Then, we can rewrite (21.15) and (21.16) to find the following inequalities:

$$g(b) = b^2 + 2(d - r)b \geq 0 \quad \text{for } b \geq 0, \tag{21.17}$$

$$g(b) = b^2 + 2(-d - r)b \geq 0 \quad \text{for } b < 0. \tag{21.18}$$

Here, the inequality (21.17) follows from the fact that $r \leq d$ and $b \geq 0$ imply $(d - r)b \geq 0$, while the one in (21.18) follows from the fact that $-d \leq r$, i.e., $-d - r \leq 0$, and $b < 0$ imply $(-d - r)b \geq 0$. That is, (21.17) and (21.18) show $g(b) \geq 0$ for any b, and the lower bound 0 is attained for

$$b = 0 \tag{21.19}$$

which is the solution for $-d \leq r \leq d$. This is illustrated in Fig. 21.1b.

Next, let us consider the cases with $r > d \geq 0$. Here, the inequality in (21.18) holds true, since $(-d - r)b > 0$ for $b < 0$. However, the inequality in (21.17) does not hold in general, since $r > d \geq 0$ and $b \geq 0$ imply $(d - r)b \leq 0$. Further, we can rewrite $g(b)$ in (21.17) as

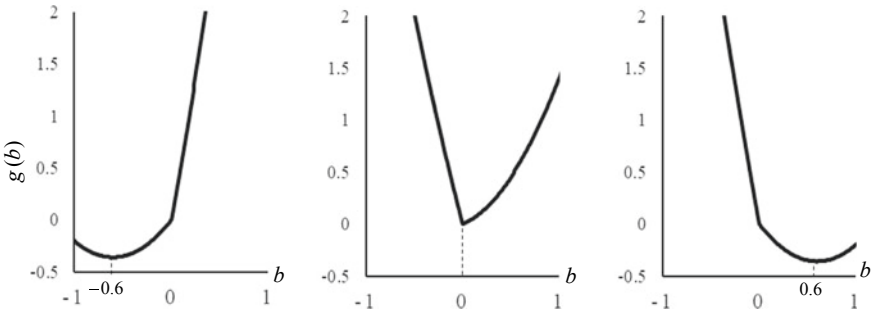
$$g(b) = b^2 - 2(r-d)b = \{b - (r-d)\}^2 - (r-d)^2 \geq -(r-d)^2 \quad \text{for } b \geq 0 \tag{21.20}$$

with $-(r - d)^2 \leq 0$. This implies that the lower bound of $g(b)$ is $-(r - d)^2$. This can be attained for

$$b = r - d, \tag{21.21}$$

which equals sign$[r](|r| - d)$ in (21.14) since of $r > d \geq 0$. It is illustrated how (21.21) is the solution for $r > d$ in Fig. 21.1c, where we can find that the minimum is attained for (21.21), i.e., $b = 1.6 - 1 = 0.6$.

Finally, let us consider the case of $r < -d \leq 0$. Since this implies $r \leq 0$ and thus $d - r \geq 0$, the inequality in (21.17) holds. However, the one in (21.18) does not hold in general, since $r < -d$, i.e., $-d - r > 0$, and $b < 0$ imply $(-d - r)b < 0$. Further, we can rewrite $g(b)$ in (21.18) as



(a) $r = -1.6 < -d = -1$ (b) $-d = -1 < r = 0.8 < d = 1$ (c) $d = 1 < r = 1.6$

Fig. 21.1 $g(b) = b^2 - 2rb + 2d|b|$ versus b (horizonal axis) for $d = 1$ and $r = -1.6, 0.8, 1.6$

$$g(b) = b^2 - 2(r+d)b = \{b - (r+d)\}^2 - (r+d)^2 \geq -(r+d)^2 \quad \text{for } b < 0, \tag{21.22}$$

implying that the lower bound of $g(b)$ is $-(r + d)^2$. This can be attained for

$$b = r + d, \tag{21.23}$$

whose equivalence to sign[r]($|r| - d$) in (21.14) follows from the fact that $r \leq 0$ implies sign[r] $\times$ ($|r| - d$) = $-|r| + d = r + d$. It is illustrated that (21.23) is the solution for $r < -d$ in Fig. 21.1a, where we can see that the minimum is attained for (21.23), i.e., $b = -1.6 + 1 = -0.6$.

21.4 Selection of Penalty Weight

The lasso algorithm in the last section provides the optimal **b** for a given *penalty weight w*: The resulting **b** *depends* on the w value. Thus, the lasso algorithm is *run multiple times* for some w values, which provides multiple solutions of **b**. Among these, the *best* **b** is *selected*. For the selection, we can use *information criteria* such as *AIC* and *BIC* introduced in Chap. 8: **b** and the corresponding BIC value are obtained for each w value, and the *solution* of **b** with the *least* BIC can be regarded as the *best* one. Here, BIC may be replaced by AIC. In the remaining parts of this subsection, we describe how the information criteria for lasso are derived and defined.

Sparse regression procedures including lasso are formulated as minimizing (21.4), which is included in the *penalized least squares method*. It differs from the *maximum likelihood* (*ML*) *method* which leads to the information criteria described in Chap. 8. However, they can be defined for (21.4), since its *minimization* can be *reformulated* as an *ML* problem. This fact is shown through the *two kinds of equivalence*, as explained in the next paragraphs.

First, it is known that the minimization of (21.4) is *equivalent* to minimizing the *least squares* (*LS*) *function* (21.2) subjected to the *inequality* constraint

$$\text{Pen}(\mathbf{b}) \leq u, \tag{21.24}$$

where the positive scalar u can be associated with the penalty weight w (Tibshirani, 1996). However, it is beyond the scope of this book to prove the equivalence and show the relationship of u to w.

Next, the *LS* problem of minimizing (21.2) is *equivalent* to the *ML* problem *of maximizing the log likelihood* derived from the model (21.1) with the supposition that **e** has the following multivariate *normal distribution*:

$$\mathbf{e} \sim N_n\left(\mathbf{0}_n, \sigma^2 \mathbf{I}_n\right). \tag{21.25}$$

The equivalence is explained next:

Note 21.3. When Maximum Likelihood and Least Squares Methods are Equivalent

Let an $n \times 1$ data vector $\mathbf{y}$ be modeled as

$$\mathbf{y} = \boldsymbol{\phi}(\boldsymbol{\theta}) + \mathbf{e}. \tag{21.26}$$

Here, $\mathbf{e}$ is an $n \times 1$ error vector and $\boldsymbol{\phi}(\boldsymbol{\theta})$ is an $n \times 1$ vector which is a function of the parameters contained in vector $\boldsymbol{\theta}$. The regression model (21.1) is a special case of (21.26) in which $\boldsymbol{\phi}(\boldsymbol{\theta}) = \mathbf{X}\mathbf{b} + c\mathbf{1}_n$ with $\boldsymbol{\theta} = [\mathbf{b}', c]'$. The following discussions hold for any model that can be expressed as (21.26).

We suppose (21.25). Then, $\mathbf{y} \sim N_n(\boldsymbol{\phi}(\boldsymbol{\theta}), \sigma^2\mathbf{I}_n)$, whose probability density function is

$$\begin{aligned} P\left(\mathbf{y}|\boldsymbol{\theta}, \sigma^2\right) &= \frac{1}{(2\pi)^{n/2}\left|\sigma^2\mathbf{I}_n\right|^{1/2}} \exp\left\{-\frac{1}{2}[\mathbf{y} - \boldsymbol{\phi}(\boldsymbol{\theta})]'\left(\sigma^2\mathbf{I}_n\right)^{-1}[\mathbf{y} - \boldsymbol{\phi}(\boldsymbol{\theta})]\right\} \\ &= \frac{1}{(2\pi)^{n/2}\left(\sigma^{2n}\right)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}\|\mathbf{y} - \boldsymbol{\phi}(\boldsymbol{\theta})\|^2\right\}. \\ &= (2\pi)^{-n/2}\left(\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\|\mathbf{y} - \boldsymbol{\phi}(\boldsymbol{\theta})\|^2\right\}. \end{aligned} \tag{21.27}$$

Here, we have used (8.10) with the fact that the *singular value decomposition* of $\sigma^2\mathbf{I}_n$ is expressed as $\sigma^2\mathbf{I}_n = \mathbf{I}_n(\sigma^2\mathbf{I}_n)\mathbf{I}_n'$: The determinant $|\sigma^2\mathbf{I}_n|$ is given by the nth power of σ^2. The logarithm of (21.27) gives the log likelihood

$$l\left(\boldsymbol{\theta}, \sigma^2\right) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\|\mathbf{y} - \boldsymbol{\phi}(\boldsymbol{\theta})\|^2. \tag{21.28}$$

The partial derivative of (21.28) with respect to σ^2 is known to be given by $\eta(\sigma^2) = \partial l(\boldsymbol{\theta}, \sigma^2)/\partial\sigma^2 = -n/(2\sigma^2) + \|\mathbf{y} - \boldsymbol{\phi}(\boldsymbol{\theta})\|^2/(2\sigma^4)$ $= [n/(2\sigma^2)]\left(n^{-1}\|\mathbf{y} - \boldsymbol{\phi}(\boldsymbol{\theta})\|^2/\sigma^2 - 1\right)$. See Appendix A.6.3 for partial derivative. We can find $\eta(\sigma^2) = 0$ for

$$\sigma^2 = \frac{1}{n}\|\mathbf{y} - \boldsymbol{\phi}(\boldsymbol{\theta})\|^2, \tag{21.29}$$

with $\eta(\sigma^2) > 0$ for $\sigma^2 < n^{-1}\|\mathbf{y} - \boldsymbol{\phi}(\boldsymbol{\theta})\|^2$ and $\eta(\sigma^2) < 0$ for $\sigma^2 > n^{-1}\|\mathbf{y} - \boldsymbol{\phi}(\boldsymbol{\theta})\|^2$. This shows that (21.28) is maximum when (21.29) for a given $\boldsymbol{\theta}$ and the solution of σ^2 must satisfy (21.29).

Substituting (21.29) into (21.28), we have

$$\begin{aligned} l(\boldsymbol{\theta}) &= -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\left(\frac{1}{n}\|\mathbf{y} - \boldsymbol{\phi}(\boldsymbol{\theta})\|^2\right) - \frac{n}{2} \\ &= -\frac{n}{2}\log\|\mathbf{y} - \boldsymbol{\phi}(\boldsymbol{\theta})\|^2 + const \end{aligned} \tag{21.30}$$

with the term *const* not depending on $\boldsymbol{\theta}$. Here, we find that the *maximization* of (21.30) over $\boldsymbol{\theta}$ is equivalent to *minimizing* the *least squares function* $\|\mathbf{y} - \boldsymbol{\phi}(\boldsymbol{\theta})\|^2$ over $\boldsymbol{\theta}$.

Now, let $\boldsymbol{\phi}(\boldsymbol{\theta})$ in (21.26) be the *regression function* $\boldsymbol{\phi}(\boldsymbol{\theta}) = \mathbf{X}\mathbf{b} + c\mathbf{1}_n$ with $\boldsymbol{\theta} = [\mathbf{b}', c]'$. Substituting this in (21.30) leads to $-(n/2)\log\|\mathbf{y} - \mathbf{X}\mathbf{b} - c\mathbf{1}_n\|^2 + const$. Its maximization is equivalent to minimizing (21.2), i.e., $\|\mathbf{y} - \mathbf{X}\mathbf{b} - c\mathbf{1}n\|^2$. Further, the c minimizing this is given by (4.9) and its substitution in (21.2) leads to (21.3). Thus, the maximum of the *log likelihood* (21.30) can be expressed as

$$l(\hat{\mathbf{b}}) = -\frac{n}{2}\log\left\|\mathbf{J}\mathbf{y} - \mathbf{J}\mathbf{X}\hat{\mathbf{b}}\right\|^2 + const \tag{21.31}$$

in *regression analysis*. Here, $\hat{\mathbf{b}}$ is the optimal $\mathbf{b}$ maximizing (21.31) or minimizing (21.3) subject to the constraint (21.24) in the *sparse regression*, while $\widehat{\mathbf{b}}$ is the optimal $\mathbf{b}$ without a constraint in the *standard regression*.

By substituting the part of (21.31) excluding *const* into $l(\widehat{\boldsymbol{\Theta}})$ in (8.25), BIC can be defined as

$$\mathrm{BIC} = n\log\left\|\mathbf{J}\mathbf{y} - \mathbf{J}\mathbf{X}\hat{\mathbf{b}}\right\|^2 + \eta\log n. \tag{21.32}$$

in regression analysis. Here, η is the *number of parameters* to be estimated. The reason why *const* in (21.31) may be ignored is that the value of *const* is equivalent among the solutions for different procedures: *const* is irrelevant for the comparison of BIC among the solutions. In *standard regression*, η is the number of coefficients plus two corresponding to c and σ^2: $\eta = p + 2$. What should be the value of η in *sparse regression*? Some authors argue that it is the number of nonzero coefficients plus two:

$$\eta = \mathrm{Card}(\hat{\mathbf{b}}) + 2 \tag{21.33}$$

(e.g., Zou, Hastie, & Tibshirani, 2007). Here, Card() stands for the cardinality of the parenthesized vector or matrix, i.e., the number of its nonzero elements. We adopt (21.33). In a parallel manner, AIC can also be defined.

The lasso solution in Table 21.1B has been selected using (21.32) with (21.33). That is, we obtained $\hat{\mathbf{b}}$ for each of w = 0.01, 0.02, 0.03, …, 9.96, 9.98, 10.00, and the $\hat{\mathbf{b}}$ for w = 7.96 which gave the least BIC is presented in Table 21.1B.

21.5 L_0 Sparse Regression

In Note 21.1, the L_{q}- norm is defined for $q > 0$. The L_0 norm, i.e., $\|\mathbf{A}\|_q$ for $q = 0$, is exceptionally defined as follows:

Note 21.4. L_0 Norm
For $q = 0$, the L_q norm in Note 21.1 cannot be defined. But, if exceptionally $1/q$ is defined as 1 and $|0|^q$ is set to 0 for $q = 0$ in Note 21.1, the L_0 *norm* of $\mathbf{A} = (a_{ij})$ $(n \times p)$ is given by

$$\|\mathbf{A}\|_0 = \mathrm{Card}(\mathbf{A}) = \sum_{i=1}^{n}\sum_{j=1}^{p} I(a_{ij} \neq 0), \tag{21.34}$$

with

$$I(a_{ij} \neq 0) = \begin{cases} 0 & \text{if } a_{ij} = 0 \\ 1 & \text{otherwise} \end{cases}. \tag{21.35}$$

In this section, we consider sparse regression with $\|\mathbf{b}\|_0 = \sum_{j=1}^{p} I(b_j \neq 0)$ used for Pen($\mathbf{b}$) in (21.4), i.e., minimizing

$$f_{L0}(\mathbf{b}, c) = \|\mathbf{Jy} - \mathbf{JXb}\|^2 + nw\|\mathbf{b}\|_0. \tag{21.36}$$

We refer to this minimization as L_0 *sparse regression*. Its algorithm can be derived by substituting $I(b_j \neq 0)$ for $|b_j|$ in the equations of Sect. 21.3.

The L_0-norm version of (21.10) is derived by substituting $I(b_j \neq 0)$ for $|b_j|$:

$$f_{L0}(\mathbf{b}) = \left\|\mathbf{r}_{(j)}\right\|^2 + \left(\mathbf{x}_j'\mathbf{J}\mathbf{x}_j\right)h_j(b_j) + nw\sum_{k\neq j} I(b_k \neq 0). \tag{21.37}$$

Here, only $h_j(b_j)$ is a function of b_j, which is expressed as

$$h_j(b_j) = b_j^2 - 2\frac{\mathbf{r}'_{(j)}\mathbf{J}\mathbf{x}_j}{\mathbf{x}_j'\mathbf{J}\mathbf{x}_j}b_j + \frac{nw}{\mathbf{x}_j'\mathbf{J}\mathbf{x}_j}I(b_j \neq 0)$$
$$= b_j^2 - 2\frac{\mathbf{r}'_{(j)}\mathbf{J}\mathbf{x}_j}{nv_{jj}}b_j + \frac{w}{v_{jj}}I(b_j \neq 0) = b_j^2 - 2r_j(\mathbf{b}_{[j]})b_j + s_j(w)I(b_j \neq 0), \tag{21.38}$$

with $v_{jj} = n^{-1}\mathbf{x}_j'\mathbf{J}\mathbf{x}_j$ the variance of the jth variable, $r_j(\mathbf{b}_{[j]})$ defined as (21.12), and

$$s_j(w) = \frac{w}{v_{jj}} \geq 0. \tag{21.39}$$

The coefficient b_j minimizing (21.38) is the optimal one that minimizes the loss function (21.36) with the other coefficients b_k ($k \neq j$) fixed. The minimizer of (21.38) is given by

$$b_j = \begin{cases} 0 & \text{if } |r_j(\mathbf{b}_{[j]})| \leq \sqrt{s_j(w)} \\ r_j(\mathbf{b}_{[j]}) & \text{otherwise} \end{cases}, \tag{21.40}$$

as explained in Note 21.5 presented below. We can obtain the optimal $\mathbf{b}$ by iterating the update of b_j through (21.40) over $j = 1, \ldots, p$. Thus, the *algorithm for the L_0 sparse regression* can be summarized as follows:

Step 1. Initialize $\mathbf{b}$
Step 2. For each of $j = 1, \ldots, p$, perform the following: Obtain $r_j(\mathbf{b}_{[j]})$ using the current b_k values in (21.12) to update b_j with (21.40).
Step 3. Finish if convergence is reached; otherwise, go back to Step 2.

Here, we should note that the current $r_j(\mathbf{b}_{[j]})$ value must be obtained before the update of b_j in Step 2, since $r_j(\mathbf{b}_{[j]})$ in (21.14) is a function of the coefficients $b_1, \ldots, b_p$ except b_j. The initialization in Step 1 is made by setting $\mathbf{b}$ to the standard regression for the computations in this chapter.

Note 21.5. Minimizing a Quadratic Function plus L_0 Norm
For the sake of simplicity, we omit $(\mathbf{b}_{[j]})$, (w), and the subscript j from the symbols in (21.38). Thus, (21.38) is simplified as $h(b) = b^2 - 2rb + sI(b \neq 0)$. It can be rewritten as

$$h(b) = \begin{cases} 0 & \text{if } b = 0 \\ (b-r)^2 + s - r^2 & \text{otherwise} \end{cases}. \tag{21.41}$$

Let us consider the cases of $|r| > s^{1/2}$, i.e., $s - r^2 < 0$. Then, (21.41) shows that the lower limit of $h(b)$ is $s - r^2$, which is attained for $b = r$. For the other

cases with $|r| \leq s^{1/2}$, i.e., $s - r^2 \geq 0$, (21.41) implies that the lower limit of $h(b)$ is 0. It can be attained for $b = 0$. These facts lead to (21.40).

A suitable penalty weight w can be selected as described in Sect. 21.4, using BIC defined as (21.32) with (21.33). Thus, the solution of the L_0 sparse regression in Table 21.1C has been obtained using (21.32) with (21.33). That is, we run the above algorithm to obtain $\mathbf{b}$ for each of $w = 0.01, 0.02, 0.03, \ldots, 9.96, 9.98, 10.00$. This results in the $\mathbf{b}$ for $w = 0.56$ giving the least BIC. The corresponding solution is presented in Table 21.1(C).

The *absolute values of the nonzero coefficients* in *lasso* cannot be greater than those in the L_0 sparse regression. This fact can be proved as follows: The lasso formula (21.14) shows the absolute value of nonzero b_j to be $|b_j| = |r_j(\mathbf{b}_{[j]})| - d_j(w)$, while the L_0 counterpart (21.40) shows $|b_j| = |r_j(\mathbf{b}_{[j]})|$. Since $d_j(w) \geq 0$, $|r_j(\mathbf{b}_{[j]})| - d_j(w)$ in lasso cannot be greater than $|r_j(\mathbf{b}_{[j]})|$ in the L_0 sparse regression. This fact can be illustrated in Table 21.1B, C: The absolute values of the nonzero solutions in lasso are smaller than the L_0 counterparts. The property of parameters being estimated so that their absolute values are smaller is referred to as *shrinkage* of parameter estimates. For this reason, "*shrinkage*" is included in the name *least absolute selection and shrinkage operator*, which lasso abbreviates as described in Sect. 21.2.

21.6 Standard Regression in Ordinary and High-Dimensional Cases

So far, we treated the cases in which regression analysis is performed for data sets with more individuals than variables, which we call *ordinary* cases. In this section, we consider the cases with *more variables than individuals*, which can be called *high-dimensional* cases. In this section, we explain how *standard regression* produces *unusable results* in *high-dimensional* cases, in order to prepare for the next section where sparse regression is shown to be useful in such cases. For the explanation, we compare properties of the *standard regression* solution between the *ordinary* cases of $n > p + 1$ (more individuals) and the *high-dimensional* cases of $n < p$ (more variables). The goal of the section is to reach the following conclusions:

[*Ordinary*] If $n > p + 1$, then the value of the loss function (21.3) is usually greater than zero. If n is sufficiently greater than $p + 1$, the solution of $\mathbf{b}$ is useful, as seen so far.

[*High-Dim*] If $n < p$, then the resulting (21.3) value is zero, i.e., a perfect fit $\mathbf{Jy} = \mathbf{JXb}$ is attained, but the solution of $\mathbf{b}$ is useless, as it is not unique

To arrive at this goal, a key point is whether rank(**JX**) is equivalent to rank([**JX**, **Jy**]) or not, with [**JX**, **Jy**] an $n \times (p + 1)$ block matrix.

We start with the fact that

$$\text{rank}(\mathbf{JX}) \le \min(n-1, p) \quad \text{and} \quad \text{rank}([\mathbf{JX}, \mathbf{Jy}]) \le \min(n-1, p+1). \tag{21.42}$$

Although (3.32) leads to rank(**JX**) $\le$ min(n, p), this n is replaced by $n - 1$ in (21.42) since $\mathbf{1}_n'\mathbf{JX} = \mathbf{0}_p'$: The n rows of **JX** are not linearly independent, thus rank (**JX**) is less than n. The inequalities in (21.42) imply

$$\text{rank}(\mathbf{JX}) \le p \quad \text{and} \quad \text{rank}([\mathbf{JX}, \mathbf{Jy}]) \le p+1 \quad \text{for } n > p+1, \tag{21.43}$$

$$\text{rank}(\mathbf{JX}) \le n-1 \quad \text{and} \quad \text{rank}([\mathbf{JX}, \mathbf{Jy}]) \le n-1 \quad \text{for } n < p. \tag{21.44}$$

Now, we suppose that rank(**JX**) and rank([**JX**, **Jy**]) attain their upper limits. Then, (21.43) leads to that

$$\begin{aligned} &\text{if } n > p+1,\ \text{rank}(\mathbf{JX}) = p \text{ and } \text{rank}([\mathbf{JX}, \mathbf{Jy}]) = p+1 \\ &\text{thus, } \ \text{rank}(\mathbf{JX}) \ne \text{rank}([\mathbf{JX}, \mathbf{Jy}]), \end{aligned} \tag{21.45}$$

while (21.44) implies that

$$\text{if } n < p, \quad \text{rank}(\mathbf{JX}) = \text{rank}([\mathbf{JX}, \mathbf{Jy}]) = n-1. \tag{21.46}$$

How (21.46) leads to the above conclusion [*High-Dim*] is explained next:

Note 21.6. High-Dimensional Regression
This title refers to the regression in the cases of $n < p$ as in (21.46). This implies that we can substitute **JX** into **X** and **Jy** into **y** in (17.9) and (17.11): For given **JX** and **Jy**, the system of linear equations,

$$\mathbf{JXb} = \mathbf{Jy}, \tag{21.47}$$

has a solution of **b** and thus the value of loss function (21.3) becomes zero. From (17.12), the solution of **b** for (21.47) is given by

$$\mathbf{b} = (\mathbf{JX})^+ \mathbf{Jy} + \{\mathbf{I}_p - (\mathbf{JX})^+ \mathbf{JX}\}\mathbf{q} \tag{21.48}$$

with **q** an arbitrary $p \times 1$ vector.

The resulting (21.3) value being 0 implies a *perfect* fit, but (21.48) shows that the solution of **b** is *not unique*, since **q** is arbitrary: *Infinitely many solutions* exist. Thus, *high-dimensional regression* is *useless*.

In the *ordinary* cases with (21.45), (21.47) does not hold: In general, the value of (21.3) is not zero. However, it is useful, as found in Chap. 4.

An illustration of *how high-dimensional regression is useless* is shown below, using the high-dimensional **JX** (5×6) and **Jy** in Table 21.2. The solution (21.48) obtained for $\mathbf{q} = \mathbf{0}_6$ is

$$\mathbf{b} = [0.010, 0.198, -0.075, -0.022, 0.025, 0.196]', \tag{21.49}$$

while (21.48) for $\mathbf{q} = [2, 2, -2, 2, 3, 2]'$ is

$$\mathbf{b} = [1.648, -0.937, -2.276, 1.414, 1.480, 0.391]'. \tag{21.50}$$

These two solutions are very *different*, but both solutions for **b** attain the *perfect* fit $f(\mathbf{b}) = \|\mathbf{Jy} - \mathbf{JXb}\|^2 = 0$ with $\mathbf{JXb} = \mathbf{Jy}$.

Indeed, the vector **Jy** in Table 21.2 has been *artificially generated* using the formula $\mathbf{y} = \mathbf{JXb} + \mathbf{e}$. Here, **b** was set to

$$\mathbf{b}_{\text{true}} = [-2, 1, 0, 0, 0, 0]', \tag{21.51}$$

JX was the same as in Table 21.2, and **e** is a centered error vector with the absolute values of its elements somewhat smaller than those of **JX**. Here, it is important to note that **e** is centered: $\mathbf{e} = \mathbf{Je}$. Thus, **y** generated by the above formula $\mathbf{y} = \mathbf{JXb} + \mathbf{e}$ satisfies $\mathbf{y} = \mathbf{Jy}$: The formula may be written as $\mathbf{Jy} = \mathbf{JXb} + \mathbf{e}$. Thus, the **b** minimizing $f(\mathbf{b}) = \|\mathbf{e}\|^2 = \|\mathbf{Jy} - \mathbf{JXb}\|^2 = 0$ can be expected to be close to (21.51). In this sense, the subscript "true" has been attached to **b** in (21.51): If a procedure provides a solution **b** close to $\mathbf{b}_{\text{true}}$, the solution can be considered right. Unfortunately, both (21.49) and (21.50) are far from $\mathbf{b}_{\text{true}}$ (21.51), which demonstrates that high-dimensional regression is unusable.

It should be noted that the third to final elements in (21.51) are *zeros*, which implies $\mathbf{Jy} = \mathbf{JX}_{[2]}\mathbf{b}_{[2]} + \mathbf{e}$, with $\mathbf{X}_{[2]}$ the 5×2 matrix containing the first two columns of **X**, and $\mathbf{b}_{[2]} = [-2,1]'$ containing the first two element of (21.51). For the final column and the first two ones in Table 21.2 which are **Jy** and $\mathbf{JX}_{[2]}$, respectively, we performed regression analysis. This can be referred to as *standard regression* in an *ordinary* case, since $n = 5 > p = 2$. The resulting coefficient vector was $\mathbf{b} = [-0.167, 0.104]'$. This is fairly similar to $\mathbf{b}_{[2]} = [-2,1]'$. The result can be restated to claim that a *useful solution* was obtained, by *excluding* the last four

Table 21.2 Example of high-dimensional data with $n = 5 < p = 6$

JX						**Jy**
1.060	0.071	−1.609	−1.923	−1.482	−1.439	−0.131
−1.385	1.763	−0.539	0.652	1.514	−0.293	0.341
−0.137	0.042	0.876	0.826	0.543	1.150	0.162
−0.726	−0.686	0.116	0.442	−0.128	1.113	0.053
1.188	−1.190	1.157	0.003	−0.447	−0.531	−0.425

columns from **JX**; in other words, selecting the first two variables among the six columns of **JX**. Performing such *variable selection computationally* is considered in the next section.

21.7 High-Dimensional Variable Selection by Sparse Regression

The last section demonstrated the following fact: In the *high-dimensional* cases of $n < p$, *standard regression* solutions are *unusable*, but *variable selection* is *useful* for selecting $q < n$ variables among p variables. *Sparse regression* can be used for *variable selection*: It is expected that the coefficients for the q useful variables are estimated as nonzero, while the remaining coefficients are estimated as zeros, i.e., corresponding variables are excluded. We illustrate this in a *simulation study* example. Indeed, the illustration with Table 21.2 in the last section also falls under the category of simulation studies. A generalized setting for such studies is introduced in the following.

Note 21.7. Simulation Studies
What the term *simulation* stands for differs across disciplines. Here, we deal with simulation studies used in statistics. These studies are often made for assessing the performance of analysis procedures, in particular, for the purpose of evaluating "how exactly the parameter values underlying data can be *recovered* by the procedures". What this phrase put in quotation marks means is explained in the next paragraphs.

Let us suppose that a procedure to be assessed is modeled as

$$\mathbf{y} = \boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{X}) + \mathbf{e} \tag{21.52}$$

for an $n \times 1$ data vector $\mathbf{y}$, Here, $\mathbf{e}$ is an $n \times 1$ error vector, and $\boldsymbol{\eta}(\boldsymbol{\theta}, \mathbf{X})$ is an $n \times 1$ vector which is a function of an unknown parameter vector $\boldsymbol{\theta}$ to be obtained and the given matrix $\mathbf{X}$ ($n \times p$) containing data not included in $\mathbf{y}$.

The simulation study for (21.52) proceeds via the following steps:

[1] The data vector $\mathbf{y}$ is *artificially generated* with model (21.52). Here, $\boldsymbol{\theta}$ is set to a given vector $\boldsymbol{\theta}_{\text{true}}$, whose elements, i.e., parameter values, are specified artificially, while the elements in $\mathbf{e}$ are *generated randomly*, that is, set to random numbers generated by machine, with the numbers following a particular probability distribution. The elements of $\mathbf{X}$ in (21.52) are specified artificially or generated randomly.

[2] The analysis procedure to be assessed is carried out for the above $\mathbf{y}$ and $\mathbf{X}$, in order to obtain the solution of $\boldsymbol{\theta}$. Let the resulting $\boldsymbol{\theta}$ be denoted by $\hat{\boldsymbol{\theta}}$.

[3] It is assessed whether $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_{\text{true}}$ are similar or not. If they are similar, for example, $\boldsymbol{\theta}_{\text{true}} = [2.0, -3.0]$ and $\hat{\boldsymbol{\theta}} = [2.1, -3.2]$, the parameters are said to be *recovered well* and the procedure is considered to be promising.

The elements of $\boldsymbol{\theta}_{\text{true}}$ are called the *true parameter values*, as the aim of the above study is the assessment of whether the solution approximates $\boldsymbol{\theta}_{\text{true}}$ or not.

Simulation studies can also be found, in which the elements of $\boldsymbol{\theta}_{\text{true}}$ are generated randomly. The studies can be useful also for assessing procedures modeled without $\mathbf{X}$, i.e., the ones modeled as $\mathbf{y} = \boldsymbol{\eta}(\boldsymbol{\theta}) + \mathbf{e}$ rather than (21.52).

We illustrate a simulation study in which (21.52) is the regression model (21.1) with the intercept c set to 0: Data are artificially generated according to $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$ as in the next paragraph.

We consider the *high-dimensional* case with $n = 100 < p = 300$. The coefficient vector $\mathbf{b}$ set to the true $\mathbf{b}_{\text{true}}$, whose elements were *filled with zeros except for three elements*, being 2.0, −3.0, and 1.5: Among the 300 true coefficients, only these three are nonzero. The elements of $\mathbf{e} = [e_1, \ldots, e_{100}]'$ were generated randomly so that they follow a normal distribution whose mean and variance are zero and σ_e^2, while the rows of $\mathbf{X}$ are generated randomly so that they follow $N_{300}(\mathbf{0}_{300}, \sigma_x^2 \mathbf{I}_{300})$. Here, the values of σ_e^2 and σ_x^2 were accommodated so that $\|\mathbf{e}\|^2/\|\mathbf{y}\|^2 \cong 0.1$. This proportion implies that the 10 percent variation present in data vector $\mathbf{y}$ can correspond to errors, roughly speaking; a full explanation is too involved to detail here.

The *lasso* and L_0 *sparse regression* procedures were carried out for the data in $\mathbf{y}$ and $\mathbf{X}$ generated as above. The values of the tuning parameter w were set as $w = 0.01, 0.02, \ldots, 0.98, 1.00$ in lasso and $w = 0.1, 0.2, \ldots, 9.8, 10.0$ in the L_0 sparse regression. As a result, the least BIC was attained for $w = 0.77$ in lasso and $w = 3.0$ in L_0. The results for those least BIC are presented in Table 21.3 with the values of the nonzero elements in $\mathbf{b}_{\text{true}}$. In the table, MIS stands for the number of coefficients whose true values are zeros, but estimates are nonzero: In lasso, two among the 297 (=300 − 3) zero elements in $\mathbf{b}_{\text{true}}$ were estimated as the nonzero

Table 21.3 Lasso and L_0 sparse regression estimates for the true values of nonzero coefficients, with MIS and BIC values

Procedure		lasso	L_0
True	2.0	1.53	1.94
	−3.0	−2.53	−2.93
	1.5	1.14	1.49
MIS		2	0
BIC		99.1	69.2

values 0.03 and 0.05. However, both procedures can be said to provide *proper solutions*, in that the estimates of nonzero coefficients in the L_0 sparse regression are *close* to the true values, and the lasso estimates of nonzero coefficients are fairly similar to the true ones, though also found to shrink. The BIC result shows that the L_0 solution is better.

21.8 Bibliographical Notes

Lasso has been treated in a number of books, among which the more recent Hastie et al. (2015) is recommended. Lasso and related methods have been also treated in Hastie, Tibshirani, and Friedman (2009). The L_0 sparse regression is treated in Bühlmann and van de Geer (2011). There, high-dimensional regression is detailed in an advanced manner.

Beside the L_0 and L_1 norms, some other penalty functions, which take much more complicated forms than L_0 and L_1 norms, have been proposed for sparse regression analysis. Fan and Li's (2010) SCAD and Zhang's (2010) MC+ are among those functions.

Exercises

21.1. Summarize the cases when *sparse regression* procedures are to be used.

21.2. In a procedure called *ridge regression* (Hoerl & Kennard, 1970), the loss function to be minimized is defined as (21.4) with $\text{Pen}(\mathbf{b}) = \|\mathbf{b}\|^2 = \sum_{j=1}^{p} b_j^2$. Describe the differences between the ridge regression and lasso.

21.3. The number of parameters (η) in (21.32) is set to (21.33) in sparse regression procedures and $\eta = p + 2$ in standard regression. Show that the number of parameters may be defined as $\eta = \text{Card}(\hat{\mathbf{b}})$ for sparse regression procedures and $\eta = p$ for standard regression.

21.4. Functions which satisfy (A.6.7) in Appendix 6 are said to be *convex*. The lasso loss function (21.7) is known to be convex (i.e., Hastie et al., 2015). The proof of this is not easy. In place of it, show that (21.11) is convex.

21.5. In a procedure called *adaptive lasso* (Zou, 2006), the loss function (21.4) with $\text{Pen}(\mathbf{b}) = \sum_{j=1}^{p} \alpha_j |b_j|$ is minimized over $\mathbf{b} = [b_1, \ldots, b_p]$, for given weights $\alpha_1, \ldots, \alpha_p$. For example, α_j is set to $\left|\hat{b}_j^{(\text{std})}\right|^{-1}$, with $\hat{b}_j^{(\text{std})}$ the solution of the coefficient for the jth explanatory variable in the standard regression. Discuss the rationality of using $\alpha_j = \left|\hat{b}_j^{(\text{std})}\right|^{-1}$.

21.6. Show that when the number of explanatory variables is small enough, for example, $p = 4$, variable selection can be attained by comparing the BIC values among the solutions for the *standard regression* analyses with *all possible subsets of the explanatory variables*.

21.7. As described in Exercise 15.7, $P(\text{Parameters})P(\text{Data}|\text{Parameters})$ is maximized over parameters in the *Bayesian method* for estimating parameters. Discuss how sparse regression analysis, i.e., minimizing (21.4) over $\mathbf{b}$, can be reformulated as a *Bayesian method*, in which $P(\mathbf{b}, \sigma^2)P(\mathbf{y}|\mathbf{b}, \sigma^2, \mathbf{X})$ is maximized for the centered block data matrix $[\mathbf{X}, \mathbf{y}]$ satisfying $[\mathbf{X}, \mathbf{y}] = \mathbf{J}[\mathbf{X}, \mathbf{y}]$, where $P(\mathbf{y}|\mathbf{b}, \sigma^2, \mathbf{X})$ is the probability density of $\mathbf{y} \sim N_n(\mathbf{X}\mathbf{b}, \sigma^2\mathbf{I}_n)$ and $P(\mathbf{b}, \sigma^2) = a \times \exp\{-(2\sigma^2)^{-1}nw\text{Pen}(\mathbf{b})\}$ with a a suitable constant.

21.8. Let us consider performing the *simulation study* introduced in Note 21.7, in order to evaluate a *multidimensional scaling* procedure (Chap. 16), in which (16.5) is minimized over $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_n]'$. In this study, (21.52) is replaced by $q_{ij} = \|\mathbf{a}_i - \mathbf{a}_j\|^2 + e_{ij}$ $(i = 1, \ldots, n-1, j = i+1, \ldots, n)$ with e_{ij} an error. Show how [1], [2], and [3] in Note 21.7 are rewritten for the study.

Chapter 22
Sparse Factor Analysis

In the last chapter, modified regression analysis procedures were presented, in which a coefficient vector is estimated so that it is sparse, i.e., includes a number of zero elements. Such *sparse estimation* can be incorporated into other multivariate analysis procedures, so as to provide sparse solutions. They can be easily interpreted, as we may only focus on their nonzero elements. As such, a number of *sparse multivariate procedures* have been developed, following the sparse estimation techniques developed in regression. The procedures include *sparse factor analysis* (*FA*) for providing a *sparse factor loading matrix*. In this chapter, we introduce the *two types* of sparse FA procedures. In one of the two, a *penalty function* is *used*, while the function is *not used* in the other type. This chapter starts by describing a *drawback of confirmatory FA* (Chap. 10) which can be *handled* by sparse FA procedures.

22.1 From Confirmatory FA to Sparse FA

Let us recall the *factor analysis* (*FA*) model in Chap. 10. We present the model (10.3) again here: A $p \times 1$ random variable vector $\mathbf{x} = [x_1, \ldots, x_p]$, whose expected vector is $\mathbf{0}_p$, is modeled as

$$\mathbf{x} = \mathbf{A}\mathbf{f} + \mathbf{e}, \tag{22.1}$$

where $\mathbf{A}$ is the p variables $\times$ m-factors loading matrix, $\mathbf{f}$ is an $m \times 1$ common factor vector, and $\mathbf{e}$ is a $p \times 1$ unique factor vector. The vectors $\mathbf{f}$ and $\mathbf{e}$ are assumed to follow multivariate (MVN) distributions as seen in (10.4) and (10.6), which leads to the MVN distribution (10.9) for $\mathbf{x}$ with its covariance matrix (10.10). Then, for sample covariance matrix $\mathbf{V} = n^{-1}\mathbf{X}'\mathbf{X}$ with $\mathbf{X}$ the n individuals $\times$ p variables centered data matrix whose ith row is the transpose of (22.1) observed for individual i, the log likelihood is defined as (10.11), i.e.,

K. Adachi, *Matrix-Based Introduction to Multivariate Data Analysis*,
https://doi.org/10.1007/978-981-15-4103-2_22

$$l^*(\mathbf{A}, \mathbf{\Psi}, \mathbf{\Phi}) = \frac{n}{2}\log\left|(\mathbf{A}\mathbf{\Phi}\mathbf{A}' + \mathbf{\Psi})^{-1}\mathbf{V}\right| - \frac{n}{2}\mathrm{tr}(\mathbf{A}\mathbf{\Phi}\mathbf{A}' + \mathbf{\Psi})^{-1}\mathbf{V}. \tag{22.2}$$

Here, $\mathbf{\Phi}$ ($m \times m$) is the factor correlation vector (10.5) and $\mathbf{\Psi}$ ($p \times p$) is the diagonal matrix (10.7) including unique variances.

In *confirmatory FA* (*CFA*) (Chap. 10), (22.2) is maximized over $\mathbf{A}$, $\mathbf{\Phi}$, and $\mathbf{\Psi}$ under the assumed relationship between variables and factors. An example of this assumption is given by

$$\underset{8\times1}{\mathbf{x}} = \underset{8\times2}{\mathbf{A}}\ \underset{2\times1}{\tilde{\mathbf{f}}} + \underset{8\times1}{\mathbf{e}}, \tag{22.3}$$

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \end{bmatrix} = \begin{bmatrix} a_{11} & 0 \\ a_{21} & a_{22} \\ a_{31} & 0 \\ 0 & a_{42} \\ 0 & a_{52} \\ a_{61} & 0 \\ a_{71} & a_{72} \\ 0 & a_{82} \end{bmatrix} \begin{bmatrix} Factor\_1 \\ Factor\_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \\ e_8 \end{bmatrix},$$

for a data set of eight variables. This is illustrated in Fig. 22.1: Factor 1 is linked to variables x_1, x_2, x_3, x_6, and x_7, while Factor 2 is linked to x_2, x_4, x_5, x_7, and x_8. The assumption can also be restated as a constraint for what loadings in $\mathbf{A}$ are to be zero: In (22.3), $\mathbf{A} = (a_{jk})$ is constrained as

$$a_{12} = a_{32} = a_{41} = a_{51} = a_{62} = a_{81} = 0. \tag{22.4}$$

Such *zero constraints*, in other words, what pairs of variables and factors should be linked, *must be specified by users*.

In CFA, the users' constraints are *subjective* and might be inadequate. This problem can be *avoided* in *sparse factor analysis* (*SFA*): Loadings which should be zero (in other words, what pairs of variables and factors should be linked) are estimated *computationally* (i.e., automatically) and *objectively*. The procedure performed by SFA can be stated in more precisely to:

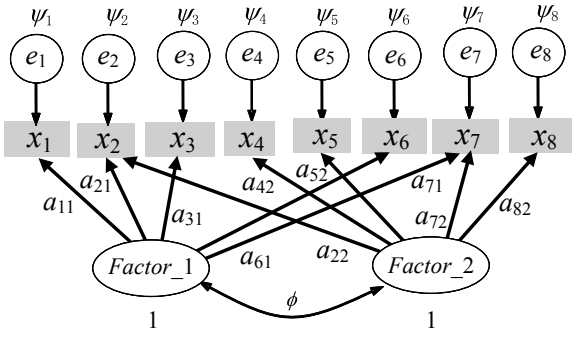


Fig. 22.1 Example of CFA models for $p = 8$ and $m = 2$

$$\begin{aligned}&\text{Optimally and jointly estimate}\\&[\text{A}]\,\text{what loadings in } \mathbf{A} \text{ are to be zero and}\\&[\text{B}]\,\text{the values of nonzero parameters}\\&\qquad\qquad\text{for a certain number of Card}(\mathbf{A}).\end{aligned} \tag{22.5}$$

Here, Card($\mathbf{A}$) (the cardinality of $\mathbf{A}$) is the number of nonzero elements in $\mathbf{A}$; for example, Card($\mathbf{A}$) = 10 in (22.3) with the constraint (22.4). Besides this, a number of other constraints exist which satisfy Card($\mathbf{A}$) = 10. In SFA, the *optimal* one can be found among these, *together with the optimal nonzero parameter values*.

In this chapter, two approaches to SFA are introduced. One is using a penalty function whose idea was introduced in the last chapter. This *penalized approach* is described in Sects. 22.2–22.5. The other approach is introduced in Sect. 22.6–22.8, in which *a penalty function is not used*.

22.2 Formulation of Penalized Sparse LVFA

We introduce an SFA procedure in which the log likelihood (22.2) is combined with a *penalty function* which penalizes nonzero valued loadings. It can be called a *penalized sparse latent variable FA* (*PS-LVFA*), as (22.2) is underlain by the *latent variable formulation* of FA, as discussed in Sect. 18.2. Out of the procedures that have been proposed so far, we introduce one by Hirose and Yamamoto (2014) as a typical PS-LVFA procedure, with only a minor modification here.

In sparse regression treated in the last chapter, the loss function of standard regression, for which a penalty function is summed, is to be minimized. On the other hand, the log likelihood (22.2) is to be *maximized* in FA, and a penalty function is instead *subtracted* from (22.2) in PS-LVFA. That is, the function to be maximized is defined as the log likelihood (22.2) *minus* penalty function Pen($\mathbf{A}$) weighted by nw:

$$f_{\text{PEN}}(\mathbf{A}, \boldsymbol{\Psi}, \boldsymbol{\Phi}) = l^*(\mathbf{A}, \boldsymbol{\Psi}, \boldsymbol{\Phi}) - nw\text{Pen}(\mathbf{A}). \tag{22.6}$$

Here, w ($\geq$ 0) serves as a penalty weight, and this is multiplied by n merely for the sake of the convenience during the subsequent derivations of equations. This is maximized over $\mathbf{A}$,$\boldsymbol{\Psi}$, and $\boldsymbol{\Phi}$, subject to $\boldsymbol{\Psi}$ being diagonal and $\boldsymbol{\Phi}$ being a correlation matrix.

Hirose and Yamamoto (2014) used the function called *MC+* (Zhang, 2010) as the penalty function (though they also have considered the L_1-norm penalty introduced in the last chapter). It is rather convenient to introduce MC+ in the form multiplied by w; i.e., $w \times$ MC+ defined as

$$wMC(a_{jk};w,u)=\begin{cases} w\left(|a_{jk}|-\frac{a_{jk}^2}{2wu}\right) & \text{if } |a_{jk}|<wu \\ \frac{w^2u}{2} & \text{otherwise} \end{cases}. \tag{22.7}$$

Here, u and w are *tuning parameters* to be specified. The function (22.7) takes a complicated form in order to maintain statistically desirable properties (Zhang, 2010), the details of which are beyond the scope of this book. The sum of (22.7) over j and k multiplied by n gives $nw\text{Pen}(\mathbf{A})=n\sum_{j=1}^{p}\sum_{k=1}^{m}wMC(a_{jk};w,u)$. Using this and (22.2) in (22.6), we have

$$\begin{aligned} f_{\text{PEN}}(\mathbf{A},\boldsymbol{\Psi},\boldsymbol{\Phi}) &= \frac{n}{2}\log\left|(\mathbf{A}\boldsymbol{\Phi}\mathbf{A}'+\boldsymbol{\Psi})^{-1}\mathbf{V}\right| - \frac{n}{2}\text{tr}(\mathbf{A}\boldsymbol{\Phi}\mathbf{A}'+\boldsymbol{\Psi})^{-1}\mathbf{V} \\ &\quad - n\sum_{j=1}^{p}\sum_{k=1}^{m}wMC(a_{jk};w,u). \end{aligned} \tag{22.8}$$

The function (22.8) is maximized over $\mathbf{A}$, $\boldsymbol{\Psi}$, and $\boldsymbol{\Phi}$ for a given $[w,u]$. Here, we should note that w and u specify the penalty function $n\sum_{j=1}^{p}\sum_{k=1}^{m}wMC(a_{jk};w,u)$, which controls Card($\mathbf{A}$) in (22.5). However, the *correspondence* of w and u values *to* Card($\mathbf{A}$) is *unknown before* maximizing (22.8): Card($\mathbf{A}$) is *found afterward* in the resulting solution. Thus, it may be more correct to rewrite the final phrase in (22.5) as in

$$\begin{aligned} &\text{Optimally and jointly estimate} \\ &[\text{A}]\text{what loadings in } \mathbf{A} \text{ are to be zero and} \\ &[\text{B}]\text{ the values of nonzero parameters,} \\ &\qquad\qquad \text{for a given } [w,c] \text{ controlling Card}(\mathbf{A}) \end{aligned} \tag{22.9}$$

for PS-LVFA.

22.3 Algorithm for Penalized Sparse LVFA

As explained in Appendix A.9.9, the *EM algorithm for penalized FA* can be used to maximize (22.8). That is, we may consider maximizing the function (A.9.43) with $g(\boldsymbol{\Theta})$ replaced by $n\sum_{j=1}^{p}\sum_{k=1}^{m}wMC(a_{jk};w,u)$ and constant c deleted:

$$\begin{aligned} \phi(\mathbf{A},\boldsymbol{\Psi},\boldsymbol{\Phi}) &= -\frac{n}{2}\log|\boldsymbol{\Psi}| - \frac{n}{2}\text{tr}(\mathbf{V}-2\mathbf{B}\mathbf{A}'+\mathbf{A}\mathbf{Q}\mathbf{A}')\boldsymbol{\Psi}^{-1} \\ &\quad - \frac{n}{2}\left(\log|\boldsymbol{\Phi}|+\text{tr}\boldsymbol{\Phi}^{-1}\mathbf{Q}\right) - n\sum_{j=1}^{p}\sum_{k=1}^{m}wMC(a_{jk};w,u). \end{aligned} \tag{22.10}$$

Here, $\mathbf{B}$ and $\mathbf{Q}$ are defined as (A.9.18) and (A.9.19), i.e.,

$$\mathbf{B} = \mathbf{VH}(\boldsymbol{\Theta}) \quad \text{and} \quad \mathbf{Q} = \mathbf{W}(\boldsymbol{\Theta}) + \mathbf{H}(\boldsymbol{\Theta})'\mathbf{VH}(\boldsymbol{\Theta}), \tag{22.11}$$

with $\mathbf{H}(\boldsymbol{\Theta})(p \times m)$ and $\mathbf{W}(\Theta)(m \times m)$ defined as (A.9.20) and (A.9.21) using (A.9.12), i.e., $\mathbf{H}(\boldsymbol{\Theta})$ and $\mathbf{W}(\boldsymbol{\Theta})$ the matrix functions of set $\boldsymbol{\Theta} = \{\mathbf{A}, \boldsymbol{\Psi}, \boldsymbol{\Phi}\}$ expressed as

$$\mathbf{H}(\boldsymbol{\Theta}) = (\mathbf{A}\boldsymbol{\Phi}\boldsymbol{A}' + \boldsymbol{\Psi})^{-1}\mathbf{A}\boldsymbol{\Phi} \text{ and } \mathbf{W}(\boldsymbol{\Theta}) = \boldsymbol{\Phi}^{1/2}\left(\mathbf{I}_m + \boldsymbol{\Phi}^{1/2}\mathbf{A}'\boldsymbol{\Psi}^{-1}\mathbf{A}\boldsymbol{\Phi}^{1/2}\right)^{-1}\boldsymbol{\Phi}^{1/2}. \tag{22.12}$$

In the EM algorithm, *E- and M-steps* are iterated until convergence is reached, as explained in Appendices A.8.5 and A.9. In the E-step, the current $\boldsymbol{\Theta} = \{\mathbf{A}, \boldsymbol{\Psi}, \boldsymbol{\Phi}\}$ values are substituted in (22.11) for providing (22.10). In the M-step, $\mathbf{A}$,$\boldsymbol{\Psi}$, and $\boldsymbol{\Phi}$ are updated so as to increase the value of (22.10). Thus, the algorithm for PS-LVFA can be summarized as follows:

Step 1. Initialize $\mathbf{A}$,$\boldsymbol{\Psi}$, and $\boldsymbol{\Phi}$
Step 2. E-step: obtain (22.11).
Step 3. M-step: update $\mathbf{A}$,$\boldsymbol{\Psi}$, and $\boldsymbol{\Phi}$ so as to increase (22.10).
Step 4. Finish if convergence is reached; otherwise, go back to Step 2.

Step 3 is detailed in the next section. Here, we describe the details for Steps 1 and 4. In Step 1, $\boldsymbol{\Phi}$ is set to $\mathbf{I}_m$. Principal component analysis followed by the varimax rotation (Chaps 5 and 13) is used for initializing $\mathbf{A}$: This is set to the matrix resulting in the varimax rotation for $\mathbf{L}_m\boldsymbol{\Delta}_m^{1/2}$. Here, $\mathbf{L}_m$ $(p \times m)$ contains the first m columns of $\mathbf{L}$, and $\boldsymbol{\Delta}_m$ is the first diagonal $m \times m$ block of $\boldsymbol{\Delta}$, with the matrices $\mathbf{L}$ and $\boldsymbol{\Delta}$ obtained by the eigenvalue decomposition in Note 6.1. The initial values of the diagonal elements in $\boldsymbol{\Psi}$ are set to those of $\mathbf{V} - \mathbf{L}_m\boldsymbol{\Delta}_m\mathbf{L}_m'$. Convergence in Step 4 is defined as the difference of the value of (22.8) $\times$ $2/n$ from the previous round being less than 0.1^8.

22.4 M-Step for Penalized Sparse LVFA

In this section, we describe how $\mathbf{A}$, $\boldsymbol{\Psi}$, and $\boldsymbol{\Phi}$ are updated so as to increase (22.10) in the M-step.

First, let us consider the updating of $\mathbf{A}$. We can rewrite (22.10) as $\phi(\mathbf{A}, \boldsymbol{\Psi}, \boldsymbol{\Phi}) = ng(\mathbf{A}) + \text{const}_{[\mathrm{A}]}$. Here, $const_{[\mathrm{A}]}$ is the constant irrelevant to $\mathbf{A} = (a_{jk})$, and

$$g(\mathbf{A}) = -\frac{1}{2}\mathrm{tr}(\mathbf{V} - 2\mathbf{B}\mathbf{A}' + \mathbf{A}\mathbf{Q}\mathbf{A}')\boldsymbol{\Psi}^{-1} - \sum_{j=1}^{p}\sum_{k=1}^{m} wMC(a_{jk}; w, u)$$
$$= -\sum_{j=1}^{p}\frac{1}{2\psi_j}\left(v_{jj} - 2\sum_{k=1}^{m} a_{jk}b_{jk} + \sum_{k=1}^{m}\sum_{l=1}^{m} q_{kl}a_{jk}a_{jl}\right) - \sum_{j=1}^{p}\sum_{k=1}^{m} wMC(a_{jk}; w, u) \tag{22.13}$$

with $\mathbf{V} = (v_{jk}), \mathbf{A} = (a_{ik}), \mathbf{B} = (b_{ik}), \mathbf{Q} = (q_{kl})$, and ψ_j the jth diagonal element of $\boldsymbol{\Psi}$. It should be noted that $\mathbf{Q}$ is symmetric with $q_{kl} = q_{lk}$. In order to rewrite (22.13) further, we use the following fact:

Note 22.1. Summation of $q_{kl}a_{jk}a_{kl}$ over k and l
The summation of $q_{kl}a_{jk}a_{jl}$ over k and l with $q_{kl} = q_{lk}$ can be rewritten as

$$\sum_{k=1}^{m}\sum_{l=1}^{m} q_{kl}a_{jk}a_{jl} = \sum_{k=1}^{m} q_{kk}a_{jk}^2 + \sum_{k=1}^{m}\sum_{l\neq k} q_{kl}a_{jk}a_{jl}$$
$$= q_{kk}a_{jk}^2 + 2\sum_{l\neq k} q_{kl}a_{jk}a_{jl} + c_{[j,k]}$$

with $c_{[i,k]} = \Sigma_{l\neq k}q_{ll}a_{jl}^2 + \Sigma_{l\neq k}\Sigma_{t\neq k}q_{lt}a_{jl}a_{jt}$ not depending on a_{jk}. This fact can be verified by the following example: for $m = 3$,

$$\sum_{k=1}^{3}\sum_{l=1}^{3} q_{kl}a_{jk}a_{jl} = \left(q_{11}a_{j1}^2 + q_{12}a_{j1}a_{j2} + q_{13}a_{j1}a_{j3}\right)$$
$$+ \left(q_{21}a_{j2}a_{j1} + q_{22}a_{j2}^2 + q_{23}a_{j2}a_{j3}\right)$$
$$+ \left(q_{31}a_{j3}a_{j1} + q_{32}a_{j3}a_{j2} + q_{33}a_{j3}^2\right)$$
$$= q_{22}a_{j2}^2 + \left(q_{12}a_{j1}a_{j2} + q_{21}a_{j2}a_{j1} + q_{23}a_{j2}a_{j3} + q_{32}a_{j3}a_{j2}\right) + c_{[j,2]}$$

with $c_{[j,2]} = q_{11}a_{j1}^2 + q_{33}a_{j3}^2 + q_{13}a_{j1}a_{j3} + q_{31}a_{j3}a_{j1}$ not depending on a_{j2}. We can use $q_{kl} = q_{lk}$ to rewrite the above equalities as

$$\sum_{k=1}^{3}\sum_{l=1}^{3} q_{kl}a_{jk}a_{jl} = q_{22}a_{j2}^2 + 2\left(q_{12}a_{j1}a_{j2} + q_{23}a_{j2}a_{j3}\right) + c_{[j,2]}$$

Using the fact in this note, (22.13) can be rewritten as a function of an element a_{jk} in $\mathbf{A}$:

$$g(a_{ik}) = -\frac{q_{kk}}{\psi_j} h\left(a_{jk}\right) + c_{[j,k]}. \tag{22.14}$$

Here, $c_{[j,k]}$ does not depend on a_{jk}, and

$$\begin{aligned} h(a_{ik}) &= \frac{1}{2q_{kk}} \left(-2a_{jk} b_{jk} + q_{kk} a_{jk}^2 + 2 \sum_{l \neq k} q_{kl} a_{jk} a_{jl} \right) + \frac{\psi_j}{q_{kk}} wMC(a_{jk}; w, u) \\ &= \frac{1}{2} \left\{ a_{jk}^2 - \frac{2}{q_{kk}} \left(a_{jk} b_{jk} - a_{jk} \sum_{l \neq k} q_{kl} a_{jl} \right) \right\} + \frac{\psi_j}{q_{kk}} wMC(a_{jk}; w, u) \\ &= \frac{1}{2} \left\{ a_{jk} - \frac{1}{q_{kk}} \left(b_{jk} - \sum_{l \neq k} q_{kl} a_{jl} \right) \right\}^2 - c^* + \frac{\psi_j}{q_{kk}} wMC(a_{jk}; w, u), \\ &= \frac{1}{2} \left\{ a_{jk} - r_{jk}(\mathbf{a}_{j[k]}) \right\}^2 - c^* + \frac{\psi_j}{q_{kk}} wMC(a_{jk}; w, u) \end{aligned} \tag{22.15}$$

with c^* irrelevant to a_{jk} and $r_{jk}\left(\mathbf{a}_{j[k]}\right) = \left(b_{jk} - \Sigma_{l \neq k} q_{kl} a_{jl}\right) / q_{kk}$ being a function of the $(m-1) \times 1$ vector $\mathbf{a}_{j[k]}$ containing $a_{j1}, \ldots, a_{jm}$ except a_{jk}. We suppose $\psi_j > 0$ and the positive-definiteness of $\mathbf{Q}$ implying $q_{kk} = \mathbf{w}_k' \mathbf{Q} \mathbf{w}_k > 0$ with $\mathbf{w}_k$ $(m \times 1)$ containing zeros except for the kth element being one (Note 8.2). Then, we can obtain the loading matrix $\mathbf{A}$ that increases (22.13) by performing the minimization of (22.15) over a_{jk} for $j = 1, \ldots, p$ and $k = 1, \ldots, m$. Using $w^* = w\psi_j / q_{kk}$ and $u^* = q_{kk} u / \psi_j$, the minimizer of (22.15) can be given by

$$a_{jk} = \begin{cases} \dfrac{\text{sign}[r_{jk}(\mathbf{a}_{j[k]})]\left(\left|r_{jk}(\mathbf{a}_{j[k]})\right| - w^*\right)_+}{1 - 1/u^*} & \text{if } \left|r_{jk}(\mathbf{a}_{j[k]})\right| \leq w^* u^* \\ r_{jk}(\mathbf{a}_{j[k]}) & \text{otherwise} \end{cases} \tag{22.16}$$

if $u^* \geq 1$; otherwise,

$$a_{jk} \text{ being the minimizer of (22.15) among } a_{jk} = 0,\ a_{jk} = r_{jk}\left(\mathbf{a}_{j[k]}\right), \text{ and } a_{jk} = \text{sign}\left[r_{jk}(\mathbf{a}_{j[k]})\right] w^* u^*. \tag{22.17}$$

Here, sign[y] and $(y)_+$ is defined for a real value y as follows: sign[y] = 1 and $(y)_+ = y$ if $y > 0$, sign[y] = −1 and $(y)_+ = 0$ if $y < 0$, and sign[y] = $(y)_+ = 0$ if $y = 0$. In this book, it is too involved to describe how (22.16) and (22.17) can be derived from (22.15). The derivation of (22.16) is explained in Zhang (2010), and that of (22.17) is found in Hirose, Ogura, and Shomodaira (2015).

Next, we consider updating the diagonal elements of $\boldsymbol{\Psi}$ so as to maximize (22.10) for given $\mathbf{A}$ and $\boldsymbol{\Phi}$. As explained in Appendix A.9.5, the update formula of those elements is given by (A.9.24), i.e.,

$$\psi_j = v_{ij} - 2\,\mathbf{b}_j'\mathbf{a}_j + \mathbf{a}_j'\mathbf{Q}\mathbf{a}_j \ (j = 1, \ldots, p) \tag{22.18}$$

Finally, let us consider updating $\boldsymbol{\Phi}$ so as to increase (22.10) for given $\mathbf{A}$ and $\boldsymbol{\Psi}$. This increment is attained by decreasing

$$\eta(\boldsymbol{\Phi}) = \log|\boldsymbol{\Phi}| + \mathrm{tr}\boldsymbol{\Phi}^{-1}\mathbf{Q}, \tag{22.19}$$

since its multiplication by $-n/2$ is only relevant to $\boldsymbol{\Phi}$ on the right side of (22.10). For decreasing (22.19), we use a procedure different from Hirose and Yamamoto (2014): $\boldsymbol{\Phi}$ is reparameterized using a $m \times m$ matrix $\mathbf{R}$ as

$$\boldsymbol{\Phi} = \mathrm{diag}(\mathbf{R}'\mathbf{R})^{-1/2}\mathbf{R}'\mathbf{R}\,\mathrm{diag}(\mathbf{R}'\mathbf{R})^{-1/2} \tag{22.20}$$

so that $\boldsymbol{\Phi}$ is a *correlation matrix*, i.e., a symmetric nonnegative definite matrix whose diagonal elements are ones. For updating $\mathbf{R} = (r_{kl})$, we use a *gradient algorithm* illustrated in Appendix A.6.3: This is iterated to update $\mathbf{R}$ to $\mathbf{R}_{\mathrm{new}}$ as

$$\mathbf{R}_{\mathrm{new}} = \mathbf{R} - s\frac{\partial\eta(\boldsymbol{\Phi})}{\partial\mathbf{R}}. \tag{22.21}$$

Here, $\partial\eta(\boldsymbol{\Phi})/\partial\mathbf{R}$ is the $m \times m$ matrix whose (k, l) element is $d\eta(\boldsymbol{\Phi})/dr_{kl}$, and s is a positive value that guarantees $\eta(\boldsymbol{\Phi}) \geq \boldsymbol{\eta}(\boldsymbol{\Phi}_{\mathrm{new}})$, with $\boldsymbol{\Phi}_{\mathrm{new}}$ the correlation matrix obtained by substituting (22.21) into $\mathbf{R}$ in (22.20). We obtain $d\eta(\boldsymbol{\Phi})/dr_{kl}$ numerically, i.e., through *numerical differentiation*, as

$$\frac{d\eta(\boldsymbol{\Phi})}{dr_{jk}} = \frac{1}{2\Delta}[(\boldsymbol{\Phi}|r_{kl} := r_{kl} + \Delta) - \eta(\boldsymbol{\Phi}|r_{kl} := r_{kl} - \Delta)]. \tag{22.22}$$

Here, Δ is a small positive value, and $\eta\left(\boldsymbol{\Phi}|r_{kl} := r_{kl}^*\right)$ denotes the (22.19) value following from the substitution of r_{kl}^* into r_{kl} with the other elements of $\mathbf{R}$ kept fixed. We use $\Delta = 0.01$ for the computations in this chapter. Thus, $\boldsymbol{\Phi}$ is updated through the following steps:

[1] Set $i_{\mathrm{R}} = 0$.
[2] Set $s = 1$ and $i_{\mathrm{s}} = 0$.
[3] Obtain (22.21) to evaluate $\eta(\boldsymbol{\Phi})$ and $\eta(\boldsymbol{\Phi}_{\mathrm{new}})$.
[4] Set $i_{\mathrm{s}} := i_{\mathrm{s}} + 1$. If $i_{\mathrm{s}} = 20$, go to [6]. If $i_{\mathrm{s}} < 20$ and $\eta(\boldsymbol{\Phi}) \geq \eta(\boldsymbol{\Phi}_{\mathrm{new}})$, go to [5]. Otherwise, set s: = $s/2$ and go back to [3].
[5] Set $i_{\mathrm{R}} := i_{\mathrm{R}} + 1$ and set $\mathbf{R} = \mathbf{R}_{\mathrm{new}}$ to provide (22.20).
[6] Finish, if $i_{\mathrm{s}} = 20$ or $i_{\mathrm{R}} = 20$ or $\eta(\boldsymbol{\Phi}) - \eta(\boldsymbol{\Phi}_{\mathrm{new}}) \leq 0.1^6$; otherwise, go back to [2].

In conclusion, the M-step is summarized as follows: For all j and k, a_{jk} is updated through (22.16) if $u^* \geq 1$ and updated through (22.17) otherwise, after $\mathbf{a}_{j[k]}$ is obtained. Next, ψ_j, $j = 1, \ldots, p$, are updated through (22.18). Finally, $\mathbf{\Phi}$ is updated through the above [1]–[6].

22.5 Using Penalized Sparse LVFA

The PS-LVFA algorithm in the last sections provides the optimal $\{\mathbf{A}, \mathbf{\Psi}, \mathbf{\Phi}\}$ for a given $[w, u]$: The resulting $\{\mathbf{A}, \mathbf{\Psi}, \mathbf{\Phi}\}$ depends on the $[w, u]$ value. In order to select a suitable $[w, u]$, Yamamoto and Hirose (2014) proposed to compare the values of BIC (8.25) for the solutions following certain $[w, u]$. BIC for PS-LVFA can be defined as

$$\mathrm{BIC} = -2l^*(\mathbf{A}, \mathbf{\Psi}, \mathbf{\Phi}) + \eta(w, u) \log n, \tag{22.23}$$

Here, $\eta(w, u) = \mathrm{Card}(\mathbf{A}; w, u) + p + m(m-1)/2$ is the number of the parameters whose values are estimated, with $\mathrm{Card}(\mathbf{A}; w, u)$ the cardinality of the resulting $\mathbf{A}$ which is a function of $[w, u]$, p the number of unique variances, and $m(m-1)/2$ that of factor correlations. The rational for this BIC-based selection of $\{w, u\}$ is explained in Zou, Hastie, and Tibshirani (2007). We also use this procedure.

For illustration, we performed PS-LVFA for the correlation matrix processed by Yanai and Ichikawa (Table 19.2), setting $m = 3$ following their approach. Here, the 50×50 combinations of w and u values were considered, with $w = 1/30, 2/30, \ldots, 50/30$, while $u = 1, 1 + 1/25, 1 + 2/25, \ldots, 1 + 49/25$. The solution for $w = 0.167$ and $u = 1.4$ gave the least BIC and is presented in Table 22.1 with the blank cells indicating estimates of zero. By noting the nonzero loadings in the table, we can interpret the three factors reasonably, as Yanai and Ichikawa (2007, p. 291) did for their solution resulting in the exploratory FA followed by rotation. The first, second, and third common factors (i.e., columns) in $\mathbf{\Lambda}$ can be interpreted as standing for *emotional instability*, *extraversion-general activity*, and *consciousness-agreeableness*, respectively. The factor correlations in Table 22.1 show that *consciousness-agreeableness* is slightly positively correlated to the other two factors though these two have a slightly negative correlation.

We also illustrate PS-LVFA using the correlation matrix in Table 22.2. Performing an FA procedure for this matrix can be considered an application of *Thurstone's* (1947) *box problem*, as explained next:

Note 22.2. Thurstone's Box Problem
In this problem, Thurstone (1947) tried to generate a data set whose variable j (= 1, …, 20) is defined as a function of the scores in common factor vector $\mathbf{f} = [x, y, z]'$. Let the function be expressed as $v_j(x, y, z)$. This is defined as in the "variable" column of Table 22.2: For example, $v_4(x, y, z) = xy$ for the

Table 22.1 PS-LVFA solution for the correlations in Table 19.2

Variable	**Λ**			**Ψ**
Extraversion		−0.41	0.54	0.60
Activity	0.30	−0.33	0.64	0.34
Empathy	0.52			0.74
Novelty			0.61	0.63
Durability	0.69			0.52
Regularity	0.79			0.37
Self-revelation			0.64	0.60
Aggressiveness		0.35	0.46	0.62
Lack of cooperativeness		0.47		0.78
Inferiority feeling		0.67	−0.35	0.49
Nervousness	0.30	0.70		0.49
Depression		0.84		0.30
Factor	**Φ**			
1	1.00			
2	−0.15	1.00		
3	0.20	0.15	1.00	

fourth variable. A data set generated with $v_j(x, y, z)$ is known as *Thurstone's box data*, as he used the heights, widths, and lengths of boxes for x, y, and z.

Thurstone considered the *ideal solution* for the box data as the one in which variables load the factor(s) used for defining the variables: For example, the fourth variable should ideally load x and y (as in Table 22.3). Thus, an FA procedure providing such a solution is regarded as promising. For this reason, performing an FA procedure for the box data has been called *Thurstone's box problem*. To date, this problem has often been used as a *cornerstone* for testing new FA procedures.

The correlations in Table 22.2 were obtained from the 400 × 20 matrix of box data, which Adachi and Trendafilov (2015) synthesized as follows: The jth variable is given by $v_j(x, y, z) + e_j$. Here, x, y, and z are generated using a random number which follows the uniform distribution for the interval [1, 10] (with its probability density being equal over the real values within the interval), and $[e_1, \ldots, e_{20}]$ follows $N_{20}(\mathbf{0}_{20}, 0.1\mathbf{I}_{20})$. See Adachi and Trendafilov (2015) for details.

In this box problem, the w and u values were selected as in the last example. As a result, $w = 0.4$ and $u = 1.04$ led to the solution with the least BIC, which is shown in Table 22.3. Here, we can see that the solution is *ideal* as explained in the above note, which demonstrates that Hirose and Yamamoto's (2014) PS-LVFA is a promising procedure.

Table 22.2 Adachi and Trendafilov's (2015) correlations for their generated box data

Variable		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	x^2	1.00																			
2	y^2	–0.05	1.00																		
3	z^2	–0.02	0.00	1.00																	
4	xy	0.62	0.57	–0.02	1.00																
5	xz	0.59	–0.03	0.60	0.42	1.00															
6	yz	–0.04	0.64	0.59	0.40	0.41	1.00														
7	$(x^2 + y^2)^{1/2}$	0.63	0.61	–0.02	0.82	0.42	0.40	1.00													
8	$(x^2 + z^2)^{1/2}$	0.65	0.01	0.61	0.47	0.81	0.40	0.46	1.00												
9	$(y^2 + z^2)^{1/2}$	–0.03	0.62	0.63	0.40	0.42	0.83	0.41	0.44	1.00											
10	$2x + 2y$	0.62	0.63	0.02	0.87	0.44	0.45	0.90	0.48	0.45	1.00										
11	$2x + 2z$	0.63	–0.02	0.63	0.44	0.86	0.42	0.45	0.89	0.45	0.46	1.00									
12	$2y + 2z$	–0.04	0.62	0.63	0.41	0.40	0.86	0.40	0.44	0.90	0.45	0.45	1.00								
13	$\log x$	0.82	–0.06	0.01	0.57	0.60	–0.03	0.59	0.62	–0.02	0.59	0.64	–0.04	1.00							
14	$\log y$	–0.05	0.80	0.02	0.57	0.00	0.58	0.57	0.02	0.60	0.59	0.00	0.63	–0.05	1.00						
15	$\log z$	–0.02	0.01	0.79	0.01	0.58	0.60	–0.02	0.55	0.62	0.02	0.61	0.64	0.01	0.02	1.00					
16	xyz	0.41	0.47	0.50	0.66	0.72	0.73	0.61	0.64	0.65	0.65	0.68	0.67	0.43	0.44	0.50	1.00				
17	$(x^2 + y^2 + z^2)^{1/2}$	0.52	0.50	0.51	0.68	0.67	0.64	0.74	0.76	0.72	0.75	0.74	0.71	0.48	0.46	0.48	0.77	1.00			
18	e^x	0.74	0.01	–0.06	0.48	0.41	–0.05	0.53	0.47	–0.04	0.50	0.42	–0.01	0.52	0.03	–0.07	0.28	0.43	1.00		
19	e^y	–0.07	0.72	0.06	0.35	–0.02	0.51	0.43	0.00	0.50	0.41	–0.03	0.45	–0.07	0.49	0.05	0.34	0.37	–0.05	1.00	
20	e^z	0.01	0.04	0.76	0.02	0.41	0.48	0.01	0.48	0.50	0.04	0.46	0.50	–0.01	0.07	0.51	0.40	0.43	–0.03	0.08	1.00

Table 22.3 PS-LVFA solution for the correlations in Table 22.2

Variable	$\boldsymbol{\Lambda}$			$\boldsymbol{\Psi}$
x^2	0.95			0.10
y^2		0.95		0.09
z^2			0.94	0.12
xy	0.67	0.63		0.18
xz	0.64		0.64	0.20
yz		0.64	0.62	0.17
$(x^2 + y^2)^{1/2}$	0.70	0.67		0.11
$(x^2 + z^2)^{1/2}$	0.68		0.65	0.13
$(y^2 + z^2)^{1/2}$		0.64	0.67	0.12
$2x + 2y$	0.69	0.69		0.08
$2x + 2z$	0.68		0.69	0.08
$2y + 2z$		0.65	0.68	0.10
$\log x$	0.89			0.21
$\log y$		0.87		0.24
$\log z$			0.89	0.22
xyz	0.47	0.49	0.54	0.25
$(x^2 + y^2 + z^2)^{1/2}$	0.58	0.53	0.54	0.11
e^x	0.72			0.48
e^y		0.70		0.52
e^z			0.71	0.49
Factor	$\boldsymbol{\Phi}$			
1	1.00	−0.05	−0.01	
2	−0.05	1.00	0.03	
3	−0.01	0.03	1.00	

22.6 Formulation of Cardinality Constrained MDFA

In this section, we introduce an SFA procedure which features the following properties: [1] it is formulated with the *matrix decomposition approach* in Chap. 18, [2] a *penalty function is not used*, and [3] Card(**A**) is *specified in advance*. From the properties [1] and [3], the procedure can be referred to as *cardinality constrained MDFA* (*CC-MDFA*).

The property [3] implies that the *cardinality* of **A** is *constrained* in CC-MDFA a priori as

$$\mathrm{Card}(\mathbf{A}) = c \tag{22.24}$$

with c a specified integer. The CC-MDFA loss function is the MDFA one (18.5), i.e.,

$$f\left(\mathbf{F},\mathbf{U},\mathbf{A},\boldsymbol{\Psi}^{1/2}\right)=f(\mathbf{Z},\mathbf{B})=\left\|\mathbf{X}-\left(\mathbf{F}\mathbf{A}+\mathbf{U}\boldsymbol{\Psi}^{1/2}\right)\right\|^2 \\ =\left\|\mathbf{X}-[\mathbf{F},\mathbf{U}]\left[\mathbf{A},\boldsymbol{\Psi}^{1/2}\right]'\right\|^2=\|\mathbf{X}-\mathbf{Z}\mathbf{B}'\|^2. \tag{22.25}$$

Here, $\mathbf{Z}=[\mathbf{F},\ \mathbf{U}]$ and $\mathbf{B}=\left[\mathbf{A},\boldsymbol{\Psi}^{1/2}\right]$ are $n\times(m+p)$ and $p\times(m+p)$ block matrices, with $\mathbf{F}$ and $\mathbf{U}$ containing common and unique factor scores, respectively. The factor score matrix $\mathbf{Z}=[\mathbf{F},\ \mathbf{U}]$ is constrained according to (18.6) and (18.7), i.e.,

$$\mathbf{1}_n'\mathbf{Z}=\mathbf{0}_{m+p}, \tag{22.26}$$

$$\frac{1}{n}\mathbf{Z}'\mathbf{Z}=\mathbf{I}_{m+p}. \tag{22.27}$$

In CC-MDFA, (22.25) is minimized over $\mathbf{Z}$ and $\mathbf{B}$ subject to (22.24), (22.26), and (22.27).
The cardinality constraint (22.24) allows the final phrase in (22.5) to be rewritten as in

Optimally and jointly estimate
[A] what loadings in $\mathbf{A}$ are to be zero and
[B]the values of nonzero parameters,
for a pre-specified Card($\mathbf{A}$). (22.28)

A *key point* in CC-MDFA is that the loss function (22.25) can decomposed using $\mathbf{S}_{\mathrm{XZ}}=n^{-1}\mathbf{X}'\mathbf{Z}$ as (18.24) i.e.,

$$f(\mathbf{S}_{\mathrm{XZ}},\mathbf{B})=\left\|\mathbf{X}-\mathbf{Z}\mathbf{S}_{\mathrm{XZ}}'\right\|^2+n\|\mathbf{S}_{\mathrm{XZ}}-\mathbf{B}\|^2, \tag{22.29}$$

where the right term $\|\mathbf{S}_{\mathrm{XZ}}-\mathbf{B}\|^2$ can be rewritten as (18.25), i.e.,

$$\|\mathbf{S}_{\mathrm{XZ}}-\mathbf{B}\|^2=\|\mathbf{S}_{\mathrm{XF}}-\mathbf{A}\|^2+\left\|\mathrm{diag}(\mathbf{S}_{\mathrm{XU}})-\boldsymbol{\Psi}^2\right\|^2+\|\mathbf{S}_{\mathrm{XU}}-\mathrm{diag}(\mathbf{S}_{\mathrm{XU}})\|^2. \tag{22.30}$$

This implies that (22.25) can be decomposed so that the part dependent on $\mathbf{A}$ is just a *simple function*

$$g(\mathbf{A})=\|\mathbf{S}_{\mathrm{XF}}-\mathbf{A}\|^2. \tag{22.31}$$

This property allows the CC-MDFA algorithm to be formed with a minor modification of the MDFA one (Chap. 18), as written in the next section.

The decomposition in (22.29) follows from the constraint (22.27): The key point in CC-MDFA, i.e., (22.29), no longer holds, without the constraint (22.27) which implies the common factors being mutually uncorrelated with $n^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_m$ as found in (18.4). Hence, the common factors must be mutually uncorrelated, i.e., *factor correlations cannot be estimated* in CC-MDFA. This is a *limitation* of CC-MDFA, which contrasts to PS-LVFA in which factor correlations can be estimated.

22.7 Algorithm for Cardinality Constrained MDFA

In CC-MDFA, only a single constraint (22.24) for $\mathbf{A}$ is added to the MDFA constraints (22.26) and (22.27). Thus, the procedures for estimating parameters excluding $\mathbf{A}$ is the same as listed in Note 18.1 (Chap. 18).

In the loss function (22.25) or (22.29) with (22.30), *only* (22.31) depends on the loading matrix $\mathbf{A}$. Thus, for finding the optimal update formula for $\mathbf{A}$, we may consider minimizing (22.31) over $\mathbf{A}$ subject to (22.24) with the other parameters fixed. This formula can be derived from the fact that (22.31) can be rewritten using $\mathbf{A} = (a_{jk})$ and $\mathbf{S}_{\mathrm{XF}} = (s_{jk})$ so that

$$g(\mathbf{A}) = \sum_{(j,k)\in\aleph_0} s_{jk}^2 + \sum_{(j,k)\in\aleph_\#} (a_{jk} - s_{jk})^2 \geq \sum_{(j,k)\in\aleph_0} s_{jk}^2. \tag{22.32}$$

Here, $\aleph_\#$ denotes the set of the c pairs of (j, k) for the loadings a_{jk} to be nonzero, $\aleph_0$ is the set of the $pm-c$ pairs of (j, k) for the a_{jk} to be zero, $\sum_{(j,k)\in\aleph_0} s_{jk}^2$ stands for the summation of s_{jk}^2 over the (j, k) contained in $\aleph_0$, and we have used $\sum_{(j,k)\in\aleph_0} (a_{jk} - s_{jk})^2 = \sum_{(j,k)\in\aleph_0} (0 - s_{jk})^2 = \sum_{(j,k)\in\aleph_0} s_{jk}^2$. The inequality in (22.32) shows that $g(\mathbf{A})$ attains its *lower limit* $\sum_{(j,k)\in\aleph_0} s_{jk}^2$, when the loading a_{jk} with $(j, k) \in \aleph_\#$ is set equal to s_{jk} so that $\sum_{(j,k)\in\aleph_\#} (a_{jk} - s_{jk})^2 = \sum_{(j,k)\in\aleph_\#} (s_{jk} - s_{jk})^2 = 0$. Furthermore, the *limit* $\sum_{(j,k)\in\aleph_0} s_{jk}^2$ is *minimal*, when $\aleph_0$ contains (j, k) for the $pm - c$ smallest s_{jk}^2 among all elements of $\mathbf{S}_{\mathrm{XF}} \odot \mathbf{S}_{\mathrm{XF}} = (s_{jk}^2)$ or equivalently, when $\aleph_\#$ contains (j, k) for the c largest s_{jk}^2. The update formula of $\mathbf{A} = (a_{jk})$ is thus given by

$$a_{jk} = \begin{cases} 0 & \text{if } s_{jk}^2 < s_{[c]}^2 \\ s_{jk} & \text{otherwise} \end{cases} \tag{22.33}$$

with $S_{[c]}^2$ the cth largest value among all elements of $\mathbf{S}_{\mathrm{XF}} \odot \mathbf{S}_{\mathrm{XF}}$.

We can find (22.33) to imply that $a_{jk}^2 = a_{jk}s_{jk} = 0$ if $s_{jk}^2 < s_{[c]}^2$ and $a_{jk}^2 = a_{jk}s_{jk} = s_{jk}^2$ otherwise, that is, $a_{jk}^2 = a_{jk}s_{jk}$ for every (j, k): $\mathbf{A} \odot \mathbf{A} = \mathbf{A} \odot \mathbf{S}_{\mathrm{XF}}$. This fact shows that (18.26) holds true also in the CC-MDFA solution, and the use of (18.26) in (18.18) leads to the standardized loss function value in CC-MDFA being given by (18.17), as this is derived in Sect. 18.6. Hence, the CC-MDFA algorithm can be stated by Note 18.1, with the replacement of Step 4 by "Update $\mathbf{A} = \left(a_{jk}\right)$ and $\boldsymbol{\Psi}$ as (22.33) and $\boldsymbol{\Psi}^{1/2}$ = diag($\mathbf{S}_{\mathrm{XU}}$), respectively". In the algorithm, $\mathbf{A}$ and $\boldsymbol{\Psi}$ are initialized as described in Sect. 22.3, and the convergence is defined as the difference of the (18.17) value from the previous round being less than 0.1^8.

22.8 Using Cardinality Constrained MDFA

The CC-MDFA algorithm in the last subsection provides the optimal $\mathbf{A}$ and $\boldsymbol{\Psi}$ for a given c in (22.24). Thus, CC-MDFA is *convenient* for users who wish to *pre-specify the cardinality of loadings*, for example, who wish to obtain a solution with half of the loadings being zero. Such pre-specification of the cardinality cannot be made in PS-LVFA.

Selecting a suitable c is also possible in Adachi and Trendafilov's (2015) procedure based on the following observation. The authors found that the CC-MDFA solutions are broadly *equivalent* to the solutions of the *likelihood-based CFA*, in which the log likelihood (22.2) is maximized subject to $\boldsymbol{\Phi} = \mathbf{I}_m$ with the locations of the zero loadings in $\mathbf{A}$ constrained to those in the CC-MDFA solutions. This suggests that *BIC* (8.25) used for CFA *can also be utilized* in CC-MDFA. The BIC value is obtained with

$$\mathrm{BIC} = -2l^*(\mathbf{A}, \boldsymbol{\Psi}, \mathbf{I}_m) + \eta(c) \log n. \tag{22.34}$$

Table 22.4 CC-MDFA solution for the correlations in Table 19.2

Variable	$\boldsymbol{\Lambda}$			$\boldsymbol{\Psi}$
Extraversion	0.24	−0.35	0.47	0.59
Activity	0.45	−0.28	0.63	0.32
Empathy	0.54			0.67
Novelty			0.64	0.58
Durability	0.67			0.54
Regularity	0.79			0.37
Self-revelation			0.62	0.59
Aggressiveness		0.41	0.51	0.54
Lack of cooperativeness		0.47		0.74
Inferiority feeling		0.62	−0.31	0.49
Nervousness	0.25	0.66		0.49
Depression		0.84		0.29

Table 22.5 CC-MDFA solution for the correlations in Table 22.2

Variable	$\mathbf{\Lambda}$			$\mathbf{\Psi}$
x^2	0.95			0.08
y^2		0.96		0.08
z^2			0.94	0.09
xy	0.67	0.61		0.17
xz	0.64		0.64	0.17
yz		0.66	0.63	0.15
$(x^2 + y^2)^{1/2}$	0.69	0.64		0.10
$(x^2 + z^2)^{1/2}$	0.68		0.64	0.12
$(y^2 + z^2)^{1/2}$		0.66	0.67	0.11
$2x + 2y$	0.68	0.67		0.08
$2x + 2z$	0.67		0.68	0.08
$2y + 2z$		0.66	0.68	0.09
$\log x$	0.89			0.19
$\log y$		0.87		0.23
$\log z$			0.88	0.21
xyz	0.47	0.49	0.54	0.22
$(x^2 + y^2 + z^2)^{1/2}$	0.57	0.52	0.54	0.10
e^x	0.71			0.48
e^y		0.68		0.52
e^z			0.71	0.49

Here, $\eta(c) = c + p$ is the number of the parameters (the nonzero loadings in $\mathbf{A}$ and the diagonal elements of $\mathbf{\Psi}$) whose values are estimated, and $l^*(\mathbf{A}, \mathbf{\Psi}, \mathbf{I}_m)$ is the (22.2) value in which the CC-MDFA solution is substituted with $\mathbf{\Phi} = \mathbf{I}_m$. Thus, we can select the solution with the best c within possible c values. They are reasonably considered as

$$c = p, \ldots, pm - m(m-1)/2. \tag{22.35}$$

Here, the lower limit has been set to p, since this prevents $\mathbf{A}$ from having an empty column if c were smaller than the limit. On the other hand, the upper limit has been set to the number of loadings minus $m(m-1)/2$, since $m(m-1)/2$ loadings can be set to zeros, without loss of generality, as discussed in Sect. 12.9.

We performed CC-MDFA for each of the correlation matrices in Tables 19.2 and 22.2, where the solution with the best c was chosen using the above method. The resulting solutions are presented in Tables 22.4 and 22.5. They are found to be very similar to the corresponding PS-LVFA solutions: The loadings in Table 22.4

can be interpreted in the same manner as those in Table 22.1, and Table 22.5 demonstrates that CC-MDFA can solve Thurstone's box problem as well as PS-LVFA.

An *advantage* of CC-MDFA over PS-LVFA is that the solution with the best BIC can be *more easily selected* as the number of solutions to be compared is smaller according to (22.35): c is an *integer* within a *restricted range*. In contrast, the solutions in PS-LVFA vary across the two tuning parameters w and u, which take *continuous real values*. Thus, it is impossible to consider the solutions with all possible w and u. Additional work is required to choose candidate values for w and u, for example, w = 1/30, 2/30, …, 50/30, and u = 1, 1 + 1/25, 1 + 2/25, …, 1 + 49/25, as in Sect. 22.5. However, factor correlations can be estimated in PS-LVFA, which cannot be done in CC-MDFA.

22.9 Sparse FA Versus Factor Rotation in Exploratory FA

As described in Chaps. 12, 13, and 18, the *exploratory FA* (*EFA*) solution has *rotational indeterminacy*: If $\mathbf{A}$ is the optimal loading matrix which optimizes the EFA objective function, $\mathbf{A}_\mathrm{T} = \mathbf{AT}$ is also optimal in that the function value remains the same even if $\mathbf{A}$ is replaced by $\mathbf{A}_\mathrm{T}$. Here, $\mathbf{T}$ is an $m \times m$ rotation matrix satisfying either (13.3) or (13.9). Thus, a *rotation* procedure follows EFA in which the matrix $\mathbf{T}$ is obtained so that the resulting $\mathbf{A}_\mathrm{T} = \mathbf{AT}$ has a desirable property. This is typically *simple structure*, as explained in Chap. 13. How this simple structure is related to the *sparseness* is shown by [1] and Table 13.2 in Sect. 13.3: [1] shows that the sparseness is a feature of simple structure and Table 13.2(A) illustrates that *ideally* simple loadings are sparse.

Table 22.6 presents an example of $\mathbf{A}_\mathrm{T}$, which was obtained by varimax rotation following EFA for the data set in Table 19.2. In Table 22.6, the loadings of the large absolute values are boldfaced. By noting these values, we can make the same interpretation we did for the sparse FA (SFA) solutions, demonstrating that *rotation following EFA* is *comparable* to *SFA*. In the next three paragraphs, we discuss some types of *differences* between the two procedures.

In general, the loadings resulting from the rotation *cannot be exactly zero or ideally simple* as in Table 13.2(A), which differs from the loadings resulting in SFA. Thus, the loadings whose absolute values are greater than 0.45 have been boldfaced in Table 22.6, for the sake of easily capturing the loadings to be noted. However, the threshold 0.45 is a subjectively selected benchmark: An *objectively defined threshold does not exist* which distinguishes which loadings should be noted. Furthermore, only noting the loadings whose magnitudes exceed the threshold implies that the other loadings should be regarded as zeros. Let us

Table 22.6 LVFA solution followed by varimax rotation in Table 19.4, with the loadings boldfaced for absolute values ≥ 0.45

Variable	$\mathbf{A}_T$			$\boldsymbol{\Psi}$
Extraversion	0.24	−0.33	**0.47**	0.61
Activity	0.42	−0.25	**0.65**	0.34
Empathy	**0.60**	−0.02	0.02	0.64
Novelty	0.04	−0.05	**0.62**	0.62
Durability	**0.66**	−0.08	0.05	0.55
Regularity	**0.71**	0.05	0.17	0.47
Self-revelation	0.03	0.16	**0.63**	0.58
Aggressiveness	−0.13	0.37	**0.55**	0.54
Lack of cooperativeness	−0.23	**0.45**	0.17	0.72
Inferiority feeling	−0.18	**0.62**	−0.30	0.49
Nervousness	0.26	**0.72**	0.02	0.41
Depression	−0.10	**0.83**	0.02	0.30

imagine that the loading matrix in Table 22.6 whose loadings are not boldfaced is replaced by zeros. Obviously, such a matrix is *not the optimally estimated solution*. In contrast, which elements are to be zero are *estimated optimally* in *SFA* (e.g., Trendafilov & Adachi, 2015).

Another difference of *rotation* from SFA is that the former *does not involve the original data*: As described in Chap. 13, the function Simp($\mathbf{A}_T$) or Comp($\mathbf{A}_T$) is optimized without a data set. This allows us to consider whether the loading matrix $\mathbf{A}_T = \mathbf{AT}$ resulting in the rotation might embody some simple structure not underlying the original data. However, $\mathbf{A}_T = \mathbf{AT}$ for any rotation matrix $\mathbf{T}$ is an *EFA solution optimally fitted to data*, as described in Chaps. 12 and 18. That is, the rotation, in which Simp($\mathbf{A}_T$) is maximized or Comp($\mathbf{A}_T$) is minimized, can be regarded as choosing *the simplest* $\mathbf{A}_T = \mathbf{AT}$ *among elements in a set of the EFA solutions* $\{\mathbf{A}_T = \mathbf{AT}: \mathbf{T} = \text{any rotation matrix}\}$.

Though the properties for the rotation following EFA that have been so far described might be considered slightly disadvantageous, some *advantages* also exist. One follows from the fact that SFA is a constrained version of EFA with a penalty function such as (22.7) or constraint (22.24). This implies that the EFA solutions fit better than the SFA *ones* to a given data set. Thus, *the loading matrix* $\mathbf{A}_T$ resulting from the *rotation* is a *better fit* to the data set than its *SFA counterpart*. Another disadvantage of SFA is that [w, u] in (22.7) or c in (22.24) *needs to be chosen* among candidates. Such a cumbersome procedure is *not required* in rotation though instead a procedure must be chosen among a variety of the rotation procedures (Browne, 2001).

22.10 Bibliographical Notes

Besides Hirose and Yamamoto (2014) referenced in this chapter, Hirose and Yamamoto (2015), Trendafilov, Fontanella, and Adachi (2017), and Jin, Moustaki, and Yang-Wallentin (2018) have proposed the procedures included in *PS-LVFA*. To the best of our knowledge, only Adachi and Trendafilov (2015) have proposed a procedure which can be considered as *CC-MDFA*. Its restrictive version called spars*est* FA has been presented by Adach and Trendafilov (2018b), in which each row of a loading matrix is constrained to zeros except for a single element, i.e., each *variable* is allowed to *load a single factor*, which implies that the resulting loading matrix can be said to be the spars*est*. The same constraint is imposed in Vichi's (2017) *disjoint FA*.

Here, we must mention that the developments in sparse FA followed on from *sparse principal component analysis* (*SPCA*) for obtaining the sparse weight matrix (Jolliffe et al. 2003; Shen & Huang, 2008; Zou et al. 2006). Sparse PCA procedures are summarized well in Hastie et al. (2015) and Trendafilov (2014). Adachi and Trendafilov (2016) have proposed a *sparse PCA* version of *CC-MDFA* for obtaining the sparse component loading matrix, and its *sparse three-way PCA* version has been presented by Ikemoto and Adachi (2016).

Exercises

22.1 Summarize in what ways *sparse FA* is superior to *confirmatory FA*.

22.2 Summarize the similarities/differences between *sparse FA* and *EFA followed by rotation*.

22.3 Let us consider PS-LVFA in which the L_1-norm of $\mathbf{A}$ (Note 21.1) is used for Pen($\mathbf{A}$) in (22.6), i.e., maximizing

$$f_{L1}(\mathbf{A}, \boldsymbol{\Psi}, \boldsymbol{\Phi}) = \frac{n}{2}\log\left|(\mathbf{A}\boldsymbol{\Phi}\mathbf{A}' + \boldsymbol{\Psi})^{-1}\mathbf{V}\right| - \frac{n}{2}\mathrm{tr}(\mathbf{A}\boldsymbol{\Phi}\mathbf{A}' + \boldsymbol{\Psi})^{-1}\mathbf{V} - nw\|\mathbf{A}\|_1. \tag{22.36}$$

over $\mathbf{A}$,$\boldsymbol{\Psi}$, and $\boldsymbol{\Phi}$. Discuss how this maximization can be attained by the EM algorithm, i.e., the alternating iteration of obtaining $\mathbf{B}$ and $\mathbf{Q}$ in the function

$$\begin{aligned}\phi_{L1}(\mathbf{A}, \boldsymbol{\Psi}, \Phi) = &-\frac{n}{2}\log|\boldsymbol{\Psi}| - \frac{n}{2}\mathrm{tr}(\mathbf{V} - 2\mathbf{B}\mathbf{A}' + \mathbf{A}\mathbf{Q}\mathbf{A}')\boldsymbol{\Psi}^{-1} \\ &-\frac{n}{2}\left(\log|\boldsymbol{\Phi}| + \mathrm{tr}\boldsymbol{\Phi}^{-1}\mathbf{Q}\right) - nw\|\mathbf{A}\|_1.\end{aligned} \tag{22.37}$$

and maximizing it over $\mathbf{A}$,$\boldsymbol{\Psi}$, and $\boldsymbol{\Phi}$. Here, $\mathbf{B}$ and $\mathbf{Q}$ are defined as (22.11) with (22.12).

22.4 Discuss how the maximization of (22.37) can be attained by the procedures in Sect. 22.4 with only the update formula of $\mathbf{A}$ being modified.

22.5 Show that (22.37) can be rewritten as $\phi_{L1}(\mathbf{A},\boldsymbol{\Psi},\boldsymbol{\Phi}) = ng_{L1}(\mathbf{A}) + const_{L1[\mathrm{A}]}$. Here, $const_{L1[\mathrm{A}]}$ is a constant independent of $\mathbf{A}$, and

$$g_{L1}(\mathbf{A}) = -\frac{1}{2}\mathrm{tr}(\mathbf{V} - 2\mathbf{B}\mathbf{A}' + \mathbf{A}\mathbf{Q}\mathbf{A}')\mathbf{\Psi}^{-1} - \sum_{j=1}^{p}\sum_{k=1}^{m} w|a_{jk}|$$
$$= -\sum_{j=1}^{p}\frac{1}{2\psi_j}\left(v_{jj} - 2\sum_{k=1}^{m} a_{jk}b_{jk} + \sum_{k=1}^{m}\sum_{l=1}^{m} q_{kl}a_{jk}a_{jl}\right) - \sum_{j=1}^{p}\sum_{k=1}^{m} w|a_{jk}| \tag{22.38}$$

with $\mathbf{V} = (v_{jk}), \mathbf{A} = (a_{jk}), \mathbf{B} = (b_{jk}), \mathbf{Q} = (q_{kl})$, and ψ_j the jth diagonal element of $\mathbf{\Psi}$.

22.6 Show that (22.38) can be rewritten as a function of an element a_{jk} in $\mathbf{A}$: $g_{L1}(a_{jk}) = -(q_{kk}/\psi_j)h_{L1}(a_{jk}) + const_{L1[j,k]}$. Here, $const_{L1[j,k]}$ is independent of a_{jk}, and

$$h_{L1}(a_{jk}) = \frac{1}{2}\left\{a_{jk}^2 - \frac{2}{q_{kk}}\left(a_{jk}b_{jk} - a_{jk}\sum_{l\neq k} q_{kl}a_{jl}\right)\right\} + \frac{\psi_j}{q_{kk}}w|a_{jk}|$$
$$= \frac{1}{2}\left\{a_{jk}^2 - 2r_{jk}(\mathbf{a}_{j[k]})a_{jk} + 2d_{jk}(w)|a_{jk}|\right\}. \tag{22.39}$$

with $\mathbf{a}_{j[k]}$ the $(m-1)\times 1$ vector $\mathbf{a}_{j[k]}$ containing $a_{j1}, \ldots, a_{jm}$ except a_{jk}, $r_{jk}(\mathbf{a}_{j[k]}) = (b_{jk} - \Sigma_{l\neq k} q_{kl}a_{jl})/q_{kk}$ (a function of $\mathbf{a}_{j[k]}$), and $d_{jk}(w) = \psi_j w/q_{kk}$.

22.7 Show that minimizing (22.39) over a_{jk} can be attained for

$$a_{jk} = \begin{cases} 0 & \text{if } -d_{jk}(w) \le r_{jk}(\mathbf{a}_{j[k]}) \le d_{jk}(w) \\ \mathrm{sign}[r_{jk}(\mathbf{a}_{j[k]})]\{|r_{jk}(\mathbf{a}_{j[k]})| - d_{jk}(w)\} & \text{otherwise} \end{cases}. \tag{22.40}$$

Hints are found in (21.11)–(21.14).

22.8 Determine the algorithm for maximizing (22.36) over $\mathbf{A}$,$\mathbf{\Psi}$, and $\mathbf{\Phi}$, by considering answers for Exercises 22.3–22.7.

22.9 The definition of *local minima* is described in Exercise 7.6 (Chap. 7). CC-MDFA is known to be sensitive to local minima, as is *k-means clustering* (Chap. 7) for the same reason. This reason can be found in the similarity between the update formulas (7.19) and (22.33). Consider and discuss how these formulas are similar.

22.10 Discuss that the possibility of a CC-MDFA solution being a local minimizer can be *reduced* with the multi-run procedure described in Exercise 7.6 (Chap. 7).

22.11 Adach and Trendafilov (2018b) proposed the constrained FA procedure, in which each row of the loading matrix $\mathbf{A}$ is constrained to zeros except a single element. Discuss how this procedure is useful for *clustering variables*.

22.12 In Adachi and Trendafilov's (2016) procedure called *unpenalized sparse loading principal component analysis* (*USLPCA*), $\|\mathbf{X} - \mathbf{P}\mathbf{C}'\|^2$ is minimized over $\mathbf{P}$ $(n \times m)$ and $\mathbf{C}$ $(p \times m)$ subject to $n^{-1}\mathbf{P}'\mathbf{P} = \mathbf{I}_m$ and $\mathrm{Card}(\mathbf{C}) = l$ (pre-specified integer), with $m < \mathrm{rank}(\mathbf{X})$. Show that the function can be decomposed as

$$\|\mathbf{X} - \mathbf{P}\mathbf{C}'\|^2 = \left\|\mathbf{X} - \mathbf{P}\mathbf{S}'_{\mathrm{XP}}\right\|^2 + n\|\mathbf{S}_{\mathrm{XP}} - \mathbf{C}\|^2 \tag{22.41}$$

with $\mathbf{S}_{\mathrm{XP}} = n^{-1}\mathbf{X}'\mathbf{P}$.

22.13 Let us consider the minimization of (22.41) over $\mathbf{C} = (c_{jk})$ subject to $\mathrm{Card}(\mathbf{C}) = l$ with $\mathbf{P}$ kept fixed. Explain that the minimization can be attained for

$$c_{jk} = \begin{cases} 0 & \text{if } s_{jk}^{\mathrm{XP}} < s_{[l]}^2 \\ s_{jk}^{\mathrm{XP}} & \text{otherwise} \end{cases}, \tag{22.42}$$

where s_{jk}^{XP} is the (j, k) element of $\mathbf{S}_{\mathrm{XP}}$ and $S[l]^2$ is the lth largest value among all elements of $\mathbf{S}_{\mathrm{XP}} \odot \mathbf{S}_{\mathrm{XP}} = \left(s_{jk}^2\right)$.

22.14 Show that (22.41) can be rewritten as $n\mathrm{tr}(\mathbf{V} - 2n^{-1}\mathbf{C}'\mathbf{X}'\mathbf{P} + \mathbf{C}\mathbf{C}')$ and minimized over $\mathbf{P}$ for

$$\mathbf{P} = \sqrt{n}\boldsymbol{\Gamma}\boldsymbol{\Xi}' = \mathbf{X}\mathbf{C}\boldsymbol{\Xi}\boldsymbol{\Delta}^{-1}\boldsymbol{\Xi}' \tag{22.43}$$

subject to $n^{-1}\mathbf{P}'\mathbf{P} = \mathbf{I}_m$ for a given $\mathbf{C}$. Here, $\mathbf{V} = n^{-1}\mathbf{X}'\mathbf{X}$, and the matrices $\boldsymbol{\Gamma}$, $\boldsymbol{\Delta}$, and $\boldsymbol{\Xi}$ are obtained through the singular value decomposition (SVD) of $n^{-1/2}\mathbf{X}\mathbf{C}$, which is defined as $n^{-1/2}\mathbf{X}\mathbf{C} = \boldsymbol{\Gamma}\boldsymbol{\Delta}\boldsymbol{\Xi}'$, with $\boldsymbol{\Gamma}'\boldsymbol{\Gamma} = \boldsymbol{\Xi}'\boldsymbol{\Xi} = \mathbf{I}_m$ and $\boldsymbol{\Delta}$ being a diagonal matrix whose diagonal elements are all positive.

22.15 Show that the loss function $\|\mathbf{X} - \mathbf{P}\mathbf{C}'\|^2$, in which the matrix $\mathbf{C} = (c_{jk})$ updated as (22.42) is substituted, can be rewritten as $\|\mathbf{X} - \mathbf{P}\mathbf{C}'\|^2 = (n\mathrm{tr}\mathbf{V}) \times \tau$ with

$$\tau = 1 - \frac{\mathrm{tr}\mathbf{C}\mathbf{C}'}{\mathrm{tr}\mathbf{V}}. \tag{22.44}$$

22.16 Show that an algorithm of *USLPCA* in Exercise 22.12 for obtaining the solution of $\mathbf{C}$ can be formed, using (22.42)–(22.44), as follows:

Step 1. Initialize $\mathbf{C}$.
Step 2. Perform the eigenvalue decomposition of $\mathbf{C}'\mathbf{V}\mathbf{C}$ defined as $\mathbf{C}'\mathbf{V}\mathbf{C} = \boldsymbol{\Xi}\boldsymbol{\Delta}^2\boldsymbol{\Xi}'$.
Step 3. Set $\mathbf{S}_{\mathrm{XP}} = \mathbf{V}\mathbf{C}\boldsymbol{\Xi}\boldsymbol{\Delta}^{-1}\boldsymbol{\Xi}'$.
Step 4. Update $\mathbf{C} = (c_{jk})$ as (22.42).

Step 5. Finish if the decrease in (22.44) from the previous round is small enough to be ignored; otherwise, go back to Step 2.

A hint is found in the fact that the substitution of (22.43) in $\mathbf{S}_{\mathrm{XP}} = n^{-1}\mathbf{X}'\mathbf{P}$ allows it to be rewritten as $\mathbf{S}_{\mathrm{XP}} = \mathbf{VC\Xi}\Delta^{-1}\mathbf{\Xi}'$.

22.17 Ikemoto and Adachi (2016) have proposed a *sparse version of the Tucker2* in Exercise 20.12. This is formulated as minimizing (20.62) over $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{H}$ subject to, $\mathbf{A}'\mathbf{A} = \mathbf{I}_P, \mathbf{B}'\mathbf{B} = \mathbf{I}_Q$, and Card($\mathbf{H}$) = c (pre-specified integer). Show that (20.62) can be decomposed as

$$f = \|\mathbf{X} - \mathbf{AH}(\mathbf{I}_K \otimes \mathbf{B}')\|^2 = \|\mathbf{X} - \mathbf{AY}(\mathbf{I}_K \otimes \mathbf{B}')\|^2 + \|\mathbf{Y} - \mathbf{H}\|^2, \quad (22.45)$$

with $\mathbf{Y} = \mathbf{A}'\mathbf{X}(\mathbf{I}_K \otimes \mathbf{B})$.

22.18 Explain that for given $\mathbf{A}$ and $\mathbf{B}$, (22.45) is minimized over $\mathbf{H} = (h_{pq})$ subject to Card($\mathbf{H}$) = c, when $h_{pq} = 0$ if $y_{pq}^2 < y_{[c]}^2$; otherwise $h_{pq} = y_{pq}$. Here, $\mathbf{Y} = (y_{pq})$ and $y_{[c]}^2$ is the cth largest value among all elements of $\mathbf{Y} \odot \mathbf{Y} = \left(y_{pq}^2\right)$.

22.19 Discuss how the update formula of h_{pq} in Exercise 22.20 is similar to the Formula (20.50) in the *three-way simplimax* rotation (Kiers, 1998a).

Appendices

The fundamentals of matrix algebra and computations for multivariate data analysis, which had not been treated in the main chapters of this book, are described in Appendices A.1–A.4. That is followed by supplements for Chaps. 8 and 15 in Appendix A.5. Iterative algorithms are summarized, and a gradient method for them is illustrated in Appendix A.6. The scale invariance of covariance structure analysis (Chaps. 9–12) is treated in Appendix A.7. That is followed by Appendix A.8 in which probability densities and expected values are detailed together with the principle of EM algorithm. This appendix serves as a preparation for Appendix A.9. Here, the EM algorithm for factor analysis is detailed which is used for the procedures treated in Chaps. 10, 12, and 22.

A.1 Geometric Understanding of Matrices and Vectors

In this appendix, fundamental properties of the vectors and matrices are described, which are considered as geometric concepts.

A.1.1 Angles Between Vectors

Vectors can be depicted as lines (with arrows) as in Fig. A.1. There, we find the triangle formed by $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{a} - \mathbf{b}$ with θ the *angle* between $\mathbf{a}$ and $\mathbf{b}$. For this triangle, the *cosine theorem*

$$\|\mathbf{a} - \mathbf{b}\|^2 = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - 2\|\mathbf{a}\|\|\mathbf{b}\| \cos\theta \tag{A.1.1}$$

K. Adachi, *Matrix-Based Introduction to Multivariate Data Analysis*,
https://doi.org/10.1007/978-981-15-4103-2

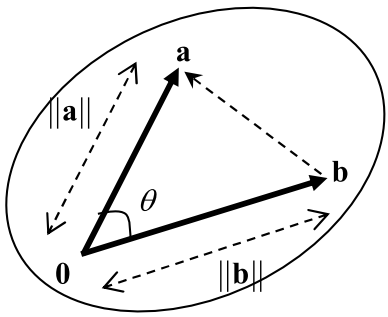


Fig A.1 Geometric illustration of vectors **a** and **b** with **0** = [0, …, 0]′

holds, which readers should have learned in high school. Its left-hand side can be expanded as $||\mathbf{a}||^2 + ||\mathbf{b}||^2 - 2\mathbf{a}'\mathbf{b}$, which implies

$$\mathbf{a}'\mathbf{b} = \|\mathbf{a}\|\|\mathbf{b}\| \cos\theta : \tag{A.1.2}$$

the inner product of two vectors is the multiplication of their lengths and the cosine of their angle. Equation (A.1.2) is rewritten as

$$\cos\theta = \frac{\mathbf{a}'\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}. \tag{A.1.3}$$

The *cosine of the angle* between two vectors equals the *division* of their *inner product* by their *lengths*.

Let the angle between vectors **s** and **t** be 90°, with their lengths not being zero. Then, **s** and **t** satisfy

$$\mathbf{s}'\mathbf{t} = 0, \tag{A.1.4}$$

because of (A.1.2) and cos 90° = 0. The two vectors in (A.1.4) are said to be mutually *orthogonal*.

A.1.2 Orthonormal Matrix

The $p \times m$ matrix $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_m]$ satisfying

$$\mathbf{W}'\mathbf{W} = \mathbf{I}_m \tag{A.1.5}$$

is said to be *column-orthonormal*, as (A.1.5) implies that the column vectors are mutually orthogonal with $\mathbf{w}_j'\mathbf{w}_k = 0$ for $j \neq k$ and of unit-length $\|\mathbf{w}_j\| = 1$ The term

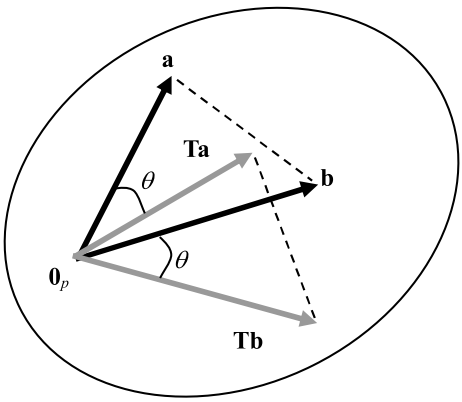


Fig A.2 Rotation of vectors by an orthonormal matrix

"orthonormal" is a composite of "orthogonal" and "normal", with the latter adjective standing for $\left\|\mathbf{w}_j\right\| = 1$.

Let a matrix $\mathbf{T}$ be a column-orthonormal and *square* of $p \times p$. It implies $\mathbf{T}$ being nonsingular and $\mathbf{T}' = \mathbf{T}^{-1}$ (the inverse matrix of $\mathbf{T}$):

$$\mathbf{T}'\mathbf{T} = \mathbf{T}\mathbf{T}' = \mathbf{I}_p. \tag{A.1.6}$$

Such a $\mathbf{T}$ is simply said to be *orthonormal*. For $p \times 1$ vectors $\mathbf{a}$ and $\mathbf{b}$,

$$\|\mathbf{T}\mathbf{a}\|^2 = \mathbf{a}'\mathbf{T}'\mathbf{T}\mathbf{a} = \mathbf{a}'\mathbf{a} = \|\mathbf{a}\|^2, \tag{A.1.7}$$

$$\|\mathbf{T}\mathbf{a} - \mathbf{T}\mathbf{b}\|^2 = (\mathbf{a} - \mathbf{b})'\mathbf{T}'\mathbf{T}(\mathbf{a} - \mathbf{b}) = (\mathbf{a} - \mathbf{b})'(\mathbf{a} - \mathbf{b}) = \|\mathbf{a} - \mathbf{b}\|^2: \tag{A.1.8}$$

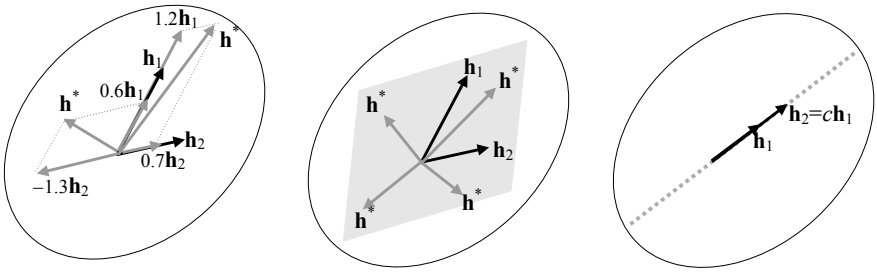
the pre-multiplication of vectors by an orthonormal matrix $\mathbf{T}$ does *not change* the *length* of the vectors or the *distance* between the vectors. This implies that the pre-multiplication simply *rotates* the vectors, as illustrated in Fig. A.2.

A.1.3 Vector Space

Let $\mathbf{H} = \left[\mathbf{h}_1, \ldots, \mathbf{h}_p\right]$ be an $n \times p$ matrix with $n > p$ and $\mathbf{b} = \left[b_1, \ldots, b_p\right]'$ a $p \times 1$ vector. The purpose of this section is to show what the *linear combination* of the column vectors in $\mathbf{H}$, i.e.,

$$\mathbf{h}^* = b_1\mathbf{h}_1 + \cdots + b_p\mathbf{h}_p = \mathbf{H}\mathbf{b}, \tag{A.1.9}$$

geometrically represents. Here, $\mathbf{H} = \left[\mathbf{h}_1, \ldots, \mathbf{h}_p\right]$ is fixed, while each element of $\mathbf{b} = \left[b_1, \ldots, b_p\right]'$ can take any real value: $-\infty < b_j < \infty$ for $j = 1, \ldots, p$.



(a) Linear combinations of $\mathbf{h}_1$ and $\mathbf{h}_2$

(b) Two-dimensional space for $[\mathbf{h}_1, \mathbf{h}_2]$ with its rank 2

(c) Uni-dimensional space for $[\mathbf{h}_1, \mathbf{h}_2]$ with its rank 1

Fig. A.3 Spaces spanned by $\mathbf{h}_1$ and $\mathbf{h}_2$

We start with the cases with $p = 2$, where (A.1.9) is simplified as

$$\mathbf{h}^* = b_1\mathbf{h}_1 + b_2\mathbf{h}_2. \tag{A.1.10}$$

In Fig. A.3a, the two vectors obtained with (A.1.10) are illustrated when $[b_1, b_2] = [0.6, -1.3]$ and when $[b_1, b_2] = [1.2, 0.7]$. Since vectors $\mathbf{h}_1$, $\mathbf{h}_2$, and $\mathbf{h}^*$ are $n \times 1$, they extend in an n-dimensional space; this is depicted as an ellipse in Fig. A.3. However, $\mathbf{h}^*$ cannot extend in arbitrary directions; they are *restricted*. As illustrated in Fig. A.3b, $\mathbf{h}^*$ can only extend on the grayed plane, i.e., on a *two*-dimensional space, on which $\mathbf{h}_1$ and $\mathbf{h}_2$ extend. This plane is formed by (A.1.10) with $-\infty < b_1 < \infty$ and $-\infty < b_2 < \infty$. Here, it should be noted that the ranges of b_1 and b_2 are $-\infty < b_1 < \infty$ and $-\infty < b_2 < \infty$, which implies the plane extends *infinitely*, though that cannot be depicted in the figure due to the limitations of the page. The plane in Fig. A.3b is called a *two-dimensional space spanned by* $\mathbf{h}_1$ and $\mathbf{h}_2$. Obviously, this space is *included* in the n-dimensional one for $n > p = 2$. Thus, the grayed plane is illustrated inside the ellipse in Fig. A.3b. The notions in this paragraph can be captured intuitively as follows:

Note A.1.1. Intuitive Understanding of Vector Spaces

Let us view the vectors $\mathbf{h}_1$ and $\mathbf{h}_2$ in Fig. 1.3b as *pencils* before our eyes, with $n = 3$. Then, we can verify that a *sheet* (or a thin *notebook*) can be located as the grayed plane in Fig. 1.3b; i.e., so that two *pencils*, $\mathbf{h}_1$ and $\mathbf{h}_2$, extend on the *sheet*.

Further, let $\mathbf{h}^*$ be another *pencil* extending in the direction satisfying (A.1.10). Then, we can verify that *pencil* $\mathbf{h}^*$ necessarily extends in the direction of the *sheet*; i.e., it cannot extend in a direction different from the *sheet*, regardless of the values b_1 and b_2 take. Here, the world in which we, $\mathbf{h}_1$, $\mathbf{h}_2$, $\mathbf{h}^*$, and the sheet exist is a three-dimensional space, but the sheet in whose direction $\mathbf{h}_1$, $\mathbf{h}_2$, and $\mathbf{h}^*$ extend is restricted to the two-dimensional space included in the three-dimensional one.

Though we have supposed so far that $\mathbf{h}_1$ and $\mathbf{h}_2$ are linearly *independent* with $\text{rank}([\mathbf{h}_1, \mathbf{h}_2]) = 2$ in Fig. A.3a, b, the case with $\mathbf{h}_2 = c\mathbf{h}_1$ (linearly *dependent*) and $\text{rank}([\mathbf{h}_1, \mathbf{h}_2]) = 1$ is illustrated in Fig. A.3c; linear dependence and the rank of a matrix were introduced in Sects. 3.9 and 3.10. Then, the space spanned by $\mathbf{h}_1$ and $\mathbf{h}_2$ is *one-dimensional*; the space is a line when $\text{rank}([\mathbf{h}_1, \mathbf{h}_2]) = 1$. It can also be ascertained that $\mathbf{h}_2 = c\mathbf{h}_1$ allows (A.1.10) to be rewritten as $\mathbf{h}^* = b_1\mathbf{h}_1 + b_2c\mathbf{h}_1 = (b_1 + b_2c)\mathbf{h}_1$ for $\mathbf{h}_2 = c\mathbf{h}_1$.

Now, let us consider the cases of $p = 3$, where (A.1.9) is expressed as

$$\mathbf{h}^* = b_1\mathbf{h}_1 + b_2\mathbf{h}_2 + b_3\mathbf{h}_3. \tag{A.1.11}$$

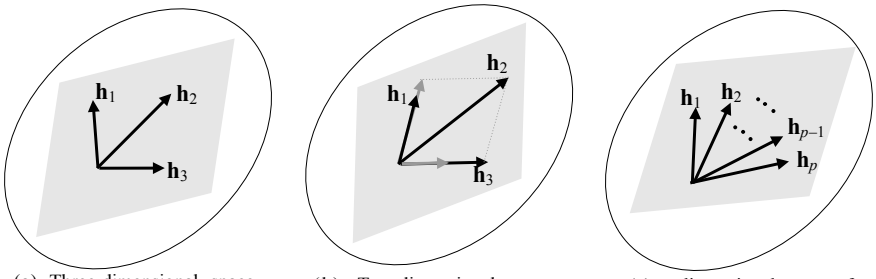
This gives the same story as in the previous paragraphs. The *three-dimensional space spanned by* $\mathbf{h}_1$, $\mathbf{h}_2$, and $\mathbf{h}_3$, which are linearly independent, is depicted as the grayed object in Fig. A.4a. Though that space (grayed object) is depicted as a "plane" in the figure, it is of three dimensions.

In Fig. A.3b, the case is illustrated in which $\mathbf{h}_1$, $\mathbf{h}_2$, and $\mathbf{h}_3$ are linearly *dependent*, with $\mathbf{h}_2 = c_1\mathbf{h}_1 + c_2\mathbf{h}_3$, but $\mathbf{h}_1$ and $\mathbf{h}_3$ are linearly independent, and $\text{rank}([\mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3]) = 2$. In this case, the space spanned by $\mathbf{h}_1$, $\mathbf{h}_2$, and $\mathbf{h}_3$ is *two*-dimensional, since (A.1.11) can be rewritten as $\mathbf{h}^* = b_1\mathbf{h}_1 + b_2(c_1\mathbf{h}_1 + c_2\mathbf{h}_3) + b_3\mathbf{h}_3 = (b_1 + b_2c_1)\mathbf{h}_1 + (b_2c_2 + b_3)\mathbf{h}_3$, which implies that the space spanned by $\mathbf{h}_1$, $\mathbf{h}_2$, and $\mathbf{h}_3$ is equivalent to the two-dimensional space spanned by $\mathbf{h}_1$ and $\mathbf{h}_3$.

The space spanned by $\mathbf{h}_1, \ldots, \mathbf{h}_p$ can be defined for $p > 3$ in the same manner as for $p = 2, 3$. This is illustrated in Fig. A.3c. That space is called the *column space* of $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_p]$ and is formally expressed as

$$\Xi(\mathbf{H}) = \{\mathbf{h}^* : \mathbf{h}^* = \mathbf{Hb} = b_1\mathbf{h}_1 + \cdots + b_p\mathbf{h}_p; -\infty < b_j < \infty, j = 1, \ldots, p\}. \tag{A.1.12}$$

The *dimensionality of the space* is equal to $r = \text{rank}(\mathbf{H})$. As $n > p$, this space is included in the n-dimensional space depicted as the ellipse in Fig. A.3c. Thus, the



(a) Three-dimensional space for [$\mathbf{h}_1$, $\mathbf{h}_2$, $\mathbf{h}_3$] with its rank 3

(b) Two-dimensional space for [$\mathbf{h}_1$, $\mathbf{h}_2$, $\mathbf{h}_3$] with its rank 2

(c) r-dimensional space for [$\mathbf{h}_1$, ... , $\mathbf{h}_p$] with its rank $r \le p$

Fig. A.4 Spaces spanned by $\mathbf{h}_1$, …, $\mathbf{h}_p$ for $p = 3$ and for $p > 3$

r-dimensional space spanned by $\mathbf{h}_1, \ldots, \mathbf{h}_p$, i.e., the column space of $\mathbf{H}$, is a *subspace* of n-dimensional space, since a space included in another space is called a subspace of the latter.

A.1.4 Projection Onto a Subspace

Let us consider a two-dimensional subspace (i.e., plane), which is included in a p-dimensional space and spanned by the $p \times 1$ vectors $\mathbf{w}_1$ and $\mathbf{w}_2$. Here, they are of *unit length* and mutually *orthogonal* with $\|\mathbf{w}_1\| = \|\mathbf{w}_2\| = 1$ and $\mathbf{w}_1'\mathbf{w}_2 = 0$. Those equations are summarized into

$$\begin{bmatrix} \mathbf{w}'_1 \\ \mathbf{w}'_2 \end{bmatrix} [\mathbf{w}_1, \mathbf{w}_2] = \mathbf{W}'\mathbf{W} = \mathbf{I}_m. \tag{A.1.13}$$

with $m = 2$ and $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2]$ $(p \times 2)$. This implies that $\mathbf{w}_1$ and $\mathbf{w}_2$ define the *orthogonal axes on the subspace*, as illustrated in Fig. A.5. Using $\mathbf{f} = [f_1, f_2]'$, whose elements can take arbitrary real values, any point on the subspace is expressed as

$$\mathbf{g} = \mathbf{W}\mathbf{f} = f_1\mathbf{w}_1 + f_2\mathbf{w}_2. \tag{A.1.14}$$

Now, we consider what values the elements of $\mathbf{f} = [f_1, f_2]'$ should take, subject to the condition that $\mathbf{g}$ $(p \times 1)$ is the *projection* of $\tilde{\mathbf{x}}_i$ $(p \times 1)$ onto the subspace (plane) spanned by $\mathbf{w}_1$ and $\mathbf{w}_2$. This condition is restated as the difference vector $\tilde{\mathbf{x}}_i - \mathbf{g}$ being *orthogonal* to the subspace, which is equivalent to $\tilde{\mathbf{x}}_i - \mathbf{g}$ being orthogonal to

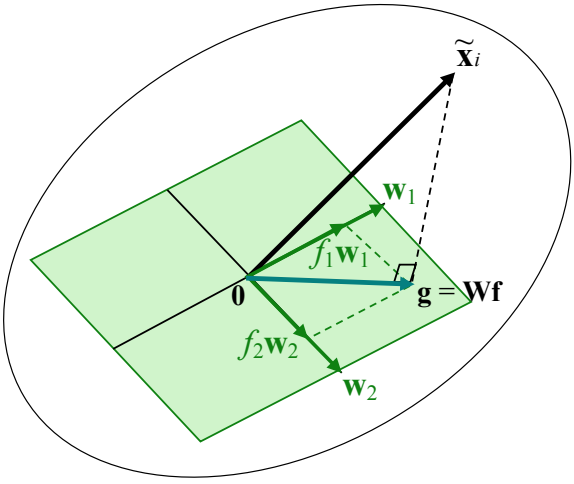


Fig A.5 Projection of a data vector on a plane

$\mathbf{w}_1$ and $\mathbf{w}_2$ with $(\tilde{\mathbf{x}}_i - \mathbf{g})'\mathbf{w}_1 = 0$ and $(\tilde{\mathbf{x}}_i - \mathbf{g})'\mathbf{w}_2 = 0$. These two equations are summarized into

$$(\tilde{\mathbf{x}}_i - \mathbf{g})'\mathbf{W} = \mathbf{0}_2'. \tag{A.1.15}$$

Substituting (A.1.14) in (A.1.15), we have $(\tilde{\mathbf{x}}_i - \mathbf{W}\mathbf{f})'\mathbf{W} = \mathbf{0}_2'$, which is rewritten as $\tilde{\mathbf{x}}'_i\mathbf{W} = \mathbf{f}'\mathbf{W}'\mathbf{W}$. In this equation, we can use (A.1.13) to get

$$\mathbf{f}' = \tilde{\mathbf{x}}_i'\mathbf{W} \quad \text{or} \quad \mathbf{f} = \mathbf{W}'\tilde{\mathbf{x}}_i. \tag{A.1.16}$$

The above discussions can be generalized to the cases with $m \geq 2$. That is, (A.1.16) expresses the *coordinates* of the *projection* of $\tilde{\mathbf{x}}_i$ onto the *subspace spanned by the columns* of $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_m]$ under the condition $\mathbf{W}'\mathbf{W} = \mathbf{I}_m$ in (A.1.13).

A.2 Decomposition of Sums of Squares

As shown in (1.31), the squared norm $||\mathbf{A}||^2 = \mathrm{tr}\mathbf{A}'\mathbf{A}$ expresses the sum of the squared elements in $\mathbf{A}$. Thus, $||\mathbf{A}||^2$ is also called a *sum of squares*. It can often be rewritten as the sum of *other* sums of squares as $\|\mathbf{A}\|^2 = \|\mathbf{B}\|^2 + \|\mathbf{C}\|^2$. Such an equality is generally called the *decomposition of the sum of squares*. The decomposition is utilized in the *least squares method* in which the parameter values are found that minimize a sum of squares.

A.2.1 Decomposition Using Averages

Let us consider the sum of squares

$$f(c) = \|\mathbf{h} - c\mathbf{1}_n\|^2, \tag{A.2.1}$$

with $\mathbf{h}$ an $n \times 1$ vector and c a scalar. We can find that (A.2.1) is minimized when c equals the *average* of the elements in $\mathbf{h}$:

$$\hat{c} = \frac{1}{n}\mathbf{1}_n'\mathbf{h}. \tag{A.2.2}$$

This result follows from the fact that (A.2.1) is decomposed as

$$\|\mathbf{h} - c\mathbf{1}_n\|^2 = \left\|\mathbf{h} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\mathbf{h}\right\|^2 + \left\|\frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\mathbf{h} - c\mathbf{1}_n\right\|^2 : \tag{A.2.3}$$

only the term $g(c) = \left\|n^{-1}\mathbf{1}_n\mathbf{1}_n'\mathbf{h} - c\mathbf{1}_n\right\|^2$ is relevant to c in the right-hand side of (A.2.3), and (A.2.2) allows $g(c)$ to attain its lower limit as $g(\hat{c}) = \left\|n^{-1}\mathbf{1}_n\mathbf{1}_n'\mathbf{h} - \mathbf{1}_n \times \hat{c}\right\|^2 = \left\|n^{-1}\mathbf{1}_n\mathbf{1}_n'\mathbf{h} - n^{-1}\mathbf{1}_n\mathbf{1}_n'\mathbf{h}\right\|^2 = 0$. The decomposition (A.2.3) is derived as follows: (A.2.1) can be rewritten as

$$\begin{aligned}\|\mathbf{h} - c\mathbf{1}_n\|^2 &= \|\mathbf{h} - \hat{c}\mathbf{1}_n + \hat{c}\mathbf{1}_n - c\mathbf{1}_n\|^2 \\ &= \|\mathbf{h} - \hat{c}\mathbf{1}_n\|^2 + \|\hat{c}\mathbf{1}_n - c\mathbf{1}_n\|^2 + 2v,\end{aligned} \tag{A.2.4}$$

with $v = (\mathbf{h} - \hat{c}\mathbf{1}_n)'(\hat{c}\mathbf{1}_n - c\mathbf{1}_n) = \hat{c}\mathbf{h}'\mathbf{1}_n - c\mathbf{h}'\mathbf{1}_n - \hat{c}^2 n + \hat{c}cn = \hat{c}(n\hat{c}) - c(n\hat{c}) - \hat{c}^2 n + \hat{c}cn = 0$ following from (A.2.2), or equivalently, $\mathbf{1}_n'\mathbf{h} = n\hat{c}$.

Next, let us consider the *sum* of the sums of squares

$$h(\mathbf{F}) = \sum_{j=1}^{J} \left\|\mathbf{F} - \mathbf{Z}_j\right\|^2, \tag{A.2.5}$$

with $\mathbf{F}$ and $\mathbf{Z}_j$ $n \times m$ matrices. We can find that (A.2.5) is minimized when $\mathbf{F}$ equals

$$\hat{\mathbf{F}} = \overline{\mathbf{Z}} = \frac{1}{J}\sum_{j=1}^{J} \mathbf{Z}_j, \tag{A.2.6}$$

using the fact that (A.2.5) is decomposed as

$$\sum_{j=1}^{J} \left\|\mathbf{F} - \mathbf{Z}_j\right\|^2 = J\left\|\mathbf{F} - \overline{\mathbf{Z}}\right\| + \sum_{j=1}^{J} \left\|\overline{\mathbf{Z}} - \mathbf{Z}_j\right\|^2. \tag{A.2.7}$$

In the right-hand side, only the term $J\left\|\mathbf{F} - \overline{\mathbf{Z}}\right\|^2$ is relevant to $\mathbf{F}$ and that term attains zero when $\mathbf{F}$ equals (A.2.6). Decomposition (A.2.7) is derived as follows: (A.2.5) can be rewritten as

$$\sum_{j=1}^{J} \left\|\mathbf{F} - \mathbf{Z}_j\right\|^2 = \sum_{j=1}^{J} \left\|\mathbf{F} - \overline{\mathbf{Z}} + \overline{\mathbf{Z}} - \mathbf{Z}_j\right\|^2 = J\left\|\mathbf{F} - \overline{\mathbf{Z}}\right\| + \sum_{j=1}^{J} \left\|\overline{\mathbf{Z}} - \mathbf{Z}_j\right\|^2 + 2\mathrm{tr}\mathbf{S}, \tag{A.2.8}$$

with

$$\begin{aligned}\mathbf{S} &= \sum_{j=1}^{J} (\mathbf{F} - \overline{\mathbf{Z}})'(\overline{\mathbf{Z}} - \mathbf{Z}_j) \\ &= \sum_{j=1}^{J} \mathbf{F}'\overline{\mathbf{Z}} - \sum_{j=1}^{J} \mathbf{F}'\mathbf{Z}_j - \sum_{j=1}^{J} \overline{\mathbf{Z}}'\overline{\mathbf{Z}} + \sum_{j=1}^{J} \overline{\mathbf{Z}}'\mathbf{Z}_j \\ &= J\mathbf{F}'\overline{\mathbf{Z}} - \mathbf{F}' \sum_{j=1}^{J} \mathbf{Z}_j - J\overline{\mathbf{Z}}'\overline{\mathbf{Z}} + \overline{\mathbf{Z}}' \sum_{j=1}^{J} \mathbf{Z}_j \\ &= J\mathbf{F}'\overline{\mathbf{Z}} - J\mathbf{F}'\overline{\mathbf{Z}} - J\overline{\mathbf{Z}}'\overline{\mathbf{Z}} + J\overline{\mathbf{Z}}'\overline{\mathbf{Z}} =_m \mathbf{O}_m,\end{aligned} \tag{A.2.9}$$

where we have used the fact that (A.2.6) implies $\sum_{j=1}^{J} \mathbf{Z}_j = J\overline{\mathbf{Z}}$.

A.2.2 Decomposition Using a Projection Matrix

The $n \times n$ matrix

$$\mathbf{P}_{\mathrm{X}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \tag{A.2.10}$$

is called a *projection matrix* for $\mathbf{X}$ ($n \times p$). Though the use of (A.2.10) allows us to generalize the discussions in A.1.4 (e.g., Banerjee and Roy, 2014; Yanai, Takeuchi, & Takane, 2011), that is beyond the scope of this book. Here, we focus only on the decomposition of sums of squares using (A.2.10).

Let us consider the sum of squares

$$f(\mathbf{B}) = \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|^2, \tag{A.2.11}$$

with $\mathbf{Y}$ and $\mathbf{B}$ being $n \times q$ and $p \times q$ matrices, respectively, and $\mathbf{X}'\mathbf{X}$ nonsingular. We find that (A.2.11) is minimized when

$$\mathbf{X}\mathbf{B} = \mathbf{P}_{\mathrm{X}}\mathbf{Y}, \text{ ie}, \mathbf{B} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}, \tag{A.2.12}$$

using the fact that (A.2.11) is decomposed as

$$\|\mathbf{Y} - \mathbf{X}\mathbf{B}\|^2 = \|\mathbf{Y} - \mathbf{P}_{\mathrm{X}}\mathbf{Y}\|^2 + \|\mathbf{P}_{\mathrm{X}}\mathbf{Y} - \mathbf{X}\mathbf{B}\|^2. \tag{A.2.13}$$

On the right-hand side, only the term $\|\mathbf{P}_{\mathrm{X}}\mathbf{Y} - \mathbf{X}\mathbf{B}\|^2$ is relevant to $\mathbf{B}$ and that term attains zero for (A.2.12). Decomposition (A.2.13) is derived as follows: (A.2.11) can be rewritten as

$$\begin{aligned}\|\mathbf{Y}-\mathbf{XB}\|^2 &= \|\mathbf{Y}-\mathbf{P}_\mathrm{X}\mathbf{Y}+\mathbf{P}_\mathrm{X}\mathbf{Y}-\mathbf{XB}\|^2 \\ &= \|\mathbf{Y}-\mathbf{P}_\mathrm{X}\mathbf{Y}\|^2+\|\mathbf{P}_\mathrm{X}\mathbf{Y}-\mathbf{XB}\|^2+2\mathrm{tr}\mathbf{C},\end{aligned} \tag{A.2.14}$$

with

$$\begin{aligned}\mathbf{C} &= (\mathbf{Y}-\mathbf{P}_\mathrm{X}\mathbf{Y})'(\mathbf{P}_\mathrm{X}\mathbf{Y}-\mathbf{XB}) \\ &= \mathbf{Y}'\mathbf{P}_\mathrm{X}\mathbf{Y}-\mathbf{Y}'\mathbf{XB}-\mathbf{Y}'\mathbf{P}_\mathrm{X}^2\mathbf{Y}+\mathbf{Y}'\mathbf{P}_\mathrm{X}'\mathbf{XB} =_q \mathrm{O}_q,\end{aligned} \tag{A.2.15}$$

where we have used $\mathbf{P}_\mathrm{X}' = \mathbf{P}_\mathrm{X}$, $\mathbf{P}_\mathrm{X}^2 = \mathbf{P}_\mathrm{X}$, and $\mathbf{P}_\mathrm{X}\mathbf{X} = \mathbf{X}$.

Solution (4.12) in Chap. 4 is obtained by setting $q = 1$ and substituting $\mathbf{JX}$ and $\mathbf{y}$ for $\mathbf{X}$ and $\mathbf{Y}$ in (A.2.12):

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{J}'\mathbf{JX})^{-1}\mathbf{X}'\mathbf{Jy} = (\mathbf{X}'\mathbf{JX})^{-1}\mathbf{X}'\mathbf{Jy}, \tag{A.2.16}$$

where $\mathbf{B}$ in (A.2.12) is replaced by $\hat{\mathbf{b}}$ $(p \times 1)$.

We should note that $n^{-1}\mathbf{1}_n\mathbf{1}_n'$ in (A.2.3) is also a projection matrix, since substituting $\mathbf{1}_n$ for $\mathbf{X}$ in (A.2.10) leads to $\mathbf{P}_{\mathbf{1}_n} = \mathbf{1}_n(\mathbf{1}_n'\mathbf{1}_n)^{-1}\mathbf{1}_n' = n^{-1}\mathbf{1}_n\mathbf{1}_n'$.

A.3 Singular Value Decomposition

The author believes that *singular value decomposition* (*SVD*) is the *most important* tool in matrix algebra, as SVD can be defined for any matrix, a number of facts can be easily derived from SVD, and it plays important roles in matrix computations as found in Appendix A.4.

A.3.1 SVD: Extended Version

Please, *learn* this theorem (SVD) by *heart* as *absolute truth*!

Theorem A.3.1. SVD (*extended version*)
Any $n \times p$ matrix $\mathbf{X}$ with $n \geq p$ can be decomposed as

$$\mathbf{X} = \tilde{\mathbf{K}}\tilde{\Lambda}\tilde{\mathbf{L}}'. \tag{A.3.1}$$

Here, $\tilde{\mathbf{K}}$ $(n \times p)$ is an $n \times p$ column-orthonormal matrix and $\tilde{\mathbf{L}}$ $(p \times p)$ is a $p \times p$ orthonormal matrix:

$$\tilde{\mathbf{K}}'\tilde{\mathbf{K}} = \tilde{\mathbf{L}}'\tilde{\mathbf{L}} = \tilde{\mathbf{L}}\tilde{\mathbf{L}}' = \mathbf{I}_p. \tag{A.3.2}$$

$\tilde{\mathbf{\Lambda}}$ is a $p \times p$ *diagonal* matrix

$$\tilde{\mathbf{\Lambda}} = \begin{bmatrix} \lambda_1 & & & & & \\ & \ddots & & & & \\ & & \lambda_r & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix} \tag{A.3.3}$$

with its diagonal elements arranged in *decreasing* order

$$\lambda_1 \geq \cdots \geq \lambda_r > 0, \tag{A.3.4}$$

the *number of the positive diagonal* elements being the *rank* of $\mathbf{X}$:

$$r = \text{rank}(\mathbf{X}), \tag{A.3.5}$$

and the blank cells standing for zero elements.

Theorem A.3.1 concerns the SVD of a matrix with the number of rows greater than or equal to that of columns. The SVD of a matrix with more columns than rows can be defined simply by *transposing* both sides of (A.3.1): Any matrix $\mathbf{X}'$ $(p \times n)$ with $p \leq n$ can be decomposed as

$$\mathbf{X}' = \tilde{\mathbf{L}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{K}}', \tag{A.3.6}$$

with (A.3.2)–(A.3.5).

Theorem A.3.1 shows that we can easily find rank($\mathbf{X}$) by *counting the number of nonzero diagonal elements* in $\tilde{\mathbf{\Lambda}}$, if the SVD of $\mathbf{X}$ is given. Further, SVD leads to the following fact: for an $n \times p$ matrix $\mathbf{X}$,

$$\text{rank}(\mathbf{X}) = \text{rank}(\mathbf{X}') = \text{rank}(\mathbf{X}\mathbf{X}') = \text{rank}(\mathbf{X}'\mathbf{X}). \tag{A.3.7}$$

Here, the first equality directly follows from (A.3.6), and the rank($\mathbf{X}$) = rank($\mathbf{X}\mathbf{X}'$) = rank($\mathbf{X}'\mathbf{X}$) follows from the fact that Theorem A.3.1 implies $\mathbf{X}'\mathbf{X} = \tilde{\mathbf{L}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{K}}'\tilde{\mathbf{K}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{L}}' = \tilde{\mathbf{L}}\tilde{\mathbf{\Lambda}}^2\tilde{\mathbf{L}}'$ and $\mathbf{X}\mathbf{X}' = \tilde{\mathbf{K}}\tilde{\mathbf{\Lambda}}^2\tilde{\mathbf{K}}'$; the SVD of $\mathbf{X}'\mathbf{X}$ and that of $\mathbf{X}\mathbf{X}'$ are $\tilde{\mathbf{L}}\tilde{\mathbf{\Lambda}}^2\tilde{\mathbf{L}}'$ and $\tilde{\mathbf{K}}\tilde{\mathbf{\Lambda}}^2\tilde{\mathbf{K}}'$, respectively.

A.3.2 SVD: Compact Version

Let us consider the same matrices as in Theorem A.3.1, and let $\mathbf{K}$ and $\mathbf{L}$ be the matrices containing the first r columns of $\tilde{\mathbf{K}}$ and $\tilde{\mathbf{L}}$, respectively, with $\tilde{\mathbf{K}} = [\mathbf{K}, \mathbf{K}_{(p-r)}]$ and $\tilde{\mathbf{L}} = [\mathbf{L}, \mathbf{L}_{(p-r)}]$ being block matrices (whose introduction is found in Sect. 14.1). Here, $\mathbf{K}_{(p-r)}$ and $\mathbf{L}_{(p-r)}$ contain the last $p - r$ columns of $\tilde{\mathbf{K}}$ and $\tilde{\mathbf{L}}$, respectively. Further, let $\mathbf{\Lambda}$ be the $r \times r$ diagonal matrix whose diagonal elements are $\lambda_1, \geq \cdots \geq \lambda_r$. Then, the right-hand side of (A.3.1) is rewritten as

$$\tilde{\mathbf{K}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{L}}' = [\mathbf{K}, \mathbf{K}_{(p-r)}] \begin{bmatrix} \mathbf{\Lambda} & \\ & {}_{p-r}\mathbf{O}_{p-r} \end{bmatrix} \begin{bmatrix} \mathbf{L}' \\ \mathbf{L}'_{p-r} \end{bmatrix} = \mathbf{K}\mathbf{\Lambda}\mathbf{L}'. \tag{A.3.8}$$

Theorem A.3.2. SVD (*compact version*)
Any $n \times p$ matrix $\mathbf{X}$ with $rank(\mathbf{X}) = r$ can be decomposed as

$$\mathbf{X} = \mathbf{K}\mathbf{\Lambda}\mathbf{L}'. \tag{A.3.9}$$

Here, $\mathbf{K}$ ($n \times r$) and $\mathbf{L}$ ($p \times r$) are *column-orthonormal* matrices with

$$\mathbf{K}'\mathbf{K} = \mathbf{L}'\mathbf{L} = \mathbf{I}_r \tag{A.3.10}$$

and $\mathbf{\Lambda}$ is the $r \times r$ *diagonal* matrix

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_r \end{bmatrix}, \tag{A.3.11}$$

whose diagonal elements are positive and arranged in *decreasing* order with

$$\lambda_1 \geq \cdots \geq \lambda_r > 0. \tag{A.3.12}$$

The diagonal matrix $\mathbf{\Lambda}$ is unique; i.e., only a single $\mathbf{\Lambda}$ exists for $\mathbf{X}$. Further, if $\lambda_1, > \cdots > \lambda_r$, $\mathbf{K}$ and $\mathbf{L}$ are also unique, except for that the signs (i.e., positive and negative) of all elements in the corresponding columns of $\mathbf{K}$ and $\mathbf{L}$ can be changed simultaneously. That is, (A.3.9) can be rewritten as $\mathbf{X} = \mathbf{K}\mathbf{\Lambda}\mathbf{L}' = (\mathbf{K}\mathbf{D}_{\pm})(\mathbf{D}_{\pm}\mathbf{\Lambda}\mathbf{D}_{\pm})(\mathbf{L}\mathbf{D}_{\pm})'$. Here, $\mathbf{D}_{\pm}$ is an $r \times r$ diagonal matrix, each of whose diagonal elements is either 1 or -1. Since $\mathbf{K}\mathbf{D}_{\pm}$ and $\mathbf{L}\mathbf{D}_{\pm}$ *can* be substituted into $\mathbf{K}$ and $\mathbf{L}$ in (A.3.10), respectively, with $\mathbf{\Lambda} = \mathbf{D}_{\pm}\mathbf{\Lambda}\mathbf{D}_{\pm}$, $\mathbf{X} = (\mathbf{K}\mathbf{D}_{\pm})(\mathbf{D}_{\pm}\mathbf{\Lambda}\mathbf{D}_{\pm})(\mathbf{L}\mathbf{D}_{\pm})'$ is also the SVD of $\mathbf{X}$.

Thus, we have the *compact* version of Theorem A.3.1.

The lth diagonal element (λ_l) of $\mathbf{\Lambda}$ is called the lth largest *singular value* of $\mathbf{X}$. The lth columns of $\mathbf{K}$ and $\mathbf{L}$ are called the left and right *singular vectors* of $\mathbf{X}$ corresponding to λ_l, respectively. Obviously, the SVD of $\mathbf{X}'$ is defined as $\mathbf{X}' = \mathbf{L}\mathbf{\Lambda}\mathbf{K}'$ with (A.3.10), (A.3.11), and (A.3.12).

Theorem A.3.2 shows that the SVD of $\mathbf{X}\mathbf{X}'$ and $\mathbf{X}'\mathbf{X}$ is defined as

$$\mathbf{X}\mathbf{X}' = \mathbf{K}\mathbf{\Lambda}^2\mathbf{K}', \tag{A.3.13}$$

$$\mathbf{X}'\mathbf{X} = \mathbf{L}\mathbf{\Lambda}^2\mathbf{L}', \tag{A.3.14}$$

respectively. The SVDs (A.3.13) and (A.3.14) lead to the *sum of squares elements* in $\mathbf{X}$ equaling the *sum of* its *squared singular values*:

$$\|\mathbf{X}\|^2 = \mathrm{tr}\mathbf{X}'\mathbf{X} = \mathrm{tr}\mathbf{X}\mathbf{X}' = \mathrm{tr}\mathbf{\Lambda}^2 = \lambda_1^2 + \cdots + \lambda_r^2, \tag{A.3.15}$$

since $\mathrm{tr}\mathbf{X}'\mathbf{X} = \mathrm{tr}\mathbf{L}\mathbf{\Lambda}\mathbf{K}'\mathbf{K}\mathbf{\Lambda}\mathbf{L}' = \mathrm{tr}\mathbf{L}\mathbf{\Lambda}\mathbf{\Lambda}\mathbf{L}' = \mathrm{tr}\mathbf{L}\mathbf{\Lambda}^2\mathbf{L}' = \mathrm{tr}\mathbf{\Lambda}^2\mathbf{L}'\mathbf{L} = \mathrm{tr}\mathbf{\Lambda}^2$. If $\mathrm{rank}(\mathbf{X}'\mathbf{X}) = p$, then it is a nonsingular square matrix and its *inverse* matrix is given by

$$(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{L}\mathbf{\Lambda}^{-2}\mathbf{L}'. \tag{A.3.16}$$

If $p = n$ and $\mathbf{X}$ is nonsingular, then

$$\mathbf{X}^{-1} = \mathbf{L}\mathbf{\Lambda}^{-1}\mathbf{K}'. \tag{A.3.17}$$

A.3.3 Other Expressions of SVD

Let us express the matrices $\mathbf{K}$ and $\mathbf{L}$ in Theorem A.3.2 as $\mathbf{K} = [\mathbf{k}_1, \ldots, \mathbf{k}_m, \mathbf{k}_{m+1}, \ldots, \mathbf{k}_r] = [\mathbf{K}_m, \mathbf{K}_{[m]}]$ and $\mathbf{L} = [\mathbf{l}_1, \ldots, \mathbf{l}_m, \mathbf{l}_{m+1}, \ldots, \mathbf{l}_r] = [\mathbf{L}_m, \mathbf{L}_{[m]}]$. Here,

$$\mathbf{K}_m = [\mathbf{k}_1, \ldots, \mathbf{k}_m] \text{ and } \mathbf{L}_m = [\mathbf{l}_1, \ldots, \mathbf{l}_m] \tag{A.3.18}$$

contain the *first m columns* of $\mathbf{K}$ and $\mathbf{L}$, respectively, while

$$\mathbf{K}_{[m]} = [\mathbf{k}_{m+1}, \ldots, \mathbf{k}_r] \text{ and } \mathbf{L}_{[m]} = [\mathbf{l}_{m+1}, \ldots, \mathbf{l}_r] \tag{A.3.19}$$

contain the $r - m$ *remaining columns* of $\mathbf{K}$ and $\mathbf{L}$, respectively. Then, (A.3.10) is rewritten as

$$\mathbf{k}_u'\mathbf{k}_u = \mathbf{l}_u'\mathbf{l}_u = 1 \text{ and } \mathbf{k}_u'\mathbf{k}_v = \mathbf{l}_u'\mathbf{l}_v = 0 \text{ for } u \neq v, \tag{A.3.20}$$

with $u = 1, \ldots, r$ and $v = 1, \ldots, r$. Further, SVD (A.3.9) can be rewritten as $\mathbf{X} = \lambda_1 \mathbf{k}_1 \mathbf{l}_1' + \cdots + \lambda_m \mathbf{k}_m \mathbf{l}_m' + \lambda_{m+1} \mathbf{k}_{m+1} \mathbf{l}_{m+1}' + \cdots + \lambda_r \mathbf{k}_r \mathbf{l}_r'$, which is expressed in matrix form as

$$\mathbf{X} = \mathbf{K}\boldsymbol{\Lambda}\mathbf{L}' = \mathbf{K}_m \boldsymbol{\Lambda}_m \mathbf{L}_m' + \mathbf{K}_{[m]} \boldsymbol{\Lambda}_{[m]} \mathbf{L}_{[m]}', \tag{A.3.21}$$

with

$$\boldsymbol{\Lambda}_m = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_m \end{bmatrix} \text{ and } \boldsymbol{\Lambda}_{[m]} = \begin{bmatrix} \lambda_{m+1} & & \\ & \ddots & \\ & & \lambda_r \end{bmatrix}; \text{ i.e., } \boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{\Lambda}_m & \\ & \boldsymbol{\Lambda}_{[m]} \end{bmatrix}. \tag{A.3.22}$$

By noting (A.3.20), we find

$$\mathbf{K}'\mathbf{K}_m = \mathbf{L}'\mathbf{L}_m = \begin{bmatrix} 1 & & & \\ & \ddots & & \\ & & & 1 \\ 0 & \cdots & 0 & \\ & & \vdots & \\ 0 & \cdots & 0 & \end{bmatrix} = \begin{bmatrix} \mathbf{I}_m \\ {}_{r-m}\mathbf{O}_m \end{bmatrix}: \tag{A.3.23}$$

$\mathbf{K}'\mathbf{K}_m = \mathbf{L}'\mathbf{L}_m$ equals the $r \times m$ matrix whose first m rows are those of $\mathbf{I}_m$ and the remaining rows are filled with zeros. Post-multiplying both sides of (A.3.9) by $\mathbf{L}_m$ and using (A.3.23) leads to

$$\begin{aligned} \mathbf{X}\mathbf{L}_m &= \mathbf{K}\boldsymbol{\Lambda}\mathbf{L}'\mathbf{L}_m = \left[\mathbf{K}_m, \mathbf{K}_{[m]}\right] \begin{bmatrix} \boldsymbol{\Lambda}_m & \\ & \boldsymbol{\Lambda}_{[m]} \end{bmatrix} \begin{bmatrix} \mathbf{I}_m \\ {}_{r-m}\mathbf{O}_m \end{bmatrix} = \left[\mathbf{K}_m, \mathbf{K}_{[m]}\right] \begin{bmatrix} \boldsymbol{\Lambda}_m \\ {}_{r-m}\mathbf{O}_m \end{bmatrix} \\ &= \mathbf{K}_m \boldsymbol{\Lambda}_m, \end{aligned}$$

that is,

$$\mathbf{K}_m \boldsymbol{\Lambda}_m = \mathbf{X}\mathbf{L}_m. \tag{A.3.24}$$

Further, post-multiplying both sides by $\mathbf{L}_m'$ gives

$$\mathbf{K}_m \boldsymbol{\Lambda}_m \mathbf{L}_m' = \mathbf{X}\mathbf{L}_m\mathbf{L}_m'. \tag{A.3.25}$$

We can also use (A.3.23) to rewrite SVD (A.3.9) as

$$\mathbf{L}_m \boldsymbol{\Lambda}_m = \mathbf{X}'\mathbf{K}_m, \tag{A.3.26}$$

which follows from

$$\mathbf{X}'\mathbf{K}_m = \mathbf{L}\boldsymbol{\Lambda}\mathbf{K}'\mathbf{K}_m = \left[\mathbf{L}_m, \mathbf{L}_{[m]}\right]\begin{bmatrix}\boldsymbol{\Lambda}_m & \\ & \boldsymbol{\Lambda}_{[m]}\end{bmatrix}\begin{bmatrix}\mathbf{I}_m \\ {}_{r-m}\mathbf{O}_m\end{bmatrix} = \left[\mathbf{L}_m, \mathbf{L}_{[m]}\right]\begin{bmatrix}\boldsymbol{\Lambda}_m \\ {}_{r-m}\mathbf{O}_m\end{bmatrix}$$
$$= \mathbf{L}_m\boldsymbol{\Lambda}_m.$$

A.3.4 SVD and Eigenvalue Decomposition for Sysmmetric Matrices

Let us define $\mathbf{C} = \mathbf{X}'\mathbf{X}$ with $\mathbf{X}$ treated in Theorem A.3.2. As shown in A.3.14, the SVD of $\mathbf{C}$ is given by $\mathbf{C} = \mathbf{L}\boldsymbol{\Lambda}^2\mathbf{L}'$. This is also the *eigenvalue decomposition* (*EVD*) of $\mathbf{C}$ as found in Note 6.1 and the next theorem:

Theorem A.3.3. EVD of Nonnegative-Definite Matrices

A symmetric matrix $\mathbf{C}$ $(p \times p)$ being nonnegative-definite is equivalent to the property of $\mathbf{C}$ that it can be rewritten as $\mathbf{C} = \mathbf{X}'\mathbf{X}$, as described in Note 8.2 (i.e., we find this fact, if the matrices $\mathbf{S}$ and $\mathbf{B}$ in Note 8.2 are rewritten as $\mathbf{C}$ and $\mathbf{X}'$).

Let the SVD of $\mathbf{X}$ $(n \times p)$ be defined as in Theorem A.3.2 with $\text{rank}(\mathbf{X}) = r$. Then, the SVD of $\mathbf{C} = \mathbf{X}'\mathbf{X}$ is expressed as

$$\mathbf{C} = \mathbf{L}\boldsymbol{\Lambda}^2\mathbf{L}', \tag{A.3.27}$$

as already shown in (A.3.14). We can also refer to (A.3.27) as the EVD or *spectral decomposition* of $\mathbf{C}$ as described in Note 6.1. Here, λ_k^2 (the kth diagonal elements of $\boldsymbol{\Lambda}^2$) is called the kth largest *eigenvalues* of $\mathbf{C}$, and the kth column of $\mathbf{L}$ is called the *eigenvector* of $\mathbf{C}$ corresponding to λ_k^2.

As shown above, the SVD and EVD are equivalent for a nonnegative definite symmetric matrix which is the product of a matrix and its transpose. However, it does *not hold* true for a symmetric matrix which is not nonnegative-definite, as shown next.

Let $\mathbf{S}$ be an arbitrary $n \times n$ symmetric matrix with $\text{rank}(\mathbf{S}) = r \leq n$. The EVD of $\mathbf{S}$ can be expressed as

$$\mathbf{S} = \mathbf{E}\boldsymbol{\Theta}\mathbf{E}'. \tag{A.3.28}$$

Here, $\mathbf{E}'\mathbf{E} = \mathbf{I}_r$, and $\mathbf{\Theta}$ is the $r \times r$ diagonal matrix with its kth diagonal element θ_k satisfying $|\theta_k| \geq |\theta_{k+1}|$. In general, θ_k, an eigenvalue of $\mathbf{S}$, can be *negative*, which implies that (A.3.28) is not the SVD of $\mathbf{S}$. Its SVD can be expressed as

$$\mathbf{S} = \mathbf{EDD\Theta E}'. \tag{A.3.29}$$

Here, $\mathbf{D}$ is the $r \times r$ diagonal matrix whose kth diagonal element is 1 if $\theta_k > 0$, but -1 otherwise. We can find that $\mathbf{D\Theta}$ is the diagonal matrix with positive diagonal elements, i.e., the singular values of $\mathbf{S}$, and the corresponding singular vectors are contained in $\mathbf{ED}$ and $\mathbf{E}$, with $(\mathbf{ED})'\mathbf{ED} = \mathbf{DE}'\mathbf{ED} = \mathbf{D}^2 = \mathbf{I}_r$.

A.4 Matrix Computations Using SVD

The purpose of this appendix is to present solutions for the problems of *maximizing* some *traces* of matrix products and *reduced rank approximations*. Their foundation is given by the Theorem in Appendix A.4.1.

A.4.1 ten Berge's Theorem with Suborthonormal Matrices

Definition A.4.1. Suborthonormal Matrix
A matrix is *suborthonormal* if it can be completed to be an orthonormal matrix by appending rows, columns, or both, or if it is orthonormal (ten Berge, 1993, pp. 27–28).

An example of a suborthonormal matrix is $\mathbf{A} = \begin{bmatrix} 0.8 & 0.0 \\ 0.0 & 0.1 \end{bmatrix}$ (ten Berge, 1993, p. 28), since we can append the row [0.6, 0.0] and the column $\begin{bmatrix} 0.6 \\ 0.0 \\ -0.8 \end{bmatrix}$ to $\mathbf{A}$ so that it can be completed to be orthonormal $\tilde{\mathbf{A}} = \begin{bmatrix} 0.8 & 0.0 & 0.6 \\ 0.0 & 0.1 & 0.0 \\ 0.6 & 0.0 & -0.8 \end{bmatrix}$ with $\tilde{\mathbf{A}}'\tilde{\mathbf{A}} = \tilde{\mathbf{A}}\tilde{\mathbf{A}}' = \mathbf{I}_3$.

$$\begin{aligned} &\text{A } p \times m \textit{ column}-\textit{orthonormal} \text{ matrix } \mathbf{B} \\ &\quad \text{and } \mathbf{B}' \text{are } \textit{suborthonormal} \text{ with } p > m, \end{aligned} \tag{A.4.1}$$

since the $p \times p$ matrices $[\mathbf{B}, \mathbf{C}]$ and $\begin{bmatrix} \mathbf{B}' \\ \mathbf{C}' \end{bmatrix}$ are orthonormal, with $\mathbf{C}$ a $p \times (p-m)$ matrix satisfying $\mathbf{B}'\mathbf{C} = {}_m\mathbf{O}_{p-m}$ and $\mathbf{C}'\mathbf{C} = \mathbf{I}_{p-m}$. A suborthonormal matrix has the following property:

$$\text{the } \textit{product} \text{ of } \textit{suborthonormal} \text{ matrices} = \text{a } \textit{suborthonormal} \text{ matrix} \qquad (A.4.2)$$

(ten Berge, 1983, 1993).

The following theorem concerning suborthonormal matrices gives the foundation for the facts shown in Appendices A.4.2–A.4.5:

Theorem A.4.1. ten Berge's (1993) Theorem

If $\mathbf{S}$ is a $p \times p$ *suborthonormal* matrix with $rank(\mathbf{S}) = m \leq p$ and $\mathbf{D} = \begin{bmatrix} d_1 & & \\ & \ddots & \\ & & d_p \end{bmatrix}$ is a $p \times p$ *diagonal* matrix with $d_1 \geq \cdots \geq d_p \geq 0$, then

$$f(\mathbf{S}) = \text{tr}\mathbf{S}\mathbf{D} \leq \text{tr}\mathbf{D}_m = d_1 + \cdots + d_m \leq \mathbf{trD}, \qquad (A.4.3)$$

with $\mathbf{D}_m = \begin{bmatrix} d_1 & & \\ & \ddots & \\ & & d_m \end{bmatrix}$ the $m \times m$ diagonal matrix whose diagonal elements are the first m ones of $\mathbf{D}$.

This theorem has been proved by ten Berge (1983, Theorem 2) in a more generalized setting. As $d_1 + \cdots + d_m \leq \text{tr}\mathbf{D}$ obviously holds, this has been added to ten Berge's (1993, p. 28) inequality in (A.4.3).

A.4.2 Maximization of Trace Functions

In this section, we consider the *maximization* problems for three forms of *trace functions*. Here, the sentence "maximize $f(\mathbf{B})$ over $\mathbf{B}$ s.t. $g(\mathbf{B}) = c$" means "obtain the matrix $\mathbf{B}$ that maximizes $f(\mathbf{B})$ subject to the constraint $g(\mathbf{B}) = c$" with "s.t." the abbreviation for "*subject to*".

Theorem A.4.2
For an $n \times p$ matrix $\mathbf{Y}$ with $rank(\mathbf{Y}) = p$, we consider the problem:

$$\text{Maximize } f(\mathbf{C}) = \text{tr}\mathbf{Y}'\mathbf{C} \text{ over } \mathbf{C}(n \times p) \text{ s.t. } \mathbf{C}'\mathbf{C} = \mathbf{I}_p. \tag{A.4.4}$$

This is attained for

$$\mathbf{C} = \mathbf{U}\mathbf{V}'. \tag{A.4.5}$$

Here, $\mathbf{U}$ $(n \times p)$ and $\mathbf{V}$ $(p \times p)$ are given by the SVD of $\mathbf{Y}$ defined as $\mathbf{Y} = \mathbf{U}\mathbf{D}\mathbf{V}'$ with $\mathbf{U}'\mathbf{U} = \mathbf{V}'\mathbf{V} = \mathbf{I}_p$ and $\mathbf{D}$ a $p \times p$ diagonal matrix whose diagonal elements are all positive.

Proof By substituting $\mathbf{Y} = \mathbf{U}\mathbf{D}\mathbf{V}'$ in $f(\mathbf{C}) = \text{tr}\mathbf{Y}'\mathbf{C}$, this is rewritten as $f(\mathbf{C}) = \text{tr}\mathbf{V}\mathbf{D}\mathbf{U}'\mathbf{C} = \text{tr}\mathbf{U}'\mathbf{C}\mathbf{V}\mathbf{D}$. The column-orthonormality of $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{C}$ implies that $\mathbf{U}'\mathbf{C}\mathbf{V}$ is suborthonormal, because of (A.4.1) and (A.4.2). Further, $r = \text{rank}(\mathbf{U}'\mathbf{C}\mathbf{V}) \leq p$, while $\mathbf{D}$ is a $p \times p$ diagonal matrix with all diagonal elements positive. Those facts and Theorem A.4.1 lead to $f(\mathbf{C}) = \text{tr}\mathbf{U}'\mathbf{C}\mathbf{V}\mathbf{D} \leq \text{tr}\mathbf{D}$. Here, the upper bound $\text{tr}\mathbf{D}$ is attained for (A.4.5) as $f(\mathbf{U}\mathbf{V}') = \text{tr}\mathbf{V}\mathbf{D}\mathbf{U}'\mathbf{U}\mathbf{V}' = \text{tr}\mathbf{D}$, with (A.4.7) satisfying the constraints in (A.4.6) as $\mathbf{C}'\mathbf{C} = \mathbf{V}\mathbf{U}'\mathbf{U}\mathbf{V}' = \mathbf{V}\mathbf{V}' = \mathbf{I}_p$, because $\mathbf{U}'\mathbf{U} = \mathbf{V}'\mathbf{V} = \mathbf{I}_p$ and $\mathbf{V}$ being $p \times p$ implies $\mathbf{V}\mathbf{V}' = \mathbf{I}_p$. □

Theorem A.4.3
For the $n \times p$ matrix $\mathbf{X}$ in Theorem A.3.2, we consider the following problem:

$$\begin{aligned} \text{Maximize } f(\mathbf{A}, \mathbf{B}) = \text{tr}\mathbf{A}'\mathbf{X}\mathbf{B} \text{ over } \mathbf{A}(n \times m) \text{ and } \mathbf{B}(p \times m) \\ \text{s.t. } \mathbf{A}'\mathbf{A} = \mathbf{B}'\mathbf{B} = \mathbf{I}_m \text{ with } m \leq r = \text{rank}(\mathbf{X}). \end{aligned} \tag{A.4.6}$$

This is attained for

$$\mathbf{A} = \mathbf{K}_m\mathbf{T} \text{ and } \mathbf{B} = \mathbf{L}_m\mathbf{T} \tag{A.4.7}$$

with $\mathbf{K}_m$ and $\mathbf{L}_m$ defined as in (A.3.18) and $\mathbf{T}$ an $m \times m$ orthonormal matrix.

Proof By substituting (A.3.9) (the SVD of $\mathbf{X}$) in $f(\mathbf{A},\mathbf{B}) = \text{tr}\mathbf{A}'\mathbf{X}\mathbf{B}$, this is rewritten as $f(\mathbf{A},\mathbf{B}) = \text{tr}\mathbf{A}'\mathbf{K}\mathbf{\Lambda}\mathbf{L}'\mathbf{B} = \text{tr}\mathbf{L}'\mathbf{B}\mathbf{A}'\mathbf{K}\mathbf{\Lambda}$. As found in (A.3.10) and (A.4.6), $\mathbf{K}$, $\mathbf{L}$, $\mathbf{A}$, and $\mathbf{B}$ are column-orthonormal, and $\mathbf{L}'\mathbf{B}\mathbf{A}'\mathbf{K}$ is suborthonormal because of (A.4.1) and (A.4.2). Further, $\text{rank}(\mathbf{L}'\mathbf{B}\mathbf{A}'\mathbf{K}) \leq m \leq r$, while $\mathbf{\Lambda}$ is an $r \times r$ diagonal matrix with all diagonal elements positive. Those facts and Theorem A.4.1 lead to $f(\mathbf{A}, \mathbf{B}) = \text{tr}\mathbf{L}'\mathbf{B}\mathbf{A}'\mathbf{K}\mathbf{\Lambda} \leq \text{tr}\mathbf{\Lambda}_m$ with $\mathbf{\Lambda}_m$ defined as (A.3.22). Here, the upper bound

$\mathrm{tr}\boldsymbol{\Lambda}_m$ is attained for (A.4.7) as $f(\mathbf{K}_m\mathbf{T}, \mathbf{L}_m\mathbf{T}) = \mathrm{tr}\mathbf{T}'\mathbf{K}_m'(\mathbf{K}\boldsymbol{\Lambda}\mathbf{L}')\mathbf{L}_m\mathbf{T} = \mathrm{tr}\mathbf{L}'\mathbf{L}_m\mathbf{T}\mathbf{T}'\mathbf{K}_m'\mathbf{K}\boldsymbol{\Lambda} = \mathrm{tr}\mathbf{K}_m'\mathbf{K}\boldsymbol{\Lambda}\mathbf{L}'\mathbf{L}_m = \mathrm{tr}\boldsymbol{\Lambda}_m$, with (A.4.7) satisfying the constraints in (A.4.6) as $\mathbf{T}'\mathbf{K}_m'\mathbf{K}_m\mathbf{T} = \mathbf{T}'\mathbf{L}_m'\mathbf{L}_m\mathbf{T} = \mathbf{I}_m$ because of (A.1.6) and (A.3.10). □

Solution (A.4.7) shows that it is not unique; we can choose an arbitrary $m \times m$ orthonormal matrix as $\mathbf{T}$. Thus, we can choose $\mathbf{T}$ in the rotation methods described in Chap. 13, after obtaining (A.4.7) with $\mathbf{T} = \mathbf{I}_m$. Solutions that can be rotated as (A.4.7) are said to have *rotational indeterminacy*. This can be avoided by adding the following constraint to (A.4.6): $\mathbf{A}'\mathbf{X}\mathbf{B}$ is a diagonal matrix whose diagonal elements are arranged in descending order. Then, the solution is restricted to $\mathbf{A} = \mathbf{K}_m$ and $\mathbf{B} = \mathbf{L}_m$, which leads to $\mathbf{A}'\mathbf{X}\mathbf{B} = \boldsymbol{\Lambda}_m$.

Theorem A.4.4

For the $n \times p$ matrix $\mathbf{X}$ in Theorem A.3.2, we consider the following problem:

$$\begin{aligned} &\text{Maximize } f(\mathbf{W}) = \mathrm{tr}\mathbf{W}'\mathbf{X}'\mathbf{X}\mathbf{W} \text{ over } \mathbf{W}(p \times m) \\ &\qquad\qquad \text{s.t. } \mathbf{W}'\mathbf{W} = \mathbf{I}_m \text{ with } m \leq r = \mathrm{rank}(\mathbf{X}). \end{aligned} \tag{A.4.8}$$

This is attained for

$$\mathbf{W} = \mathbf{L}_m\mathbf{T}, \tag{A.4.9}$$

with $\mathbf{L}_m$ defined as in (A.3.18) and $\mathbf{T}$ an $m \times m$ orthonormal matrix. The matrix $\mathbf{L}_m$ can also be defined through the EVD of $\mathbf{X}'\mathbf{X}$ as described in Theorem A.3.3.

Proof By substituting (A.3.9) in $f(\mathbf{W})$, it is rewritten as $f(\mathbf{W}) = \mathrm{tr}\mathbf{W}'\mathbf{L}\boldsymbol{\Lambda}^2\mathbf{L}'\mathbf{W} = \mathrm{tr}\mathbf{L}'\mathbf{W}\mathbf{W}'\mathbf{L}\boldsymbol{\Lambda}^2$. As found in (A.3.9) and (A.4.8), $\mathbf{L}$ and $\mathbf{W}$ are column-orthonormal, and $\mathbf{L}'\mathbf{W}\mathbf{W}'\mathbf{L}$ is suborthonormal because of (A.4.1) and (A.4.2). Further, $\mathrm{rank}(\mathbf{L}'\mathbf{W}\mathbf{W}'\mathbf{L}) \leq m \leq r$, while $\boldsymbol{\Lambda}^2$ is an $r \times r$ diagonal matrix with all diagonal elements positive. This fact and Theorem A.4.1 lead to $f(\mathbf{W}) = \mathrm{tr}\mathbf{W}'\mathbf{X}'\mathbf{X}\mathbf{W} \leq \mathrm{tr}\boldsymbol{\Lambda}_m^2$, with $\boldsymbol{\Lambda}_m$ defined as (A.3.22). Here, the upper bound is attained for (A.4.9) as $f(\mathbf{L}_m\mathbf{T}) = \mathrm{tr}\mathbf{L}'\mathbf{L}_m\mathbf{T}\mathbf{T}'\mathbf{L}_m'\mathbf{L}\boldsymbol{\Lambda}^2 = \mathrm{tr}\boldsymbol{\Lambda}_m^2$, with (A.4.9) satisfying the constraint in (A.4.8) as $\mathbf{T}'\mathbf{L}_m'\mathbf{L}_m\mathbf{T} = \mathbf{I}_m$. □

Solution (A.4.9) also has rotational indeterminacy, which can be avoided by adding the constraint (6.4) (in Chap. 6) to (A.4.8).

A.4.3 Reduced Rank Approximation

In Chap. 5, *principal component analysis* (*PCA*) is introduced as a problem of obtaining the matrix product $\mathbf{FA}'$ that well approximates a data matrix $\mathbf{X}$, subject to the number of the columns of $\mathbf{F}$ and that of $\mathbf{A}$ being smaller than the rank of $\mathbf{X}$. Such a problem can be restated as approximating $\mathbf{X}$ by another matrix of lower rank and is called *reduced rank approximation*. The theorem for the approximation is presented next:

Theorem A.4.5. Reduced Rank Approximation
For the $n \times p$ matrix $\mathbf{X}$ in Theorem A.3.2, we consider the following problem:

$$\text{Minimize } f(\mathbf{M}) = \|\mathbf{X} - \mathbf{M}\|^2 \text{ over } \mathbf{M} \text{ s.t. } \text{rank}(\mathbf{M}) \leq m \leq \text{rank}(\mathbf{X}). \tag{A.4.10}$$

This is attained for

$$\mathbf{M} = \mathbf{K}_m \mathbf{\Lambda}_m \mathbf{L}_m'. \tag{A.4.11}$$

Here, it should be noted that the constraint in (A.4.10) is *rank*($\mathbf{M}$) equaling or being less than m, but solution (A.4.11) is restricted to *rank*($\mathbf{M}$) = m.

Proof Using the extended version of SVD (Theorem A.3.1) for $\mathbf{M}$, it is expressed as $\mathbf{M} = \mathbf{P\Omega Q}'$, with $\mathbf{P}'\mathbf{P} = \mathbf{Q}'\mathbf{Q} = \mathbf{I}_m$ and $\mathbf{\Omega}$ an $m \times m$ diagonal matrix whose elements are nonnegative. Then, $f(\mathbf{M})$ is rewritten as

$$\begin{aligned} f(\mathbf{P\Omega Q}') &= \|\mathbf{X} - \mathbf{P\Omega Q}'\|^2 \\ &= \|\mathbf{X} - \mathbf{XQQ}' + \mathbf{XQQ}' - \mathbf{P\Omega Q}'\|^2 \\ &= \|\mathbf{X} - \mathbf{XQQ}'\|^2 + \|\mathbf{XQQ}' - \mathbf{P\Omega Q}'\|^2 + 2c. \end{aligned} \tag{A.4.12}$$

Here, we can use $\mathbf{Q}'\mathbf{Q} = \mathbf{I}_m$ to get

$$\begin{aligned} c &= \text{tr}(\mathbf{X} - \mathbf{XQQ}')'(\mathbf{XQQ}' - \mathbf{P\Omega Q}') \\ &= \text{tr}\mathbf{X}'\mathbf{XQQ}' - \text{tr}\mathbf{X}'\mathbf{P\Omega Q}' - \text{tr}\mathbf{QQ}'\mathbf{X}'\mathbf{XQQ}' + \text{tr}\mathbf{QQ}'\mathbf{X}'\mathbf{P\Omega Q} \\ &= \text{tr}\mathbf{Q}'\mathbf{X}'\mathbf{XQ} - \text{tr}\mathbf{X}'\mathbf{P\Omega Q}' - \text{tr}\mathbf{Q}'\mathbf{X}'\mathbf{XQ} + \text{tr}\mathbf{X}'\mathbf{P\Omega Q}' = 0 \end{aligned} \tag{A.4.13}$$

and

$$\|\mathbf{X}-\mathbf{XQQ}'\|^2=\|\mathbf{X}\|^2-2\mathrm{tr}\mathbf{X}'\mathbf{XQQ}'+\mathrm{tr}\mathbf{QQ}'\mathbf{X}'\mathbf{XQQ}'=\|\mathbf{X}\|^2-\mathrm{tr}\mathbf{Q}'\mathbf{X}'\mathbf{XQ}. \tag{A.4.14}$$

Using (A.4.13) and (A.4.14) in (A.4.12), this is further rewritten as

$$f(\mathbf{P\Omega Q}')=\|\mathbf{X}\|^2-\mathrm{tr}\mathbf{Q}'\mathbf{X}'\mathbf{XQ}+\|\mathbf{XQQ}'-\mathbf{P\Omega Q}'\|^2. \tag{A.4.15}$$

This function can be minimized, if $\mathbf{P}$, $\mathbf{\Omega}$, and $\mathbf{Q}$ are found that simultaneously maximize $\mathrm{tr}\mathbf{Q}'\mathbf{X}'\mathbf{XQ}$ and minimize $\|\mathbf{XQQ}'-\mathbf{P\Omega Q}'\|^2$. Such $\mathbf{P}$, $\mathbf{\Omega}$, and $\mathbf{Q}$ are given by

$$\mathbf{P}=\mathbf{K}_m,\quad \mathbf{\Omega}=\mathbf{\Lambda}_m,\ \text{ and }\ \mathbf{Q}=\mathbf{L}_m, \tag{A.4.16}$$

which is shown as follows: (A.4.16) allows $\|\mathbf{XQQ}'-\mathbf{P\Omega Q}'\|^2$ to attain its lower limit, zero, as $\left\|\mathbf{XL}_m\mathbf{L}_m'-\mathbf{K}_m\mathbf{\Lambda}_m\mathbf{L}_m'\right\|^2=0$ because of (A.3.25), while $\mathbf{Q}=\mathbf{L}_m$ in (A.4.16) maximizes $\mathrm{tr}\mathbf{Q}'\mathbf{X}'\mathbf{XQ}$ subject to $\mathbf{Q}'\mathbf{Q}=\mathbf{I}_m$ because of Theorem A.4.4. The substitution of (A.4.16) in $\mathbf{M}=\mathbf{P\Omega Q}'$ leads to (A.4.11). □

Matrix $\mathbf{M}$ in Theorem A.4.5 can be replaced by

$$\mathbf{M}=\mathbf{FA}', \tag{A.4.17}$$

with $\mathbf{F}$ and $\mathbf{A}$ being $n\times m$ and $p\times m$ matrices, respectively. This replacement gives the formulation of principal component analysis in Chap. 5.

Theorem A.4.5 is referred to as Eckart and Young's (1936) theorem in some of the literature. The theorem has been proved in another manner by Takane (2014).

A.4.4 Modified Reduced Rank Approximation

In this section, we treat the *reduced rank approximation* problems for *generalized canonical correlation analysis* (*GCCA*) and *multiple correspondence analysis* (*MCA*). In this and the following sections, we use

$$\mathrm{rank}(\mathbf{PQR})=\mathrm{rank}(\mathbf{Q})\ \text{ if }\ \mathbf{P}\ \text{ and }\ \mathbf{R}\ \text{ are nonsingular} \tag{A.4.18}$$

(e.g., Lütkepohl, 1996).

Theorem A.4.6. GCCA Problems

For a given $n \times p$ block matrix $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_J]$ with its jth block $\mathbf{X}_j$ ($n \times p_j$), we consider the following problem:

$$\begin{aligned} \text{Minimize } \eta(\mathbf{F}, \mathbf{C}) = \sum_{j=1}^{J} \left\| \mathbf{F} - \mathbf{X}_j \mathbf{C}_j \right\|^2 \text{over } \mathbf{F} \text{ and } \mathbf{C} \\ \text{s.t. } \frac{1}{n}\mathbf{F}'\mathbf{F} = \mathbf{I}_m \text{ with } m \leq r = \text{rank}(\mathbf{X}). \end{aligned} \tag{A.4.19}$$

Here, $\mathbf{C} = \begin{bmatrix} \mathbf{C}_1 \\ \vdots \\ \mathbf{C}_J \end{bmatrix}$ is the $p \times m$ block matrix with its jth block, $\mathbf{C}_j$, being $p_j \times m$.

Problem (A.4.19) is equivalent to

$$\begin{aligned} \text{Minimize } f(\mathbf{F}, \mathbf{C}) = \left\| \mathbf{X}\mathbf{D}_{\mathrm{X}}^{-1/2} - \frac{1}{n}\mathbf{F}\mathbf{C}'\mathbf{D}_{\mathrm{X}}^{1/2} \right\|^2 \text{ over } \mathbf{F} \text{ and } \mathbf{C} \\ \text{s.t.} \frac{1}{n}\mathbf{F}'\mathbf{F} = \mathbf{I}_m, \text{ with } m \leq r = \text{rank}(\mathbf{X}). \end{aligned} \tag{A.4.20}$$

Here, $\mathbf{D}_{\mathrm{X}} = \begin{bmatrix} \mathbf{X}'_1\mathbf{X}_1 & & \\ & \ddots & \\ & & \mathbf{X}'_J\mathbf{X}_J \end{bmatrix}$ is the $p \times p$ nonsingular block diagonal matrix.

Those problems are solved through the SVD of $\mathbf{X}\mathbf{D}_{\mathrm{X}}^{-1/2}$, defined as

$$\mathbf{X}\mathbf{D}_{\mathrm{X}}^{-1/2} = \mathbf{N}\mathbf{\Phi}\mathbf{M}', \tag{A.4.21}$$

with $\mathbf{N}'\mathbf{N} = \mathbf{M}'\mathbf{M} = \mathbf{I}_r$ and $\mathbf{\Phi}$ a diagonal matrix whose diagonal elements are arranged in descending order. The minimization in (A.4.19) and (A.4.20) is attained for

$$\mathbf{F} = \sqrt{n}\mathbf{N}_m\mathbf{T} \text{ and } \mathbf{C} = \sqrt{n}\mathbf{D}_{\mathrm{X}}^{-1/2}\mathbf{M}_m\mathbf{\Phi}_m\mathbf{T}, \tag{A.4.22}$$

where $\mathbf{M}_m$ and $\mathbf{N}_m$ contain the first m columns of $\mathbf{M}$ and $\mathbf{N}$, respectively, $\mathbf{\Phi}_m$ is the first $m \times m$ diagonal block of $\mathbf{\Phi}$, and $\mathbf{T}$ is an $m \times m$ orthonormal matrix.

Proof The loss function in (A.4.19) can be expanded as

$$\begin{aligned}\eta(\mathbf{F},\mathbf{C}) &= J\mathrm{tr}\mathbf{F}'\mathbf{F} - 2\mathrm{tr}\mathbf{F}'\sum_{j=1}^{J}\mathbf{X},\mathbf{C}_j + \mathrm{tr}\sum_{j=1}^{J}\mathbf{C}_j'\mathbf{X}_j'\mathbf{X},\mathbf{C}_j \\ &= nmJ - 2\mathrm{tr}\mathbf{F}'\mathbf{X}\mathbf{C} + \mathrm{tr}\mathbf{C}'\mathbf{D}\mathrm{x}\mathbf{C},\end{aligned} \tag{A.4.23}$$

and the function in (A.4.20) multiplied by n is expanded as

$$\begin{aligned}n\times f(\mathbf{F},\mathbf{C}) &= n\mathrm{tr}\mathbf{X}\mathbf{D}_{\mathrm{X}}^{-1}\mathbf{X}' - 2\mathrm{tr}\mathbf{D}_{\mathrm{X}}^{-1/2}\mathbf{X}'\mathbf{F}\mathbf{C}'\mathbf{D}_{\mathrm{X}}^{1/2} + \frac{1}{n}\mathrm{tr}\mathbf{D}_{\mathrm{X}}^{1/2}\mathbf{C}\mathbf{F}'\mathbf{F}\mathbf{C}'\mathbf{D}_{\mathrm{X}}^{1/2} \\ &= n\mathrm{tr}\mathbf{X}\mathbf{D}_{\mathrm{X}}^{-1}\mathbf{X}' - 2\mathrm{tr}\mathbf{X}'\mathbf{F}\mathbf{C}' + \mathrm{tr}\mathbf{C}'\mathbf{D}_{\mathrm{X}}\mathbf{C},\end{aligned} \tag{A.4.24}$$

where the constraint $n^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_m$ has been used. Since the parts relevant to $\mathbf{F}$ and $\mathbf{C}$ in (A.4.23) are the same as those in (A.4.24), the problems (A.4.19) and (A.4.20) with the same constraints are equivalent.

Because of (A.4.18), $r = \mathrm{rank}(\mathbf{X}) = \mathrm{rank}(\mathbf{X}\mathbf{D}_{\mathrm{X}}^{-1/2})$, while $\mathrm{rank}(n^{-1}\mathbf{F}\mathbf{C}'\mathbf{D}_{\mathrm{X}}^{1/2}) \leq m \leq r$. Thus, problem (A.4.20) is the reduced rank approximation of $\mathbf{X}\mathbf{D}_{\mathrm{X}}^{-1/2}$ by $n^{-1}\mathbf{F}'\mathbf{C}'\mathbf{D}_{\mathrm{X}}^{1/2}$ as the approximation of $\mathbf{X}$ by $\mathbf{M}$ in Theorem A.4.5; the minimization in (A.4.20) is attained for

$$\frac{1}{n}\mathbf{F}\mathbf{C}'\mathbf{D}_{\mathrm{X}}^{1/2} = \mathbf{N}_m\boldsymbol{\Phi}_m\mathbf{M}_m'. \tag{A.4.25}$$

Matrices $\mathbf{F}$ and $\mathbf{C}$ in (A.4.22) satisfy (A.4.25) and the constraints in (A.4.19) and (A.4.20). □

The constraint of $\mathbf{F}$ being *centered* is added to the above problems in those that follow:

Theorem A.4.7. MCA Problems

Let us suppose that an $n \times K$ block matrix $\mathbf{G} = [\mathbf{G}_1, \ldots, \mathbf{G}_J]$ is given, with its jth block $\mathbf{G}_j$ ($n \times K_j$) defined as (14.33) with (14.34). For $\mathbf{G}$, we consider the following problem:

$$\begin{aligned}&\text{Minimize } \eta(\mathbf{F},\mathbf{C}) = \sum_{j=1}^{J}\left\|\mathbf{F} - \mathbf{G}_j\mathbf{C}_j\right\|^2 \text{ over } \mathbf{F} \text{ and } \mathbf{C} \\ &\qquad\text{s.t. } \frac{1}{n}\mathbf{F}'\mathbf{F} = \mathbf{I}_m, \mathbf{J}\mathbf{F} = \mathbf{F}, \text{and } m \leq r = \mathrm{rank}(\mathbf{J}\mathbf{G}).\end{aligned} \tag{A.4.26}$$

Here, $\mathbf{J}$ is defined as (2.10) and $\mathbf{C} = \begin{bmatrix} \mathbf{C}_1 \\ \vdots \\ \mathbf{C}_J \end{bmatrix}$ is the $K \times m$ block matrix with its jth block, $\mathbf{C}_j$ ($K_j \times m$).

Problem (A.4.26) is equivalent to

$$\text{Minimize } f(\mathbf{F}, \mathbf{C}) = \left\| \mathbf{J}\mathbf{G}\mathbf{D}_{\mathrm{G}}{}^{-1/2} - \frac{1}{n}\mathbf{F}\mathbf{C}'\mathbf{D}_{\mathrm{G}}^{1/2} \right\|^2 \text{ over } \mathbf{F} \text{ and } \mathbf{C}$$
$$\text{s.t.} \frac{1}{n}\mathbf{F}'\mathbf{F} = \mathbf{I}_m, \mathbf{J}\mathbf{F} = \mathbf{F}, \text{and } m \leq r = \text{rank}(\mathbf{J}\mathbf{G}). \tag{A.4.27}$$

Here, $\mathbf{D}_{\mathrm{G}} = \begin{bmatrix} \mathbf{G}'_1\mathbf{G}_1 & & \\ & \ddots & \\ & & \mathbf{G}'_J\mathbf{G}_J \end{bmatrix}$ is the $K \times K$ nonsingular block diagonal matrix, which is a simply diagonal matrix as explained Sect. 14.5.

Those problems are solved through the SVD of $\mathbf{J}\mathbf{G}\mathbf{D}_{\mathrm{G}}^{-1/2}$, defined as

$$\mathbf{J}\mathbf{G}\mathbf{D}_{\mathrm{G}}^{-1/2} = \mathbf{S}\mathbf{\Theta}\mathbf{P}', \tag{A.4.28}$$

with $\mathbf{S}'\mathbf{S} = \mathbf{P}'\mathbf{P} = \mathbf{I}_r$ and $\mathbf{\Theta}$ a diagonal matrix whose diagonal elements are arranged in descending order. The minimization in (A.4.26) and (A.4.27) is attained for

$$\mathbf{F} = \sqrt{n}\mathbf{S}_m\mathbf{T} \text{ and } \mathbf{C} = \sqrt{n}\mathbf{D}_{\mathrm{G}}^{-1/2}\mathbf{P}_m\mathbf{\Theta}_m\mathbf{T}, \tag{A.4.29}$$

where $\mathbf{S}_m$ and $\mathbf{P}_m$ contain the first m columns of $\mathbf{S}$ and $\mathbf{P}$, respectively, $\mathbf{\Theta}_m$ is the first $m \times m$ diagonal block of $\mathbf{\Theta}$, and $\mathbf{T}$ is an $m \times m$ orthonormal matrix.

Proof The loss function in (A.4.26) can be expanded as

$$\begin{aligned} \eta(\mathbf{F}, \mathbf{C}) &= J\text{tr}\mathbf{F}'\mathbf{F} - 2\text{tr}\mathbf{F}'\sum_{j=1}^{J}\mathbf{G}_j\mathbf{C}_j + \text{tr}\sum_{j=1}^{J}\mathbf{C}'_j\mathbf{G}'_j\mathbf{G}_j\mathbf{C}_j \\ &= nmJ - 2\text{tr}\mathbf{F}'\mathbf{J}\mathbf{G}\mathbf{C} + \text{tr}\mathbf{C}'\mathbf{D}_{\mathrm{G}}\mathbf{C}, \end{aligned} \tag{A.4.30}$$

where we have used the constraints $n^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_m$ and $\mathbf{J}\mathbf{F} = \mathbf{F}$. On the other hand, (A.4.27) multiplied by n is expanded as

$$
\begin{aligned}
n \times f(\mathbf{F}, \mathbf{C}) &= n\mathrm{tr}\mathbf{J}\mathbf{G}\mathbf{D}_{\mathrm{G}}^{-1}\mathbf{G}\mathbf{J}' - 2\mathrm{tr}\mathbf{D}_{\mathrm{G}}^{-1/2}\mathbf{G}'\mathbf{J}\mathbf{F}\mathbf{C}'\mathbf{D}_{\mathrm{G}}^{1/2} \\
&\quad + \frac{1}{n}\mathrm{tr}\mathbf{D}_{\mathrm{G}}^{1/2}\mathbf{C}\mathbf{F}'\mathbf{F}\mathbf{C}'\mathbf{D}_{\mathrm{G}}^{1/2}\mathbf{D}_{\mathrm{G}}^{1/2} \\
&= n\mathrm{tr}\mathbf{J}\mathbf{G}\mathbf{D}_{\mathbf{G}}^{-1}\mathbf{G}\mathbf{J}' - 2\mathrm{tr}\mathbf{G}'\mathbf{J}\mathbf{F}\mathbf{C}' + \mathrm{tr}\mathbf{C}'\mathbf{D}_{\mathrm{G}}\mathbf{C},
\end{aligned}
\tag{A.4.31}
$$

where the constraint $n^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_m$ has been used. Since the parts relevant to $\mathbf{F}$ and $\mathbf{C}$ in (A.4.30) are the same as those in (A.4.31), problems (A.4.26) and (A.4.27) with the same constraints are equivalent.

Because of (A.4.18), $r = \mathrm{rank}(\mathbf{J}\mathbf{G}) = \mathrm{rank}\left(\mathbf{J}\mathbf{G}\mathbf{D}_{\mathrm{G}}^{-1/2}\right)$, while $\mathrm{rank}(n^{-1}\mathbf{F}'\mathbf{C}'\mathbf{D}_{\mathrm{G}}^{1/2}) \leq m \leq r$. Thus, problem (A.4.27) is the reduced rank approximation of $\mathbf{J}\mathbf{G}\mathbf{D}_{\mathrm{G}}^{-1/2}$ by $n^{-1}\mathbf{F}\mathbf{C}'\mathbf{D}_{\mathrm{G}}^{1/2}$ as the approximation of $\mathbf{X}$ by $\mathbf{M}$ in Theorem A.4.5; the minimization in (A.4.27) is attained for

$$
\frac{1}{n}\mathbf{F}\mathbf{C}'\mathbf{D}_{\mathrm{G}}^{1/2} = \mathbf{S}_m\mathbf{\Theta}_m\mathbf{P}'_m. \tag{A.4.32}
$$

The $\mathbf{F}$ and $\mathbf{C}$ in (A.4.29) satisfy (A.4.32) and the constraints in (A.4.26) and (A.4.27), where $\mathbf{J}\mathbf{F} = \mathbf{F}$ follows from the fact that $\mathbf{F} = \sqrt{n}\mathbf{S}_m\mathbf{T}$, in (A.4.29), can be rewritten as $\mathbf{F} = \sqrt{n}\mathbf{S}\mathbf{\Theta}\mathbf{P}'\mathbf{P}_m\mathbf{\Theta}_m^{-1}\mathbf{T} = \sqrt{n}\mathbf{J}\mathbf{G}\mathbf{D}_{\mathrm{G}}^{-1/2}\mathbf{P}_m\mathbf{\Theta}_m^{-1}\mathbf{T}$ with (2.12). □

The GCCA and MCA solutions (A.4.22) and (A.4.29) show that they have rotational indeterminacy. This can be avoided, if the constraint

$$
\begin{aligned}
&\mathbf{C}'\mathbf{D}_{\mathrm{G}}\mathbf{C} \text{ being a diagonal matrix whose} \\
&\text{diagonal elements are arranged in descending order}
\end{aligned}
\tag{A.4.33}
$$

is added to (A.4.26) and (A.4.27) for the MCA solution. Since (A.4.29) leads to $\mathbf{C}'\mathbf{D}_{\mathrm{G}}\mathbf{C} = \mathbf{T}'\mathbf{\Theta}_m^2\mathbf{T}$, (A.4.33) requires $\mathbf{T} = \mathbf{I}_m$. The indeterminacy of the GCCA solution can also be avoided, by adding the constraint (A.4.33) with $\mathbf{D}_{\mathrm{G}}$ replaced by $\mathbf{D}_{\mathrm{X}}$ to (A.4.19) and (A.4.20), so that the GCCA solution is unique. Then, $\mathbf{T}$ in (A.4.22) is fixed to $\mathbf{I}_m$.

A.4.5 Modified Versions of Maximizing Trace Functions

In A.4.2, the parameter matrix $\mathbf{C}$ was constrained as $\mathbf{C}'\mathbf{C}$ being the identity matrix. In this section, $\mathbf{C}$ is constrained rather as $\mathbf{C}'\mathbf{V}\mathbf{C}$ being the identity matrix with $\mathbf{V}$ a *given positive definite matrix* (Note 8.2), and the *symmetric square roots* $\mathbf{V}^{1/2}$ and $\mathbf{V}^{-1/2}$ are used that satisfy $\mathbf{V}^{1/2}\mathbf{V}^{1/2} = \mathbf{V}$ and $\mathbf{V}^{-1/2}\mathbf{V}^{-1/2} = \mathbf{V}^{-1}$, respectively. How to obtain $\mathbf{V}^{1/2}$ and $\mathbf{V}^{-1/2}$ from $\mathbf{V}$ is described in the following section.

Theorem A.4.8.
Let us define matrices as $\mathbf{V}_{11}$ $(p_1 \times p_1)$, $\mathbf{V}_{22}$ $(p_2 \times p_2)$, and $\mathbf{V}_{12}$ $(p_1 \times p_2)$, with $\mathbf{V}_{11}$ and $\mathbf{V}_{22}$ symmetric and positive definite. We consider the following problem:

$$\begin{aligned} &\text{Maximize } \mathrm{tr}\mathbf{C}_1'\mathbf{V}_{12}\mathbf{C}_2 \text{ over } \mathbf{C}_1(p_1 \times m) \text{ and } \mathbf{C}_2(p_2 \times m) \\ &\text{s.t. } \mathbf{C}_1'\mathbf{V}_{11}\mathbf{C}_1 = \mathbf{C}_2'\mathbf{V}_{22}\mathbf{C}_2 = \mathbf{I}_m \text{ with } m \le r = \mathrm{rank}(\mathrm{V}_{12}). \end{aligned} \tag{A.4.34}$$

It is solved through the SVD of $\mathbf{V}_{11}^{-1/2}\mathbf{V}_{12}\mathbf{V}_{22}^{-1/2}$ defined as

$$\mathbf{V}_{11}^{-1/2}\mathbf{V}_{12}\mathbf{V}_{22}^{-1/2} = \mathbf{H}\boldsymbol{\Omega}\mathbf{R}', \tag{A.4.35}$$

with $\mathbf{H}'\mathbf{H} = \mathbf{R}'\mathbf{R} = \mathbf{I}_r$ and $\boldsymbol{\Omega}$ the diagonal matrix whose diagonal elements are arranged in descending order. The maximization in (A.4.34) is attained for

$$\mathbf{C}_1 = \mathbf{V}_{11}^{-1/2}\mathbf{H}_m\mathbf{T} \text{ and } \mathbf{C}_2 = \mathbf{V}_{22}^{-1/2}\mathbf{R}_m\mathbf{T}, \tag{A.4.36}$$

where $\mathbf{H}_m$ and $\mathbf{R}_m$ contain the first m columns of $\mathbf{H}$ and those of $\mathbf{R}$, respectively, and $\mathbf{T}$ is an $m \times m$ orthonormal matrix.

Proof By defining $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{Y}$ as

$$\mathbf{A} = \mathbf{V}_{11}^{1/2}\mathbf{C}_1, \quad \mathbf{B} = \mathbf{V}_{22}^{1/2}\mathbf{C}_2, \tag{A.4.37}$$

$$\mathbf{Y} = \mathbf{V}_{11}^{-1/2}\mathbf{V}_{12}\mathbf{V}_{22}^{-1/2}, \tag{A.4.38}$$

(A.4.34) can be transformed into the equivalent problem:

$$\begin{aligned} &\text{Maximize } \mathrm{tr}\mathbf{A}'\mathbf{Y}\mathbf{B} \text{ over } \mathbf{A}(p_1 \times m) \text{ and } \mathbf{B}(p_2 \times m) \\ &\text{s.t. } \mathbf{A}'\mathbf{A} = \mathbf{B}'\mathbf{B} = \mathbf{I}_m \text{ with } m \le r = \mathrm{rank}(\mathbf{Y}), \end{aligned} \tag{A.4.39}$$

where we have used r = rank($\mathbf{V}_{12}$) = rank($\mathbf{Y}$), following from (A.4.18). Since problem (A.4.39) is equivalent to (A.4.6) in Theorem A.4.3, the solution for (A.4.39) is given by

$$\mathbf{A} = \mathbf{H}_m\mathbf{T} \text{ and } \mathbf{B} = \mathbf{R}_m\mathbf{T}, \tag{A.4.40}$$

when the SVD of (A.4.38) is defined as (A.4.35). Using (A.4.37) in (A.4.40), we have (A.4.36). □

A related theorem is given next:

Theorem A.4.9.
Let $\mathbf{V}$ be a $p \times p$ symmetric positive definite matrix and $\mathbf{M}$ be a $p \times p$ symmetric and nonnegative definite matrix with its rank r. We consider the following problem:

$$\text{Maximize}\,\text{tr}\,\mathbf{B}'\mathbf{M}\mathbf{B}\ \text{over}\ \mathbf{B}(p \times m)\text{s.t.}\mathbf{B}'\mathbf{V}\mathbf{B} = \mathbf{I}_m \text{with}\ m \leq r = \text{rank}(\mathbf{M}). \tag{A.4.41}$$

This is solved through the EVD of $\mathbf{V}^{-1/2}\mathbf{M}\mathbf{V}^{-1/2}$ defined as

$$\mathbf{V}^{-1/2}\mathbf{M}\mathbf{V}^{-1/2} = \mathbf{Q}\boldsymbol{\Theta}^2\mathbf{Q}', \tag{A.4.42}$$

with $\mathbf{Q}'\mathbf{Q} = \mathbf{I}_r$ and $\boldsymbol{\Theta}$ the diagonal matrix whose diagonal elements are arranged in descending order. The maximization in (A.4.41) is attained for

$$\mathbf{B} = \mathbf{V}^{-1/2}\mathbf{Q}_m\mathbf{T}, \tag{A.4.43}$$

where $\mathbf{Q}_m$ contains the first m columns of $\mathbf{Q}$, and $\mathbf{T}$ is an $m \times m$ orthonormal matrix.

Proof By defining $\mathbf{W}$ and $\mathbf{Y}$ as

$$\mathbf{W} = \mathbf{V}^{1/2}\mathbf{B}, \tag{A.4.44}$$

$$\mathbf{Y} = \mathbf{V}^{-1/2}\mathbf{M}\mathbf{V}^{-1/2}, \tag{A.4.45}$$

(A.4.41) can be transformed into the equivalent problem:

$$\text{Maximize}\,\text{tr}\,\mathbf{W}'\mathbf{Y}\mathbf{W}\ \text{over}\ \mathbf{W}\,(p \times m)\,\text{s.t.}\mathbf{W}'\mathbf{W} = \mathbf{I}_m\,\text{with}\ m \leq r = \text{rank}(\mathbf{Y}), \tag{A.4.46}$$

where we have used r = rank($\mathbf{M}$) = rank($\mathbf{Y}$), following from (A.4.18). Since (A.4.46) is equivalent to (A.4.8) in Theorem A.4.4, the solution for (A.4.46) is given by

$$\mathbf{W} = \mathbf{Q}_m\mathbf{T}, \tag{A.4.47}$$

when the EVD of (A.4.45) is defined as (A.4.42). Using (A.4.44) in (A.4.47), we have (A.4.43). □

The solution of (A.4.36) is found to have rotational indeterminacy. Also, it is possessed by (A.4.43). This indeterminacy is avoided by adding the following constraint to (A.4.41): $\mathbf{B}'\mathbf{M}\mathbf{B}$ is a diagonal matrix whose diagonal elements are arranged in descending order. Then, the solution is restricted to $\mathbf{B} = \mathbf{V}^{-1/2}\mathbf{Q}_m$. This solution has been used for the canonical discriminant analysis in Chap. 15.

A.4.6 Obtaining Symmetric Square Roots of Matrices

Let $\mathbf{V} = \mathbf{U}\mathbf{U}'$ be a $p \times p$ *positive definite symmetric* matrix. As in (A.3.13), the SVD of $\mathbf{V} = \mathbf{U}\mathbf{U}'$ can be defined as $\mathbf{V} = \boldsymbol{\Gamma}\boldsymbol{\Delta}^2\boldsymbol{\Gamma}'$ with $\boldsymbol{\Gamma}'\boldsymbol{\Gamma} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}' = \mathbf{I}_p$ and $\boldsymbol{\Delta}^2$ a $p \times p$ diagonal matrix. The p diagonal elements of $\boldsymbol{\Delta}^2$ are all positive with $\mathrm{rank}(\mathbf{V}) = p$, since any positive definite matrix is nonsingular as written in Note 8.2.

The *symmetric square root* of $\mathbf{V}$ is given by

$$\mathbf{V}^{1/2} = \boldsymbol{\Gamma}\boldsymbol{\Delta}\boldsymbol{\Gamma}', \tag{A.4.48}$$

with each diagonal element of $\boldsymbol{\Delta}$ being the square root of the corresponding one of $\boldsymbol{\Delta}^2$. We can easily verify that $\mathbf{V}^{1/2}\mathbf{V}^{1/2} = \boldsymbol{\Gamma}\boldsymbol{\Delta}\boldsymbol{\Gamma}'\boldsymbol{\Gamma}\boldsymbol{\Delta}\boldsymbol{\Gamma}' = \boldsymbol{\Gamma}\boldsymbol{\Delta}\boldsymbol{\Delta}\boldsymbol{\Gamma}' = \boldsymbol{\Gamma}\boldsymbol{\Delta}^2\boldsymbol{\Gamma}' = \mathbf{V}$.

The *inverse* matrix of $\mathbf{V}$ is expressed as $\mathbf{V}^{-1} = \boldsymbol{\Gamma}\boldsymbol{\Delta}^{-2}\boldsymbol{\Gamma}'$. Its symmetric square root is given by

$$\mathbf{V}^{-1/2} = \boldsymbol{\Gamma}\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}', \tag{A.4.49}$$

with each diagonal element of $\boldsymbol{\Delta}^{-1}$ being the reciprocal of the square root of the corresponding element in $\boldsymbol{\Delta}^2$. We can easily verify that $\mathbf{V}^{-1/2}\mathbf{V}^{-1/2} = \boldsymbol{\Gamma}\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}'\boldsymbol{\Gamma}\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}' = \boldsymbol{\Gamma}\boldsymbol{\Delta}^{-1}\boldsymbol{\Delta}^{-1} = \boldsymbol{\Gamma}\boldsymbol{\Delta}^{-2}\boldsymbol{\Gamma}' = \mathbf{V}^{-1}$.

Next, we consider the *symmetric square root* of the *block diagonal* matrix $\mathbf{D} = \begin{bmatrix} \mathbf{V}_1 & & \\ & \ddots & \\ & & \mathbf{V}_J \end{bmatrix} = \begin{bmatrix} \mathbf{U}_1'\mathbf{U}_1 & & \\ & \ddots & \\ & & \mathbf{U}_J'\mathbf{U}_J \end{bmatrix}$, which is symmetric and positive definite. These properties imply that the diagonal blocks $\mathbf{V}_j = \mathbf{U}_j'\mathbf{U}_j$ ($p_j \times p_j$) ($j = 1, \ldots, J$) are also symmetric and positive definite. Thus, the SVD of $\mathbf{V}_j$ can be defined as $\mathbf{V}_j = \boldsymbol{\Gamma}_j\boldsymbol{\Delta}_j^2\boldsymbol{\Gamma}_j'$, with $\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j = \boldsymbol{\Gamma}_j\boldsymbol{\Gamma}_j' = \mathbf{I}_{p_j}$ and $\boldsymbol{\Delta}_j^2$ the $p_j \times p_j$ diagonal matrix whose p_j diagonal elements are positive.

The symmetric square root of $\mathbf{D}$ is given by

$$\mathbf{D}^{1/2} = \begin{bmatrix} \mathbf{V}_1^{1/2} & & \\ & \ddots & \\ & & \mathbf{V}_J^{1/2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Gamma}_1\boldsymbol{\Delta}_1\boldsymbol{\Gamma}_1' & & \\ & \ddots & \\ & & \boldsymbol{\Gamma}_J\boldsymbol{\Delta}_J\boldsymbol{\Gamma}_J' \end{bmatrix} \tag{A.4.50}$$

and the root of $\mathbf{D}^{-1}$ is given by

$$\mathbf{D}^{1/2} = \begin{bmatrix} \mathbf{V}_1^{-1/2} & & \\ & \ddots & \\ & & \mathbf{V}_J^{-1/2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Gamma}_1\boldsymbol{\Delta}_1^{-1}\boldsymbol{\Gamma}_1' & & \\ & \ddots & \\ & & \boldsymbol{\Gamma}_J\boldsymbol{\Delta}_J^{-1}\boldsymbol{\Gamma}_J' \end{bmatrix}. \tag{A.4.51}$$

We can verify $\mathbf{D}^{1/2}\mathbf{D}^{1/2} = \mathbf{D}$ and $\mathbf{D}^{-1/2}\mathbf{D}^{-1/2} = \mathbf{D}^{-1}$ from the fact that $\boldsymbol{\Gamma}_j\boldsymbol{\Delta}_j\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j\boldsymbol{\Delta}_j\boldsymbol{\Gamma}_j' = \boldsymbol{\Gamma}_j\boldsymbol{\Delta}_j^2\boldsymbol{\Gamma}_j' = \mathbf{V}_j$ and $\boldsymbol{\Gamma}_j\boldsymbol{\Delta}_j^{-1}\boldsymbol{\Gamma}_j'\boldsymbol{\Gamma}_j\boldsymbol{\Delta}_j^{-1}\boldsymbol{\Gamma}_j' = \boldsymbol{\Gamma}_j\boldsymbol{\Delta}_j^{-2}\boldsymbol{\Gamma}_j' = \mathbf{V}_j^{-1}$.

Since $\mathbf{D}_{\mathrm{G}} = \begin{bmatrix} \mathbf{G}'_1\mathbf{G}_1 & & \\ & \ddots & \\ & & \mathbf{G}'_J\mathbf{G}_J \end{bmatrix}$ in Theorem A.4.7 is diagonal, its square root $\mathbf{D}_{\mathrm{G}}^{1/2}$ is simply the diagonal matrix whose diagonal elements are the square roots of the corresponding ones in $\mathbf{D}_{\mathrm{G}}$. On the other hand, the root of $\mathbf{D}_{\mathrm{G}}^{-1}$ is given by $\mathbf{D}_{\mathrm{G}}^{-1/2}$, whose diagonal elements are the reciprocals of the square roots of the corresponding elements in $\mathbf{D}_{\mathrm{G}}$.

For the $p \times p$ symmetric positive definite matrix $\mathbf{V} = \mathbf{U}\mathbf{U}'$ which appeared first in this section, $\mathbf{U}$ is called the *square root* of $\mathbf{V}$. It is given by $\mathbf{U} = \boldsymbol{\Gamma}\boldsymbol{\Delta}$, using $\mathbf{V} = \boldsymbol{\Gamma}\boldsymbol{\Delta}^2\boldsymbol{\Gamma}'$. The root $\mathbf{U}$ can also be used for solving the problems in the previous appendices. However, we must be careful about whether $\mathbf{U}$ or $\mathbf{U}'$ is used in solutions, as $\mathbf{U} = \boldsymbol{\Gamma}\boldsymbol{\Delta}$ is *not symmetric*, which differs from the symmetric matrices in (A.4.48–A.4.51). Therefore, we chose to use the *symmetric* roots in this book.

A.5 Normal Maximum Likelihood Estimates

We derive the *maximum likelihoods* of *mean* vectors and *covariance* matrices for the *multivariate normal distributions*, which are used in Chaps. 8 and 15.

A.5.1 Estimates of Means and Covariances

Log likelihood (8.20) is presented again here:

$$l(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{l=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}), \tag{A.5.1}$$

In this appendix, it is shown that the maximum likelihood estimates (MLE) of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ maximizing (A.5.1) are given by (8.21) and (8.22), i.e.,

$$\hat{\boldsymbol{\mu}} = \overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i, \tag{A.5.2}$$

$$\hat{\boldsymbol{\Sigma}} = \mathbf{V} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})', \tag{A.5.3}$$

respectively, on the supposition of $\boldsymbol{\Sigma}$ being positive definite.

For proving that (A.5.2) is the MLE of $\boldsymbol{\mu}$, we can use an extension of the *decomposition of the sum of squares* treated in Appendix A.2.1, as follows: In the right-hand side of (A.5.1), only the second term is relevant to $\boldsymbol{\mu}$; thus, its maximum likelihood estimate is the $\boldsymbol{\mu}$ minimizing that term multiplied by -2:

$$\begin{aligned}\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}) &= \sum_{i=1}^{n}(\mathbf{x}_i-\overline{\mathbf{x}}+\overline{\mathbf{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\overline{\mathbf{x}}+\overline{\mathbf{x}}-\boldsymbol{\mu})\\ &= \sum_{i=1}^{n}(\mathbf{x}_i-\overline{\mathbf{x}})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\overline{\mathbf{x}})+n(\overline{\mathbf{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\overline{\mathbf{x}}-\boldsymbol{\mu})+2c.\end{aligned} \tag{A.5.4}$$

Here, c is found to be zero as

$$\begin{aligned}c &= \sum_{i=1}^{n}(\mathbf{x}_i-\overline{\mathbf{x}})'\boldsymbol{\Sigma}^{-1}(\overline{\mathbf{x}}-\boldsymbol{\mu})\\ &= \sum_{i=1}^{n}\mathbf{x}'_i\boldsymbol{\Sigma}^{-1}\overline{\mathbf{x}}-\sum_{i=1}^{n}\mathbf{x}'_i\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\sum_{i=1}^{n}\overline{\mathbf{x}}'\boldsymbol{\Sigma}^{-1}\mathbf{x}+\sum_{i=1}^{n}\overline{\mathbf{x}}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\\ &= n\overline{\mathbf{x}}'\boldsymbol{\Sigma}^{-1}\overline{\mathbf{x}}-n\overline{\mathbf{x}}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-n\overline{\mathbf{x}}'\boldsymbol{\Sigma}^{-1}\overline{\mathbf{x}}+n\overline{\mathbf{x}}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}=0.\end{aligned} \tag{A.5.5}$$

This implies that the term relevant to $\boldsymbol{\mu}$ in (A.5.4) is only $n(\overline{\mathbf{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\overline{\mathbf{x}}-\boldsymbol{\mu})$, which attains the lower limit, zero, for (A.5.2); it gives the MLE for $\boldsymbol{\mu}$.

Substituting (A.5.2) in (A.5.1), it is rewritten as

$$\begin{aligned}l(\boldsymbol{\Sigma}) &= -\frac{n}{2}\log|\boldsymbol{\Sigma}|-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i-\overline{\mathbf{x}})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\overline{\mathbf{x}})\\ &= -\frac{n}{2}\log|\boldsymbol{\Sigma}|-\frac{n}{2}\mathrm{tr}\boldsymbol{\Sigma}^{-1}\left\{\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\overline{\mathbf{x}})(\mathbf{x}_i-\overline{\mathbf{x}})'\right\}\\ &= -\frac{n}{2}\log|\boldsymbol{\Sigma}|-\frac{n}{2}\mathrm{tr}\boldsymbol{\Sigma}^{-1}\mathbf{V}=-\frac{n}{2}\left(\log|\boldsymbol{\Sigma}|+\mathrm{tr}\boldsymbol{\Sigma}^{-1}\mathbf{V}\right).\end{aligned} \tag{A.5.6}$$

This shows that our remaining task is to minimize $\log|\boldsymbol{\Sigma}|+\mathrm{tr}\boldsymbol{\Sigma}^{-1}\mathbf{V}=\mathrm{tr}\boldsymbol{\Sigma}^{-1}\mathbf{V}-\log\left|\boldsymbol{\Sigma}^{-1}\right|$ over $\boldsymbol{\Sigma}$, which is equivalent to minimizing

$$\begin{aligned}g(\boldsymbol{\Sigma}) &= \mathrm{tr}\boldsymbol{\Sigma}^{-1}\mathbf{V}-\log\left|\boldsymbol{\Sigma}^{-1}\right|-\log|\mathbf{V}|=\mathrm{tr}\boldsymbol{\Sigma}^{-1}\mathbf{V}-\log\left|\boldsymbol{\Sigma}^{-1}\mathbf{V}\right|\\ &= \mathrm{tr}\mathbf{V}^{1/2}\boldsymbol{\Sigma}^{-1}\mathbf{V}^{1/2}-\log\left|\mathbf{V}^{1/2}\boldsymbol{\Sigma}^{-1}\mathbf{V}^{1/2}\right|.\end{aligned} \tag{A.5.7}$$

where we have used (8.11).

From Note 8.2, we can express $\boldsymbol{\Sigma}^{-1}$ as $\boldsymbol{\Sigma}^{-1} = \mathbf{U}\mathbf{U}'$. Let the SVD of $\mathbf{U}'\mathbf{V}^{1/2}$ be defined as $\mathbf{U}'\mathbf{V}^{1/2} = \boldsymbol{\Xi}\boldsymbol{\Omega}\boldsymbol{\Gamma}'$, which implies

$$\mathbf{V}^{1/2}\boldsymbol{\Sigma}^{-1}\mathbf{V}^{1/2} = \boldsymbol{\Gamma}\boldsymbol{\Omega}^2\boldsymbol{\Gamma}' \tag{A.5.8}$$

with $\boldsymbol{\Gamma}\boldsymbol{\Gamma}' = \boldsymbol{\Gamma}'\boldsymbol{\Gamma} = \mathbf{I}_p$ and $\boldsymbol{\Omega}^2$ a diagonal matrix whose jth diagonal element is $\omega_j > 0$. Using (A.5.8) in (A.5.7), this can be rewritten as

$$\begin{aligned} g(\boldsymbol{\Sigma}) &= \mathrm{tr}\boldsymbol{\Omega}^2 - \log\left(\left|\boldsymbol{\Omega}^2\right| \times |\boldsymbol{\Gamma}| \times |\boldsymbol{\Gamma}'|\right) \\ &= \mathrm{tr}\boldsymbol{\Omega}^2 - \log\left|\boldsymbol{\Omega}^2\right| = \sum_{j=1}^{p} \omega_j - \sum_{j=1}^{p} \log\omega_j = \sum_{j=1}^{p} h(\omega_j), \end{aligned} \tag{A.5.7$'$}$$

with $h(\omega_j) = \omega_j - \log\,\omega_j$. Here, we have used the fact that $\boldsymbol{\Gamma}' = \boldsymbol{\Gamma}^{-1}$ and (8.12) leads to $|\boldsymbol{\Gamma}| \times |\boldsymbol{\Gamma}'| = 1$. It is known that the *differentiation* of $h(\omega_j)$ with respect to ω_j is given by $h'(\omega_j) = \mathrm{d}h(\omega_j)/\mathrm{d}\omega_j = 1 - 1/\omega_j$, which is found to satisfy

$$\begin{aligned} &h'(\omega_j) < 0 \ \text{ for } \ 0 < \omega_j < 1, \\ &h'(\omega_j) = 0 \ \text{ for } \ \omega_j = 1, \\ &h'(\omega_j) > 0 \ \text{ for } \ \omega_j > 1. \end{aligned} \tag{A.5.9}$$

This shows that (A.5.7′) is minimized for $\omega_j = 1 (j = 1, \ldots, p)$, i.e., $\boldsymbol{\Omega}^2 = \mathbf{I}_p$. Using this, (A.5.8) is rewritten as

$$\mathbf{V}^{1/2}\boldsymbol{\Sigma}^{-1}\mathbf{V}^{1/2} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}' = \mathbf{I}_p, \tag{A.5.10}$$

which leads to (A.5.3).

A.5.2 Multiple Groups with Homogeneous Covariances

Let us consider an $n \times p$ block data matrix $\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_K \end{bmatrix}$, whose kth block is an $n_k \times p$ matrix $\mathbf{X}_k$ ($k = 1, \ldots, K$) with its ith row being $\mathbf{x}_{ki}'$. We suppose $\mathbf{x}_{ki} \sim N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$, i.e., that the probability density of $\mathbf{x}_{ki}$ observed is given by

$$P(\mathbf{x}_{ki}|\boldsymbol{\mu}_k, \Sigma) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}_{ki} - \boldsymbol{\mu}_k)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_{ki} - \boldsymbol{\mu}_k)\right\}. \tag{A.5.11}$$

Here, it should be noted that $\boldsymbol{\mu}_k$ has subscript k, but $\boldsymbol{\Sigma}$ does not, which implies that the *mean* vectors of the distributions *differ* across K blocks, while their *covariance* matrices are *homogeneous* across them.

We further suppose the rows of $\mathbf{X}$ to be mutually independently observed. Then, the likelihood for $\mathbf{X}$ is expressed as the product of (A.5.11) over $k = 1, \ldots, K$, and $i = 1, \ldots, n_k$:

$$\begin{aligned} P(\mathbf{X}|\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \Sigma) &= \prod_{k=1}^{K} \prod_{i=1}^{n_k} \left\{ \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{ -\frac{1}{2}(\mathbf{x}_{ki} - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_{ki} - \boldsymbol{\mu}_k) \right\} \right\} \\ &= \frac{1}{(2\pi)^{np/2}|\boldsymbol{\Sigma}|^{n/2}} \exp\left\{ -\frac{1}{2} \sum_{k=1}^{K} \sum_{i=1}^{n_k} (\mathbf{x}_{ki} - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_{ki} - \boldsymbol{\mu}_k) \right\}. \end{aligned} \tag{A.5.12}$$

This leads to the log likelihood

$$l(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}) = -\frac{n}{2} \log|\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{k=1}^{K} \sum_{i=1}^{n_k} (\mathbf{x}_{ki} - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_{ki} - \boldsymbol{\mu}_k), \tag{A.5.13}$$

where the terms irrelevant to $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$, and $\boldsymbol{\Sigma}$ have been omitted. Log likelihood (A.5.13) is maximized for

$$\hat{\boldsymbol{\mu}}_k = \overline{\mathbf{x}}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \mathbf{x}_{ki}, \tag{A.5.14}$$

$$\hat{\boldsymbol{\Sigma}} = \mathbf{W} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} (\mathbf{x}_{ki} - \overline{\mathbf{x}}_k)(\mathbf{x}_{ki} - \overline{\mathbf{x}}_k)', \tag{A.5.15}$$

which is proved in the following paragraphs.

Let us rewrite (A.5.13) as $-(2/n) \log|\boldsymbol{\Sigma}| - 2 \sum_{k=1}^{K} m(\boldsymbol{\mu}_k)$ with

$$\begin{aligned} m(\boldsymbol{\mu}_k) &= \sum_{i=1}^{n_k} (\mathbf{x}_{ki} - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_{ki} - \boldsymbol{\mu}_k) \\ &= \sum_{i=1}^{n_k} (\mathbf{x}_{ki} - \overline{\mathbf{x}}_k)' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_{ki} - \overline{\mathbf{x}}_k) + n(\overline{\mathbf{x}}_k - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}^{-1} (\overline{\mathbf{x}}_k - \boldsymbol{\mu}_k). \end{aligned} \tag{A.5.16}$$

Only this function is relevant to $\boldsymbol{\mu}_k$ in (A.5.13), and the right side of (A.5.16) can be derived as (A.5.4) and (A.5.5) are derived. This fact shows that (A.5.14) is MLE for $\boldsymbol{\mu}_k$.

Substituting (A.5.14) into $\boldsymbol{\mu}_k$ in (A.5.13), we can rewrite it as

$$\begin{aligned}\log l(\boldsymbol{\Sigma}) &= -\frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{k=1}^{K}\sum_{l=1}^{m}(\mathbf{x}_h - \overline{\mathbf{x}}_k)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_{kl} - \overline{\mathbf{x}}_k) \\ &= -\frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{n}{2}\mathrm{tr}\boldsymbol{\Sigma}^{-1}\mathrm{W} = -\frac{n}{2}\left(\log|\boldsymbol{\Sigma}| + \mathrm{tr}\boldsymbol{\Sigma}^{-1}\mathrm{W}\right),\end{aligned} \tag{A.5.17}$$

which is *equivalent* to (A.5.6) if $\mathbf{W}$ is replaced by $\mathbf{V}$. Thus, MLE of $\boldsymbol{\Sigma}$ is given by (A.5.15), i.e., (A.5.3) with $\mathbf{V}$ replaced by $\mathbf{W}$.

A.6 Iterative Algorithms

In this appendix, *iterative algorithms* used in statistical data analysis are first outlined, followed by an illustration of a gradient algorithm. Though the descriptions in this book are very elementary, more advanced and exhaustive descriptions of a variety of iterative algorithms used in statistical computing are found in Lange (2010). Further, matrix-intensive descriptions of the algorithms are found in Hansen, Pereyra, and Scherer (2013) and Absil, Mahony, and Sepulchre (2008).

A.6.1 General Methodology

Let us use $\phi(\boldsymbol{\theta})$ for a function of parameter vector $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_q]'$ to be *minimized* and $\hat{\boldsymbol{\theta}}$ for the solution, i.e., the vector $\boldsymbol{\theta}$ minimizing $\phi(\boldsymbol{\theta})$. For log likelihood $l(\boldsymbol{\theta})$, we can set $\phi(\boldsymbol{\theta}) = -l(\boldsymbol{\theta})$ so that the *maximum likelihood method* for maximizing $l(\boldsymbol{\theta})$ is equivalent to *minimizing* $\phi(\boldsymbol{\theta})$. The following stories hold for any optimization including least squares and maximum likelihood methods.

If the solution $\hat{\boldsymbol{\theta}}$ is not explicitly given, we must find $\hat{\boldsymbol{\theta}}$ by using an iterative algorithm in which the update of $\boldsymbol{\theta}$ is iterated. By expressing the vector $\boldsymbol{\theta}$ at the tth iteration as $\boldsymbol{\theta}_{[t]}$, any iterative algorithm can be described with the following steps:

Note A.6.1. General Expression of Iterative Algorithm

Step 1. Set $\boldsymbol{\theta}$ to an initial value vector $\boldsymbol{\theta}_{[t]}$ with $t = 0$.

Step 2. Update $\boldsymbol{\theta}_{[t]}$ to $\boldsymbol{\theta}_{[t+1]}$ so that $\phi(\boldsymbol{\theta}_{[t+1]}) \leq \phi(\boldsymbol{\theta}_{[t]})$.

Step 3. Regard $\boldsymbol{\theta}_{[t+1]}$ as $\hat{\boldsymbol{\theta}}$ if convergence is reached; otherwise, increase t by one and go back to Step 2.

Here, the convergence in Step 3 can be defined as $\phi(\boldsymbol{\theta}_{[t]}) - \phi(\boldsymbol{\theta}_{[t+1]})$, $\left\|\theta_{[t]} - \theta_{[t+1]}\right\|$, or the maximum of $\theta_k^{[t]} - \theta_k^{[t+1]}$ over k is small enough to be ignored, with $\theta_k^{[t]}$ the kth element of $\boldsymbol{\theta}_{[t]}$.

There are various types of iterative algorithms. They can be roughly classified into three groups: parameter partition, auxiliary function, and gradient algorithms. They differ with respect to the way in which the update in Step 2 in Note A.6.1 is performed. We can also combine the three groups of algorithms or two of the three to form a whole algorithm.

In the *parameter partition algorithms*, the elements of $\boldsymbol{\theta}$ are partitioned into subsets as $\boldsymbol{\theta}' = [\boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_s', \ldots, \boldsymbol{\theta}_S']$ with the sth subset vector $\boldsymbol{\theta}_s'$ at the tth iteration expressed as $\theta_s^{[t]}$. Then, Step 2 in Note A.6.1 is divided into the substeps, in each of which $\phi(\boldsymbol{\theta})$ is minimized over only $\boldsymbol{\theta}_s$, with the other parameter sets kept fixed, and the resulting $\boldsymbol{\theta}_s$ gives $\theta_s^{[t+1]}$. In the most simple case, with $S = 2$ and $\boldsymbol{\theta} = [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2]$, Step 2 in Note A.6.1 consists of the following substeps:

Step 2.1 Minimize $\phi(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$ over $\boldsymbol{\theta}_1$ with $\boldsymbol{\theta}_2$ fixed at $\theta_2^{[t]}$ and the resulting $\boldsymbol{\theta}_1$ (that minimizes $\phi(\boldsymbol{\theta}_1, \theta_2^{[t]})$) gives $\theta_1^{[t+1]}$.

Step 2.2. Minimize $\phi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ over $\boldsymbol{\theta}_2$ with $\boldsymbol{\theta}_1$ fixed at $\theta_1^{[t+1]}$ and the resulting $\boldsymbol{\theta}_2$ (that minimizes $\phi(\theta_1^{[t+1]}, \boldsymbol{\theta}_2)$) gives $\theta_2^{[t+1]}$.

This approach is useful for cases in which it is easy to minimize $\phi(\boldsymbol{\theta})$ over a subset of parameters with the other parameters being fixed. The parameter partition algorithms are also referred to as *coordinate descending algorithms* as described in Chap. 21. In particular, such algorithms for least squares problems are known as *alternating least squares* (*ALS*) *algorithms* (e.g., Young, 1981). Their examples have been shown in Chaps. 7, 18, 20, and 21.

In *auxiliary function algorithms*, a different function $\eta(\boldsymbol{\theta})$ is used, which satisfies $\phi(\boldsymbol{\theta}_{[t]}) = \eta(\boldsymbol{\theta}_{[t]}) \geq \eta(\boldsymbol{\theta}_{[t+1]}) \geq \phi(\boldsymbol{\theta}_{[t+1]})$ with $\eta(\boldsymbol{\theta})$ being easier to handle than $\phi(\boldsymbol{\theta})$. Here, the update of $\boldsymbol{\theta}_{[t]}$ leading to $\eta(\boldsymbol{\theta}_{[t]}) \geq \eta(\boldsymbol{\theta}_{[t+1]})$ implies $\phi(\boldsymbol{\theta}_{[t]}) \geq \phi(\boldsymbol{\theta}_{[t+1]})$. One of the auxiliary function algorithms is the *EM algorithm* originally presented by Dempster, Laird, and Rubin (1977). Its principle is introduced in A.8.5, and the EM algorithm specially designed for factor analysis is detailed in A.9. A book-length description of the EM algorithm is found in McLachlan and Krishnan (2008). The auxiliary function algorithm also includes the *majorization algorithm* introduced in Chap. 16. Majorization algorithms useful for some multivariate analysis procedures can be found in Kiers (2002).

In *gradient algorithms*, the differential of $\phi(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ is used. This type of algorithm is illustrated in the remaining sections.

A.6.2 Gradient Algorithm for Single Parameter Cases

For introducing the gradient algorithms, we consider an example of $\phi(\boldsymbol{\theta})$ to be minimized:

$$\phi(\theta) = 16\theta^4 - 192\theta^3 + 880\theta^2 - 1824\theta + 1444, \tag{A.6.1}$$

which is a function of a *single* parameter $\boldsymbol{\theta} = [\theta]$. Figure A.6 shows $\phi(\theta)$ values against θ, where (A.6.1) is found to attain its minimum at $\theta = 3$, i.e., the solution $\hat{\theta} = 3$. However, we suppose that only formula (A.6.1) is given and $\hat{\theta}$ is unknown. Then, a *gradient algorithm* can be used, in which the *derivative* of $\phi(\theta)$ with respect to θ,

$$\begin{aligned}\frac{\mathrm{d}\phi(\theta)}{\mathrm{d}\theta} &= 16 \times 4\theta^3 - 192 \times 3\theta^2 + 880 \times 2\theta - 1824 \\ &= 64\theta^3 - 576\theta^2 + 1760\theta - 1824,\end{aligned} \tag{A.6.2}$$

is noted. The value of (A.6.2) with θ set to a specific value, $\theta_{[t]}$, that is,

$$\nabla\phi\left(\theta_{[t]}\right) = 64\theta_{[t]}^3 - 576\theta_{[t]}^2 + 1760\theta_{[t]} - 1824 \tag{A.6.3}$$

is called the *gradient* of $\phi(\theta)$ at $\theta = \theta_{[t]}$.

For example,

$$\text{if } \theta_{[t]} = 2, \text{ then } \nabla\phi(2)blank; = 64 \times 2^3 - 576 \times 2^2 + 1760 \times 2 - 1824 = -96, \tag{A.6.4}$$

$$\text{If } \theta_{[t]} = 3.8, \text{ then } \nabla\phi(3.8) = 64 \times 3.8^3 - 576 \times 3.8^2 + 1760 \times 3.8 - 1824 = 58.4. \tag{A.6.5}$$

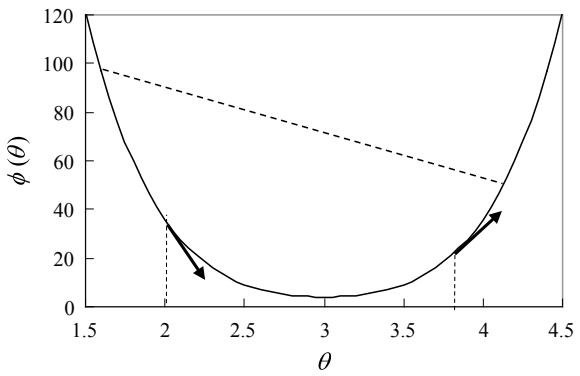


Fig. A.6 Function $\phi(\theta)$ against θ with arrows expressing the gradients of $\phi(\theta)$ and a dotted line for illustrating the convexity of $\phi(\theta)$

These values show the directions of the tangential lines represented as the arrows in Fig. A.6. Let us note that $\theta_{[t]} = 2$, giving the negative value (=−96) as (A.6.4), is less than the solution $\hat{\theta}$ (=3), while $\theta_{[t]} = 3.8$, leading to the positive value (A.6.5), is greater than $\hat{\theta}$. These relationships of $\theta_{[t]}$ to the solution $\hat{\theta}$ generally hold for (A.6.3); $\nabla\phi(\theta_t)$ is *negative* when $\theta_{[t]} < \hat{\theta}$; thus, $\theta_{[t]}$ should be updated to a *larger* value so as to approach the solution $\hat{\theta}$, while $\nabla\phi(\theta_{[t]})$ is *positive* for $\theta_{[t]} > \hat{\theta}$; thus, $\theta_{[t]}$ is to be updated to a *smaller* value to approach $\hat{\theta}$. This implies that $\theta_{[t]}$ is to be updated in the direction of $-1 \times \nabla(\theta_{[t]})$, i.e., in the *opposite direction of the sign* of $\nabla\phi(\theta_{[t]})$. This update is formally expressed as

$$\theta_{[t+1]} = \theta_{[t]} - s\nabla\phi\left(\theta_{[t]}\right), \tag{A.6.6}$$

with s a suitable positive value. The resulting $\theta_{[t+1]}$ can be closer to $\hat{\theta}$ than $\theta_{[t]}$, with $\phi(\theta_{[t+1]}) \leq \phi(\theta_{[t]})$, if s is suitably chosen.

We find that whether update (A.6.6) is *successful or not* depends on which s is chosen. One unsuccessful example is if $s = 1$ is chosen for $\theta_{[t]} = 2$. Then, (A.6.4) and (A.6.6) show $\theta_{[t+1]} = \theta_{[t]} - s\nabla\phi(\theta_{[t]}) = 2 - (-96) = 98$; the updated $\theta_{[t+1]}$ far exceeds $\hat{\theta}$ and $\phi(\theta_{[t+1]}) > \phi(\theta_{[t]})$. However, such cases can be avoided by choosing s with the following steps:

Step 2.1. Set s to 1.
Step 2.2. Obtain $\theta_{[t+1]}$ with (A.6.6).
Step 2.3. Finish if $\phi(\theta_{[t+1]}) \leq \phi(\theta_{[t]})$; otherwise, set $s := s/2$ and go back to Step 2.2.

Here, "$s := s/2$" stands for reduce the s value to half; s is reduced as 1, 1/2, $1/2^2$, $1/2^3$, … . When $\theta_{[t]} = 2$, returning to Step 2.2 seven times leads to $\theta_{[t+1]} = \theta_{[t]} - s\nabla\phi(\theta_{[t]}) = 2 - (1/2^7) \times 98 = 2 - 1/128 \times (-96) = 2.75$, which is close to $\hat{\theta}$.

The three steps in Sect. A.6.1, with Step 2 in Note A.6.1 replaced by the above Steps 2.1, 2.2, and 2.3, allow us to find $\hat{\theta}$ if $\phi(\theta)$ is *convex*. Here, the adjective "convex", roughly speaking, stands for the fact that the curve of $\phi(\theta)$ is not a zigzag. The exact definition, with multiple parameters considered, is as follows: $\phi(\boldsymbol{\theta})$ is said to be *convex*,

$$\phi(\alpha\theta_1 + (1-\alpha)\theta_2) \leq \alpha\phi(\theta_1) + (1-\alpha)\phi(\theta_2), \tag{A.6.7}$$

for every pair of $q \times 1$ vectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, and every α taking a value within the range from 0 to 1. This implies that, as a dotted line in Fig. A.6, the line connecting the two points of $\phi(\boldsymbol{\theta})$ is not lower than $\phi(\boldsymbol{\theta})$.

Although more efficient procedures than the one in the above Step 2.3 have been developed for choosing s (e.g., Boyed & Vandenberghe, 2004), they are beyond the scope in this book.

A.6.3 Gradient Algorithm for Multiple Parameter Cases

For *multiple* parameter cases with $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_q]'$, update formula (A.6.6) is extended as

$$\boldsymbol{\theta}_{[t+1]} = \boldsymbol{\theta}_{[t]} - s\nabla\phi\left(\boldsymbol{\theta}_{[t]}\right), \tag{A.6.8}$$

with $\nabla\phi(\boldsymbol{\theta}_{[t]})$ the $q \times 1$ gradient vector, which is the vector $\partial\phi(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ at $\boldsymbol{\theta} = \boldsymbol{\theta}_{[t]}$. Here, $\partial\phi(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ denotes the *partial derivative* of $\phi(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. That is, $\partial\phi(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ is the $q \times 1$ vector, and its kth element is the derivative of $\phi(\boldsymbol{\theta})$ with respect to θ_k (the kth element of $\boldsymbol{\theta}$), where $\phi(\boldsymbol{\theta})$ is regarded as a function of *only* θ_k with $\phi(\boldsymbol{\theta}) = \phi(\theta_k)$ and θ_l ($l \neq k$) treated as a *fixed constant*. For example, when $q = 3$ and

$$\phi(\boldsymbol{\theta}) = 3\theta_1^2 + 6\theta_2^2 - 4\theta_1\theta_3 + 5\theta_2\theta_3 - 7\theta_2 + 9\theta_3, \tag{A.6.9}$$

its partial derivative is

$$\frac{\partial\phi(\theta)}{\partial\theta} = \begin{bmatrix} \mathrm{d}\phi(\theta_1)/\mathrm{d}\theta_1 \\ \mathrm{d}\phi(\theta_2)/\mathrm{d}\theta_2 \\ d\phi(\theta_3)/\mathrm{d}\theta_3 \end{bmatrix} = \begin{bmatrix} 6\theta_1 - 4\theta_3 \\ 12\theta_2 + 5\theta_3 - 7 \\ -4\theta_1 + 5\theta_2 + 9 \end{bmatrix}. \tag{A.6.10}$$

Note its second element. There, (A.6.9) has been regarded as a function of only θ_2, i.e., $\phi(\theta_2) = (6)\theta_2^2 + (5\theta_3 - 7)\theta_2 + (3\theta_1^2 - 4\theta_1\theta_3 + 9\theta_3)$, with the parenthesized terms being treated as fixed constants.

In multiple parameter cases, the three steps in the last section are simply replaced by their vector versions:

Step 2.1. Set s to 1.
Step 2.2. Obtain $\boldsymbol{\theta}_{[t+1]}$ with (A.6.8).
Step 2.3. Finish if $\phi(\boldsymbol{\theta}_{[t+1]}) \leq \phi(\boldsymbol{\theta}_{[t]})$; otherwise, set $s := s/2$ and go back to Step 2.2.

The three steps in Note A.6.1, with Step 2 in Note A.6.1 replaced by the above steps, allow us to find $\hat{\boldsymbol{\theta}}$, if $\phi(\boldsymbol{\theta})$ is convex with (A.6.7). The algorithm with (A.6.8) is illustrated in Fig. A.7.

Though we have only introduced a procedure using the (first) derivative, more effective procedures, including one in which first and *second* derivatives are used, have been developed, which are beyond the scope of this book. Advanced theories for gradient and related algorithms are detailed in Absil, Mahony, and Sepulchre (2008), and Boyed and Vandenberghe (2004). One republication of a classical book dealing with such theories is Ortega and Rheinboldt (2000).

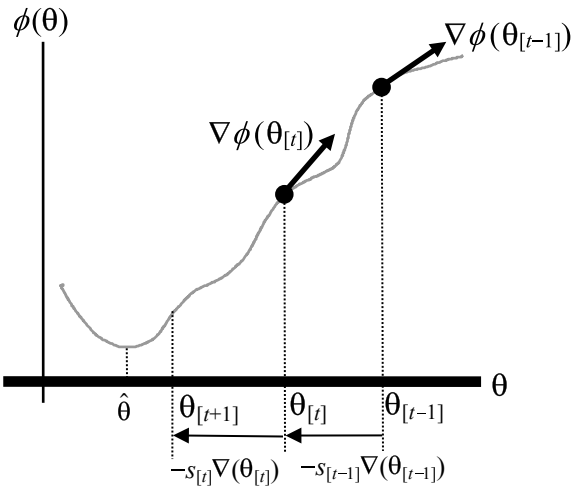


Fig. A.7 Illustration of $\boldsymbol{\theta}_{[t]}$ approaching the solution $\hat{\boldsymbol{\theta}}$ with an increase in t, where the horizontal axis represents the q-dimensional space for $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_q]'$ and subscript t is attached to s, as the s value chosen for t differs from the one for $t - 1$

A.7 Scale Invariance of Covariance Structure Analysis

The procedures treated in Chaps. 9–12 can be generally referred to as *covariance structure analysis* (*CSA*), as explained in Sect. 9.4 with Note 9.2. These procedures have the property that the value of the objective function for the unstandardized solution is equivalent to the value for the corresponding standardized solution. This property is called *scale invariance*: CSA is said to be *scale invariant*. In this Appendix, the invariance is defined exactly, and we show that *path analysis* treated in Chap. 9 and *factor analysis* in Chaps. 10 and 12 are scale invariant. Furthermore, it is also shown that their unstandardized solutions can be *straightforwardly transformed* into the corresponding standardized ones. The scale invariance of *structural equation modeling* (Chap. 11) is too involved to be treated in this book.

A.7.1 Definition of Scale Invariance

Let $\mathbf{X}$ be an n-individuals $\times$ p-variables data matrix centered as $\mathbf{X} = \mathbf{JX}$ with $\mathbf{J} = \mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n'$. As described in Sect. 9.4, CSA for $\mathbf{X}$ is formulated as maximizing (9.15), i.e.,

$$l^*(\boldsymbol{\Sigma}|\mathbf{V}) = \frac{n}{2}\log\left|\boldsymbol{\Sigma}^{-1}\mathbf{V}\right| - \frac{n}{2}\mathrm{tr}\boldsymbol{\Sigma}^{-1}\mathbf{V} \tag{A.7.1}$$

over the parameters in the *covariance structure* (i.e., model-based covariance matrix) $\boldsymbol{\Sigma}$ ($p \times p$) subject to constraints for $\boldsymbol{\Sigma}$. Here, "$|\mathbf{V}$" has been added to $l^*(\boldsymbol{\Sigma})$, in

order to indicate that the log likelihood (A.7.1) is defined for a given inter-variable sample covariance matrix $\mathbf{V} = n^{-1}\mathbf{X}'\mathbf{X}$. A set of the resulting parameter estimates is referred to as the *unstandardized solution*.

Now, let us consider performing CSA for the transformed data matrix

$$\mathbf{Z} = \mathbf{X}\mathbf{D}^{-1} \tag{A.7.2}$$

with $\mathbf{D}$ a $p \times p$ diagonal matrix whose diagonal elements $d_1, \ldots, d_p$ are all positive. The inter-variable covariance matrix for (A.7.2) is expressed as

$$\mathbf{R} = \frac{1}{n}\mathbf{Z}'\mathbf{Z} = \frac{1}{n}\mathbf{D}^{-1}\mathbf{X}'\mathbf{X}\mathbf{D}^{-1} = \mathbf{D}^{-1}\mathbf{V}\mathbf{D}^{-1}. \tag{A.7.3}$$

Thus, in CSA for $\mathbf{Z}$, the log likelihood is defined by substituting $\mathbf{R}$ into $\mathbf{V}$ in (A.7.1):

$$l^*(\tilde{\mathbf{\Sigma}}|\mathbf{R}) = \frac{n}{2}\log|\tilde{\mathbf{\Sigma}}^{-1}\mathbf{R}| - \frac{n}{2}\mathrm{tr}\tilde{\mathbf{\Sigma}}^{-1}\mathbf{R}, \tag{A.7.4}$$

This maximization over the parameters in the *covariance structure* $\tilde{\mathbf{\Sigma}}$ $(p \times p)$ under the constraints for $\tilde{\mathbf{\Sigma}}$ provides the CSA solution for (A.7.2) or (A.7.3). Here, we have attached the tilde symbol ($\sim$) to $\mathbf{\Sigma}$, as the matrix $\tilde{\mathbf{\Sigma}}$ maximizing (A.7.4) differs from the matrix $\mathbf{\Sigma}$ maximizing (A.7.1).

The following theorem can be used to show that (A.7.1) equals (A.7.4) under certain conditions:

Theorem A.7.1.
Let $\aleph(\mathbf{M})$ denote a set of allowable values for the elements of a matrix $\mathbf{M}$. If

$$\aleph\left(\mathbf{D}^{-1}\mathbf{\Sigma}\mathbf{D}^{-1}\right) = \aleph(\tilde{\mathbf{\Sigma}}) \tag{A.7.5}$$

holds true, then (A.7.1) is equivalent to (A.7.4):

$$l^*(\mathbf{\Sigma}|\mathbf{V}) = l^*(\tilde{\mathbf{\Sigma}}|\mathbf{R}), \tag{A.7.6}$$

which can be rewritten as

$$l^*(\mathbf{\Sigma}|\mathbf{V}) = l^*\left(\mathbf{D}^{-1}\mathbf{\Sigma}\mathbf{D}^{-1}|\mathbf{D}^{-1}\mathbf{V}\mathbf{D}^{-1}\right). \tag{A.7.7}$$

Scale invariance is defined by (A.7.7): An analysis procedure whose objective function satisfies (A.7.7) is said to be *scale invariant*.

Proof Using (4.16), (8.11), and $\mathbf{V} = \mathbf{DRD}$ that follows from (A.7.3), (A.7.1) can be rewritten as

$$\begin{aligned} l^*(\boldsymbol{\Sigma}|\mathbf{V}) &= \frac{n}{2}\log\left|\boldsymbol{\Sigma}^{-1}\mathbf{DRD}\right| - \frac{n}{2}\mathrm{tr}\boldsymbol{\Sigma}^{-1}\mathbf{DRD} \\ &= \frac{n}{2}\log\left|\left(\mathbf{D}\boldsymbol{\Sigma}^{-1}\mathbf{D}\right)\mathbf{R}\right| - \frac{n}{2}\mathrm{tr}\left(\mathbf{D}\boldsymbol{\Sigma}^{-1}\mathbf{D}\right)\mathbf{R} \\ &= \frac{n}{2}\log\left|\left(\mathbf{D}^{-1}\boldsymbol{\Sigma}\mathbf{D}^{-1}\right)^{-1}\mathbf{R}\right| - \frac{n}{2}\mathrm{tr}\left(\mathbf{D}^{-1}\boldsymbol{\Sigma}\mathbf{D}^{-1}\right)^{-1}\mathbf{R}. \end{aligned} \tag{A.7.8}$$

Comparing this equation with (A.7.4), we find that if (A.7.5) holds true, (A.7.1) or (A.7.8) is equivalent to (A.7.4) with

$$\tilde{\boldsymbol{\Sigma}} = \mathbf{D}^{-1}\boldsymbol{\Sigma}\mathbf{D}^{-1}. \tag{A.7.9}$$

By substituting (A.7.3) and (A.7.9) in (A.7.6), this can be rewritten as (A.7.7). □

The CSA solution for (A.7.2) is called the *standardized solution*, when (A.7.2) contains the standard scores of $\mathbf{X}$; that is, the jth diagonal elements of $\mathbf{D}$ are the standard deviation of the jth variable in centered $\mathbf{X}$. Thus, scale invariance (A.7.7) implies that the likelihood value for the unstandardized solution equals that for the corresponding standardized one.

If (A.7.5) holds, we have (A.7.9). This and (A.7.3) lead to

$$\mathrm{tr}\tilde{\boldsymbol{\Sigma}}^{-1}\mathbf{R} = \mathrm{tr}\left(\mathbf{D}^{-1}\boldsymbol{\Sigma}\mathbf{D}^{-1}\right)^{-1}\mathbf{D}^{-1}\mathbf{V}\mathbf{D}^{-1} = \mathrm{tr}\mathbf{D}\boldsymbol{\Sigma}^{-1}\mathbf{D}\mathbf{D}^{-1}\mathbf{V}\mathbf{D}^{-1} = \mathrm{tr}\boldsymbol{\Sigma}^{-1}\mathbf{V}$$

and

$$1 - \frac{\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{V} - \mathbf{I}_p)^2}{\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{V})^2} = 1 - \frac{\mathrm{tr}(\tilde{\boldsymbol{\Sigma}}^{-1}\mathbf{R} - \mathbf{I}_p)^2}{\mathrm{tr}(\tilde{\boldsymbol{\Sigma}}^{-1}\mathbf{R})^2}. \tag{A.7.10}$$

This equality implies that the *GFI index* (9.18) shows the same value between unstandardized and standardized solutions, if (A.7.5) holds.

The above theorem shows that the procedures in CSA are scale invariant under the condition (A.7.5). In the next two sections, it is shown that (A.7.5) holds for the procedures treated in Chaps. 9, 10, and 12.

A.7.2 Scale Invariance of Factor Analysis

To show that (A.7.5) holds true in *confirmatory factor analysis* (*CFA*) (Chap. 10), we start by considering the constraints imposed on the covariance structure $\boldsymbol{\Sigma} = (\sigma_{jk})$. Among them, the main constraint is (10.10), i.e.,

$$\boldsymbol{\Sigma} = \mathbf{A}\boldsymbol{\Phi}\mathbf{A}' + \boldsymbol{\Psi}. \tag{A.7.11}$$

Further, $\mathbf{A} = (a_{jk})$, $\boldsymbol{\Phi}$, and $\boldsymbol{\Psi}$ on the right side are constrained as follows:

$$a_{jk} = 0 \text{ if } (j,k) \in \Re_0; \text{otherwise any real value}, \tag{A.7.12}$$

$$\boldsymbol{\Psi} \text{ is a diagonal matrix}, \tag{A.7.13}$$

and $\boldsymbol{\Phi}$ is a correlation matrix. Here, $\Re_0$ denotes a set of pairs (j, k) with the corresponding a_{jk} set to zero. Those constraints specify $\aleph(\mathbf{D}^{-1}\boldsymbol{\Sigma}\mathbf{D}^{-1})$ and $\aleph(\tilde{\boldsymbol{\Sigma}})$ in (A.7.5).

The constraint (A.7.11) leads to

$$\mathbf{D}^{-1}\boldsymbol{\Sigma}\mathbf{D}^{-1} = \mathbf{D}^{-1}\mathbf{A}\boldsymbol{\Phi}\mathbf{A}'\mathbf{D}^{-1} + \mathbf{D}^{-1}\boldsymbol{\Psi}\mathbf{D}^{-1} = \tilde{\mathbf{A}}\tilde{\boldsymbol{\Phi}}\tilde{\mathbf{A}}' + \tilde{\boldsymbol{\Psi}}. \tag{A.7.14}$$

with

$$\tilde{\mathbf{A}} = \mathbf{D}^{-1}\mathbf{A}, \quad \tilde{\boldsymbol{\Phi}} = \boldsymbol{\Phi}, \quad \text{and} \quad \tilde{\boldsymbol{\Psi}} = \mathbf{D}^{-1}\boldsymbol{\Psi}\mathbf{D}^{-1}. \tag{A.7.15}$$

Here, (A.7.11) and (A.7.14) take an identical form. It implies that (A.7.5) holds true with $\tilde{\boldsymbol{\Sigma}} = \tilde{\mathbf{A}}\tilde{\boldsymbol{\Phi}}\tilde{\mathbf{A}}' + \tilde{\boldsymbol{\Psi}}$, if $\aleph(\mathbf{A}) = \aleph(\tilde{\mathbf{A}})$, $\aleph(\boldsymbol{\Phi}) = \aleph(\tilde{\boldsymbol{\Phi}})$, and $\aleph(\boldsymbol{\Psi}) = \aleph(\tilde{\boldsymbol{\Psi}})$. These three identities are shown in the next paragraph.

We can derive $\aleph(\mathbf{A}) = \aleph(\tilde{\mathbf{A}})$ and $\aleph(\boldsymbol{\Psi}) = \aleph(\tilde{\boldsymbol{\Psi}})$ from the fact that $\mathbf{D}$ in (A.7.15) is diagonal with its diagonal elements d_j ($j = 1, \ldots, p$) all positive. These properties of $\mathbf{D}$ imply that the (j, k) elements of $\tilde{\mathbf{A}} = (\tilde{a}_{jk})$ are expressed as $\tilde{a}_{jk} = a_{jk}/d_j$ and $\tilde{\boldsymbol{\Psi}}$ is the diagonal matrix whose jth diagonal elements are ψ_j/d_j^2 for $j = 1, \ldots, p$. Thus, $\tilde{a}_{jk}$ and $\tilde{\boldsymbol{\Psi}}$ can be substituted into a_{jk} in (A.7.12) and $\boldsymbol{\Psi}$ in (A.7.13), respectively. This implies $\aleph(\mathbf{A}) = \aleph(\tilde{\mathbf{A}})$ and $\aleph(\boldsymbol{\Psi}) = \aleph(\tilde{\boldsymbol{\Psi}})$. Obviously, $\aleph(\boldsymbol{\Phi}) = \aleph(\tilde{\boldsymbol{\Phi}})$ follows from (A.7.15). These results lead to (A.7.5) in CFA, and Theorem A.7.1 implies the scale invariance of CFA. Furthermore, (A.7.15) show that the standardized solution is transformed from the unstandardized one by (10.12).

Scale invariance of the *exploratory FA* (*EFA*) treated in Chap. 12 follows straightforwardly from the invariance of CFA. As found in (12.8), $\boldsymbol{\Sigma}$ and $\mathbf{D}^{-1}\boldsymbol{\Sigma}\mathbf{D}^{-1}$ in EFA are given by (A.7.11) and (A.7.14), respectively, with $\boldsymbol{\Phi} = \tilde{\boldsymbol{\Phi}}$ fixed to $\mathbf{I}_m$ and without constraint (A.7.12). Thus, the equality (A.7.5) with $\tilde{\boldsymbol{\Sigma}} = \tilde{\mathbf{A}}\tilde{\mathbf{A}}' + \tilde{\boldsymbol{\Psi}}$ can be found from $\aleph(\mathbf{A}) = \aleph(\tilde{\mathbf{A}})$ and $\aleph(\boldsymbol{\Psi}) = \aleph(\tilde{\boldsymbol{\Psi}})$ derived in the last paragraph.

A.7.3 Scale Invariance of Path Analysis

In *path analysis* (Chap. 9), $\mathbf{\Sigma} = (\sigma_{jk})$ is constrained through (9.12), i.e.,

$$\mathbf{\Sigma} = \left(\mathbf{I}_p - \mathbf{B}\right)^{-1}\Omega\left(\mathbf{I}_p - \mathbf{B}\right)^{-1\prime} \tag{A.7.16}$$

Further, $\mathbf{B} = (b_{jk})$ $(p \times p)$ and $\mathbf{\Omega} = (\omega_{kl})$ $(p \times p)$ on the right side have the following constraints:

$$b_{jk} = 0 \text{ if } (j,k) \in \Re_0; \text{ otherwise any real value,} \tag{A.7.17}$$

$$\Omega = (\omega_{kl}) \text{ is a covariance matrix among variables for which } \omega_{kl} = 0 \text{ if } (k,l) \in \Im_0; \text{ otherwise any real value.} \tag{A.7.18}$$

Here, $\Re_0$ and $\Im_0$ denote the sets of (j, k) and (k, l) with their corresponding elements equal to zeros, respectively.

The constraint (A.7.16) leads to

$$\begin{aligned}
\mathbf{D}^{-1}\Sigma\mathbf{D}^{-1} &= \mathbf{D}^{-1}\left(\mathbf{I}_p - \mathbf{B}\right)^{-1}\Omega\left(\mathbf{I}_p - \mathbf{B}\right)^{-1\prime}\mathbf{D}^{-1} \\
&= \left\{\left(\mathbf{I}_p - \mathbf{B}\right)\mathbf{D}\right\}^{-1}\Omega\left\{\mathbf{D}\left(\mathbf{I}_p - \mathbf{B}\right)'\right\}^{-1} \\
&= (\mathbf{D} - \mathbf{BD})^{-1}\Omega(\mathbf{D} - \mathbf{DB}')^{-1} \\
&= \left\{\mathbf{D}\left(\mathbf{I}_p - \mathbf{D}^{-1}\mathbf{BD}\right)\right\}^{-1}\Omega\left\{\left(\mathbf{I}_p - \mathbf{D}^{-1}\mathbf{BD}\right)'\mathbf{D}\right\}^{-1} \\
&= \left(\mathbf{I}_p - \mathbf{D}^{-1}\mathbf{BD}\right)^{-1}\mathbf{D}^{-1}\Omega\mathbf{D}^{-1}\left(\mathbf{I}_p - \mathbf{D}^{-1}\mathbf{BD}\right)'^{-1} \\
&= (\mathbf{I}_p - \tilde{\mathbf{B}})^{-1}\tilde{\mathbf{\Omega}}(\mathbf{I}_p - \tilde{\mathbf{B}})^{-1\prime}.
\end{aligned} \tag{A.7.19}$$

with

$$\tilde{\mathbf{B}} = \mathbf{D}^{-1}\mathbf{BD} \text{ and } \tilde{\mathbf{\Omega}} = \mathbf{D}^{-1}\Omega\mathbf{D}^{-1}. \tag{A.7.20}$$

Here, (A.7.16) and (A.7.19) take an identical form. It implies that (A.7.5) holds true with $\tilde{\mathbf{\Sigma}} = (\mathbf{I}_p - \tilde{\mathbf{B}})^{-1}\tilde{\mathbf{\Omega}}(\mathbf{I}_p - \tilde{\mathbf{B}})^{-1\prime}$, if $\aleph(\mathbf{B}) = \aleph(\tilde{\mathbf{B}})$ and $\aleph(\Omega) = \aleph\left(\tilde{\mathbf{\Omega}}\right)$. These two identities are derived in the next paragraph.

We can show $\aleph(\mathbf{B}) = \aleph(\tilde{\mathbf{B}})$ from the fact that the (j, k) element of $\tilde{\mathbf{B}}$ being $\tilde{b}_{jk} = b_{jk}\left(d_k/d_j\right)$ from (A.7.20), with $d_k/d_j > 0$: $\tilde{b}_{jk}$ can be substituted into b_{jk} in (A.7.17). We can also derive $\aleph(\Omega) = \aleph\left(\tilde{\mathbf{\Omega}}\right)$ from the following two properties: [1] a matrix being a covariance matrix among variables can be rewritten in the form $n^{-1}\mathbf{Y}'\mathbf{JY}$, with $\mathbf{J} = \mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_{n'}$. Thus, we have $\mathbf{\Omega} = n^{-1}\mathbf{Y}'\mathbf{JY}$, and from (A.7.20) $\tilde{\mathbf{\Omega}}$ can also be written in the identical form $n^{-1}\tilde{\mathbf{Y}}'\mathbf{J}\tilde{\mathbf{Y}}$ with $\tilde{\mathbf{Y}} = \mathbf{YD}^{-1}$: Hence $\tilde{\mathbf{\Omega}}$ is

also a covariance matrix. [2] (A.7.20) also shows that the (k, l) element of $\tilde{\boldsymbol{\Omega}}$ is $\tilde{\omega}_{kl} = \omega_{kl}/(d_k d_l)$, with $d_k d_l > 0$. Those properties allow us to find that $\tilde{\boldsymbol{\Omega}}$ and $\tilde{\omega}_{kl}$ can be substituted into $\boldsymbol{\Omega}$ and ω_{kl} in (A.7.18), respectively.

The results in the last paragraph demonstrate the scale invariance of path analysis. Furthermore, (A.7.20) shows that the standardized solution is transformed from the unstandardized one by (9.20).

A.8 Probability Densities and Expected Values with EM Algorithm

In this appendix, we describe some details of probability densities and introduce expected values to explain the foundations of the EM algorithm.

A.8.1 Joint, Conditional, and Marginal Probability Densities

Let $\mathbf{x}$ $(p \times 1)$ and $\mathbf{f}$ $(m \times 1)$ be the random vectors whose probability density functions (PDF) are denoted as $P(\mathbf{x}|\boldsymbol{\Theta})$ and $P(\mathbf{f}|\boldsymbol{\Theta})$, with $\boldsymbol{\Theta}$ a set of the parameters specifying those PDF. As illustrated by Fig. 8.3a in Chap. 8, the area below the PDF expresses a probability. This implies that the integral of any PDF over all possible values is one: $\int P(\mathbf{x}|\Theta)\mathrm{d}\mathbf{x} = 1$. Integral calculus for probabilities is detailed in Khuri (2003).

The PDF $P(\mathbf{x},\mathbf{f}|\boldsymbol{\Theta})$ stands for the PDF of $\mathbf{x}$ and $\mathbf{f}$ observed jointly, with $\iint P(\mathbf{x},\mathbf{f}|\Theta)\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{f} = 1$. In particular, $P(\mathbf{x}, \mathbf{f}|\boldsymbol{\Theta})$ is called the *joint PDF* of $\mathbf{x}$ and $\mathbf{f}$, for distinguishing it from $P(\mathbf{x}|\boldsymbol{\Theta})$ or $P(\mathbf{f}|\boldsymbol{\Theta})$ which is a function of a single vector. In Fig. A.8, $P(\mathbf{x},\mathbf{f}|\boldsymbol{\Theta})$ is illustrated by a three-dimensional *mountain*-like object, where $\mathbf{x}$ and $\mathbf{f}$ values are represented by the width and depth, respectively. The integral of the joint PDF $P(\mathbf{x},\mathbf{f}|\boldsymbol{\Theta})$ over all possible $\mathbf{f}$ leads to $P(\mathbf{x}|\boldsymbol{\Theta})$:

$$P(\mathbf{x}|\Theta) = \int P(\mathbf{x},\mathbf{f}|\Theta)\mathrm{d}\mathbf{f}. \tag{A.8.1}$$

Similarly, $P(\mathbf{f}|\boldsymbol{\Theta})$ is the integral of $P(\mathbf{x},\mathbf{f}|\boldsymbol{\Theta})$ with respect to $\mathbf{x}$: $P(\mathbf{f}|\Theta) = \int P(\mathbf{x},\mathbf{f}|\Theta)\mathrm{d}\mathbf{x}$. The PDF (A.8.1) is illustrated "behind" the mountain in Fig. A.8. Here, "behind" can be rephrased as "in a marginal territory". In this sense, (A.8.1) is called the *marginal PDF* of $P(\mathbf{x}, \mathbf{f}|\boldsymbol{\Theta})$ for $\mathbf{x}$.

PDF $P(\mathbf{f}|\mathbf{x},\boldsymbol{\Theta})$ expresses the PDF of $\mathbf{f}$ conditional on $\mathbf{x}$ being a particular vector. Thus, $P(\mathbf{f}|\mathbf{x},\boldsymbol{\Theta})$ is called the *conditional PDF* of $\mathbf{f}$, in particular, to distinguish it from the marginal and joint PDF. The conditional PDF $P(\mathbf{f}|\mathbf{x},\boldsymbol{\Theta})$ is illustrated on the right in Fig. A.8 As seen there, we may consider $P(\mathbf{f}|\mathbf{x},\boldsymbol{\Theta})$ as the cross section of the *mountain* in $P(\mathbf{x},\mathbf{f}|\boldsymbol{\Theta})$ corresponding to a particular $\mathbf{x}$ value, indicated by a white real line in Fig. A.8. Similarly, we can consider the PDF $P(\mathbf{x}|\mathbf{f},\boldsymbol{\Theta})$, i.e., the PDF of

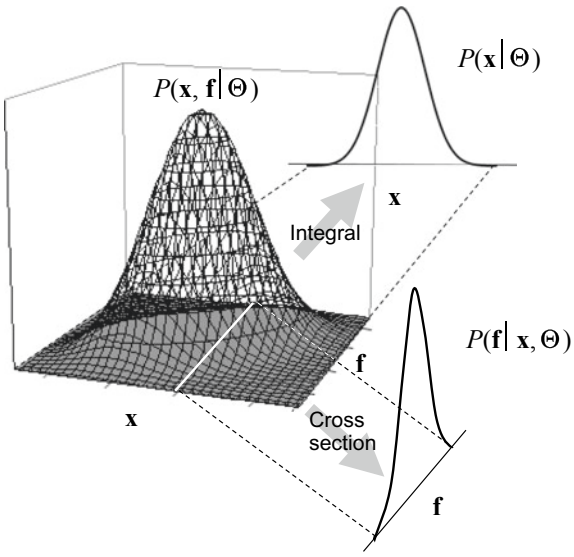


Fig. A.8 Joint, marginal, and conditional distributions

x conditional on **f**. An important fact regarding probabilities is that the *joint PDF* of two vectors is given by the *product* of the *conditional PDF* of one of the two vectors and the *marginal PDF* of the other vector:

$$P(\mathbf{x}, \mathbf{f}|\Theta) = P(\mathbf{f}|\mathbf{x}, \Theta)P(\mathbf{x}|\Theta) = P(\mathbf{x}|\mathbf{f}, \Theta)P(\mathbf{f}|\Theta). \tag{A.8.2}$$

This implies

$$P(\mathbf{f}|\mathbf{x}, \boldsymbol{\Theta}) = \frac{P(\mathbf{x}, \mathbf{f}|\boldsymbol{\Theta})}{P(\mathbf{x}|\boldsymbol{\Theta})} = \frac{P(\mathbf{x}, \mathbf{f}|\boldsymbol{\Theta})}{\int P(\mathbf{x}, \mathbf{f}|\boldsymbol{\Theta})\mathrm{d}\mathbf{f}} = \frac{P(\mathbf{x}|\mathbf{f}, \boldsymbol{\Theta})P(\mathbf{f}|\boldsymbol{\Theta})}{\int P(\mathbf{x}|\mathbf{f}, \boldsymbol{\Theta})P(\mathbf{f}|\boldsymbol{\Theta})\mathrm{d}\mathbf{f}}, \tag{A.8.3}$$

which is *Bayes' theorem* (15.15) extended to continuous variables.

A.8.2 Expected Values

An *expected value* refers to the *average* of a random variable derived theoretically using probabilities. For example, let x denote a number shown by a dice. The average of the numbers shown by the dice rolled many times is expected to be $E[x] = \sum_{x=1}^{6} xP(x)$ with $P(x)$ (x = 1, …, 6) the probability that the dice is rolled to show the number x. The above $E[x]$ defines the expected value of x, which is a discrete random variable. If x is a continuous random variable taking any real value, the definition of the *expected value* can be straightforwardly generalized by

replacing the summation by an integral: $E[x] = \int xP(x)\,\mathrm{d}x$ with $P(x)$ the PDF of x. The expected value of an $m \times 1$ random vector $\mathbf{f} = [f_1, \ldots, f_m]'$, which may also be called an *expected vector*, is defined as

$$E[\mathbf{f}] = \int \mathbf{f} P(\mathbf{f}|\boldsymbol{\Theta})\mathrm{d}\mathbf{f}, \tag{A.8.4}$$

which is an $m \times 1$ vector with $E[\mathbf{f}] = [E[f_1], \ldots, E[f_m]]'$.

Let us denote a matrix function of $\mathbf{f}$ as $\mathbf{M}(\mathbf{f})$ ($n \times m$), where the term *matrix function* refers to a function providing a matrix. Since vectors are included in matrices and a scalar is a 1×1 matrix, the properties for a matrix function hold true even if the function produces a scalar or vector. The expected value of $\mathbf{M}(\mathbf{f})$ is expressed as

$$E[\mathbf{M}(\mathbf{f})] = \int \mathbf{M}(\mathbf{f}) P(\mathbf{f}|\Theta)\mathrm{d}\mathbf{f}, \tag{A.8.5}$$

which is an $n \times m$ matrix with its (i, j) element being the expected value of the counterpart of $\mathbf{M}(\mathbf{f})$. The expected value of $\mathbf{M}(\mathbf{f})$ pre-multiplied by a fixed matrix $\mathbf{Y}$ satisfies

$$E[\mathbf{Y}\mathbf{M}(\mathbf{f})] = \mathbf{Y}E[\mathbf{M}(\mathbf{f})], \quad \text{or equivalently,} \quad E\left[\mathbf{M}(\mathbf{f})'\mathbf{Y}'\right] = E[\mathbf{M}(\mathbf{f})]'\mathbf{Y}'. \tag{A.8.6}$$

Let $\mathbf{N}(\mathbf{f})$ ($n \times m$) be a matrix function other than $\mathbf{M}(\mathbf{f})$. These functions satisfy

$$E[\mathbf{Y}\mathbf{M}(\mathbf{f}) + \mathbf{Z}\mathbf{N}(\mathbf{f}) + \mathbf{C}] = \mathbf{Y}E[\mathbf{M}(\mathbf{f})] + \mathbf{Z}E[\mathbf{N}(\mathbf{f})] + \mathbf{C} \tag{A.8.7}$$

with $\mathbf{Y}$, $\mathbf{Z}$, and $\mathbf{C}$ fixed matrices.

Now, let us denote a matrix function of two vectors as $\mathbf{H}(\mathbf{x}, \mathbf{f})$. Its expected value over $\mathbf{f}$ with $\mathbf{x}$ a particular fixed vector is expressed as

$$E[\mathbf{H}(\mathbf{x}, \mathbf{f})|\mathbf{x}] = \int \mathbf{H}(\mathbf{x}, \mathbf{f}) P(\mathbf{f}|\mathbf{x}, \Theta)\mathrm{d}\mathbf{f}. \tag{A.8.8}$$

Here, $\mathbf{x}$ being fixed is indicated by the fact that $\mathbf{H}(\mathbf{f},\mathbf{x})$ is followed by $|\mathbf{x}$ on the left side and the PDF on the right side is a conditional PDF $P(\mathbf{f}|\mathbf{x}, \boldsymbol{\Theta})$.

A.8.3 Covariances as Expected Values

Let us consider the *covariance defined theoretically using probabilities*. The covariance between two random variables x and y is defined as the *expected value* of the product of $x - E[x]$ and $y - E[y]$, that is, $E[(x - E[x])\,(y - E[y])]$. Thus, the

$m \times m$ covariance matrix among the m continuous random variables in $\mathbf{f} = [f_1, \ldots, f_m]'$ is defined as

$$V[\mathbf{f}] = E\big[(\mathbf{f} - E[\mathbf{f}])(\mathbf{f} - E[\mathbf{f}])'\big] = \int (\mathbf{f} - E[\mathbf{f}])(\mathbf{f} - E[\mathbf{f}])' P(\mathbf{f}|\boldsymbol{\Theta}) \mathrm{d}\mathbf{f}, \quad \text{(A.8.9)}$$

whose (j, k) element is $E[(f_j - E[f_j])\,(f_k - E[f_k])]$, i.e., the covariance between f_j and f_k.

Covariance matrix (A.8.9) can be rewritten as

$$V[\mathbf{f}] = E[\mathbf{f}\mathbf{f}'] - E[\mathbf{f}]E[\mathbf{f}]', \quad \text{(A.8.10)}$$

since (A.8.9) can be rewritten as follows:

$$\begin{aligned} V[\mathbf{f}] &= \int \big(\mathbf{f}\mathbf{f}' - \mathbf{f}E[\mathbf{f}]' - E[\mathbf{f}]\mathbf{f}' - E[\mathbf{f}]E[\mathbf{f}]'\big) P(\mathbf{f}|\boldsymbol{\Theta}) \mathrm{d}\mathbf{f} \\ &= \int \mathbf{f}\mathbf{f}' P(\mathbf{f}|\boldsymbol{\Theta}) \mathrm{d}\mathbf{f} - \int \mathbf{f} P(\mathbf{f}|\boldsymbol{\Theta}) \mathrm{d}\mathbf{f} E[\mathbf{f}]' \\ &\quad - E[\mathbf{f}] \int \mathbf{f}' P(\mathbf{f}|\boldsymbol{\Theta}) \mathrm{d}\mathbf{f} + E[\mathbf{f}]E[\mathbf{f}]' \int P(\mathbf{f}|\boldsymbol{\Theta}) \mathrm{d}\mathbf{f} \\ &= E[\mathbf{f}\mathbf{f}'] - E[\mathbf{f}]E[\mathbf{f}]' - E[\mathbf{f}]E[\mathbf{f}]' + E[\mathbf{f}]E[\mathbf{f}]', \end{aligned} \quad \text{(A.8.11)}$$

where we have used (A.8.4) and $\int P(\mathbf{f}|\boldsymbol{\Theta}) \mathrm{d}\mathbf{f} = 1$.

The $p \times m$ covariance matrix between $\mathbf{x} = [x_1, \ldots, x_p]'$ and $\mathbf{f}$ is defined as

$$E\big[(\mathbf{x} - E[\mathbf{x}])(\mathbf{f} - E[\mathbf{f}])'\big] = \iint (\mathbf{x} - E[\mathbf{x}])(\mathbf{f} - E[\mathbf{f}])' P(\mathbf{x}, \mathbf{f}|\boldsymbol{\Theta}) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{f}. \quad \text{(A.8.12)}$$

The $m \times m$ covariance matrix among the m random variables in $\mathbf{f} = [f_1, \ldots, f_m]'$, which is conditional on $\mathbf{x}$ being a particular vector, is expressed as

$$V[\mathbf{f}|\mathbf{x}] = \int (\mathbf{f} - E[\mathbf{f}|\mathbf{x}])(\mathbf{f} - E[\mathbf{f}|\mathbf{x}])' P(\mathbf{f}|\mathbf{x}, \boldsymbol{\Theta}) \mathrm{d}\mathbf{f} = E[\mathbf{f}\mathbf{f}'|\mathbf{x}] - E[\mathbf{f}|\mathbf{x}]E[\mathbf{f}|\mathbf{x}]'. \quad \text{(A.8.13)}$$

Here, the last identity can be derived by (A.8.11) whose $E[\mathbf{f}]$, $E[\mathbf{f}\mathbf{f}']$, and $P(\mathbf{f}|\boldsymbol{\Theta})$ are replaced by $E[\mathbf{f}|\mathbf{x}], E[\mathbf{f}\mathbf{f}'|\mathbf{x}]$, and $P(\mathbf{f}|\mathbf{x}, \boldsymbol{\Theta})$, respectively.

A.8.4 Expected Values and Covariances of Multivariate Normal Variables

Let us suppose that a random vector $\mathbf{x}$ follows the *multivariate normal* (*MVN*) *distribution* whose PDF is given by (8.9) (Chap. 8), i.e.,

$$P(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}. \tag{A.8.14}$$

Then, the expected vector and covariance matrix for $\mathbf{x}$ are known to be

$$E[\mathbf{x}] = \boldsymbol{\mu} \text{ and } V[\mathbf{x}] = \boldsymbol{\Sigma} \tag{A.8.15}$$

(e.g., Anderson 2003). For this reason, (A.8.14) is described as a PDF with *mean* vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Here, "*mean*" is a synonym of "average" and "expected value".

A.8.5 EM Algorithm

The equations of (A.8.1)–(A.8.8) still hold true, if the vectors in $\mathbf{x}$ and $\mathbf{f}$ are replaced by scalars and matrices: For example, the matrix version of (A.8.1) is $P(\mathbf{X}|\boldsymbol{\Theta}) = \int P(\mathbf{X}, \mathbf{F}|\boldsymbol{\Theta})\mathrm{d}\mathbf{F}$.

Let us consider the *maximum likelihood method*, in which the *log likelihood* $\log P(\mathbf{X}|\boldsymbol{\Theta})$ is maximized over parameter matrix $\boldsymbol{\Theta}$ for a given $n \times p$ data matrix $\mathbf{X}$, as in Chap. 8. For this maximization, the EM algorithm is useful (Dempster, Laird and Rubin, 1977), when an $n \times m$ matrix $\mathbf{F}$ is associated with $\mathbf{X}$, and the expectation $\log P(\mathbf{X}, \mathbf{F}|\boldsymbol{\Theta})$ for a given $\mathbf{X}$, i.e.,

$$E[\log P(\mathbf{X}, \mathbf{F}|\boldsymbol{\Theta})|\mathbf{X}] = \int \log P(\mathbf{X}, \mathbf{F}|\boldsymbol{\Theta}) P(\mathbf{F}|\mathbf{X}, \boldsymbol{\Theta})\mathrm{d}\mathbf{F}, \tag{A.8.16}$$

is easier to handle than $\log P(\mathbf{X}|\boldsymbol{\Theta})$. The algorithm deals with iteratively obtaining (A.8.16) for the current $\boldsymbol{\Theta}$ and updating $\boldsymbol{\Theta}$ so that(A.8.16) increases. Why this leads to the maximization of $\log P(\mathbf{X}|\boldsymbol{\Theta})$ can be explained by the following inequality, which Danish mathematician Johan Jensen (1859–1925) presented:

Theorem A.8.1. A Special Case of Jensen's Inequality

$$\log E[y] \geq E[\log y]. \tag{A.8.17}$$

The theorem leads to

$$\log E\left[\frac{P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta}_{\text{new}})}{P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta})}|\mathbf{X}\right] \geq E\left[\log \frac{P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta}_{\text{new}})}{P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta})}|\mathbf{X}\right] \tag{A.8.18}$$

with $\boldsymbol{\Theta}_{\text{new}}$ the updated version of $\boldsymbol{\Theta}$. As proved in this section below, the left and right sides of (A.8.18) can be rewritten as

$$\log E\left[\frac{P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta}_{\text{new}})}{P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta})}|\mathbf{X}\right] = \log P(\mathbf{X}|\boldsymbol{\Theta}_{\text{new}}) - \log P(\mathbf{X}|\boldsymbol{\Theta}), \tag{A.8.19}$$

$$E\left[\log \frac{P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta}_{\text{new}})}{P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta})}|\mathbf{X}\right] = E[\log P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta}_{\text{new}})|\mathbf{X}] - E[\log P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta})|\mathbf{X}], \tag{A.8.20}$$

respectively. Thus, we have

$$\begin{aligned} &\log P(\mathbf{X}|\boldsymbol{\Theta}_{\text{new}}) - \log P(\mathbf{X}|\boldsymbol{\Theta}) \\ &\geq E[\log P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta}_{\text{new}})|\mathbf{X}] - E[\log P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta})|\mathbf{X}], \end{aligned} \tag{A.8.21}$$

by substituting (A.8.19) and (A.8.20) in (A.8.18).

In (A.8.21), we can find that $\log P(\mathbf{X}|\boldsymbol{\Theta}_{\text{new}}) \geq \log P(\mathbf{X}|\boldsymbol{\Theta})$ is guaranteed, if $\boldsymbol{\Theta}$ is updated to $\boldsymbol{\Theta}_{\text{new}}$ so that $E[\log P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta}_{\text{new}})|\mathbf{X}] \geq E[\log P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta})|\mathbf{X}]$. Thus, $\log P(\mathbf{X}|\boldsymbol{\Theta})$ can reach its maximum, by iterating the following two steps:

$$E\text{-}step: \text{ Obtain } E[\log P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta})|\mathbf{X}]; \tag{A.8.22}$$

$$\begin{aligned} M\text{-}step: &\text{ Update } \boldsymbol{\Theta} \text{ so as to increase } E[\log P(\mathrm{X},\mathbf{F}|\boldsymbol{\Theta})|\mathrm{X}], \\ &\text{i.e., lead to } E[\log P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta}_{\text{new}})] \geq E[\log P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta})]. \end{aligned} \tag{A.8.23}$$

Here, the *E*- and *M-steps* are the abbreviations for *expectation* and *maximization* steps, respectively. The name *M-step* follows from the fact that "maximize" can be substituted for "increase" in that step and indicates the intent.

Let us prove (A.8.19) and (A.8.20). The latter directly is derived using (A.8.7):

$$E\left[\log\frac{P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta}_{\text{new}})}{P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta})}|\mathbf{X}\right] = E[\log P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta}_{\text{new}}) - \log P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta})|\mathbf{X}]$$
$$= E[\log P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta}_{\text{new}})|\mathbf{X}] - E[\log P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta})|\mathbf{X}].$$

On the other hand, (A.8.19) is derived by starting from its right side. It can be rewritten as

$$\log\frac{P(\mathbf{X}|\boldsymbol{\Theta}_{\text{new}})}{P(\mathbf{X}|\boldsymbol{\Theta})} = \log\frac{\int P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta}_{\text{new}})\mathrm{d}\mathbf{F}}{P(\mathbf{X}|\boldsymbol{\Theta})} = \log\int\left\{\frac{P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta}_{\text{new}})}{P(\mathbf{X}|\boldsymbol{\Theta})}\right\}\mathrm{d}\mathbf{F}, \quad \text{(A.8.24)}$$

since of (A.8.1) and $\int\phi(\mathbf{F})\mathrm{d}\mathbf{F}/\eta(\mathbf{X}) = \int\{\phi(\mathbf{F})/\eta(\mathbf{X})\}\mathrm{d}\mathbf{F}$, with $\eta(\mathbf{X})$ a function of $\mathbf{X}$ irrelevant to $\mathbf{F}$. The parenthesized fraction in (A.8.24) can be rewritten as

$$\frac{P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta}_{\text{new}})}{P(\mathbf{X}|\boldsymbol{\Theta})} = \frac{P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta}_{\text{new}})}{P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta})}\frac{P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta})}{P(\mathbf{X}|\boldsymbol{\Theta})} = \frac{P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta}_{\text{new}})}{P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta})}P(\mathbf{F}|\mathbf{X},\boldsymbol{\Theta}), \quad \text{(A.8.25)}$$

where the last identity is derived using (A.8.3). We can substitute $P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta}_{\text{new}})/P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta})$ and $P(\mathbf{F}|\mathbf{X},\boldsymbol{\Theta})$ in (A.8.25) into $\mathbf{H}(\mathbf{x},\mathbf{f})$ and $P(\mathbf{f}|\mathbf{x},\boldsymbol{\Theta})$ in the right side of (A.8.8), respectively, so as to have

$$\int\frac{P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta}_{\text{new}})}{P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta})}P(\mathbf{F}|\mathbf{X},\boldsymbol{\Theta})\mathrm{d}\mathbf{F} = E\left[\frac{P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta}_{\text{new}})}{P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta})}|\mathbf{X}\right]. \quad \text{(A.8.26)}$$

This logarithm, i.e., (A.8.24), leads to the left side of (A.8.19).

A.9 EM Algorithm for Factor Analysis

The *EM algorithm*, whose foundation is introduced in Appendix A.8.5, can be used for estimating the *factor analysis* (*FA*) solution. Here, FA is categorized into *confirmatory* (*CFA*), *exploratory* (*EFA*), and *sparse* (*SFA*), which are treated in Chaps. 10, 12, 18, and 22, though the formulation of EFA in Chap. 18 is not related to the EM algorithm. The descriptions for the EM algorithm in Sects. A.9.1–A.9.5 are common among *EFA*, *CFA*, and *SFA*. The procedures specific to *EFA* and *CFA* are treated in A.9.6–A.9.8, while those for *SFA* are introduced in A.9.9.

A.9.1 Outline

Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]'$ be an n-individual $\times$ p-variables centered data matrix to be analyzed by FA, with $\mathbf{V} = n^{-1}\mathbf{X}'\mathbf{X}$ the inter-variable covariance matrix. The log likelihood to be maximized in FA can be expressed as (10.11), i.e.,

$$\log P(\mathbf{X}|\boldsymbol{\Theta}) \propto l^*(\boldsymbol{\Theta}) = \frac{n}{2}\log\left|(\mathbf{A}\boldsymbol{\Phi}\mathbf{A}' + \Psi)^{-1}\mathbf{V}\right| - \frac{n}{2}\mathrm{tr}(\mathbf{A}\boldsymbol{\Phi}\mathbf{A}' + \Psi)^{-1}\mathbf{V} \quad \text{(A.9.1)}$$

with $\boldsymbol{\Theta} = \{\mathbf{A}, \boldsymbol{\Psi}, \boldsymbol{\Phi}\}$ a set of the parameters to be estimated. Here, $\log P(\mathbf{X}|\boldsymbol{\Theta}) \propto l(\boldsymbol{\Theta})$ stands for both of its sides being mutually *proportional*: as found from (8.19) and (8.20), $\log P(\mathbf{X}|\boldsymbol{\Theta})$ for the multivariate normal (MVN) distribution is the sum of log likelihood and a constant that is not dependent on $\boldsymbol{\Theta}$. In (A.9.1), $\mathbf{A}$ is the $p \times m$ matrix containing factor loadings, $\boldsymbol{\Psi}$ is the $p \times p$ diagonal matrix whose diagonal elements are unique variances, and $\boldsymbol{\Phi}$ is an $m \times m$ factor correlation matrix, which can be fixed to $\mathbf{I}_m$ in exploratory FA (EFA) as explained in Chap. 12.

In the EM algorithm, $E[\log P(\mathbf{X}, \mathbf{F}|\boldsymbol{\Theta})|\mathbf{X}]$ is considered in order to maximize (A.9.1), as explained in A.8.5. In this appendix, $\mathbf{F} = [\mathbf{f}_1, \ldots, \mathbf{f}_n]$ is the n-individuals $\times$ m-factors matrix of common factor scores, with $\mathbf{f}_i$ the score vector for individual i (= 1, …, n). We can view $\log P(\mathbf{X}, \mathbf{F}|\boldsymbol{\Theta})$ as the log likelihood of $\boldsymbol{\Theta}$ for $\mathbf{F}$ supposed to be a data set observed together with $\mathbf{X}$ (though $\mathbf{F}$ is not observed in reality). In this sense, $\log P(\mathbf{X}, \mathbf{F}|\boldsymbol{\Theta})$ and $E[\log P(\mathbf{X}, \mathbf{F}|\boldsymbol{\Theta})|\mathbf{X}]$ are referred to as the *complete data log likelihood* and *expected complete data log likelihood*, respectively, in some literature.

The *E-step* (A.8.22) and *M-step* (A.8.23) are iterated until convergence is reached in the EM algorithm, in order to obtain the FA parameters in $\boldsymbol{\Theta}$. Before describing those two steps, we must express the *complete data log likelihood* $\log P(\mathbf{X}, \mathbf{F}|\boldsymbol{\Theta})$ and its *expected value* $E[\log P(\mathbf{X}, \mathbf{F}|\boldsymbol{\Theta})|\mathbf{X}]$ *explicitly*. In the next two sections, $\log P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta})$ and $E[\log P(\mathbf{X}, \mathbf{F}|\boldsymbol{\Theta})|\mathbf{X}]$ are treated subsequently.

A.9.2 Complete Data Log Likelihood

Let us explicitly express $\log P(\mathbf{X}, \mathbf{F}|\boldsymbol{\Theta})$ for the FA model (10.3), with (10.4) and (10.6). The matrix version of (A.8.2), $P(\mathbf{X}, \mathbf{F}|\boldsymbol{\Theta}) = P(\mathbf{X}|\mathbf{F}, \boldsymbol{\Theta})P(\mathbf{F}|\boldsymbol{\Theta})$, leads to

$$\log P(\mathbf{X}, \mathbf{F}|\boldsymbol{\Theta}) = \log P(\mathbf{X}|\mathbf{F}, \boldsymbol{\Theta}) + \log P(\mathbf{F}|\boldsymbol{\Theta}). \quad \text{(A.9.2)}$$

Supposing the mutual independence among $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and that among $\mathbf{f}_1, \ldots, \mathbf{f}_n$ lead to $P(\mathbf{X}|\mathbf{F}, \boldsymbol{\Theta}) = \prod_{i=1}^{n} P(\mathbf{x}_i|\mathbf{f}_i, \boldsymbol{\Theta})$ and $P(\mathbf{F}|\boldsymbol{\Theta}) = \prod_{i=1}^{n} P(\mathbf{f}_i|\boldsymbol{\Theta})$. Here, the latter logarithm is found to be

$$\log P(\mathbf{F}|\boldsymbol{\Theta}) = \sum_{i=1}^{n} \log P(\mathbf{f}_i|\boldsymbol{\Theta}) = -\frac{nm}{2}\log 2\pi - \frac{n}{2}\log|\boldsymbol{\Phi}| - \frac{1}{2}\sum_{i=1}^{n} \mathbf{f}_i'\boldsymbol{\Phi}^{-1}\mathbf{f}_i, \tag{A.9.3}$$

since of (10.4) with (8.19), while the former one can be derived using the next theorem:

Theorem A.9.1. A Property of the Multivariate Normal Distribution
If $\mathbf{x}$ is a $p \times 1$ vector of random variables satisfying $\mathbf{x} = \mathbf{t} + \mathbf{u}$, with $\mathbf{u} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Omega})$, then $\mathbf{x}$ follows $N_p(\mathbf{t} + \boldsymbol{\mu}, \boldsymbol{\Omega})$ conditional on $\mathbf{t}$ being a particular vector. This is expressed as

$$\mathbf{x}|\mathbf{t} \sim N_p(\mathbf{t} + \boldsymbol{\mu}, \boldsymbol{\Omega}), \tag{A.9.4}$$

with the probability density function (PDF) of $N_p(\mathbf{t} + \boldsymbol{\mu}, \boldsymbol{\Omega})$ denoted by $\mathrm{P}(\mathbf{x}|\mathbf{t})$ (e.g., Anderson, 2003).

By taking account of (10.3) and (10.6) in this theorem, we can find $P(\mathbf{x}_i|\mathbf{f}_i, \boldsymbol{\Theta})$ to be the PDF of $N_p(\mathbf{A}\mathbf{f}_i, \boldsymbol{\Psi})$. This fact is expressed as

$$\mathbf{x}_i|\mathbf{f}_i \sim N_p(\mathbf{A}\mathbf{f}_i, \boldsymbol{\Psi}). \tag{A.9.5}$$

Thus, the logarithm of $P(\mathbf{X}|\mathbf{F}, \boldsymbol{\Theta}) = \prod_{i=1}^{n} P(\mathbf{x}_i|\mathbf{f}_i, \boldsymbol{\Theta})$ is expressed, using (A.9.5) in (8.19), as

$$\begin{aligned}\log P(\mathbf{X}|\mathbf{F}, \boldsymbol{\Theta}) &= \sum_{i=1}^{n} \log P(\mathbf{x}_i|\mathbf{f}_i, \boldsymbol{\Theta}) \\ &= -\frac{np}{2}\log 2\pi - \frac{n}{2}\log|\boldsymbol{\Psi}| - \frac{1}{2}\sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{A}\mathbf{f}_i)'\boldsymbol{\Psi}^{-1}(\mathbf{x}_i - \mathbf{A}\mathbf{f}_i).\end{aligned} \tag{A.9.6}$$

Using (A.9.3) and (A.9.6) in (A.9.2) with $c = -2^{-1}n(p + m)\log 2\pi$ and $\mathbf{V} = n^{-1}\mathbf{X}'\mathbf{X}$, we have

$$\begin{aligned}\log P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta}) &= -\frac{n}{2}\log|\boldsymbol{\Psi}| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i-\mathbf{A}\mathbf{f}_i)'\boldsymbol{\Psi}^{-1}(\mathbf{x}_i-\mathbf{A}\mathbf{f}_i) - \frac{n}{2}\log|\boldsymbol{\Phi}| - \frac{1}{2}\sum_{i=1}^{n}\mathbf{f}_i'\boldsymbol{\Phi}^{-1}\mathbf{f}_i + c \\ &= -\frac{n}{2}\log|\boldsymbol{\Psi}| - \frac{n}{2}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}'_i\boldsymbol{\Psi}^{-1}\mathbf{x}_i - \frac{2}{n}\sum_{i=1}^{n}\mathbf{f}_i'\mathbf{A}'\boldsymbol{\Psi}^{-1}\mathbf{x}_i + \frac{1}{n}\sum_{i=1}^{n}\mathbf{f}_i'\mathbf{A}'\boldsymbol{\Psi}^{-1}\mathbf{A}\mathbf{f}_i\right) \\ &\quad -\frac{n}{2}\left(\log|\boldsymbol{\Phi}| + \frac{1}{n}\sum_{i=1}^{n}\mathbf{f}_i'\boldsymbol{\Phi}^{-1}\mathbf{f}_i\right) + c \\ &= -\frac{n}{2}\log|\boldsymbol{\Psi}| - \frac{n}{2}\left(\mathrm{tr}\mathbf{V}\boldsymbol{\Psi}^{-1} - \frac{2}{n}\mathrm{tr}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{f}_i'\mathbf{A}'\boldsymbol{\Psi}^{-1} + \frac{1}{n}\mathrm{tr}\mathbf{A}\sum_{i=1}^{n}\mathbf{f}_i\mathbf{f}_i'\mathbf{A}'\boldsymbol{\Psi}^{-1}\right) \\ &\quad -\frac{n}{2}\left(\log|\boldsymbol{\Phi}| + \frac{1}{n}\mathrm{tr}\sum_{i=1}^{n}\mathbf{f}_i\mathbf{f}_i'\boldsymbol{\Phi}^{-1}\right) + c \\ &= -\frac{n}{2}\log|\boldsymbol{\Psi}| - \frac{n}{2}\mathrm{tr}\left(\mathbf{V} - \frac{2}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{f}_i'\mathbf{A}' + \frac{1}{n}\mathbf{A}\sum_{i=1}^{n}\mathbf{f}_i\mathbf{f}_i'\mathbf{A}'\right)\boldsymbol{\Psi}^{-1} \\ &\quad -\frac{n}{2}\left(\log|\boldsymbol{\Phi}| + \frac{1}{n}\mathrm{tr}\boldsymbol{\Phi}^{-1}\sum_{i=1}^{n}\mathbf{f}_i\mathbf{f}_i'\right) + c.\end{aligned} \tag{A.9.7}$$

A.9.3 Expected Complete Data Log Likelihood

In this section, we consider the *expected value* of (A.9.7) for a given $\mathbf{X}$, i.e., $E[\log P(\mathbf{X}, \mathbf{F}|\boldsymbol{\Theta})|\mathbf{X}]$. Using (A.8.7) and (A.8.8), the expected value is expressed as

$$\begin{aligned}E[\log P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta})|\mathbf{X}] &= \int \log P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta})P(\mathbf{F}|\mathbf{X},\boldsymbol{\Theta})\mathrm{d}\mathbf{F} \\ &= -\frac{n}{2}\log|\boldsymbol{\Psi}| - \frac{n}{2}\mathrm{tr}\left(\mathbf{V} - 2E\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{f}_i'|\mathbf{X}\right]\mathbf{A}' + \mathbf{A}E\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{f}_i\mathbf{f}_i'|\mathbf{X}\right]\mathbf{A}'\right)\boldsymbol{\Psi}^{-1} \\ &\quad -\frac{n}{2}\left(\log|\boldsymbol{\Phi}| + \mathrm{tr}\boldsymbol{\Phi}^{-1}E\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{f}_i\mathbf{f}_i'|\mathbf{X}\right]\right) + c \\ &= -\frac{n}{2}\log|\boldsymbol{\Psi}| - \frac{n}{2}\mathrm{tr}(\mathbf{V} - 2\mathbf{B}\mathbf{A}' + \mathbf{A}\mathbf{Q}\mathbf{A}')\boldsymbol{\Psi}^{-1} - \frac{n}{2}\left(\log|\boldsymbol{\Phi}| + \mathrm{tr}\boldsymbol{\Phi}^{-1}\mathbf{Q}\right) + c,\end{aligned} \tag{A.9.8}$$

with

$$\mathbf{B} = E\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{f}_i'|\mathbf{X}\right] = \frac{1}{n}\sum_{i=1}^{n}E\left[\mathbf{x}_i\mathbf{f}_i'|\mathbf{X}\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_iE[\mathbf{f}_i|\mathbf{X}]' = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_iE[\mathbf{f}_i|\mathbf{x}_i]', \tag{A.9.9}$$

$$\mathbf{Q} = E\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{f}_i\mathbf{f}_i'|\mathbf{X}\right] = \frac{1}{n}\sum_{i=1}^{n}E\left[\mathbf{f}_i\mathbf{f}_i'|\mathbf{X}\right] = \frac{1}{n}\sum_{i=1}^{n}E\left[\mathbf{f}_i\mathbf{f}_i'|\mathbf{x}_i\right]. \tag{A.9.10}$$

Here, we have used $E[\mathbf{f}_i|\mathbf{X}] = E[\mathbf{f}_i|\mathbf{x}_i]$ and $E\left[\mathbf{f}_i\mathbf{f}_i'|\mathbf{X}\right] = E\left[\mathbf{f}_i\mathbf{f}_i'|\mathbf{x}_i\right]$ which follow from the mutual independence supposed among $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and among $\mathbf{f}_1, \ldots, \mathbf{f}_n$.

Thus, the *E-step* (A.8.22) for FA is restated as obtaining (A.9.9) and (A.9.10). The step is followed by the *M-step*, in which $\mathbf{A}$, $\boldsymbol{\Psi}$, and $\boldsymbol{\Phi}$ are updated separatory so as to increase (A.9.8). In the next section, we detail the E-step, followed by the sections regarding the M-step. Here, it should be cautioned that the E-step and how to update $\boldsymbol{\Psi}$ in the M-step are *common* for EFA, CFA, and SFA. On the other hand, the updating of $\mathbf{A}$ and $\boldsymbol{\Phi}$ in the M-step *differs* among EFA, CFA, and SFA: Their update in *EFA* and *CFA* is described in Sects. A.9.6–A.9.8, while that for *SFA* is treated in A.9.9.

A.9.4 E-Step

For obtaining (A.9.9) and (A.9.10), it is required to find how $E[\mathbf{f}_i|\mathbf{x}_i]$ and $E\left[\mathbf{f}_i\mathbf{f}_i'|\mathbf{x}_i\right]$ are explicitly expressed under the FA model (10.3), (10.4), (10.6), and independence of (10.4) to (10.6). They can be summarized as

$$\begin{bmatrix}\mathbf{f}\\ \mathbf{x}\end{bmatrix} \sim N_{p+m}\left(\begin{bmatrix}\mathbf{0}_m\\ \mathbf{0}_p\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Phi} & \boldsymbol{\Phi}\mathbf{A}'\\ \mathbf{A}\boldsymbol{\Phi} & \boldsymbol{\Sigma}_\Theta\end{bmatrix}\right). \tag{A.9.11}$$

with

$$\boldsymbol{\Sigma}_\Theta = \mathbf{A}\boldsymbol{\Phi}\mathbf{A}' + \Psi \tag{A.9.12}$$

as found in (10.10). Here, the covariance matrix between $\mathbf{x}$ and $\mathbf{f}$ being $\mathbf{A}\boldsymbol{\Phi}$ is derived from the fact that $E[(\mathbf{e} - E[\mathbf{e}])(\mathbf{f} - E[\mathbf{f}])]' = E[\mathbf{e}\mathbf{f}'] =_p \mathbf{O}_m$ and thus $E[(\mathbf{x} - E[\mathbf{x}])(\mathbf{f} - E[\mathbf{f}])]' = E[\mathbf{x}\mathbf{f}'] = E[(\mathbf{A}\mathbf{f} + \mathbf{e})\mathbf{f}']' = \mathbf{A}E[\mathbf{f}\mathbf{f}']' = \mathbf{A}\boldsymbol{\Phi}$, since (10.3), (10.4), (10.9), (A.8.15), and the mutual independence of $\mathbf{f}$ and $\mathbf{e}$.

We can derive $E[\mathbf{f}_i|\mathbf{x}_i]$ and $E\left[\mathbf{f}_i\mathbf{f}_i'|\mathbf{x}_i\right]$ in (A.9.9) and (A.9.10) from (A.9.11), using the next theorem:

Theorem A.9.2. Conditional Multivariate Normal Distribution
Let us suppose that the $(q + r) \times 1$ vector $[\mathbf{y}', \mathbf{z}']'$, which consists of $\mathbf{y}$ ($q \times 1$) and $\mathbf{z}$ ($r \times 1$), follows an MVN distribution: $\begin{bmatrix}\mathbf{y}\\ \mathbf{z}\end{bmatrix} \sim N_{q+r}\left(\begin{bmatrix}\boldsymbol{\mu}_y\\ \boldsymbol{\mu}_z\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yz}\\ \boldsymbol{\Sigma}_{yz}' & \boldsymbol{\Sigma}_{zz}\end{bmatrix}\right)$ with $\boldsymbol{\mu}_y$, $\boldsymbol{\mu}_z$, and $\boldsymbol{\Sigma}_{yz}$ being $q \times 1$, $r \times 1$, and $q \times r$, respectively. Then,

$$\mathbf{y}|\mathbf{z} \sim N_q\left(\boldsymbol{\mu}_y + \boldsymbol{\Sigma}_{yz}\boldsymbol{\Sigma}_{zz}^{-1}(\mathbf{z} - \boldsymbol{\mu}_z), \boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}_{yz}\boldsymbol{\Sigma}_{zz}^{-1}\boldsymbol{\Sigma}_{yz}'\right) \tag{A.9.13}$$

(e.g., Anderson, 2003).

Comparing this theorem with (A.9.11), we can find

$$\mathbf{f}_i|\mathbf{x}_i \sim N_m\left(\boldsymbol{\Phi}\mathbf{A}'\boldsymbol{\Sigma}_\Theta^{-1}\mathbf{x}_i, \boldsymbol{\Phi} - \boldsymbol{\Phi}\mathbf{A}'\boldsymbol{\Sigma}_\Theta^{-1}\mathbf{A}\boldsymbol{\Phi}\right). \tag{A.9.14}$$

This and (A.8.15) imply

$$E[\mathbf{f}_i|\mathbf{x}_i] = \boldsymbol{\Phi}\mathbf{A}'\boldsymbol{\Sigma}_\Theta^{-1}\mathbf{x}_i, \tag{A.9.15}$$

$$V[\mathbf{f}_i|\mathbf{x}_i] = \boldsymbol{\Phi} - \boldsymbol{\Phi}\mathbf{A}'\boldsymbol{\Sigma}_\Theta^{-1}\mathbf{A}\boldsymbol{\Phi}. \tag{A.9.16}$$

Using these two equations in $E[\mathbf{f}\mathbf{f}'|\mathbf{x}] = V[\mathbf{f}|\mathbf{x}] + E[\mathbf{f}|\mathbf{x}]E[\mathbf{f}|\mathbf{x}]'$ following from (A.8.13), we can find

$$\begin{aligned} E\left[\mathbf{f}_i\mathbf{f}_i'|\mathbf{x}_i\right] &= V[\mathbf{f}_i|\mathbf{x}_i] + E[\mathbf{f}_i|\mathbf{x}_i]E[\mathbf{f}_i|\mathbf{x}_i]' \\ &= \boldsymbol{\Phi} - \boldsymbol{\Phi}\mathbf{A}'\boldsymbol{\Sigma}_\Theta^{-1}\mathbf{A}\boldsymbol{\Phi} + \boldsymbol{\Phi}\mathbf{A}'\boldsymbol{\Sigma}_\Theta^{-1}\mathbf{x}_i\mathbf{x}_i'\,\boldsymbol{\Sigma}_\Theta^{-1}\mathbf{A}\boldsymbol{\Phi}. \end{aligned} \tag{A.9.17}$$

By substituting (A.9.15) and (A.9.17) in (A.9.9) and (A.9.10), these two equations can be rewritten as

$$\mathbf{B} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\boldsymbol{\Sigma}_\Theta^{-1}\mathbf{A}\boldsymbol{\Phi} = \mathbf{V}\mathbf{H}(\boldsymbol{\Theta}), \tag{A.9.18}$$

$$\begin{aligned} \mathbf{Q} &= \boldsymbol{\Phi} - \boldsymbol{\Phi}\mathbf{A}'\,\boldsymbol{\Sigma}_\Theta^{-1}\mathbf{A}\boldsymbol{\Phi} + \boldsymbol{\Phi}\mathbf{A}'\,\boldsymbol{\Sigma}_\Theta^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\right)\boldsymbol{\Sigma}_\Theta^{-1}\mathbf{A}\boldsymbol{\Phi} \\ &= \mathbf{W}(\boldsymbol{\Theta}) + \mathbf{H}(\boldsymbol{\Theta})'\mathbf{V}\mathbf{H}(\boldsymbol{\Theta}) \end{aligned} \tag{A.9.19}$$

with $\mathbf{V} = n^{-1}\mathbf{X}'\mathbf{X}$. Here, $\mathbf{H}(\boldsymbol{\Theta})$ ($p \times m$) and $\mathbf{W}(\boldsymbol{\Theta})$ ($p \times m$) are matrix functions of $\boldsymbol{\Theta} = \{\mathbf{A}, \boldsymbol{\Psi}, \boldsymbol{\Phi}\}$:

$$\mathbf{H}(\boldsymbol{\Theta}) = \boldsymbol{\Sigma}_\Theta^{-1}\mathbf{A}\boldsymbol{\Phi}, \tag{A.9.20}$$

$$\mathbf{W}(\boldsymbol{\Theta}) = \boldsymbol{\Phi} - \boldsymbol{\Phi}\mathbf{A}'\boldsymbol{\Sigma}_\Theta^{-1}\mathbf{A}\boldsymbol{\Phi} = \boldsymbol{\Phi}^{1/2}\left(\mathbf{I}_m + \boldsymbol{\Phi}^{1/2}\mathbf{A}'\boldsymbol{\Psi}^{-1}\mathbf{A}\boldsymbol{\Phi}^{1/2}\right)^{-1}\boldsymbol{\Phi}^{1/2}, \tag{A.9.21}$$

with $\mathbf{\Phi}^{1/2}$ the *symmetric square root* of $\mathbf{\Phi}$: $\mathbf{\Phi}^{1/2}\mathbf{\Phi}^{1/2} = \mathbf{\Phi}$ and $\mathbf{\Phi}^{1/2\prime} = \mathbf{\Phi}^{1/2}$. The last identity in (A.9.21) has been derived by Adachi (2013, Lemma 1), using the following relation:

Theorem A.9.3. Inverse of a Sum of Matrices (Seber, 2008, p. 309)
Let $\mathbf{M}$ and $\mathbf{N}$ be nonsingular matrices. Then, we have

$$\mathbf{M}^{-1} - \mathbf{M}^{-1}\mathbf{U}\left(\mathbf{Y}\mathbf{M}^{-1}\mathbf{U} + \mathbf{N}^{-1}\right)^{-1}\mathbf{Y}\mathbf{M}^{-1} = (\mathbf{M} + \mathbf{U}\mathbf{N}\mathbf{Y})^{-1}. \tag{A.9.22}$$

We can use (A.9.12) to rewrite $\mathbf{\Phi} - \mathbf{\Phi}\mathbf{A}'\mathbf{\Sigma}_{\Theta}^{-1}\mathbf{A}\mathbf{\Phi}$ in (A.9.21) as $\mathbf{\Phi}^{1/2}\mathbf{\Omega}\mathbf{\Phi}^{1/2}$ with $\mathbf{\Omega} = \mathbf{I}_m - \mathbf{\Phi}^{1/2}\mathbf{A}'\mathbf{\Sigma}_{\Theta}^{-1}\mathbf{A}\mathbf{\Phi}^{1/2} = \mathbf{I}_m - \mathbf{\Phi}^{1/2}\mathbf{A}'(\mathbf{A}\mathbf{\Phi}\mathbf{A}' + \mathbf{\Psi})^{-1}\mathbf{A}\mathbf{\Phi}^{1/2}$. By setting $\mathbf{M} = \mathbf{I}_m$, $\mathbf{U} = \mathbf{\Phi}^{1/2}\mathbf{A}'$, $\mathbf{Y} = \mathbf{A}\mathbf{\Phi}^{1/2}$, and $\mathbf{N} = \mathbf{\Psi}$ in the left side of (A.9.22), its right side is found to be $\mathbf{\Omega} = \left(\mathbf{I}_m + \mathbf{\Phi}^{1/2}\mathbf{A}'\mathbf{\Psi}^{-1}\mathbf{A}\mathbf{\Phi}^{1/2}\right)^{-1}$. This substitution in $\mathbf{W}(\mathbf{\Theta}) = \mathbf{\Phi}^{1/2}\mathbf{\Omega}\mathbf{\Phi}^{1/2}$ leads to (A.9.21). Using (A.9.19) with (A.9.21), Adachi (2013) has shown that $\mathbf{Q}$ is *positive-definite*, if $\mathbf{\Psi}$ and $\mathbf{\Phi}$ are *positive-definite* and $\mathbf{V}$ is *nonnegative-definite*. See Note 8.2 for the nonnegative- and positive-definiteness.

A.9.5 Updating Unique Variances in M-Step

Let us consider maximizing (A.9.8) over the diagonal $\mathbf{\Psi}$ with $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_p]'$ and $\mathbf{\Phi}$ kept fixed. We should notice that (A.9.8) can be rewritten as $-(n/2)\sum_{j=1}^{p} h_j(\psi_j) + const_{[\Psi]}$. Here, ψ_j is the jth diagonal element of $\mathbf{\Psi}$, $const_{[\Psi]}$ is independent of $\mathbf{\Psi}$, and

$$h_j(\psi_j) = \log\psi_j + \frac{1}{\psi_j}\left(v_{jj} - 2\mathbf{b}_j'\mathbf{a}_j + \mathbf{a}_j'\mathbf{Q}\mathbf{a}_j\right) = \log\psi_j + \frac{u_j}{\psi_j} \tag{A.9.23}$$

with $\mathbf{b}_j'$ the jth row of (A.9.18), v_{jj} the jth diagonal element of $\mathbf{V}$, and $u_j = v_{ij} - 2\mathbf{b}_j'\mathbf{a}_j + \mathbf{a}_j'\mathbf{Q}\mathbf{a}_j$.

Thus, the maximization can be attained by minimizing (A.9.23) over ψ_j for $j = 1, \ldots, p$. The minimizer is given by

$$\psi_j = u_j = v_{jj} - 2\mathbf{b}_j'\mathbf{a}_j + \mathbf{a}_j'\mathbf{Q}\mathbf{a}_j, \tag{A.9.24}$$

from the following fact: The *differentiation* of (A.9.23) with respect to ψ_j is known to be given by $h_j'(\psi_j) = \mathrm{d}h_j(\psi_j)/\mathrm{d}\psi_j = 1/\psi_j - u_j/\psi_j^2 = (\psi_j - u_j)/\psi_j^2$. This shows that $h_j'(\psi_j) < 0$ for $\psi_j < u_j$, $h_j'(\psi_j) = 0$ for (A.9.24), and $h_j'(\psi_j) > 0$ for $\psi_j > u_j$.

A.9.6 Updating Factor Covariance in M-Step for CFA

In this section, we consider updating $\boldsymbol{\Phi}$. This update is *skipped* in *EFA* with $\boldsymbol{\Phi}$ fixed to $\mathbf{I}_m$. Thus, only the case for *CFA* is treated here: We consider how (A.9.8) is maximized over $\boldsymbol{\Phi}$ with $\mathbf{A}$ and $\boldsymbol{\Psi}$ fixed in CFA.

Below (A.9.1), we described that $\boldsymbol{\Phi}$ is a factor *correlation* matrix, whose diagonal elements are restricted to ones. However, $\boldsymbol{\Phi} = (\phi_{jk})$ may be regarded *simply* as a *covariance* matrix *without* the *restriction*, on the supposition of $\phi_{jj} > 0$ ($j = 1, \ldots, m$). This follows from the fact that the log likelihood (A.9.1) can be rewritten as

$$l(\mathbf{A}, \boldsymbol{\Psi}, \boldsymbol{\Phi}) = \frac{n}{2}\log\left|\left(\mathbf{A}_{\mathrm{R}}\boldsymbol{\Phi}_{\mathrm{R}}\mathbf{A}_{\mathrm{R}}' + \boldsymbol{\Psi}\right)^{-1}\mathbf{V}\right| - \frac{n}{2}\mathrm{tr}\left(\mathbf{A}_{\mathrm{R}}\boldsymbol{\Phi}_{\mathrm{R}}\mathbf{A}_{\mathrm{R}}' + \boldsymbol{\Psi}\right)^{-1}\mathbf{V}. \tag{A.9.25}$$

Here,

$$\mathbf{A}_{\mathrm{R}} = \mathbf{A}\mathrm{diag}(\boldsymbol{\Phi})^{1/2} \tag{A.9.26}$$

$$\boldsymbol{\Phi}_{\mathrm{R}} = \mathrm{diag}(\boldsymbol{\Phi})^{-1/2}\boldsymbol{\Phi}\mathrm{diag}(\boldsymbol{\Phi})^{-1/2} \tag{A.9.27}$$

with $\mathrm{diag}(\boldsymbol{\Phi})^{-1/2}$ the $m \times m$ diagonal matrix whose jth diagonal element is $\phi_{jj}^{-1/2}$. Here, we should note that the elements of (A.9.26) are zeros whose counterparts in $\mathbf{A}$ are constrained to be zeros in CFA and (A.9.27) is a correlation matrix whose diagonal elements are ones. Those points imply that the maximum attainable value of (A.9.1) does *not depend* on whether $\boldsymbol{\Phi}$ is treated as a *correlation* matrix or a *covariance* matrix. Thus, we choose treating $\boldsymbol{\Phi}$ as the latter, as a *covariance* matrix which does not have the restriction possessed by a correlation one is easier to deal with.

On the right side of (A.9.8) to be maximized over $\boldsymbol{\Phi}$, the term relevant to $\boldsymbol{\Phi}$ is $-(n/2)\left(\log|\boldsymbol{\Phi}| + \mathrm{tr}\boldsymbol{\Phi}^{-1}\mathbf{Q}\right)$. This is found to be equivalent to (A.5.6), if $\boldsymbol{\Phi}$ and $\mathbf{Q}$ are replaced by $\boldsymbol{\Sigma}$ and $\mathbf{V}$, respectively. Thus, (A.9.8) is maximized for

$$\boldsymbol{\Phi} = \mathbf{Q} \tag{A.9.28}$$

as (A.5.6) is maximized for (A.5.3). Here, it must be kept in mind that $\boldsymbol{\Phi}$ is treated as a *covariance* matrix: A factor correlation matrix and the corresponding loading matrix are given by (A.9.27) and (A.9.26), respectively. Thus, the matrices $\boldsymbol{\Phi}$ and $\mathbf{A}$ resulting in the EM algorithm must finally be *transformed* into the factor correlation matrix (A.9.27) and corresponding loading matrix (A.9.26).

A.9.7 Updating Loadings in M-Step for EFA and CFA

We consider updating the loading matrix $\mathbf{A}$ separately in *CFA* with the constraint on $\mathbf{A}$ and in *EFA* without the constraint.

First, let us consider maximizing (A.9.8) over $\mathbf{A}$ with $\boldsymbol{\Psi}$ fixed and $\boldsymbol{\Phi} = \mathbf{I}_m$ in *EFA*. The function (A.9.8) can be rewritten as $-(n/2)f(\mathbf{A}) + const_{[\mathrm{A}]}$ with $const_{[\mathrm{A}]}$ independent of $\mathbf{A}$ and

$$\begin{aligned} f(\mathbf{A}) &= \mathrm{tr}\mathbf{A}\mathbf{Q}\mathbf{A}'\boldsymbol{\Psi}^{-1} - 2\mathrm{tr}\mathbf{B}\mathbf{A}'\boldsymbol{\Psi}^{-1} \\ &= \left\|\boldsymbol{\Psi}^{-1/2}\mathbf{A}\mathbf{Q}^{1/2} - \boldsymbol{\Psi}^{-1/2}\mathbf{B}\mathbf{Q}^{-1/2}\right\|^2 - \left\|\boldsymbol{\Psi}^{-1/2}\mathbf{B}\mathbf{Q}^{-1/2}\right\|^2. \end{aligned} \tag{A.9.29}$$

Thus, the maximization of (A.9.8) amounts to minimizing (A.9.29). The minimizer is given by

$$\mathbf{A} = \mathbf{B}\mathbf{Q}^{-1}, \tag{A.9.30}$$

since only the term $\left\|\boldsymbol{\Psi}^{-1/2}\mathbf{A}\mathbf{Q}^{1/2} - \boldsymbol{\Psi}^{-1/2}\mathbf{B}\mathbf{Q}^{-1/2}\right\|^2$ is dependent on $\mathbf{A}$ on the right side of (A.9.29) and that term attains the lower limit zero for (A.9.30).

Next, let us consider the case of *CFA* subject to some elements in $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_p]'$ constrained to zeros. We can use a procedure of updating $\mathbf{A}$ *row-wise*, in which the maximization of (A.9.8) over $\mathbf{a}_j$ with the other parameters fixed is performed for $j = 1, \ldots, p$. This follows from the fact that (A.9.8) can be rewritten as $-(n/2)\sum_{j=1}^{p} f_j(\mathbf{a}_j)/\psi_j + const_{[j]}$ with $const_{[j]}$ independent of $\mathbf{a}_j$ and

$$f_j(\mathbf{a}_j) = \mathbf{a}_j'\mathbf{Q}\mathbf{a}_j - 2\mathbf{b}_j'\mathbf{a}_j = \mathbf{a}_j'\mathbf{H}_j'\mathbf{H}_j\mathbf{Q}\mathbf{H}_j'\mathbf{H}_j\mathbf{a}_j - 2\mathbf{b}_j'\mathbf{H}_j'\mathbf{H}_j\mathbf{a}_j. \tag{A.9.31}$$

Here, $\mathbf{H}_j$ is the $m_j \times m$ binary matrix satisfying $\mathbf{a}_j^{\#} = \mathbf{H}_j\mathbf{a}_j$, with m_j being the number of the unconstrained elements in $\mathbf{a}_j$, and $\mathbf{a}_j^{\#}$ being the $m_j \times 1$ vector obtained by deleting the constrained elements from $\mathbf{a}_j$: For example, if $\mathbf{a}_j = \begin{bmatrix} a_{j1} \\ 0 \\ a_{j3} \end{bmatrix}$ with the first and third loadings unconstrained, then $\mathbf{H}_j = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, which leads to the vector $\mathbf{a}_j^{\#} = \mathbf{H}_j\mathbf{a}_j = \begin{bmatrix} a_{j1} \\ a_{j3} \end{bmatrix}$ containing only unconstrained loadings. The last identity in (A.9.31) follows from

$$\mathbf{a}_j = \mathbf{H}_j'\mathbf{H}_j\mathbf{a}_j. \tag{A.9.32}$$

We can further rewrite (A.9.31) using $\mathbf{Q}_{\mathrm{H}} = \mathbf{H}_j \mathbf{Q} \mathbf{H}_j' \, (m_j \times m_j)$ as

$$f_j(\mathbf{A}) = \left\| \mathbf{Q}_{\mathrm{H}}^{1/2} \mathbf{H}_j \mathbf{a}_j - \mathbf{Q}_{\mathrm{H}}^{-1/2} \mathbf{H}_j \mathbf{b}_j \right\|^2 - \left\| \mathbf{Q}_{\mathrm{H}}^{-1/2} \mathbf{H}_j \mathbf{b}_j \right\|^2. \tag{A.9.33}$$

Here, the elements of $\mathbf{Q}_{\mathrm{H}}$ and $\mathbf{H}_j \mathbf{b}_j$ are restricted to the ones of $\mathbf{Q}$ and $\mathbf{b}_j$ corresponding to the unconstrained loadings, respectively. We can find that (A.9.33) is minimized for $\mathbf{H}_j \mathbf{a}_j = \mathbf{Q}_{\mathrm{H}}^{-1} \mathbf{H}_j \mathbf{b}_j$. Then, the vector $\mathbf{a}_j$ to be obtained is

$$\mathbf{a}_j = \mathbf{H}_j' \mathbf{Q}_{\mathrm{H}}^{-1} \mathbf{H}_j \mathbf{b}_j, \tag{A.9.34}$$

since of (A.9.32).

A.9.8 Whole Steps in EFA and CFA

The EM algorithm for *EFA* with $\boldsymbol{\Phi} = \mathbf{I}_m$ consists of the following steps:

Step 1. Initialize $\mathbf{A}$ and $\boldsymbol{\Psi}$.

Step 2. Obtain (A.9.18) and (A.9.19), with $\mathbf{H}(\boldsymbol{\Theta}) = (\mathbf{A}\mathbf{A}' + \boldsymbol{\Psi})^{-1}\mathbf{A}$ and $\mathbf{W}(\boldsymbol{\Theta}) = \left(\mathbf{I}_m + \mathbf{A}'\boldsymbol{\Psi}^{-1}\mathbf{A}\right)^{-1}$, i.e., the versions of (A.9.20) and (A.9.21) simplified using $\boldsymbol{\Phi} = \mathbf{I}_m$.

Step 3. Update $\mathbf{A}$ and the diagonal elements of $\boldsymbol{\Psi}$ through (A.9.30) and (A.9.24) respectively.

Step 4. Finish if the increase in the value of (A.9.1) with $\boldsymbol{\Phi} = \mathbf{I}_m$ from the previous round can be ignored; otherwise, go back to Step 2.

On the other hands, the EM algorithm for *CFA* consists of the following steps:

Step 1. Initialize $\mathbf{A}$, $\boldsymbol{\Phi}$, and $\boldsymbol{\Psi}$.

Step 2. Obtain (A.9.18) and (A.9.19) using (A.9.20) and (A.9.21)

Step 3. Update $\boldsymbol{\Phi}$, the diagonal elements of $\boldsymbol{\Psi}$, and each row of $\mathbf{A}$ through (A.9.28), (A.9.24), and (A.9.34), respectively.

Step 4. Go to Step 5 if the increase in the (A.9.1) value from the previous round can be ignored; otherwise, go back to Step 2.

Step 5. Finish with the loading and factor correlation matrices set to (A.9.26) and (A.9.27), respectively.

A.9.9 Algorithm for Penalized Factor Analysis

Let $g(\boldsymbol{\Theta})$ be a *penalty function* of $\boldsymbol{\Theta}$ penalizing particular values of parameters in $\boldsymbol{\Theta}$, with $g(\boldsymbol{\Theta})$ being independent of the random variables in $\mathbf{X}$ and $\mathbf{F}$. A *penalized FA* (*PFA*) can be formulated as maximizing (A.9.1) minus $g(\boldsymbol{\Theta})$, i.e.,

$$\begin{aligned}&\log P(\mathbf{X}|\boldsymbol{\Theta}) - g(\boldsymbol{\Theta}) \propto l^*(\boldsymbol{\Theta}) - g(\boldsymbol{\Theta})\\ &\quad = \tfrac{n}{2}\log\left|(\mathbf{A}\boldsymbol{\Phi}\mathbf{A}' + \boldsymbol{\Psi})^{-1}\mathbf{V}\right| - \tfrac{n}{2}\mathrm{tr}(\mathbf{A}\boldsymbol{\Phi}\mathbf{A}' + \boldsymbol{\Psi})^{-1}\mathbf{V} - g(\boldsymbol{\Theta}),\end{aligned} \tag{A.9.35}$$

over $\boldsymbol{\Theta}$. PFA includes the *penalized sparse latent variable FA* (*PS-LVFA*) treated in Chap. 22.

To show that the EM algorithm can also be used for PFA, we define

$$P^{\#}(\mathbf{X}|\boldsymbol{\Theta}) = P(\mathbf{X}|\boldsymbol{\Theta})\exp\{-g(\boldsymbol{\Theta})\}, \tag{A.9.36}$$

$$P^{\#}(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta}) = P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta})\exp\{-g(\boldsymbol{\Theta})\}. \tag{A.9.37}$$

PFA can be regarded as maximizing (A.9.36), since the logarithm of (A.9.36) equals the left side of (A.9.35), and a penalized version of (A.8.21), i.e.,

$$\begin{aligned}&\log P^{\#}(\mathbf{X}|\boldsymbol{\Theta}_{\text{new}}) - \log P^{\#}(\mathbf{X}|\boldsymbol{\Theta})\\ &\quad \geq E\left[\log P^{\#}(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta}_{\text{new}})|\mathbf{X}\right] - E\left[\log P^{\#}(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta})|\mathbf{X}\right],\end{aligned} \tag{A.9.38}$$

holds true, as shown in the next paragraph.

Let us substitute (A.9.36), (A.9.37), $P^{\#}(\mathbf{X}|\boldsymbol{\Theta}_{\text{new}})$, and $P^{\#}(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta}_{\text{new}})$ for $P(\mathbf{X}|\boldsymbol{\Theta})$, $P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta})$, $P(\mathbf{X}|\boldsymbol{\Theta}_{\text{new}})$, and $P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta}_{\text{new}})$, and $P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta}_{\text{new}})$ on the left sides of (A.8.19) and (A.8.20). Then, we have the following equations:

$$\begin{aligned}\log E\left[\frac{P^{\#}(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta}_{\text{new}})}{P^{\#}(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta})}|\mathbf{X}\right] &= \log E\left[\frac{P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta}_{\text{new}})\exp\{-g(\boldsymbol{\Theta}_{\text{new}})\}}{P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta})\exp\{-g(\boldsymbol{\Theta})\}}|\mathbf{X}\right]\\ &= \log\left\{E\left[\frac{P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta}_{\text{new}})}{P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta})}|\mathbf{X}\right] \times \frac{\exp\{-g(\boldsymbol{\Theta}_{\text{new}})\}}{\exp\{-g(\boldsymbol{\Theta})\}}\right\}\\ &= \log E\left[\frac{P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta}_{\text{new}})}{P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta})}|\mathbf{X}\right] + \log\frac{\exp\{-g(\boldsymbol{\Theta}_{\text{new}})\}}{\exp\{-g(\boldsymbol{\Theta})\}}\\ &= \log P(\mathbf{X}|\boldsymbol{\Theta}_{\text{new}}) - \log P(\mathbf{X}|\boldsymbol{\Theta}) + \log\frac{\exp\{-g(\boldsymbol{\Theta}_{\text{new}})\}}{\exp\{-g(\boldsymbol{\Theta})\}}\\ &= \log P^{\#}(\mathbf{X}|\boldsymbol{\Theta}_{\text{new}}) - \log P^{\#}(\mathbf{X}|\boldsymbol{\Theta}),\end{aligned} \tag{A.9.39}$$

$$E\left[\log\frac{P^{\#}(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta}_{\text{new}})}{P^{\#}(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta})}|\mathbf{X}\right]=E\left[\log P^{\#}(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta}_{\text{new}})|\mathbf{X}\right]-E\left[\log P^{\#}(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta})|\mathbf{X}\right]. \tag{A.9.40}$$

Here, the fourth identity in (A.9.39) is derived using (A.8.19). Theorem A.8.1 shows (A.9.39) $\geq$ (A.9.40), which implies (A.9.38).

As found from the fact that (A.8.21) leads to the EM algorithm with (A.8.22) and (A.8.23), the inequality in (A.9.38) implies that the EM algorithm, in which $P(\mathbf{X}|\boldsymbol{\Theta})$ and $P(\mathbf{X}, \mathbf{F}|\boldsymbol{\Theta})$ are replaced by (A.9.36) and (A.9.37), respectively, can be used for PFA. That is, (A.9.35) can reach its maximum, by iterating the penalized versions of (A.8.22) and (A.8.23),

$$\textit{E-step}\text{: Obtain } E\left[\log P^{\#}(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta})|\mathbf{X}\right], \tag{A.9.41}$$

$$\begin{aligned}\textit{M-step}:\ &\text{Update } \boldsymbol{\Theta} \text{ so as to increase } E\left[\log P^{\#}(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta})|\mathrm{X}\right]\\ &\text{i.e., lead to } E\left[\log P^{\#}(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta}_{\text{new}})\right]\geq E\left[\log P^{\#}(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta})\right].\end{aligned} \tag{A.9.42}$$

Here,

$$\begin{aligned}E[\log P^{*}(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta})|\mathbf{X}]&=E[\log P(\mathbf{X},\mathbf{F}|\boldsymbol{\Theta})|\mathbf{X}]-g(\boldsymbol{\Theta})\\&=-\frac{n}{2}\log|\boldsymbol{\Psi}|-\frac{n}{2}\mathrm{tr}(\mathbf{V}-2\mathbf{B}\mathbf{A}'+\mathbf{A}\mathbf{Q}\mathbf{A}')\boldsymbol{\Psi}^{-1}\\&\quad-\frac{n}{2}\left(\log|\boldsymbol{\Phi}|+\mathrm{tr}\boldsymbol{\Phi}^{-1}\mathbf{Q}\right)+c-g(\boldsymbol{\Theta}):\end{aligned} \tag{A.9.43}$$

(A.9.8) minus $g(\boldsymbol{\Theta})$.

The procedures for the *E-step* (A.9.41) are the same as in Sect. A.9.4. On the other hand, how $\boldsymbol{\Theta}$ is updated in the *M-step* (A.9.42) differs in general from standard FA without penalty function. Care must be taken that $\boldsymbol{\Phi}$ cannot be treated as a *covariance* matrix as in Sect. A.9.6, *unless* (A.9.35) can be rewritten as

$$l^{*}(\boldsymbol{\Theta})-g(\boldsymbol{\Theta})=\tfrac{n}{2}\log\left|\left(\mathbf{A}_{\mathrm{R}}\boldsymbol{\Phi}_{\mathrm{R}}\mathbf{A}_{\mathrm{R}}'+\boldsymbol{\Psi}\right)^{-1}\mathbf{V}\right|-\tfrac{n}{2}\mathrm{tr}\left(\mathbf{A}_{\mathrm{R}}\boldsymbol{\Phi}_{\mathrm{R}}\mathbf{A}_{\mathrm{R}}'+\boldsymbol{\Psi}\right)^{-1}\mathbf{V}-g(\boldsymbol{\Theta}_{\mathrm{R}}). \tag{A.9.44}$$

with (A.9.26), (A.9.27), and $\boldsymbol{\Theta}_{\mathrm{R}} = \{\mathbf{A}_{\mathrm{R}}, \boldsymbol{\Phi}_{\mathrm{R}}, \boldsymbol{\Psi}\}$. PS-LVFA in Chap. 22 is one case where (A.9.44) does not hold true. Thus, $\boldsymbol{\Phi}$ must be *constrained to a correlation matrix* in Chap. 22. However, the PS-LVFA procedure for updating $\boldsymbol{\Psi}$ in (A.9.42) is the same as in A.9.5.

References

Absil, P.-A., Mahony, R., & Sepulchre, R. (2008). *Optimization algorithms on matrix manifolds*. Princeton, NJ: Princeton University Press.

Adachi, K. (2004). Correct classification rates in multiple correspondence analysis. *Journal of the Japanese Society of Computational Statistics, 17,* 1–20.

Adachi, K. (2006). *Multivariate data analysis: An introduction for psychology, pedagogy, and sociology courses*. Nakanishiya-Shuppan (in Japanese): Kyoto.

Adachi, K. (2009). Joint Procrustes analysis for simultaneous nonsingular transformation of component score and loading matrices. *Psychometrika, 74,* 667–683.

Adachi, K. (2011). Constrained principal component analysis of standardized data for biplots with unit-length variable vectors. *Advances in Data Analysis and Classification, 5,* 23–36.

Adachi, K. (2012). Some contributions to data-fitting factor analysis with empirical comparisons to covariance-fitting factor analysis. *Journal of the Japanese Society of Computational Statistics, 25,* 25–38.

Adachi, K. (2013). Factor analysis with EM algorithm never gives improper solutions when sample covariance and initial parameter matrices are proper. *Psychometrika, 78,* 380–394.

Adachi, K. (2014). Sparse path analysis: Computational identification of causality between explanatory and dependent variables. In *Proceedings of the 28th Symposium of the Japanese Society of Computational Statistics* (pp. 223–226).

Adachi, K. (2016). Three-way principal component analysis with its applications to psychology. In T. Sakata (Ed.), *Applied matrix and tensor variate data analysis* (pp. 1–21). Tokyo: Springer.

Adachi, K. (2019). Factor analysis: Latent variable, matrix decomposition, and constrained uniqueness formulations. In *WIREs computational statistics*. https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.1458.

Adachi, K., & Murakami, T. (2011). *Nonmetric multivariate analysis: From principal component analysis to multiple correspondence analysis*. Tokyo: Asakura-Shoten. (in Japanese).

Adachi, K., & Trendafilov, N.T. (2015). Sparse orthogonal factor analysis. In E. Carpita, E. Brentari, & Qannari, E. M. (Eds.), *Advances in latent variable: Methods, models, and Applications* (pp. 227–239). Cham, Switzerland: Springer.

Adachi, K., & Trendafilov, N. T. (2016). Sparse principal component analysis subject to prespecified cardinality of loadings. *Computational Statistics, 31,* 1403–1427.

Adachi, K., & Trendafilov, N. T. (2018a). Some mathematical properties of the matrix decomposition solution in factor analysis. *Psychometrika, 83,* 407–424.

Adachi, K., & Trendafilov, N. T. (2018b). Sparsest factor analysis for clustering variables: A matrix decomposition approach. *Advances in Data Analysis and Classification, 12,* 559–585.

K. Adachi, *Matrix-Based Introduction to Multivariate Data Analysis*,
https://doi.org/10.1007/978-981-15-4103-2

Adachi, K., & Trendafilov, N. T. (2019). Some inequalities contrasting principal component and factor analyses solutions. *Japanese Journal of Statistics and Data Science, 2,* 31–47.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19,* 716–723.

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis* (3rd ed.). New York: Wiley.

Anderson, T. W., & Rubin, H. (1956). Statistical inference in factor analysis. In J. Neyman (Ed.), *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 5, pp. 111–150). Berkeley, CA: University of California Press.

Banerjee, S., & Roy, A. (2014). *Linear algebra and matrix analysis for statistics*. Boca Raton, FL: CRC Press.

Bartholomew, D., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis*: *A unified approach* (3rd ed.). Chichester: West Sussex: Wiley.

Bentler, P. M. (1985). *Theory and implementation of EQS: A structural equation program* (manual for program version 2.0). Los Angels: BMDP Statistical Software.

Benzécri, J.-P. (1992). *Correspondence analysis handbook*. New York: Marcel Dekker.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Borg, I., & Groenen, P. J. F. (2005). *Modern multidimensional scaling: Theory and applications* (2nd ed.). New York: Springer.

Borg, I., Groenen, P. J. F., & Mair, P. (2013). *Applied multidimensional scaling*. Heidelberg: Springer.

Boyed, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge: Cambridge University Press.

Browne, M. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research, 36,* 111–150.

Bühlmann, P., & van de Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory and applications*. Berlin: Springer.

Carroll, J. D., Green, P. E., & Chaturvedi, A. (1997). *Mathematical tools for applied multivariate analysis* (rev. ed.). San Diego, California: Academic.

Committee for Guiding Psychological Experiments. (1985). *Experiments and tests*: *The foundations of psychology* (Explanation version). Tokyo: Bifukan. (in Japanese).

Cox, T. F., & Cox, M. A. A. (2000). *Multidimensional scaling* (2nd ed.). London: Chapman & Hall.

de Leeuw, J. (1977). Application of convex analysis to multidimensional scaling. In J. R. Barra, F. Brodeau, G. Rominer, & B. Van Custem (Eds.), *Recent developments in statistics* (pp. 133–145). Amsterdam: North-Holland.

de Leeuw, J. (2004). Least squares optimal scaling of partially observed linear systems. In K. van Montfort, J. Oud, & A. Satorra (Eds.), *Recent developments of structural equation models: Theory and applications* (pp. 121–134). Dordrecht: Kluwer Academic Publishers.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39,* 1–38.

De Soete, G., & Carroll, J. D. (1994). K-means clustering in a low-dimensional Euclidean space. In E. Diday, Y. Lechevallier, M. Schader., P. Bertrand, & B. Burtschy (Eds.), *New approaches in classification and data analysis* (pp. 212–219). Berlin: Springer.

Eckart, C., & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika, 1,* 211–218.

Eldén, L. (2007). *Matrix methods in data mining and pattern recognition*. PA, Philadelphia: Society of Industrial and Applied Mathematics (SIAM).

Everitt, B. S. (1993). *Cluster analysis* (3rd ed.). London: Edward Arnold.

Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. (2013). *Regression: Models, methods, and applications*. Heidelberg: Springer.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association, 96,* 1348–1360.
Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 7,* 179–188.
Gan, G., Ma, C., & Wu, J. (2007) *Data clustering: Theory, algorithms, and applications*. Philadelphia, PA: Society of Industrial and Applied Mathematics (SIAM).
Gifi, A. (1990). *Nonlinear multivariate analysis*. Chichester: Wiley.
Gentle, J. E. (2017). *Matrix algebra: Theory, computations and applications in statistics* (2nd ed.). New York: Springer.
Golub, G. H., & Van Loan, C. F. (2013). *Matrix computations* (3rd ed.). Baltimore: Johns Hopkins University Press.
Gower, J. C., Lubbe, S., & le Roux, N. (2011). *Understanding biplots*. Chichester: Wiley.
Gower, J. C., & Dijksterhuis, G. B. (2004). *Procrustes problems*. Oxford: Oxford University Press.
Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*. London: Academic.
Greenacre, M. J. (2007). *Correspondence analysis in practice* (2nd ed.). Boca Raton, FL: CRC Press/Taylor & Francis Group.
Groenen, P. J. F. (1993). *The majorization approach to multidimensional scaling: Some problems and extensions*. Leiden, The Netherlands: DSWO.
Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrika, 11,* 1–12.
Hand, D. J. (1997). *Construction and assessment of classification rules*. Chichester: Wiley.
Hansen, P. C., Pereyra, V., & Scherer, G. (2013). *Least square data fitting with applications*. Baltimore, MD: The John Hopkins University Press.
Harman, H. H. (1976). *Modern factor analysis* (3rd ed.). Chicago: The University of Chicago Press.
Harman, H. H., & Jones, W. H. (1966). Factor analysis by minimizing residuals (Minres). *Psychomerika, 31,* 351–369.
Hartigan, J. A., & Wang, M. A. (1979). A k-means clustering algorithm. *Applied Statistics, 28,* 100–108.
Harville, D. A. (1997). *Matrix algebla from a statistician's perspective*. New York: Springer.
Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York: Springer.
Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: The lasso and generalizations*. Boca Raton, FL: CRC Press/Taylor & Francis Group.
Harshman, R.A. (1970). Foundations of the PARAFAC procedure: Models and conditions for an "exploratory" multi-mode factor analysis. *UCLA Working Papers in Phonetics*, *16*, 1–84.
Hayashi, C. (1952). On the prediction of phenomena from qualitative data and the quantification of qualitative data from mathematico-statistical point of view. *Annals of the Institute of Mathematical Statistics, 3,* 69–98.
Heiser, W. J. (1991). A generalized majorization method for least squares multidimensional scaling of pseudo distances that may be negative. *Psychometrika, 56,* 7–27.
Hempel, C. (1966). *Philosophy in natural science*. Englewood Cliffs, New Jersey: Prentice Hall.
Hitchcock, F. L. (1927). Multiple invariants and generalized rank of a p-way matrix or tensor. *Journal of Mathematics and Physics, 7,* 39–79.
Hirose, K., Ogura, Y., & Shimodaira, H. (2015). Estimating scale-free networks via the exponentiation of minimax concave penalty. *Journal of the Japanese Society of Computational Statistics, 28,* 139–154.
Hirose, K., & Yamamoto, M. (2014). Estimation of an oblique structure via penalized likelihood factor analysis. *Computational Statistics and Data Analysis, 79,* 120–132.
Hirose, K., & Yamamoto, M. (2015). Sparse estimation via nonconcave penalized likelihood in factor analysis model. *Statistics and Computing, 25,* 863–875.

Hoerl, A. E., & Kennard, R. (1970). Ridge regression: Biased estimation for non-orthogonal roblems. *Technometrics*, *12*, 55-67. (Reprinted from *Technometrics*, *42* (2000), 80–86).

Hogg, R. V., McKean, J. W., & Craig, A. T. (2019). *Introduction to mathematical statistics* (8th ed.). Boston, MA: Pearson.

Holzinger, K. J., & Swineford, F. (1939). *Supplementary Educational Monographs No. 48. A study in factor analysis: The stability of a bi-factor solution*, University of Chicago

Horn, R. A., & Johonson, C. R. (2013). *Matrix analysis* (2nd ed.). Cambridge: Cambridge University Press.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Statistics, 24,* 417–441.

Hotelling, H. (1936). Relations between sets of variables. *Biometrika, 28,* 321–377.

Huley, J. R., & Cattell, R. B. (1962). The Procrustes program: Producing direct rotation to test a hypothesized factor structure. *Behavioral Science, 7,* 258–262.

Ihara, M., & Kano, Y. (1986). A new estimator of the uniqueness in factor analysis. *Psychometrika, 51,* 563–566.

Ikemoto, H., & Adachi, K. (2016). Sparse Tucker2 analysis of three-way data subject to a constrained number of zero elements in a core array. *Computational Statistics and Data Analysis, 98,* 1–18.

Izenman, A. J. (2008). *Modern multivariate statistical techniques: Regression, classification, and manifold learning*. New York, NY: Springer.

Jennrich, R. I. (2001). A simple general method for orthogonal rotation. *Psychometrika, 66,* 289–306.

Jennrich, R. I. (2002). A simple general method for oblique rotation. *Psychometrika, 67,* 7–20.

Jennrich, R. I., & Sampson, P. F. (1966). Rotation for simple loadings. *Psychometrika, 31,* 313–323.

Jin, S., Moustaki, I., & Yang-Wallentin, F. (2018). Approximated penalized maximum likelihood for exploratory factor analysis: An orthogonal case. *Psychometrika, 83,* 628–649.

Jolliffe, I. T. (2002). *Principal component analysis* (2nd ed.). New York: Springer.

Jolliffe, I. T., Trendafilov, N. T., & Uddin, M. (2003). A modified principal component technique based on the LASSO. *Journal of Computational and Graphical Statistics, 12,* 531–547.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34,* 183–202.

Jöreskog, K. G. (1970). A general method for analysis of covariance structures. *Biometrika, 57,* 239–251.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika, 23,* 187–200.

Kano, Y. (1990). Noniterative estimation and the choice of the number of factors in exploratory factor analysis. *Psychometrika, 55,* 277–291.

Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions*. Thousand Oaks, California: Sage Publications.

Kettenring, J. R. (1971). Canonical analysis of several sets of variables. *Biometrika, 58,* 433–460.

Khuri, A. I. (2003). *Advanced calculus with applications in statistics* (2nd ed.). Hoboken, New Jersey: Wiley.

Kiers, H. A. L. (1994). Simplimax: Oblique rotation to an optimal target with simple structure. *Psychometrika, 59,* 567–579.

Kiers, H. A. L. (1998a). Three-way SIMPLIMAX for oblique rotation of the three-mode factor analysis core to simple structure. *Computational Statistics and Data Analysis, 28,* 307–324.

Kiers, H. A. L. (1998b). Joint orthomax rotation of the core and component matrices resulting from three-mode principal component analysis. *Journal of Classification, 15,* 245–263.

Kiers, H. A. L. (2002). Setting up alternating least squares and iterative majorization algorithms for solving various matrix optimization problems. *Computational Statistics and Data Analysis, 41,* 157–170.

Koch, I. (2014). *Analysis of multivariate and high-dimensional data*. Cambridge: Cambridge University Press.
Konishi, S. (2014). *Introduction to multivariate analysis: Linear and nonlinear modeling*. Boca Raton, FL: CRC Press.
Konishi, S., & Kitagawa, G. (2007). *Information criteria and statistical modeling*. New York: Springer.
Kroonenberg, P. M. (2008). *Applied multiway data analysis*. Hoboken: Wiley.
Lange, K. (2010). *Numerical analysis for statisticians* (2nd ed.). New York: Springer.
Lattin, J., Carroll, J. D., & Green, P. E. (2003). *Analyzing multivariate data*. Pacific Grove: CA, Thomson Learning Inc.
Lütkepohl, H. (1996). *Handbook of matrices*. Chichester: Wiley.
McLachlan, G. J., & Krishnan, T. (2008). *The EM algorithm and extensions* (2nd ed.). New York: Wiley.
McLachlan, G. J. (1992). *Discriminant analysis and statistical pattern recognition*. New York: Wiley.
MacQueen, J.B. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium*, *1*, 281–297.
Magnus, J. R., & Neudecker, H. (2019). *Matrix differential calculus with an applications in statistics and econometrics* (3rd ed.). Chichester: Wily.
Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to regression analysis* (5th ed.). Hoboken, New Jersey: Wiley.
Mosier, C. I. (1939). Determining a simple structure when loadings for certain tests are known. *Psychometrika, 4,* 149–162.
Mulaik, S. A. (2010). *Foundations of factor analysis* (2nd ed.). Boca Raton: CRC Press.
Nishisato, S. (1980). *Analysis of categorical data: Dual scaling and its applications*. Toronto: University of Toronto Press.
Ogasawara, H. (2000). Some relationships between factors and components. *Psychometrika, 65,* 167–185.
Ortega, J. M., & Rheinboldt, W. C. (2000). *Iterative solution of nonlinear equations in several variables*. Philadelphia, PA: Society of Industrial and Applied Mathematics (SIAM).
Osgood, C. E., & Luria, Z. (1954). A blind analysis of a case of multiple personality. *Journal of Abnormal and Social Psychology, 49,* 579–591.
Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazines, 2,* 559–572.
Rao, C. R. (1973). *Linear statistical inference and its applications*. New York: Wiley.
Rao, C. R. (2001). *Linear statistical inference and its applications* (2nd ed.). New York: Wiley.
Rao, C. R., & Mitra, S. K. (1971). *Generalized inverse of matrices and its applications*. New York: Wiley.
Rao, C. R., & Rao, M. B. (1998). *Matrix algebra and its applications to statistics and econometrics*. Singapore: World Scientific Publishing.
Reyment, R., & Jöreskog, K. G. (1996). *Applied factor analysis in the natural sciences*. Cambridge: Cambridge University Press.
Rencher, A. C., & Christensen, W. F. (2012). *Methods of multivariate analysis* (3rd ed.). Hoboken, New Jersey: Wiley.
Rubin, D. B., & Thayer, D. T. (1982). EM algorithms for ML factor analysis. *Psychometrika, 47,* 69–76.
Schneeweiss, H., & Mathes, H. (1995). Factor analysis and principal components. *Journal of Multivariate Analysis, 55,* 105–124.
Schott, J. R. (2005). *Matrix analysis for statistics* (2nd ed.). Hoboken, New Jersey: Wiley.
Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461–464.
Searle, R. S., & Khuri, A. I. (2017). *Matrix algebra useful for statistics* (2nd ed.). Hoboken, New Jersey: Wiley.
Seber, G. A. F. (1984). *Multivariate observations*. New York: Wiley.

Seber, G. A. F. (2008). *A matrix handbook for statisticians*. Hoboken, New Jersey: Wiley.
Shen, H., & Huang, J. Z. (2008). Sparse principal component analysis via regularized lower rank approximation. *Journal of Multivariate Analysis, 99,* 1015–1034.
Shimodaira, H. (2016). Cross-validation of matching correlation analysis by resampling matching weights. *Neural Networks, 75,* 126–140.
Smilde, A., Bro, R., & Geladi, P. (2004). *Multi-way analysis: Applications in the chemical sciences*. Chichester: Wiley.
Sočan, G. (2003). *The incremental value of minimum rank factor analysis*. Ph.D. thesis, University of Groningen, Groningen.
Spearman, C. (1904). "General intelligence" objectively determined and measured. *American Journal of Psychology, 15,* 201–293.
Stegeman, A. (2016). A new method for simultaneous estimation of the factor model parameters, factor scores, and unique parts. *Computational Statistics and Data Analysis, 99,* 189–203.
Takane, Y. (2014). *Constrained principal component analysis and related techniques*. Boca Raton, FL: CRC Press.
Takane, Y., Young, F. W., & de Leeuw, J. (1977). Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika, 42,* 7–67.
Tanaka, Y., & Tarumi, T. (1995). *Handbook for statistical analysis: Multivariate analysis (windows version)*. Tokyo: Kyoritsu-Shuppan. (in Japanese).
ten Berge, J. M. F. (1983). A generalization of Kristof's theorem on the trace of certain matrix products. *Psychometrika, 48,* 519–523.
ten Berge, J. M. F. (1993). *Least squares optimization in multivariate analysis*. Leiden: DSWO Press.
ten Berge, J. M. F., & Kiers, H. A. L. (1996). Optimality criteria for principal component analysis and generalizations. *British Journal of Mathematical and Statistical Psychology, 49,* 335–345.
ten Berge, J. M. F., Knol, D. L., & Kiers, H. A. L. (1988). A treatment of the orthomax rotation family in terms of diagonalization, and a re-examination of a singular value approach to varimax rotation. *Computational Statistics Quarterly, 3,* 207–217.
Thurstone, L. L. (1935). *The vectors of mind*. Chicago: University of Chicago Press.
Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago: University of Chicago Press.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 58,* 267–288.
Timm, N. H. (2002). *Applied multivariate analysis*. New York: Springer.
Tipping, M. E., & Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 61,* 611–622.
Torgerson, W. S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika, 17,* 401–419.
Toyoda, H. (1988). *Covariance structure analysis (structural equation modeling): Introductory part*. Tokyo: Asakura-Shoten. (in Japanese).
Trendafilov, N. T. (2014). From simple structure to sparse components: A review. *Computational Statistics, 29,* 431–454.
Trendafilov, N. T., & Adachi (2015). Sparse versus simple structure loadings. *Psychometrika*, *80,* 776–790.
Trendafilov, N. T., Fontanella, S., & Adachi, K. (2017). Sparse exploratory factor analysis. *Psychometrika, 82,* 778–794.
Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika, 31,* 279–311.
Uno, K., Adachi, K., & Trendafilov, N. T. (2019). Clustered common factor exploration in factor analysis. *Psychometrika, 84,* 1048-1067.
van de Geer, J. P. (1984). Linear relations among *k* sets of variables. *Psychometrika, 49,* 79–94.

Velicer, W. F., & Jackson, D. N. (1990). Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure. *Multivariate Behavioral Research, 25,* 1–28.

Vichi, M. (2017). Disjoint factor analysis with cross-loadings. *Advances in Data Analysis and Classification, 11,* 563–591.

Vichi, M., & Kiers, H. A. L. (2001). Factorial k -means analysis for two-way data. *Computational Statistics and Data Analysis, 37,* 49–64.

Wright, S. (1918). On the nature of size factors. *Genetics, 3,* 367–374.

Unkel, S., & Trendafilov, N. T. (2010). Simultaneous parameter estimation in exploratory factor analysis: An expository review. *International Statistical Review, 78,* 363–382.

Wang, J., & Wang, X. (2012). *Structural equation modeling: Applications using Mplus*. Chichester, West Sussex: Wiley.

Wright, S. (1960). Path coefficients and path regressions: Alternative or complementary cocepts? *Biometrics, 16,* 189–202.

Yanai, H., & Ichikawa, M. (2007). Factor analysis. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 257–296). Amsterdam: Elsevier.

Yanai, H., Takeuchi, K., & Takane, Y. (2011). *Projection matrices, generalized inverse matrices, and singular value decomposition*. New York: Springer.

Yates, A. (1987). *Multivariate exploratory data analysis: A perspective on exploratory factor analysis*. Albany: State University of New York Press.

Yeung, K. Y., & Ruzzo, W. L. (2001). Principal component analysis for clustering gene expression data. *Bioinformatics, 17,* 763–774.

Young, F. W. (1981). Quantitative analysis of qualitative data. *Psychometrika, 46,* 357–388.

Zhang, C. (2010). Nearly unbiased variable selection under minimax cocave penalty. *Annals of Statistics, 38,* 894–942.

Zou, D. M., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics, 15,* 265–286.

Zou, H. (2006). The adaptive lasso and its oracle property. *Journal of the American Statistical Association, 101,* 1418–1429.

Zou, H., Hastie, T., & Tibshirani, R. (2007). On the degree of freedom of the lasso. *Annals of Statistics, 35,* 2173–2192.

Index

K. Adachi, *Matrix-Based Introduction to Multivariate Data Analysis*,
https://doi.org/10.1007/978-981-15-4103-2