# Factor Analysis of Data Matrices: New Theoretical and Computational Aspects With Applications

Thesis

UNRESTRICTED

# Factor Analysis of Data Matrices: New Theoretical and Computational Aspects with Applications

A thesis
submitted to The Open University
in fulfilment of the requirements for the degree of
Doctor of Philosophy (PhD)

Department of Mathematics and Statistics
Faculty of Mathematics, Computing & Technology
The Open University
Walton Hall, Milton Keynes, MK7 6AA
United Kingdom

by

Steffen Unkel,
MSc in Statistics, Diplom-Volkswirt, Diplom-Kaufmann

November 2009

DATE OF SUBMISSION: 1 SEPT 2009

DATE OF AWARD: 17 NOV 2009

# Abstract

The classical fitting problem in exploratory factor analysis (EFA) is to find estimates for the factor loadings matrix and the matrix of unique factor variances which give the best fit to the sample covariance or correlation matrix with respect to some goodness-of-fit criterion. Predicted factor scores can be obtained as a function of these estimates and the data. In this thesis, the EFA model is considered as a specific data matrix decomposition with fixed unknown matrix parameters. Fitting the EFA model directly to the data yields simultaneous solutions for both loadings and factor scores. Several new algorithms are introduced for the least squares and weighted least squares estimation of all EFA model unknowns. The numerical procedures are based on the singular value decomposition, facilitate the estimation of both common and unique factor scores, and work equally well when the number of variables exceeds the number of available observations.

Like EFA, noisy independent component analysis (ICA) is a technique for reduction of the data dimensionality in which the interrelationships among the observed variables are explained in terms of a much smaller number of latent factors. The key difference between EFA and noisy ICA is that in the latter model the common factors are assumed to be both independent and non-normal. In contrast to EFA, there is no rotational indeterminacy in noisy ICA. In this thesis, noisy ICA is viewed as a method of factor rotation in EFA. Starting from an initial EFA solution, an orthogonal rotation matrix is sought that minimizes the dependence between the common factors. The idea of rotating the scores towards independence is also employed in three-mode factor analysis to analyze data sets having a three-way structure.

The new theoretical and computational aspects contained in this thesis are illustrated by means of several examples with real and artificial data.

Dedicated solely to my parents.

*"A work of this kind is never really finished; one only calls it finished because one has done all that is possible in the time and the circumstances."*

Johann Wolfgang von Goethe (Italian Journey: Caserta, 16th of March 1787)

# Acknowledgements

This thesis is the result of three years of work in which I have been accompanied and supported by various people. I am taking this opportunity to thank all those who assisted me in one way or another.

First and foremeost, I would like to express my deep and sincere gratitude to Dr. Nickolay T. Trendafilov and Professor Dr. C. Paddy Farrington for giving me the opportunity to study towards a doctoral degree at the Open University.

Dr. Nickolay T. Trendafilov has been my principal supervisor. His logical reasoning, his patience of a saint and his unique sense of Bulgarian humour have been of great value to me throughout my studies.

I am very grateful to Professor Dr. M. Chris Jones for proofreading this manuscript prior to submission, although of course he cannot be held responsible for any errors in the final version.

Special thanks are due to Professor Dr. Paul H. Garthwaite and Professor Dr. Frank Critchley for helpful comments on two papers which contain material presented herein.

It is most appreciated that Dr. S. Karen Vines and Professor John C. Gower were willing to assess my probation report after the first year of study.

I am very pleased that Professor John C. Gower and Professor Dr. Wojtek J. Krzanowski were willing to examine this thesis. Professor Dr. Kevin J. McConway kindly agreed

to chair the examination panel in the viva voce.

I would like to thank all members of staff in the Department of Mathematics and Statistics for providing an excellent and inspiring working atmosphere. In particular, Dr. Karim A. Anaya-Izquierdo was always available when I needed support or encouragement.

Finally, I am forever indebted to my parents Marlies and Wilhelm Unkel for their love.

# Contents

# List of Figures

# List of Tables

# Part I

# Setting the Scene

# Chapter 1

# Introduction and Preliminaries

## 1.1 Motivation

Multivariate data are often viewed as indirect measurements arising from underlying sources or latent variables which cannot be directly measured. One is forced to examine the hidden sources by collecting data on manifest variables which are considered indicators of the concepts of real interest (e.g., Bartholomew, Steele, Moustaki, and Galbraith, 2002).

Consider the following example (Stone, 2004): Electroencephalogram brain scans measure the neuronal activity in various parts of the brain indirectly via electromagnetic signals recorded at sensors placed at different positions on the head. Since each signal contains contributions from many different brain regions, observed signals are a mixture of the hidden sources. The aim is to extract such sources and to provide information on which parts of the brain are activated at a given time or by a given task.

Exploratory factor analysis (EFA) is a statistical model which addresses this issue of extracting the sources underlying a set of measured signal mixtures. Some key references are Bartholomew and Knott (1999), Harman (1976), Lawley and Maxwell (1971) and Mulaik (1972). The model of EFA aims to explain the interrelationships among $p$ manifest variables by $k$ ($\ll p$) latent variables called common factors. To allow for some variation in each observed variable that remains unaccounted for by the common factors, $p$ additional latent variables called unique factors are introduced, each of which accounts for the unique variance in its associated manifest variable.

The classical fitting problem in EFA is to find estimates for the factor loadings matrix and the matrix of unique factor variances which give the best fit, for some specified value of $k$, to the sample correlation matrix with respect to some goodness-of-fit criterion. One may then construct factor scores for the $n$ observations on the $k$ common factors as a function of these estimates and the data.

In this thesis, new approaches for fitting the EFA model are presented. Without passing via an estimate for the model correlation matrix, the EFA model is fitted directly to the data. That is, the EFA model is considered as a specific data matrix decomposition with fixed unknown matrix parameters. Unlike the factorization of a correlation matrix, fitting the EFA model to the data yields factor loadings and common factor scores simultaneously (Horst, 1965; Jöreskog, 1962; Lawley, 1942; McDonald, 1979; Whittle, 1952; Young, 1941).

De Leeuw (2004, 2008) proposed simultaneous estimation of all EFA model unknowns by optimizing a least squares (LS) loss function. The algorithms of De Leeuw (2004, 2008) are further developed in this thesis. However, these approaches are designed for the classical case of 'vertical' data matrices with $n > p$. In a number of modern applications, the number of available observations is less than the number of variables, such as for example in microarray genomic analysis or in atmospheric science. This thesis introduces a couple of novel methods for the simultaneous estimation of all EFA model unknowns which are able to fit the EFA model to 'horizontal' data matrices with $p \geq n$. New assumptions are imposed on the EFA model parameters which necessarily require the acceptance of unique factors having zero variance.

Principal component analysis (PCA) (e.g., Jolliffe, 2002) is a descriptive statistical technique that replaces a set of $p$ observed variables by $k$ ($\ll p$) uncorrelated variables called principal components whilst retaining as much as possible of the total sample variance. Despite the differences between PCA and EFA (e.g., Jolliffe, 2002, pp. 158-161), both methods aim to reduce the dimensionality of a data set. It is of interest to find conditions under which PCA and EFA solutions can or cannot be close for a particular data set (Rao, 1996). Therefore, in this thesis PCA is viewed as a special case of EFA with the error term resembling the EFA one. Based on an initial PCA solution, the error term is then decomposed to achieve an EFA-like factorization of the data. This specific EFA-like PCA construction helps to compare the numerical solutions obtained by PCA and EFA. A new approach to accomplish PCA by means of the QR factorization of the data matrix (e.g., Golub and Van Loan, 1996) is introduced.

Classical EFA techniques taking input data in the form of correlations are very vulnerable to the presence of outliers. One may either use some robust modification of the sample correlation matrix to overcome the outlier problem in the data (e.g., Pison, Rousseeuw, Filzmoser, and Croux, 2003) or look for alternative techniques working directly with the data matrix. Croux, Filzmoser, Pison, and Rousseeuw (2003) proposed robust factorization of the data matrix into a loadings matrix and a matrix of factor scores by optimizing a resistant alternating regression scheme. However, since estimates of the unique factor variances are obtained after the corresponding loss function has already been optimized, the approach of Croux, Filzmoser, Pison, and Rousseeuw (2003) resembles a robust PCA solution to EFA rather than 'truly' robust EFA.

In this thesis, an algorithm for robust simultaneous estimation of all EFA model unknowns is introduced. The EFA model is fitted to the data matrix by minimizing a certain weighted least squares (WLS) goodness-of-fit measure. By imposing weights on the residuals of the unweighted least squares (ULS) fitting, the WLS loss function is a generalization of the one used by De Leeuw (2004, 2008). Kiers (1997b) introduced a very general approach for fitting a model to a data matrix by WLS. The WLS fitting problem is reduced to iteratively solving a corresponding ULS problem, by using a majorization approach (e.g., Heiser, 1995). In this thesis, the majorizing function of Kiers (1997b) is used in a procedure for iteratively reweighted least squares in which the weights depend on the residuals and are updated after each cycle of updating the model parameters. The influence of large residuals on the loss function is curbed using Huber's criterion (Huber, 1981). This procedure leads to robust EFA that can resist the effect of outliers in the data.

Like EFA, independent component analysis (ICA) (e.g., Hyvärinen, Karhunen, and Oja, 2001) is a statistical model for reduction of the data dimensionality in which the interrelationships among the observed variables are explained in terms of a much smaller number of latent sources. The method of ICA is based on the assumption that if different sources stem from different physical processes, then those sources are statistically mutually independent. Accordingly, ICA tries to separate signal mixtures into independent sources and if independent components can be found they are identified with the hidden sources. In addition, at most one component in ICA is allowed to be normally distributed.

The ICA formulation is closely related to PCA. Whereas PCA only decorrelates the data, ICA looks for components that are mutually independent and non-normal. This is achieved by optimizing criteria that involve measures of departure from normality and/or independence using supplementary information not contained in the sample covariance or correlation matrix (e.g., Hyvärinen, Karhunen, and Oja, 2001). The ICA approach can also be viewed as a special case of exploratory projection pursuit (Friedman and Tukey, 1974). Whereas in projection pursuit the maximally non-normal projections of the data are considered interesting from the viewpoint of visualization and exploratory data analysis, ICA seeks non-normal projections of the data which produce independent components.

The vast majority of the literature treats the classical (noise-free) ICA model without allowing for unique factors. In most applications, it might be more realistic to assume that the manifest variables contain some kind of specific variance as well as measurement error. The introduction of unique factors in the ICA framework has led to the development of so-called noisy ICA models (e.g., Davies, 2004; Hyvärinen, Karhunen, and Oja, 2001). The noisy ICA formulation is very similar to EFA. The key difference between EFA and noisy ICA is that in the latter model the common factors are assumed to be both independent and non-normal. This assumption solves the rotational indeterminacy of the EFA model (Mooijaart, 1985). The loading matrix can be identified up to trivial ambiguities and unlike EFA there is no need for further factor rotation. In fact, noisy ICA can be considered as one particular method for factor rotation, along with the traditional 'simple structure' rotation methods (e.g. Varimax) which originated in psychometrics (Hastie, Tibshirani, and Friedman, 2009).

This thesis contributes to this area by exploiting the link between noisy ICA and EFA with factor rotation. Starting from an initial EFA solution, an orthogonal rotation matrix is sought that minimizes the dependence between the common factors. The optimal rotation matrix found is then applied to the initial loading matrix to compensate for the rotation of the scores. This procedure fits the noisy ICA model and coincidentally finds uniquely identified factor loadings in EFA. In contrast to the standard noisy ICA model with random latent sources, it is assumed in this thesis that the common factors are fixed matrix parameters. This new method is named independent exploratory factor analysis.

Finally, the idea of rotating the scores towards independence is employed in three-mode factor analysis. Standard ICA is based on two-way data matrices. Sometimes three-way data emerge, for instance, if $n$ subjects are measured on $p$ variables on $t$ occasions. For analyzing three-way data sets, Beckmann and Smith (2005) and De Vos, De Lathauwer, and Van Huffel (2007) combined ICA and the three-way model of PARallel FACtor analysis (PARAFAC) (Harshman, 1970), also known as the CANonical DECOMPosition (CANDECOMP) model (Carroll and Chang, 1970). In this thesis, an alternative approach to ICA for three-way data is considered. The rotational freedom of the three-mode factor analysis (Tucker3) model (Kroonenberg and De Leeuw, 1980; Tucker, 1966) is exploited to implement ICA in one mode of the data.

The new theoretical and computational aspects contained in this thesis are illustrated by means of several examples with real and artificial data.

Computations in this thesis are carried out using the software package MATLAB 7.7.0 (The MathWorks, 2008) on a PC under the Windows XP operating system with an Intel Pentium 4 CPU having 2.4 GHz clock frequency and 1 GB of RAM. All computer code used in the numerical experiments is available upon request.

## 1.2 Outline of the thesis

The thesis is organized as follows. Chapter 2 briefly outlines the continuous-time projected gradient method (Trendafilov, 2006) which is occasionally used in this thesis for studying and/or solving constrained matrix optimization problems. Chapter 3 describes the four related statistical techniques for dimensionality reduction of data which are referred to in this thesis: PCA, projection pursuit, noisy ICA, and EFA.

Part II entitled "EFA as Data Matrix Decomposition" begins in Chapter 4 with a literature review of procedures for fitting the EFA model with fixed common factors. Chapter 5 discusses algorithms for simultaneous estimation of all EFA model unknowns in the classical case of vertical data matrices ($n > p$). By means of the projected gradient approach, first-order necessary conditions for the existence of the minimizers of the corresponding loss functions are established. Methods for EFA-like PCA for the case $n > p$ are proposed in Chapter 6. A comparison of the optimality conditions derived for EFA-like PCA to the ones for simultaneous EFA sheds light on when one can expect similar EFA and PCA solutions. The algorithms developed in Chapter 5 and Chapter 6 are illustrated numerically with Harman's five socio-economic variables data (Harman, 1976).

Chapter 7 covers the case of horizontal data matrices ($p \geq n$). Novel approaches for simultaneous estimation of all EFA model unknowns are introduced and an algorithm for EFA-like PCA is presented. The new approaches are illustrated with Thurstone's 26-variable box data (Thurstone, 1947) and a real large high-dimensional data set from atmospheric science. In Chapter 8, a majorization algorithm for simultaneous parameter estimation in robust EFA is presented. An application to European health and fertility data shows the performance of the proposed approach. Chapter 9 concludes Part II by summarizing the main findings.

Part III entitled "Rotation Towards Independence in Factor Analysis" begins in Chapter 10 with a literature review of procedures for fitting the noisy ICA model. The independent exploratory factor analysis method is introduced in Chapter 11. A fitting solution for the new method is obtained by rotating an initial EFA solution such that the common factor scores are approximately independent. This is done using an appropriate rotation criterion and the projected gradient method. An application to Thurstone's 26-variable box problem is presented.

In Chapter 12 a novel approach to ICA for three-way data is considered. By exploiting the rotational freedom of the Tucker3 model, ICA is implemented in one mode of the data. A simulation experiment illustrates the performance of the proposed approach under different conditions. Chapter 13 concludes Part III.

The current Chapter continues with defining some notation and introducing the preliminary concepts of centring and scaling multivariate data.

## 1.3 Notation

Matrices and vectors are denoted by uppercase and lowercase letters in bold-faced type, $\mathbf{A}$ and $\mathbf{a}$, respectively, and scalars by italics. A vector $\mathbf{a}$ is considered to be a column vector. The transpose of $\mathbf{A}$ is denoted by $\mathbf{A}^\top$; thus $\mathbf{a}^\top$ is a row vector. If $\mathbf{A}$ is square, its determinant is written $\det(\mathbf{A})$ and the sum of its diagonal elements, the trace, is written $\mathrm{trace}(\mathbf{A})$. If $\mathbf{A}$ is nonsingular, its inverse is denoted by $\mathbf{A}^{-1}$.

Subscripts to lowercase letters indicate elements of a matrix. For example, the $(i, j)$-th element of $\mathbf{A}$ is indicated by $a_{ij}$. An $n \times n$ diagonal matrix with elements $d_1, d_2, \ldots, d_n$ on its main diagonal is represented by $\text{diag}(d_1, d_2, \ldots, d_n)$. An identity matrix of order $n$ is denoted by $\mathbf{I}_n$ and a vector of $n$ ones by $\mathbf{1}_n$. Analogously, an $n \times p$ matrix of zeros is denoted by $\mathbf{O}_{n \times p}$ and a vector of $n$ zeros by $\mathbf{0}_n$.

The symbol $\mathbb{R}$ represents the set of real numbers. The vector space of all $m \times n$ real matrices is denoted by $\mathbb{R}^{m \times n}$. For any matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ the rank of $\mathbf{A}$ is written $\text{rank}(\mathbf{A})$, the range space of $\mathbf{A}$ is denoted by $\text{range}(\mathbf{A})$ and the null space of $\mathbf{A}$ by $\text{null}(\mathbf{A})$.

Random variables and their realizations are not distinguished by using uppercase and lowercase letters. This is because in the multivariate case the reader will be more concerned about whether a vector or matrix of data is involved than with the distinction between random variables and their observed values.

If $\mathbf{x}$ is a random vector, then $\text{E}(\mathbf{x})$ represents the expectation or mean of $\mathbf{x}$. If $\mathbf{x}$ is a $p$-dimensional random vector which is normally distributed with mean $\boldsymbol{\mu} = \text{E}(\mathbf{x})$ and covariance matrix $\boldsymbol{\Sigma} = \text{E}\left[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top\right]$, this is abbreviated as $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Further symbols and definitions are introduced when necessary to clarify the presentation.

## 1.4 Multivariate data and preprocessing

Let $\mathbf{x} = (x_1, \ldots, x_p)^\top$ be a $p$-dimensional random vector with population mean $\boldsymbol{\mu}$ and population covariance matrix $\boldsymbol{\Sigma}$. Throughout this thesis it is assumed that all variables are continuous.

Suppose that a sample of $n$ realizations of $\mathbf{x}$ is available. These $np$ measurements $x_{ij}$ $(i = 1, \ldots, n; j = 1, \ldots, p)$ can be collected in a data matrix $\mathbf{X} = (\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(n)})^\top =$

$(\mathbf{x}_1, \ldots, \mathbf{x}_p) \in \mathbb{R}^{n \times p}$ with $\mathbf{x}_{(i)}^\top = (x_{i1}, \ldots, x_{ip})$ being the $i$-th observation vector ($i = 1, \ldots, n$) and $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^\top$ being the vector of the $n$ measurements on the $j$-th variable ($j = 1, \ldots, p$).

It will be useful in this and the subsequent Chapters to preprocess $\mathbf{x}$ so that its components have commensurate means. This is done by centring $\mathbf{x}$, that is, $\mathbf{x} \leftarrow \mathbf{x} - \boldsymbol{\mu}$. For the transformed vector $\mathbf{x}$ it holds that $\mathrm{E}(\mathbf{x}) = \mathbf{0}_p$. In a sample setting, the centred data matrix in which all columns have zero mean can be computed as

$$\mathbf{X} \leftarrow \mathbf{C}_n \mathbf{X} \ , \tag{1.1}$$

where $\mathbf{C}_n = (\mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^\top)$ is the centring matrix. Unless specified otherwise, it is always assumed in the sequel that both $\mathbf{x}$ and $\mathbf{X}$ are mean-centred.

One can transform a mean-centred random vector or mean-centred data further such that its variables have commensurate scales. Let $\boldsymbol{\Delta}$ be the $p \times p$ diagonal matrix whose elements on the main diagonal are the same as those of $\Sigma$. The standardized random vector $\mathbf{z}$ with components having unit variance can be obtained as

$$\mathbf{z} = \boldsymbol{\Delta}^{-1/2}\mathbf{x} \ , \tag{1.2}$$

where $\boldsymbol{\Delta}^{-1/2}$ is the diagonal matrix whose diagonal entries are the inverses of the square roots of those of $\boldsymbol{\Delta}$. This procedure carries over to the sample case in a straightforward fashion. Let $\mathrm{S}_\mathbf{X} = \mathbf{X}^\top\mathbf{X}/(n-1)$ be the sample covariance matrix of $\mathbf{X}$ and let $\mathbf{D}$ denote the $p \times p$ diagonal matrix whose elements on the main diagonal are the same as those of $\mathrm{S}_\mathbf{X}$. The standardized data matrix $\mathbf{Z}$ with all its columns having variance equal to one can be computed as

$$\mathbf{Z} = \mathbf{X}\mathbf{D}^{-1/2} \ . \tag{1.3}$$

Thus, $\mathbf{Z}^{\mathsf{T}}\mathbf{Z}/(n-1)$ is the sample correlation matrix. It will be found convenient in the later development to introduce a different form of scaling such that the variables are normalized to have unit length. One can obtain such a normalized matrix $\mathbf{Z}$ as

$$\mathbf{Z} = \frac{1}{\sqrt{n-1}}\mathbf{X}\mathbf{D}^{-1/2} \ , \tag{1.4}$$

in which the columns have variance equal to $1/(n-1)$. One advantage of the scaling in (1.4) is that now $\mathbf{Z}^{\mathsf{T}}\mathbf{Z}$ is the matrix of observed correlations so that division by $(n-1)$ is not required.

# Chapter 2

# The Dynamical System Approach to Optimization

Problems in multivariate statistics are often concerned with the optimization of matrix functions of structured (e.g. orthogonal) matrix unknowns. The dynamical system approach (Trendafilov, 2006) is a natural way of solving such optimization problems as it is especially designed to follow the geometry of the matrix parameters. It is a specific continuous-time method based on the classical gradient approach and modified for analyzing and solving constrained optimization problems. In Section 2.1 some rationale for the dynamical system approach is given. The four constrained manifolds used in this thesis are briefly discussed in Section 2.2.

## 2.1 Rationale

Let $\mathbf{Y}$ be an arbitrary real matrix and let $\mathcal{F}(\mathbf{Y})$ denote an objective function to be minimized. In a continuous-time setting the gradient descent method for unconstrained optimization of $\mathcal{F}(\mathbf{Y})$ can be expressed by the following gradient dynamical system (Hirsh and Smale, 1974):

$$\dot{\mathbf{Y}}(t) = \frac{d\mathbf{Y}(t)}{dt} = -\nabla\mathcal{F}(\mathbf{Y}(t)) \tag{2.1}$$

together with the initial condition $\mathbf{Y}(0) = \mathbf{Y}_0$, where $t \geq 0$ is a real variable interpreted as time and $\nabla\mathcal{F}$ is the gradient of the objective function $\mathcal{F}$. The solution of the initial value problem for the matrix ordinary differential equation (ODE) of first order in (2.1) gives the curve (hereafter referred to as 'flow') $\mathbf{Y}(t)$ along the steepest descent direction leading to a minimizer of $\mathcal{F}$.

Assume that $\mathbf{Y}(t)$ is restricted to move on a certain Riemannian manifold $\mathcal{M}$ (Helmke and Moore, 1994). Since the gradient in (2.1) is determined only by $\mathcal{F}$ and not by the constraint manifold imposed, $\nabla\mathcal{F}(\mathbf{Y}(t))$ may move the flow $\mathbf{Y}(t)$ out of $\mathcal{M}$. In this case, the gradient projection method can be used instead (Chu and Driessel, 1990). Its aim is to keep the flow $\mathbf{Y}(t)$ following the steepest descent direction and moving on the constrained manifold simultaneously. Unlike (2.1), the projected gradient method is concerned with the following dynamical system (Trendafilov, 2006):

$$\dot{\mathbf{Y}}(t) = -\pi(\nabla\mathcal{F}(\mathbf{Y}(t))) \ , \tag{2.2}$$

where $\pi(\nabla\mathcal{F})$ denotes the projection of the gradient $\nabla\mathcal{F}(\mathbf{Y}(t))$ onto the tangent space of the feasible set $\mathcal{M}$. Chu and Driessel (1990) showed that $\pi(\nabla\mathcal{F})$ is monotonically and globally decreasing along $\mathbf{Y}(t)$, i.e. convergence to a (local) minimizer is reached independently of the initial state $\mathbf{Y}_0$.

## 2.2 Constrained manifolds

Let the feasible set $\mathcal{M}$ be the manifold of all real $p \times k$ matrices with orthonormal columns, that is,

$$\mathcal{M} = \mathcal{O}(p,k) := \{\mathbf{T} \in \mathbb{R}^{p \times k} | \mathbf{T}^\top\mathbf{T} = \mathbf{I}_k\} \quad (k \leq p) \ ,$$

of which Stiefel (1935-1936) first studied the topological properties. The feasible set of all $p \times k$ column-wise orthonormal matrices $\mathcal{O}(p,k)$ forms a smooth compact submanifold in $\mathbb{R}^{p\times k}$ whose dimension is given by $\dim(\mathcal{O}(p,k)) = pk - (k+1)k/2$ (Edelman, Arias, and Smith, 1998). Note that the compact Stiefel manifold of all $p \times k$ orthonormal matrices is distinct from the non-compact Stiefel manifold of all $p \times k$ matrices whose columns are linearly independent (Absil, Mahony, and Sepulchre, 2008; Helmke and Moore, 1994):

$$\mathcal{ST}(p,k) := \{\mathbf{T} \in \mathbb{R}^{p\times k} |\, \mathrm{rank}(\mathbf{T}) = k\} \quad (k \leq p) \ .$$

The compact Stiefel manifold $\mathcal{O}(p,k)$ can be regarded as being embedded in the $pk$-dimensional Euclidean space $\mathbb{R}^{p\times k}$ equipped with the Frobenius inner product:

$$\langle \mathbf{A}, \mathbf{B} \rangle := \mathrm{trace}(\mathbf{A}^\top \mathbf{B}) \tag{2.3}$$

for any $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{p\times k}$. Then, $||\mathbf{A}||_F = \langle \mathbf{A}, \mathbf{A} \rangle^{1/2} = \sqrt{\mathrm{trace}(\mathbf{A}^\top \mathbf{A})}$ is the induced Frobenius matrix norm of $\mathbf{A}$. Suppose that $\mathbf{T}$ depends on $t$ such that, for all $t \geq 0$, $\mathbf{T}(t)$ forms a one-parameter family of $p \times k$ orthonormal matrices. Thus, $\mathbf{T}(t)$ can be regarded as a curve evolving on $\mathcal{O}(p,k)$. To facilitate notation, $\mathbf{T}(t)$ is abbreviated to $\mathbf{T}$ hereafter.

In the projected gradient approach the crucial step is to project the gradient of the objective function onto the feasible set of the optimization problem. The projection of an arbitrary $\mathbf{G} \in \mathbb{R}^{p\times k}$ onto the tangent space $\mathcal{T}_\mathbf{T}\mathcal{O}(p,k)$ at $\mathbf{T} \in \mathcal{O}(p,k)$ is given by (Edelman, Arias, and Smith, 1998):

$$\pi_\mathcal{T}(\mathbf{G}) = \mathbf{T}\frac{\mathbf{T}^\top\mathbf{G} - \mathbf{G}^\top\mathbf{T}}{2} + (\mathbf{I}_p - \mathbf{T}\mathbf{T}^\top)\mathbf{G} \ . \tag{2.4}$$

For orthogonal matrices with $p = k$, let

$$\mathcal{O}(k) := \{\mathbf{T} \in \mathbb{R}^{k\times k} | \mathbf{T}^\top\mathbf{T} = \mathbf{T}\mathbf{T}^\top = \mathbf{I}_k\} \tag{2.5}$$

be defined as the set of all $k \times k$ orthogonal matrices which forms a smooth manifold with $\dim(\mathcal{O}(k)) = k(k-1)/2$ in $\mathbb{R}^{k \times k}$. The projection of an arbitrary $\mathbf{G} \in \mathbb{R}^{k \times k}$ onto the tangent space $\mathcal{T}_{\mathbf{T}}\mathcal{O}(k)$ at $\mathbf{T} \in \mathcal{O}(k)$ is given by (Jennrich, 2001):

$$\pi_{\mathcal{T}}(\mathbf{G}) = \mathbf{T}\frac{\mathbf{T}^\top\mathbf{G} - \mathbf{G}^\top\mathbf{T}}{2} \ . \tag{2.6}$$

Finally, let $\mathcal{OB}(k)$ be the set of all non-singular $k \times k$ matrices $\mathbf{T}$ with columns of length one, that is,

$$\mathcal{OB}(k) := \{\mathbf{T} \in \mathbb{R}^{k \times k} | \mathrm{diag}(\mathbf{T}^\top\mathbf{T}) = \mathbf{I}_k\} \ . \tag{2.7}$$

The set $\mathcal{OB}(k)$ is the set of square oblique matrices $\mathbf{T}$ which forms a smooth manifold with $\dim(\mathcal{OB}(k)) = k(k-1)$ in $\mathbb{R}^{k \times k}$. The projection of an arbitrary $\mathbf{G} \in \mathbb{R}^{k \times k}$ onto the tangent space $\mathcal{T}_{\mathbf{T}}\mathcal{OB}(k)$ at $\mathbf{T} \in \mathcal{OB}(k)$ is given by (Jennrich, 2002):

$$\pi_{\mathcal{T}}(\mathbf{G}) = \mathbf{G} - \mathbf{T}\ \mathrm{diag}(\mathbf{T}^\top\mathbf{G}) \ . \tag{2.8}$$

The dynamical system approach can also give qualitative information about the optimization problem under consideration. For example, first-order necessary conditions for the existence of stationary points can easily be derived.

The computational procedures require numerical integrators for solving initial value problems for matrix ODEs. They are implemented in MATLAB (The MathWorks, 2008). Throughout the thesis, the solver used for the initial value problems is **ode15s** from the MATLAB in-built ODE suite (Shampine and Reichelt, 1997). The code **ode15s** is a quasi-constant step size implementation of the Klopfenstein-Shampine family of numerical differential formulae for stiff systems (Shampine and Reichelt, 1997). The integration of the matrix ODEs is terminated when the relative improvement of the objective function between two consecutive output points is less than $10^{-7}$, indicating that a local minimizer has been found. This stopping criterion is used to control the accuracy in following the solution path.

# Chapter 3

# Dimensionality Reduction Techniques in Unsupervised Learning

All the techniques described in this Chapter aim to reduce the dimensionality of a set of data. Given a set of $p$ observed variables, the aim is to find a $k$-dimensional ($k \ll p$) representation of it, while retaining as much as possible of the information present in the original set, according to some criterion. Since all the techniques have in common that the representation of the data is explored without an a priori output measure, they can be classified as unsupervised learning (Hastie, Tibshirani, and Friedman, 2009). Both PCA and projection pursuit are viewed as merely descriptive methods concerned with summarizing a data matrix in a manner which expresses its structure in a smaller number of dimensions. In contrast, EFA and noisy ICA are both latent variable models, that is, they attempt to achieve a reduction from $p$ to $k$ dimensions by invoking a statistical model relating the $p$ observed variables to $k$ latent variables.

All four methods can be described in a population or a sample setting. For PCA and projection pursuit the descriptions are given in a sample setting in Section 3.1. In Section 3.2, the model-based techniques noisy ICA and EFA are outlined in both population and sample settings.

## 3.1 Descriptive methods

### 3.1.1 Principal component analysis

Principal component analysis (e.g., Jolliffe, 2002) is the most popular multivariate technique for reducing the dimensionality of a data set. Let $\mathbf{X} = (\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(n)})^\top \in \mathbb{R}^{n \times p}$ be a given data matrix with sample covariance matrix $\mathbf{S}_\mathbf{X}$. The aim of PCA is to derive $k$ ($\ll p$) uncorrelated linear combinations of the $p$-dimensional observation vectors $\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(n)}$, called the sample principal components (PCs), which retain most of the total variation present in the data. This is achieved by taking those $k$ components that successively have maximum variance, that is, PCA looks for $k$ vectors $\mathbf{e}_j \in \mathbb{R}^{p \times 1}$ $(j = 1, \ldots, k)$ which

$$\begin{aligned} &\text{maximize} && \mathbf{e}_j^\top \mathbf{S}_\mathbf{X} \mathbf{e}_j \\ &\text{subject to} && \mathbf{e}_j^\top \mathbf{e}_j = 1 \quad \text{for } j = 1, \ldots, k \quad \text{and} && (3.1) \\ & && \mathbf{e}_i^\top \mathbf{e}_j = 0 \quad \text{for } i = 1, \ldots, j-1 \quad (j \geq 2) \ . && (3.2) \end{aligned}$$

It turns out that $\mathbf{y}_j = \mathbf{X}\mathbf{e}_j$ is the $j$-th sample PC with zero mean and variance $\omega_j$, where $\mathbf{e}_j$ is an eigenvector of $\mathbf{S}_\mathbf{X}$ corresponding to its $j$-th largest eigenvalue $\omega_j$ $(j = 1, \ldots, k)$. The condition (3.2) ensures that the sample PCs are uncorrelated.

The sample PCs can be found efficiently using the singular value decomposition (SVD) of $\mathbf{X}$ (e.g., Golub and Van Loan, 1996). Assume that $\mathbf{X}$ has rank $r$ with $r \leq \min\{n, p\}$. Expressing $\mathbf{X}$ by its SVD gives

$$\mathbf{X} = \mathbf{V}\mathbf{D}\mathbf{E}^\top = \sum_{j=1}^{r} \sigma_j \mathbf{v}_j \mathbf{e}_j^\top \ , \tag{3.3}$$

where $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_r) \in \mathbb{R}^{n \times r}$ and $\mathbf{E} = (\mathbf{e}_1, \ldots, \mathbf{e}_r) \in \mathbb{R}^{p \times r}$ are orthonormal matrices such that $\mathbf{V}^\top\mathbf{V} = \mathbf{E}^\top\mathbf{E} = \mathbf{I}_r$, and $\mathbf{D} \in \mathbb{R}^{r \times r}$ is a diagonal matrix with the singular values of $\mathbf{X}$ sorted in decreasing order, $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$, on its main diagonal. The matrix $\mathbf{E}$ in (3.3) is the matrix of coefficients or loadings and the matrix of component scores $\mathbf{Y} \in \mathbb{R}^{n \times r}$ is given by $\mathbf{Y} = \mathbf{VD}$. Since it holds that $\mathbf{E}^\top \mathbf{E} = \mathbf{I}_r$ and $\mathbf{Y}^\top \mathbf{Y}/(n-1) = \mathbf{D}^2/(n-1)$, the loadings are orthogonal and the sample PCs are uncorrelated. The variance of the $j$-th sample PC is $\sigma_j^2/(n-1)$ which is equal to the $j$-th largest eigenvalue, $\omega_j$, of $\mathbf{S_X}$ $(j = 1, \ldots, r)$. In practice, the leading $k$ components with $k \ll r$ usually account for a substantial proportion, $(\omega_1 + \cdots + \omega_k)/\text{trace}(\mathbf{S_X})$, of the total variance in the data (say 80%) and the sum in (3.3) is therefore truncated after the first $k$ terms. If so, PCA comes down to finding a matrix $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_k) \in \mathbb{R}^{n \times k}$ of component scores of the $n$ samples on the $k$ components and a matrix $\mathbf{E} = (\mathbf{e}_1, \ldots, \mathbf{e}_k) \in \mathbb{R}^{p \times k}$ of coefficients whose $k$-th column is the vector of loadings for the $k$-th component. Due to the least squares property of the SVD (Eckart and Young, 1936), PCA can be defined as the minimization of

$$||\mathbf{X} - \mathbf{Y}\mathbf{E}^\top||_F^2 \ . \tag{3.4}$$

Note that PCA is not scale-invariant. When variables are measured on different scales or on a common scale with widely differing ranges, the data are often standardized prior to PCA. Basically, the sample PCs are then obtained from an eigenvalue decomposition of the sample correlation matrix. These components are not equal to those derived from $\mathbf{S_X}$ and knowledge of one set does not allow simple transformation to the other set (e.g., Krzanowski, 1988).

To enhance interpretation of the sample PCs, it is common in PCA to rotate the matrix of loadings by optimizing a certain 'simplicity' criterion (e.g., Richman, 1986, see also Chapter 11 in this thesis). The method of rotation emerged in EFA and was motivated both by solving the rotational indeterminacy problem and by facilitating the factors' interpretation (Browne, 2001). Rotation can be performed either in an orthogonal or an oblique (non-orthogonal) fashion. Several analytic orthogonal and oblique rotation criteria exist in the literature (e.g., Browne, 2001; Richman, 1986). To aid interpretation, all criteria are designed to make the coefficients as simple as possible in some sense, with most loadings made to have values either 'close to zero' or 'far from zero', and with as few as possible of the coefficients taking intermediate values. However, after rotation, either one or both of the properties possessed by PCA, that is, orthogonality of the loadings and uncorrelatedness of the component scores, is lost.

### 3.1.2 Projection pursuit

Principal component analysis provides a computationally efficient way of projecting the $p$-dimensional data cloud orthogonally onto a $k$-dimensional subspace. However, the variance-maximization projection accomplished by PCA does not necessarily afford the most informative view of the structure of multivariate data (Bolton and Krzanowski, 1999). Various other possibilities exist to provide representations of the data based on subspace projection. Projection pursuit (Friedman and Tukey, 1974) is concerned with finding 'interesting' low-dimensional projections of multivariate data in which features such as clusters or outliers can be detected (see also Friedman, 1987; Huber, 1985; Jones and Sibson, 1987). Since in practice a graphical display of the projections is the output of choice, the projections are typically 1-, 2-, or 3-dimensional (Nason, 1992). To form a linear projection of the $n \times p$ data matrix $\mathbf{X}$ onto the real line, a $p$-dimensional vector $\mathbf{a}$ is specified with the constraint $\mathbf{a}^{\mathsf{T}}\mathbf{a} = 1$. The projected data is formed by $\mathbf{Xa}$.

For an orthogonal projection onto a space of $k$ ($k \ll p$) dimensions greater than one, a $p \times k$ matrix $\mathbf{A}$ with $\mathbf{A}^\top \mathbf{A} = \mathbf{I}_k$ is required. The projected data is formed by $\mathbf{XA}$. Projection pursuit works by associating a function value $\mathcal{I}(\mathbf{XA})$ to each and every low-dimensional projection. This function value is, say, large for projections revealing interesting structure, and small for uninteresting ones. Interesting projections are then revealed by optimizing the criterion over all possible projections. The function $\mathcal{I}$ is called the projection index. A special case of a projection pursuit technique is PCA in which the index of interestingness is the variance of the projected data. This is one of the few available indices that can be optimized analytically.

Since PCA investigates the covariance structure of the data, there is no need for projection pursuit to do the same. Therefore, before the application of projection pursuit, it is common practice (e.g., Nason, 1992) to preprocess the data $\mathbf{X}$ by a linear transformation to have identity sample covariance matrix. This transformation is called sphering (or whitening) (Tukey and Tukey, 1981) and can be carried out as follows.

Let $\mathbf{S}_\mathbf{X}$ be positive semi-definite with $\text{rank}(\mathbf{S}_\mathbf{X}) = r$ ($r \leq p$) and let the eigenvalue decomposition of $\mathbf{S}_\mathbf{X}$ be

$$\mathbf{S}_\mathbf{X} = \mathbf{E}\boldsymbol{\Omega}\mathbf{E}^\top = \sum_{i=1}^{r} \omega_i \mathbf{e}_i \mathbf{e}_i^\top \ , \tag{3.5}$$

where $\boldsymbol{\Omega} = \text{diag}(\omega_1, \ldots, \omega_r)$ is an $r \times r$ diagonal matrix containing the positive eigenvalues of $\mathbf{S}_\mathbf{X}$, $\omega_1 \geq \cdots \geq \omega_r > 0$, on its main diagonal and $\mathbf{E} \in \mathbb{R}^{p \times r}$ is an orthonormal matrix whose columns $\mathbf{e}_1, \ldots, \mathbf{e}_r$ are the corresponding unit-norm eigenvectors of $\omega_1, \ldots, \omega_r$.

The sphered data matrix $\tilde{\mathbf{X}} \in \mathbb{R}^{n \times r}$ can be obtained by

$$\tilde{\mathbf{X}} = \mathbf{X}\mathbf{E}\boldsymbol{\Omega}^{-1/2} = \mathbf{X}\mathbf{W} \ , \tag{3.6}$$

where $\mathbf{W} = \mathbf{E}\boldsymbol{\Omega}^{-1/2} \in \mathbb{R}^{p\times r}$ is the sphering matrix. Let $\mathbf{S}_{\tilde{\mathbf{X}}} = \tilde{\mathbf{X}}^{\mathsf{T}}\tilde{\mathbf{X}}/(n-1)$ be the covariance matrix of $\tilde{\mathbf{X}}$. For $\mathbf{S}_{\tilde{\mathbf{X}}}$ it holds that

$$\mathbf{S}_{\tilde{\mathbf{X}}} = \boldsymbol{\Omega}^{-1/2}\mathbf{E}^{\mathsf{T}}\mathbf{S}_{\mathbf{X}}\mathbf{E}\boldsymbol{\Omega}^{-1/2} = \boldsymbol{\Omega}^{-1/2}\mathbf{E}^{\mathsf{T}}\mathbf{E}\boldsymbol{\Omega}\mathbf{E}^{\mathsf{T}}\mathbf{E}\boldsymbol{\Omega}^{-1/2} = \mathbf{I}_r \ , \tag{3.7}$$

as desired. Since all orthogonal projections of sphered centered data inherit the properties of zero mean and identity covariance matrix, $\mathbf{W} = \mathbf{E}\boldsymbol{\Omega}^{-1/2}$ is by no means the only choice for a sphering matrix.

The transformation in (3.6) is equivalent to computing the sample PCs of $\mathbf{X}$ and then rescaling each of the sample PCs to have unit variance. To reduce the dimensionality of the data to $k$ ($\ll r$) dimensions, only the first $k$ sphered components need to be retained and the sum in (3.5) is truncated after $k$ terms, where $k$ is chosen to explain a certain proportion of the total variance. Sphering then comes down to finding a matrix $\mathbf{W} \in \mathbb{R}^{p\times k}$ and a transformed data matrix $\tilde{\mathbf{X}} \in \mathbb{R}^{n\times k}$ with $\mathbf{S}_{\tilde{\mathbf{X}}} = \mathbf{I}_k$.

A more detailed discussion of sphering for projection pursuit can be found in Jones and Sibson (1987) and Nason (1992) (see also the comments of Gower and of Hastie and Tibshirani in the discussion of Jones and Sibson, 1987).

Once a sphered data matrix $\tilde{\mathbf{X}}$ has been obtained, projection pursuit essentially consists of the following two-step procedure:

1. Choose a projection index $\mathcal{I}$ to judge the merit of a particular $k$-dimensional projection of $\tilde{\mathbf{X}}$.

2. Use an optimization algorithm to find the local optima of $\mathcal{I}$ chosen in step 1 over all $k$-dimensional projections of $\tilde{\mathbf{X}}$. This step determines the most informative $k$-dimensional projection of the data.

Most techniques in projection pursuit start from the premise that normality represents the notion of 'uninterestingness'. This is due to the fact that most low-dimensional projections of high-dimensional data look approximately normally distributed (Diaconis and Freedman, 1984). The projection pursuit indices are thus optimized to find projections showing departures from normality. Examples of projection pursuit indices frequently used include cumulant-based indices and negative entropy (negentropy) (see also Chapter 10).

## 3.2 Latent variable models

### 3.2.1 Noisy independent component analysis

Independent component analysis (e.g., Hyvärinen, Karhunen, and Oja, 2001) is a model that seeks to uncover latent sources underlying a set of observed signal mixtures. In particular, ICA offers a methodology for doing what is called 'blind source separation' (e.g., Izenman, 2008, Chapter 15). The blind source separation problem consists in the recovery of unknown independent signals from observed linear combinations of them. The adjective 'blind' signifies that almost no information about the sources' distribution and the mixing process is known. An illustration of blind source separation is the 'cocktail-party problem' (e.g., Hyvärinen, Karhunen, and Oja, 2001) in which the speech of several people is received from microphones present in the room. The task is to recover the speech of the individual speakers from the overlapped talk.

Assume that $\mathbf{x} \in \mathbb{R}^{p\times 1}$ is a random vector of manifest variables. First, consider the noise-free ICA model (e.g., Comon, 1994) which states that $\mathbf{x}$ can be modeled as

$$\mathbf{x} = \mathbf{M}\boldsymbol{\xi} \ , \tag{3.8}$$

where $\boldsymbol{\xi} \in \mathbb{R}^{k \times 1}$ is a random vector of $k \leq p$ latent variables called sources or components and $\mathbf{M} \in \mathbb{R}^{p \times k}$ is a mixing matrix of fixed coefficients. Assume that the mixing matrix $\mathbf{M}$ has full column rank. Furthermore, suppose that the components of $\boldsymbol{\xi}$ have zero mean and unit variance. Finally, let $\boldsymbol{\xi}$ consist of mutually independent sources of which at most one is Gaussian, and whose densities are square integrable.

Given a data matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p) \in \mathbb{R}^{n \times p}$ of $n$ observations on $\mathbf{x}$, the noise-free ICA model holds if $\mathbf{X}$ can be written as

$$\mathbf{X} = \Xi \mathbf{M}^\top \ , \tag{3.9}$$

where $\Xi = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_k) \in \mathbb{R}^{n \times k}$ is the unknown matrix of scores for the $k$ sources on the $n$ observations. The notation here follows the convention established in Section 1.4. The aim of ICA is to estimate $\mathbf{M}$ and hence recover $\Xi$. In noise-free ICA the elements of $\mathbf{M}$ (and hence $\Xi$) can only be identified up to ambiguities in permutation and sign (Robitzsch, 2003). In other words, a separating (unmixing) matrix $\mathbf{B} \in \mathbb{R}^{k \times p}$ is sought such that

$$\mathbf{BM} = \mathbf{\Pi} \ , \tag{3.10}$$

where $\mathbf{\Pi} \in \mathbb{R}^{k \times k}$ is a generalized permutation matrix (e.g., Lütkepohl, 1996) whose nonzero entries are ±1. If the sources are not standardized to have equal variance, then $\mathbf{M}$ is unique up to permutation and scaling ambiguities (Robitzsch, 2003).

Before carrying out ICA, the data are typically sphered (e.g., Hyvärinen, Karhunen, and Oja, 2001) as in projection pursuit and the dimensionality of the data is reduced from $p$ to $k$ dimensions. Since sphering is essentially decorrelation followed by scaling, PCA is used for this preprocessing step. Postmultiplying $\mathbf{X}$ by the sphering matrix

$\mathbf{W} \in \mathbb{R}^{p \times k}$ introduced in the previous Section, (3.9) becomes

$$\tilde{\mathbf{X}} = \Xi \mathbf{M}^{\mathsf{T}} \mathbf{W} = \Xi \tilde{\mathbf{M}}^{\mathsf{T}} \ , \tag{3.11}$$

where $\tilde{\mathbf{M}} = \mathbf{W}^{\mathsf{T}} \mathbf{M}$ is the new orthogonal mixing matrix. Indeed:

$$\mathbf{S}_{\tilde{\mathbf{X}}} = \frac{1}{n-1} \tilde{\mathbf{X}}^{\mathsf{T}} \tilde{\mathbf{X}} = \tilde{\mathbf{M}} \frac{1}{n-1} \Xi^{\mathsf{T}} \Xi \tilde{\mathbf{M}}^{\mathsf{T}} = \tilde{\mathbf{M}} \mathbf{I}_k \tilde{\mathbf{M}}^{\mathsf{T}} = \tilde{\mathbf{M}} \tilde{\mathbf{M}}^{\mathsf{T}} = \mathbf{I}_k \ . \tag{3.12}$$

Thus, the search for the mixing matrix is confined to the set of orthogonal matrices. In other words, the search for the separating matrix $\mathbf{B}$ amounts to looking for an orthogonal matrix $\mathbf{T}$ such that $\mathbf{T}^{\mathsf{T}} \tilde{\mathbf{M}} = \mathbf{\Pi}$.

Sphering aids understanding of why ICA is not able to find $\mathbf{T}$ for normally distributed data. In Figure 3.1, 1000 realizations of two independent components both generated from a standard normal distribution are linearly mixed using a random mixing matrix $\mathbf{M}$. Since the sphered data has circular symmetry, the orthogonal matrix $\mathbf{T}$ cannot be found for normal data. In other words, the ICA model is not identifiable for independent Gaussian sources. This phenomenon is related to the property that jointly uncorrelated normal random variables are independent. It turns out that for $\mathbf{T}$ to be identifiable at most one source is allowed to be normally distributed (Robitzsch, 2003).

The sphering operation also shows the relation between PCA and ICA. Principal component analysis decorrelates the data using second-order information contained in the sample covariance matrix. That is, PCA is able to transform any linear mixture of independent components into uncorrelated components. However, the sphering matrix $\mathbf{W}$ only determines the sources up to an orthogonal transformation. Unlike PCA, ICA not only decorrelates the data, i.e. it goes one step further and finds the orthogonal transformation $\mathbf{T}$ that is left after decorrelation.

One main estimation principle for finding $\mathbf{T}$ is 'non-linear decorrelation' (Hyvärinen,

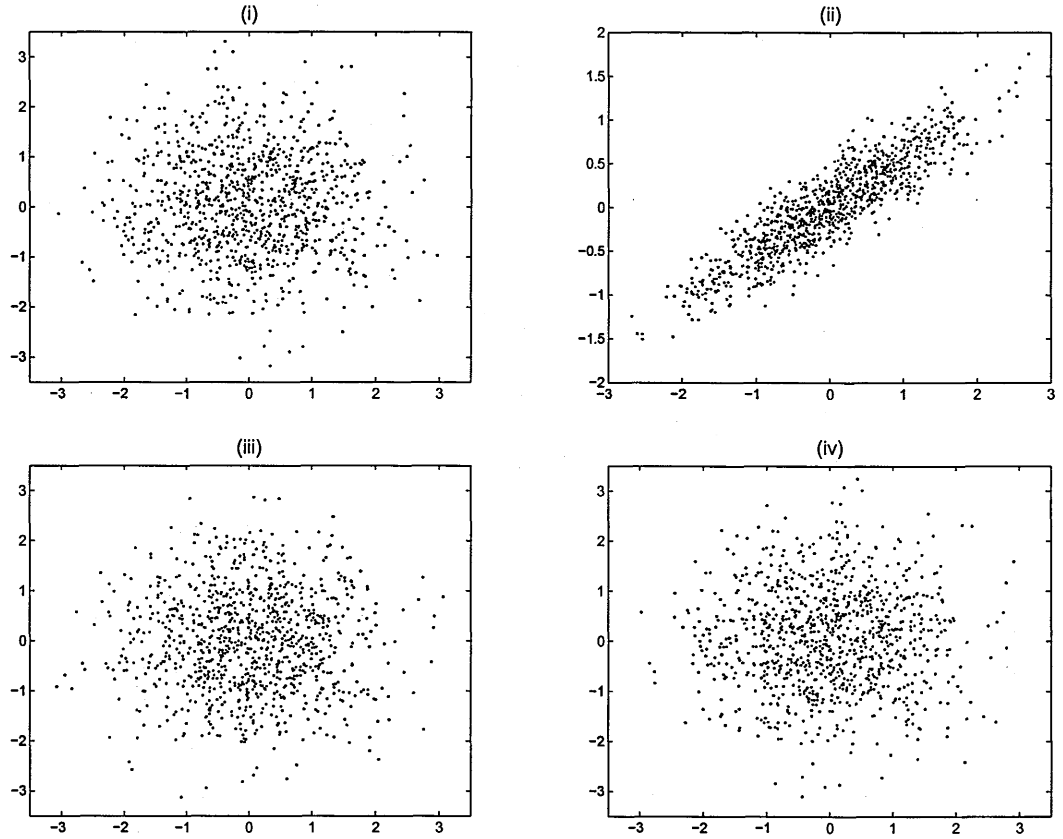

Figure 3.1: ICA of Gaussian data: (i) original data; (ii) after mixing; (iii) sphered data; (iv) reconstruction by ICA.

Karhunen, and Oja, 2001). Let $\xi_1, \ldots, \xi_k$ be a sequence of independent univariate random variables. For the $\xi_i$ $(i = 1, \ldots, k)$ it holds that (e.g., Casella and Berger, 2002, pp. 154-155):

$$\mathrm{E}[g(\xi_i)h(\xi_j)] = \mathrm{E}[g(\xi_i)]\mathrm{E}[h(\xi_j)] \quad \text{for } i \neq j \ ,$$

where $g(\xi_i)$ and $h(\xi_j)$ are any absolutely integrable functions of $\xi_i$ and $\xi_j$, respectively. Hence, independence implies non-linear uncorrelatedness. Thus, one could attempt to implement ICA by a stronger form of decorrelation where the recovered components are uncorrelated even after some non-linear transformation. If the non-linearities are chosen properly, the independent components can be recovered approximately (Hyvärinen, Karhunen, and Oja, 2001). Using the sample covariance matrix one can only decorrelate the data in the linear sense. Following the principle of non-linear decorrelation one has to use some form of supplementary (higher-order) information not contained in the sample covariance matrix.

Another main estimation principle is 'maximization of non-normality'. This principle is based on Lyapunov's central limit theorem (e.g., Pawitan, 2001, pp. 233-234). Assume that $\xi_1, \ldots, \xi_k$ are independent but not necessarily identically distributed random variables. Suppose that each $\xi_i$ $(i = 1, \ldots, k)$ has finite expected value $\mu_i$ and finite variance $\sigma_i^2$, with at least one $\sigma_i^2 > 0$. Suppose that the third absolute central moments, $\mathrm{E}(|X_i - \mu_i|^3)$, are finite for each $i$ and that Lyapunov's condition is satisfied:

$$\lim_{k\to\infty} \frac{(\sum_{i=1}^k \mathrm{E}(|X_i - \mu_i|^3))^{1/3}}{(\sum_{i=1}^k \sigma_i^2)^{1/2}} = 0 \ .$$

Let the random variable $S_K = \xi_1 + \cdots + \xi_K$ denote the $K$-th partial sum of $\xi_1, \ldots, \xi_k$. Then, for the normalized partial sum it holds that

$$\lim_{K\to\infty} Z_K = \frac{S_K - \sum_{i=1}^K \mu_i}{\sqrt{\sum_{i=1}^K \sigma_i^2}} \xrightarrow{d} \mathcal{N}(0,1) \ ,$$

suggesting that a sum of two or more independent non-normal random variables is closer to the normal distribution than the original ones (see also Diaconis and Freedman, 1984). In ICA, take a linear combination of the observed mixture variables which, in turn, is also a linear combination of the independent components. This linear combination of two or more independent sources is usually closer to the normal distribution than any of the original sources and will be maximally non-normal if it is in fact one of the independent components. Estimating the independent components can therefore be accomplished by finding the linear combination of the mixture variables which maximizes its non-normality. It has been argued that even for a fairly small number of sources (say, $k = 10$) the distribution of the linear mixture is usually close to the normal. This seems to hold even if the densities of the sources are far from each other and are far from being normal (Hyvärinen, Karhunen, and Oja, 2001, p. 35). This principle of maximization of non-normality shows the close connection between ICA and projection pursuit. Whereas projection pursuit looks merely for non-normal projections of the data, ICA seeks for both non-normal and independent components.

Solving the ICA problem is usually performed by specifying a criterion (called the objective or contrast function) for measuring the departure from normality and/or independence and then constructing an algorithm for optimizing this criterion. Criteria for measuring independence/non-normality will be discussed in Chapter 10. For a concise summary of optimization algorithms in noise-free ICA, the reader is referred to Izenman (2008) and the references therein.

The vast majority of the relevant literature treats the noise-free ICA model. However, often it is more realistic to assume that observations consist of a mixture of signals contaminated by some kind of measurement error and/or that the manifest variables have some specific variance which cannot be accounted for by the latent sources. In ICA, measurement error is referred to as noise and the introduction of noise in the ICA framework has led to the development of so-called noisy ICA models (e.g., Davies, 2004; Hyvärinen, Karhunen, and Oja, 2001). The linear noisy ICA model is defined as the following latent variable model (e.g., Hyvärinen, Karhunen, and Oja, 2001):

$$\mathbf{x} = \mathbf{M}\boldsymbol{\xi} + \mathbf{u} \ , \tag{3.13}$$

where $\mathbf{u} \in \mathbb{R}^{p\times 1}$ is a random vector of observational noise. In the sequel, to emphasize that the components of $\mathbf{u}$ contain noise specific to the corresponding observed variable,

$\mathbf{u}$ will be referred to as a vector of unique factors. Assume that $\mathbf{u} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Psi}^2)$, where $\mathbf{\Psi}^2$ is assumed to be a positive definite diagonal matrix. Finally, suppose that $\mathrm{E}(\boldsymbol{\xi}\mathbf{u}^\top) = \mathbf{O}_{k\times p}$. That is, the unique factors are normally distributed with diagonal covariance matrix $\mathbf{\Psi}^2$ and are uncorrelated from the latent sources. In the ICA literature it is often assumed that the unique factors have homoscedastic variance $\sigma^2$ and hence $\mathbf{\Psi}^2$ is simply of the form $\sigma^2\mathbf{I}_p$ (Hyvärinen, Karhunen, and Oja, 2001). For many applications this may be too restrictive. Here, the unique variances are allowed to vary across the manifest variables.

In noisy ICA the mixing matrix $\mathbf{M}$ is still unique up to ambiguities in permutation and sign (Robitzsch, 2003). In sample form the noisy ICA model can be written as

$$\mathbf{X} = \Xi\mathbf{M}^\top + \mathbf{U} \ , \tag{3.14}$$

where $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_p) \in \mathbb{R}^{n\times p}$ is the unknown matrix of scores of the $n$ observations on the $p$ unique factors. The notation again follows the convention established in Section 1.4. The aim of noisy ICA is to estimate $\mathbf{M}$ and recover $\Xi$. However, due to the existence of $\mathbf{U}$ in (3.14), knowledge of $\mathbf{M}$ does not give direct access to $\Xi$ as in noise-free ICA. This means that apart from estimating the mixing matrix $\mathbf{M}$ one requires a method for estimating the realizations of the independent components $\Xi$. This makes the noisy ICA problem much more difficult to solve. Procedures for fitting the noisy ICA model will be discussed in Chapter 10. The noisy ICA model is closely related to EFA which is discussed in the following.

### 3.2.2 Exploratory factor analysis

It is customary to work with standardized variables in this context, so we shall do so in the sequel. Let $\mathbf{z} \in \mathbb{R}^{p\times 1}$ be a random vector of standardized observed variables as defined in (1.2). Suppose that the EFA model (e.g., Lawley and Maxwell, 1971) holds which states that $\mathbf{z}$ can be written in the form:

$$\mathbf{z} = \Lambda\mathbf{f} + \mathbf{u} \ , \tag{3.15}$$

where $\mathbf{f} \in \mathbb{R}^{k\times 1}$ is a vector of $k$ ($k \ll p$) common factors, $\Lambda \in \mathbb{R}^{p\times k}$ with $\operatorname{rank}(\Lambda) = k$ is a matrix of fixed coefficients referred to as factor loadings, and $\mathbf{u} \in \mathbb{R}^{p\times 1}$ is a vector of unique factors. The choice of $k$ in EFA is subject to some limitations (Ledermann, 1937) which will not be discussed here. Assume that $\mathrm{E}(\mathbf{f}) = \mathbf{0}_k$ and $\mathrm{E}(\mathbf{u}) = \mathbf{0}_p$. Furthermore, let $\mathrm{E}(\mathbf{u}\mathbf{u}^\top) = \Psi^2$, where $\Psi^2$ is assumed to be a positive definite diagonal matrix. Finally, suppose that $\mathrm{E}(\mathbf{f}\mathbf{f}^\top) = \mathbf{I}_k$ and $\mathrm{E}(\mathbf{f}\mathbf{u}^\top) = \mathbf{O}_{k\times p}$. Hence, all factors are uncorrelated with one another. Under these assumptions, the model in (3.15) represents an EFA model with uncorrelated or orthogonal (random) common factors.

Unlike the above EFA model with random factors, the fixed EFA model considers $\mathbf{f}$ to be a vector of non-random quantities or parameters which vary from one case to another (Anderson and Rubin, 1956; Lawley, 1942).

The idea behind model (3.15) is that the common factors account for the covariance structure among the set of manifest variables, while each unique factor corresponds to that portion of a particular observed variable that cannot be accounted for by the common factors. As such, a unique factor contains the specificity of that variable as well as errors in measurement or noise.

For the random EFA model, it is often convenient to assume that $\mathbf{u}$ and $\mathbf{f}$ and hence

$\mathbf{z}$ are multinormally distributed. This assumption is usually made in order that maximum likelihood estimation can be used and for purposes of statistical inference (Mardia, Kent, and Bibby, 1979). As the elements of $\mathbf{f}$ are uncorrelated, the assumption of normality means that they are statistically independent random variables.

In contrast, noisy ICA assumes that the $k$ sources are both mutually independent and non-normal, or that at least all but one of them are non-normal. Apart from this difference, the EFA model (3.15) is virtually identical to the noisy ICA model (3.13), where the common factors $\mathbf{f}$ correspond to the sources $\boldsymbol{\xi}$ and the factor loadings $\mathbf{\Lambda}$ to the mixing matrix $\mathbf{M}$.

For the purposes of this thesis an alternative representation of the EFA model is employed in the sequel which is also used in the psychometric literature (e.g., Harman, 1976; Mulaik, 1972; Yates, 1987). The standard EFA model (3.15) can be rewritten as

$$\mathbf{z} = \mathbf{\Lambda}\mathbf{f} + \mathbf{\Psi}\mathbf{u} \ . \tag{3.16}$$

Assume that $\mathrm{E}(\mathbf{u}\mathbf{u}^\top) = \mathbf{I}_p$ and $\mathbf{\Psi}$ is a diagonal matrix of fixed coefficients called uniquenesses. Then, the model representation (3.16) is equivalent to model (3.15) in which $\mathrm{E}(\mathbf{u}\mathbf{u}^\top) = \mathbf{\Psi}^2$.

The EFA model (3.16) and the associated assumptions imply the following model correlation structure $\mathbf{\Theta}$ for the observed variables:

$$\mathbf{\Theta} = \mathbf{\Lambda}\mathbf{\Lambda}^\top + \mathbf{\Psi}^2 \ . \tag{3.17}$$

The converse also holds. If $\mathbf{\Theta}$ can be decomposed into the form (3.17) then the $k$-factor model holds for $\mathbf{z}$ (Mardia, Kent, and Bibby, 1979).

If the $k$-factor model holds then it also holds if the factors are rotated. If $\mathbf{T}$ is an arbitrary orthogonal $k \times k$ matrix, (3.16) may be rewritten as

$$\mathbf{z} = \mathbf{\Lambda}\mathbf{T}\mathbf{T}^{\mathsf{T}}\mathbf{f} + \mathbf{\Psi}\mathbf{u} \ , \tag{3.18}$$

which is a model with loading matrix $\mathbf{\Lambda}\mathbf{T}$ and common factors $\mathbf{T}^{\mathsf{T}}\mathbf{f}$. The assumptions about the variables that make up the original model are not violated by this transformation. Thus, if (3.18) holds, $\mathbf{\Theta}$ can be written as $\mathbf{\Theta} = (\mathbf{\Lambda}\mathbf{T})(\mathbf{T}^{\mathsf{T}}\mathbf{\Lambda}^{\mathsf{T}}) + \mathbf{\Psi}^2$, that is, for fixed $\mathbf{\Psi}$ and $k > 1$ there is a rotational indeterminacy in the decomposition of $\mathbf{\Theta}$ in terms of $\mathbf{\Lambda}$ and $\mathbf{\Psi}$. This means that there is an infinite number of factor loadings satisfying the original assumptions of the model. In other words, the parameters of the EFA model cannot be identified uniquely from second-order cross products (covariances or correlations) only.

Consequently, to ensure a unique solution for the model unknowns $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ additional constraints such as e.g. $\mathbf{\Lambda}^{\mathsf{T}}\mathbf{\Lambda}$ or $\mathbf{\Lambda}^{\mathsf{T}}\mathbf{\Psi}^{-2}\mathbf{\Lambda}$ being a diagonal matrix are imposed on the parameters in the original model (Jöreskog, 1977). These constraints eliminate the indeterminacy in (3.17), but such solutions are usually difficult to interpret. Instead, the parameter estimation is usually followed by some kind of rotation of $\mathbf{\Lambda}$ to some structure with specific features (e.g., Browne, 2001, see also Chapter 11 in this thesis).

Suppose that a sample of $n$ observations on $\mathbf{z}$ is available. Collect these measurements in a data matrix $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_p) \in \mathbb{R}^{n \times p}$ in which $\mathbf{z}_j = (z_{1j}, \ldots, z_{nj})^{\mathsf{T}}$ $(j = 1, \ldots, p)$. The $k$-factor model holds if $\mathbf{Z}$ can be written in the form:

$$\mathbf{Z} = \mathbf{F}\mathbf{\Lambda}^{\mathsf{T}} + \mathbf{U}\mathbf{\Psi} \ , \tag{3.19}$$

where $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_k) \in \mathbb{R}^{n \times k}$ and $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_p) \in \mathbb{R}^{n \times p}$ denote the unknown matrices of factor scores for the $k$ common factors and $p$ unique factors on $n$ observations, respectively. The notation again follows the convention established in Section

1.4. Without changing notation, assume that the columns of $\mathbf{Z}$, $\mathbf{F}$ and $\mathbf{U}$ are scaled to have unit length. Suppose that $\text{rank}(\mathbf{\Lambda}) = k$, $\mathbf{F}^\top\mathbf{F} = \mathbf{I}_k$, $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_p$, $\mathbf{U}^\top\mathbf{F} = \mathbf{O}_{p\times k}$ and that $\mathbf{\Psi}$ is a diagonal matrix.

In standard EFA (with random common factors), a pair $\{\mathbf{\Lambda}, \mathbf{\Psi}\}$ is sought which gives the best fit, for some specified value of $k$, to the sample correlation matrix $\mathbf{Z}^\top\mathbf{Z}$ with respect to some discrepancy measure. The process of finding this pair is called factor extraction. Various factor extraction methods have been proposed (e.g., Harman, 1976; Mulaik, 1972). If the data are assumed normally distributed the maximum likelihood principle is preferred. Then the factor extraction problem can be formulated as optimization of a certain log-likelihood function which is equivalent to the following fitting problem (Magnus and Neudecker, 1988):

$$\min_{\mathbf{\Lambda},\mathbf{\Psi}} \log(\det(\mathbf{\Lambda}\mathbf{\Lambda}^\top + \mathbf{\Psi}^2)) + \text{trace}((\mathbf{\Lambda}\mathbf{\Lambda}^\top + \mathbf{\Psi}^2)^{-1}(\mathbf{Z}^\top\mathbf{Z})) \ , \tag{3.20}$$

referred to as maximum likelihood (ML) factor analysis. It is worth mentioning that the loadings found by ML factor analysis for a correlation matrix are equivalent to those for the corresponding covariance matrix, that is, in contrast to PCA, ML factor analysis is scale invariant (Mardia, Kent, and Bibby, 1979).

If nothing is assumed about the distribution of the data, (3.20) can still be used as one way of measuring the discrepancy between the model and the sample correlation matrix. There are a number of other discrepancy measures which are used in place of (3.20). A natural choice is the least squares approach for fitting the EFA model. It can be formulated as the following general class of WLS problems (Bartholomew and Knott, 1999):

$$\min_{\mathbf{\Lambda},\mathbf{\Psi}} ||(\mathbf{Z}^\top\mathbf{Z} - \mathbf{\Lambda}\mathbf{\Lambda}^\top - \mathbf{\Psi}^2)\mathbf{\Gamma}||_F^2 \ , \tag{3.21}$$

where $\mathbf{\Gamma}$ is a matrix of weights. The case $\mathbf{\Gamma} = \mathbf{\Theta}^{-1}$ is known as generalized least squares (GLS) factor analysis. If $\mathbf{\Gamma} = \mathbf{I}_p$, (3.21) reduces to an unweighted LS optimization problem. The standard numerical solutions of the optimization problems (3.20) and (3.21) are iterative, usually based on a Newton-Raphson procedure. Alternatively, the appropriate discrepancy function is minimized over one of the unknowns of the problem, keeping the other one fixed (Jöreskog, 1967, 1977; Lawley and Maxwell, 1971). An expectation maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977) for solving (3.20) was developed by Rubin and Thayer (1982).

Suppose that a pair $\{\mathbf{\Lambda}, \mathbf{\Psi}\}$ is obtained by solving the factor extraction problem stated above. Then, common factor scores can be computed as a function of $\mathbf{Z}$, $\mathbf{\Lambda}$ and possibly $\mathbf{\Psi}$ in a number of ways (e.g., Harman, 1976; Mulaik, 1972). The most popular one minimizes the discrepancies between the true and estimated factor scores in a least squares sense (Thurstone, 1935) and leads to linear regression of $\mathbf{F}$ on the data $\mathbf{Z}$, i.e.:

$$\mathbf{F}_R = \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{\Lambda} \; . \tag{3.22}$$

Equation (3.22) may be put in the alternative form (Ledermann, 1939):

$$\mathbf{F}_L = \mathbf{Z}\mathbf{\Psi}^{-2}\mathbf{\Lambda}(\mathbf{I}_k + \mathbf{\Lambda}^\top\mathbf{\Psi}^{-2}\mathbf{\Lambda})^{-1} \; , \tag{3.23}$$

which requires computing the inverse of a $k \times k$ matrix instead of a $p \times p$ matrix as in (3.22). An alternative to the regression method has been proposed by Bartlett (1937). It requires minimization of the sum of squares of the unique factors weighted by the reciprocal of their variances, i.e. it minimizes:

$$\text{trace}(\mathbf{Z} - \mathbf{F}\mathbf{\Lambda}^\top)\mathbf{\Psi}^{-2}(\mathbf{Z} - \mathbf{F}\mathbf{\Lambda}^\top)^\top \; , \tag{3.24}$$

which gives

$$\mathbf{F}_B = \mathbf{Z}\mathbf{\Psi}^{-2}\mathbf{\Lambda}(\mathbf{\Lambda}^\top\mathbf{\Psi}^{-2}\mathbf{\Lambda})^{-1} \; . \tag{3.25}$$

It seems rather disappointing to obtain non-orthogonal common factors $\mathbf{F}_R, \mathbf{F}_L$ and $\mathbf{F}_B$ when considering an EFA model with orthogonal factors. Orthogonality of the factor scores can be achieved for $\mathbf{F}_B$ if (3.24) is minimized subject to the constraint $\mathbf{F}^\top\mathbf{F} = \mathbf{I}_k$. This leads to the modification of Bartlett's factor scores proposed by Anderson and Rubin (1956):

$$\mathbf{F}_{AR} = \mathbf{Z}\Psi^{-2}\Lambda(\Lambda^\top\Psi^{-2}(\mathbf{Z}^\top\mathbf{Z})\Psi^{-2}\Lambda)^{-1/2} \, , \tag{3.26}$$

which satisfies the correlation-preserving constraint $\mathbf{F}_{AR}^\top\mathbf{F}_{AR} = \mathbf{I}_k$. Note that (3.26) and (3.24) are undefined if $\Psi$ is singular, a situation not uncommon in practice.

Strictly speaking, the term 'estimation' when applied to common and unique factors means that they cannot be identified uniquely, rather than obtaining them in a standard procedure for finding particular sample statistics. This form of indeterminacy is known as 'factor indeterminacy' (e.g., Mulaik, 1972, 2005). This indeterminacy is due to the fact that the EFA model postulates the existence of $k$ common and $p$ unique factors such that the $p$ observed variables can be represented as their linear combinations. Thus, the scores of the $n$ observations on the common and unique factors are not uniquely identifiable. Guttman (1955) showed that an infinite set of scores for the common and unique factors can be constructed satisfying the EFA model equation and its constraints (see also Kestelman, 1952). Following Guttman's approach and assuming that the common factors are orthogonal one can consider (Mulaik, 2005):

$$\mathbf{F}_G = \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\Lambda + \mathbf{S}\mathbf{G} \tag{3.27}$$

and

$$\mathbf{U}_G = \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\Psi - \mathbf{S}\mathbf{G}\Lambda^\top\Psi^{-1} \, , \tag{3.28}$$

where S is an arbitrary $n \times k$ matrix satisfying (McDonald, 1979):

$$\mathbf{S}^\top\mathbf{S} = \mathbf{I}_k \quad \text{and} \quad \mathbf{S}^\top\mathbf{Z} = \mathbf{O}_{p\times k} \ . \tag{3.29}$$

In other words, **S** is an orthonormal matrix orthogonal in $\mathbb{R}^n$ to the data **Z**, i.e. the subspace spanned by **S** is orthogonal to the subspace spanned by **Z**. One way to find such **S** is by the QR decomposition of **Z** (e.g., Golub and Van Loan, 1996):

$$\mathbf{Z} = \mathbf{QR} = [\mathbf{Q}_{n\times p} \ \ \mathbf{Q}_\perp]\mathbf{R} \ ,$$

where $\mathbf{Q} \in \mathbb{R}^{n\times n}$ is orthogonal, $\mathbf{R} \in \mathbb{R}^{n\times p}$ is upper triangular, the columns of $\mathbf{Q}_{n\times p}$ form an orthonormal basis for range(**Z**) and the columns of the $n \times (n-p)$ matrix $\mathbf{Q}_\perp$ form an orthonormal basis for null(**Z**). Then, **S** can be formed by taking any $k$ columns of $\mathbf{Q}_\perp$, assuming that $k < n - p$.

The matrix **G** in (3.27) and (3.28) is "any $k \times k$ Gram factor of the residual covariance matrix for the common factors after the parts of them predictable by linear regression from the observed variables have been partialed out" (Mulaik, 2005, p. 181), i.e.:

$$\mathbf{G}^\top\mathbf{G} = \mathbf{I}_k - \mathbf{F}_R^\top\mathbf{F}_R = \mathbf{I}_k - \mathbf{\Lambda}^\top(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{\Lambda} \ , \tag{3.30}$$

where $\mathbf{I}_k - \mathbf{\Lambda}^\top(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{\Lambda}$ is assumed positive semi-definite (McDonald, 1979).

One can check by direct substitution of (3.27) and (3.28) into (3.19) that the EFA model equation is satisfied. Also, according to the EFA model requirements the following properties are fulfilled: $\mathbf{F}_G^\top\mathbf{F}_G = \mathbf{I}_k$, $\mathbf{U}_G^\top\mathbf{U}_G = \mathbf{I}_p$ and $\mathbf{F}_G^\top\mathbf{U}_G = \mathbf{O}_{k\times p}$. For example:

$$\begin{aligned}
\mathbf{U}_G^\top\mathbf{U}_G &= \left[\mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{\Psi} - \mathbf{S}\mathbf{G}\mathbf{\Lambda}^\top\mathbf{\Psi}^{-1}\right]^\top \left[\mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{\Psi} - \mathbf{S}\mathbf{G}\mathbf{\Lambda}^\top\mathbf{\Psi}^{-1}\right] \\
&= \mathbf{\Psi}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{\Psi} + \mathbf{\Psi}^{-1}\mathbf{\Lambda}\mathbf{G}^\top\mathbf{G}\mathbf{\Lambda}^\top\mathbf{\Psi}^{-1} \\
&= \mathbf{\Psi}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{\Psi} + \mathbf{\Psi}^{-1}\mathbf{\Lambda}\mathbf{\Lambda}^\top\mathbf{\Psi}^{-1} - \mathbf{\Psi}^{-1}\mathbf{\Lambda}\mathbf{\Lambda}^\top(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{\Lambda}\mathbf{\Lambda}^\top\mathbf{\Psi}^{-1} \\
&= \mathbf{\Psi}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{\Psi} + \mathbf{\Psi}^{-1}\mathbf{\Lambda}\mathbf{\Lambda}^\top\mathbf{\Psi}^{-1} - \mathbf{\Psi}^{-1}(\mathbf{Z}^\top\mathbf{Z} - \mathbf{\Psi}^2)(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{\Lambda}\mathbf{\Lambda}^\top\mathbf{\Psi}^{-1} \\
&= \mathbf{\Psi}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{\Psi} + \mathbf{\Psi}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{\Lambda}\mathbf{\Lambda}^\top\mathbf{\Psi}^{-1} \\
&= \mathbf{I}_p\,.
\end{aligned}$$

Note that the expressions in (3.27) and (3.28) imply $\mathbf{Z}^\top\mathbf{F}_G = \mathbf{\Lambda}$ and $\mathbf{Z}^\top\mathbf{U}_G = \mathbf{\Psi}$ (and thus diagonal). Guttman's approach to finding factor scores $\mathbf{F}_G$ does not suffer from the common weakness of $\mathbf{F}_{AR}$ which requires $\mathbf{\Psi}$ being nonsingular. Unfortunately, this requirement is still needed to find the unique factors $\mathbf{U}_G$.

In the sequel, the common factors are assumed to be fixed population parameters and the EFA model is considered as a specific data matrix decomposition. Unlike factoring a correlation matrix, a decomposition of the data matrix yield simultaneous solutions for both loadings and factor scores. In the next Chapter, a review of procedures for fitting the EFA model with fixed common factors is presented.

# Part II

# EFA as Data Matrix Decomposition

# Chapter 4

# Fitting the fixed EFA model

Lawley (1942) introduced an EFA model in which both the common factors and the factor loadings are treated as fixed unknown quantities. To fit the EFA model with fixed common factors, Lawley (1942) proposed to maximize the log-likelihood of the data (see also Young, 1941):

$$\mathfrak{L}_1 = -\frac{n}{2}\left[\log(2\pi) + \log(\det(\Psi^2)) + \mathrm{trace}(\mathbf{Z} - \mathbf{F}\Lambda^\top)\Psi^{-2}(\mathbf{Z} - \mathbf{F}\Lambda^\top)^\top\right] \ . \qquad (4.1)$$

Instead of maximizing (4.1), one might try to minimize the function

$$\begin{aligned}\mathfrak{L}_2 &= \frac{1}{n}\mathfrak{L}_1 + \frac{1}{2}\log(2\pi) \ , \\ &= \frac{1}{2}\left[\log(\det(\Psi^2)) + \mathrm{trace}(\mathbf{Z} - \mathbf{F}\Lambda^\top)\Psi^{-2}(\mathbf{Z} - \mathbf{F}\Lambda^\top)^\top\right] \ . \qquad (4.2)\end{aligned}$$

Anderson and Rubin (1956) showed that the fixed EFA model cannot be fitted to the data by the standard maximum likelihood approach as the corresponding log-likelihood loss function (4.2) to be minimized is unbounded below. Hence, maximum-likelihood estimators do not exist for the fixed EFA model.

Attempts to find estimators for loadings and factor scores based on the likelihood have persisted (Whittle, 1952; Jöreskog, 1962), based partly on the conjecture that the loadings for the fixed EFA model would resemble those of the random EFA model (Basilevsky, 1994). McDonald (1979) circumvented the difficulty noted by Anderson and Rubin (1956) in the original treatment of the fixed EFA model by Lawley (1942). He proposed to minimize the logarithm of the ratio of the likelihood under the hypothesized model to the likelihood under the alternative hypothesis that the error covariance matrix is any positive definite matrix:

$$\mathfrak{L}_3 = \frac{1}{2}\left[\log(\det(\mathrm{diag}(\mathbf{E}^\top\mathbf{E}))) - \log(\det(\mathbf{E}^\top\mathbf{E}))\right] \quad , \tag{4.3}$$

where $\mathbf{E} = \mathbf{Z} - \mathbf{F}\mathbf{\Lambda}^\top$. McDonald (1979) showed that (4.3) is bounded below by zero, a bound which is attained only if $\mathbf{E}^\top\mathbf{E}$ is diagonal. Thus, minimizing (4.3) yields maximum-likelihood-ratio estimators (see also Etezadi-Amoli and McDonald, 1983). Moreover, McDonald (1979) proved that the likelihood-based estimators of the factor loadings and uniquenesses are the same as in the random EFA model, while estimators of the common factor scores are the same as the arbitrary solutions given by Guttman (1955) discussed in the previous Chapter.

McDonald (1979) also studied LS fitting of the fixed EFA model. Consider the following objective function to be minimized:

$$\mathcal{F}_{McD}(\mathbf{F}, \mathbf{\Lambda}, \mathbf{\Psi}) = ||(\mathbf{Z} - \mathbf{F}\mathbf{\Lambda}^\top)^\top(\mathbf{Z} - \mathbf{F}\mathbf{\Lambda}^\top) - \mathbf{\Psi}^2||_F^2 \, . \tag{4.4}$$

Unlike the log-likelihood loss function (4.2), the LS loss function (4.4) is bounded below (Golub and Van Loan, 1996, p. 605). McDonald (1979) showed that the parameter estimates found by minimizing (4.4) can be compared to the standard EFA least squares estimates (with random common factors) obtained by minimizing (e.g. Jöreskog, 1977):

$$\mathcal{F}_{LS}(\mathbf{\Lambda}, \mathbf{\Psi}) = ||\mathbf{Z}^\top\mathbf{Z} - \mathbf{\Lambda}\mathbf{\Lambda}^\top - \mathbf{\Psi}^2||_F^2 \, . \tag{4.5}$$

Indeed, the gradients of $\mathcal{F}_{LS}(\mathbf{\Lambda}, \mathbf{\Psi})$ with respect to the unknowns $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ are (for convenience the objective function (4.5) is multiplied by .25):

$$\begin{aligned} \nabla_{\mathbf{\Lambda}}^{LS} &= -(\mathbf{Z}^\top\mathbf{Z} - \mathbf{\Lambda}\mathbf{\Lambda}^\top - \mathbf{\Psi}^2)\mathbf{\Lambda} \ , \\ \nabla_{\mathbf{\Psi}}^{LS} &= -[\mathrm{diag}(\mathbf{Z}^\top\mathbf{Z} - \mathbf{\Lambda}\mathbf{\Lambda}^\top) - \mathbf{\Psi}^2]\mathbf{\Psi} \ . \end{aligned}$$

McDonald (1979) found that the gradients of $\mathcal{F}_{McD}(\mathbf{\Lambda}, \mathbf{\Psi}, \mathbf{F})$ with respect to the unknowns $\mathbf{\Lambda}, \mathbf{\Psi}$ and $\mathbf{F}$ can be written as (for convenience the objective function (4.4) is multiplied by .25):

$$\begin{aligned} \nabla_{\mathbf{\Lambda}}^{McD} &= -[(\mathbf{Z} - \mathbf{F}\mathbf{\Lambda}^\top)^\top(\mathbf{Z} - \mathbf{F}\mathbf{\Lambda}^\top) - \mathbf{\Psi}^2](\mathbf{Z} - \mathbf{F}\mathbf{\Lambda}^\top)^\top\mathbf{F} \ , \\ \nabla_{\mathbf{\Psi}}^{McD} &= -\mathrm{diag}((\mathbf{Z} - \mathbf{F}\mathbf{\Lambda}^\top)^\top(\mathbf{Z} - \mathbf{F}\mathbf{\Lambda}^\top) - \mathbf{\Psi}^2)\mathbf{\Psi} \ , \\ \nabla_{\mathbf{F}}^{McD} &= -(\mathbf{Z} - \mathbf{F}\mathbf{\Lambda}^\top)[(\mathbf{Z} - \mathbf{F}\mathbf{\Lambda}^\top)^\top(\mathbf{Z} - \mathbf{F}\mathbf{\Lambda}^\top) - \mathbf{\Psi}^2]\mathbf{\Lambda} \ . \end{aligned}$$

The values of the gradients are then calculated at $\mathbf{F} = \mathbf{F}_G$ from (3.27):

$$\begin{aligned} \nabla_{\mathbf{\Lambda}}^{McD} &= -(\mathbf{Z}^\top\mathbf{Z} - \mathbf{\Lambda}\mathbf{\Lambda}^\top - \mathbf{\Psi}^2)(\mathbf{Z} - \mathbf{F}_G\mathbf{\Lambda}^\top)^\top\mathbf{F}_G = \mathbf{O}_{p\times k} \ , \\ \nabla_{\mathbf{\Psi}}^{McD} &= -[\mathrm{diag}(\mathbf{Z}^\top\mathbf{Z} - \mathbf{\Lambda}\mathbf{\Lambda}^\top) - \mathbf{\Psi}^2]\mathbf{\Psi} = \nabla_{\mathbf{\Psi}}^{LS} \ , \\ \nabla_{\mathbf{F}}^{McD} &= -(\mathbf{Z} - \mathbf{F}_G\mathbf{\Lambda}^\top)(\mathbf{Z}^\top\mathbf{Z} - \mathbf{\Lambda}\mathbf{\Lambda}^\top - \mathbf{\Psi}^2)\mathbf{\Lambda} = (\mathbf{Z} - \mathbf{F}_G\mathbf{\Lambda}^\top)\nabla_{\mathbf{\Lambda}}^{LS} \ . \end{aligned}$$

While calculating the gradients of $\mathcal{F}_{McD}(\mathbf{\Lambda}, \mathbf{\Psi}, \mathbf{F})$ at $\mathbf{F} = \mathbf{F}_G$ one simply makes use of the features $\mathbf{F}_G^\top\mathbf{F}_G = \mathbf{I}_k$ and $\mathbf{Z}^\top\mathbf{F}_G = \mathbf{\Lambda}$. Of course, any other common factors $\mathbf{F}$ satisfying these conditions and $\mathbf{F}^\top\mathbf{U} = \mathbf{O}_{k\times p}$ would produce the same results.

Thus, McDonald (1979) established that the LS approach for fitting the fixed EFA model gives a minimum of the loss function as well as estimators of the factor loadings and uniquenesses which are the same as the corresponding ones in the random EFA model. The estimators of the common factor scores are the same as those given by the expressions of Guttman (1955).

A LS procedure for finding the matrix of common factor scores $\mathbf{F}$ is also outlined in Horst (1965). He wrote: "Having given some arbitrary factor loading matrix, whether centroid, multiple group, or principal axis, we may wish to determine that factor score matrix which, when post-multiplied by the transpose of the factor loading matrix, yields a product which is the least squares approximation to the data matrix. This means that the sums of squares of elements of the residual matrix will be a minimum." (Horst, 1965, p. 471). Following this strategy, the suggested LS factor score matrix is sought to minimize

$$\mathcal{F}_H(\mathbf{F}) = ||\mathbf{Z} - \mathbf{F}\Lambda^\top||_F^2 \ , \tag{4.6}$$

which is simply given by

$$\mathbf{F}_H = \mathbf{Z}\Lambda(\Lambda^\top\Lambda)^{-1} \tag{4.7}$$

for an arbitrary factor loading matrix $\Lambda$ (Horst, 1965, p. 479).

Horst (1965) also proposed a rank reduction algorithm for factoring a data matrix $\mathbf{Z}$. For some starting approximation $\Lambda_0$ of the factor loadings, let $\mathbf{L}_0$ be the $k \times k$ lower triangular matrix obtained from the Cholesky decomposition (e.g., Golub and Van Loan, 1996) $\mathbf{L}_0\mathbf{L}_0^\top$ of $\Lambda_0^\top\mathbf{Z}^\top\mathbf{Z}\Lambda_0$. Then, the successive approximation $\Lambda_1$ of the factor loadings is found as (Horst, 1965, p. 274):

$$\Lambda_1 = \mathbf{Z}^\top\mathbf{Z}\Lambda_0(\mathbf{L}_0^\top)^{-1} \ .$$

It follows from

$$\mathbf{Z}^\top\mathbf{Z} - \Lambda_1\Lambda_1^\top = \mathbf{Z}^\top\mathbf{Z} - \mathbf{Z}^\top\mathbf{Z}\Lambda_0(\Lambda_0^\top\mathbf{Z}^\top\mathbf{Z}\Lambda_0)^{-1}\Lambda_0^\top\mathbf{Z}^\top\mathbf{Z} \ ,$$

that the successive approximation $\Lambda_1$ is always a rank reducing matrix for $\mathbf{Z}^\top\mathbf{Z}$. After convergence of the algorithm, the final $\Lambda$ found is the matrix of factor loadings. The common factor scores are obtained as $\mathbf{F} = \mathbf{Z}\mathbf{\Lambda}\text{diag}(\mathbf{\Lambda}^\top\mathbf{Z}^\top\mathbf{Z}\mathbf{\Lambda})^{-1/2}$, i.e. $\mathbf{F}$ is an oblique matrix with $\text{diag}(\mathbf{F}^\top\mathbf{F}) = \mathbf{I}_k$.

Both algorithms in Horst (1965) find a pair $\{\mathbf{\Lambda}, \mathbf{F}\}$. No care is taken to obtain unique factor scores $\mathbf{U}$ or the uniquenesses $\mathbf{\Psi}$. In this sense, the proposed procedures resemble PCA rather more than EFA.

# Chapter 5

# Simultaneous Estimation of all EFA Model Unknowns

In formulating EFA models with random or fixed common factors, the standard approach is to embed the data in a replication framework by assuming the observations are realizations of random variables. In the sequel, the EFA model is formulated directly in terms of the data instead and all model unknowns $\mathbf{\Lambda}, \mathbf{\Psi}, \mathbf{F}$ and $\mathbf{U}$ are assumed to be fixed matrix parameters.

For $n > p$, De Leeuw (2004, 2008) proposed to minimize the following LS loss function:

$$\mathcal{F}_{DeL}(\mathbf{\Lambda}, \mathbf{\Psi}, \mathbf{F}, \mathbf{U}) = \left\| \mathbf{Z} - [\mathbf{F} \quad \mathbf{U}] \begin{bmatrix} \mathbf{\Lambda}^\top \\ \mathbf{\Psi} \end{bmatrix} \right\|_F^2 , \tag{5.1}$$

subject to $\mathrm{rank}(\mathbf{\Lambda}) = k$, $\mathbf{F}^\top\mathbf{F} = \mathbf{I}_k$, $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_p$, $\mathbf{U}^\top\mathbf{F} = \mathbf{O}_{p\times k}$ and $\mathbf{\Psi}$ being a $p \times p$ diagonal matrix. Minimizing (5.1) amounts to minimizing the sum of the squares of the residuals defined as the differences between the observed and predicted standardized values of the scores on the variables. The loss function $\mathcal{F}_{DeL}$ defined in (5.1) is bounded below (Golub and Van Loan, 1996, p. 605). The idea is that for given or estimated $\mathbf{\Lambda}$ and $\mathbf{\Psi}$, the common and unique factor scores $\mathbf{F}$ and $\mathbf{U}$ can be found as a solution of a Procrustes problem (De Leeuw, 2004, 2008).

The EFA model (3.19) and the imposed constraints imply the following identities:

$$\mathbf{F}^\top\mathbf{Z} = \mathbf{F}^\top\mathbf{F}\mathbf{\Lambda}^\top + \mathbf{F}^\top\mathbf{U}\mathbf{\Psi} \Longrightarrow \mathbf{F}^\top\mathbf{Z} = \mathbf{\Lambda}^\top, \tag{5.2}$$

$$\mathbf{U}^\top\mathbf{Z} = \mathbf{U}^\top\mathbf{F}\mathbf{\Lambda}^\top + \mathbf{U}^\top\mathbf{U}\mathbf{\Psi} \Longrightarrow \mathbf{U}^\top\mathbf{Z} = \mathbf{\Psi} \text{ (and thus diagonal) }. \tag{5.3}$$

The identities (5.2) and (5.3) also imply that

$$\begin{aligned}(\mathbf{I}_n - \mathbf{F}\mathbf{F}^\top)\mathbf{Z} &= \mathbf{Z} - \mathbf{F}\mathbf{F}^\top\mathbf{Z} = \mathbf{Z} - \mathbf{F}\mathbf{F}^\top(\mathbf{F}\mathbf{\Lambda}^\top + \mathbf{U}\mathbf{\Psi})\ , \\ &= \mathbf{Z} - \mathbf{F}\mathbf{F}^\top\mathbf{F}\mathbf{\Lambda}^\top - \mathbf{F}\mathbf{F}^\top\mathbf{U}\mathbf{\Psi} = \mathbf{Z} - \mathbf{F}\mathbf{\Lambda}^\top = \mathbf{U}\mathbf{\Psi} \end{aligned} \tag{5.4}$$

and thus

$$\mathbf{Z}^\top(\mathbf{I}_n - \mathbf{F}\mathbf{F}^\top)\mathbf{Z} = \mathbf{\Psi}^2\ . \tag{5.5}$$

Any proper EFA solution should fulfil (5.2) – (5.4) and most likely they would appear as optimality conditions of the problem.

## 5.1 Dynamical system approach and optimality conditions

Before discussing efficient methods for minimizing $\mathcal{F}_{DeL}$ in (5.1), such a solution for simultaneous estimation of all EFA matrix parameters $\{\mathbf{\Lambda}, \mathbf{\Psi}, \mathbf{F}, \mathbf{U}\}$ can be explored by making use of the continuous-time projected gradient approach as described in Chapter 2. It does not rely on an alternating solution for certain parameters while the rest are kept fixed.

The gradients of $\mathcal{F}_{DeL}$ with respect to the unknowns $\mathbf{\Lambda}$, $\mathbf{\Psi}$, $\mathbf{F}$ and $\mathbf{U}$ are (for convenience the objective function (5.1) is multiplied by .5):

$$\begin{aligned}
\nabla_{\mathbf{\Lambda}}^{DeL} &= -(\mathbf{Z} - \mathbf{F}\mathbf{\Lambda}^{\top} - \mathbf{U}\mathbf{\Psi})^{\top}\mathbf{F} \ , \\
\nabla_{\mathbf{\Psi}}^{DeL} &= -\mathrm{diag}\left((\mathbf{Z} - \mathbf{F}\mathbf{\Lambda}^{\top} - \mathbf{U}\mathbf{\Psi})^{\top}\mathbf{U}\right) \ , \\
\nabla_{\mathbf{F}}^{DeL} &= -(\mathbf{Z} - \mathbf{F}\mathbf{\Lambda}^{\top} - \mathbf{U}\mathbf{\Psi})\mathbf{\Lambda} \ , \\
\nabla_{\mathbf{U}}^{DeL} &= -(\mathbf{Z} - \mathbf{F}\mathbf{\Lambda}^{\top} - \mathbf{U}\mathbf{\Psi})\mathbf{\Psi} \ .
\end{aligned}$$

First order optimality conditions are readily available if the gradients of $\mathcal{F}_{DeL}$ are set equal to zero. Somehow more informative first order conditions for the existence of the minimizers of $\mathcal{F}_{DeL}$ can be obtained by applying the dynamical system approach.

By projecting the gradients of $\mathcal{F}_{DeL}$ onto the corresponding constrained manifolds of $\mathbf{\Lambda}, \mathbf{\Psi}, \mathbf{F}$ and $\mathbf{U}$, respectively, the following matrix ODEs of first order are obtained:

$$\dot{\mathbf{\Lambda}} = .5\mathbf{\Lambda}(\mathbf{\Lambda}^{\top}\mathbf{Z}^{\top}\mathbf{F} - \mathbf{F}^{\top}\mathbf{Z}\mathbf{\Lambda}) + (\mathbf{I}_p - \mathbf{\Lambda}(\mathbf{\Lambda}^{\top}\mathbf{\Lambda})^{-1}\mathbf{\Lambda}^{\top})\mathbf{Z}^{\top}\mathbf{F} \ , \tag{5.6}$$

$$\dot{\mathbf{\Psi}} = \mathrm{diag}(\mathbf{Z}^{\top}\mathbf{U}) - \mathbf{\Psi} \ , \tag{5.7}$$

$$\dot{\mathbf{F}} = .5\mathbf{F}(\mathbf{F}^{\top}\mathbf{Z}\mathbf{\Lambda} - \mathbf{\Lambda}^{\top}\mathbf{Z}^{\top}\mathbf{F}) + (\mathbf{I}_n - \mathbf{F}\mathbf{F}^{\top})(\mathbf{Z} - \mathbf{U}\mathbf{\Psi})\mathbf{\Lambda} \ , \tag{5.8}$$

$$\dot{\mathbf{U}} = .5\mathbf{U}(\mathbf{U}^{\top}\mathbf{Z}\mathbf{\Psi} - \mathbf{\Psi}\mathbf{Z}^{\top}\mathbf{U}) + (\mathbf{I}_n - \mathbf{U}\mathbf{U}^{\top})(\mathbf{Z} - \mathbf{F}\mathbf{\Lambda}^{\top})\mathbf{\Psi} \ . \tag{5.9}$$

The dynamical system (5.6) – (5.9) governs simultaneous steepest descent flows for $\mathbf{\Lambda}, \mathbf{\Psi}, \mathbf{F}$ and $\mathbf{U}$ leading to the minimum of $\mathcal{F}_{DeL}$. The right hand side of equation (5.6) is the projection of $\nabla_{\mathbf{\Lambda}}^{DeL}$ onto the non-compact Stiefel manifold of all $p \times k$ matrices with rank exactly $k$.

Let $\mathcal{D}(p)$ denote the linear subspace of all $p \times p$ diagonal matrices. Using the fact that the tangent space of $\mathcal{D}(p)$ is $\mathcal{D}(p)$ itself, the right hand side of equation (5.7) is the projection of $\nabla_{\mathbf{\Psi}}^{DeL}$ onto $\mathcal{D}(p)$.

The right hand sides of the equations (5.8) and (5.9) are the corresponding projections of $\nabla_{\mathbf{F}}^{DeL}$ and $\nabla_{\mathbf{U}}^{DeL}$ onto the compact Stiefel manifolds of all $n \times k$ and $n \times p$ column-wise orthonormal matrices, respectively.

For more accurate estimation of $\mathbf{\Lambda}, \mathbf{\Psi}, \mathbf{F}$ and $\mathbf{U}$ one can consider alternatively the following system of three matrix ODEs:

$$\dot{\mathbf{\Lambda}} = .5\mathbf{\Lambda}(\mathbf{\Lambda}^\top\mathbf{Z}^\top\mathbf{F} - \mathbf{F}^\top\mathbf{Z}\mathbf{\Lambda}) + (\mathbf{I}_p - \mathbf{\Lambda}(\mathbf{\Lambda}^\top\mathbf{\Lambda})^{-1}\mathbf{\Lambda}^\top)\mathbf{Z}^\top\mathbf{F} \quad , \tag{5.10}$$

$$\dot{\mathbf{\Psi}} = \mathrm{diag}(\mathbf{Z}^\top\mathbf{U}) - \mathbf{\Psi} \ , \tag{5.11}$$

$$\dot{\mathbf{B}} = .5\mathbf{B}(\mathbf{B}^\top\mathbf{Z}\mathbf{A} - \mathbf{A}^\top\mathbf{Z}^\top\mathbf{B}) + (\mathbf{I}_n - \mathbf{B}\mathbf{B}^\top)\mathbf{Z}\mathbf{A} \ , \tag{5.12}$$

where $\mathbf{B} := [\mathbf{F} \quad \mathbf{U}]$ and $\mathbf{A} := [\mathbf{\Lambda} \quad \mathbf{\Psi}]$ are block matrices with dimensions $n \times (k+p)$ and $p \times (k+p)$, respectively, and $\dot{\mathbf{B}}$ is the projection of the gradient of $\mathcal{F}_{DeL}$ with respect to $\mathbf{B}$ onto the compact Stiefel manifold of all $n \times (k+p)$ orthonormal matrices. The problem with this approach is that the projection onto the non-compact Stiefel manifold (5.10) involves computation of the inverse matrix $(\mathbf{\Lambda}^\top\mathbf{\Lambda})^{-1}$ which may not be efficient and/or stable. To avoid this, using (5.2), one can replace (5.10) simply by

$$\dot{\mathbf{\Lambda}} = \mathbf{Z}^\top\dot{\mathbf{F}} \ ,$$

as $\dot{\mathbf{F}}$ should be calculated in (5.12) anyway.

By means of (5.10) – (5.12), the following first order optimality conditions for the existence of the minimizers of $\mathcal{F}_{DeL}$ are obtained:

$$\mathbf{\Lambda}^\top\mathbf{Z}^\top\mathbf{F} \text{ must be symmetric } , \tag{5.13}$$

$$\mathbf{Z}^\top\mathbf{F} = \mathbf{\Lambda}(\mathbf{\Lambda}^\top\mathbf{\Lambda})^{-1}\mathbf{\Lambda}^\top\mathbf{Z}^\top\mathbf{F} \ , \tag{5.14}$$

$$\mathbf{\Psi} = \mathrm{diag}(\mathbf{Z}^\top\mathbf{U}) \ , \tag{5.15}$$

$$\mathbf{B}^\top\mathbf{Z}\mathbf{A} = \mathbf{A}^\top\mathbf{Z}^\top\mathbf{B} \ , \tag{5.16}$$

$$(\mathbf{I}_n - \mathbf{B}\mathbf{B}^\top)\mathbf{Z}\mathbf{A} = \mathbf{O}_{n\times(p+k)} \ . \tag{5.17}$$

Condition (5.14) states that $\mathbf{Z}^\top\mathbf{F}$ is entirely in the range of $\Lambda$, i.e. there exists a nonzero $k \times k$ matrix $\mathbf{G}$ such that $\mathbf{Z}^\top\mathbf{F} = \Lambda\mathbf{G}$. Then, by making use of (5.13) one finds that $\Lambda^\top\Lambda\mathbf{G} = \mathbf{G}^\top\Lambda^\top\Lambda$ which can be true only if $\mathbf{G}$ is diagonal with equal entries.

The optimality condition (5.16) implies that $\mathbf{B}^\top\mathbf{Z}\mathbf{A}$ must be a symmetric matrix. Writing this in detail gives that

$$\begin{bmatrix} \mathbf{F}^\top \\ \mathbf{U}^\top \end{bmatrix} \mathbf{Z}[\Lambda \quad \Psi] = \begin{bmatrix} \mathbf{F}^\top\mathbf{Z}\Lambda & \mathbf{F}^\top\mathbf{Z}\Psi \\ \mathbf{U}^\top\mathbf{Z}\Lambda & \mathbf{U}^\top\mathbf{Z}\Psi \end{bmatrix} \tag{5.18}$$

must be a symmetric matrix. This leads to the following first order optimality conditions for $\mathbf{F}$ and $\mathbf{U}$:

$$\mathbf{F}^\top\mathbf{Z}\Lambda = \Lambda^\top\mathbf{Z}^\top\mathbf{F} \ , \tag{5.19}$$

$$\mathbf{U}^\top\mathbf{Z}\Psi = \Psi\mathbf{Z}^\top\mathbf{U} \ , \tag{5.20}$$

$$\mathbf{U}^\top\mathbf{Z}\Lambda = \Psi\mathbf{Z}^\top\mathbf{F} \ . \tag{5.21}$$

Condition (5.19) is equivalent to (5.13). Condition (5.20) and $\Psi$ being diagonal with different entries imply that $\mathbf{U}^\top\mathbf{Z}$ must be diagonal and using (5.15): $\Psi = \mathbf{U}^\top\mathbf{Z}$. Condition (5.21) enforces that $\Lambda = \mathbf{Z}^\top\mathbf{F}$. The optimality condition (5.17) implies that

$$(\mathbf{I}_n - \mathbf{F}\mathbf{F}^\top - \mathbf{U}\mathbf{U}^\top)\mathbf{Z}\mathbf{A} = \mathbf{O}_{n\times(p+k)} \ ,$$

that is, the linear subspace spanned by $\mathbf{Z}\mathbf{A} = [\mathbf{Z}\Lambda \quad \mathbf{Z}\Psi]$ is orthogonal in $\mathbb{R}^n$ to the linear subspace spanned by $[\mathbf{F} \quad \mathbf{U}]$. Writing this in detail gives

$$(\mathbf{I}_n - \mathbf{F}\mathbf{F}^\top - \mathbf{U}\mathbf{U}^\top)\mathbf{Z}\Lambda = \mathbf{O}_{n\times k} \tag{5.22}$$

and simultaneously

$$(\mathbf{I}_n - \mathbf{F}\mathbf{F}^\top - \mathbf{U}\mathbf{U}^\top)\mathbf{Z}\Psi = \mathbf{O}_{n\times p} \ . \tag{5.23}$$

Applying $\mathbf{\Psi} = \mathbf{U}^\top\mathbf{Z}$, (5.22) turns into

$$[(\mathbf{I}_n - \mathbf{F}\mathbf{F}^\top)\mathbf{Z} - \mathbf{U}\mathbf{\Psi}]\mathbf{\Lambda} = \mathbf{O}_{n\times k} \ , \tag{5.24}$$

which generalizes the identity (5.4). By making use of $\mathbf{\Lambda}^\top = \mathbf{F}^\top\mathbf{Z}$ and $\mathbf{\Psi} = \mathbf{U}^\top\mathbf{Z}$, (5.24) turns into

$$(\mathbf{Z} - \mathbf{F}\mathbf{\Lambda}^\top - \mathbf{U}\mathbf{\Psi})\mathbf{\Lambda} = \mathbf{O}_{n\times k} \tag{5.25}$$

and (5.23) turns into

$$(\mathbf{Z} - \mathbf{F}\mathbf{\Lambda}^\top - \mathbf{U}\mathbf{\Psi})\mathbf{\Psi} = \mathbf{O}_{n\times p} \ . \tag{5.26}$$

If $\mathbf{\Psi}$ is nonsingular, (5.26) leads to: $\mathbf{Z} = \mathbf{F}\mathbf{\Lambda}^\top + \mathbf{U}\mathbf{\Psi}$. Combining (5.21) with (5.19) implies that at the minimum of $\mathcal{F}_{DeL}$:

$$(\mathbf{F} - \mathbf{U}\mathbf{\Psi}^{-1}\mathbf{\Lambda})^\top\mathbf{Z}\mathbf{\Lambda} = \mathbf{O}_{n\times k} \ , \tag{5.27}$$

which means that $\mathbf{F} - \mathbf{U}\mathbf{\Psi}^{-1}\mathbf{\Lambda}$ belongs to null$(\mathbf{\Lambda}^\top\mathbf{Z}^\top)$. For an arbitrary $n \times k$ matrix $\mathbf{\Omega}$ one can express $\mathbf{F}$ as

$$\mathbf{F} = \mathbf{U}\mathbf{\Psi}^{-1}\mathbf{\Lambda} + (\mathbf{I}_n - \mathbf{Z}\mathbf{\Lambda}(\mathbf{\Lambda}^\top\mathbf{Z}^\top\mathbf{Z}\mathbf{\Lambda})^{-1}\mathbf{\Lambda}^\top\mathbf{Z}^\top)\mathbf{\Omega} \ . \tag{5.28}$$

Premultiplying $\mathbf{F}$ by its transpose, (5.28) turns into

$$\mathbf{F}^\top\mathbf{F} = \mathbf{I}_k = \mathbf{\Lambda}^\top\mathbf{\Psi}^{-2}\mathbf{\Lambda} + \ldots \ , \tag{5.29}$$

which implies that, in general, $\mathbf{\Lambda}^\top\mathbf{\Psi}^{-2}\mathbf{\Lambda}$ is not a diagonal matrix as is assumed in ML factor analysis (e.g., Mulaik, 1972). The condition (5.27) is considerably more general than the assumption $\mathbf{\Lambda}^\top\mathbf{\Psi}^{-2}\mathbf{\Lambda}$ being a diagonal matrix.

The factor loadings $\mathbf{\Lambda}$ can be any $p \times k$ matrix of full column rank. For example, any matrix $\mathbf{\Lambda V}$, where $\mathbf{V}$ is an arbitrary $k \times k$ orthogonal matrix, gives the same model fit if one compensates for this rotation in the scores. For interpretational reasons and to avoid this rotational indeterminacy the property $\text{rank}(\mathbf{\Lambda}) = k$ can be accomplished by having $\mathbf{\Lambda}$ in the form of a $p \times k$ lower triangular matrix $\mathbf{L}$, with a triangle of $k(k-1)/2$ zeros (Unkel and Trendafilov, 2009a). Consider the following slightly modified loss function:

$$\mathcal{F}_{DeL}(\mathbf{L}, \mathbf{\Psi}, \mathbf{F}, \mathbf{U}) = \left\| \mathbf{Z} - [\mathbf{F} \quad \mathbf{U}] \begin{bmatrix} \mathbf{L}^\top \\ \mathbf{\Psi} \end{bmatrix} \right\|_F^2 , \tag{5.30}$$

which should be minimized subject to the usual EFA constraints ($\mathbf{F}^\top\mathbf{F} = \mathbf{I}_k, \mathbf{U}^\top\mathbf{U} = \mathbf{I}_p, \mathbf{U}^\top\mathbf{F} = \mathbf{O}_{p\times k}$ and $\mathbf{\Psi}$ being a $p \times p$ diagonal matrix).

Let $\mathcal{L}(p,k)$ denote the linear subspace of all $p \times k$ lower triangular matrices and note that the tangent space of $\mathcal{L}(p,k)$ is $\mathcal{L}(p,k)$ itself. Then, the parameter matrices $\mathbf{L}, \mathbf{\Psi}, \mathbf{F}$ and $\mathbf{U}$ that minimize (5.30) can be found by integrating the following matrix ODEs simultaneously:

$$\dot{\mathbf{L}} = \mathtt{tril}(\mathbf{Z}^\top\mathbf{F}) - \mathbf{L} , \tag{5.31}$$

$$\dot{\mathbf{\Psi}} = \text{diag}(\mathbf{Z}^\top\mathbf{U}) - \mathbf{\Psi} , \tag{5.32}$$

$$\dot{\mathbf{B}} = .5\mathbf{B}(\mathbf{B}^\top\mathbf{ZA} - \mathbf{A}^\top\mathbf{Z}^\top\mathbf{B}) + (\mathbf{I}_n - \mathbf{BB}^\top)\mathbf{ZA} , \tag{5.33}$$

where `tril()` is the operator taking the lower triangular part of its argument, that is, $\mathtt{tril}(\mathbf{Z}^\top\mathbf{F})$ is composed of the elements of $\mathbf{Z}^\top\mathbf{F}$ with the upper triangle of $k(k-1)/2$ elements replaced by zeros. Note that (5.32) and (5.33) are equivalent to (5.11) and (5.12), respectively.

Thus, the first order optimality conditions for $\mathbf{L}, \mathbf{\Psi}, \mathbf{F}$ and $\mathbf{U}$ that minimize $\mathcal{F}_{DeL}$ in

(5.30) are (5.19)–(5.23) with $\mathbf{\Lambda}$ replaced by $\mathbf{L}$ and

$$\begin{aligned}\mathbf{L} &= \texttt{tril}(\mathbf{Z}^\top\mathbf{F}) \ , \\ \mathbf{\Psi} &= \text{diag}(\mathbf{Z}^\top\mathbf{U}) \ .\end{aligned}$$

Of course, if $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ are available from a standard EFA solution, finding a block-orthonormal matrix $[\mathbf{F} \quad \mathbf{U}]$ is just another method for factor score estimation.

## 5.2 Iterative algorithms

De Leeuw (2004, 2008) proposes an algorithm to optimize (5.1) that finds $\mathbf{F}$ and $\mathbf{U}$ simultaneously by solving an augmented Procrustes problem for the block-orthonormal matrix $[\mathbf{F} \quad \mathbf{U}]$. By making use of the block matrices $\mathbf{B} = [\mathbf{F} \quad \mathbf{U}]$ and $\mathbf{A} = [\mathbf{\Lambda} \quad \mathbf{\Psi}]$ defined in the previous Section, (5.1) can be rewritten as

$$\mathcal{F}_{DeL} = \left|\left|\mathbf{Z} - \mathbf{B}\mathbf{A}^\top\right|\right|_F^2 = ||\mathbf{Z}||_F^2 + \text{trace}(\mathbf{A}\mathbf{B}^\top\mathbf{B}\mathbf{A}^\top) - 2\,\text{trace}(\mathbf{B}^\top\mathbf{Z}\mathbf{A}) \ . \qquad (5.34)$$

Thus, as with the standard Procrustes problem (Gower and Dijksterhuis, 2004; Golub and Van Loan, 1996), the minimization of $\mathcal{F}_{DeL}$ in (5.34) subject to

$$\mathbf{B}^\top\mathbf{B} = \begin{bmatrix} \mathbf{F}^\top \\ \mathbf{U}^\top \end{bmatrix} \begin{bmatrix} \mathbf{F} & \mathbf{U} \end{bmatrix} = \begin{bmatrix} \mathbf{F}^\top\mathbf{F} & \mathbf{F}^\top\mathbf{U} \\ \mathbf{U}^\top\mathbf{F} & \mathbf{U}^\top\mathbf{U} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_k & \mathbf{O}_{k\times p} \\ \mathbf{O}_{p\times k} & \mathbf{I}_p \end{bmatrix} = \mathbf{I}_{k+p} \ ,$$

is equivalent to the maximization of $\text{trace}(\mathbf{B}^\top\mathbf{Z}\mathbf{A})$ (for given or estimated $\mathbf{A}$).

Assume that $n > p + k$ and let

$$\mathbf{Y} = \mathbf{Z}\mathbf{A} = \begin{bmatrix} \mathbf{Z}\mathbf{\Lambda} & \mathbf{Z}\mathbf{\Psi} \\ n\times k & n\times p \end{bmatrix} \ . \qquad (5.35)$$

The SVD of $\mathbf{Y}$ can be expressed in the form $\mathbf{Y} = \mathbf{VDT}^\top$, where $\mathbf{V}, \mathbf{T}$ and $\mathbf{D}$ are partitioned as follows:

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_1 & \mathbf{V}_2 \\ n \times k & n \times p \end{bmatrix}, \mathbf{T} = \begin{bmatrix} \mathbf{T}_{11} & \mathbf{T}_{12} \\ k \times k & k \times p \\ \mathbf{T}_{21} & \mathbf{T}_{22} \\ p \times k & p \times p \end{bmatrix}, \mathbf{D} = \begin{bmatrix} \mathbf{D}_{11} & \mathbf{O}_{k\times p} \\ \mathbf{O}_{p\times k} & \mathbf{D}_{22} \end{bmatrix}. \tag{5.36}$$

Then, the minimum of $\mathcal{F}_{DeL}$ in (5.34) is achieved by

$$\begin{aligned} \mathbf{B} &= \mathbf{VT}^\top = [\mathbf{V}_1 \;\; \mathbf{V}_2] \begin{bmatrix} \mathbf{T}_{11} & \mathbf{T}_{12} \\ \mathbf{T}_{21} & \mathbf{T}_{22} \end{bmatrix}^\top, \\ &= [\mathbf{V}_1\mathbf{T}_{11}^\top + \mathbf{V}_2\mathbf{T}_{12}^\top \;\; \mathbf{V}_1\mathbf{T}_{21}^\top + \mathbf{V}_2\mathbf{T}_{22}^\top], \end{aligned} \tag{5.37}$$

which means that

$$\mathbf{F} = \mathbf{V}_1\mathbf{T}_{11}^\top + \mathbf{V}_2\mathbf{T}_{12}^\top \tag{5.38}$$

and

$$\mathbf{U} = \mathbf{V}_1\mathbf{T}_{21}^\top + \mathbf{V}_2\mathbf{T}_{22}^\top . \tag{5.39}$$

One can easily check that $\mathbf{B}^\top\mathbf{B} = \mathbf{TT}^\top = \mathbf{I}_{k+p}$. Indeed:

$$\begin{aligned} \mathbf{B}^\top\mathbf{B} &= \begin{bmatrix} \mathbf{T}_{11}\mathbf{T}_{11}^\top + \mathbf{T}_{12}\mathbf{T}_{12}^\top & \mathbf{T}_{11}\mathbf{T}_{21}^\top + \mathbf{T}_{12}\mathbf{T}_{22}^\top \\ \mathbf{T}_{21}\mathbf{T}_{11}^\top + \mathbf{T}_{22}\mathbf{T}_{12}^\top & \mathbf{T}_{21}\mathbf{T}_{21}^\top + \mathbf{T}_{22}\mathbf{T}_{22}^\top \end{bmatrix}, \\ &= \begin{bmatrix} \mathbf{I}_k & \mathbf{O}_{k\times p} \\ \mathbf{O}_{p\times k} & \mathbf{I}_p \end{bmatrix} = \mathbf{I}_{k+p}, \end{aligned} \tag{5.40}$$

which follows from the SVD of $\mathbf{Y}$. Thus, the algorithm finds block-orthonormal $[\mathbf{F} \;\; \mathbf{U}]$ for given or estimated $\mathbf{\Lambda}$ and $\mathbf{\Psi}$, and then updates them as $\mathbf{\Lambda} = \mathbf{Z}^\top\mathbf{F}$ and $\mathbf{\Psi} = \text{diag}(\mathbf{U}^\top\mathbf{Z})$.

The matrix of factor loadings $\mathbf{\Lambda}$ in the EFA model is required to have full column rank $k$. Assuming that $\text{rank}(\mathbf{Z}) \geq k$, the alternating algorithm preserves the rank of $\mathbf{\Lambda}$ by constructing it as $\mathbf{\Lambda} = \mathbf{Z}^{\top}\mathbf{F}$ which gives $\text{rank}(\mathbf{\Lambda}) \leq \min\{\text{rank}(\mathbf{Z}), \text{rank}(\mathbf{F})\} = k$. The whole alternating least squares (ALS) process of finding $\{\mathbf{F}, \mathbf{U}\}$ and $\{\mathbf{\Lambda}, \mathbf{\Psi}\}$ continues until the loss function (5.1) cannot be reduced further.

The approach by De Leeuw (2004, 2008) is equivalent to a method developed by Henk A. L. Kiers in some unpublished notes (H. A. L. Kiers, personal communication, 2009). Sočan (2003) called this approach Direct-simple factor analysis and gives a description in some detail.

Since premultiplying (3.19) by $\mathbf{Z}^{\top}$ gives

$$\mathbf{Z}^{\top}\mathbf{Z} = (\mathbf{F}\mathbf{\Lambda}^{\top})^{\top}\mathbf{F}\mathbf{\Lambda}^{\top} + (\mathbf{U}\mathbf{\Psi})^{\top}\mathbf{U}\mathbf{\Psi} = \mathbf{\Lambda}\mathbf{\Lambda}^{\top} + \mathbf{\Psi}^{2} \ ,$$

it can be seen that optimizing the loss function (5.1) is just another (orthogonally invariant) way to optimize (4.5), that is, to measure how similar the sample correlation matrix $\mathbf{Z}^{\top}\mathbf{Z}$ is to the model correlation matrix $\mathbf{\Theta} = \mathbf{\Lambda}\mathbf{\Lambda}^{\top} + \mathbf{\Psi}^{2}$ in the least squares sense.

De Leeuw (2004) also outlines an algorithm to optimize (5.1) that updates $\mathbf{F}$ and $\mathbf{U}$ successively:

(i) for given $\mathbf{\Lambda}$, $\mathbf{\Psi}$, and $\mathbf{U}$ find orthonormal $\mathbf{F}$ which minimizes $||(\mathbf{Z} - \mathbf{U}\mathbf{\Psi}) - \mathbf{F}\mathbf{\Lambda}^{\top}||_F^2$ ,

(ii) for given $\mathbf{\Lambda}$, $\mathbf{\Psi}$, and $\mathbf{F}$ find orthonormal $\mathbf{U}$ which minimizes $||(\mathbf{Z} - \mathbf{F}\mathbf{\Lambda}^{\top}) - \mathbf{U}\mathbf{\Psi}||_F^2$ ,

(iii) for given $\mathbf{F}$ and $\mathbf{U}$, find $\mathbf{\Lambda} = \mathbf{Z}^{\top}\mathbf{F}$ and $\mathbf{\Psi} = \text{diag}(\mathbf{U}^{\top}\mathbf{Z})$ .

However, no indication is given in De Leeuw (2004) how the orthonormal $\mathbf{F}$ and $\mathbf{U}$ constructed this way could fulfill $\mathbf{U}^{\top}\mathbf{F} = \mathbf{O}_{p \times k}$. Furthermore, $\mathbf{Z} - \mathbf{F}\mathbf{\Lambda}^{\top}$ in step (ii) is always, by construction, rank deficient. A method for solving such modified Procrustes problems can be obtained as follows.

Assume that $\mathbf{F}$ has full column rank $k$. Let the columns of the $n \times (n-k)$ matrix $\mathbf{F}_\perp$ form an orthonormal basis of null($\mathbf{F}$) in $\mathbb{R}^n$. One way to find such $\mathbf{F}_\perp$ is by means of the QR factorization of $\mathbf{F}$ which has the following simple form:

$$\mathbf{F} = \mathbf{QR} = \mathbf{Q}\begin{bmatrix} \mathbf{I}_k \\ \mathbf{O}_{(n-k)\times k} \end{bmatrix} = \begin{bmatrix} \mathbf{F} & \mathbf{F}_\perp \end{bmatrix}\begin{bmatrix} \mathbf{I}_k \\ \mathbf{O}_{(n-k)\times k} \end{bmatrix} , \tag{5.41}$$

where $\mathbf{Q} \in \mathbb{R}^{n\times n}$ is orthogonal and $\mathbf{R} \in \mathbb{R}^{n\times k}$ is upper triangular. Let $\mathbf{U} = \mathbf{F}_\perp\tilde{\mathbf{U}}$ and note that $\mathbf{U}^\top\mathbf{U} = \tilde{\mathbf{U}}^\top\mathbf{F}_\perp^\top\mathbf{F}_\perp\tilde{\mathbf{U}} = \tilde{\mathbf{U}}^\top\tilde{\mathbf{U}} = \mathbf{I}_p$. Using (5.41), the constraint $\mathbf{F}^\top\mathbf{U} = \mathbf{O}_{k\times p}$, and the fact that $\mathbf{F}^\top(\mathbf{Z} - \mathbf{F}\Lambda^\top) = \mathbf{O}_{k\times p}$, the function in (ii) can be transformed into

$$\begin{aligned} ||\mathbf{Q}^\top(\mathbf{Z} - \mathbf{F}\Lambda^\top - \mathbf{U}\Psi)||_F^2 &= \left\|\begin{matrix} \mathbf{F}^\top(\mathbf{Z} - \mathbf{F}\Lambda^\top - \mathbf{U}\Psi) \\ \mathbf{F}_\perp^\top(\mathbf{Z} - \mathbf{F}\Lambda^\top - \mathbf{U}\Psi) \end{matrix}\right\|_F^2 , \\ &= ||\mathbf{F}_\perp^\top(\mathbf{Z} - \mathbf{F}\Lambda^\top) - \tilde{\mathbf{U}}\Psi||_F^2 . \end{aligned} \tag{5.42}$$

Thus, the modified Procrustes problem (ii) is reduced to the standard Procrustes problem of minimizing (5.42) subject to $\tilde{\mathbf{U}}^\top\tilde{\mathbf{U}} = \mathbf{I}_p$. Suppose $\tilde{\mathbf{U}}$ is the $(n-k)\times p$ orthonormal matrix found by minimizing (5.42). Then, the original $\mathbf{U}$ of the Procrustes-like problem (ii) is computed as $\mathbf{U} = \mathbf{F}_\perp\tilde{\mathbf{U}}$. The constraint $\mathbf{F}^\top\mathbf{U} = \mathbf{O}_{k\times p}$ is fulfilled as $\mathbf{F}^\top\mathbf{F}_\perp\tilde{\mathbf{U}} = \mathbf{O}_{k\times p}$. The alternating procedure (i) – (iii) is continued until the loss function (5.1) cannot be reduced further.

For reasons explained above one can look for a $p \times k$ lower triangular matrix of factor loadings $\mathbf{L}$, with a triangle of $k(k-1)/2$ zeros. Then, the updating formula $\Lambda = \mathbf{Z}^\top\mathbf{F}$ should simply be replaced by $\mathbf{L} = \texttt{tril}(\mathbf{Z}^\top\mathbf{F})$ and thus one obtains an alternative solution of the Procrustes problems discussed above.

# Chapter 6

# EFA-like PCA for $n > p$

A statistical technique that is frequently used as a synonym for EFA is PCA. Despite the differences between PCA and EFA (e.g., Jolliffe, 2002, pp. 158-161), both methods aim to reduce the dimensionality of a set of data. It is of interest to find conditions under which PCA and EFA solutions can or cannot be close for a particular data set (Rao, 1996). For this reason, in this Chapter PCA is viewed as a special case of EFA with the error term resembling the EFA one. Based on an initial PCA solution, the error term is then decomposed to achieve an EFA-like factorization of the data. This specific EFA-like PCA construction helps to compare the numerical solutions obtained by PCA and EFA. In Section 6.1, aside from the standard PCA based on the SVD, a new procedure to accomplish PCA by means of the QR factorization of the data is introduced. Numerical procedures to achieve an EFA-like factorization of the error term are presented in Section 6.2. The algorithms developed in Chapter 5 and Chapter 6 are illustrated numerically with Harman's five socio-economic variables data (Harman, 1976) in Section 6.3.

## 6.1 PCA based on the SVD and QR factorization of the data

As a matrix decomposition, PCA is based on the SVD (Golub and Van Loan, 1996) of the data $\mathbf{Z}$ which is quite different from the EFA matrix decomposition of $\mathbf{Z}$. To appreciate the difference between EFA and PCA, consider the SVD of $\mathbf{Z}$ which has the form:

$$\mathbf{Z} = \mathbf{V}\mathbf{D}\mathbf{T}^\top \ , \tag{6.1}$$

where $\mathbf{V} \in \mathbb{R}^{n \times p}$ is orthonormal, $\mathbf{T} \in \mathbb{R}^{p \times p}$ is orthogonal and $\mathbf{D} \in \mathbb{R}^{p \times p}$ is a diagonal matrix with the singular values of $\mathbf{Z}$ sorted in decreasing order, $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p \geq 0$, on its main diagonal. After partitioning, (6.1) can be written as

$$\mathbf{Z} = \mathbf{V}_1\mathbf{D}_1\mathbf{T}_1^\top + \mathbf{V}_2\mathbf{D}_2\mathbf{T}_2^\top \ , \tag{6.2}$$

where $\mathbf{D}_1 = \text{diag}(\sigma_1, ..., \sigma_k)$, $\mathbf{D}_2 = \text{diag}(\sigma_{k+1}, ..., \sigma_p)$ and $\mathbf{V}_1, \mathbf{V}_2, \mathbf{T}_1$, and $\mathbf{T}_2$ are the corresponding orthonormal matrices of left and right singular vectors with sizes $n \times k$, $n \times (p-k)$, $p \times k$, and $p \times (p-k)$, respectively. The norm of the error term $\mathbf{E}_{SVD} = \mathbf{V}_2\mathbf{D}_2\mathbf{T}_2^\top$ is

$$||\mathbf{E}_{SVD}||_F = ||\mathbf{D}_2||_F = \sqrt{\sum_{i=k+1}^{p} \sigma_i^2} \ .$$

By defining $\mathbf{F} := \mathbf{V}_1$ and $\mathbf{\Lambda} := \mathbf{T}_1\mathbf{D}_1$, one obtains the first (common) part $\mathbf{F}\mathbf{\Lambda}^\top$ in the EFA decomposition of $\mathbf{Z}$. Moreover, $\mathbf{F}$ $(= \mathbf{V}_1)$ in both EFA and PCA is orthogonal to the second ('error') term. However, the error term $\mathbf{E}_{SVD}$ in the PCA decomposition has a very different structure from $\mathbf{U}\mathbf{\Psi}$ in EFA. It will be demonstrated that the form of $\mathbf{E}_{SVD}$ gives only superficial differences between EFA and PCA. What really matters is the underlying model of the EFA matrix decomposition (3.19) which is assumed a priori. In other words, the EFA algorithms look for pairs of unknowns $\{\mathbf{\Lambda}, \mathbf{\Psi}\}$ and $\{\mathbf{F}, \mathbf{U}\}$, whereas EFA-like PCA looks for $\{\mathbf{\Lambda}, \mathbf{F}\}$ and $\{\mathbf{\Psi}, \mathbf{U}\}$.

Formally speaking, to get an EFA-like decomposition from the PCA one, $\mathbf{E}_{SVD}$ should be further decomposed as a product of orthonormal and diagonal matrices $\mathbf{U}$ and $\mathbf{\Psi}$ of sizes $n \times p$ and $p \times p$, respectively. This can be formulated as the following LS optimization problem:

$$\min_{\mathbf{U},\mathbf{\Psi}} ||\mathbf{E}_{SVD} - \mathbf{U}\mathbf{\Psi}||_F^2 \ , \tag{6.3}$$

subject to the constraints $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_p$ and $\mathbf{\Psi}$ being a diagonal matrix. In addition, $\mathbf{U}$ should be orthogonal to $\mathbf{F}$ already found in (6.2).

Traditionally, PCA accomplished by the SVD reduced-rank approximation is considered as the optimal method for the reduction of the dimensionality of the data. This is due to the LS property of the SVD stated in (3.4). However, as a rank-reducing method, the SVD can be expensive to compute.

Recently, it has been shown that "if any reduced-rank approximation is accurate then it contains good approximations to the singular vectors corresponding to large singular values" (Berry, Pulatova, and Stewart, 2005, Theorem 6.1).

This is true, in particular, for the QR decomposition (Golub and Van Loan, 1996) which possesses a number of attractive numerical properties that the SVD lacks (Stewart, 1998). For computational and interpretational reasons, one can perform a PCA-like analysis based on the QR factorization of $\mathbf{Z}$ which is given by

$$\mathbf{Z} = \mathbf{Q}\mathbf{R} \ , \tag{6.4}$$

where $\mathbf{Q} \in \mathbb{R}^{n \times n}$ is orthogonal and $\mathbf{R} \in \mathbb{R}^{n \times p}$ is upper triangular. After partitioning, (6.4) can be written as

$$\mathbf{Z} = \begin{bmatrix} \mathbf{Q}_1 & \mathbf{Q}_2 \end{bmatrix} \begin{bmatrix} \mathbf{R}_1 \\ \\ \mathbf{R}_2 \end{bmatrix} = \mathbf{Q}_1\mathbf{R}_1 + \mathbf{Q}_2\mathbf{R}_2 \ , \tag{6.5}$$

where $\mathbf{Q}_1 \in \mathbb{R}^{n \times k}$ and $\mathbf{Q}_2 \in \mathbb{R}^{n \times (n-k)}$ are orthonormal and $\mathbf{R}_1 \in \mathbb{R}^{k \times p}$ and $\mathbf{R}_2 \in \mathbb{R}^{(n-k) \times p}$ are upper triangular. By defining $\mathbf{F} := \mathbf{Q}_1$ and $\mathbf{L} := \mathbf{R}_1^{\top}$, (6.5) turns into

$$\mathbf{Z} = \mathbf{F}\mathbf{L}^{\top} + \mathbf{E}_{QR} \ , \tag{6.6}$$

where $\mathbf{L} \in \mathbb{R}^{p \times k}$ is a lower triangular matrix and $\mathbf{E}_{QR} = \mathbf{Q}_2\mathbf{R}_2$ is the error term. Note that $\mathbf{F}^{\top}\mathbf{E}_{QR} = \mathbf{O}_{k \times p}$. The norm of $\mathbf{E}_{QR}$ is $||\mathbf{E}_{QR}||_F = ||\mathbf{R}_2||_F$, that is, the size of the error equals the sum of squares of the elements of an upper $(p-k) \times (p-k)$ triangular submatrix which are the only non-zero entries in $\mathbf{R}_2$.

To get a further EFA-like decomposition of $\mathbf{Z}$ one needs to solve the optimization problem:

$$\min_{\mathbf{U},\mathbf{\Psi}} ||\mathbf{E}_{QR} - \mathbf{U}\mathbf{\Psi}||_F^2 \ , \tag{6.7}$$

subject to $\mathbf{U}^{\top}\mathbf{U} = \mathbf{I}_p$, $\mathbf{U}^{\top}\mathbf{F} = \mathbf{O}_{p \times k}$ and $\mathbf{\Psi}$ being a diagonal matrix.

Thus, whether an initial PCA solution is available from either the SVD or the QR factorization of $\mathbf{Z}$, a Procrustes-like problem should be solved in both cases to obtain the corresponding EFA-like solution.

## 6.2 EFA-like decomposition of the error term

### 6.2.1 Dynamical system approach and optimality conditions

Before finding an efficient method for solving (6.3) and (6.7), such EFA-like solutions can be explored by considering the following dynamical system:

$$\dot{\mathbf{U}} = .5\mathbf{U}(\mathbf{U}^\top\mathbf{E}\boldsymbol{\Psi} - \boldsymbol{\Psi}\mathbf{E}^\top\mathbf{U}) + (\mathbf{I}_n - \mathbf{U}\mathbf{U}^\top)\mathbf{E}\boldsymbol{\Psi} \ , \tag{6.8}$$

$$\dot{\boldsymbol{\Psi}} = \text{diag}(\mathbf{E}^\top\mathbf{U}) - \boldsymbol{\Psi} \ , \tag{6.9}$$

where $\mathbf{E}$ denotes either $\mathbf{E}_{SVD}$ or $\mathbf{E}_{QR}$. The descent gradient flow $\dot{\mathbf{U}}$ starts from a random orthonormal $\mathbf{U}_0$ orthogonal to the corresponding $\mathbf{F}$ and remains orthogonal to $\mathbf{F}$ until convergence. In practice, since $\mathbf{E}$ has not full column rank, one can rewrite (6.8) for $\tilde{\mathbf{U}} = \mathbf{F}_\perp^\top\mathbf{U}$ and replace $\mathbf{E}$ by $\tilde{\mathbf{E}} = \mathbf{F}_\perp^\top\mathbf{E}$, where $\mathbf{F}_\perp$ is obtained from the QR decomposition of $\mathbf{F}$ in (5.41).

First order optimality conditions for $\mathbf{U}$ and $\boldsymbol{\Psi}$ that minimize either (6.3) or (6.7) are available from (6.8) and (6.9):

$$\mathbf{U}^\top\mathbf{E}\boldsymbol{\Psi} = \boldsymbol{\Psi}\mathbf{E}^\top\mathbf{U} \ , \tag{6.10}$$

$$(\mathbf{I}_n - \mathbf{U}\mathbf{U}^\top)\mathbf{E}\boldsymbol{\Psi} = \mathbf{O}_{n\times p} \ , \tag{6.11}$$

$$\boldsymbol{\Psi} = \text{diag}(\mathbf{U}^\top\mathbf{E}) \ . \tag{6.12}$$

Condition (6.10) states that $\mathbf{U}^\top\mathbf{E}\boldsymbol{\Psi}$ is a symmetric matrix. Since $\boldsymbol{\Psi}$ is diagonal, $\mathbf{U}^\top\mathbf{E}\boldsymbol{\Psi}$ is also diagonal. If $\boldsymbol{\Psi}$ is nonsingular, then $\mathbf{U}^\top\mathbf{E}$ is necessarily diagonal.

As $\mathbf{U}^\top\mathbf{F} = \mathbf{O}_{p\times k}$, the optimality conditions (6.10) – (6.12) are equivalent to

$$\mathbf{U}^\top\mathbf{Z}\boldsymbol{\Psi} = \boldsymbol{\Psi}\mathbf{Z}^\top\mathbf{U} \ , \tag{6.13}$$

$$(\mathbf{I}_n - \mathbf{U}\mathbf{U}^\top)\mathbf{Z}\boldsymbol{\Psi} = \mathbf{F}\boldsymbol{\Lambda}^\top\boldsymbol{\Psi} \ , \tag{6.14}$$

$$\boldsymbol{\Psi} = \text{diag}(\mathbf{U}^\top\mathbf{Z}) \ . \tag{6.15}$$

The conditions (6.13) and (6.15) together with the condition $\mathbf{\Lambda} = \mathbf{Z}^\top\mathbf{F}$, where the latter follows from the construction of either (6.2) or (6.4), are identical to the optimality conditions for simultaneous EFA established in (5.19) – (5.21). Clearly, the simultaneous EFA solutions will coincide with the EFA-like PCA solutions if they have identical $\mathbf{F}$ and $\mathbf{\Lambda}$ obtained from either (6.2) or (6.6). If $\mathbf{\Psi}$ is nonsingular, then (6.14) leads to

$$\mathbf{Z} = \mathbf{F}\mathbf{\Lambda}^\top + \mathbf{U}\mathbf{U}^\top\mathbf{Z} \ , \tag{6.16}$$

which together with (6.15) gives the following optimality condition:

$$\mathbf{Z} = \mathbf{F}\mathbf{\Lambda}^\top + \mathbf{U}\mathbf{\Psi} \ . \tag{6.17}$$

One is tempted to say that those $\mathbf{\Lambda}, \mathbf{F}, \mathbf{U}$ and $\mathbf{\Psi}$ that are solutions of the EFA-like PCA, are also solutions of the simultaneous EFA, despite the fact that $\mathbf{\Lambda}$ and $\mathbf{F}$ are found beforehand by the SVD or the QR factorization of $\mathbf{Z}$. However, in contrast to simultaneous EFA, the EFA-like PCA solutions do *not* satisfy the condition (5.25):

$$(\mathbf{Z} - \mathbf{F}\mathbf{\Lambda}^\top - \mathbf{U}\mathbf{\Psi})\mathbf{\Lambda} = \mathbf{O}_{n\times k} \ .$$

The optimality conditions derived for simultaneous EFA and EFA-like PCA shed light on in which cases one can expect similar EFA and PCA solutions. This addresses the following comments made by Rao (1996): "Some conditions under which the factor scores and principal components are close to each other have been given by Schneeweiss and Mathes (1995). It would be of interest to pursue such theoretical investigations and also examine in individual data sets the actual differences between principal components and factor scores." Both EFA and PCA solutions will be similar if the PCA solution meets the EFA optimality condition (5.25).

### 6.2.2 Iterative algorithm

The application of EFA-like PCA to $\mathbf{Z}$ requires an efficient method for solving the Procrustes-like problem:

$$\min_{\mathbf{U},\mathbf{\Psi}} ||\mathbf{E} - \mathbf{U}\mathbf{\Psi}||_F^2 \ , \tag{6.18}$$

subject to $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_p$ and $\mathbf{\Psi}$ being a diagonal matrix. In addition, $\mathbf{U}$ should be orthogonal to $\mathbf{F}$, for which $\mathbf{F}^\top\mathbf{E} = \mathbf{O}_{k\times p}$, that is, $\mathbf{F}^\top\mathbf{U} = \mathbf{O}_{k\times p}$. The major problem with transforming (6.18) into a standard Procrustes problem is that $\mathbf{E}$ is always, by construction, rank deficient.

Using $\mathbf{E} = \mathbf{F}_\perp\tilde{\mathbf{E}}$ and $\mathbf{U} = \mathbf{F}_\perp\tilde{\mathbf{U}}$, where $\mathbf{F}_\perp$ is obtained from the QR decomposition of $\mathbf{F}$ in (5.41), one notes that $\mathbf{U}^\top\mathbf{U} = \tilde{\mathbf{U}}^\top\mathbf{F}_\perp^\top\mathbf{F}_\perp\tilde{\mathbf{U}} = \tilde{\mathbf{U}}^\top\tilde{\mathbf{U}} = \mathbf{I}_p$. Then, the objective function in (6.18) can be transformed into

$$||\mathbf{E} - \mathbf{U}\mathbf{\Psi}||_F^2 = ||\mathbf{Q}^\top(\mathbf{E} - \mathbf{U}\mathbf{\Psi})||_F^2 = \left|\left| \begin{matrix} \mathbf{F}^\top(\mathbf{E} - \mathbf{U}\mathbf{\Psi}) \\ \mathbf{F}_\perp^\top(\mathbf{E} - \mathbf{U}\mathbf{\Psi}) \end{matrix} \right|\right|_F^2 = ||\tilde{\mathbf{E}} - \tilde{\mathbf{U}}\mathbf{\Psi}||_F^2 \ , \tag{6.19}$$

where $\mathbf{Q} = \begin{bmatrix} \mathbf{F} & \mathbf{F}_\perp \end{bmatrix}$, $\mathbf{F}^\top\mathbf{U} = \mathbf{O}_{k\times p}$, and $\mathbf{F}^\top\mathbf{E} = \mathbf{O}_{k\times p}$. Thus, the modified Procrustes problem (6.18) is reduced to the following standard Procrustes problem:

$$\min_{\tilde{\mathbf{U}},\mathbf{\Psi}} ||\tilde{\mathbf{E}} - \tilde{\mathbf{U}}\mathbf{\Psi}||_F^2 \ , \tag{6.20}$$

subject to $\tilde{\mathbf{U}}^\top\tilde{\mathbf{U}} = \mathbf{I}_p$ and $\mathbf{\Psi}$ being a diagonal matrix. Hence, the Procrustes-like problem (6.18) can be solved by an alternating procedure of solving (6.20) and updating $\mathbf{\Psi} = \mathrm{diag}(\mathbf{U}^\top\mathbf{E}) = \mathrm{diag}(\tilde{\mathbf{U}}^\top\tilde{\mathbf{E}})$ until convergence.

## 6.3 Application to Harman's five socio-economic variables data

To illustrate the iterative algorithms developed in Chapter 5 and Chapter 6, a well-known and frequently studied data set in factor analysis is employed next: Harman's five socio-economic variables data (Harman, 1976). The raw data are given in Table 6.1 (Harman, 1976, Table 2.1, p. 14). Only $n = 12$ observations and $p = 5$ variables are analyzed. The twelve observations are census tracts - small areal subdivisions of the city of Los Angeles. The five socio-economic variables are 'total population' (POPULATION), 'median school years' (SCHOOL), 'total employment' (EMPLOYMENT),

| Tract | POPULATION | SCHOOL | EMPLOYMENT | SERVICES | HOUSE |
|---:|---:|---:|---:|---:|---:|
| 1 | 5700 | 12.8 | 2500 | 270 | 25000 |
| 2 | 1000 | 10.9 | 600 | 10 | 10000 |
| 3 | 3400 | 8.8 | 1000 | 10 | 9000 |
| 4 | 3800 | 13.6 | 1700 | 140 | 25000 |
| 5 | 4000 | 12.8 | 1600 | 140 | 25000 |
| 6 | 8200 | 8.3 | 2600 | 60 | 12000 |
| 7 | 1200 | 11.4 | 400 | 10 | 16000 |
| 8 | 9100 | 11.5 | 3300 | 60 | 14000 |
| 9 | 9900 | 12.5 | 3400 | 180 | 18000 |
| 10 | 9600 | 13.7 | 3600 | 390 | 25000 |
| 11 | 9600 | 9.6 | 3300 | 80 | 12000 |
| 12 | 9400 | 11.4 | 4000 | 100 | 13000 |

Table 6.1: Raw data for Harman's five socio-economic variables.

'miscellaneous professional services' (SERVICES) and 'median house value' (HOUSE). The raw data are preprocessed such that the variables have zero mean and unit length. The preprocessed measurements are collected in a $12 \times 5$ matrix $\mathbf{Z}$ and are fitted by an EFA model with two common factors ($k = 2$) in terms of different LS loss functions.

First, standard EFA least squares solutions $\{\mathbf{\Lambda}, \mathbf{\Psi}\}$ are obtained by minimizing $\mathcal{F}_{LS}$ in (4.5). To make the solutions comparable to the ones obtained in the sequel, these are found by defining the LS fitting problem of minimizing $\mathcal{F}_{LS}$ according to an eigenvalue decomposition (EVD) and a lower triangular (LT) reparameterization of the EFA model, respectively (Trendafilov, 2003, 2005). It would be helpful to recall briefly the idea of the EVD and LT reparameterization.

Consider the EVD of the positive semi-definite matrix $\mathbf{\Lambda}\mathbf{\Lambda}^\top$ of rank at most $k$ in (3.17), that is, let $\mathbf{\Lambda}\mathbf{\Lambda}^\top = \mathbf{Q}\mathbf{D}^2\mathbf{Q}^\top$, where $\mathbf{D}^2$ is a $k \times k$ diagonal matrix composed of the largest (non-negative) $k$ eigenvalues of $\mathbf{\Lambda}\mathbf{\Lambda}^\top$ arranged in descending order and $\mathbf{Q}$ is a $p \times k$ column-wise orthonormal matrix containing the corresponding eigenvectors. Then, the model correlation structure (3.17) can be rewritten as

$$\mathbf{\Theta} = \mathbf{Q}\mathbf{D}^2\mathbf{Q}^\top + \mathbf{\Psi}^2 \ .$$

Thus, instead of a pair $\{\mathbf{\Lambda}, \mathbf{\Psi}\}$, a triple $\{\mathbf{Q}, \mathbf{D}, \mathbf{\Psi}\}$ will be sought and $\mathbf{\Lambda}$ is decomposed as $\mathbf{QD}$.

Let $\mathbf{L}$ be a $p \times k$ lower triangular matrix, with a triangle of $k(k-1)/2$ zeros. Then $\mathbf{\Lambda}\mathbf{\Lambda}^\top$ can be reparameterized by $\mathbf{L}\mathbf{L}^\top$. Hence, for the LT reparameterization, (3.17) can be rewritten as

$$\mathbf{\Theta} = \mathbf{L}\mathbf{L}^\top + \mathbf{\Psi}^2 \ .$$

For both reparameterizations, the corresponding LS fitting problems are solved by making use of the projected gradient approach (Trendafilov, 2003, 2005). The LS solutions $\{\Lambda, \Psi^2\}$ are given in Table 6.2.

| | EVD reparameterization | | | LT reparameterization | | |
|---|---|---|---|---|---|---|
| Variable | $\Lambda$ | | $\Psi^2$ | $\mathbf{L}$ | | $\Psi^2$ |
| POPULATION | .62 | .78 | .0117 | 1.00 | 0 | .0101 |
| SCHOOL | .70 | -.52 | .2344 | .03 | .87 | .2347 |
| EMPLOYMENT | .70 | .68 | .0347 | .97 | .13 | .0439 |
| SERVICES | .88 | -.14 | .2029 | .43 | .78 | .2029 |
| HOUSE | .78 | -.60 | .0260 | .01 | .99 | .0251 |

Table 6.2: Standard LS solutions for Harman's five socio-economic variables data.

Then, LS solutions for estimating $\{\mathbf{F}, \Lambda, \mathbf{U}, \Psi\}$ simultaneously are obtained by minimizing $\mathcal{F}_{DeL}$ in (5.1), making use of the new iterative algorithm discussed in Section 5.2 for updating $\mathbf{F}$ and $\mathbf{U}$ successively. To reduce the chance of mistaking a locally optimal solution for a globally optimal one, the algorithm was run twenty times, each with different randomly chosen column-wise orthonormal matrices $\mathbf{F}$ and $\mathbf{U}$. The algorithm was stopped when successive function values differed by less than $\epsilon = 10^{-6}$.

The corresponding results for $\{\Lambda, \Psi^2\}$ applying two parameterizations for the loadings are provided in Table 6.3. The results reported are the 'best' obtained after the twenty random starts. By 'best' a solution employing the full column rank (FCR) preserving formula $\Lambda = \mathbf{Z}^\top\mathbf{F}$ is meant which resembles the lower triangular one $\mathbf{L} = \mathtt{tril}(\mathbf{Z}^\top\mathbf{F})$ most.

| | $\Lambda = \mathbf{Z}^\top\mathbf{F}$ | | | $\mathbf{L} = \text{tril}(\mathbf{Z}^\top\mathbf{F})$ | | |
|---|---|---|---|---|---|---|
| | error of fit = .002835 | | | error of fit = .002836 | | |
| Variable | $\Lambda$ | | $\Psi^2$ | $\mathbf{L}$ | | $\Psi^2$ |
| POPULATION | .99 | .05 | .0150 | 1.00 | 0 | .0173 |
| SCHOOL | -.01 | .88 | .2292 | .03 | .88 | .2307 |
| EMPLOYMENT | .97 | .16 | .0182 | .98 | .11 | .0158 |
| SERVICES | .40 | .80 | .2001 | .44 | .78 | .2009 |
| HOUSE | -.03 | .98 | .0318 | .02 | .98 | .0292 |

Table 6.3: Simultaneous EFA solutions for Harman's five socio-economic variables data.

For both algorithms the twenty runs led to the same minimum of $\mathcal{F}_{DeL}$, up to the fourth decimal place. Numerical experiments revealed that the algorithm employing a lower triangular matrix $\mathbf{L}$ is slower but yields pretty stable loadings. In contrast, the algorithm employing $\Lambda = \mathbf{Z}^\top\mathbf{F}$ is faster, but converges to quite different $\Lambda$.

The iterative algorithm gives the same $\Psi^2$ and goodness-of-fit for both types of loadings. It is worth mentioning that their $\Psi^2$ values are similar to those produced by the standard EFA solutions in Table 6.2. Moreover, for the lower triangular reparameterization the loadings of the two solutions are almost identical.

It is of interest to compare the simultaneous EFA solutions in Table 6.3 and the standard EFA solutions in Table 6.2 with the ones obtained by means of EFA-like PCA based on the SVD and the QR decomposition of $\mathbf{Z}$. The corresponding algorithms minimizing (6.19) were run twenty times each and were stopped when successive function values differed by less than $\epsilon = 10^{-6}$. The initial value for $\Psi$ was simply taken to be

| | SVD<br>error of fit = .059281 | | | QR decomposition<br>error of fit = .029820 | | |
|---|---|---|---|---|---|---|
| Variable | $\mathbf{\Lambda}$ | | $\mathbf{\Psi}^2$ | $\mathbf{L}$ | | $\mathbf{\Psi}^2$ |
| POPULATION | .58 | .81 | .0000 | 1.00 | 0 | .0000 |
| SCHOOL | .77 | -.54 | .0945 | .01 | 1.00 | .0000 |
| EMPLOYMENT | .67 | .73 | .0095 | .97 | .14 | .0314 |
| SERVICES | .93 | -.10 | .1019 | .44 | .69 | .3114 |
| HOUSE | .79 | -.56 | .0055 | .02 | .86 | .2211 |

Table 6.4: EFA-like solutions for Harman's five socio-economic variables data.

a diagonal matrix with diagonal entries randomly drawn from a uniform distribution on the unit interval. The EFA-like solutions obtained from the two types of PCA are shown in Table 6.4. The most striking feature of both EFA-like PCA solutions compared to the simultaneous EFA solutions given in Table 6.3 is that the fits attained by the former are considerably worse.

One can further assess the difference between the simultaneous EFA solutions and the EFA-like PCA solutions by substituting them into the optimality condition (5.25) which is not satisfied for EFA-like PCA, that is, by calculating

$$E = \frac{||(\mathbf{Z} - \mathbf{F}\mathbf{\Lambda}^{\top} - \mathbf{U}\mathbf{\Psi})\mathbf{\Lambda}||_F^2}{nk} \; . \tag{6.21}$$

This gives a value of 0.0079 for EFA-like PCA based on the SVD and $4.5080 \times 10^{-8}$ for simultaneous EFA. For the LT reparameterization, the values are 0.0015 for EFA-like PCA based on the QR factorization and $2.0241 \times 10^{-8}$ for simultaneous EFA, respectively.

The SVD based EFA-like solution in Table 6.4 has loadings which are similar to the ones of standard EFA using $\mathbf{\Lambda} = \mathbf{Z}^\top\mathbf{F}$. The corresponding uniquenesses are smaller than the ones obtained by both standard EFA and simultaneous EFA. The QR based EFA-like PCA solution has loadings which are very similar to the ones in Table 6.2 and Table 6.3 employing the lower triangular reparametrization but the uniquenesses differ.

Numerical experiments revealed that both EFA-like PCA procedures are faster than the iterative algorithms for simultaneous parameter estimation.

Estimated common factor scores for the five socio-economic variables data are shown in Table 6.5. The first two pairs of columns are the scores obtained by using the formula (3.26) proposed by Anderson and Rubin (1956). They are denoted by $\mathbf{F}_{AR_{EVD}}$ and $\mathbf{F}_{AR_{LT}}$, respectively, as they are calculated from the two types of loadings shown in Table 6.2. For both parameterizations of the loadings, the next two pairs of columns are the factor scores $\mathbf{F}_{FCR}$ and $\mathbf{F}_{LT}$ found by the iterative algorithm for simultaneous parameter estimation. The last two pairs of columns $\mathbf{F}_{SVD}$ and $\mathbf{F}_{QR}$ show component scores obtained by PCA based on the SVD and the QR decomposition, respectively. Note that for all sets of scores it holds that $\mathbf{F}^\top\mathbf{F} = \mathbf{I}_k$. It can be seen from Table 6.5 that the factor scores $\mathbf{F}_{AR_{LT}}$ and $\mathbf{F}_{LT}$ as well as the component scores $\mathbf{F}_{QR}$, all obtained from a lower triangular parameterization of the loadings matrix, are quite similar.

In contrast, Guttman's arbitrary constructions for the common factor scores discussed in Section 3.2.2 may not always be applicable in practice. Indeed, for both loading matrices in Table 6.2, one can easily check that $\mathbf{I}_k - \mathbf{\Lambda}^\top(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{\Lambda}$ in (3.30) is not positive semi-definite as required. For example, using the parametrization $\mathbf{\Lambda} = \mathbf{Z}^\top\mathbf{F}$

| Tract | Standard EFA | | | | Simultaneous EFA | | | | PCA | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\mathbf{F}_{AR_{EVD}}$ | | $\mathbf{F}_{AR_{LT}}$ | | $\mathbf{F}_{FCR}$ | | $\mathbf{F}_{LT}$ | | $\mathbf{F}_{SVD}$ | | $\mathbf{F}_{QR}$ | |
| 1 | .27 | .28 | -.05 | .38 | .04 | .38 | .02 | .38 | .29 | .21 | -.05 | .37 |
| 2 | -.51 | .18 | -.46 | -.29 | -.41 | -.32 | -.41 | -.32 | -.40 | .23 | -.46 | -.30 |
| 3 | -.45 | -.04 | -.25 | -.38 | -.28 | -.37 | -.27 | -.37 | -.44 | -.03 | -.25 | -.44 |
| 4 | .16 | .40 | -.21 | .37 | -.21 | .39 | -.20 | .38 | .14 | .39 | -.21 | .37 |
| 5 | .16 | .38 | -.20 | .36 | -.22 | .38 | -.21 | .38 | .10 | .35 | -.20 | .37 |
| 6 | -.11 | -.31 | .17 | -.28 | .13 | -.25 | .12 | -.25 | -.21 | -.35 | .17 | -.27 |
| 7 | -.31 | .32 | -.44 | -.04 | -.47 | -.03 | -.48 | -.04 | -.31 | .35 | -.44 | -.00 |
| 8 | .04 | -.29 | .25 | -.15 | .25 | -.15 | .24 | -.15 | .02 | -.26 | .25 | -.15 |
| 9 | .24 | -.22 | .32 | .05 | .25 | .04 | .25 | .04 | .24 | -.17 | .32 | .05 |
| 10 | .50 | .02 | .29 | .40 | .26 | .37 | .27 | .37 | .57 | .02 | .29 | .38 |
| 11 | -.02 | -.39 | .29 | -.26 | .27 | -.25 | .27 | -.25 | -.07 | -.40 | .29 | -.25 |
| 12 | .04 | -.33 | .28 | -.17 | .39 | -.19 | .41 | -.19 | .08 | -.35 | .28 | -.18 |

Table 6.5: Common factor scores for Harman's five socio-economic variables data.

one finds:

$$\mathbf{G}^{\top}\mathbf{G} = \mathbf{I}_k - \mathbf{\Lambda}^{\top}(\mathbf{Z}^{\top}\mathbf{Z})^{-1}\mathbf{\Lambda} = \begin{pmatrix} .0155 & .0038 \\ .0038 & -.0177 \end{pmatrix}.$$

From this point of view the estimation procedures developed in this thesis for obtaining common factor scores present more reliable alternatives to Guttman's expressions.

# Chapter 7

# EFA of Data Matrices with $p \geq n$

In a number of modern applications, the number of available observations is less than the number of variables. Consider for example data arising from experiments in genome research. The data from such experiments are usually in the form of large horizontal matrices of expression levels of $p$ genes (variables) under $n$ experimental conditions (observations) such as different times, cells or tissues. Another discipline where high-dimensional data with $p \gg n$ typically occur is in atmospheric science, where a meteorological variable is measured at $p$ spatial locations at $n$ different points in time.

This Chapter covers the case of EFA of horizontal data matrices with $p \geq n$ (Unkel and Trendafilov, 2009a). Novel numerical procedures for simultaneous estimation of all EFA model unknowns are introduced in Section 7.1. An algorithm for EFA-like PCA is presented in Section 7.2. In Section 7.3, the new procedures are illustrated with Thurstone's 26-variable box data (Thurstone, 1947) and a real large high-dimensional data set from atmospheric science.

## 7.1 Simultaneous Estimation of all EFA Model Unknowns

If $p \geq n$, the sample covariance/correlation matrix is singular. Then, the most common factor extraction methods, such as ML factor analysis or GLS factor analysis, cannot be applied. Robertson and Symons (2007) consider maximum likelihood fitting of such rank-deficient correlation matrices by the EFA correlation structure $\Theta = \mathbf{\Lambda}\mathbf{\Lambda}^\top + \mathbf{\Psi}^2$. In other words, they look to approximate a singular symmetric matrix by a positive definite one having the specific form $\mathbf{\Lambda}\mathbf{\Lambda}^\top + \mathbf{\Psi}^2$ imposed by the EFA model and assuming $\mathbf{\Psi}^2$ positive definite.

Alternatively, one can minimize the LS loss function (4.5), which does not require $\mathbf{Z}^\top\mathbf{Z}$ to be invertible. However, there is a conceptual difficulty in adopting the approach to EFA introduced by Robertson and Symons (2007) or minimizing (4.5).

It is demonstrated in Section 3.2.2 that the EFA correlation structure $\Theta = \mathbf{\Lambda}\mathbf{\Lambda}^\top + \mathbf{\Psi}^2$ is a consequence of the accepted EFA model (3.19):

$$\mathbf{Z} = \mathbf{F}\mathbf{\Lambda}^\top + \mathbf{U}\mathbf{\Psi} \ ,$$

and the assumptions made for its parameters. When $p > n$, there is no problem to assume that the rank of the loading matrix $\mathbf{\Lambda}$ is $k$ ($k \ll p$). The $k$-factor model can still assume that $\mathbf{F}^\top\mathbf{F} = \mathbf{I}_k$ and $\mathbf{U}^\top\mathbf{F} = \mathbf{O}_{p\times k}$. Unfortunately, the classical constraint $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_p$ cannot be fulfilled as $\mathbf{U}^\top\mathbf{U}$ has at most rank $n$ ($< p$). With $\mathbf{U}^\top\mathbf{U} \neq \mathbf{I}_p$, the EFA correlation structure can be written as

$$\Theta = \mathbf{\Lambda}\mathbf{\Lambda}^\top + \mathbf{\Psi}\mathbf{U}^\top\mathbf{U}\mathbf{\Psi} \ .$$

In order to preserve the standard EFA correlation structure (3.17), the more general constraint $\mathbf{U}^\top\mathbf{U}\mathbf{\Psi} = \mathbf{\Psi}$ is introduced. In other words, $\mathbf{\Psi}$ can have at most $n$ nonzero entries. Then, the EFA model and the new constraint imposed imply the same identities as for the classical case ($n > p$):

$$\mathbf{F}^\top\mathbf{Z} = \mathbf{F}^\top\mathbf{F}\boldsymbol{\Lambda}^\top + \mathbf{F}^\top\mathbf{U}\boldsymbol{\Psi} \Longrightarrow \mathbf{F}^\top\mathbf{Z} = \boldsymbol{\Lambda}^\top, \tag{7.1}$$

$$\mathbf{U}^\top\mathbf{Z} = \mathbf{U}^\top\mathbf{F}\boldsymbol{\Lambda}^\top + \mathbf{U}^\top\mathbf{U}\boldsymbol{\Psi} \Longrightarrow \mathbf{U}^\top\mathbf{Z} = \boldsymbol{\Psi} \text{ (and thus diagonal)} , \tag{7.2}$$

which can be used to find $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ for given or estimated $\mathbf{F}$ and $\mathbf{U}$.

The immediate consequence of the new constraint $\mathbf{U}^\top\mathbf{U}\boldsymbol{\Psi} = \boldsymbol{\Psi}$ is that the existence of unique factors with zero variances should be acceptable in the EFA model when $p \geq n$. There is a long standing debate in classical EFA ($n > p$) about the acceptance of zero entries in $\boldsymbol{\Psi}^2$ which are commonly referred to as Heywood cases (Bartholomew and Knott, 1999; Jöreskog, 1977). While Bartholomew and Knott (1999) argue that in such situations the Heywood case variable is explained entirely by the common factors, Anderson (1984) finds it unsatisfactory for interpretational reasons and requires $\boldsymbol{\Psi}^2$ to be strictly positive definite. Either way, since the diagonal entries in $\boldsymbol{\Psi}^2$ are interpreted as variances, negative values are inadmissible and $\boldsymbol{\Psi}^2$ must be non-negative definite. It seems that a universal EFA model covering both cases $p > n$ and $n \geq p$ should accept $\boldsymbol{\Psi}^2$ being positive *semi*-definite.

By making use of the block matrix $\mathbf{B} = [\mathbf{F} \;\; \mathbf{U}]$, simultaneous estimation of the EFA parameters can be performed again by solving an augmented Procrustes problem by minimizing:

$$\mathcal{F}_{DeL} = ||\mathbf{Z} - \mathbf{B}\mathbf{A}^\top||_F^2$$

subject to the following new constraint:

$$\mathbf{B}\mathbf{B}^\top = \begin{bmatrix} \mathbf{F} & \mathbf{U} \end{bmatrix} \begin{bmatrix} \mathbf{F}^\top \\ \mathbf{U}^\top \end{bmatrix} = \mathbf{F}\mathbf{F}^\top + \mathbf{U}\mathbf{U}^\top = \mathbf{I}_n \ . \tag{7.3}$$

This modified EFA problem will fit the singular covariance/correlation $p \times p$ matrix of rank at most $n$ by the sum $\mathbf{\Lambda}\mathbf{\Lambda}^\top + \mathbf{\Psi}^2$ of two positive semi-definite $p \times p$ matrices with ranks $k$ and $n-k$, respectively. A similar problem is studied by Grubišić and Pietersz (2007) where a low-rank approximation to a singular correlation matrix is sought but without specifying the form of the approximation matrix.

The new constraint (7.3) simply gives the following identity:

$$\mathbf{U} = \mathbf{U}\mathbf{U}^\top\mathbf{U} \ ,$$

which postmultiplied by $\mathbf{\Psi}$ leads to

$$\mathbf{U}\mathbf{\Psi} = \mathbf{U}\mathbf{U}^\top\mathbf{U}\mathbf{\Psi} \Rightarrow \mathbf{U}(\mathbf{\Psi} - \mathbf{U}^\top\mathbf{U}\mathbf{\Psi}) = \mathbf{O}_{n\times p} \ . \tag{7.4}$$

Since $\mathbf{U}$ has not full column rank, (7.4) demonstrates that the constraint $\mathbf{U}^\top\mathbf{U}\mathbf{\Psi} = \mathbf{\Psi}$ does not necessarily follow from (7.3) and must be imposed separately. For $p \geq n$, the EFA loss function $\mathcal{F}_{DeL}$ has the following form:

$$\mathcal{F}_{DeL} = ||\mathbf{Z} - \mathbf{B}\mathbf{A}^\top||_F^2 = ||\mathbf{Z}||_F^2 + \mathrm{trace}(\mathbf{B}^\top\mathbf{B}\mathbf{A}^\top\mathbf{A}) - 2\,\mathrm{trace}(\mathbf{B}^\top\mathbf{Z}\mathbf{A}) \ , \tag{7.5}$$

which is different from (5.34), because $\mathbf{B}^\top\mathbf{B}$ is not an identity matrix in this case. Nevertheless, one can see that:

$$\begin{aligned}
\mathrm{trace}(\mathbf{B}^\top\mathbf{B}\mathbf{A}^\top\mathbf{A}) &= \mathrm{trace}\left\{\begin{bmatrix}\mathbf{F}^\top \\ \mathbf{U}^\top\end{bmatrix}\begin{bmatrix}\mathbf{F} & \mathbf{U}\end{bmatrix}\begin{bmatrix}\mathbf{\Lambda}^\top \\ \mathbf{\Psi}\end{bmatrix}\begin{bmatrix}\mathbf{\Lambda} & \mathbf{\Psi}\end{bmatrix}\right\} \ , \\
&= \mathrm{trace}\left\{\begin{bmatrix}\mathbf{I}_k & \mathbf{O}_{k\times p} \\ \mathbf{O}_{p\times k} & \mathbf{U}^\top\mathbf{U}\end{bmatrix}\begin{bmatrix}\mathbf{\Lambda}^\top\mathbf{\Lambda} & \mathbf{\Lambda}^\top\mathbf{\Psi} \\ \mathbf{\Psi}\mathbf{\Lambda} & \mathbf{\Psi}^2\end{bmatrix}\right\} \ , \\
&= \mathrm{trace}\left\{\begin{bmatrix}\mathbf{\Lambda}^\top\mathbf{\Lambda} & \mathbf{\Lambda}^\top\mathbf{\Psi} \\ \mathbf{U}^\top\mathbf{U}\mathbf{\Psi}\mathbf{\Lambda} & \mathbf{U}^\top\mathbf{U}\mathbf{\Psi}^2\end{bmatrix}\right\} \ , \\
&= \mathrm{trace}(\mathbf{\Lambda}^\top\mathbf{\Lambda}) + \mathrm{trace}(\mathbf{\Psi}^2) \ ,
\end{aligned}$$

showing that trace($\mathbf{B}^\top\mathbf{B}\mathbf{A}^\top\mathbf{A}$) does not depend on $\mathbf{F}$ and $\mathbf{U}$. Thus, for given or estimated $\mathbf{A}$, the minimization of $\mathcal{F}_{DeL}$ remains equivalent to the maximization of trace($\mathbf{B}^\top\mathbf{Z}\mathbf{A}$) and simply requires the SVD of $\mathbf{A}^\top\mathbf{Z}^\top$. After solving the Procrustes problem for $\mathbf{B} = [\mathbf{F}\ \ \mathbf{U}]$, one can update the values of $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ by making use of the identities (7.1) and (7.2). The whole alternating process of finding $\{\mathbf{F}, \mathbf{U}\}$ and $\{\mathbf{\Lambda}, \mathbf{\Psi}\}$ continues until the loss function (7.5) cannot be reduced further.

Another way to find $\mathbf{F}$ and $\mathbf{U}$ such that $\mathbf{F}^\top\mathbf{F} = \mathbf{I}_k$ and $\mathbf{U}^\top\mathbf{F} = \mathbf{O}_{p\times k}$ is by updating $\mathbf{F}$ and $\mathbf{U}$ successively. The first step of the algorithm is as follows:

(i) for given $\mathbf{\Lambda}, \mathbf{\Psi}$ and $\mathbf{U}$, find $\mathbf{F}$ that minimizes $\left|\left|(\mathbf{Z} - \mathbf{U}\mathbf{\Psi}) - \mathbf{F}\mathbf{\Lambda}^\top\right|\right|_F^2$ , subject to $\mathbf{F}^\top\mathbf{F} = \mathbf{I}_k$ .

To update $\mathbf{U}$, recall the QR decomposition of $\mathbf{F}$ in (5.41). Since $\mathbf{Q}\mathbf{Q}^\top = \mathbf{F}\mathbf{F}^\top + \mathbf{F}_\perp\mathbf{F}_\perp^\top = \mathbf{I}_n$, it follows from (7.3) that $\mathbf{U}\mathbf{U}^\top = \mathbf{F}_\perp\mathbf{F}_\perp^\top$. As $\mathbf{U} = \mathbf{F}_\perp\tilde{\mathbf{U}}$, one also notes that $\mathbf{U}\mathbf{U}^\top = \mathbf{F}_\perp\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top\mathbf{F}_\perp^\top$. Thus, only $\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top = \mathbf{I}_{n-k}$ ensures that $\mathbf{U}\mathbf{U}^\top = \mathbf{F}_\perp\mathbf{F}_\perp^\top$ and hence the new constraint $\mathbf{F}\mathbf{F}^\top + \mathbf{U}\mathbf{U}^\top = \mathbf{I}_n$ is fulfilled. The loss function $\mathcal{F}_{DeL}$ is transformed into $\left|\left|\mathbf{F}_\perp^\top(\mathbf{Z} - \mathbf{F}\mathbf{\Lambda}^\top) - \tilde{\mathbf{U}}\mathbf{\Psi}\right|\right|_F^2$. Then, the second step of the algorithm is as follows:

(ii) for given $\mathbf{\Lambda}, \mathbf{\Psi}$ and $\mathbf{F}$, find $\tilde{\mathbf{U}}$ that minimizes $\left|\left|\mathbf{F}_\perp^\top(\mathbf{Z} - \mathbf{F}\mathbf{\Lambda}^\top) - \tilde{\mathbf{U}}\mathbf{\Psi}\right|\right|_F^2$ , subject to $\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top = \mathbf{I}_{n-k}$ .

Expanding the loss function in (ii) gives

$$||\mathbf{F}_\perp^\top(\mathbf{Z} - \mathbf{F}\mathbf{\Lambda}^\top)||_F + \mathrm{trace}(\mathbf{\Psi}\tilde{\mathbf{U}}^\top\tilde{\mathbf{U}}\mathbf{\Psi}) - 2\,\mathrm{trace}(\mathbf{\Psi}(\mathbf{Z} - \mathbf{F}\mathbf{\Lambda}^\top)^\top\mathbf{F}_\perp\tilde{\mathbf{U}}) \ .$$

Using the new constraint $\mathbf{U}^\top\mathbf{U}\mathbf{\Psi} = \mathbf{\Psi}$, the middle term, trace($\mathbf{\Psi}\tilde{\mathbf{U}}^\top\tilde{\mathbf{U}}\mathbf{\Psi}$), turns into

$$\mathrm{trace}(\mathbf{\Psi}\tilde{\mathbf{U}}^\top\tilde{\mathbf{U}}\mathbf{\Psi}) = \mathrm{trace}(\mathbf{\Psi}\mathbf{U}^\top\mathbf{F}_\perp\mathbf{F}_\perp^\top\mathbf{U}\mathbf{\Psi}) = \mathrm{trace}(\mathbf{\Psi}\mathbf{U}^\top\mathbf{U}\mathbf{U}^\top\mathbf{U}\mathbf{\Psi}) = \mathrm{trace}(\mathbf{\Psi}^2) \ ,$$

which shows that trace($\boldsymbol{\Psi}\tilde{\mathbf{U}}^\top\tilde{\mathbf{U}}\boldsymbol{\Psi}$) does not depend on $\tilde{\mathbf{U}}$. Thus, the minimization of $||\mathbf{F}_\perp^\top(\mathbf{Z}-\mathbf{F}\boldsymbol{\Lambda}^\top)-\tilde{\mathbf{U}}\boldsymbol{\Psi}||_F^2$ is equivalent to the maximization of trace($\boldsymbol{\Psi}(\mathbf{Z}-\mathbf{F}\boldsymbol{\Lambda}^\top)^\top\mathbf{F}_\perp\tilde{\mathbf{U}}$) and simply requires the SVD of $\boldsymbol{\Psi}(\mathbf{Z}-\mathbf{F}\boldsymbol{\Lambda}^\top)^\top\mathbf{F}_\perp$. After solving this Procrustes problem, the original $\mathbf{U}$ is computed as $\mathbf{U} = \mathbf{F}_\perp\tilde{\mathbf{U}}$. Finally:

(iii) for given $\mathbf{F}$ and $\mathbf{U}$, find $\boldsymbol{\Lambda} = \mathbf{Z}^\top\mathbf{F}$ and $\boldsymbol{\Psi} = \text{diag}(\mathbf{U}^\top\mathbf{Z})$ .

The alternating procedure (i) – (iii) is continued until the loss function (7.5) cannot be reduced further.

## 7.2 EFA-like PCA

If $p > n$, the rank of the unique part $\mathbf{U}\boldsymbol{\Psi}$ of the EFA model (3.19) can be at most $n$. As mentioned in the previous Section, this implies that for such cases insisting on positive definite $\boldsymbol{\Psi}^2$ is not reasonable. Since $\mathbf{U}^\top\mathbf{U} \neq \mathbf{I}_p$, the constraint $\boldsymbol{\Psi} = \mathbf{U}^\top\mathbf{U}\boldsymbol{\Psi}$ is imposed. Instead of (6.18), the following optimization problem is solved:

$$\min_{\mathbf{U},\boldsymbol{\Psi}} ||\mathbf{E} - \mathbf{U}\boldsymbol{\Psi}||_F^2 \ , \tag{7.6}$$

subject to the constraints: $\mathbf{F}\mathbf{F}^\top + \mathbf{U}\mathbf{U}^\top = \mathbf{I}_n$, $\mathbf{U}^\top\mathbf{F} = \mathbf{O}_{p\times k}$ and $\boldsymbol{\Psi}$ being a diagonal matrix.

Recall the QR decomposition of $\mathbf{F}$ in (5.41) and again set $\mathbf{U} = \mathbf{F}_\perp\tilde{\mathbf{U}}$, $\mathbf{E} = \mathbf{F}_\perp\tilde{\mathbf{E}}$. Since $\mathbf{Q}\mathbf{Q}^\top = \mathbf{F}\mathbf{F}^\top + \mathbf{F}_\perp\mathbf{F}_\perp^\top = \mathbf{I}_n$, it follows that $\mathbf{U}\mathbf{U}^\top = \mathbf{F}_\perp\mathbf{F}_\perp^\top$. As $\mathbf{U}\mathbf{U}^\top = \mathbf{F}_\perp\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top\mathbf{F}_\perp^\top$, only $\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top = \mathbf{I}_{n-k}$ ensures that $\mathbf{U}\mathbf{U}^\top = \mathbf{F}_\perp\mathbf{F}_\perp^\top$ and hence $\mathbf{F}\mathbf{F}^\top + \mathbf{U}\mathbf{U}^\top = \mathbf{I}_n$. By making use of $\mathbf{F}^\top\mathbf{U} = \mathbf{O}_{k\times p}$ and $\mathbf{F}^\top\mathbf{E} = \mathbf{O}_{k\times p}$, the optimization problem (7.6) can then be reduced to the following one:

$$\min_{\tilde{\mathbf{U}},\boldsymbol{\Psi}} ||\tilde{\mathbf{E}} - \tilde{\mathbf{U}}\boldsymbol{\Psi}||_F^2 \ , \tag{7.7}$$

subject to $\tilde{\mathbf{U}}\tilde{\mathbf{U}}^{\top} = \mathbf{I}_{n-k}$ and $\mathbf{\Psi}$ being a diagonal matrix. Hence, the Procrustes-like problem (7.6) can be solved by an alternating procedure solving (7.7) for $\tilde{\mathbf{U}}$ and updating $\mathbf{\Psi} = \mathrm{diag}(\mathbf{U}^{\top}\mathbf{E}) = \mathrm{diag}(\tilde{\mathbf{U}}^{\top}\tilde{\mathbf{E}})$ until convergence.

## 7.3 Applications

### 7.3.1 Thurstone's 26-variable box data

Thurstone (1947) collected a random sample of 20 boxes and measured their three dimensions $x$ (length), $y$ (width) and $z$ (height). In this data set, the boxes constitute the observational units. Table 7.1 shows the three dimensions $x$, $y$ and $z$ for each box.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $x$ | 3 | 3 | 3 | 3 | 3 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| $y$ | 2 | 2 | 3 | 3 | 3 | 2 | 2 | 3 | 3 | 3 | 4 | 4 | 4 | 2 | 2 | 3 | 3 | 4 | 4 | 4 |
| $z$ | 1 | 2 | 1 | 2 | 3 | 1 | 2 | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 2 | 3 | 1 | 2 | 3 |

Table 7.1: Dimensions $x$ (length), $y$ (width) and $z$ (height) of Thurstone's twenty boxes.

The variables of the example are twenty-six functions of these dimensions: $x$, $y$, $z$, $xy$, $xz$, $yz$, $x^2y$, $xy^2$, $x^2z$, $xz^2$, $y^2z$, $yz^2$, $x/y$, $y/x$, $x/z$, $z/x$, $y/z$, $z/y$, $2x+2y$, $2x+2z$, $2y+2z$, $\sqrt{x^2+y^2}$, $\sqrt{x^2+z^2}$, $\sqrt{y^2+z^2}$, $xyz$ and $\sqrt{x^2+y^2+z^2}$. Some of the manifest variables are non-linear functions of the dimensions of the boxes. However, the linear regression over the values $x$, $y$ and $z$ shown in Table 7.1 which were used to generate the data is quite satisfying (Jennrich and Trendafilov, 2005). Therefore, the assumption of linearity made in EFA is only mildly violated.

The observed variables are mean-centered and scaled to have unit norm. The result is expressed in a $20 \times 26$ data matrix $\mathbf{Z}$. The first few eigenvalues of the sample correlation $\mathbf{Z}^{\mathsf{T}}\mathbf{Z}$ sorted in decreasing order are 12.4217, 7.1807, 5.5386, and 0.2963. As expected, three eigenvalues are considerably greater than one, which is Kaiser's solution (Kaiser, 1958) for the number of common factors.

The data matrix $\mathbf{Z}$ is fitted by an EFA model in terms of different LS loss functions, making use of the new iterative algorithms for simultaneous EFA and EFA-like PCA discussed in Section 7.1 and Section 7.2, respectively. As in Section 6.3, standard EFA least squares solutions $\{\boldsymbol{\Lambda}, \boldsymbol{\Psi}\}$ are obtained first and shown in Table 7.2.

Since for both the EVD and the LT reparameterization the algorithms of Trendafilov (2003, 2005) result in $\boldsymbol{\Psi}^2$ staying on the cone of positive definite diagonal matrices, the standard LS approach of minimizing (4.5) leads to uniquenesses being strictly positive. In contrast, the simultaneous EFA as well as the EFA-like PCA procedures both allow the unique factors to have zero variance. The corresponding solutions are given in Table 7.3 and Table 7.4, respectively.

Table 7.3 and Table 7.4 show that for the 26-variable box data the fit attained by the EFA-like solutions is worse than by the simultaneous EFA solutions. The values for $E$ in (6.21) are 0.0049 for EFA-like PCA based on the SVD and $1.4743 \times 10^{-8}$ for simultaneous EFA, respectively. Applying the LT parameterization, the values are 0.0142 for EFA-like PCA based on the QR factorization and $1.1754 \times 10^{-7}$ for simultaneous EFA, respectively.

For both EFA-like solutions, the loadings in Table 7.4 are virtually identical to the corresponding ones in Table 7.2 and therefore the PCA loadings can be used as ade-

| | EVD reparameterization | | | | LT reparameterization | | | |
|---|---|---|---|---|---|---|---|---|
| Formula | $\Lambda$ | | | $\Psi^2$ | **L** | | | $\Psi^2$ |
| $x$ | .50 | .53 | .68 | .0054 | 1.00 | 0 | 0 | .0011 |
| $y$ | .47 | .70 | -.53 | .0058 | .25 | .97 | 0 | .0015 |
| $z$ | -.63 | .78 | -.00 | .0058 | .10 | .23 | .97 | .0022 |
| $xy$ | .61 | .79 | -.07 | .0089 | .68 | .73 | -.00 | .0079 |
| $xz$ | -.35 | .89 | .27 | .0108 | .49 | .20 | .84 | .0103 |
| $yz$ | -.29 | .92 | -.22 | .0132 | .19 | .60 | .77 | .0130 |
| $x^2y$ | .61 | .76 | .17 | .0292 | .82 | .54 | -.00 | .0293 |
| $xy^2$ | .58 | .76 | -.25 | .0254 | .52 | .84 | -.03 | .0256 |
| $x^2z$ | -.14 | .87 | .42 | .0454 | .68 | .16 | .68 | .0455 |
| $xz^2$ | -.44 | .86 | .14 | .0449 | .33 | .25 | .88 | .0449 |
| $y^2z$ | -.08 | .92 | -.30 | .0570 | .25 | .73 | .59 | .0570 |
| $yz^2$ | -.42 | .87 | -.14 | .0540 | .16 | .46 | .84 | .0541 |
| $x/y$ | -.06 | -.30 | .93 | .0420 | .44 | -.87 | -.04 | .0423 |
| $y/x$ | .07 | .27 | -.94 | .0319 | -.47 | .87 | .01 | .0322 |
| $x/z$ | .80 | -.47 | .23 | .0927 | .31 | -.15 | -.89 | .0929 |
| $z/x$ | -.80 | .46 | -.30 | .0665 | -.36 | .20 | .87 | .0666 |
| $y/z$ | .86 | -.28 | -.34 | .0727 | .05 | .39 | -.88 | .0728 |
| $z/y$ | -.85 | .30 | .33 | .0789 | -.04 | -.37 | .88 | .0791 |
| $2x + 2y$ | .61 | .78 | .09 | .0064 | .79 | .61 | .00 | .0032 |
| $2x + 2z$ | -.09 | .88 | .46 | .0071 | .74 | .16 | .65 | .0042 |
| $2y + 2z$ | -.09 | .93 | -.34 | .0066 | .22 | .76 | .61 | .0033 |
| $(x^2 + y^2)^{1/2}$ | .61 | .75 | .23 | .0102 | .87 | .49 | -.00 | .0094 |
| $(x^2 + z^2)^{1/2}$ | .18 | .79 | .58 | .0162 | .90 | .11 | .40 | .0163 |
| $(y^2 + z^2)^{1/2}$ | .09 | .90 | -.42 | .0133 | .24 | .86 | .44 | .0132 |
| $xyz$ | -.11 | .98 | -.00 | .0289 | .47 | .54 | .68 | .0290 |
| $(x^2 + y^2 + z^2)^{1/2}$ | .37 | .90 | .20 | .0142 | .80 | .52 | .28 | .0142 |

Table 7.2: Standard LS solutions for Thurstone's 26-variable box data.

quate surrogates for the corresponding EFA loadings. Compared to the standard LS solutions the uniquenesses are smaller for EFA-like PCA, but the former ones are very similar to the ones obtained by simultaneous EFA.

All algorithms employing a LT parameterization give virtually identical loadings. Moreover, the loadings exhibit an interpretable and contextually meaningful relation between the observed variables and the common factors. If one ignores all loadings with magnitude .25 or less in the LT loading matrices in Table 7.3 and Table 7.4, the remaining loadings perfectly identify which of the box dimensions $x, y$ and $z$ were used to generate each of the variables. Using the results for EFA-like PCA based on the QR decomposition in Table 7.2, this can be done by ignoring all loadings with magnitudes of .26 or less.

| | $\Lambda = \mathbf{Z}^{\top}\mathbf{F}$ | | | | $\mathbf{L} = \mathrm{tril}(\mathbf{Z}^{\top}\mathbf{F})$ | | | |
|---|---|---|---|---|---|---|---|---|
| | error of fit = .175174 | | | | error of fit = .175184 | | | |
| Formula | $\Lambda$ | | | $\Psi^2$ | $\mathbf{L}$ | | | $\Psi^2$ |
| $x$ | .99 | .17 | .02 | .0000 | 1.00 | 0 | 0 | .0000 |
| $y$ | .10 | .90 | .43 | .0000 | .25 | .97 | 0 | .0000 |
| $z$ | .12 | -.20 | .97 | .0000 | .10 | .23 | .96 | .0000 |
| $xy$ | .55 | .76 | .33 | .0000 | .68 | .73 | -.00 | .0000 |
| $xz$ | .50 | -.11 | .85 | .0000 | .49 | .20 | .84 | .0000 |
| $yz$ | .14 | .22 | .96 | .0000 | .20 | .59 | .77 | .0000 |
| $x^2y$ | .72 | .62 | .25 | .0191 | .82 | .54 | -.00 | .0191 |
| $xy^2$ | .38 | .84 | .35 | .0001 | .52 | .84 | -.03 | .0000 |
| $x^2z$ | .69 | -.05 | .69 | .0198 | .68 | .15 | .68 | .0198 |
| $xz^2$ | .34 | -.12 | .92 | .0000 | .33 | .24 | .90 | .0000 |
| $y^2z$ | .16 | .43 | .86 | .0298 | .25 | .73 | .60 | .0298 |
| $yz^2$ | .13 | .06 | .97 | .0000 | .16 | .45 | .85 | .0000 |
| $x/y$ | .57 | -.68 | -.42 | .0279 | .44 | -.87 | -.05 | .0279 |
| $y/x$ | -.59 | .68 | .39 | .0290 | -.46 | .87 | .02 | .0290 |
| $x/z$ | .27 | .30 | -.86 | .0811 | .31 | -.15 | -.89 | .0811 |
| $z/x$ | -.33 | -.26 | .87 | .0476 | -.36 | .20 | .88 | .0476 |
| $y/z$ | -.07 | .74 | -.61 | .0566 | .04 | .40 | -.87 | .0566 |
| $z/y$ | .08 | -.72 | .62 | .0651 | -.03 | -.38 | .88 | .0651 |
| $2x + 2y$ | .68 | .67 | .28 | .0000 | .79 | .61 | .00 | .0000 |
| $2x + 2z$ | .75 | -.02 | .66 | .0000 | .74 | .15 | .65 | .0000 |
| $2y + 2z$ | .14 | .44 | .88 | .0000 | .23 | .76 | .61 | .0000 |
| $(x^2 + y^2)^{1/2}$ | .78 | .58 | .23 | .0000 | .87 | .49 | -.01 | .0000 |
| $(x^2 + z^2)^{1/2}$ | .90 | .07 | .42 | .0001 | .91 | .10 | .39 | .0001 |
| $(y^2 + z^2)^{1/2}$ | .13 | .61 | .77 | .0000 | .25 | .86 | .44 | .0000 |
| $xyz$ | .41 | .26 | .86 | .0017 | .47 | .54 | .68 | .0017 |
| $(x^2 + y^2 + z^2)^{1/2}$ | .72 | .47 | .49 | .0001 | .80 | .52 | .28 | .0001 |

Table 7.3: Simultaneous EFA solutions for Thurstone's 26-variable box data.

| | SVD | | | | QR decomposition | | | |
|---|---|---|---|---|---|---|---|---|
| | error of fit = .198038 | | | | error of fit = .222478 | | | |
| Formula | $\Lambda$ | | | $\Psi^2$ | $\mathbf{L}$ | | | $\Psi^2$ |
| $x$ | .50 | .53 | .68 | .0000 | 1.00 | 0 | 0 | .0000 |
| $y$ | .47 | .70 | -.53 | .0000 | .26 | .97 | 0 | .0000 |
| $z$ | -.62 | .78 | -.00 | .0000 | .10 | .23 | .97 | .0000 |
| $xy$ | .61 | .79 | -.06 | .0000 | .68 | .73 | -.01 | .0000 |
| $xz$ | -.34 | .89 | .27 | .0000 | .49 | .20 | .83 | .0000 |
| $yz$ | -.29 | .92 | -.22 | .0000 | .20 | .59 | .76 | .0000 |
| $x^2y$ | .61 | .76 | .17 | .0157 | .82 | .54 | -.00 | .0285 |
| $xy^2$ | .59 | .76 | -.25 | .0014 | .52 | .84 | -.03 | .0000 |
| $x^2z$ | -.14 | .87 | .42 | .0177 | .68 | .15 | .67 | .0450 |
| $xz^2$ | -.44 | .86 | .14 | .0000 | .33 | .24 | .88 | .0004 |
| $y^2z$ | -.08 | .92 | -.30 | .0290 | .24 | .73 | .58 | .0572 |
| $yz^2$ | -.42 | .87 | -.14 | .0000 | .16 | .45 | .84 | .0000 |
| $x/y$ | -.06 | -.30 | .94 | .0149 | .45 | -.87 | -.04 | .0210 |
| $y/x$ | .07 | .27 | -.95 | .0169 | -.46 | .87 | .02 | .0196 |
| $x/z$ | .81 | -.47 | .24 | .0613 | .31 | -.16 | -.88 | .0768 |
| $z/x$ | -.80 | .46 | -.31 | .0343 | -.36 | .20 | .89 | .0123 |
| $y/z$ | .86 | -.28 | -.34 | .0394 | .04 | .40 | -.87 | .0559 |
| $z/y$ | -.86 | .30 | .34 | .0459 | -.04 | -.37 | .90 | .0348 |
| $2x + 2y$ | .61 | .78 | .09 | .0000 | .79 | .61 | .00 | .0000 |
| $2x + 2z$ | -.09 | .88 | .46 | .0000 | .74 | .16 | .65 | .0000 |
| $2y + 2z$ | -.09 | .93 | -.34 | .0000 | .22 | .76 | .61 | .0000 |
| $(x^2 + y^2)^{1/2}$ | .61 | .75 | .23 | .0000 | .87 | .49 | -.01 | .0000 |
| $(x^2 + z^2)^{1/2}$ | .18 | .79 | .58 | .0001 | .91 | .10 | .40 | .0001 |
| $(y^2 + z^2)^{1/2}$ | .09 | .90 | -.42 | .0000 | .24 | .86 | .44 | .0000 |
| $xyz$ | -.10 | .98 | -.00 | .0016 | .46 | .53 | .67 | .0057 |
| $(x^2 + y^2 + z^2)^{1/2}$ | .37 | .90 | .20 | .0001 | .80 | .52 | .28 | .0001 |

Table 7.4: EFA-like solutions for Thurstone's 26-variable box data.

### 7.3.2 Atmospheric science data

Climate is a natural system that is characterized by complex and high-dimensional phenomena. To improve understanding of the physical behaviour of the system, it is often useful to reduce the dimensionality of the data. This requires the development and use of statistical techniques in atmospheric science for describing patterns of meteorological variables over a large spatial area in low-dimensional space.

Empirical orthogonal function (EOF) analysis, known in statistics as PCA, is among the most widely used methods in atmospheric science (Hannachi, Jolliffe, and Stephenson, 2007; Jolliffe, 2002). Given any space-time meteorological data set, EOF analysis finds a set of orthogonal spatial patterns (EOFs), in PCA referred to as loadings, along with a set of associated uncorrelated time series or principal components, such that the first few PCs account for as much as possible of the total sample variance.

Unlike PCA, the use of EFA in atmospheric science is quite rare (M. B. Richman, personal communication, 2008). Most publications which claim to make use of EFA in their studies actually apply PCA/EOF analysis. Regarding the use of EFA in the literature, we are aware of the publications by Bukantis (2002), Carter and Elsner (1997), Bärring (1987), Walsh, Richman, and Allen (1982), and Walsh and Richman (1981).

At this point, EFA is applied to atmospheric science data from the National Center for Environmental Prediction/National Center for Atmospheric Research (NCEP/NCAR) reanalysis project. The data set was kindly provided by Dr. Abdel Hannachi who is currently affiliated with the Department of Meteorology, King Abdulaziz University, Jeddah, Saudi Arabia. The data set consists of winter monthly sea-level pressures (SLP) over the Northern Hemisphere north of 20°N. The winter season is conventionally defined by the months December, January and February (DJF) (e.g., Hannachi, Jolliffe, and Stephenson, 2007; Wallace and Gutzler, 1981). The data set spans the period December 1948 to February 2006 ($n = 174$ observations) and is available on a regular grid with a 2.5°latitude × 2.5°longitude resolution ($p = 29 \times 144 = 4176$ variables representing gridpoints).

Prior to the analysis the data were preprocessed as follows. First, the mean annual cycle was calculated by averaging the monthly data over the years. Anomalies were then computed as departures from the mean annual cycle. To account for the converging longitudes poleward, an area weighting was finally performed by multiplying each grid point by the square root of the cosine of the corresponding latitude. These weighted SLP anomalies are the data to which methods are applied.

Applying the new approach for fitting the EFA model in the case $p \geq n$ which updates $\mathbf{F}$ and $\mathbf{U}$ simultaneously, five factors are extracted which account for 60.2% of the total variance in the data. This choice is dictated by the need for a balance between explained variance and spatial scales. Extracting more factors increases the explained variance but includes more small scales. Five factors are found to provide a good balance.

For $k = 5$ and twenty random starts, the procedure required on average 90 iterations, taking about 20 minutes to converge. The algorithm was stopped when successive function values differed by less than $\epsilon = 10^{-3}$. Using a higher accuracy criterion such as $\epsilon = 10^{-6}$ needed considerably more CPU time but did not change the quality of the solution. Numerical experiments revealed that the algorithm converges to the same minimum of the loss function, up to the second decimal place.

For comparison, factorizing a 4176 × 4176 covariance matrix and finding a numerical solution for minimizing the LS loss function $\mathcal{F}_{LS}$ in (4.5) based on an iterative Newton-Raphson procedure takes about 2.5 hours.

By means of the loading matrix, EFA provides a method of describing spatial patterns of winter sea-level pressures. For each factor, there is a loading for each manifest variable, and because variables are gridpoints it is possible to plot each loading on a map at its corresponding gridpoint, and then draw contours through geographical locations having the same coefficient values. Compared to a loading matrix with 4176 loadings for each factor, this spatial map representation introduced by Maryon (1979) greatly aids interpretation, as is illustrated in Figure 7.1.

For the winter SLP data, the plots represent the first (i) and second (ii) column of the $4176 \times 5$ loading matrix. These plots give the maps of loadings, arbitrarily renormalized to give 'round numbers' on the contours. Winter months having large positive scores for the factors will tend to have high SLP values, where loadings on the map are positive, and low SLP values at gridpoints where the coefficients are negative. The first and second common factor explains 14% and 13% of the total sample variance, respectively.

The first pattern (i) shows the North Atlantic Oscillation (NAO). The NAO is a climatic phenomenon in the North Atlantic Ocean of fluctuations in the difference of sea-level pressure between the Icelandic low and the Azores high (Hannachi, Jolliffe, and Stephenson, 2007). The second EFA pattern (ii) yields the North Pacific Oscillation (NPO) or Pacific pattern, a monopolar structure sitting over the North Pacific (Hannachi, Jolliffe, and Stephenson, 2007). For the twenty different random starts, the obtained EFA loadings look similar.

It is of interest to compare the spatial patterns obtained by EFA to the ones obtained

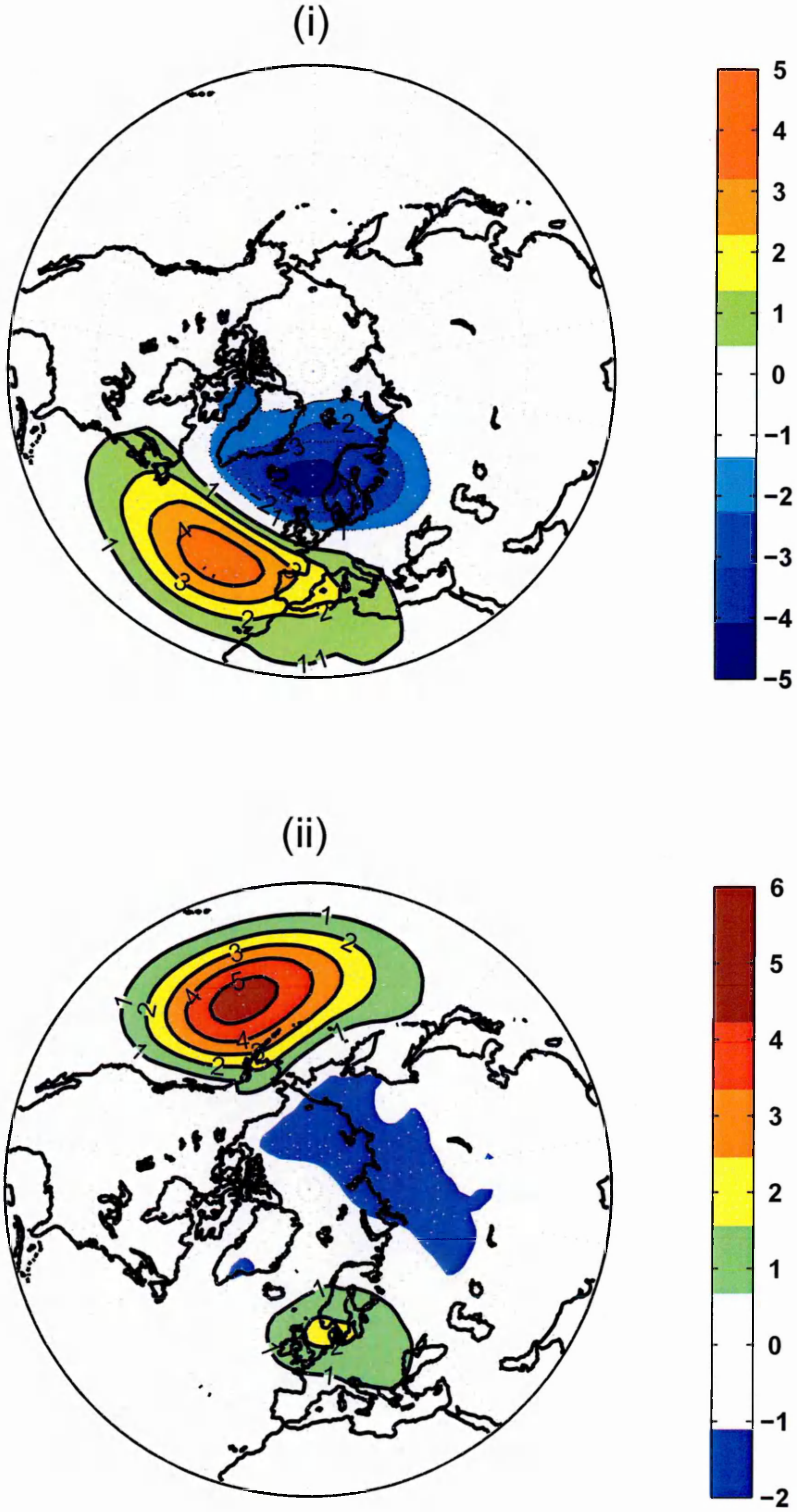

Figure 7.1: Spatial map representations of the first (i) and second (ii) column of the EFA loading matrix for winter SLP data ($k = 5$).

by PCA/EOF analysis. Figure 7.2 shows the two leading modes of variability of the winter monthly SLP. They explain 21% (1st EOF) and 13% (2nd EOF) of the total winter variance. The spatial map (i) shows a low-pressure centre over the polar region and two high-pressure centres over the Mediterranean/North-east Atlantic and over the North Pacific, respectively. This tripolar structure corresponds to the familiar Annular Oscillation (AO) (Hannachi, Jolliffe, and Stephenson, 2007). Like EFA pattern (ii), the EOF2 has the NPO with a polar high over the North Pacific but in addition it also has a low centre over the North-east Atlantic.

Finally, the effect of increasing the number of extracted factors was also studied. With more extracted factors, the scale of the spatial patterns becomes smaller and more concentrated. In particular, the NAO pattern starts to lose its structure.

One is tempted to ask for the use of statistical tools to validate the EFA model and to compare EFA with PCA. Note that EFA as well as PCA are not usually validated as such but examined to see whether they give an insightful low-dimensional representation of the data. In particular for this type of atmospheric science data, there is less interest in the quality of the fit than whether the factors found can be usefully interpreted in terms of the underlying physics of the process. That is, the emphasis is on physical interpretation, not on statistical significance. In any case the model fit can be improved by increasing the number of extracted factors. Thus, results are evaluated and compared by the map representations of the EFA and PCA loading matrices to find spatial patterns which can be interpreted in a meteorological sense. One can be content with the comparison if the maps look the same, but one may also be content if they are different, provided that both sets of maps can be usefully interpreted.

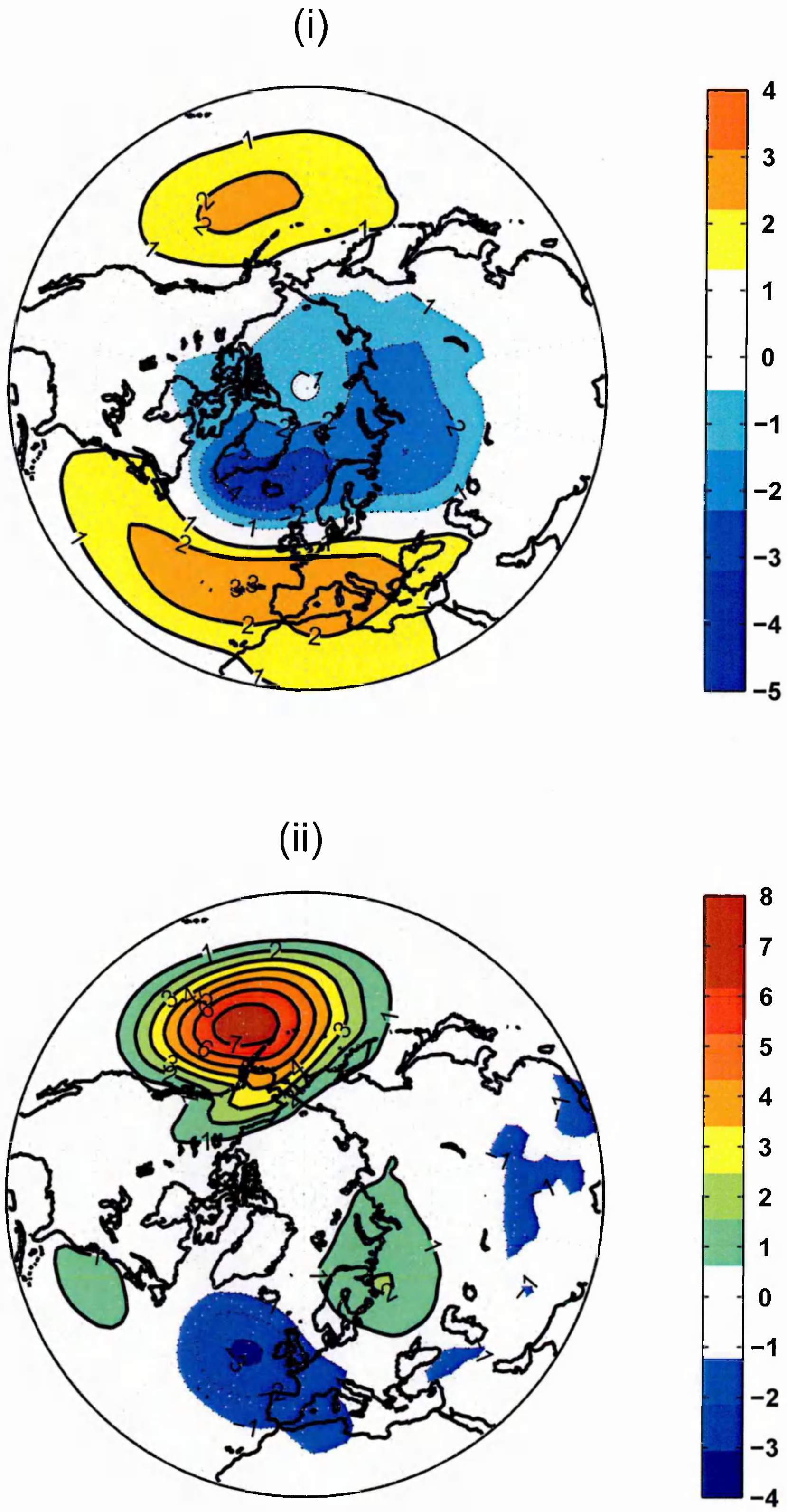

Figure 7.2: Spatial map representations of the two leading EOFs one (i) and two (ii) for winter SLP data ($k = 5$). The EOFs have been multiplied by 100.

# Chapter 8

# Simultaneous Estimation of all Model Unknowns in Robust EFA

Classical EFA techniques take input data in the form of a matrix of second-order cross products, that is, of correlations or covariances. Since the influence of outliers in the data is multiplied by the use of product moments these approaches are not robust. In the context of EFA, a variety of robust estimates of the multivariate scatter matrix have been proposed. Among them are the multivariate M-estimator (Kosfeld, 1996; Huber, 1981), the minimum volume ellipsoid estimator (Filzmoser, 1999; Rousseeuw, 1985), and the minimum covariance determinant estimator (Pison, Rousseeuw, Filzmoser, and Croux, 2003; Rousseeuw, 1985). Mavridis and Moustaki (2008) performed outlier detection in factor analysis models using a forward search algorithm instead of using some robust modification of the sample correlation matrix.

In the current Chapter, an alternative approach to resist the effect of outliers is presented. Without passing via an estimate of the model correlation matrix, the EFA model is fitted directly to the data matrix (Unkel and Trendafilov, 2009c).

Croux, Filzmoser, Pison, and Rousseeuw (2003) proposed robust factorization of the data matrix into a pair of estimates $\{\hat{\mathbf{\Lambda}}, \hat{\mathbf{F}}\}$ by optimizing a resistant alternating (criss-cross) regression scheme (Wold, 1966; Gabriel and Zamir, 1979). Croux, Filzmoser, Pison, and Rousseeuw (2003) apply an 'ignoring errors' strategy to factor analysis. The unique factors are not incorporated in the corresponding loss function but treated as residuals to be minimized and an estimate for $\mathbf{\Psi}^2$ is obtained after the discrepancy measure has already been optimized. This is not satisfactory as $\mathbf{U\Psi}$ is part of the EFA model (3.19). By neglecting this part and computing the fitted values $\hat{\mathbf{Z}}$ simply as $\hat{\mathbf{Z}} = \hat{\mathbf{F}}\hat{\mathbf{\Lambda}}^\top$, the approach of Croux, Filzmoser, Pison, and Rousseeuw (2003) resembles a robust PCA solution to EFA with additional interpretational tools rather than 'truly' robust EFA. Moreover, monotonic convergence of the regression algorithm has not been proven.

In this Chapter, a robust approach for simultaneous estimation of $\mathbf{\Lambda}$, $\mathbf{F}$, $\mathbf{\Psi}$ and $\mathbf{U}$ is presented (Unkel and Trendafilov, 2009c). The EFA model is fitted to the data matrix by minimizing a certain WLS goodness-of-fit measure. By imposing weights on the residuals of the ULS fitting, the WLS loss function considered is a generalization of the ULS one used by De Leeuw (2004, 2008).

Kiers (1997b) introduced a very general approach for fitting a model to a data matrix by WLS. It consists of iteratively performing steps of an existing algorithm for ULS fitting of the same model. The approach is based on minimizing an auxiliary function that majorizes the WLS loss function. In this Chapter, the majorizing function of Kiers (1997b) is used in a procedure for iteratively reweighted least squares (IRLS) in which the weights depend on the residuals and are updated after each cycle of updating the model parameters. Monotonic convergence of the IRLS algorithm is guaranteed. To down-weight the effect of outliers in the data, the Huber criterion is used as a robustifier (Huber, 1981). Optimizing Huber's function by the IRLS algorithm leads to robust EFA.

In Section 8.1, the ULS function of De Leeuw (2004, 2008) is generalized by setting up a robust WLS discrepancy measure. A reweighted least squares algorithm which monotonically improves the value of the WLS objective function using iterative majorization is presented in Section 8.2. Section 8.3 illustrates the performance of the proposed robust EFA approach on real data.

## 8.1 Weighted least squares loss function and choice of robustifier

The approach of De Leeuw (2004, 2008) can be generalized by imposing weights on the ULS residuals. Assume that $n > p$ and consider the following WLS objective function:

$$f_{WLS} = ||(\mathbf{Z} - \mathbf{B}\mathbf{A}^\top) \odot \mathbf{W}||_F^2 = \sum_{i=1}^{n} \sum_{j=1}^{p} w_{ij}^2 (z_{ij} - \mathbf{b}_i^\top \mathbf{a}_j)^2 = \sum_{i=1}^{n} \sum_{j=1}^{p} w_{ij}^2 e_{ij}^2 \ , \quad (8.1)$$

where $\mathbf{b}_i$ and $\mathbf{a}_j$ are column vectors of length $k + p$ containing the elements of row $i$ of $\mathbf{B} = [\mathbf{F} \quad \mathbf{U}]$ and of row $j$ of $\mathbf{A} = [\mathbf{\Lambda} \quad \mathbf{\Psi}]$, respectively, $\mathbf{W}$ is an $n \times p$ matrix of non-negative weights $w_{ij}$ attached to each residual $e_{ij}$, and $\odot$ denotes the elementwise (Hadamard) matrix product. For given $\mathbf{Z}$ and $\mathbf{W}$, the aim is to minimize (8.1) over $\mathbf{B}$ and $\mathbf{A}$ subject to the constraints on the EFA model parameters, that is, rank$(\mathbf{\Lambda}) = k$, $\mathbf{F}^\top\mathbf{F} = \mathbf{I}_k$, $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_p$, $\mathbf{U}^\top\mathbf{F} = \mathbf{O}_{p\times k}$, and $\mathbf{\Psi}$ being a $p \times p$ diagonal matrix.

The WLS goodness-of-fit criterion is typically associated with robust statistical methods. The idea of robust methods is to down-weight cases with large residuals relative to cases with small residuals. For many statistical techniques the standard ULS goodness-of-fit criterion is very sensitive to large deviations between the model and the data (Verboon, 1994). When the ULS criterion is trying to compensate for these large errors, the solution of the problem may be 'shifted' towards an incorrect one.

What remains is to make a careful choice of the weights attached to each residual in order to implement a robust form of simultaneous parameter estimation in EFA. Assume that the data matrix $\mathbf{Z}$ contains some outlying points giving rise to high residuals. Then, loss functions which try to curb the influence of those outliers are desirable. Several possibilities for downweighting are available, for example the $\ell_1$ matrix norm, where the sum of moduli of errors is minimized (e.g., Rousseeuw and Leroy, 1987, Chapter 1), the biweight function (Mosteller and Tukey, 1977) or Huber's robust estimator (Huber, 1981). Consider the Huber function:

$$f_H(\mathbf{B},\mathbf{A}) = \sum_{i=1}^{n}\sum_{j=1}^{p} f_H(e_{ij}) \ , \tag{8.2}$$

$$\text{where} \quad f_H(e_{ij}) = \begin{cases} e_{ij}^2 & \text{for} \quad |e_{ij}| < \gamma \ , \\ 2\gamma|e_{ij}| - \gamma^2 & \text{for} \quad |e_{ij}| \geq \gamma \ , \end{cases}$$

and $\gamma$ is a given 'tuning constant', which distinguishes small residuals from large ones. The Huber function is symmetric and for $|e_{ij}| \geq \gamma$, the residuals have less effect than they would have with ULS. The basic idea is that outliers, yielding large residuals when the structure of the majority of the data points is fitted well, will have a less disturbing effect upon the solution than in the ULS case. When $\gamma$ is chosen very large, (8.2) becomes equal to the ULS function; when $\gamma$ is close to zero, (8.2) reduces to the least sum of absolute residuals criterion. Therefore, the loss function is a hybrid $\ell_1 - \ell_2$ error measure and the algorithm to minimize it can also be used for solving the $\ell_1$ problem. The criterion (8.2) can be formulated in a WLS form with (squared) weights

$$w_{ij}^2 = \begin{cases} 1 & \text{for} \quad |e_{ij}| < \gamma \ , \\ 2\gamma/|e_{ij}| - \gamma^2/e_{ij}^2 & \text{for} \quad |e_{ij}| \geq \gamma \ . \end{cases} \tag{8.3}$$

Of course, the residuals and hence the weights are not known and have to be given initial estimates. After fitting the model to the data, new estimates of the residuals will be available which can be used to compute new weights. Thus, $f_H(\mathbf{B}, \mathbf{A})$ or equivalently $f_{WLS}$ with weights defined in (8.3) can be minimized by an algorithm that is based on IRLS.

One may resort to standard numerical procedures like Newton or conjugate gradient methods to solve the WLS problem. However, these techniques become computationally slow when the matrices of unknowns are large (Kiers, 2002). In the next Section, a monotonic convergent IRLS algorithm is derived by means of the iterative majorization approach to optimization instead.

## 8.2 Iterative reweighted least squares algorithm

### 8.2.1 Iterative majorization approach to optimization

The basic idea of iterative majorization (e.g., Heiser, 1995) is that in each iteration a complicated objective function is substituted by a simple (for instance, linear or quadratic) auxiliary function called a majorizing function.

Assume a model parameter matrix $\mathbf{X}$ varies in some domain $\Omega$. The aim is to minimize an objective function, $f(\mathbf{X})$, by minimizing a majorizing function $m(\mathbf{X}|\mathbf{X}^c)$, where $\mathbf{X}^c$ denotes the current estimate of $\mathbf{X}$ (called the supporting point). The majorizing function must meet the following requirements:

$$m(\mathbf{X}|\mathbf{X}^c) \geq f(\mathbf{X}) \quad \forall \quad \mathbf{X} \ , \tag{8.4}$$

and

$$m(\mathbf{X}^c|\mathbf{X}^c) = f(\mathbf{X}^c) \ . \tag{8.5}$$

Thus, the values of the majorizing function must never be smaller than the values of the original loss function and the values of both functions must coincide at the supporting point.

Iterative majorization consists of finding an update $\mathbf{X}^{\mathrm{u}}$ for $\mathbf{X}$ (called the successor point) which minimizes the majorizing function $m(\mathbf{X}|\mathbf{X}^{\mathrm{c}})$, that is,

$$\mathbf{X}^{\mathrm{u}} = \arg\min_{\mathbf{X}\in\Omega} m(\mathbf{X}|\mathbf{X}^{\mathrm{c}}) \ .$$

Using the conditions (8.4) and (8.5) as well as the fact that the search for $\mathbf{X}^{\mathrm{u}}$ is such that $m(\mathbf{X}^{\mathrm{u}}|\mathbf{X}^{\mathrm{c}}) \leq m(\mathbf{X}^{\mathrm{c}}|\mathbf{X}^{\mathrm{c}})$, the following chain of inequalities holds:

$$f(\mathbf{X}^{\mathrm{u}}) \leq m(\mathbf{X}^{\mathrm{u}}|\mathbf{X}^{\mathrm{c}}) = \min_{\mathbf{X}\in\Omega} m(\mathbf{X}|\mathbf{X}^{\mathrm{c}}) \leq m(\mathbf{X}^{\mathrm{c}}|\mathbf{X}^{\mathrm{c}}) = f(\mathbf{X}^{\mathrm{c}}) \ . \tag{8.6}$$

Hence, by iteratively minimizing $m(\mathbf{X}|\mathbf{X}^{\mathrm{c}})$, a sequence of monotonically decreasing loss function values is obtained. If the loss function $f(\mathbf{X})$ is bounded below, the iterative procedure will stop at a stationary point which is not necessarily a local optimum.

Note that optimizing $f_H(\mathbf{B}, \mathbf{A})$ or the WLS fitting problem in (8.1) involves two block matrices or parameter sets, that is, $\mathbf{B}$ and $\mathbf{A}$. However, these optimization problems can be solved by a combination of a majorization and a block relaxation (De Leeuw, 1994) algorithm. The only complex step here is to find an update $\mathbf{B}^{\mathrm{u}}$ using a solution for minimizing $f_H(\mathbf{B}, \mathbf{A})$ over $\mathbf{B}$ (keeping $\mathbf{A}$ fixed). This is done using the majorization approach. In the next section, a simple majorizing function $m(\mathbf{B}, \mathbf{A}|\mathbf{B}^{\mathrm{c}}, \mathbf{A}^{\mathrm{c}})$ is presented which can be used to decrease the objective function $f_H(\mathbf{B}, \mathbf{A})$.

### 8.2.2 Majorizing function and optimization algorithm

Note that $f_H(\mathbf{B}, \mathbf{A})$ in (8.2) is separable, that is, $f_H(\mathbf{B}, \mathbf{A}) = \sum_{i=1}^{n}\sum_{j=1}^{p} f_H(e_{ij})$. De Leeuw and Lange (2009) show that the 'sharpest' (best possible) quadratic majorizer of $f_H(e_{ij})$ is

$$m_H(e_{ij}|e^c_{ij}) = \begin{cases} e^2_{ij} & \text{for} \quad |e^c_{ij}| < \gamma \ , \\ \frac{\gamma}{|e^c_{ij}|}e^2_{ij} + \gamma|e^c_{ij}| - \gamma^2 & \text{for} \quad |e^c_{ij}| \geq \gamma \ , \end{cases} \tag{8.7}$$

where $e^c_{ij}$ denotes the estimate of residual $e_{ij}$ found in the previous iteration.

If $m_H(e_{ij}|e^c_{ij})$ majorizes $f_H(e_{ij})$ for $i = 1, \ldots, n$ and $j = 1, \ldots, p$, then $m_H(\mathbf{B}, \mathbf{A}|\mathbf{B}^c, \mathbf{A}^c)$ majorizes $f_H(\mathbf{B}, \mathbf{A})$, that is,

$$f_H(\mathbf{B}, \mathbf{A}) \leq m_H(\mathbf{B}, \mathbf{A}|\mathbf{B}^c, \mathbf{A}^c)$$

and

$$f_H(\mathbf{B}^c, \mathbf{A}^c) = m_H(\mathbf{B}^c, \mathbf{A}^c|\mathbf{B}^c, \mathbf{A}^c) \ .$$

By choosing weights $w^2_{ij}$ as

$$w^2_{ij} = \begin{cases} 1 & \text{for} \quad |e^c_{ij}| < \gamma \ , \\ \gamma/|e^c_{ij}| & \text{for} \quad |e^c_{ij}| \geq \gamma \ , \end{cases} \tag{8.8}$$

the piecewise function (8.7) turns into (Verboon and Heiser, 1992):

$$m_H(e_{ij}|e^c_{ij}) = \begin{cases} w^2_{ij}e^2_{ij} & \text{for} \quad |e^c_{ij}| < \gamma \ , \\ w^2_{ij}e^2_{ij} + \gamma|e^c_{ij}| - \gamma^2 & \text{for} \quad |e^c_{ij}| \geq \gamma \ . \end{cases} \tag{8.9}$$

From (8.9), it can be seen that $m_H(\mathbf{B}, \mathbf{A}|\mathbf{B}^c, \mathbf{A}^c)$ is (up to a constant for $|e^c_{ij}| \geq \gamma$) a WLS function in the residuals $e_{ij}$ and thus in $\mathbf{B}$ and $\mathbf{A}$. Therefore, it suffices to look at the WLS problem $f_{WLS}$ in (8.1) with weights defined in (8.8).

Kiers (1997b) introduced an iterative majorization algorithm to solve any WLS fitting problem in cases where an algorithm for the corresponding ULS fitting is already available. Using the method of Kiers (1997b), $f_{WLS}$ in (8.1) is majorized and touched by

$$m_K(\mathbf{B}, \mathbf{A}|\mathbf{B}^c, \mathbf{A}^c) = \alpha + w^2_m||\tilde{\mathbf{Z}} - \mathbf{B}\mathbf{A}^\top||^2_F \ , \tag{8.10}$$

where $\tilde{\mathbf{Z}} = \mathbf{B}^c\mathbf{A}^{c^\top} + w_m^{-2}(\mathbf{W} \odot \mathbf{W} \odot (\mathbf{Z} - \mathbf{B}^c\mathbf{A}^{c^\top}))$, $w_m^2$ is the maximum of the squared elements of $\mathbf{W}$, and $\alpha$ is a constant. Thus, by setting-up an iterative majorization algorithm, the WLS problem can be solved by iteratively solving the corresponding ULS problem of fitting the same model to $\tilde{\mathbf{Z}}$ instead of $\mathbf{Z}$, that is, by minimizing $||\tilde{\mathbf{Z}} - \mathbf{B}\mathbf{A}^\top||_F^2$.

Keeping $\mathbf{A}$ fixed and using the fact that $||\tilde{\mathbf{Z}} - \mathbf{B}\mathbf{A}^\top||_F^2 = ||\tilde{\mathbf{Z}}||_F^2 + ||\mathbf{A}||_F^2 - 2\text{trace}(\mathbf{B}^\top\tilde{\mathbf{Z}}\mathbf{A})$, the minimization of $||\tilde{\mathbf{Z}} - \mathbf{B}\mathbf{A}^\top||_F^2$ is equivalent to the maximization of trace$(\mathbf{B}^\top\tilde{\mathbf{Z}}\mathbf{A})$ over $\mathbf{B}$. The maximizing $\mathbf{B}$ can be found as a solution to a Procrustes problem which can be solved analytically via the SVD of $\tilde{\mathbf{Z}}\mathbf{A}$.

Once optimal factor scores $\mathbf{F}$ and $\mathbf{U}$ are found, the parameter matrices $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ are updated by $\mathbf{\Lambda} = \tilde{\mathbf{Z}}^\top\mathbf{F}$ and $\mathbf{\Psi} = \text{diag}(\mathbf{U}^\top\tilde{\mathbf{Z}})$ which immediately follows from replacing $\mathbf{Z}$ by $\tilde{\mathbf{Z}}$ in the identities (5.2) and (5.3). This alternating procedure continues until the original loss function cannot be reduced further. Hence, the optimization algorithm is a combination of an iterative majorization and a block relaxation approach.

Note that in Kiers (1997b), the weights are considered fixed. However, the iterative majorization approach can be used in an IRLS setting in which the weights depend on the residuals and are updated according to (8.8) after each cycle of updating the model parameters.

For WLS fitting of matrix decomposition problems, Groenen, Giaquinto, and Kiers (2003) proposed the so called weighted majorization algorithm (see also Groenen, Giaquinto, and Kiers, 2005). Using the method of Groenen, Giaquinto, and Kiers (2003), $f_{WLS}$ in (8.1) is majorized and touched by

$$m_G(\mathbf{B}, \mathbf{A}|\mathbf{B}^c, \mathbf{A}^c) = \beta + \text{trace}(\tilde{\mathbf{Z}} - \mathbf{B}\mathbf{A}^\top)^\top\mathbf{D}_m^2(\tilde{\mathbf{Z}} - \mathbf{B}\mathbf{A}^\top) \ , \qquad (8.11)$$

where $\beta$ is a constant, $\mathbf{D}_m^2$ is an $n \times n$ diagonal matrix containing the squared maximum row values $m_i^2$ $(i = 1, \dots, n)$ of $\mathbf{W}$, and $\tilde{\mathbf{Z}} = \mathbf{B}^c\mathbf{A}^{c^\mathrm{T}} + \mathbf{D}_m^{-2}(\mathbf{W} \odot \mathbf{W} \odot (\mathbf{Z} - \mathbf{B}^c\mathbf{A}^{c^\mathrm{T}}))$ with elements

$$\tilde{z}_{ij} = \left[1 - \frac{w_{ij}^2}{m_i^2}\right] \mathbf{b}_i^{c^\mathrm{T}} \mathbf{a}_j^c + \frac{w_{ij}^2}{m_i^2} z_{ij} \ . \tag{8.12}$$

Minimizing $m_G(\mathbf{B}, \mathbf{A}|\mathbf{B}^c, \mathbf{A}^c)$ is a WLS problem in a diagonal metric instead of a ULS problem for the minimization of $m_K(\mathbf{B}, \mathbf{A}|\mathbf{B}^c, \mathbf{A}^c)$.

If there is a single weight $w_{ij}$ that is much larger than all the other weights, the effect is limited only to the single row to which the large weight belongs. This is because the term $w_{ij}^2/m_i^2$ in (8.12) depends on the largest squared weight per row, while in Kiers (1997b) it depends on the overall largest (squared) weight $w_m^2$. For the special case of $\mathbf{D}_m^2 = w_m^2\mathbf{I}_n$ and thus all maximum row weights equal, the weighted majorization approach coincides with the method of Kiers (1997b).

In rows that have their weights close to the largest row weight, the weighted majorization algorithm will fit $\mathbf{b}_i^\mathrm{T}\mathbf{a}_j$ to a large extent to $z_{ij}$. For rows with a large single weight it will fit $\mathbf{b}_i^\mathrm{T}\mathbf{a}_j$ mostly to $\mathbf{b}_i^{c^\mathrm{T}}\mathbf{a}_j^c$ and to a minor extent to $z_{ij}$. Since $w_{ij}^2/m_i^2 \geq w_{ij}^2/w_m^2$, in weighted majorization $\mathbf{b}_i^\mathrm{T}\mathbf{a}_j$ is fitted more to the data than to the values of the previous iteration compared to the method of Kiers (1997b). The consequence is that the more deviant the largest weight is from all the other weights, the slower the algorithm of Kiers (1997b) is expected to be compared to the one by Groenen, Giaquinto, and Kiers (2003).

The weighted majorization method can be applied to any WLS matrix decomposition model with differential non-negative weights for each residual (Groenen, Giaquinto, and Kiers, 2005). Naturally, it only improves those optimization problems for which a diagonally WLS solution can be obtained easily. Unfortunately, by imposing a diagonal weight matrix $\mathbf{D}_m^2$ in (8.10) it becomes much more difficult to minimize (8.1) by iterative majorization. Indeed, expanding the non-constant term in (8.11) gives

$$||\mathbf{D}_m(\tilde{\mathbf{Z}}-\mathbf{B}\mathbf{A}^\top)||_F^2 = ||\mathbf{D}_m\tilde{\mathbf{Z}}||_F^2+\text{trace}(\mathbf{A}\mathbf{B}^\top\mathbf{D}_m^2\mathbf{B}\mathbf{A}^\top)-2\text{trace}(\mathbf{B}^\top\mathbf{D}_m^2\tilde{\mathbf{Z}}\mathbf{A}) \quad , \quad (8.13)$$

in which the second term on the right side of (8.13) is not constant. Hence, for given or estimated $\mathbf{A}$ the minimization of $m_G(\mathbf{B},\mathbf{A}|\mathbf{B}^c,\mathbf{A}^c)$ is not equivalent to the maximization of $\text{trace}(\mathbf{B}^\top\mathbf{D}_m^2\tilde{\mathbf{Z}}\mathbf{A})$ over $\mathbf{B}$. No closed-form solution for this optimization problem exists. Since it only makes sense to consider an algorithm if the minima of the majorizing functions in the substeps can be obtained readily, the majorization algorithm of Kiers (1997b) is used in the sequel.

Summarizing, the set-up for the proposed IRLS algorithm for minimizing $f_H(\mathbf{B},\mathbf{A})$ over $\mathbf{B}$ and $\mathbf{A}$ is as follows:

1. Set convergence criterion $\epsilon$ to some small value, say $10^{-6}$. Initialize $\mathbf{B}$ and $\mathbf{A}$ as $\mathbf{B}^c$ and $\mathbf{A}^c$, respectively. Set the iteration counter $c = 0$.

2. For given $\gamma$, compute $f_H^c = f_H(\mathbf{B}^c,\mathbf{A}^c)$ and weight matrix $\mathbf{W}^c$ according to (8.8).

3. Compute $\tilde{\mathbf{Z}}^c = \mathbf{B}^c\mathbf{A}^{c^\top} + w_m^{-2}(\mathbf{W}^c \odot \mathbf{W}^c \odot (\mathbf{Z} - \mathbf{B}^c\mathbf{A}^{c^\top}))$.

4. Keeping $\mathbf{A}^c$ fixed, find $\mathbf{B}^{c+1}$ that minimizes $||\tilde{\mathbf{Z}} - \mathbf{B}\mathbf{A}^{c^\top}||_F^2$ over $\mathbf{B}$.

5. Partition the block matrix $\mathbf{B}^{c+1}$ into $\mathbf{F}^{c+1}$ and $\mathbf{U}^{c+1}$.

   Update $\mathbf{A}^{c+1} = \left[\mathbf{\Lambda}^{c+1} : \mathbf{\Psi}^{c+1}\right]$ using $\mathbf{\Lambda}^{c+1} = \tilde{\mathbf{Z}}^\top\mathbf{F}^{c+1}$ and $\mathbf{\Psi}^{c+1} = \text{diag}(\mathbf{U}^{c+1^\top}\tilde{\mathbf{Z}})$.

6. Compute $f_H^{c+1} = f_H(\mathbf{B}^{c+1},\mathbf{A}^{c+1})$ and new weight matrix $\mathbf{W}^{c+1}$ according to (8.8).

7. If $(f_H^c - f_H^{c+1}) > \epsilon f_H^c$, set $c = c + 1$ and go to step 3; else consider the algorithm converged.

Essentially, the IRLS algorithm for minimizing $f_H(\mathbf{B}, \mathbf{A})$ consists of two main steps. In one step, a WLS problem is solved for a fixed set of weights by means of the majorization approach, and in the other the weights are chosen as a monotonically decreasing function of the absolute values of the residuals from the previous step.

## 8.3 Application to European health and fertility data

The proposed approach is applied to data originating from the statistical office of the European Union (Eurostat). Nine variables ($p = 9$) related to health and fertility are measured for 16 European countries ($n = 16$). The data set can be downloaded via *http://www.statistik.tuwien.ac.at/public/filz/data/europop* and is reported in Table 8.1. Croux, Filzmoser, Pison, and Rousseeuw (2003) used these data in the context of robust factor analysis.

The variables are average population growth from the year 1986 to 2000 (`pop_growth`), percentage of women of an age able to give birth (`give_birth`), percentage of women of all ages per hundred men (`women%`), life expectancy of women (`lifeexp_f`) and of men (`lifeexp_m`), infant mortality rate (`inf_mort`), number of inhabitants per physician (`inhab/doc`), daily consumption of calories per capita (`calories`), and percentage of babies which are underweight at birth (`baby_underw`). The observations are Austria (A), Albania (AL), Bulgaria (BG), Switzerland (CH), Czechoslovakia (CS), German Democratic Republic (GDR), Hungary (H), Norway (N), Poland (PL), Romania (RO), Sweden (S), Finland (F), Soviet Union (SU), Turkey (TR), Yugoslavia (YU), and the European Community (EC). These sixteen countries correspond to their configuration in the year 1986.

| | pop_growth | give_birth | women% | lifeexp_f | lifeexp_m | inf_mort | inhab/doc | calories | baby_underw |
|---|---|---|---|---|---|---|---|---|---|
| A | -0.1 | 48 | 110 | 77 | 70 | 10 | 440 | 3440 | 6 |
| AL | 1.8 | 50 | 97 | 75 | 68 | 41 | 2100 | 2716 | 7 |
| BG | 0.2 | 47 | 101 | 75 | 69 | 15 | 400 | 3593 | 6 |
| CH | 0.0 | 44 | 103 | 80 | 74 | 7 | 390 | 3406 | 5 |
| CS | 0.3 | 46 | 105 | 75 | 66 | 14 | 350 | 3473 | 6 |
| GDR | 0.0 | 47 | 110 | 75 | 68 | 9 | 490 | 3769 | 6 |
| H | -0.1 | 46 | 106 | 75 | 67 | 19 | 390 | 3544 | 10 |
| N | 0.2 | 48 | 101 | 80 | 74 | 9 | 460 | 3171 | 4 |
| PL | 0.6 | 48 | 104 | 76 | 68 | 18 | 550 | 3224 | 8 |
| RO | 0.5 | 47 | 102 | 73 | 68 | 26 | 700 | 3413 | 6 |
| S | 0.0 | 47 | 101 | 80 | 74 | 6 | 410 | 3007 | 4 |
| F | 0.2 | 47 | 107 | 79 | 72 | 6 | 460 | 2961 | 4 |
| SU | 0.7 | 48 | 112 | 73 | 64 | 30 | 270 | 3332 | 6 |
| TR | 1.9 | 49 | 97 | 67 | 62 | 79 | 1530 | 3218 | 8 |
| YU | 0.5 | 51 | 103 | 74 | 68 | 27 | 700 | 3499 | 7 |
| EC | 0.22 | 48.4 | 103.9 | 78.3 | 72.6 | 10.2 | 509.1 | 3420.5 | 5.4 |

Table 8.1: Eurostat data.

To underline the necessity of a robust analysis, potential outliers in the data are identified first. Assume that the raw data from Table 8.1 are stored in a matrix $\mathbf{X} = (\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(n)})^\top \in \mathbb{R}^{n \times p}$. Robust distances are used to detect whether an observation is an outlier or not. The robust distance of an observation $i$ is defined as

$$\mathrm{RD}_i = \sqrt{(\mathbf{x}_{(i)} - \hat{\boldsymbol{\mu}}_{\mathrm{MVE}})^\top \hat{\boldsymbol{\Sigma}}_{\mathrm{MVE}}^{-1} (\mathbf{x}_{(i)} - \hat{\boldsymbol{\mu}}_{\mathrm{MVE}})} \ , \tag{8.14}$$

where $\hat{\boldsymbol{\mu}}_{\mathrm{MVE}}$ and $\hat{\boldsymbol{\Sigma}}_{\mathrm{MVE}}$ denote the minimum volume ellipsoid (MVE) estimates of the location vector $\boldsymbol{\mu}$ and the scatter matrix $\boldsymbol{\Sigma}$ for the $p$ variables, respectively (Rousseeuw, 1985). Becker and Gather (2001) proposed to use the MVE estimator as an outlier identification tool. The MVE looks for the ellipsoid with smallest volume that covers at least $h$ data points. To ensure that $\hat{\boldsymbol{\mu}}_{\mathrm{MVE}}$ and $\hat{\boldsymbol{\Sigma}}_{\mathrm{MVE}}$ have maximum breakdown values, $h$ can be taken equal to $[(n+p+1)/2]$, where $[a]$ denotes the integer part of $a \in \mathbb{R}$. The location estimator $\hat{\boldsymbol{\mu}}_{\mathrm{MVE}}$ is defined as the center of this ellipsoid. The corresponding covariance estimator $\hat{\boldsymbol{\Sigma}}_{\mathrm{MVE}}$ is given by the ellipsoid itself, multiplied by a suitable factor to obtain Fisher consistency at the multivariate normal distribution. The robust distance (8.14) is a robustification of the Mahalanobis distance defined as

$$\mathrm{MD}_i = \sqrt{(\mathbf{x}_{(i)} - \bar{\mathbf{x}})^\top \mathbf{S}^{-1} (\mathbf{x}_{(i)} - \bar{\mathbf{x}})} \ , \tag{8.15}$$

which uses the sample mean $\bar{\mathbf{x}}$ and sample covariance matrix $\mathbf{S}$ as estimates of location and scatter. For multivariate normally distributed data the values in (8.15) are approximately distributed according to $\chi^2_p$. Using the robust distances in (8.14), an observation is declared as an outlier if the RD for an observation $i$ is larger than a cut-off value, say $\sqrt{\chi^2_{p,0.975}}$. For the Eurostat data, Figure 8.1 displays for each observation its robust distance. A horizontal line is drawn at the cut-off value $\sqrt{\chi^2_{p,0.975}} = \sqrt{19.023} = 4.3615$. Two outlying observations can clearly be identified: Albania and Turkey.

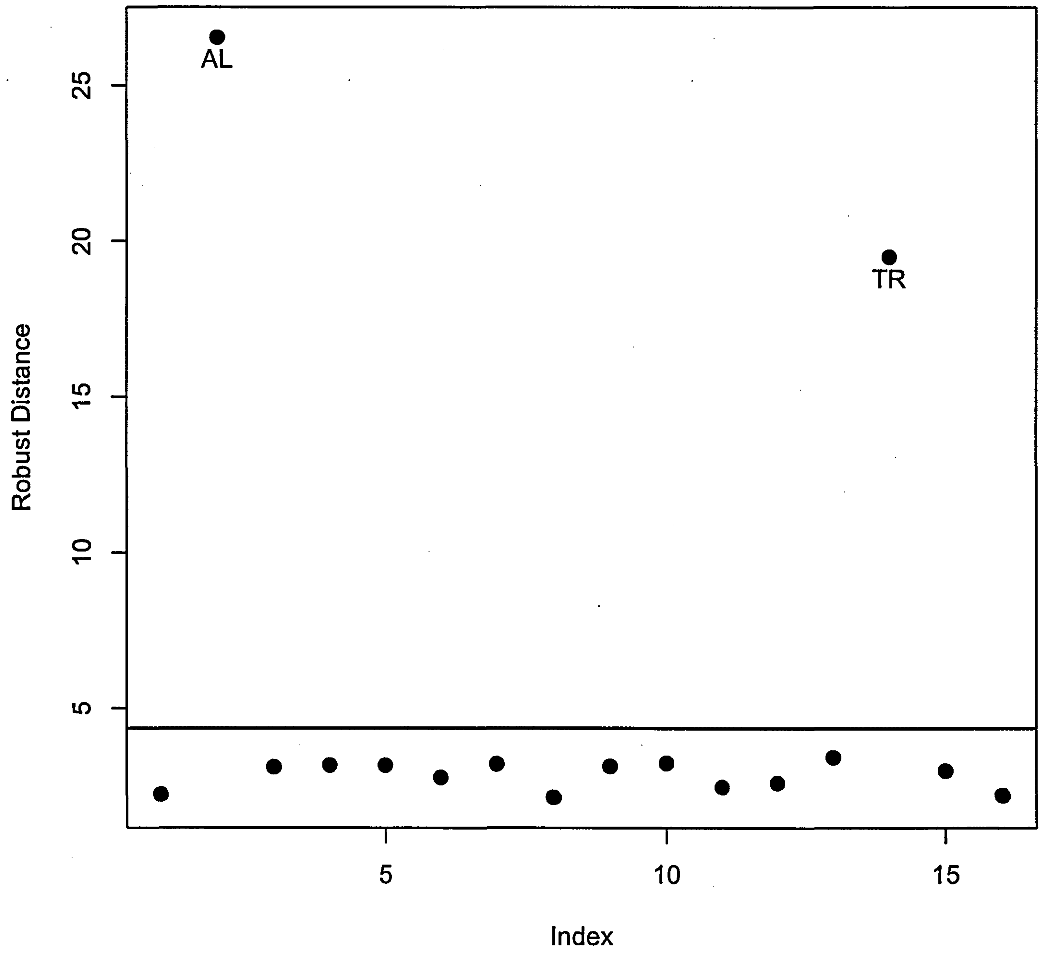

Figure 8.1: Identification of outliers by robust distances.

The data are standardized next so that the $p$ variables have commensurate means and scales. Since the data contain outliers, the data mean as well as its standard deviation are no longer reliable estimates, and therefore, robust data preprocessing is required. For this purpose the coordinatewise median (med) is used for the estimation of the location and the median absolute deviation (MAD) for scale estimation. Both estimators attain a maximum breakdown value of 50% (Rousseeuw and Leroy, 1987). In mathematical terms, the robust preprocessed data are obtained as

$$z_{ij} = \frac{x_{ij} - \mathrm{med}(x_{1j}, \ldots, x_{nj})}{\mathrm{MAD}(x_{1j}, \ldots, x_{nj})} \quad \text{for} \quad j = 1, \ldots, p \ , \tag{8.16}$$

where $\mathrm{MAD}(x_{1j}, \ldots, x_{nj}) = 1.4826 \times \mathrm{med}_i\{|x_{ij} - \mathrm{med}(x_{1j}, \ldots, x_{nj})|\}$. The centered and scaled observations $z_{ij}$ $(i = 1, \ldots, n; j = 1, \ldots, p)$ are stored in the matrix $\mathbf{Z} \in \mathbb{R}^{n \times p}$ and are the data to which the proposed method is applied.

A two-factor-model is considered in the following. The model parameter matrices **F** and **U** are initialized randomly in an orthonormal block matrix **B**. To avoid local optima, the algorithm was run twenty times and it was stopped when successive function values differed by less than $\epsilon = 10^{-6}$. For $k = 2$, the procedure required on average 920 iterations, taking about 0.41 seconds. The twenty runs led to the same function value, up to the third decimal place, which was deemed adequate. Using a higher accuracy criterion such as $\epsilon = 10^{-9}$ needed considerably more CPU time but did not change the quality of the solution.

Thus, it can be concluded that, whereas the algorithm takes a large number of iterations, the analysis of the present data set can be carried out very quickly on a currently standard computer, and the IRLS algorithm is numerically stable. To give some insight into the iteration process, the function value has been plotted for two of the twenty randomly started runs against the iteration number in Figure 8.2. It can be seen that for both runs the decrease of the objective function is rather gradual. Whereas in the first example (left), the decreases are consistently decreasing, in the second example (right) the process seems to have converged after about 500 iterations, but then jumps down after which the process gradually decreases until convergence. In both cases, the monotonically decreasing function value stabilized at the same height.

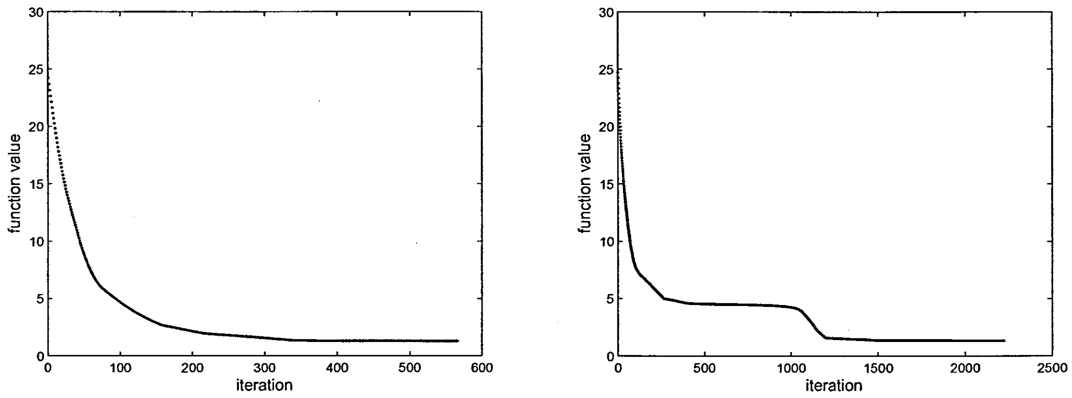

Figure 8.2: Function value plotted against iteration number for two randomly started runs of the IRLS algorithm.

After convergence, not only estimates for the block parameter matrices **A** and **B** but also a set of weights $w_{ij} \in ]0,1]$ $(i = 1,\ldots,n; j = 1,\ldots,p)$ are available that can be used to examine the outlyingness of the residuals. Weights equal to 1 are assigned to data that fit the model well, while outliers will have small weights.

Recall that **W** is an $n \times p$ matrix of possibly different weights attached to each residual, that is, elementwise weighting is performed. The loss function is applied to each residual element, $e_{ij}$, and then these loss values are summed up to obtain the overall loss. Instead of elementwise weighting, rowwise weighting aggregates the residuals over the rows and the loss function is then applied to these $n$ aggregated values. Since small weights could be assigned to separate scores of an observation, leaving its other scores unaffected, elementwise weighting is more flexible than rowwise weighting which considers whole objects as possible outliers (Verboon, 1994).

Figure 8.3 is a checkerboard plot, a flat surface plot with its view set to directly above, of the final $16 \times 9$ matrix of weights **W**. The values of the elements of **W** specify the colour in each cell of the plot, ranging from yellow to red. Note that the scale of the

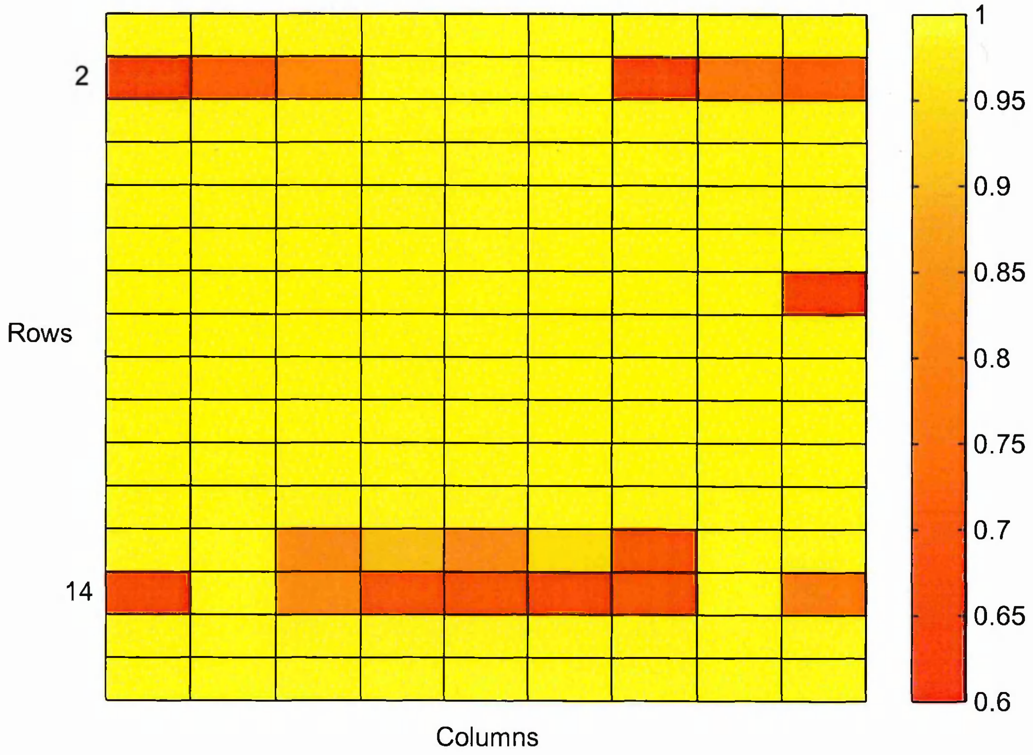

Figure 8.3: Checkerboard plot of the 16 × 9 matrix of weights **W** obtained after the IRLS algorithm has converged for the Eurostat data.

colourbar in Figure 8.3 is upside-down, that is, red rectangular faces of the surface correspond to small weights and hence large residuals. As Figure 8.3 shows, the proposed IRLS algorithm detects that the observations Albania (row 2) and Turkey (row 14) are outlying in most of the variables. As desired, the IRLS algorithm fits the data in such a way that residuals corresponding to outliers are relatively large after applying the procedure.

The Huber function depends on the tuning constant $\gamma$ which must be determined before running the IRLS algorithm. For the Eurostat data, the Huber constant was chosen as $\gamma = 0.05$. This value appeared to yield the most satisfactory results, which means that the outliers were well distinguished from the other points.

The purpose of a robust procedure is to fit the majority of the data or the proper (non-outlying) data points well. In other words, the outliers should not be capable of obscuring the main structure in the data. Fitting the EFA model by the proposed approach yields $\hat{\mathbf{Z}} = \hat{\mathbf{F}}\hat{\mathbf{\Lambda}}^{\top} + \hat{\mathbf{U}}\hat{\mathbf{\Psi}}$. From $\hat{\mathbf{Z}}$ one finds the matrix $\hat{\mathbf{X}}$ containing the fitted values $\hat{x}_{ij}$ $(i = 1 \ldots, n; j = 1, \ldots, p)$ expressed in the location and scale of the original values of the variables as

$$\hat{\mathbf{X}} = \mathbf{1}_n\hat{\boldsymbol{\mu}}^{\top} + \mathbf{1}_n\hat{\boldsymbol{\sigma}}^{\top} \odot \hat{\mathbf{Z}} \ ,$$

where $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\sigma}}$ are the vectors of (robust) estimates of center and scale for the $p$ variables, respectively.

As in Croux, Filzmoser, Pison, and Rousseeuw (2003), the fit of the majority of the data is evaluated by computing $\sum\limits_{\substack{i=1 \\ i\neq 2,14}}^{n} \sum\limits_{j=1}^{p}(x_{ij} - \hat{x}_{ij})^2$, that is, the sum of the squared differences between the observed and fitted values with the index $i$ running over all rows except 2 and 14 (the rows representing the outliers Albania and Turkey). For three different approaches, the corresponding values are given in Table 8.2. With the ULS approach of De Leeuw (2004, 2008), the effect of the outliers on the model fit is clearly disastrous. It yields large residuals for non-outlying data points. The robust IRLS procedure outperforms the non-robust ULS approach and the majority of the data is fitted well. The results from Table 8.2 imply that for the presentation of the parameter estimates the proposed IRLS approach is preferable.

Finally, as Table 8.2 reveals, the IRLS algorithm fits the proper data points better than the robust approach to factor analysis proposed in Croux, Filzmoser, Pison, and Rousseeuw (2003). Recall that in Croux, Filzmoser, Pison, and Rousseeuw (2003) the data matrix is factorized into a pair $\{\hat{\mathbf{\Lambda}}, \hat{\mathbf{F}}\}$ by optimizing a weighted $\ell_1$ alternating

| Method | $\sum_{\substack{i=1 \\ i\neq 2,14}}^{n} \sum_{j=1}^{p} (x_{ij} - \hat{x}_{ij})^2$ |
|---|---|
| Least squares decomposition [De Leeuw (2004, 2008)] | 48653 |
| Weighted $\ell_1$ alternating regression [Croux, Filzmoser, Pison, and Rousseeuw (2003)] | 575 |
| IRLS approach | 420 |

Table 8.2: Quality of fit for three different EFA factorizations applied to Eurostat data.

regression procedure. Estimates for the matrix of unique variances $\mathbf{\Psi}^2$ are obtained from the residuals $\mathbf{Z} - \hat{\mathbf{F}}\hat{\mathbf{\Lambda}}^\top$ after the algorithm has already converged. Additional weights which would account for the unequal variances among the unique factors are not included in the regression scheme because it "can affect the stability of the algorithm" (Filzmoser, 2002). However, $\mathbf{U}\mathbf{\Psi}$ is part of the EFA model (3.19). If the EFA model (3.19) holds for a particular set of data, then a procedure which estimates all EFA model unknowns $\{\hat{\mathbf{F}}, \hat{\mathbf{\Lambda}}, \hat{\mathbf{U}}, \hat{\mathbf{\Psi}}\}$ simultaneously can be expected to obtain a better fit than an approach which ignores the unique part of the EFA model in the loss function.

# Chapter 9

# Discussion

Classical EFA fitting techniques factorize the sample covariance or correlation matrix into a factor loadings matrix and a matrix of unique factor variances with respect to some goodness-of-fit criterion. In Part II of this thesis, the EFA model was considered as a specific data matrix decomposition with fixed unknown matrix parameters. Several new algorithms were introduced for the LS and WLS estimation of all EFA model unknowns.

For vertical data matrices with $n > p$, a new iterative algorithm for the simultaneous estimation of all EFA model parameters $\{\mathbf{F}, \mathbf{\Lambda}, \mathbf{U}, \mathbf{\Psi}\}$ was introduced in Chapter 5 which updates the common and unique factor scores successively.

Furthermore, a reparameterization of the EFA model was proposed to produce solutions with a lower triangular matrix $\mathbf{L}$ of loadings. Whereas the parameter matrix $\mathbf{\Lambda}$ in the classical EFA formulation (3.19) admits an infinite number of orthogonal rotations, the lower triangular reparameterization removes the rotational indeterminacy of the EFA model and leads to solutions already having an interpretable pattern. Moreover, the new parameters are subject to the constraint $\mathrm{rank}(\mathbf{L}) = k$ expressing the nature of the EFA model, rather than facilitating the numerical method for their estimation (as is the case with the condition $\mathbf{\Lambda}^{\top}\mathbf{\Lambda}$ or $\mathbf{\Lambda}^{\top}\mathbf{\Psi}^{-2}\mathbf{\Lambda}$ being diagonal for the standard EFA estimation procedures).

In Chapter 6, EFA was viewed as a special case of PCA with the error term resembling the EFA one. This EFA-like PCA construction helped to provide conditions under which for a particular data set the PCs and their coefficients can be used as adequate surrogates for the common factors and their loadings. It was demonstrated by examples that the PCA and EFA solutions can look very similar, despite the fact that the EFA model provides a better fit to the data than PCA. As an alternative to the SVD, the QR decomposition was proposed for rank reducing approximation of the data. Whereas the iterative algorithms for simultaneous EFA look for pairs $\{\mathbf{\Lambda}, \mathbf{\Psi}\}$ and $\{\mathbf{F}, \mathbf{U}\}$, EFA-like PCA looks for pairs $\{\mathbf{\Lambda}, \mathbf{F}\}$ and $\{\mathbf{\Psi}, \mathbf{U}\}$.

For horizontal data matrices with $p \geq n$, an extension of the EFA model was proposed in Chapter 7 and novel algorithms for the estimation of all EFA model unknowns were presented. The new model requires $\mathbf{\Psi}^2$ being positive semi-definite. For a number of modern applications the data are often high-dimensional with $n \ll p$. Iterative algorithms factorizing a $p \times p$ correlation matrix $\mathbf{Z}^{\top}\mathbf{Z}$ may become computationally slow if $p$ is huge. In the case that $n \ll p$, taking an $n \times p$ data matrix $\mathbf{Z}$ as an input for EFA seems a reasonable alternative.

Classical EFA techniques take input data in the form of covariances/correlations and are very vulnerable to the presence of outliers. To overcome the outlier problem, the EFA model was considered as a weighted data matrix decomposition in Chapter 8. The weights in the WLS loss function are used to diminish the outliers' influence in the data. Each entry in the data matrix is separately weighted and the weights are related to the residuals obtained after each cycle of updating the model parameters. This yields an iteratively reweighted least squares (IRLS) scheme. An iterative majorization approach to optimization is employed to provide a monotonically convergent IRLS algorithm for the robust estimation of all EFA model parameters.

In the proposed IRLS scheme, the weights are chosen according to the Huber function. However, by merely choosing the weights differently, the iterative majorization approach presented can be used for a variety of resistant loss function. Examples are the biweight function (Mosteller and Tukey, 1977) or simply a trimming function, which assigns 0 to residuals larger than a particular value and 1 otherwise. This makes the IRLS algorithm widely applicable.

Due to the factor score indeterminacy in EFA, the common and unique factor scores are not uniquely estimable. However, the non-uniqueness of the factor scores is not a problem for the numerical algorithms that find estimates for all matrix parameters. From this point of view, the numerical procedures developed in this thesis avoid the conceptual problem of factor score indeterminacy and facilitate the estimation of both **F** and **U**. Far more important than this, the new algorithms facilitate the application of EFA for analyzing multivariate data because they are based on the computationally efficient and well-studied numerical procedure of the SVD of data matrices, as PCA is. The main drawback of the proposed decomposition models with fixed matrix parameters is that it is not possible to test them by statistical methods. Nevertheless, as De Leeuw (2008) points out, the notions of monotonic convergence of the algorithms, stability of solutions and goodness-of-fit continue to apply.

# Part III

# Rotation Towards Independence in Factor Analysis

# Chapter 10

# Fitting the noisy ICA model

Recall the noisy ICA model (3.13):

$$\mathbf{x} = \mathbf{M}\boldsymbol{\xi} + \mathbf{u} \ . \tag{10.1}$$

The problem in noisy ICA is to estimate the mixing matrix $\mathbf{M}$ and to obtain the realizations of the independent components $\boldsymbol{\xi}$ from $n$ available observations on $\mathbf{x}$ only. Due to the existence of $\mathbf{u}$ in (10.1), knowledge of $\mathbf{M}$ does not give direct access to the independent components. This implies that apart from a procedure for estimating the mixing matrix one requires a method for obtaining the realizations of the independent components. This Chapter provides an account that is intended as a review of ICA in the presence of normally distributed noise. Objective functions, originally devised in noise-free ICA, which measure the departure of the recovered components from independence are presented in Section 10.1. How these criteria are embedded in algorithms for fitting the noisy ICA model is discussed in Section 10.2.

## 10.1 Criteria for measuring departure from independence

The assumption underlying all ICA models is that the latent sources are independent, that is, the $k$-dimensional joint probability density function (pdf) of $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_k)^\top$,

$f_{\boldsymbol{\xi}}(\boldsymbol{\xi})$, factorizes into the product of their marginal densities as

$$f_{\boldsymbol{\xi}}(\boldsymbol{\xi}) = \prod_{i=1}^{k} f_{\xi_i}(\xi_i) \ . \tag{10.2}$$

Let $\hat{\boldsymbol{\xi}} = (\hat{\xi}_1, \ldots, \hat{\xi}_k)^{\mathrm{T}}$ denote a vector of 'estimates'. Note that $\xi_1, \ldots, \xi_k$ cannot be estimated in the usual statistical sense since they are not parameters which can be inferred from sample statistics but values ascribed to unobservable variates.

If the joint pdf of $\hat{\boldsymbol{\xi}}$ also factorizes, then $\hat{\xi}_1, \ldots, \hat{\xi}_k$ are independent and the ICA problem of separating observed linear mixtures of signals into independent sources is solved.

An objective (also called contrast) function is a scalar measure, $\varphi : \mathbb{R}^k \to \mathbb{R}$, which serves as a criterion for measuring the deviation of $\hat{\xi}_1, \ldots, \hat{\xi}_k$ from independence (Cardoso, 1998).

### 10.1.1 Contrast functions defined in terms of differential entropies

Contrast functions (or contrasts for short) are operating on a pdf and are designed such that source separation is obtained when they reach their optimal value, that is,

$$\varphi(f_{\hat{\boldsymbol{\xi}}}(\hat{\boldsymbol{\xi}})) \geq \varphi(f_{\boldsymbol{\xi}}(\boldsymbol{\xi})) \ , \tag{10.3}$$

where $f_{\hat{\boldsymbol{\xi}}}(\hat{\boldsymbol{\xi}})$ denotes the joint pdf of $\hat{\boldsymbol{\xi}}$. Equality (10.3) holds if and only if $\hat{\boldsymbol{\xi}}$ is a copy of $\boldsymbol{\xi}$, that is, the entries of $\hat{\boldsymbol{\xi}}$ are identical to those of $\boldsymbol{\xi}$ up to permutation and scaling ambiguities. To facilitate the notation, (10.3) can be rewritten as

$$\varphi[\hat{\boldsymbol{\xi}}] \geq \varphi[\boldsymbol{\xi}] \ , \tag{10.4}$$

where square brackets are used to emphasize that the contrasts depend on the pdfs of $\hat{\boldsymbol{\xi}}$ and $\boldsymbol{\xi}$ rather than directly on $\hat{\boldsymbol{\xi}}$ and $\boldsymbol{\xi}$. Following Cardoso (1998), the canonical form of an ICA contrast may be regarded as being derived from the information-theoretic concept of mutual information.

Mutual information can be derived in terms of differential entropies as follows (see Cover and Thomas, 1991, Chapter 9). Differential entropy is a concept in information theory which extends the idea of Shannon entropy, a measure of the uncertainty associated with a discrete random variable, to continuous probability distributions. Differential entropy of a random vector x with pdf $f_{\mathrm{x}}(\mathrm{x})$ is defined as

$$H[\mathrm{x}] = -\int_{\mathfrak{X}} f_{\mathrm{x}}(\mathrm{x}) \log f_{\mathrm{x}}(\mathrm{x})\, d\mathrm{x} \ . \tag{10.5}$$

where $\mathfrak{X}$ denotes the support set (Casella and Berger, 2002, p. 50) of the distribution of x. In this and all subsequent examples involving an integral it is assumed that the integral exists and that integration is carried out over the support set of the corresponding distribution. Thus, for example, the differential entropy of a $p$-variate random vector $\mathrm{x} \sim \mathcal{N}_p(\mathbf{0}, \Sigma)$ is (Cover and Thomas, 1991, pp. 230-231):

$$H(\mathcal{N}_p(\mathbf{0}, \Sigma)) = \frac{1}{2} \ln \left[(2\pi \mathrm{e})^p \det(\Sigma)\right] \ , \tag{10.6}$$

where ln denotes the natural logarithm. Since it only changes the measurement scale, the choice of the log base is not important. Joint entropy, $H[\mathrm{x}, \mathrm{y}]$, of two random vectors x and y with joint density, $f_{\mathrm{x,y}}(\mathrm{x}, \mathrm{y})$, is

$$H[\mathrm{x}, \mathrm{y}] = -\int f_{\mathrm{x,y}}(\mathrm{x}, \mathrm{y}) \log f_{\mathrm{x,y}}(\mathrm{x}, \mathrm{y})\, d\mathrm{x}\, d\mathrm{y} \tag{10.7}$$

and the conditional entropy of x given y is given by

$$H[\mathrm{x}|\mathrm{y}] = -\int f_{\mathrm{x,y}}(\mathrm{x}, \mathrm{y}) \log f_{\mathrm{x}|\mathrm{y}}(\mathrm{x}|\mathrm{y})\, d\mathrm{x}\, d\mathrm{y} \ , \tag{10.8}$$

where $f_{\mathrm{x}|\mathrm{y}}(\mathrm{x}|\mathrm{y})$ denotes the conditional density function of x given y. Using (10.5), (10.7) and (10.8), it follows that

$$H[\mathrm{x}, \mathrm{y}] = H[\mathrm{x}] + H[\mathrm{y}|\mathrm{x}] = H[\mathrm{y}] + H[\mathrm{x}|\mathrm{y}] \ . \tag{10.9}$$

Mutual information is then defined as

$$\begin{aligned} I[\mathrm{x};\mathrm{y}] &= \int f_{\mathrm{x,y}}(\mathrm{x},\mathrm{y}) \log \frac{f_{\mathrm{x,y}}(\mathrm{x},\mathrm{y})}{f_{\mathrm{x}}(\mathrm{x}) f_{\mathrm{y}}(\mathrm{y})} \, d\mathrm{x} \, d\mathrm{y} \ , \\ &= H[\mathrm{x}] + H[\mathrm{y}] - H[\mathrm{x},\mathrm{y}] \ . \end{aligned} \tag{10.10}$$

Thus, $I[\mathrm{x};\mathrm{y}]$ is the difference in the information that is obtained by observing $\mathbf{x}$ and $\mathbf{y}$ separately and jointly. Using (10.9), (10.10) becomes

$$I[\mathrm{x};\mathrm{y}] = H[\mathrm{x}] - H[\mathrm{x}|\mathrm{y}] = H[\mathrm{y}] - H[\mathrm{y}|\mathrm{x}] \ . \tag{10.11}$$

Since $H[\mathrm{x}|\mathrm{y}] \leq H[\mathrm{x}]$ and $H[\mathrm{y}|\mathrm{x}] \leq H[\mathrm{y}]$, $I[\mathrm{x};\mathrm{y}] \geq 0$ where equality holds if and only if $\mathbf{x}$ and $\mathbf{y}$ are mutually independent.

Using entropies, the mutual information between $\hat{\xi}_1, \ldots, \hat{\xi}_k$ can be written as

$$\begin{aligned} I[\hat{\boldsymbol{\xi}}] &= \sum_{i=1}^{k} H[\hat{\xi}_i] - H[\hat{\boldsymbol{\xi}}] \ , \\ &= \int f_{\hat{\boldsymbol{\xi}}}(\hat{\boldsymbol{\xi}}) \log \frac{f_{\hat{\boldsymbol{\xi}}}(\hat{\boldsymbol{\xi}})}{\prod_{i=1}^{k} f_{\hat{\xi}_i}(\hat{\xi}_i)} \, d\hat{\boldsymbol{\xi}} \ . \end{aligned} \tag{10.12}$$

which is the information common to $\hat{\xi}_1, \ldots, \hat{\xi}_k$. Thus, $I[\hat{\boldsymbol{\xi}}] = 0$ if and only if $\hat{\xi}_1, \ldots, \hat{\xi}_k$ are mutually independent, so that the joint density factorizes. Hence, source separation can be evaluated by the following contrast:

$$\varphi_{\mathrm{MI}}[\hat{\boldsymbol{\xi}}] = I[\hat{\boldsymbol{\xi}}] \ , \tag{10.13}$$

where $\varphi_{\mathrm{MI}}[\hat{\boldsymbol{\xi}}]$ is minimized when the source separation into independent sources is successful, that is, if $\varphi_{\mathrm{MI}}[\hat{\boldsymbol{\xi}}] = \varphi_{\mathrm{MI}}[\boldsymbol{\xi}] = 0$.

Alternatively, mutual information may be interpreted as a distance between two pdfs $f_{\mathrm{x}}(\mathrm{x})$ and $g_{\mathrm{x}}(\mathrm{x})$ using the Kullback-Leibler divergence which is defined as (Cover and Thomas, 1991, p. 231):

$$D_{\mathrm{KL}}\left(f_{\mathrm{x}}(\mathrm{x}) || g_{\mathrm{x}}(\mathrm{x})\right) = \int f_{\mathrm{x}}(\mathrm{x}) \log \frac{f_{\mathrm{x}}(\mathrm{x})}{g_{\mathrm{x}}(\mathrm{x})} \, d\mathrm{x} \ . \tag{10.14}$$

Note that $D_{\text{KL}}\left(f_{\mathbf{x}}(\mathbf{x})||g_{\mathbf{x}}(\mathbf{x})\right)$ is finite only if the support set of $f_{\mathbf{x}}(\mathbf{x})$ is contained in the support set of $g_{\mathbf{x}}(\mathbf{x})$.

The Kullback-Leibler divergence is not symmetric and therefore it does not satisfy the distance axioms (e.g., Mardia, Kent, and Bibby, 1979, pp. 375-376). However, since $D_{\text{KL}}\left(f_{\mathbf{x}}(\mathbf{x})||g_{\mathbf{x}}(\mathbf{x})\right) \geq 0$ with equality if and only if $f_{\mathbf{x}}(\mathbf{x})$ is a copy of $g_{\mathbf{x}}(\mathbf{x})$, it can be used as a measure quantifying the closeness of two distributions. Comparison between (10.12) and (10.14) reveals that the mutual information between $\hat{\xi}_1, \ldots, \hat{\xi}_k$ is identical to the Kullback-Leibler divergence between $f_{\hat{\boldsymbol{\xi}}}(\hat{\boldsymbol{\xi}})$ and its version for independent $\hat{\xi}_i$ $(i = 1, \ldots, k)$, that is, $\varphi_{\text{MI}}[\hat{\boldsymbol{\xi}}] = D_{\text{KL}}\left(f_{\hat{\boldsymbol{\xi}}}(\hat{\boldsymbol{\xi}})||\prod_{i=1}^{k} f_{\hat{\xi}_i}(\hat{\xi}_i)\right)$.

Among all distributions with fixed covariance structure, the normal distribution maximizes the differential entropy. It serves therefore as an upper bound which is stated in the following Theorem 10.1.

**Theorem 10.1.** Let $\mathbf{x} \in \mathbb{R}^{p \times 1}$ be a random vector with zero mean and covariance matrix $\boldsymbol{\Sigma}$. Then, $H[\mathbf{x}] \leq \frac{1}{2}\log[(2\pi \mathrm{e})^p \det(\boldsymbol{\Sigma})]$ with equality if and only if $\mathbf{x} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$.

*Proof.* (see Cover and Thomas, 1991, p. 234). □

Source separation may be achieved by exploiting this property of the normal distribution. Assume that $\hat{\xi}_1, \ldots, \hat{\xi}_k$ are uncorrelated. In light of (10.12), minimizing the mutual information between $\hat{\xi}_1, \ldots, \hat{\xi}_k$ is equivalent to minimizing the marginal entropies $\sum_{i=1}^{k} H[\hat{\xi}_i]$, which in turn, according to Theorem 10.1, amounts to maximizing their departure from normality.

The deviation from normality of $\hat{\boldsymbol{\xi}}$ may be quantified conveniently in terms of the negentropy measure (e.g., Lee, 1998). Let $f_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}})$ be a Gaussian pdf with entropy $H[\tilde{\mathbf{x}}]$

having the same mean and covariance matrix as $f_{\hat{\boldsymbol{\xi}}}(\hat{\boldsymbol{\xi}})$. Negentropy is defined as

$$J(\hat{\boldsymbol{\xi}}) = D_{\text{KL}}\left(f_{\hat{\boldsymbol{\xi}}}(\hat{\boldsymbol{\xi}}) || f_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}})\right) = H[\tilde{\mathbf{x}}] - H[\hat{\boldsymbol{\xi}}] \geq 0 \ . \tag{10.15}$$

If $\hat{\xi}_1, \ldots, \hat{\xi}_k$ are uncorrelated, negentropy is related to mutual information by

$$\varphi_{\text{MI}}[\hat{\boldsymbol{\xi}}] = J(\hat{\boldsymbol{\xi}}) - \sum_{i=1}^{k} J(\hat{\xi}_i) \ . \tag{10.16}$$

Like mutual information, the negentropy is non-negative and it is zero if and only if $\hat{\boldsymbol{\xi}}$ is normally distributed.

The use of differential entropy or negentropy as a contrast requires estimates of the densities involved. This is computationally rather complicated and/or error prone. Therefore, approximations to the contrast functions are often used in practice instead (e.g., Izenman, 2008). These approximations are either based on higher-order cumulants using polynomial density expansions or based on non-polynomial functions.

### 10.1.2 Approximations to contrast functions

Given a $p$-dimensional random vector $\mathbf{x} = (x_1, \ldots, x_p)^\top$ with pdf $f_{\mathbf{x}}(\mathbf{x})$ and a vector $\mathbf{t} = (t_1, \ldots, t_p)^\top \in \mathbb{R}^{p \times 1}$, the joint moment generating function of $\mathbf{x}$ is (e.g., Mood, Graybill, and Boes, 1974, pp. 78-80):

$$M_{\mathbf{x}}(\mathbf{t}) = \int_{-\infty}^{\infty} \mathrm{e}^{\mathbf{t}^\top \mathbf{x}} f_{\mathbf{x}}(\mathbf{x})\, d\mathbf{x} = \mathrm{E}\left(\mathrm{e}^{\mathbf{t}^\top \mathbf{x}}\right) \ , \tag{10.17}$$

if the expectation exists for all values of $t_1, \ldots, t_p$ such that $-h < t_i < h$ for some $h > 0$ $(i = 1, \ldots, p)$. For any set of non-negative integers $r_1, \ldots, r_p$, let $m_{r_1,\ldots,r_p}$ denote the $(r_1, \ldots, r_p)$-th joint moment of $\mathbf{x}$ which is the coefficient of $(t_1^{r_1} \cdots t_p^{r_p})/(r_1! \cdots r_p!)$ in the Taylor series expansion of $M_{\mathbf{x}}(\mathbf{t})$ around $\mathbf{t} = \mathbf{0}_p$. This implies that

$$m_{r_1,\ldots,r_p} = \mathrm{E}(x_1^{r_1} \cdots x_p^{r_p}) = \frac{\partial^{r_1+\cdots+r_p}}{\partial t_1^{r_1} \cdots \partial t_p^{r_p}} M_{\mathbf{x}}(\mathbf{t})\bigg|_{\mathbf{t}\,=\,\mathbf{0}_p} \ . \tag{10.18}$$

Using (10.18) to generate the first-order moments $m_{r_1}, \ldots, m_{r_p}$, the $(r_1, \ldots, r_p)$-th joint central moment of $\mathbf{x}$ is $\mu_{r_1,\ldots,r_p} = \mathrm{E}[(x_1 - m_{r_1})^{r_1} \cdots (x_p - m_{r_p})^{r_p}]$.

The joint cumulant generating function is defined as

$$K_{\mathbf{X}}(\mathbf{t}) = \ln\left(M_{\mathbf{X}}(\mathbf{t})\right) \ . \tag{10.19}$$

Let $\kappa_{r_1,\ldots,r_p}$ denote the $(r_1, \ldots, r_p)$-th joint cumulant of $\mathbf{x}$ which is the coefficient of $(t_1^{r_1} \cdots t_p^{r_p})/(r_1! \ldots r_p!)$ in the Taylor series expansion of $K_{\mathbf{X}}(\mathbf{t})$ around $\mathbf{t} = \mathbf{0}_p$. Using (10.18) and (10.19), $\kappa_{r_1,\ldots,r_p}$ can be expressed as

$$\kappa_{r_1,\ldots,r_p} = \frac{\partial^{r_1+\cdots+r_p}}{\partial t_1^{r_1} \cdots \partial t_p^{r_p}} K_{\mathbf{X}}(\mathbf{t})\bigg|_{\mathbf{t}=\mathbf{0}_p} \ . \tag{10.20}$$

Thus, for example, the first four univariate cumulants for a single random variable $x$ with mean $\mu$ and variance $\sigma^2$ are

$$\begin{aligned} \kappa_1 &= m_1 = \mu \ , \\ \kappa_2 &= \mu_2 = m_2 - m_1^2 = \sigma^2 \ , \\ \kappa_3 &= \mu_3 = m_3 - 3\mu m_2 + 2\mu^3 \ , \\ \kappa_4 &= \mu_4 - 3\mu_2^2 = \mu_4 - 3\kappa_2^2 \ , \end{aligned}$$

where $\kappa_3$ and $\kappa_4$ can be used to define the skewness and kurtosis of a distribution, which measure the asymmetry and peakedness of a pdf, respectively. Skewness, $\gamma_1$, and (excess) kurtosis, $\gamma_2$, are defined as

$$\gamma_1 = \frac{\kappa_3}{\kappa_2^{3/2}} = \frac{\mu_3}{\sigma^3} \ , \tag{10.21}$$

$$\gamma_2 = \frac{\kappa_4}{\kappa_2^2} = \frac{\mu_4}{\sigma^4} - 3 \ . \tag{10.22}$$

For symmetric distributions such as the normal, $\gamma_1 = 0$. Distributions with positive skewness (negative skewness) are called right-skewed (left-skewed).

Distributions for which $\gamma_2 = 0$ are called mesokurtic and those for which $\gamma_2 > 0$ ($\gamma_2 < 0$)

are named leptokurtic (platykurtic). Leptokurtic distributions have a sharper peak and heavier tails than the normal curve. Since for the normal $\gamma_2 = 0$ and for most non-normal distributions $\gamma_2 \neq 0$, kurtosis can also be used as a measure of non-gaussianity and hence as an objective function in ICA.

Consider now the multivariate case. According to Hinich (1994), if $r_1, \ldots, r_p$ are all equal to one, the joint cumulants are called simple and the 2nd and 4th-order (simple) joint cumulants are

$$\begin{aligned}
\kappa_{1,1}(x_i, x_j) &= \mu_{1,1} = \mathrm{E}(x_i x_j) \quad (i \neq j) \ , \\
\kappa_{1,1,1,1}(x_i x_j x_k x_l) &= \mathrm{E}(x_i x_j x_k x_l) - \mathrm{E}(x_i x_j)\mathrm{E}(x_k x_l) \\
&\quad -\mathrm{E}(x_i x_k)\mathrm{E}(x_j x_l) - \mathrm{E}(x_i x_l)\mathrm{E}(x_j x_k) \quad (i \neq j \neq k \neq l) \ .
\end{aligned}$$

If all random variables are identical, $\kappa_{1,1} = \kappa_2 = \sigma^2$ and $\kappa_{1,1,1,1} = \kappa_4$, that is, the $n$th-order (simple) joint cumulant turns into the $n$th-order univariate cumulant.

For the following reasons it is preferable to express higher-order statistics through cumulants rather than through moments (McCullagh, 1987, p. 25):

(i) Most statistical calculations using cumulants are simpler than the corresponding calculations using moments.

(ii) For independent random variables, the cumulants of a sum are the sum of the cumulants.

(iii) For independent random variables, the joint cumulants are zero.

(iv) If $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$, then

$$K_{\mathbf{x}}(\mathbf{t}) = \mathbf{t}^{\mathsf{T}}\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}^{\mathsf{T}}\Sigma\mathbf{t} \ .$$

Hence, unlike higher-order moments, the cumulants with order greater than two vanish for the normal distribution.

(v) Polynomial density expansions are conveniently expressed in terms of cumulants.

As stated in (v), approximations of the contrasts through cumulants can be obtained using polynomial representations of the pdfs in an orthonormal series expansion (Stuart and Ord, 1994).

Using multivariate Edgeworth series expansion up to fourth-order and the assumption that $x_1, \ldots, x_p$ are uncorrelated, $\varphi_{\text{MI}}[\mathrm{x}]$ can be approximated by (Hyvärinen, 1999c):

$$\varphi_{\text{MI}}[\mathrm{x}] \approx c + \frac{1}{48} \sum_{i=1}^{p} \left[ 4\kappa_3^2(x_i) + \kappa_4^2(x_i) + 7\kappa_4^4(x_i) - 6\kappa_3^2(x_i)\kappa_4(x_i) \right] \quad , \tag{10.23}$$

where $c$ is a constant. The approximation (10.23), in a slightly altered guise, forms the basis of many objective functions in ICA. For example, Comon (1994) proposed simply to maximize the criterion $\sum_{i=1}^{p} \kappa_4^2(x_i)$. Comon (1994) also suggested a criterion testing the independence between the components by summing all squared joint (simple) cumulants. Cardoso and Souloumiac (1993) proposed a similar criterion:

$$\sum_{i,j,k,l \neq i,i,k,l} \kappa_{1,1,1,1}^2(x_i, x_j, x_k, x_l) \quad , \tag{10.24}$$

also being a sum of the squared joint (simple) cumulants of the components, where the notation indicates that the sum is taken over all the quadruples of indices with $i \neq j$. In the context of projection pursuit, Jones and Sibson (1987) approximate the departure from normality measured by negentropy using higher-order cumulants as follows:

$$J(x) \approx \frac{1}{12}\kappa_3^2(x) + \frac{1}{48}\kappa_4^2(x) \quad , \tag{10.25}$$

where the random variable $x$ is assumed to be of zero mean and unit variance. For symmetric distributions the first term in (10.25) vanishes, and this approximation leads to the use of kurtosis as a measure of non-normality. In fact, the same projection pursuit indexes (higher-order cumulants, polynomial-based indexes) are often used as objective functions in ICA (Izenman, 2008).

Cumulant-based approximations simplify considerably the use of mutual information and negentropy. However, the main difficulty of using cumulant-based indexes arises from their lack of robustness against outliers (e.g. Jones and Sibson, 1987). This led to the development and use of approximations based on non-polynomial functions. Hyvärinen (1999a) uses an approximation to negentropy of the form:

$$J(G(x)) \approx \beta \left[\mathrm{E}\{G(x)\} - \mathrm{E}\{G(z)\}\right]^2 \ , \tag{10.26}$$

where $\beta$ is a positive constant and $z \sim \mathcal{N}(0,1)$. Note that the objective function $G$ must not be quadratic because otherwise (10.26) is zero for all distributions. This approximation is a generalization of the cumulant-based approximation (10.25), if $x$ has a symmetric distribution in which the first term in (10.25) vanishes. Indeed, taking $G(x) = x^4$, one obtains a kurtosis-based approximation. By choosing $G(x)$ carefully one can obtain approximations of negentropy that are more robust performers. Hyvärinen (1999a) proposed the following choices of $G(x)$:

$$G_1(x) = \frac{1}{\alpha} \log\cosh(\alpha x) \ , \tag{10.27}$$

$$G_2(x) = -e^{-x^2/2} \ , \tag{10.28}$$

where in (10.27) $1 \leq \alpha \leq 2$ is some suitable constant (usually, $\alpha = 1$).

After choosing a measure of independence, one needs a practical algorithm in which this measure is embedded and that can be used for fitting the noisy ICA model.

## 10.2 Estimation procedures in noisy ICA

Since cumulants of order greater than two are unaffected by Gaussian noise, it is possible to identify the mixing matrix $\mathbf{M}$ by optimizing an independence measure composed of higher-order cumulants only (e.g., Lathauwer, Moor, and Vanderwalle, 1996). An advantage of this approach is that one does not need to know or estimate the noise covariance matrix $\mathbf{\Psi}^2$. However, in Lathauwer, Moor, and Vanderwalle (1996) no indication is given as to how the realizations of the sources $\boldsymbol{\xi}$ should be estimated after an estimate for $\mathbf{M}$ is found.

Another attempt to estimate the noisy ICA model is to modify ordinary noise-free ICA methods such that the effect of the noise is removed or at least reduced. Suppose the noise covariance matrix $\mathbf{\Psi}^2$ is known. Then this information can be taken into account to correct the second-order statistics of the observed data. This can be done either in conjunction with higher-order cumulants or some non-linear measures of independence which are immune to Gaussian noise (Hyvärinen, 1999b). However, the method of Hyvärinen (1999b) requires prior knowledge of $\mathbf{\Psi}^2$ and, again, there is no mention about how the additional problem of estimating $\boldsymbol{\xi}$ is solved.

Estimates of $\mathbf{M}$ and $\boldsymbol{\xi}$ can be obtained within a maximum likelihood framework. Methods which belong to this category are presented next.

### 10.2.1 Methods within a maximum-likelihood framework

Assuming that the noise covariance matrix $\mathbf{\Psi}^2$ is known, Hyvärinen (1998) maximizes the joint likelihood of the mixing matrix and the realizations of the independent components. A more popular approach proposed by Attias (1999) and Moulines, Cardoso, and Gassiat (1997) is to introduce a generative parametric density model for the distributions of the independent sources based on mixtures of Gaussians (MoG). Unlike Attias (1999), Moulines, Cardoso, and Gassiat (1997) do not discuss the reconstruction of the sources once an estimate of the mixing matrix is obtained. The approach by Attias (1999) is a general model named independent factor analysis in which $\mathbf{\Psi}^2$ is not necessarily diagonal. It reduces to EFA when $\mathbf{\Psi}^2$ is diagonal and the model sources are normally distributed. Independent factor analysis is described in the sequel.

Let the independent sources $\xi_i$ $(i = 1, \ldots, k)$ have arbitrary densities $f_{\xi_i}(\xi_i|\theta_i)$, where the $i$-th source density is parameterized by $\theta_i$. Then, the unknown parameters of the noisy ICA model (10.1) are $\Theta = \{\mathrm{M}, \mathbf{\Psi}^2, \boldsymbol{\theta}\}$, where $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_k\}$. The resulting model density for the observed signals is

$$\begin{aligned} f_{\mathrm{x}}(\mathrm{x}|\Theta) &= \int f_{\mathrm{x}|\boldsymbol{\xi}}(\mathrm{x}|\boldsymbol{\xi}) f_{\boldsymbol{\xi}}(\boldsymbol{\xi}|\boldsymbol{\theta})\, d\boldsymbol{\xi} \ , \\ &= \int \phi(\mathrm{x}; \mathrm{M}\boldsymbol{\xi}, \mathbf{\Psi}^2) \prod_{i=1}^{k} f_{\xi_i}(\xi_i|\theta_i)\, d\boldsymbol{\xi} \ , \end{aligned} \tag{10.29}$$

where $\phi(\bullet; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the multivariate normal pdf with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ and $d\boldsymbol{\xi} = \prod_i d\xi_i$. To choose a parametric form for $f_{\xi_i}(\xi_i|\theta_i)$, which is both sufficiently general to model arbitrary source densities and allows one to solve the integral in (10.29) analytically, a factorized MoG model for the sources is adopted (Pawitan, 2001, pp. 349-352). Assume that source $i$ is a mixture of $m_i$ Gaussians with means $\mu_{i,q_i}$, variances $v_{i,q_i}$, and mixing proportions $\pi_{i,q_i}$ $(q_i = 1, \ldots, m_i)$:

$$f_{\xi_i}(\xi_i|\theta_i) = \sum_{q_i=1}^{m_i} \pi_{i,q_i} \phi(\xi_i; \mu_{i,q_i}, v_{i,q_i}) \ , \tag{10.30}$$

where $\theta_i = \{\pi_{i,q_i}, \mu_{i,q_i}, v_{i,q_i}\}$ and $\sum_{q_i=1}^{m_i} \pi_{i,q_i} = 1$ for each source. The parametric form in (10.30) provides a probabilistic generative description of the sources in which the different Gaussians play the role of hidden states. Viewed in $k$-dimensional space, the joint source density $f_{\boldsymbol{\xi}}(\boldsymbol{\xi}|\boldsymbol{\theta})$ is itself a MoG. Its collective hidden states $\mathbf{q} = (q_1, \ldots, q_k)^\top$ run over all $\prod_{i=1}^k m_i$ possible combinations of source states. Each state $\mathbf{q}$ corresponds to a $k$-dimensional Gaussian density whose mixing proportions $\boldsymbol{\pi}_\mathbf{q} = \prod_{i=1}^k \pi_{i,q_i}$, mean $\boldsymbol{\mu}_\mathbf{q} = (\mu_{1,q_1}, \ldots, \mu_{k,q_k})^\top$, and covariance matrix $\boldsymbol{\Upsilon}_\mathbf{q} = \text{diag}(\upsilon_{1,q_1}, \ldots, \upsilon_{k,q_k})$ are determined by those of the constituent source state. Hence,

$$f_{\boldsymbol{\xi}}(\boldsymbol{\xi}|\boldsymbol{\theta}) = \prod_{i=1}^{k} f_{\xi_i}(\xi_i|\theta_i) = \sum_{\mathbf{q}} \pi_\mathbf{q} \phi(\boldsymbol{\xi}; \boldsymbol{\mu}_\mathbf{q}, \boldsymbol{\Upsilon}_\mathbf{q}) \ , \tag{10.31}$$

where the sum over all collective states $\mathbf{q}$ represents summing over all the individual source states, that is, $\sum_\mathbf{q} = \sum_{q_1=1}^{m_1} + \cdots + \sum_{q_k=1}^{m_k}$.

Using (10.29) and (10.31), the likelihood $\mathcal{L}$ of the parameter set $\Theta$ given the independent and identically distributed data $\mathcal{X} = \{\mathrm{x}_1, \ldots, \mathrm{x}_n\}$ is

$$\begin{aligned}
\mathcal{L}(\Theta|\mathcal{X}) &= \prod_{i=1}^{n} f_{\mathrm{x}_i}(\mathrm{x}_i|\Theta) \ , \\
&= \prod_{i=1}^{n} \sum_{\mathbf{q}} \int \pi_\mathbf{q} \phi(\mathrm{x}_i; \mathrm{M}\boldsymbol{\xi}, \Psi^2) \phi(\boldsymbol{\xi}; \boldsymbol{\mu}_\mathbf{q}, \boldsymbol{\Upsilon}_\mathbf{q}) \, d\boldsymbol{\xi} \ , \\
&= \prod_{i=1}^{n} \sum_{\mathbf{q}} \int \pi_\mathbf{q} \phi(\mathrm{x}_i; \mathrm{M}\boldsymbol{\mu}_\mathbf{q}, \mathrm{M}\boldsymbol{\Upsilon}_\mathbf{q}\mathrm{M}^\top + \Psi^2) \, d\boldsymbol{\xi} \ .
\end{aligned} \tag{10.32}$$

In Attias (1999), the model parameters are chosen to minimize a contrast such as the Kullback-Leibler distance (10.14) which measures the distance between the model and the observed sensor densities. Minimizing the KL distance is equivalent to finding $\Theta$ that maximizes $\mathcal{L}$ or $\log \mathcal{L}$, that is,

$$\hat{\Theta}_{\text{ML}} = \arg\max_{\Theta} \log \mathcal{L}(\Theta|\mathcal{X}) \ . \tag{10.33}$$

The model parameters can be found by the iterative EM algorithm (Dempster, Laird, and Rubin, 1977). The EM algorithm is suited for problems where the data is incomplete or some parts are missing. In the context of noisy ICA, $\boldsymbol{\xi}$ and $\mathbf{q}$ represent the missing data and $\mathcal{X}$ is called the incomplete data. The likelihood in (10.32), $\mathcal{L}(\Theta|\mathcal{X})$, is called the incomplete likelihood. Let $\mathcal{Y} = \{\boldsymbol{\xi}, \mathbf{q}\}$ denote the missing information. If it is possible to 'fill in' $\mathcal{Y}$, the analysis of the complete likelihood $\mathcal{L}(\Theta|\mathcal{X}, \mathcal{Y})$ is relatively simple. The EM algorithm consists of two steps. In the first step (E-step), one finds the expected value of $\log \mathcal{L}(\Theta|\mathcal{X}, \mathcal{Y})$ with respect to the unknown data $\mathcal{Y}$ given $\mathcal{X}$ and the current parameter estimates:

$$Q(\Theta, \Theta^{(i-1)}) = \mathrm{E}\left[\log \mathcal{L}(\Theta|\mathcal{X}, \mathcal{Y})|\mathcal{X}, \Theta^{(i-1)})\right] \ .$$

The parameters obtained in the previous iteration that are used to evaluate the expectation are denoted by $\Theta^{(i-1)}$ and $\Theta$ are the new parameters that are to be optimized to increase $Q$. The second step (M-step) is to maximize the expectation computed in the first step, that is,

$$\Theta^{(i)} = \arg\max_{\Theta} Q(\Theta, \Theta^{(i-1)}) \ .$$

The two steps are repeated until convergence is achieved. Each iteration is guaranteed to increase the log-likelihood.

Once the estimates of the parameters have been obtained, estimates of the latent variables can be constructed from the determined density model. The two most common ways are to use the minimum mean squared error (MMSE) estimator or the maximum a posteriori (MAP) estimator. Each satisfies a different optimality criterion.

The MMSE minimizes $\mathrm{E}[(\boldsymbol{\xi} - \hat{\boldsymbol{\xi}})^{\mathsf{T}}(\boldsymbol{\xi} - \hat{\boldsymbol{\xi}})]^2$ and the optimal estimate is given by the conditional mean of the sources given the observed data:

$$\hat{\boldsymbol{\xi}}_{\text{MMSE}} = \mathrm{E}(\boldsymbol{\xi}|\mathbf{x}) = \int \boldsymbol{\xi} f_{\boldsymbol{\xi}|\mathbf{x}}(\boldsymbol{\xi}|\mathbf{x}, \Theta)\, d\boldsymbol{\xi} \ , \tag{10.34}$$

where $f_{\boldsymbol{\xi}|\mathbf{x}}(\boldsymbol{\xi}|\mathbf{x}, \Theta)$ is the posterior density of the sources which depends on the generative parameters $\Theta$. The conditional mean (10.34) has already been calculated in the

E-step of the EM algorithm (see Attias, 1999, Appendix).

The MAP estimator maximizes the source posterior:

$$f_{\boldsymbol{\xi}|\mathbf{x}}(\boldsymbol{\xi}|\mathbf{x}) = \frac{f_{\mathbf{x}|\boldsymbol{\xi}}(\mathbf{x}|\boldsymbol{\xi}) f_{\boldsymbol{\xi}}(\boldsymbol{\xi}|\boldsymbol{\theta})}{f_{\mathbf{x}}(\mathbf{x})} \ . \tag{10.35}$$

For given $\mathbf{x}$, maximizing the source posterior is equivalent to maximizing the numerator in (10.35) or its logarithm, that is,

$$\hat{\boldsymbol{\xi}}_{\text{MAP}} = \arg\max_{\boldsymbol{\xi}} \left( \log f_{\mathbf{x}|\boldsymbol{\xi}}(\mathbf{x}|\boldsymbol{\xi}) + \sum_{i=1}^{k} \log f_{\xi_i}(\xi_i|\theta_i) \right) \ . \tag{10.36}$$

To find $\hat{\boldsymbol{\xi}}_{\text{MAP}}$ the quantity $\log f_{\mathbf{x}|\boldsymbol{\xi}}(\mathbf{x}|\boldsymbol{\xi}) + \sum_{i=1}^{k} \log f_{\xi_i}(\xi_i|\theta_i)$ in (10.36) can be maximized iteratively by means of the gradient ascent method (Attias, 1999).

Both the MAP and the MMSE estimator are non-linear functions of the data. For normally distributed sources, they are equal and reduce to the linear estimator (3.23) for predicting common factor scores in EFA devised by Ledermann (1939).

Independent factor analysis has the benefit of being able to estimate all model unknowns $\mathbf{M}$, $\boldsymbol{\Xi}$ and $\boldsymbol{\Psi}^2$. However, this approach has a couple of drawbacks. One problem is that it is not clear how to choose the number of Gaussians for modelling the source distributions. Moreover, EM algorithms are notorious for being slow to converge (Pawitan, 2001, p. 348), which limit their use to analyzing data sets having low-dimensionality.

### 10.2.2 Noisy ICA as a method of factor rotation

A more practical approach for fitting the noisy ICA model is to transform the problem of obtaining approximately independent realizations of the factors into a specific EFA task. This amounts to separating the estimation procedure into two parts.

The first part is to decorrelate the data and to reduce its dimensionality. This can be achieved by sphering the observations. In the spirit of Section 3.2.1, the sphering operation in noisy ICA is called quasi-sphering (Hyvärinen, Karhunen, and Oja, 2001). Since in preliminary quasi-sphering of the noisy ICA model the effect of the noise covariance matrix $\mathbf{\Psi}^2$ has to be taken into account, EFA is employed instead of PCA. Recall the noisy ICA model in its sample form (3.14):

$$\mathbf{X} = \mathbf{\Xi}\mathbf{M}^\top + \mathbf{U} \ . \tag{10.37}$$

Quasi-sphering of $\mathbf{X}$ in (10.37) can be performed as follows. Assume that as a result of an EFA solution, a pair of estimates $\{\hat{\mathbf{\Lambda}}, \hat{\mathbf{\Psi}}^2\}$ for the matrix of factor loadings and the matrix of unique factor variances is available. Ikeda and Toyama (2000) suggest to use ML factor analysis for this factor extraction problem. Unkel and Trendafilov (2007) and Stegeman and Mooijaart (2008) propose (unweighted) LS fitting.

Hence, by means of EFA, the noisy ICA mixing matrix $\mathbf{M}$ in (10.37) can be identified with the ambiguity of an orthogonal rotation. Let the eigenvalue decomposition of the noise-free (reduced) sample covariance matrix $\mathbf{S_X} - \hat{\mathbf{\Psi}}^2$ of rank $r$ $(r \leq p)$ be

$$\mathbf{S_X} - \hat{\mathbf{\Psi}}^2 = \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^\top = \mathbf{E}\mathbf{\Omega}\mathbf{E}^\top = \sum_{i=1}^{r} \omega_i \mathbf{e}_i \mathbf{e}_i^\top \ , \tag{10.38}$$

where $\mathbf{\Omega} = \text{diag}(\omega_1, \ldots, \omega_r)$ is an $r \times r$ diagonal matrix containing the positive eigenvalues of $\mathbf{S_X} - \hat{\mathbf{\Psi}}^2$, $\omega_1 \geq \cdots \geq \omega_r > 0$, on its main diagonal and $\mathbf{E} \in \mathbb{R}^{p \times r}$ is an orthonormal matrix whose columns $\mathbf{e}_1, \ldots, \mathbf{e}_r$ are the corresponding unit-norm eigenvectors of $\omega_1, \ldots, \omega_r$. To reduce the dimensionality of the data to $k$ $(\ll r)$ dimensions, assume that the sum in (10.38) is truncated after $k$ terms. Postmultiplying $\mathbf{X}$ by $\mathbf{Q}^\top = \mathbf{E}\mathbf{\Omega}^{-1/2} \in \mathbb{R}^{p \times k}$ gives

$$\tilde{\mathbf{X}} = \mathbf{X}\mathbf{E}\mathbf{\Omega}^{-1/2} = \mathbf{X}\mathbf{Q}^\top \ , \tag{10.39}$$

where $\tilde{\mathbf{X}} = (\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_k) \in \mathbb{R}^{n \times k}$ is the quasi-sphered data matrix. Since

$$\tilde{\mathbf{X}} = \Xi\mathbf{M}^\top\mathbf{Q}^\top + \mathbf{U}\mathbf{Q}^\top = \Xi\tilde{\mathbf{M}}^\top + \tilde{\mathbf{U}} \ , \tag{10.40}$$

the quasi-sphered data matrix follows a noisy ICA model as well with square mixing matrix $\tilde{\mathbf{M}} = \mathbf{Q}\mathbf{M} \in \mathbb{R}^{k \times k}$ and linear transform of the unique factor matrix $\tilde{\mathbf{U}} = \mathbf{U}\mathbf{Q}^\top \in \mathbb{R}^{p \times k}$. Since

$$\begin{aligned} \mathbf{S}_{\tilde{\mathbf{X}}} &= \frac{1}{n-1}\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} = \frac{1}{n-1}\tilde{\mathbf{M}}\Xi^\top\Xi\tilde{\mathbf{M}}^\top + \frac{1}{n-1}\tilde{\mathbf{U}}^\top\tilde{\mathbf{U}} \ , \\ &= \mathbf{Q}\mathbf{M}\mathbf{M}^\top\mathbf{Q}^\top + \mathbf{Q}\Psi^2\mathbf{Q}^\top \ , \\ &= \Omega^{-1/2}\mathbf{E}^\top\mathbf{E}\Omega\mathbf{E}^\top\mathbf{E}\Omega^{-1/2} + \mathbf{Q}\Psi^2\mathbf{Q}^\top \ , \\ &= \mathbf{I}_k + \mathbf{Q}\Psi^2\mathbf{Q}^\top \ , \end{aligned} \tag{10.41}$$

the new mixing matrix $\tilde{\mathbf{M}}$ is orthogonal. As (10.41) reveals, the covariance matrix of the quasi-sphered data is not the identity matrix. In other words, after preprocessing the data by $\mathbf{Q}^\top$ only the part of $\tilde{\mathbf{X}}$ due to the common factors is uncorrelated whereas the part of $\tilde{\mathbf{X}}$ due to the noise remains correlated. The transformed data $\tilde{\mathbf{X}}$ has noise covariance matrix $\mathbf{Q}\Psi^2\mathbf{Q}^\top = \Omega^{-1/2}\mathbf{E}^\top\Psi^2\mathbf{E}\Omega^{-1/2}$.

Ikeda and Toyama (2000) proposed the following alternative approach for performing quasi-sphering. Let $\mathbf{Q} = (\hat{\Lambda}^\top\hat{\Psi}^{-2}\hat{\Lambda})^{-1}\hat{\Lambda}^\top\hat{\Psi}^{-2} \in \mathbb{R}^{k \times p}$ be a generalized inverse matrix of $\hat{\Lambda}$ satisfying the condition (e.g., Lütkepohl, 1996):

$$\hat{\Lambda}\mathbf{Q}\hat{\Lambda} = \hat{\Lambda}(\hat{\Lambda}^\top\hat{\Psi}^{-2}\hat{\Lambda})^{-1}\hat{\Lambda}^\top\hat{\Psi}^{-2}\hat{\Lambda} = \hat{\Lambda} \ . \tag{10.42}$$

The quasi-sphered data matrix $\tilde{\mathbf{X}} = \mathbf{X}\mathbf{Q}^\top = \Xi\tilde{\mathbf{M}}^\top + \tilde{\mathbf{U}}$ has sample covariance matrix

$$\begin{aligned} \mathbf{S}_{\tilde{\mathbf{X}}} &= \mathbf{Q}\hat{\Lambda}\hat{\Lambda}^\top\mathbf{Q}^\top + \mathbf{Q}\Psi^2\mathbf{Q}^\top \ , \\ &= (\hat{\Lambda}^\top\hat{\Psi}^{-2}\hat{\Lambda})^{-1}\hat{\Lambda}^\top\hat{\Psi}^{-2}\hat{\Lambda}\hat{\Lambda}^\top\hat{\Psi}^{-2}\hat{\Lambda}(\hat{\Lambda}^\top\hat{\Psi}^{-2}\hat{\Lambda})^{-1} + (\hat{\Lambda}^\top\hat{\Psi}^{-2}\hat{\Lambda})^{-1} \ , \\ &= \mathbf{I}_k + (\hat{\Lambda}^\top\hat{\Psi}^{-2}\hat{\Lambda})^{-1} \ , \end{aligned} \tag{10.43}$$

and noise covariance matrix $(\hat{\mathbf{\Lambda}}^\top \hat{\mathbf{\Psi}}^{-2} \hat{\mathbf{\Lambda}})^{-1}$. Note that preprocessing $\mathbf{X}$ by $\mathbf{Q}^\top = \hat{\mathbf{\Psi}}^{-2} \hat{\mathbf{\Lambda}}(\hat{\mathbf{\Lambda}}^\top \hat{\mathbf{\Psi}}^{-2} \hat{\mathbf{\Lambda}})^{-1}$ is equivalent to obtaining the common factor scores (3.25) (Bartlett, 1937).

The generalized inverse of $\hat{\mathbf{\Lambda}}$ is not unique. One could also choose $\mathbf{Q} = (\hat{\mathbf{\Lambda}}^\top \hat{\mathbf{\Lambda}})^{-1} \hat{\mathbf{\Lambda}}^\top \in \mathbb{R}^{k \times p}$ (Stegeman and Mooijaart, 2008) which corresponds to the unique Moore-Penrose inverse of $\hat{\mathbf{\Lambda}}$ satisfying the four conditions (Lütkepohl, 1996, pp. 34-35):

$$\hat{\mathbf{\Lambda}}\mathbf{Q}\hat{\mathbf{\Lambda}} = \hat{\mathbf{\Lambda}}(\hat{\mathbf{\Lambda}}^\top \hat{\mathbf{\Lambda}})^{-1} \hat{\mathbf{\Lambda}}^\top \hat{\mathbf{\Lambda}} = \hat{\mathbf{\Lambda}} \ . \tag{10.44}$$

$$\begin{aligned} \mathbf{Q}\hat{\mathbf{\Lambda}}\mathbf{Q} &= (\hat{\mathbf{\Lambda}}^\top \hat{\mathbf{\Lambda}})^{-1} \hat{\mathbf{\Lambda}}^\top \hat{\mathbf{\Lambda}}(\hat{\mathbf{\Lambda}}^\top \hat{\mathbf{\Lambda}})^{-1} \hat{\mathbf{\Lambda}}^\top \ , \\ &= (\hat{\mathbf{\Lambda}}^\top \hat{\mathbf{\Lambda}})^{-1} \hat{\mathbf{\Lambda}}^\top = \mathbf{Q} \ . \end{aligned} \tag{10.45}$$

$$\begin{aligned} (\hat{\mathbf{\Lambda}}\mathbf{Q})^\top &= (\hat{\mathbf{\Lambda}}(\hat{\mathbf{\Lambda}}^\top \hat{\mathbf{\Lambda}})^{-1} \hat{\mathbf{\Lambda}}^\top)^\top \ , \\ &= \hat{\mathbf{\Lambda}}(\hat{\mathbf{\Lambda}}^\top \hat{\mathbf{\Lambda}})^{-1} \hat{\mathbf{\Lambda}}^\top = \hat{\mathbf{\Lambda}}\mathbf{Q} \ . \end{aligned} \tag{10.46}$$

$$(\mathbf{Q}\hat{\mathbf{\Lambda}})^\top = \mathbf{Q}\hat{\mathbf{\Lambda}} \ , \tag{10.47}$$

where (10.47) follows from the fact that $\text{rank}(\hat{\mathbf{\Lambda}}) = k \Leftrightarrow \mathbf{Q}\hat{\mathbf{\Lambda}} = \mathbf{I}_k$.

Note that postmultiplying $\mathbf{X}$ by $\mathbf{Q}^\top = \hat{\mathbf{\Lambda}}(\hat{\mathbf{\Lambda}}^\top \hat{\mathbf{\Lambda}})^{-1}$ is equivalent to obtaining the common factor scores (4.6) (Horst, 1965). For this choice of $\mathbf{Q}$, again it holds that $\mathbf{S}_{\tilde{\mathbf{X}}} \neq \mathbf{I}_k$. Summarizing, the choices of $\mathbf{Q}$ made in the ICA literature lead to preprocessed data $\tilde{\mathbf{X}}$ which are still correlated.

To obtain preprocessed data with no second-order correlations, Unkel and Trendafilov (2007) proposed to use $\mathbf{Q} = (\hat{\mathbf{\Lambda}}^\top \hat{\mathbf{\Psi}}^{-2} \mathbf{S}_\mathbf{X} \hat{\mathbf{\Psi}}^{-2} \hat{\mathbf{\Lambda}})^{-1/2} \hat{\mathbf{\Lambda}}^\top \hat{\mathbf{\Psi}}^{-2} \in \mathbb{R}^{k \times p}$ instead. Postmultiplying $\mathbf{X}$ by $\mathbf{Q}^\top$ is equivalent to obtaining the common factor scores (3.26) (Anderson and Rubin, 1956) and leads to uncorrelated data $\tilde{\mathbf{X}}$ because

$$\begin{aligned}
\mathbf{S}_{\tilde{\mathbf{X}}} &= \mathbf{Q}\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^\top\mathbf{Q}^\top + \mathbf{Q}\mathbf{\Psi}^2\mathbf{Q}^\top \ , \\
&= (\hat{\mathbf{\Lambda}}^\top\hat{\mathbf{\Psi}}^{-2}\mathbf{S}_\mathbf{X}\hat{\mathbf{\Psi}}^{-2}\hat{\mathbf{\Lambda}})^{-1/2}\hat{\mathbf{\Lambda}}^\top\hat{\mathbf{\Psi}}^{-2}\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^\top\hat{\mathbf{\Psi}}^{-2}\hat{\mathbf{\Lambda}}(\hat{\mathbf{\Lambda}}^\top\hat{\mathbf{\Psi}}^{-2}\mathbf{S}_\mathbf{X}\hat{\mathbf{\Psi}}^{-2}\hat{\mathbf{\Lambda}})^{-1/2} \\
&\quad +(\hat{\mathbf{\Lambda}}^\top\hat{\mathbf{\Psi}}^{-2}\mathbf{S}_\mathbf{X}\hat{\mathbf{\Psi}}^{-2}\hat{\mathbf{\Lambda}})^{-1/2}\hat{\mathbf{\Lambda}}^\top\hat{\mathbf{\Psi}}^{-2}\hat{\mathbf{\Psi}}^2\hat{\mathbf{\Psi}}^{-2}\hat{\mathbf{\Lambda}}(\hat{\mathbf{\Lambda}}^\top\hat{\mathbf{\Psi}}^{-2}\mathbf{S}_\mathbf{X}\hat{\mathbf{\Psi}}^{-2}\hat{\mathbf{\Lambda}})^{-1/2} \ , \\
&= (\hat{\mathbf{\Lambda}}^\top\hat{\mathbf{\Psi}}^{-2}\mathbf{S}_\mathbf{X}\hat{\mathbf{\Psi}}^{-2}\hat{\mathbf{\Lambda}})^{-1/2}\hat{\mathbf{\Lambda}}^\top\hat{\mathbf{\Psi}}^{-2}\mathbf{S}_\mathbf{X}\hat{\mathbf{\Psi}}^{-2}\hat{\mathbf{\Lambda}}(\hat{\mathbf{\Lambda}}^\top\hat{\mathbf{\Psi}}^{-2}\mathbf{S}_\mathbf{X}\hat{\mathbf{\Psi}}^{-2}\hat{\mathbf{\Lambda}})^{-1/2} \ , \\
&= \mathbf{I}_k \ . \qquad (10.48)
\end{aligned}$$

Once quasi-sphered data have been obtained, the second part of the estimation procedure is to find an orthogonal $k \times k$ matrix $\mathbf{T}$ which rotates $\tilde{\mathbf{X}}$ towards independence, that is,

$$\hat{\Xi} = \tilde{\mathbf{X}}\mathbf{T} \ , \qquad (10.49)$$

where $\hat{\Xi}$ is an estimate of the matrix $\Xi$ of independent common factor scores of the $n$ observations on $k$ common factors.

Noisy ICA starts then essentially from an EFA solution and seeks a rotation matrix which minimizes the dependence between the common factors. From this point of view, noisy ICA is another method of factor rotation along with the well-known simple structure rotation methods such as e.g. Varimax which originated in psychometrics (Hastie, Tibshirani, and Friedman, 2009). The difference between EFA and noisy ICA is that the rotation criteria involve factor loadings for EFA and factor scores for noisy ICA.

To find $\mathbf{T}$, one needs a rotation criterion and an optimization algorithm. Ikeda and Toyama (2000) optimize the criterion (10.24) proposed by Cardoso and Souloumiac (1993). The popularity of this criterion stems mainly from the fact that it can be optimized by means of simultaneous diagonalization of a set of fourth-order cumulant matrices. Let $\mathbf{V}$ be an arbitrary $k \times k$ orthogonal matrix. Suppose that $\{\mathbf{N}_i | i = 1, \ldots, k^2\}$ is any basis for the $k^2$-dimensional linear space of $k \times k$ matrices and let $\mathbf{C}(\mathbf{N}_i)$ $(i = 1, \ldots, k^2)$ denote the matrix of fourth-order cumulants in which the $(p, q)$-th element is defined as

$$\sum_{r=1}^{k} \sum_{s=1}^{k} \kappa_{1,1,1,1}(\tilde{\mathbf{x}}_p, \tilde{\mathbf{x}}_q, \tilde{\mathbf{x}}_r, \tilde{\mathbf{x}}_s)\, n_{rs}^{(i)} \ , \tag{10.50}$$

where $n_{rs}^{(i)}$ is the $(r, s)$-th element of $\mathbf{N}_i$. One simple way to choose the set $\{\mathbf{N}_i | i = 1, \ldots, k^2\}$ is as $\{\boldsymbol{\iota}_r \boldsymbol{\iota}_s^\top | 1 \leq r, s \leq k\}$, where $\boldsymbol{\iota}_r$ is a $k$-dimensional column vector with a single entry of one in the $r$-th position and zeros elsewhere (Cardoso, 1999).

Minimizing (10.24) is then equivalent to maximizing (Cardoso and Souloumiac, 1993):

$$\mathcal{F}_{JADE}(\mathbf{V}) = \sum_{i=1}^{k^2} \text{trace}\left[\text{diag}(\mathbf{V}^\top \mathbf{C}(\mathbf{N}_i)\mathbf{V})\right]^2 \ . \tag{10.51}$$

Maximizing (10.51) over all orthogonal matrices $\mathbf{V}$ gives $\mathbf{T}$. The algorithm that maximizes (10.51) is called JADE (Joint Approximate Diagonalization of Eigenmatrices) and is a Jacobi-type algorithm (Cardoso and Souloumiac, 1993, 1996). The idea is to parameterize $\mathbf{V}$ by a product of plane rotations using Jacobi rotation matrices and then optimize with respect to the Givens angle, $\theta$, involved in each rotation. The $(i, j)$-th plane rotation on the whole set of cumulant matrices is performed by a Jacobi rotation matrix $\mathbf{R}(i, j, \theta)$ which is an identity matrix where the $(i, i)$ and $(j, j)$ entries are replaced by $\cos(\theta)$, the $(i, j)$ entry is replaced by $-\sin(\theta)$, and the $(j, i)$ entry is replaced by $\sin(\theta)$. At each step the Givens angle is found by updating $\mathbf{V} \leftarrow \mathbf{VR}$ and solving (10.51). Givens rotations are orthogonal and multiplication by $\mathbf{R}$ amounts to a rotation of $\theta$ radians on the $(i, j)$-th coordinate plane. The aim of JADE is to make the cumulant matrices as diagonal as possible which coincides with making the columns of $\tilde{\mathbf{X}}$ as independent as possible. Once the optimal rotation matrix $\mathbf{T}$ and hence $\hat{\Xi}$ is found, an estimate for the ICA mixing matrix $\mathbf{M}$ is obtained by $\hat{\mathbf{M}} = \hat{\mathbf{\Lambda}}\mathbf{T}$.

In addition to algorithms diagonalizing fourth-order cumulant matrices, Stegeman and Mooijaart (2008) used the Newton-type FastICA algorithm (Hyvärinen, 1999a) to maximize the sample analogue of the approximation to negentropy (10.26) to find $\mathbf{T}$.

Kano, Miyamoto, and Shimizu (2003) proposed a rotation criterion derived from the Crawford-Ferguson family of rotation criteria in EFA (Crawford and Ferguson, 1970):

$$\begin{aligned} \mathcal{F}_{CF}(\mathbf{V}) &= (1-\tau)\,\mathrm{trace}(\hat{\mathbf{\Lambda}}\mathbf{V}\odot\hat{\mathbf{\Lambda}}\mathbf{V})^{\top}(\hat{\mathbf{\Lambda}}\mathbf{V}\odot\hat{\mathbf{\Lambda}}\mathbf{V})(\mathbf{1}_k\mathbf{1}_k^{\top}-\mathbf{I}_k) \\ &\quad +\tau\,\mathrm{trace}(\hat{\mathbf{\Lambda}}\mathbf{V}\odot\hat{\mathbf{\Lambda}}\mathbf{V})^{\top}(\mathbf{1}_p\mathbf{1}_p^{\top}-\mathbf{I}_p)\ , \end{aligned} \tag{10.52}$$

where $\hat{\mathbf{\Lambda}}$ is an initial EFA loading matrix. The function $\mathcal{F}_{CF}(\mathbf{V})$ provides a family of orthogonal and oblique rotation criteria by choosing different values of $\tau$. For example, when restricted to orthogonal rotation, the Crawford-Ferguson family with $\tau = 1/p$ yields the Varimax criterion (Kaiser, 1958).

Assume that in (10.52) $\hat{\mathbf{\Lambda}}$ is replaced by $\tilde{\mathbf{X}}$. The idea of Varimax, maximizing the variance of the squared entries in each column of $\tilde{\mathbf{X}}\mathbf{V}$, is closely related to the maximization of the fourth-order cumulant of each column of $\tilde{\mathbf{X}}\mathbf{V}$. In fact, by setting $\tau = 3/p$, the criterion $\mathcal{F}_{CF}(\mathbf{V})$ is proportional to the criterion $\sum_{j=1}^{k}\kappa_4(j)$, where $\kappa_4(j)$ denotes the fourth-order cumulant of the $j$-th column of $\tilde{\mathbf{X}}\mathbf{V}$. Minimizing (10.52) over all orthogonal matrices $\mathbf{V}$ gives $\mathbf{T}$. However, in Kano, Miyamoto, and Shimizu (2003) no indication is given which algorithm is actually used to optimize (10.52).

Moreover, although Kano, Miyamoto, and Shimizu (2003) explore the connection between noisy ICA and EFA with factor rotation, their procedure is actually used to fit the noise-free ICA model. As a pre-analysis, PCA is used to sphere the data. Then

(10.52) is optimized to obtain the separation matrix.

Jennrich and Trendafilov (2005) also considered the noise-free case. The authors introduced a criterion being a sum of squared fourth-order statistics formed by covariances computed from squared components. Optimization of the criterion is carried out by means of the continuous-time projected gradient method as described in Chapter 2. This specific rotation criterion will be discussed in more detail in the next Chapter.

Unkel and Trendafilov (2007) used the criterion introduced by Jennrich and Trendafilov (2005) to fit the noisy ICA model taking explicitly unique factors into account. Instead of solving ODEs, an iterative scheme proposed by Jennrich (2001) is used to keep the gradient flow following the steepest descent direction and moving on the constrained manifold of orthogonal matrices simultaneously (see also Bernaards and Jennrich, 2005).

# Chapter 11

# Independent Exploratory Factor Analysis

In contrast to the standard noisy ICA model with random latent sources, it is assumed in this Chapter that the underlying factors are nonrandom quantities or parameters to be estimated. This new method is named independent exploratory factor analysis (IEFA) (Unkel, Trendafilov, Hannachi, and Jolliffe, 2009; Unkel and Trendafilov, 2009b).

A fitting solution for IEFA is obtained by exploiting the link between IEFA and EFA with factor rotation. That is, starting from an initial EFA solution an orthogonal rotation matrix is sought such that the common factor scores are approximately independent. The rotation matrix found is then applied to the EFA loading matrix to compensate for the rotation of the scores.

To obtain the initial EFA solution, the iterative algorithms for simultaneous LS estimation of all EFA model unknowns developed in Part II are employed. Unlike the methods discussed in the previous Chapter for obtaining an EFA solution and quasi-sphered data, the algorithms used in IEFA are based on the SVD of data matrices. The SVD of data matrices is computationally efficient, facilitates the computation of sphered factor scores, and works well when the number of variables exceeds the number of observations. Section 11.1 describes the rotation criterion and the optimization algorithm. In Section 11.2, IEFA is applied to Thurstone's 26-variable box problem in psychometrics.

## 11.1 Rotation criterion and optimization algorithm

Assume that an estimate $\hat{\mathbf{F}}$ for sphered common factor scores is obtained by means of the numerical iterative procedures discussed in Chapter 5 (for $n > p$) or Chapter 7 (for $p \geq n$). To solve the corresponding IEFA problem one needs to go one step further. The common factor scores should be independent. For this reason, they are rotated towards independence, that is,

$$\hat{\Xi} = \hat{\mathbf{F}}\mathbf{T} \ , \tag{11.1}$$

for some $k \times k$ orthogonal matrix $\mathbf{T}$.

To find the matrix $\mathbf{T}$ that leads to approximately independent factor scores $\hat{\Xi}$, an appropriate rotation criterion is set up next which resembles the simple structure rotation criteria in EFA.

If the common factors are independent their squares are also independent. Thus, the model covariance matrix of the squared factors is diagonal. Let $\mathbf{V}$ be an arbitrary orthogonal $k \times k$ matrix and let

$$\mathbf{G} = \hat{\mathbf{F}}\mathbf{V} \ . \tag{11.2}$$

The sample covariance matrix between the element-wise squares of $\mathbf{G}$ is

$$\mathbf{S}_{\mathbf{H}} = \frac{1}{n-1}\mathbf{H}^{\top}\mathbf{C}_n\mathbf{H} \ , \tag{11.3}$$

where $\mathbf{H} = \mathbf{G} \odot \mathbf{G}$ and $\mathbf{C}_n$ is the centring matrix introduced in (1.1). The matrix $\mathbf{S_H}$ is the covariance matrix of the squared orthogonally transformed factor scores $\hat{\mathbf{F}}$.

The problem is to find the $k \times k$ orthogonal matrix $\mathbf{T}$ which makes the covariance matrix $\mathbf{S_H}$ in (11.3) as close as possible to a diagonal matrix. This is equivalent to reducing the off-diagonal elements of $\mathbf{S_H}$ as much as possible. The approximate diagonalization of $\mathbf{S_H}$ can be achieved by minimizing the sum of the squared off-diagonal elements of $\mathbf{S_H}$. For this reason, the following rotation criterion is defined (Jennrich and Trendafilov, 2005; Unkel and Trendafilov, 2007):

$$\mathcal{F}(\mathbf{V}) = \text{trace}(\mathbf{1}_k\mathbf{1}_k^\top - \mathbf{I}_k)(\mathbf{S_H} \odot \mathbf{S_H}) \; . \tag{11.4}$$

The aim is to minimize the sum of the squared off-diagonal elements of $\mathbf{S_H}$ over all orthogonal matrices $\mathbf{V}$ or, equivalently, over all orthogonal rotations $\mathbf{V}$ of $\hat{\mathbf{F}}$. Since $\mathbf{S_H}$ is symmetric, it is sufficient to minimize the elements below or above the diagonal, respectively. Therefore, the criterion (11.4) is multiplied by 1/2. In other words, one needs to solve the following optimization problem:

$$\min_{\mathbf{V} \in \mathcal{O}(k)} \frac{1}{2} \text{trace}(\mathbf{1}_k\mathbf{1}_k^\top - \mathbf{I}_k)(\mathbf{S_H} \odot \mathbf{S_H}) \; , \tag{11.5}$$

where $\mathcal{O}(k)$ denotes the set of $k \times k$ orthogonal matrices as defined in (2.5). Solving (11.5) over all orthogonal matrices $\mathbf{V}$ gives $\mathbf{T}$.

The continuous-time projected gradient approach (see Chapter 2) can be used to find $\mathbf{T}$. In order to apply this approach, one has to construct the projection of the gradient of the objective function (11.4) onto the feasible set $\mathcal{O}(k)$. Thus, one needs to know the gradient $\nabla_\mathbf{V}$ of $\mathcal{F}(\mathbf{V})$ at $\mathbf{V}$ with respect to the Frobenius inner product (2.3). The differential $d\mathcal{F}$ of $\mathcal{F}(\mathbf{V})$ at $\mathbf{V}$ can be expressed in the form (Jennrich, 2001):

$$d\mathcal{F} = \langle \nabla_\mathbf{V}, d\mathbf{V} \rangle \; . \tag{11.6}$$

Thus, a formula for $\nabla_{\mathbf{V}}$ can be found by expressing the differential $d\mathcal{F}$ of $\mathcal{F}(\mathbf{V})$ at $\mathbf{V}$ in this form. The differential of the criterion function in (11.4) is

$$d\mathcal{F} = \mathrm{trace}(\mathbf{1}_k\mathbf{1}_k^\top - \mathbf{I}_k)(\mathbf{S_H} \odot (d\mathbf{S_H})) \ . \tag{11.7}$$

After substituting $d\mathbf{S_H} = (d\mathbf{H})^\top\mathbf{C}_n\mathbf{H} + \mathbf{H}^\top\mathbf{C}_n(d\mathbf{H})$ in (11.7) and making use of the following identity (Magnus and Neudecker, 1988):

$$\mathrm{trace}\mathbf{A}^\top(\mathbf{B} \odot \mathbf{D}) = \mathrm{trace}(\mathbf{A}^\top \odot \mathbf{B}^\top)\mathbf{D} \ ,$$

where $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{D}$ are $n \times k$ matrices, one finds that

$$\begin{aligned} d\mathcal{F} &= \mathrm{trace}(\mathbf{1}_k\mathbf{1}_k^\top - \mathbf{I}_k)\left[\mathbf{S_H} \odot \left[(d\mathbf{H})^\top\mathbf{C}_n\mathbf{H} + \mathbf{H}^\top\mathbf{C}_n(d\mathbf{H})\right]^\top\right] \ , \\ &= 2\,\mathrm{trace}(d\mathbf{H})^\top\mathbf{C}_n\mathbf{H}\left[(\mathbf{1}_k\mathbf{1}_k^\top - \mathbf{I}_k) \odot \mathbf{S_H}\right] \ . \end{aligned} \tag{11.8}$$

Since $\mathbf{H} = \mathbf{G} \odot \mathbf{G}$, it follows that

$$d\mathbf{H} = d\mathbf{G} \odot \mathbf{G} + \mathbf{G} \odot d\mathbf{G} = 2(d\mathbf{G} \odot \mathbf{G}) \ . \tag{11.9}$$

Using (11.9), (11.8) can be written as

$$\begin{aligned} d\mathcal{F} &= 4\,\mathrm{trace}(d\mathbf{G} \odot \mathbf{G})^\top\mathbf{C}_n\mathbf{H}\left[(\mathbf{1}_k\mathbf{1}_k^\top - \mathbf{I}_k)\mathbf{S_H}\right] \ , \\ &= 4\,\mathrm{trace}(d\mathbf{G})^\top\left[\mathbf{G} \odot \mathbf{C}_n\mathbf{H}\left[(\mathbf{1}_k\mathbf{1}_k^\top - \mathbf{I}_k)\mathbf{S_H}\right]\right] \ . \end{aligned} \tag{11.10}$$

Since $\mathbf{G} = \hat{\mathbf{F}}\mathbf{V}$ and $d\mathbf{G} = \hat{\mathbf{F}}d\mathbf{V}$, one can finally express $d\mathcal{F}$ as

$$\begin{aligned} d\mathcal{F} &= 4\,\mathrm{trace}(d\mathbf{V})^\top\hat{\mathbf{F}}^\top(\mathbf{G} \odot \left[\mathbf{C}_n\mathbf{H}\left[(\mathbf{1}_k\mathbf{1}_k^\top - \mathbf{I}_k)\mathbf{S_H}\right]\right]) \ , \\ &= 4\,\langle\hat{\mathbf{F}}^\top(\mathbf{G} \odot \left[\mathbf{C}_n\mathbf{H}\left[(\mathbf{1}_k\mathbf{1}_k^\top - \mathbf{I}_k)\mathbf{S_H}\right]\right]), d\mathbf{V}\rangle \ . \end{aligned} \tag{11.11}$$

Hence, using (11.6), the gradient of the objective function (11.4) with respect to the Frobenius inner product (2.3) is

$$\nabla_{\mathbf{V}} = 4\,\hat{\mathbf{F}}^\top(\mathbf{G} \odot \left[\mathbf{C}_n\mathbf{H}\left[(\mathbf{1}_k\mathbf{1}_k^\top - \mathbf{I}_k) \odot \mathbf{S_H}\right]\right]) \ . \tag{11.12}$$

Then, the projection $\dot{\mathbf{V}}$ of $\nabla_{\mathbf{V}}$ onto the tangent space $\mathcal{T}_{\mathbf{V}}\mathcal{O}(k)$ at $\mathbf{V} \in \mathcal{O}(k)$ is

$$\dot{\mathbf{V}} = \mathbf{V}\frac{\nabla_{\mathbf{V}}^{\top}\mathbf{V} - \mathbf{V}^{\top}\nabla_{\mathbf{V}}}{2} . \tag{11.13}$$

A solution of the minimization problem (11.5) can then be found as a limit point $\mathbf{V}_{\infty}$ of the gradient flow $\mathbf{V}(t)$ evolving on $\mathcal{O}(k)$ and defined by the dynamical system (11.13). The rotation criterion is monotonically decreasing and the algorithm converges from any starting point to a stationary point. At a stationary point of $\mathcal{F}$ restricted to $\mathcal{O}(k)$, the Frobenius norm of the gradient after projection onto the plane tangent to $\mathcal{O}(k)$ at the current value of $\mathbf{V}$ is zero. The algorithm stops when the Frobenius norm of the gradient after projection is less than some prescribed precision, say $10^{-6}$.

Jennrich (2001) proposed an iterative scheme to keep the gradient flow 'nailed' to the manifold of orthogonal matrices (see also Bernaards and Jennrich, 2005). Jennrich (2004a) introduced a modification of the gradient projection algorithms of Jennrich (2001, 2002). The gradients are replaced by numerical approximations, that is, the algorithm of Jennrich (2004a) only requires the definition of the criterion. However, the use of numerical gradients generally needs more CPU time than using exact gradients. Once the optimal rotation matrix $\mathbf{T}$ and hence $\hat{\Xi}$ has been found, the IEFA mixing matrix is obtained by $\hat{\mathbf{M}} = \hat{\Lambda}\mathbf{T}$. Summarizing, the proposed IEFA method is as follows:

1. Set up the number of common factors, $k$, prescribed or estimated.

2. Estimate all EFA model unknowns by optimizing $||\mathbf{Z} - \mathbf{F}\Lambda^{\top} - \mathbf{U}\Psi||_F^2$.

3. Find an orthogonal matrix $\mathbf{T}$ that solves (11.5).

4. Calculate approximately independent factor scores as $\hat{\Xi} = \hat{\mathbf{F}}\mathbf{T}$.

5. Obtain the IEFA mixing matrix by $\hat{\mathbf{M}} = \hat{\Lambda}\mathbf{T}$.

## 11.2 Application to Thurstone's box problem

Developing analytical methods for factor rotation has a long history (Browne, 2001). It is motivated by both solving the rotational indeterminacy problem in EFA and facilitating the factors' interpretation. The aim for analytic rotation is to find loadings with 'simple structure' in an objective manner. Thurstone has set forth a number of general principles which, vaguely stated, say that a loading matrix with many small values and a small number of larger values is simpler than one with mostly intermediate values (Thurstone, 1947, p. 335; Yates, 1987, p. 34). Thurstone's 26-variable box problem (Thurstone, 1947) was notorious for being difficult to solve by any analytic rotation method. The problem is to find simple loadings which identify the dimensions of the boxes.

As in Jennrich and Trendafilov (2005), seven additional boxes, whose dimensions are given in Table 11.1, are added to the twenty boxes in Table 7.1 to form a set of boxes whose dimensions $x$, $y$ and $z$ are independent. This data set is in turn well-suited to be analyzed by IEFA.

| | 21 | 22 | 23 | 24 | 25 | 26 | 27 |
|---|---|---|---|---|---|---|---|
| $x$ | 3 | 3 | 3 | 3 | 4 | 5 | 5 |
| $y$ | 4 | 4 | 4 | 2 | 2 | 3 | 2 |
| $z$ | 1 | 2 | 3 | 3 | 3 | 1 | 3 |

Table 11.1: Dimensions $x$ (length), $y$ (width) and $z$ (height) of the seven additional boxes for Thurstone's 26-variable box data.

Assume that the three dimensions of the extended set of 27 boxes constitute the column vectors $\xi_1$, $\xi_2$ and $\xi_3$ of the factor score matrix $\Xi$. As $\xi_1$, $\xi_2$ and $\xi_3$ are independent, the box problem seems quite appropriate to be attacked by IEFA instead of simple structure analytic rotation.

As in Section 7.3.1, the twenty-six functions of Thurstone (1947) are used to generate the observed variables. The columns of the resulting 27 × 26 data matrix are then mean-centered and scaled to unit norm to obtain a data matrix $\mathbf{Z}$.

The IEFA method described in the previous Section was applied to get $\hat{\Xi} = \hat{\mathbf{F}}\mathbf{T}$. The columns $\hat{\xi}_1$, $\hat{\xi}_2$, and $\hat{\xi}_3$ of $\hat{\Xi}$ are the rotated factor scores and estimates of the standardized form of the three dimensions $\xi_1$, $\xi_2$ and $\xi_3$ used to generate the mixtures. According to Table 11.2, $\hat{\xi}_1$, $\hat{\xi}_2$ and $\hat{\xi}_3$ are quite independent. The off-diagonal elements of the correlation matrix for the element-wise squares of $\hat{\xi}_1$, $\hat{\xi}_2$ and $\hat{\xi}_3$ are all very small (below $3 \times 10^{-5}$). The value of the rotation criterion (11.4) at the minimum is $3.77 \times 10^{-10}$. Figure 11.1 displays that IEFA has quite accurately recovered the dimensions for each of the 27 boxes.

| | | |
|---|---|---|
| .48441 | -.00001 | .00001 |
| -.00003 | .48847 | -.00000 |
| .00002 | -.00001 | .49531 |

Table 11.2: Covariances (diagonal and above) and correlations (below diagonal) between the element-wise squares of $\hat{\xi}_1$, $\hat{\xi}_2$ and $\hat{\xi}_3$ for the 27 × 26 box data.

In ICA, the performance of an algorithm is often quantified in terms of the following error measure (Amari, Cichocki, and Yang, 1996):

$$\sum_{i=1}^{k}\left(\sum_{j=1}^{k}\frac{|p_{ij}|}{\max_l |p_{il}|}-1\right)+\sum_{j=1}^{k}\left(\sum_{i=1}^{k}\frac{|p_{ij}|}{\max_l |p_{lj}|}-1\right) , \qquad (11.14)$$

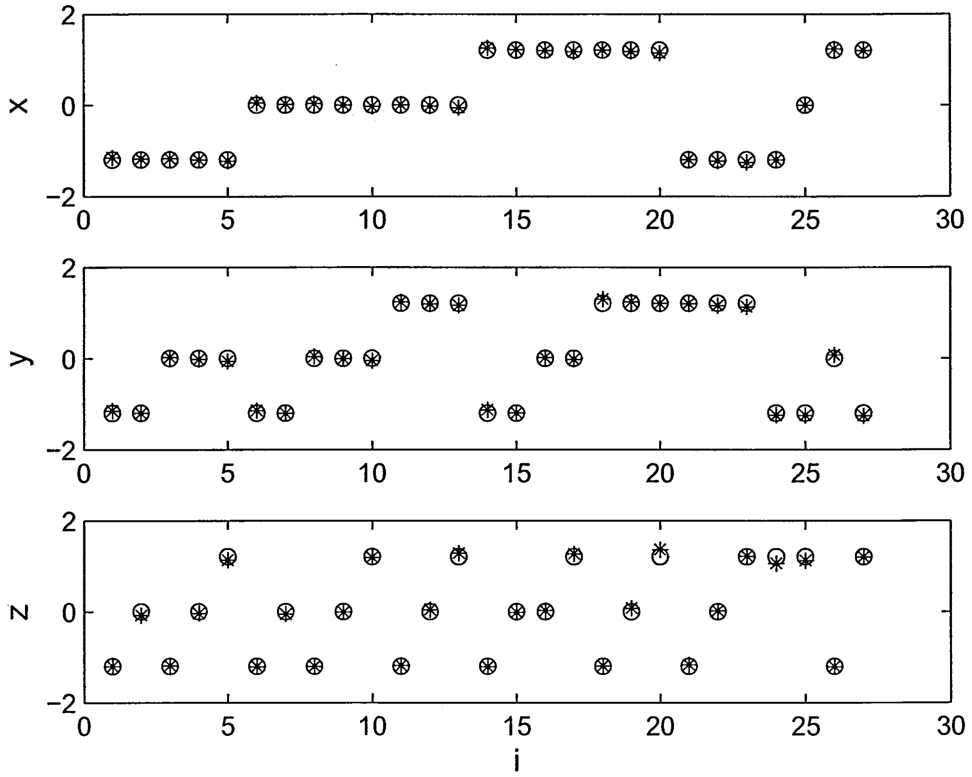

Figure 11.1: Standardized box dimensions 'o' and their estimates '*' for each dimension $x$ (length) (upper panel), $y$ (width) (middle panel) and $z$ (height) (lower panel) and each box $i$ $(i = 1, \ldots, 27)$ for the $27 \times 26$ box problem.

where the $p_{ij}$ $(i, j = 1, \ldots, k)$ are the elements of the performance matrix $\mathbf{P} = \mathbf{BM}$ and $\mathbf{B} = (\hat{\mathbf{M}}^{\mathrm{T}}\hat{\mathbf{M}})^{-1}\hat{\mathbf{M}}^{\mathrm{T}} \in \mathbb{R}^{k \times p}$ denotes the Moore-Penrose inverse of the estimated mixing matrix $\hat{\mathbf{M}}$. The index (11.14) evaluates unbiasedness of the estimation of the mixing matrix $\mathbf{M}$ apart from permutation and sign ambiguities. In a noise-less setting the estimate of $\Xi$ is simply obtained by $\hat{\Xi} = \mathbf{XB}$. Hence, unbiased estimation of the mixing matrix is equivalent to unbiased estimation of $\Xi$ and (11.14) can be considered as an appropriate measure of evaluating source separation. As discussed in Section 3.2.1 this simple relationship does not hold any more in noisy ICA. The following error measure is proposed instead:

$$E = \frac{||\Xi - \hat{\Xi}||_F}{||\Xi||_F} \ , \tag{11.15}$$

which evaluates the performance of IEFA by means of the normalized Frobenius norm of the difference between the true sources and the recovered sources. The relative error measure $E$ indicates good performance in accurateness of separation by low values and vanishes if the dimensions of the boxes are recovered perfectly. For the 27 × 26 box problem $E = .0473$ confirming that IEFA has recovered the dimensions of the boxes very well.

The IEFA loading matrix is obtained by $\hat{\mathbf{M}} = \hat{\mathbf{\Lambda}}\mathbf{T}$. The simple structure achieved in $\hat{\mathbf{M}}$ can be compared to the ones obtained by the more sophisticated rotation-to-simplicity methods originated in psychometrics such as Varimax, Minimum entropy, Quartimin, and Geomin. Whereas Varimax and Minimum entropy are orthogonal rotation methods, Quartimin and Geomin use oblique rotations.

Recall from Section 10.2.2 that the Varimax criterion (Kaiser, 1958) can be derived by restricting the Crawford-Ferguson family of criteria (10.52) to orthogonal rotation and setting $\tau = 1/p$.

McCammon (1966) suggested an orthogonal rotation criterion based on the entropy function of information theory. The simplest entropy criterion (Jennrich, 2004b) is

$$\mathcal{F}(\mathbf{V}) = -\text{trace}\left[(\hat{\mathbf{\Lambda}}\mathbf{V} \odot \hat{\mathbf{\Lambda}}\mathbf{V})^\top \ln(\hat{\mathbf{\Lambda}}\mathbf{V} \odot \hat{\mathbf{\Lambda}}\mathbf{V})\right] \ . \tag{11.16}$$

Unlike orthogonal rotation, oblique rotation methods seek a non-orthogonal and non-singular rotation matrix $\mathbf{V} \in \mathbb{R}^{k \times k}$ with columns having unit length, such that the oblique rotated loadings $\hat{\mathbf{\Lambda}}(\mathbf{V}^\top)^{-1}$ optimize a particular criterion $\mathcal{F}(\mathbf{V})$. Oblique rotations give extra flexibility and often produce a better simple structure than orthogonal rotations. If the Crawford-Ferguson family is optimized subject to $\mathbf{V}$ being a non-singular matrix with columns of length one, then (10.52) with $\tau = 0$ yields the Quartimin criterion of Carroll (1953).

Finally, the following modified version of the Geomin criterion (Yates, 1987) proposed by Browne (2001) is considered:

$$\mathcal{F}(\mathbf{V}) = \mathbf{1}_p^\top \exp\left[\frac{1}{k}\left(\ln\left((\mathbf{\Lambda V} \odot \mathbf{\Lambda V}) + \epsilon\right)\right)\mathbf{1}_k\right] , \tag{11.17}$$

where $\epsilon$ is a small constant, say $\epsilon = .01$ (Browne, 2001), to eliminate problems arising from loadings equal to zero.

The continuous-time projected gradient approach was used to optimize all four rotation-to-simplicity criteria. The gradients for the Varimax, Minimum entropy, Quartimin and Geomin methods can be found in Bernaards and Jennrich (2005). The feasible set for orthogonal criteria is $\mathcal{O}(k)$ in (2.5) and for oblique criteria is $\mathcal{OB}(k)$ in (2.7).

The criteria considered might have multiple local minima. As in Browne (2001) and Jennrich (2004b), this is dealt with by arbitrarily defining the best rotation produced from 20 random starts for the initial rotation matrix to be the operational minimizer of the criterion under consideration. For the orthogonal rotation criteria orthogonal random starts are used (e.g., Browne, 2001). An orthogonal random start is a matrix that is randomly selected from the uniform distribution on the group of orthogonal matrices (Jennrich, 2004b). For the oblique rotation methods oblique random starts are used as advocated by Rozeboom (1991). An oblique random start is a matrix whose columns are independently generated and randomly selected from a unit sphere of appropriate dimension.

For Varimax, Minimum entropy and Quartimin (as well as for IEFA) an identity start and 20 random starts gave the same criterion value. Apparently for the 27 × 26 box data and those methods random starts are not required. Geomin failed to produce a global minimum when started from an identity matrix. It gave the minimum criterion value using 15 of the 20 random starts.

The IEFA loadings and the solutions of the Varimax, Minimum entropy, Quartimin and Geomin rotations of $\hat{\Lambda}$ are given in Table 11.3 and Table 11.4, respectively.

If one ignores all loadings with magnitude .05 or less in the IEFA loading matrix, the remaining loadings perfectly identify the box dimensions $x, y$ and $z$ used to generate the mixtures. The simple structure achieved by IEFA is nearly as good as the ones obtained by the best rotation-to-simplicity methods known for Thurstone's box problem, namely the Minimum entropy and Geomin criteria (Bernaards and Jennrich, 2005). With the latter two methods if one ignores all entries with magnitude .01 and .02 or less, respectively, the remaining loadings perfectly identify the dimensions that are used to generate the manifest variables. Both the Varimax and the Quartimin method fail to identify some of the corresponding box dimensions.

In Table 11.3 results for the noise-free ICA approach of Jennrich and Trendafilov (2005) are also given. These rotated loadings are also obtained by optimizing the rotation-to-independence criterion (11.4) but the initial sphered factor scores are derived by means of PCA instead of EFA. For this approach one has to ignore all loadings with magnitude .15 or less to identify the box dimensions.

One must look fairly hard at the aligned rotated loading matrices in Table 11.3 and Table 11.4 to form an opinion about the simple structure achieved. Two or more rotated loading matrices can be more easily compared using sorted absolute loadings (SAL) plots as advocated by Jennrich (2004b). The SAL plot does not need to produce the

| | IEFA | | | Varimax | | | Minimum entropy | | | Noise-free ICA | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Function | (orthogonal) | | | (orthogonal) | | | (orthogonal) | | | (orthogonal) | | |
| $x$ | 1.00 | .02 | .03 | .89 | -.45 | .04 | 1.00 | -.01 | .00 | .99 | .08 | .09 |
| $y$ | -.02 | 1.00 | .04 | .45 | .89 | -.09 | .01 | 1.00 | .00 | -.09 | .99 | .11 |
| $z$ | -.03 | -.04 | 1.00 | .00 | .10 | .99 | .00 | .00 | .99 | -.08 | -.12 | .99 |
| $xy$ | .58 | .81 | .04 | .89 | .44 | -.05 | .60 | .79 | -.01 | .52 | .83 | .14 |
| $xz$ | .42 | -.03 | .90 | .40 | -.12 | .90 | .45 | -.01 | .89 | .37 | -.07 | .92 |
| $yz$ | -.04 | .52 | .84 | .25 | .57 | .77 | .00 | .55 | .82 | -.12 | .44 | .87 |
| $x^2y$ | .79 | .58 | .04 | .97 | .14 | -.02 | .81 | .56 | -.01 | .75 | .62 | .14 |
| $xy^2$ | .35 | .92 | .04 | .74 | .64 | -.07 | .38 | .91 | -.01 | .29 | .93 | .14 |
| $x^2z$ | .65 | -.02 | .72 | .60 | -.24 | .73 | .67 | -.01 | .70 | .61 | -.03 | .76 |
| $xz^2$ | .26 | -.03 | .94 | .25 | -.04 | .94 | .28 | .00 | .94 | .20 | -.09 | .95 |
| $y^2z$ | -.04 | .72 | .64 | .34 | .72 | .54 | .00 | .75 | .61 | -.12 | .67 | .69 |
| $yz^2$ | -.03 | .32 | .92 | .16 | .41 | .87 | .00 | .36 | .91 | -.11 | .25 | .94 |
| $x/y$ | .58 | -.78 | -.01 | .14 | -.96 | .10 | .56 | -.80 | .00 | .64 | -.74 | -.03 |
| $y/x$ | -.61 | .76 | .02 | -.18 | .95 | -.09 | -.59 | .77 | .00 | -.67 | .72 | .03 |
| $x/z$ | .41 | .05 | -.86 | .35 | -.25 | -.85 | .39 | .00 | -.87 | .46 | .15 | -.83 |
| $z/x$ | -.47 | -.04 | .84 | -.40 | .29 | .83 | -.44 | .01 | .86 | -.52 | -.14 | .81 |
| $y/z$ | .02 | .52 | -.80 | .23 | .34 | -.85 | .01 | .48 | -.81 | .03 | .58 | -.76 |
| $z/y$ | .00 | -.59 | .75 | -.24 | -.42 | .82 | .01 | -.55 | .78 | .00 | -.65 | .71 |
| $2x + 2y$ | .69 | .72 | .05 | .95 | .31 | -.03 | .71 | .70 | .00 | .64 | .75 | .14 |
| $2x + 2z$ | .69 | -.02 | .72 | .63 | -.25 | .73 | .71 | .00 | .70 | .64 | -.03 | .76 |
| $2y + 2z$ | -.03 | .68 | .73 | .32 | .69 | .64 | .01 | .71 | .70 | -.12 | .61 | .78 |
| $\sqrt{x^2 + y^2}$ | .78 | .61 | .04 | .98 | .17 | -.02 | .80 | .59 | .00 | .73 | .65 | .14 |
| $\sqrt{x^2 + z^2}$ | .87 | .00 | .47 | .79 | -.36 | .48 | .89 | .00 | .44 | .84 | .02 | .52 |
| $\sqrt{y^2 + z^2}$ | -.03 | .79 | .58 | .38 | .77 | .47 | .01 | .82 | .54 | -.11 | .74 | .63 |
| $xyz$ | .34 | .48 | .78 | .56 | .35 | .72 | .37 | .50 | .76 | .26 | .44 | .84 |
| $\sqrt{x^2 + y^2 + z^2}$ | .71 | .55 | .41 | .91 | .19 | .35 | .74 | .55 | .36 | .65 | .56 | .49 |

Table 11.3: Orthogonally rotated loading matrices obtained by IEFA, Varimax, Minimum Entropy and noise-free ICA for the 27 × 26 box problem.

| | IEFA | | | Quartimin | | | Geomin ($\epsilon = .01$) | | |
|---|---|---|---|---|---|---|---|---|---|
| Function | (orthogonal) | | | (oblique) | | | (oblique) | | |
| $x$ | 1.00 | .02 | .03 | .89 | -.57 | -.06 | 1.00 | -.02 | -.02 |
| $y$ | -.02 | 1.00 | .04 | .49 | .83 | -.07 | -.02 | 1.00 | -.01 |
| $z$ | -.03 | -.04 | 1.00 | -.09 | .01 | 1.00 | -.01 | -.01 | 1.00 |
| $xy$ | .58 | .81 | .04 | .92 | .32 | -.09 | .58 | .78 | -.02 |
| $xz$ | .42 | -.03 | .90 | .32 | -.24 | .87 | .44 | -.02 | .88 |
| $yz$ | -.04 | .52 | .84 | .20 | .47 | .79 | -.02 | .54 | .82 |
| $x^2y$ | .79 | .58 | .04 | .99 | .01 | -.09 | .79 | .55 | -.02 |
| $xy^2$ | .35 | .92 | .04 | .78 | .54 | -.09 | .36 | .90 | -.02 |
| $x^2z$ | .65 | -.02 | .72 | .53 | -.37 | .67 | .66 | -.02 | .69 |
| $xz^2$ | .26 | -.03 | .94 | .17 | -.15 | .93 | .27 | -.01 | .93 |
| $y^2z$ | -.04 | .72 | .64 | .31 | .63 | .57 | -.02 | .74 | .60 |
| $yz^2$ | -.03 | .32 | .92 | .10 | .31 | .89 | -.02 | .35 | .90 |
| $x/y$ | .58 | -.78 | -.01 | .11 | -.98 | .03 | .58 | -.81 | .00 |
| $y/x$ | -.61 | .76 | .02 | -.15 | .98 | -.02 | -.61 | .78 | .01 |
| $x/z$ | .41 | .05 | -.86 | .43 | -.23 | -.90 | .40 | .01 | -.88 |
| $z/x$ | -.47 | -.04 | .84 | -.47 | .27 | .89 | -.46 | .01 | .87 |
| $y/z$ | .02 | .52 | -.80 | .32 | .38 | -.86 | .01 | .49 | -.82 |
| $z/y$ | .00 | -.59 | .75 | -.33 | -.45 | .82 | .01 | -.56 | .78 |
| $2x + 2y$ | .69 | .72 | .05 | .98 | .18 | -.09 | .70 | .69 | -.02 |
| $2x + 2z$ | .69 | -.02 | .72 | .57 | -.39 | .67 | .70 | -.02 | .69 |
| $2y + 2z$ | -.03 | .68 | .73 | .29 | .60 | .66 | -.02 | .70 | .70 |
| $\sqrt{x^2 + y^2}$ | .78 | .61 | .04 | 1.00 | .04 | -.09 | .78 | .58 | -.02 |
| $\sqrt{x^2 + z^2}$ | .87 | .00 | .47 | .75 | -.49 | .40 | .88 | -.02 | .43 |
| $\sqrt{y^2 + z^2}$ | -.03 | .79 | .58 | .36 | .68 | .49 | -.01 | .81 | .54 |
| $xyz$ | .34 | .48 | .78 | .51 | .22 | .71 | .36 | .49 | .74 |
| $\sqrt{x^2 + y^2 + z^2}$ | .71 | .55 | .41 | .89 | .05 | .29 | .72 | .53 | .34 |

Table 11.4: Obliquely rotated loading matrices obtained by Quartimin and Geomin compared to IEFA for the 27 × 26 box problem.

column permutations and sign changes often required to align loadings matrices for comparison. Let $m = pk$ and let $|\lambda_1| \leq |\lambda_2| \leq \cdots \leq |\lambda_m|$ denote the absolute values of the rotated loadings sorted in increasing order. The SAL plot is a plot of $|\lambda_j|$ against $j$ for $j = 1, \ldots, m$. Generally, the greater the number of small loadings, the simpler the loading matrix.

Figure 11.2 (i) and (ii) are SAL plots of the IEFA rotated loadings compared to the orthogonal criteria Varimax and Minimum entropy as well as to the oblique criteria Quartimin and Geomin, respectively. The plots reveal that Minimum Entropy, Geomin and IEFA all have the smallest 27 loadings very close to zero and seem to encourage small loadings much more than Varimax and Quartimin. The smallest 27 loadings for IEFA are only slightly larger than the ones of Minimum entropy and Geomin.

Clearly, IEFA outperforms Quartimin and Varimax in terms of simplicity achieved. Results for the noise-free ICA approach of Jennrich and Trendafilov (2005) are also shown in Figure 11.2 (i). It is worth noting that IEFA is able to obtain a better simple structure than the noise-free approach. Since both methods optimize the same criterion, results suggest that it is the initial EFA decomposition taking unique factors into account that causes this simpler structure. Note that while easier to compare than a set of aligned loading matrices, the SAL plot contains less information. Only the distributions of the absolute loadings are displayed and not the locations and the signs within a loading matrix.

It is worth investigating whether IEFA recovers the dimensions of the boxes accurately and produces simple loadings if the factors are correlated rather than independent. To discover this, the previous analysis was carried out using the original set of Thurstone's twenty boxes from Table 7.1. Assume that the three dimensions constitute the column

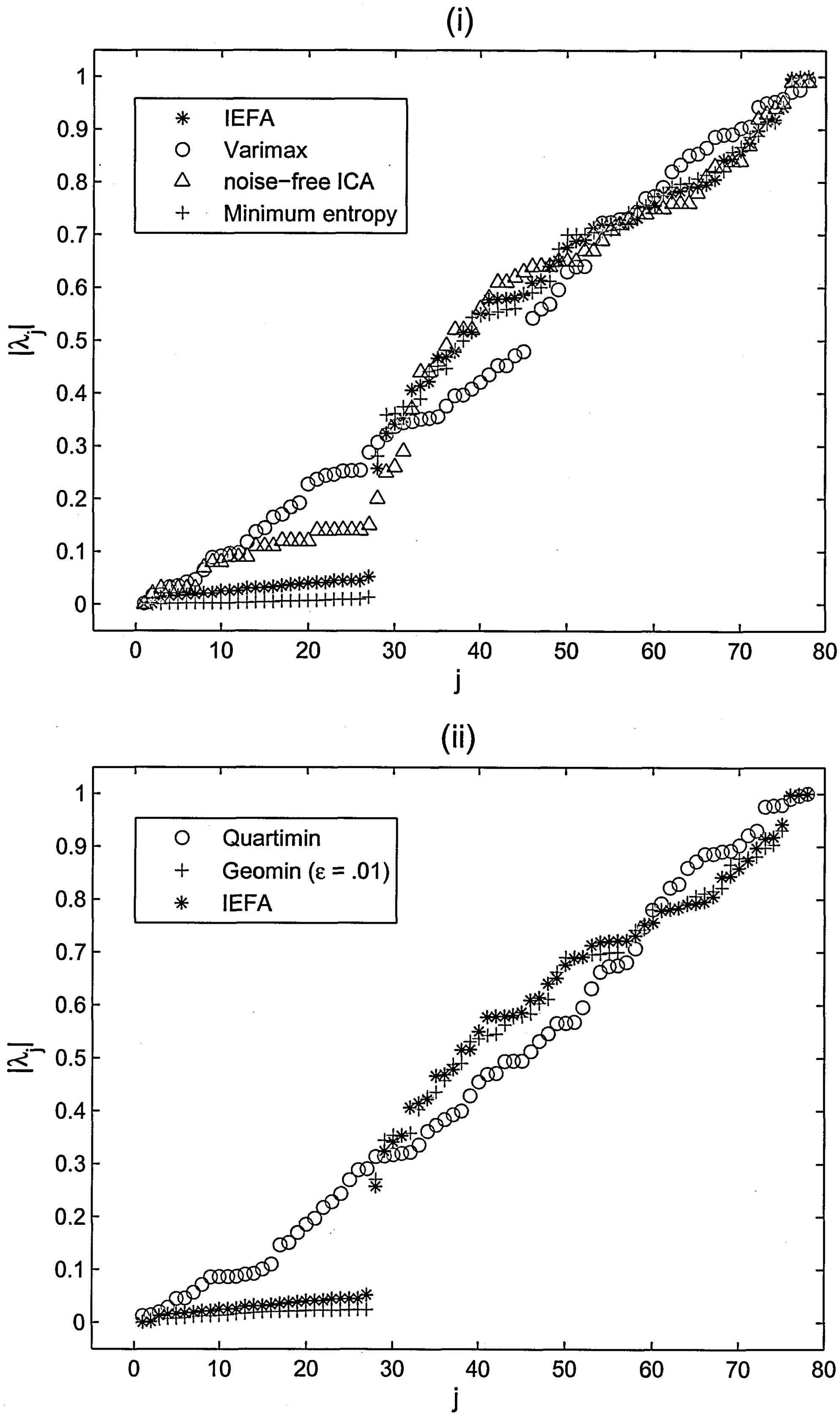

Figure 11.2: SAL plots of IEFA, noise-free ICA as well as orthogonal (i) and oblique (ii) rotation-to-simplicity criteria applied to the 27 × 26 box problem.

vectors $\xi_1$, $\xi_2$ and $\xi_3$ of the factor score matrix $\Xi$. The columns have the intercorrelations .25 between $\xi_1$ and $\xi_2$, .10 between $\xi_1$ and $\xi_3$, and .25 between $\xi_2$ and $\xi_3$, respectively. Thus, the factors are dependent. The off-diagonal elements of the correlation matrix for the element-wise squares of $\hat{\xi}_1$, $\hat{\xi}_2$ and $\hat{\xi}_3$ in Table 11.5 indicate that the recovered factors are also not independent. Nevertheless, Figure 11.3 displays that for the 20 × 26 box problem IEFA recovers the dimensions of the boxes fairly accurately. This is confirmed by the low value .1720 for the error measure $E$.

| | | |
|---|---|---|
| .73159 | -.10079 | -.04575 |
| -.14018 | .69673 | -.04569 |
| -.06301 | -.06450 | .72041 |

Table 11.5: Covariances (diagonal and above) and correlations (below diagonal) between the element-wise squares of $\hat{\xi}_1$, $\hat{\xi}_2$ and $\hat{\xi}_3$ for the 20 × 26 box data.

The simplicity of the rotated loadings can be assessed from Table 11.6. For IEFA, Minimum entropy and Geomin the factors are clearly related in an appropriate way to the box dimensions. The smallest absolute loadings can be associated with the missing dimensions in the formula for the corresponding variable. This is not the case for Quartimin, which fails to identify the dimensions $x$, $y$ and $z$ used to generate some of the mixtures. With dependent factors, the oblique Geomin criterion performs best. If one ignores all loadings with magnitude .04 or less, the remaining loadings perfectly identify the box dimensions used to generate the observed variables. For IEFA and Minimum entropy, this can be done by ignoring loadings with magnitudes of .20 and .31 or less, respectively. Hence, the simple structure achieved by IEFA is even better than the one obtained by the best orthogonal rotation-to simplicity method known for

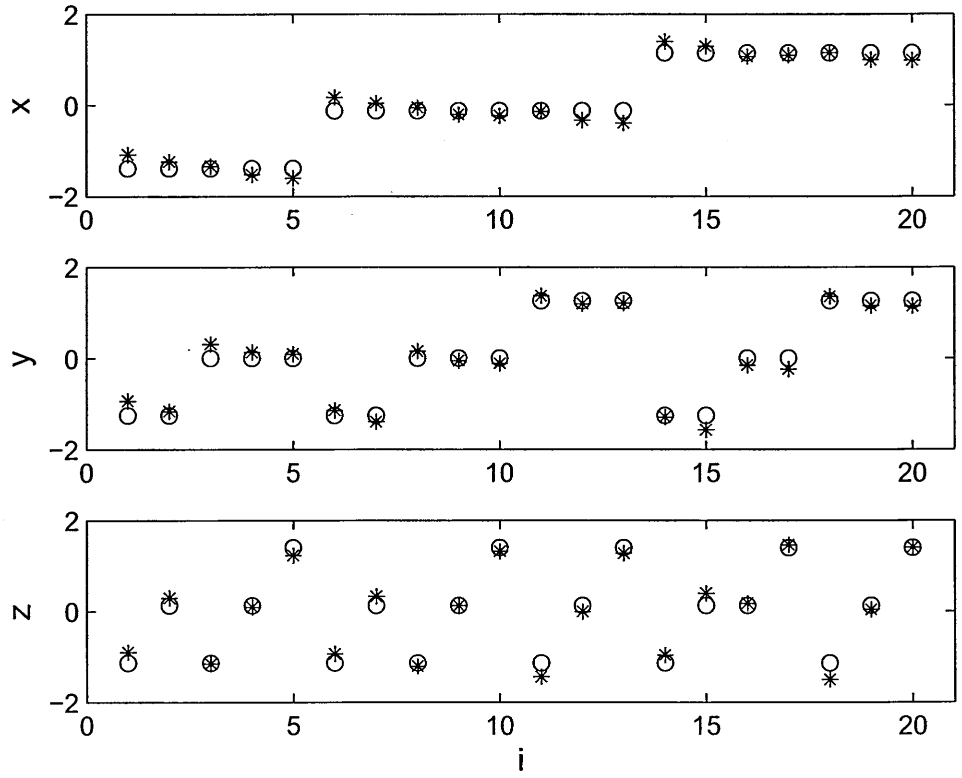

Figure 11.3: Standardized box dimensions 'o' and their estimates '*' for each dimension $x$ (length) (upper panel), $y$ (width) (middle panel) and $z$ (height) (lower panel) and each box $i$ ($i = 1, \ldots, 20$) for the $20 \times 26$ box problem.

solving Thurstone's box problem, namely Minimum entropy.

| | IEFA | | | Quartimin | | | Minimum entropy | | | Geomin ($\epsilon = .01$) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Function | (orthogonal) | | | (oblique) | | | (orthogonal) | | | (oblique) | | |
| $x$ | .99 | .15 | .05 | .97 | -.48 | -.09 | .96 | .27 | .04 | 1.00 | -.02 | .00 |
| $y$ | .10 | .99 | .10 | .58 | .74 | -.05 | -.02 | 1.00 | .05 | .00 | 1.00 | -.01 |
| $z$ | .02 | .15 | .99 | .05 | .08 | .97 | .01 | .20 | .98 | -.03 | .01 | 1.00 |
| $xy$ | .56 | .82 | .10 | .91 | .32 | -.09 | .46 | .88 | .05 | .49 | .75 | -.01 |
| $xz$ | .42 | .19 | .88 | .43 | -.12 | .81 | .39 | .28 | .87 | .38 | .00 | .87 |
| $yz$ | .07 | .54 | .83 | .29 | .38 | .75 | .01 | .59 | .80 | -.01 | .43 | .79 |
| $x^2y$ | .73 | .66 | .09 | .99 | .09 | -.09 | .65 | .74 | .05 | .69 | .55 | -.01 |
| $xy^2$ | .39 | .90 | .08 | .80 | .50 | -.10 | .28 | .95 | .03 | .31 | .87 | -.04 |
| $x^2z$ | .62 | .19 | .73 | .63 | -.25 | .64 | .60 | .29 | .72 | .60 | -.02 | .71 |
| $xz^2$ | .26 | .20 | .92 | .29 | -.01 | .88 | .23 | .28 | .92 | .20 | .03 | .93 |
| $y^2z$ | .11 | .70 | .67 | .41 | .49 | .57 | .02 | .74 | .64 | .01 | .62 | .61 |
| $yz^2$ | .05 | .40 | .89 | .20 | .26 | .85 | .01 | .44 | .88 | -.02 | .27 | .88 |
| $x/y$ | .57 | -.79 | -.09 | .14 | -.99 | -.06 | .66 | -.72 | -.06 | .67 | -.91 | -.03 |
| $y/x$ | -.59 | .78 | .06 | -.16 | .99 | .03 | -.68 | .71 | .03 | -.69 | .91 | .00 |
| $x/z$ | .36 | -.02 | -.88 | .36 | -.22 | -.93 | .36 | -.02 | -.88 | .41 | .04 | -.91 |
| $z/x$ | -.42 | .06 | .87 | -.39 | .28 | .92 | -.42 | .05 | .87 | -.47 | .01 | .90 |
| $y/z$ | .02 | .48 | -.84 | .29 | .40 | -.91 | -.04 | .44 | -.85 | .01 | .62 | -.91 |
| $z/y$ | -.02 | -.46 | .84 | -.28 | -.39 | .92 | .04 | -.42 | .86 | .00 | -.60 | .92 |
| $2x + 2y$ | .68 | .72 | .10 | .98 | .16 | -.09 | .60 | .80 | .06 | .64 | .62 | .00 |
| $2x + 2z$ | .68 | .20 | .70 | .69 | -.27 | .59 | .66 | .31 | .68 | .66 | -.01 | .67 |
| $2y + 2z$ | .08 | .72 | .68 | .40 | .52 | .58 | .00 | .76 | .65 | -.01 | .64 | .62 |
| $\sqrt{x^2 + y^2}$ | .78 | .61 | .09 | 1.00 | .02 | -.10 | .70 | .70 | .05 | .75 | .49 | -.01 |
| $\sqrt{x^2 + z^2}$ | .86 | .20 | .45 | .86 | -.38 | .32 | .83 | .32 | .44 | .85 | .00 | .41 |
| $\sqrt{y^2 + z^2}$ | .09 | .84 | .52 | .47 | .61 | .40 | -.01 | .87 | .48 | -.01 | .79 | .44 |
| $xyz$ | .35 | .54 | .75 | .55 | .20 | .64 | .29 | .61 | .72 | .28 | .39 | .70 |
| $\sqrt{x^2 + y^2 + z^2}$ | .70 | .60 | .37 | .92 | .05 | .20 | .62 | .70 | .33 | .65 | .46 | .29 |

Table 11.6: Rotated loading matrices obtained by IEFA, Quartimin, Minimum Entropy and Geomin for the 20 × 26 box problem.

# Chapter 12

# Implementation of ICA in Three-mode Factor Analysis

Statistical methods like EFA or ICA are used to analyze two-way data matrices. Three-way data emerge, for instance, in multivariate longitudinal studies where $I$ subjects are measured on $J$ variables on $K$ occasions. The three ways pertain to three different sets of entities named 'modes' of the data. Three-mode models are explicitly designed to handle such data. For a comprehensive survey of three-mode models and recent advances the reader is referred to Acar and Yener (2007), Kroonenberg (2008) and the references therein.

The most prominent three-mode models are the three-mode factor analysis (Tucker3) model introduced by Tucker (1966) and the PARAFAC/CANDECOMP (CP) model introduced independently by Harshman (1970) and Carroll and Chang (1970).

Both models are supposed to be extensions of bilinear factor analysis to trilinear data (Harshman, 1970; Tucker, 1966). Since neither of these two techniques estimate unique factors or unique variances, both models are generally referred to in the literature as three-mode component models (Kroonenberg, 1983).

Beckmann and Smith (2005) and De Vos, De Lathauwer, and Van Huffel (2007) combined the CP model and ICA. These approaches are reviewed in the next Section. In

Section 12.2, an alternative approach to ICA for analyzing three-way data is considered combining ICA and the Tucker3 model. The performance of the proposed approach is evaluated by numerical experiments in Section 12.3.

## 12.1 Combining ICA and the CP model

Let a three-dimensional array $\mathcal{X} \in \mathbb{R}^{I\times J\times K}$ be defined as the collection of elements $\{x_{ijk}|\, i,\ldots,I; j=1,\ldots,J; k=1,\ldots,K\}$ which are placed in $\mathcal{X}$ such that the indices $i, j$, and $k$ run along the vertical, horizontal, and depth axes, respectively. A three-way data array of order $I \times J \times K$ is sometimes called a 3rd order 'tensor' in $\mathbb{R}^{I\times J\times K}$ (e.g., Acar and Yener, 2007). Each of the three sets of indices $i, j$ and $k$ designates one mode of the data. Then, the three-mode CP model with $R$ components is defined as (Carroll and Chang, 1970; Harshman, 1970):

$$x_{ijk} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr} + e_{ijk} \quad (i=1,\ldots,I; j=1,\ldots,J; k=1,\ldots,K) \ , \tag{12.1}$$

where $a_{ir}$, $b_{jr}$ and $c_{kr}$ denote the elements of the component matrices $\mathbf{A} \in \mathbb{R}^{I\times R}$, $\mathbf{B} \in \mathbb{R}^{J\times R}$ and $\mathbf{C} \in \mathbb{R}^{K\times R}$, respectively, and $e_{ijk}$ are the elements of the three-way error array $\mathcal{E} \in \mathbb{R}^{I\times J\times K}$.

Let $\mathbf{a}_r \in \mathbb{R}^{I\times 1}$, $\mathbf{b}_r \in \mathbb{R}^{J\times 1}$, and $\mathbf{c}_r \in \mathbb{R}^{K\times 1}$ denote the $r$-th columns of $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$, respectively. Then, the CP model (12.1) can be expressed in a concise form as (Kiers, 2000; Kroonenberg, 2008):

$$\begin{aligned} \mathcal{X} &= \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r + \mathcal{E} & (12.2)\\ \Leftrightarrow \mathrm{vec}(\mathcal{X}) &= (\mathbf{C}| \otimes |\mathbf{B}| \otimes |\mathbf{A})\mathbf{1}_R + \mathrm{vec}(\mathcal{E}) \ , & (12.3) \end{aligned}$$

where the symbol $\circ$ denotes the vector outer product, vec is the column stacking operator and $| \otimes |$ denotes the column-wise Kronecker or Khatri-Rao matrix product.

The expressions (12.2) and (12.3) display the symmetry of the CP model. The term $(\mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r)$ in (12.2) is called a rank-1 array. The rank of the three-way array $\mathcal{X}$ is defined as the minimum number of rank-1 arrays sufficient to fully decompose $\mathcal{X}$ additively (Kiers, 2000).

The aim of the CP model is to find matrices $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$ which minimize the sum of squares of the elements of the error array $\mathcal{E}$. In this sense, the CP model tries to find the best rank-$R$ approximation to $\mathcal{X}$. This can be done by means of an ALS algorithm in which each component matrix is sequentially optimized, keeping the other two component matrices fixed. For an overview and comparison of several algorithms for fitting the CP model, the reader is referred to Tomasi and Bro (2006).

The most attractive feature of the CP model is its uniqueness under the following (mild) sufficient condition (Kruskal, 1977):

$$\mathrm{rank}_k(\mathbf{A}) + \mathrm{rank}_k(\mathbf{B}) + \mathrm{rank}_k(\mathbf{C}) \geq 2R + 2 \ , \qquad (12.4)$$

where $\mathrm{rank}_k(\cdot)$ denotes the Kruskal rank or $k$-rank of a matrix. The $k$-rank of a matrix is the maximal number $r$ such that any set of $r$ columns of the matrix is linearly independent. By fixing the columns of two of the three component matrices to unit length, a CP solution for $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$ is unique up to sign and permutation ambiguities in the component matrices if (12.4) holds.

Beckmann and Smith (2005) introduced the idea of combining the CP model and ICA for analyzing three-way data. Their method is a three-way extension of the noisy ICA model (3.13) with the noise covariance matrix being proportional to the identity matrix. The noisy ICA model with homoscedastic noise variance is called a probabilistic ICA (pICA) model (Penny, Roberts, and Everson, 2001). Beckmann and Smith (2005) named their method 'tensor pICA'. In tensor pICA, it is assumed that $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$ obeys the structure of a CP model.

It will be convenient to write the CP model in matrix form. Let $\mathbf{X}_A, \mathbf{E}_A \in \mathbb{R}^{I \times JK}$ be unfolded matrices formed by the $K$ frontal 'slices' of $\mathcal{X}$ and $\mathcal{E}$, respectively, and arranged next to each other. Then, (12.3) can be rewritten as:

$$\mathbf{X}_A = \mathbf{A}(\mathbf{C}| \otimes |\mathbf{B})^{\top} + \mathbf{E}_A \ . \tag{12.5}$$

Since the CP model treats the parameter matrices in a symmetric way, matrix equations similar to (12.5) for $\mathbf{X}_B$ or $\mathbf{X}_C$ can be formulated if $\mathcal{X}$ is sliced into $I$ horizontal or $J$ lateral slices, respectively. Assume that the components in the first mode are independent, so that the $I \times R$ matrix $\mathbf{A}$ contains the values for the independent components and the mixing matrix $\mathbf{M} \in \mathbb{R}^{JK \times R}$ equals $\mathbf{C}| \otimes |\mathbf{B}$. Estimation of the tensor pICA model is done by means of the following iterative algorithm (Beckmann and Smith, 2005):

1. Ignore the structure of the mixing matrix in (12.5). Decompose the data $\mathbf{X}_A \approx \mathbf{A}\mathbf{M}^{\top}$ into a compound mixing matrix $\mathbf{M}$ and an associated matrix $\mathbf{A}$ using an approach for estimating the two-way pICA model (e.g., Beckmann and Smith, 2004; Stegeman, 2007).

2. Decompose $\mathbf{M}$ such that $\mathbf{M} \approx \mathbf{C}| \otimes |\mathbf{B}$. Map each column $r$ $(r = 1, \ldots, R)$ of $\mathbf{M}$ into a $J \times K$ matrix $\mathbf{M}_r$ which has rank 1 according to model (12.5). The matrix $\mathbf{M}_r$ contains $K$ scaled repetitions of a single column of $\mathbf{B}$. Perform an SVD on $\mathbf{M}_r$ which leads to the best rank-1 approximation of $\mathbf{M}_r$. Compute estimates of the $r$-th columns of $\mathbf{C}$ and $\mathbf{B}$ which are given by the dominant left and right singular vector of $\mathbf{M}_r$, respectively.

3. Go to step 1 if two successive estimates for $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ are not sufficiently alike and use $(\mathbf{C}| \otimes |\mathbf{B})$ to perform another iteration in the decomposition of $\mathbf{X}_A$; else consider the algorithm converged.

Note that in tensor pICA the three-way structure is imposed after the independent components are obtained. Instead, De Vos, De Lathauwer, and Van Huffel (2007) introduced 'ICA-CP' which imposes the CP structure during the ICA computation. As was mentioned in Chapter 10, the standard noisy ICA problem for two-way data can be solved by diagonalizing the fourth-order cumulant tensor of the observed data $\mathbf{X}$ in (10.37). Using the fact that all higher-order cumulants of the independent factors are diagonal tensors and that all higher-order cumulants greater than two vanish for Gaussian distributed noise, the fourth-order cumulant tensor of $\mathbf{X}$ in (10.37) has the following CP decomposition of rank $k$ (De Lathauwer, De Moor, and Vandewalle, 2000):

$$\mathcal{C}_{\mathbf{X}}^{(4)} = \sum_{r=1}^{k} \kappa_4(r)\, \mathbf{m}_r \circ \mathbf{m}_r \circ \mathbf{m}_r \circ \mathbf{m}_r \ , \tag{12.6}$$

where $\kappa_4(r)$ corresponds to the fourth-order cumulant of the $r$-th source or common factor and $\mathbf{m}_r$ denotes the $r$-th column of the mixing matrix $\mathbf{M}$.

For data sets having a three-way structure according to model (12.5), De Vos, De Lathauwer, and Van Huffel (2007) proposed to solve the ICA problem by diagonalizing the fourth-order cumulant of $\mathbf{X}_A$ in (12.5). With a mixing matrix $\mathbf{M} = (\mathbf{C}| \otimes |\mathbf{B})$, this fourth-order cumulant can be expressed as an eighth-order tensor of rank $R$ with the following CP structure:

$$\mathcal{C}_{\mathbf{X}_A}^{(8)} = \sum_{r=1}^{R} \kappa_4(r) \mathbf{c}_r \circ \mathbf{b}_r \circ \mathbf{c}_r \circ \mathbf{b}_r \circ \mathbf{c}_r \circ \mathbf{b}_r \circ \mathbf{c}_r \circ \mathbf{b}_r \ . \tag{12.7}$$

For the computation of the CP decomposition (12.7), De Vos, De Lathauwer, and Van Huffel (2007) proposed a simultaneous generalized Schur decomposition (Golub and Van Loan, 1996) of a set of matrices which is considered in De Lathauwer, De Moor, and Vandewalle (2004).

The computation of the CP decomposition in De Vos, De Lathauwer, and Van Huffel (2007) requires that $R \leq \min\{I, J\}$. Once estimates for **C** and **B** and hence **M** have been obtained, the independent sources **A** can be estimated from equation (12.5).

Both tensor pICA and ICA-CP impose a CP structure for the three-way array $\mathcal{X}$ to implement the ICA in one mode of the data. In the next Section, an alternative approach to ICA of three-way data is considered combining ICA and the Tucker3 model.

## 12.2 Combining ICA and the Tucker3 model

### 12.2.1 Tucker3 model

The Tucker3 model (Tucker, 1966) factorizes $\mathcal{X} = \{x_{ijk}\} \in \mathbb{R}^{I\times J\times K}$ such that for $i = 1,\ldots,I; j = 1,\ldots,J; k = 1,\ldots,K$:

$$x_{ijk} = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} a_{ip}b_{jq}c_{kr}g_{pqr} + e_{ijk} \quad , \tag{12.8}$$

where as before $a_{ip}$, $b_{jq}$, and $c_{kr}$ denote the elements of the component matrices $\mathbf{A} \in \mathbb{R}^{I\times P}$, $\mathbf{B} \in \mathbb{R}^{J\times Q}$ and $\mathbf{C} \in \mathbb{R}^{K\times R}$, respectively. The $g_{pqr}$ are the elements of the three-way core array $\mathcal{G} \in \mathbb{R}^{P\times Q\times R}$ which describe the interactions between the components in $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$. The core $\mathcal{G}$ can be referred to as containing the weights of all possible triads. That is, the core elements represent the importance of the respective factor combinations. The largest squared elements of the core will indicate what the most important factors are in the model of $\mathcal{X}$.

In Tucker (1966) the component matrices **A**, **B** and **C** are not subject to almost any constraints; they are required to be simply full column rank matrices. Both for computational and uniqueness purposes, Kroonenberg and De Leeuw (1980) describe the Tucker3 model in terms of column-wise orthonormal matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$.

The Tucker3 model can be expressed in a concise form as follows (Kiers, 2000; Kroonenberg, 2008):

$$\begin{aligned} \mathcal{X} &= \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} g_{pqr}\left(\mathbf{a}_p \circ \mathbf{b}_q \circ \mathbf{c}_r\right) + \mathcal{E} && (12.9) \\ \Leftrightarrow \mathrm{vec}(\mathcal{X}) &= (\mathbf{C}\otimes\mathbf{B}\otimes\mathbf{A})\mathrm{vec}(\mathcal{G}) + \mathrm{vec}(\mathcal{E}) \ , && (12.10) \end{aligned}$$

where $\otimes$ denotes the Kronecker matrix product. Expression (12.10) gives insight into the role of the elements of the core as regression weights for the columns of $\mathbf{C}\otimes\mathbf{B}\otimes\mathbf{A}$. The Tucker3 model in unfolded matrix form is:

$$\mathbf{X}_A = \mathbf{A}\mathbf{G}_A(\mathbf{C}\otimes\mathbf{B})^{\mathsf{T}} + \mathbf{E}_A \ , \qquad (12.11)$$

where the unfolded core $\mathbf{G}_A \in \mathbb{R}^{P\times QR}$ is formed by the frontal slices of $\mathcal{G}$. Because of the symmetry of the Tucker3 model displayed by the expressions (12.9) and (12.10), (12.11) can also be formulated in terms of the two remaining unfoldings $\mathbf{X}_B$ and $\mathbf{X}_C$. In contrast to the Tucker3 model, the Tucker2 model leaves one mode in $\mathcal{X}$ uncompressed (Kroonenberg and De Leeuw, 1980). Hence, the Tucker2 model uses only two component matrices, but still a full core. Consequently, unlike CP and Tucker3, the Tucker2 model is not symmetric. Leaving the third mode unreduced, the Tucker2 model can be formulated as

$$\mathbf{X}_A = \mathbf{A}\bar{\mathbf{G}}_A(\mathbf{I}_K\otimes\mathbf{B})^{\mathsf{T}} + \mathbf{E}_A \ , \qquad (12.12)$$

where $\bar{\mathbf{G}}_A \in \mathbb{R}^{P\times QK}$ denotes the 'extended core matrix'. Note that $\bar{\mathbf{G}}_A$ has full dimensionality $K$ in the uncompressed mode. The Tucker2 model does exploit fully the three-way structure of the data. An application that requires this feature is the multivariate analysis of time series data, where in general no useful meaning can be attached to the components in the time mode. In this thesis only the Tucker3 model, in which components are computed for all three modes will be considered.

The hierarchy between the Tucker3 and the CP model can be revealed by using for the unfolded CP model (12.5) a notation equivalent to (12.11), namely

$$\mathbf{X}_A = \mathbf{A}\mathbf{H}(\mathbf{C} \otimes \mathbf{B})^\top + \mathbf{E}_A \ , \tag{12.13}$$

where the matrix $\mathbf{H}$ is the $R \times R^2$ unfolded version of a unit superdiagonal array $\mathcal{H}$, that is, an array with $h_{pqr} = 1$ if $p = q = r$, and $h_{pqr} = 0$ otherwise.

Equation (12.13) shows that the CP model is a constrained version of Tucker3 in which all cross-relations (multi-collinearities) between the components in different modes are eliminated. In other words, the CP model assumes that the components in different modes only interact factor-wise. Moreover, as a consequence of a superdiagonal core, an equal number of components are extracted in each mode.

Hence, the CP model is considerably more restrictive than the Tucker3 model. However, if this restriction is tenable, it implies that the CP has a unique solution under a fairly general condition and the components found are to be interpreted without recourse to rotation. By contrast, the Tucker3 model allows for extraction of different numbers of factors in each of the three modes and any factor in a certain mode is allowed to interact with any factor in the other two modes. The Tucker3 model has no unique solutions. This is due to the full-core array structure $\mathcal{G}$. Indeed, for any non-singular square matrices $\mathbf{T} \in \mathbb{R}^{P\times P}$, $\mathbf{Q} \in \mathbb{R}^{Q\times Q}$, and $\mathbf{P} \in \mathbb{R}^{R\times R}$, one finds that (Kiers, 1992):

$$\begin{aligned}
\mathbf{X}_A &= \mathbf{A}\mathbf{G}_A(\mathbf{C}^{\mathrm{T}} \otimes \mathbf{B}^{\mathrm{T}}) + \mathbf{E}_A \ , \\
&= \mathbf{A}\mathbf{T}\mathbf{T}^{-1}\mathbf{G}_A\left[(\mathbf{C}\mathbf{P}\mathbf{P}^{-1})^{\mathrm{T}} \otimes (\mathbf{B}\mathbf{Q}\mathbf{Q}^{-1})^{\mathrm{T}}\right] + \mathbf{E}_A \ , \\
&= (\mathbf{A}\mathbf{T})\mathbf{T}^{-1}\mathbf{G}_A(\mathbf{P}^{-1} \otimes \mathbf{Q}^{-1})^{\mathrm{T}}\left[(\mathbf{C}\mathbf{P})^{\mathrm{T}} \otimes (\mathbf{B}\mathbf{Q})^{\mathrm{T}}\right] + \mathbf{E}_A \ , \\
&= \hat{\mathbf{A}}\hat{\mathbf{G}}_A(\hat{\mathbf{C}}^{\mathrm{T}} \otimes \hat{\mathbf{B}}^{\mathrm{T}}) + \mathbf{E}_A \ ,
\end{aligned} \tag{12.14}$$

where $\hat{\mathbf{A}} = \mathbf{AT}, \hat{\mathbf{B}} = \mathbf{BQ}, \hat{\mathbf{C}} = \mathbf{CP}$, and $\hat{\mathbf{G}}_A = \mathbf{T}^{-1}\mathbf{G}_A(\mathbf{P}^{-1}\otimes\mathbf{Q}^{-1})^{\mathrm{T}}$. Equation (12.14) shows that rotation of the factors leaves the fit of the model unchanged provided that such transformations are compensated in the core array $\mathcal{G}$. Hence, the Tucker3 model suffers from rotational indeterminacy and the parameter matrices can only be determined up to a rotation.

The solutions of the Tucker3 model are usually difficult to interpret, due to the interactions between the components given by $\mathcal{G}$. The rotational freedom can be exploited to enhance the interpretability of the solution (see Kroonenberg, 2008, Chapter 10). Component matrices or the core can be rotated towards a specific target (Kiers, 1992). Furthermore, orthogonal and oblique transformations of the component matrices **A**, **B** and **C**, or of the core, or both, towards a simple structure can be considered.

Kiers (1997a) proposed a procedure that aims at core array simplicity by optimizing the Orthomax criterion (Jennrich, 1970). Kiers (1998a) discusses a method for joint orthogonal rotation of the core and the component matrices so as to optimize any desired weighted sum of simplicity values for the component matrices and the core. Oblique transformations of the core to obtain a simple structure are introduced in Kiers (1998b). Although simplicity of each of the component matrices can be optimized independently, for the core to be simple one has to strike compromises between simplicity of the core and of the component matrices. As Kiers and Van Mechelen (2001) point out, the desired simplicity of each of the component matrices and of the core may differ between situations.

The rotational freedom of the Tucker3 model is now exploited to implement ICA in one mode of the data. That is, based on an initial Tucker3 solution, one of the component matrices is rotated towards independence. To begin with, a Tucker3 solution needs to be obtained.

### 12.2.2 ALS solution and rotation towards independence

To find estimates for the parameters in the Tucker3 model, consider minimizing the following loss function (Kroonenberg and De Leeuw, 1980):

$$\mathcal{F}_{K\&DeL}(\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{G}) = ||\mathbf{X}_A - \mathbf{A}\mathbf{G}_A(\mathbf{C}\otimes\mathbf{B})^{\mathsf{T}}||_F^2 \ , \tag{12.15}$$

s.t. **A**, **B** and **C** are orthonormal. The loss function (12.15) can be also formulated in terms of the two remaining unfoldings $\mathbf{X}_B$ and $\mathbf{X}_C$ by cyclically permuting the letters that indicate the modes.

Kroonenberg and De Leeuw (1980) showed that for fixed **A**, **B** and **C**, the $\mathbf{G}_A$ which minimizes (12.15) is uniquely defined as:

$$\mathbf{G}_A = \mathbf{A}^{\mathsf{T}}\mathbf{X}_A(\mathbf{C}\otimes\mathbf{B}) \ . \tag{12.16}$$

Hence, the loss function (12.15) depends only upon **A**, **B** and **C**. Substituting (12.16), the loss function (12.15) can be rewritten as:

$$\begin{aligned}\mathcal{F}_{K\&DeL}(\mathbf{A},\mathbf{B},\mathbf{C}) &= ||\mathbf{X}_A - \mathbf{A}\mathbf{A}^{\mathsf{T}}\mathbf{X}_A(\mathbf{C}\mathbf{C}^{\mathsf{T}}\otimes\mathbf{B}\mathbf{B}^{\mathsf{T}})||_F^2 \ , \\ &= ||\mathbf{X}_A||_F^2 - \mathrm{trace}(\mathbf{A}^{\mathsf{T}}\{\mathbf{X}_A(\mathbf{C}\mathbf{C}^{\mathsf{T}}\otimes\mathbf{B}\mathbf{B}^{\mathsf{T}})\mathbf{X}_A^{\mathsf{T}}\}\mathbf{A}) \ . \end{aligned} \tag{12.17}$$

Hence minimizing (12.17) over $\mathbf{A}$ (keeping $\mathbf{B}$ and $\mathbf{C}$ fixed) is equivalent to maximizing trace($\mathbf{A}^\top\{\mathbf{X}_A(\mathbf{CC}^\top \otimes \mathbf{BB}^\top)\mathbf{X}_A^\top\}\mathbf{A}$). In a completely parallel fashion, it can be shown that minimizing (12.17) over $\mathbf{B}$ (keeping $\mathbf{A}$ and $\mathbf{C}$ fixed) is equivalent to maximizing trace($\mathbf{B}^\top\{\mathbf{X}_B(\mathbf{AA}^\top \otimes \mathbf{CC}^\top)\mathbf{X}_B^\top\}\mathbf{B}$) and minimizing (12.17) over $\mathbf{C}$ (keeping $\mathbf{A}$ and $\mathbf{B}$ fixed) is equivalent to maximizing trace($\mathbf{C}^\top\{\mathbf{X}_C(\mathbf{BB}^\top \otimes \mathbf{AA}^\top)\mathbf{X}_C^\top\}\mathbf{C}$).

To minimize (12.17), Kroonenberg and De Leeuw (1980) developed an ALS algorithm, the TUCKALS3 algorithm, in which in each main iteration step, $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ are updated in turn, while keeping the other two parameter matrices fixed (see also Kroonenberg, 1983; ten Berge, De Leeuw, and Kroonenberg, 1987).

To initialize the TUCKALS3 algorithm, $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ are chosen according to Tucker's algebraic solution (Tucker, 1966). That is, initially $\mathbf{A}$ consists of the principal $P$ eigenvectors of $\mathbf{X}_A\mathbf{X}_A^\top$; $\mathbf{B}$ consists of the principal $Q$ eigenvectors of $\mathbf{X}_B\mathbf{X}_B^\top$; and $\mathbf{C}$ consists of the principal $R$ eigenvectors of $\mathbf{X}_C\mathbf{X}_C^\top$.

In each A-substep of the main iterative procedure one could carry out an eigendecomposition of the $I \times I$ matrix $\mathbf{X}_A(\mathbf{CC}^\top \otimes \mathbf{BB}^\top)\mathbf{X}_A^\top$ to find an update for $\mathbf{A}$. Analogously, in each B-substep (C-substep) one can compute an eigendecomposition of the $J \times J$ matrix $\mathbf{X}_B(\mathbf{AA}^\top \otimes \mathbf{CC}^\top)\mathbf{X}_B^\top$ (of the $K \times K$ matrix $\mathbf{X}_C(\mathbf{BB}^\top \otimes \mathbf{AA}^\top)\mathbf{X}_C^\top$) to find an update for $\mathbf{B}$ (for $\mathbf{C}$).

Such a procedure is likely to become computationally burdensome because of the sizes of the corresponding matrices. To avoid this, Kroonenberg and De Leeuw (1980) used one step in the iterative routine of Bauer-Rutishauser (Rutishauser, 1969) for computing eigenvectors and eigenvalues of a matrix, to find an update for $\mathbf{A}$ (and analogously for $\mathbf{B}$ and $\mathbf{C}$). The advantage of this approach is that in the A-substep only an eigendecomposition of a $P \times P$ matrix is required to find an update for $\mathbf{A}$, where usually

$P \ll I$. Correspondingly, in the B-substep (C-substep) only an eigendecomposition of a $Q \times Q$ ($R \times R$) matrix is required for finding an update for $\mathbf{B}$ (for $\mathbf{C}$), where in practice $Q \ll J$ and $R \ll K$.

After all parameter matrices have been estimated once, the main iterative step is repeated again and again until convergence. The iterative procedure terminates if two successive estimates for $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ are sufficiently alike or if the differences between successive iterations with respect to the loss function are below some arbitrary small value. The loss function can be shown to converge in a monotone fashion to at least a local optimum. Once estimates for the component matrices are found, an estimate for the core matrix can be computed via (12.16).

Further improvements of the original TUCKALS3 algorithm which led to reduction of the computational load are proposed by Kroonenberg, ten Berge, Brouwer, and Kiers (1989) and Kiers, Kroonenberg, and ten Berge (1992).

Assume that independence constraints are imposed on the components of the first mode. Finding orthonormal $\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{C}}$ and a core $\hat{\mathbf{G}}_A$ is an ALS problem. To implement ICA in the $A$-mode of the data, one needs to go one step further. The component scores $\mathbf{A}$ should be independent. For this reason $\hat{\mathbf{A}}$ is rotated towards independence, that is, $\tilde{\mathbf{A}} = \hat{\mathbf{A}}\mathbf{T}$ for some orthogonal $P \times P$ matrix $\mathbf{T}$. To find the matrix $\mathbf{T}$ that leads to approximately independent component scores $\tilde{\mathbf{A}}$, one can solve for example the optimization problem (11.5) by means of the projected gradient approach. In fact, any of the standard ICA criteria and algorithms discussed in Chapter 10 such as FastICA or JADE can be used to achieve the ICA goal.

Once the optimal rotation matrix has been found, the core has to be rotated as well to maintain the LS fit of the model. Summarizing, the proposed method to implement the ICA data sets having a three-way structure is as follows:

1. Set up the number of components in each mode $[P, Q, R]$, prescribed or using a model-selection procedure (see Kroonenberg, 2005).
2. Obtain estimates $\hat{\mathbf{A}}$, $\hat{\mathbf{B}}$ and $\hat{\mathbf{C}}$ solving (12.15) by means of an ALS algorithm.
3. Compute $\hat{\mathbf{G}}_A$ via (12.16).
4. Find an orthogonal matrix $\mathbf{T}$ that solves the optimization problem (11.5), where $\mathbf{S_H}$ in (11.5) is the covariance matrix of the squared orthogonally transformed component scores $\hat{\mathbf{A}}$.
5. Calculate approximately independent components in the first mode as $\tilde{\mathbf{A}} = \hat{\mathbf{A}}\mathbf{T}$.
6. Obtain a counter-rotated core by $\tilde{\mathbf{G}}_A = \mathbf{T}^\top \hat{\mathbf{G}}_A$.

## 12.3 Simulation experiment

The performance of the proposed approach to ICA (called Tucker3-ICA) of three-mode data shall be evaluated by means of a simulation study. A matricized version $\mathbf{X}_A$ of a three-way array $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$ with known model structure

$$\mathbf{X}_A = \mathbf{A}\mathbf{G}_A(\mathbf{C}^\top \otimes \mathbf{B}^\top) + \mathbf{E}_A \tag{12.18}$$

and independent components in the first mode is constructed as follows. Recall the 27 × 26 box data in Chapter 11. The dimensions of the 27 boxes in Table 7.1 and Table 11.1 are independent and hence well-suited for an ICA analysis. The three dimensions constitute the column vectors of the true component matrix $\mathbf{A} \in \mathbb{R}^{27 \times 3}$ and hence the factors the proposed approach aims to recover. The entries of the other two modes (in $\mathbf{B} \in \mathbb{R}^{20\times3}$ and $\mathbf{C} \in \mathbb{R}^{50\times3}$) are drawn from a standard normal distribution. Performance of the algorithm is investigated under different degrees of multi-collinearity in the core and various levels of additive observational noise. The elements of the unfolded core, $\mathbf{G}_A \in \mathbb{R}^{3\times9}$, are drawn randomly from the uniform distribution on the interval $[0, 1]$ in the 'low' multi-collinearity case, and from the uniform distribution on the interval $[0.5, 1.5]$ in the 'high' multi-collinearity case.

The values of $\mathbf{E}_A$ were drawn randomly from a zero-mean normal distribution with noise variances $\sigma_N^2 = 1, \ldots, 20$. Furthermore, two more conditions are added to the analysis to provide 'zero-point' references, namely the case of a unit superdiagonal core and the case of complete absence of noise.

To eliminate unwanted differences in level and scale, all elements $x_{ijk}$ $(i = 1, \ldots, I; j = 1, \ldots, J; k = 1, \ldots, K)$ in $\mathbf{X}_A$ are preprocessed by centring across the entries of the $A$-mode and normalizing within the $B$-mode, that is,

$$x_{ijk} \leftarrow \frac{x_{ijk} - x_{.jk}}{\sqrt{\sum_{i=1}^{I} \sum_{k=1}^{K} (x_{ijk} - x_{.jk})^2}}, \tag{12.19}$$

where the subscript dot is used to indicate the mean across $i = 1, \ldots, I$. Preprocessing by (12.19) centres all vertical 'fibers' by subtracting the fiber means and normalizes the complete lateral slices by dividing by the square root of the sum of squares in each slice (Kiers, 2000).

To obtain an initial Tucker3 solution the loss function (12.15) is optimized by means of the TUCKALS3 algorithm. For running the TUCKALS3 algorithm, the MATLAB $N$-way toolbox version 3.1 is used (Andersson and Bro, 2000).

Once an initial Tucker3 solution has been obtained, the procedure described above is used to implement the ICA. Monte-Carlo simulations are conducted consisting of

500 replications. The MATLAB code for running the simulation study is available upon request. The performance of the proposed approach is evaluated by means of the normalized Frobenius norm of the difference between the true (standardized) source matrix **A** and the recovered (estimated) matrix $\tilde{\mathbf{A}}$:

$$E = \frac{||\mathbf{A} - \tilde{\mathbf{A}}||_F}{||\mathbf{A}||_F} \ .$$

Figure 12.1 displays the mean value of $E$ against noise levels $\sigma_N^2 = 1, \ldots, 20$ for Tucker3-ICA and three different cores. With respect to robustness against additive Gaussian

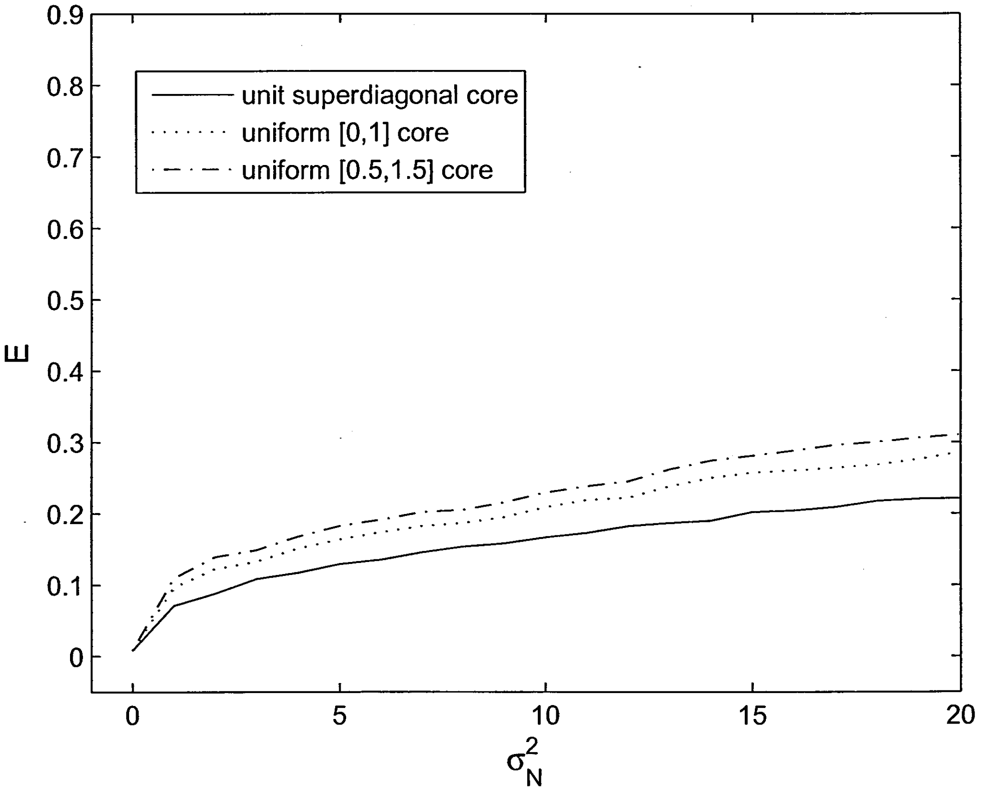

Figure 12.1: Mean value of $E$ against noise levels $\sigma_N^2 = 1, \ldots, 20$ assuming three different cores for Tucker3-ICA based on 500 replications.

noise, Tucker3-ICA performs best if the components in the three modes are only allowed to interact factorwise. However, Figure 12.1 reveals that the proposed method is also quite robust in the cases of correlated cores. The gap in performance between the unit superdiagonal core solution and the two multi-collinear cases increases with increasing noise level. Irrespective of the core, **A** is recovered accurately in the case of complete absence of noise.

Finally, the effect of varying the sizes of the data array and the core has been studied. Results (not shown) indicate that if more components are extracted in the second and third mode, the overall fit of the model is improved. The mean value of $E$ decreases with increasing the size of the core. Another feature is that the mean value of $E$ increases if the number of entries in the second and third mode, $J$ and $K$, are reduced.

# Chapter 13

# Discussion

In Chapter 11 the IEFA method was introduced for recovering independent latent sources from their observed mixtures. To implement IEFA, the new model was viewed as a method of factor rotation in EFA. First, estimates for all EFA model parameters were obtained simultaneously by means of the numerical procedures presented in Part II. Then, an orthogonal rotation matrix was sought that minimizes the dependence between the common factors. The rotation criterion used requires minimization of squared fourth-order statistics formed by covariances computed from squared components. It is easily optimized using the projected gradient approach. It was chosen because its appropriateness was easily motivated. Of course, the IEFA method can be used in combination with any other ICA rotation criterion.

Since the initial EFA decomposition is based on the computationally efficient procedure of the SVD of data matrices, the IEFA method facilitates the application of noisy ICA. In particular for high-dimensional data, this initial EFA decomposition seems to be a reasonable choice for decorrelating the common factors before applying any ICA rotation criteria.

The new approach was applied to the notorious Thurstone's 26-variable box problem. By rotating the factors towards independence, the dimensions of the boxes were revealed accurately and a simple structure of the loadings was achieved.

Note that there is an important difference between simplicity rotation criteria and IEFA. The application of the IEFA rotation is based on the assumption that the underlying sources for the process to be analyzed are independent. Then, rotating the scores towards independence produces patterns of loadings in which the obtained simplicity reflects the physics of the underlying process rather than a formal simplicity criterion. In contrast, for rotation-to simplicity methods there is no clear guidance on which one to use to enhance interpretation.

Thurstone's box data is an entirely artificial data set, though. The example was chosen as one illustration of the usefulness of IEFA. Noisy ICA is a rapidly evolving method that is currently finding applications in various disciplines, e.g. atmospheric science. It will be of great interest to explore the application of IEFA to real (correlated) data. Especially for high-dimensional data one might produce with the IEFA rotation a simple structure in the loadings in a computationally more efficient fashion. With $p \gg n$, rotating a $p \times k$ matrix of initial loadings towards simplicity needs more CPU time than rotating an $n \times k$ matrix of scores towards independence.

In Chapter 12 an extension of common two-way ICA to data sets having a third mode was presented. The ICA was implemented by exploiting the rotational freedom of the Tucker3 model. After obtaining an initial Tucker3 solution, one of the component matrices was rotated orthogonally towards independence to implement ICA in one mode of the data. However, obtaining approximately independent factor scores in one mode may spoil the simplicity and hence the interpretation of the core. Rotation towards independence in one mode can be combined with rotation-to-simplicity methods to enhance interpretation of the remaining component matrices and/or the core.

In the conducted numerical experiments Tucker3-ICA was shown to be quite robust against normally distributed noise. As indispensable stages in a complete three-way analysis process, one should carry out a more detailed study of model fit and the residuals in practice. This not only includes a careful choice of the number of components for each mode but also the choice between different models, e.g. between a Tucker3 and a Tucker2 model (Kroonenberg, 2005). Furthermore, the performance of the new approach to ICA for three-mode data has to be compared to the existing algorithms tensor pICA and ICA-CP both of which impose independence constraints in the CP model.

The Tucker3 model does not include the concept of unique factors. A (stochastic) three-mode common factor model with unique variances for combination variables was proposed by Bloxom (1968) and further developed by Bentler and Lee (1978, 1979). These techniques differ mainly from the Tucker3 model in that they are modelling covariances rather than the raw data (see also Kroonenberg, 2003). Due to aggregation over the entities in one mode no estimate for one of the component matrices is given. Extending the analysis in Part II to develop an algorithm for fitting the three-mode common factor model directly to the data shall be the subject of further work.

# Bibliography

ABSIL, P.-A., R. MAHONY, AND R. SEPULCHRE (2008): *Optimization Algorithms on Matrix Manifolds.* Princeton University Press: Princeton.

ACAR, E., AND B. YENER (2007): "Unsupervised multiway data analysis: A literature survey," Department of Computer Science, Rensselaer Polytechnic Institute, Troy, New York.

AMARI, S., A. CICHOCKI, AND H. H. YANG (1996): "A new learning algorithm for blind signal separation," in *Advances in Neural Information Processing Systems, Vol. 8*, ed. by D. Touretzky, M. Mozer, and M. Hasselmo, pp. 757–763. MIT Press: Cambridge, Massachusetts.

ANDERSON, T. W. (1984): *An Introduction to Multivariate Statistical Analysis.* John Wiley & Sons: New York, 2nd edn.

ANDERSON, T. W., AND H. RUBIN (1956): "Statistical inference in factor analysis," in *Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability, Vol. V*, ed. by J. Neyman, pp. 111–150. University of California Press: Berkeley.

ANDERSSON, C. A., AND R. BRO (2000): "The $N$-way toolbox for MATLAB," *Chemometrics and Intelligent Laboratory Systems*, 52, 1–4.

ATTIAS, H. (1999): "Independent factor analysis," *Neural Computation*, 11, 803–852.

BÄRRING, L. (1987): "Spatial patterns of daily rainfall in central Kenya: Application of principal component analysis, common factor analysis and spatial correlation," *International Journal of Climatology*, 7, 267–289.

BARTHOLOMEW, D. J., AND M. KNOTT (1999): *Latent Variable Models and Factor Analysis.* Edward Arnold: London, 2nd edn.

BARTHOLOMEW, D. J., F. STEELE, I. MOUSTAKI, AND J. I. GALBRAITH (2002): *The Analysis and Interpretation of Multivariate Data for Social Scientists.* Chapman & Hall/CRC: Boca Raton, Florida.

BARTLETT, M. S. (1937): "The statistical conception of mental factors," *British Journal of Psychology*, 28, 97–104.

BASILEVSKY, A. (1994): *Statistical Factor Analysis and Related Methods: Theory and Applications.* John Wiley & Sons: New York.

BECKER, C., AND U. GATHER (2001): "The largest nonidentifiable outlier: a comparison of multivariate simultaneous outlier identification rules," *Computational Statistics & Data Analysis*, 36, 119–127.

BECKMANN, C. F., AND S. M. SMITH (2004): "Probabilistic independent component analysis for functional magnetic resonance imaging," *IEEE Transactions on Medical Imaging*, 24, 137–152.

——— (2005): "Tensorial extensions of independent component analysis for multi-subject FMRI analysis," *Neuroimage*, 25, 294–311.

BENTLER, P. M., AND S.-Y. LEE (1978): "Statistical aspects of a three-mode factor analysis model," *Psychometrika*, 43(3), 343–352.

——— (1979): "A statistical development of three-mode factor analysis," *British Journal of Mathematical and Statistical Psychology*, 32, 87–104.

BERNAARDS, C. A., AND R. I. JENNRICH (2005): "Gradient projection algorithms and software for arbitrary rotation criteria in factor analysis," *Educational and Psychological Measurement*, 65, 676–696.

BERRY, M. W., S. A. PULATOVA, AND G. W. STEWART (2005): "Algorithm 844: Computing sparse reduced-rank approximations to sparse matrices," *ACM Transactions on Mathematical Software (TOMS)*, 31, 252–269.

BLOXOM, B. (1968): "A note on invariance in three-mode factor analysis," *Psychometrika*, 33, 347–350.

BOLTON, R. A., AND W. J. KRZANOWSKI (1999): "A characterization of principal components for projection pursuit," *The American Statistician*, 53, 108–109.

BROWNE, M. W. (2001): "An overview of analytic rotation in exploratory factor analysis," *Multivariate Behavioral Research*, 36, 111–150.

BUKANTIS, A. (2002): "Application of factor analysis for quantification of climate-forming processes in the eastern part of the Baltic Sea region," *Climate Research*, 20, 135–140.

CARDOSO, J.-F. (1998): "Blind signal separation: statistical principles," *Proceedings of the IEEE. Special Issue on Blind Identification and Estimation*, 9, 2009–2025.

——— (1999): "High-order contrasts for independent component analysis," *Neural Computation*, 11, 157–192.

Cardoso, J.-F., and A. Souloumiac (1993): "Blind beamforming for non-Gaussian signals," *IEE Proceedings-F*, 140, 362–370.

——— (1996): "Jacobi angles for simultaneous diagonalization," *(SIAM) Journal on Matrix Analysis and Applications*, 17, 161–164.

Carroll, J. B. (1953): "An analytic solution for approximating simple structure in factor analysis," *Psychometrika*, 18, 23–28.

Carroll, J. D., and J. J. Chang (1970): "Analysis of individual differences in multidimensional scaling via an $n$-way generalization of 'Eckart-Young' decomposition," *Psychometrika*, 35, 283–319.

Carter, M. M., and J. B. Elsner (1997): "A statistical method for forecasting rainfall over Puerto Rico," *Weather and Forecasting*, 12, 515–525.

Casella, G., and R. Berger (2002): *Statistical Inference.* Duxbury: Pacific Grove, California, 2nd edn.

Chu, M. T., and K. R. Driessel (1990): "The projected gradient method for least squares matrix approximations with spectral constraints," *SIAM Journal on Numerical Analysis*, 27, 1050–1060.

Comon, P. (1994): "Independent component analysis, a new concept?," *Signal Processing*, 36, 287–314.

COVER, T. M., AND J. A. THOMAS (1991): *Elements of Information Theory.* John Wiley & Sons: New York.

CRAWFORD, C. B., AND G. A. FERGUSON (1970): "A general rotation criterion and its use in orthogonal rotation," *Psychometrika*, 35, 321–332.

CROUX, C., P. FILZMOSER, G. PISON, AND P. J. ROUSSEEUW (2003): "Fitting multiplicative models by robust alternating regressions," *Statistics and Computing*, 13, 23–36.

DAVIES, M. (2004): "Identifiability issues in noisy ICA," *IEEE Signal Processing Letters*, 11, 470–473.

DE LATHAUWER, L., B. DE MOOR, AND J. VANDEWALLE (2000): "An introduction to independent component analysis," *Journal of Chemometrics*, 14, 123–149.

——— (2004): "Computation of the canonical decomposition by means of a simultaneous generalized schur decomposition," *SIAM Journal on Matrix Analysis and Applications*, 26, 295–327.

DE LEEUW, J. (1994): "Block relaxation algorithms in statistics," in *Information Systems and Data Analysis*, ed. by H.-H. Bock, W. Lensi, and M. M. Richter, pp. 308–324. Springer: Berlin.

DE LEEUW, J. (2004): "Least squares optimal scaling of partially observed linear systems," in *Recent Developments on Structural Equation Models: Theory and Applications*, ed. by K. van Montfort, J. Oud, and A. Satorra, pp. 121–134. Kluwer Academic Publishers: Dordrecht.

——— (2008): "Factor analysis as matrix decomposition," Preprint series: Department of Statistics, University of California, Los Angeles.

De Leeuw, J., and K. Lange (2009): "Sharp quadratic majorization in one dimension," *Computational Statistics & Data Analysis*, 53, 2471–2484.

De Vos, M., L. De Lathauwer, and S. Van Huffel (2007): "Imposing independence constraints in the CP model," in *Proceedings of the 7th International Conference on Independent Component Analysis and Signal Separation, ICA 2007, London, UK*, ed. by M. E. Davies, and et al., pp. 33–40. Springer: Berlin, Heidelberg.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977): "Maximum-likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B*, 39, 1–38.

Diaconis, P., and D. A. Freedman (1984): "Asymptotics of graphical projection pursuit," *The Annals of Statistics*, 12, 793–815.

Eckart, C., and G. Young (1936): "The approximation of one matrix by another of lower rank," *Psychometrika*, 1, 211–218.

Edelman, A., T. A. Arias, and S. T. Smith (1998): "The geometry of algorithms with orthogonality constraints," *SIAM Journal on Matrix Analysis and Applications*, 20, 303–353.

Etezadi-Amoli, J., and R. P. McDonald (1983): "A second generation nonlinear factor analysis," *Psychometrika*, 48, 315–342.

Filzmoser, P. (1999): "Robust principal components and factor analysis in the geostatistical treatment of environmental data," *Environmetrics*, 10, 363–375.

——— (2002): "Robust factor analysis: Methods and applications," in *Latent Variable and Latent Structure Models*, ed. by G. A. Marcoulides, and I. Moustaki, pp. 153–194. Lawrence Erlbaum Associates: Mahwah, New Jersey.

FRIEDMAN, J. H. (1987): "Exploratory projection pursuit," *Journal of the American Statistical Association*, 82, 249–266.

FRIEDMAN, J. H., AND J. W. TUKEY (1974): "A projection pursuit algorithm for exploratory data analysis," *IEEE Transactions on Computers*, 23, 881–890.

GABRIEL, K. R., AND S. ZAMIR (1979): "Lower rank approximation of matrices by least squares with any choice of weights," *Technometrics*, 21, 489–498.

GOLUB, G. H., AND C. F. VAN LOAN (1996): *Matrix Computations.* The John Hopkins University Press: Baltimore, Maryland, 3rd edn.

GOWER, J. C., AND G. B. DIJKSTERHUIS (2004): *Procrustes Problems.* Oxford University Press: Oxford.

GROENEN, P. J. F., P. GIAQUINTO, AND H. A. L. KIERS (2003): "Weighted majorization algorithms for weighted least squares decomposition models," Econometric Institute Report EI 2003-09: Erasmus University Rotterdam.

——— (2005): "An improved majorization algorithm for robust procrustes analysis," in *New Developments in Classification and Data Analysis*, ed. by M. Vichi, P. Monari, S. Mignani, and A. Montanari, pp. 151–158. Springer: Berlin.

GRUBIŠIĆ, I., AND R. PIETERSZ (2007): "Efficient rank reduction of correlation matrices," *Linear Algebra and its Applications*, 422, 629–653.

GUTTMAN, L. (1955): "The determinacy of factor score matrices with implications for five other basic problems of common-factor theory," *British Journal of Statistical Psychology*, 8, 65–81.

HANNACHI, A., I. T. JOLLIFFE, AND D. B. STEPHENSON (2007): "Empirical orthogonal functions and related techniques in atmospheric science: A review," *International Journal of Climatology*, 27, 1119–1152.

HARMAN, H. H. (1976): *Modern Factor Analysis*. University of Chicago Press: Chicago, 3rd edn.

HARSHMAN, R. A. (1970): "Foundations of the PARAFAC procedure: Models and methods for an 'explanatory' multi-mode factor analysis," *UCLA Working Papers in Phonetics*, 16, 1–84.

HASTIE, T., R. TIBSHIRANI, AND J. H. FRIEDMAN (2009): *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer: New York, 2nd edn.

HEISER, W. J. (1995): "Convergent computation by iterative majorization: Theory and applications in multidimensional data analysis," in *Recent Advances in Descriptive Multivariate Analysis*, ed. by W. J. Krzanowski, pp. 157–189. Oxford University Press: Oxford.

HELMKE, U., AND J. B. MOORE (1994): *Optimization and Dynamical Systems*. Springer: Berlin.

HINICH, M. J. (1994): "Higher order cumulants and cumulant spectra," *Circuits, Systems, and Signal Processing*, 13, 391–402.

HIRSH, M. W., AND S. SMALE (1974): *Differential Equations, Dynamical Systems, and Linear Algebra.* Academic Press: San Diego.

HORST, P. (1965): *Factor Analysis of Data Matrices.* Holt, Rinehart and Winston: New York.

HUBER, P. J. (1981): *Robust Statistics.* John Wiley & Sons: New York.

——— (1985): "Projection pursuit," *The Annals of Statistics*, 13, 435–475.

HYVÄRINEN, A. (1998): "Independent component analysis in the presence of gaussian noise by maximizing joint likelihood," *Neurocomputing*, 22, 49–67.

——— (1999a): "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Transactions on Neural Networks*, 10, 626–634.

——— (1999b): "Gaussian moments for noisy independent component analysis," *IEEE Signal Processing Letters*, 6, 145–147.

——— (1999c): "Survey on independent component analysis," *Neural Computing Surveys*, 2, 94–128.

HYVÄRINEN, A., J. KARHUNEN, AND E. OJA (2001): *Independent Component Analysis.* John Wiley & Sons: New York.

IKEDA, S., AND K. TOYAMA (2000): "Independent component analysis for noisy data - MEG data analysis," *Neural Networks*, 13, 1063–1074.

IZENMAN, A. J. (2008): *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning.* Springer: New York.

JENNRICH, R. I. (1970): "Orthogonal rotation algorithms," *Psychometrika*, 35, 229–235.

——— (2001): "A simple general procedure for orthogonal rotation," *Psychometrika*, 66, 289–306.

——— (2002): "A simple general method for oblique rotation," *Psychometrika*, 67, 7–20.

——— (2004a): "Derivative free gradient projection algorithms for rotation," *Psychometrika*, 69, 475–480.

——— (2004b): "Rotation to simple loadings using component loss functions: The orthogonal case," *Psychometrika*, 69, 257–273.

JENNRICH, R. I., AND N. T. TRENDAFILOV (2005): "Independent component analysis as a rotation method: A very different solution to Thurstone's box problem," *British Journal of Mathematical and Statistical Psychology*, 58, 199–208.

JOLLIFFE, I. T. (2002): *Principal Component Analysis.* Springer: New York, 2nd edn.

JONES, M. C., AND R. SIBSON (1987): "What is projection pursuit?," *Journal of the Royal Statistical Society, Series A*, 150, 1–36.

JÖRESKOG, K. G. (1962): "On the statistical treatment of residuals in factor analysis," *Psychometrika*, 27, 335–354.

——— (1967): "Some contributions to maximum-likelihood factor analysis," *Psychometrika*, 32, 443–482.

——— (1977): "Factor analysis by least-squares and maximum likelihood methods," in *Mathematical methods for digital computers*, ed. by K. Enslein, A. Ralston, and H. S. Wilf, pp. 125–153. John Wiley & Sons: New York.

KAISER, H. F. (1958): "The varimax criterion for analytic rotation in factor analysis," *Psychometrika*, 23, 187–200.

KANO, Y., Y. MIYAMOTO, AND S. SHIMIZU (2003): "Factor rotation and ICA," in *Proceedings of the 4th International Symposium on Independent Component Analysis and Blind Source Separation (ICA 2003)*, ed. by S. Amari, A. Cichocki, S. Makino, and N. Murata, pp. 101–105. Nara, Japan.

KESTELMAN, H. (1952): "The fundamental equation of factor analysis," *British Journal of Psychology, Statistics Section*, 5, 1–6.

KIERS, H. A. L. (1992): "TUCKALS core rotations and constrained TUCKALS modelling," *Statistica Applicata*, 4, 659–667.

——— (1997a): "Three-mode orthomax rotation," *Psychometrika*, 62, 579–598.

——— (1997b): "Weighted least squares fitting using ordinary least squares algorithms," *Psychometrika*, 62, 251–266.

——— (1998a): "Joint orthomax rotation of the core and component matrices resulting from three-mode principal component analysis," *Journal of Classification*, 15, 245–263.

——— (1998b): "Three-way SIMPLIMAX for oblique rotation of the three-mode factor analysis core to simple structure," *Computational Statistics & Data Analysis*, 28, 307–324.

——— (2000): "Towards a standardized notation and terminology in multiway analysis," *Journal of Chemometrics*, 14, 105–122.

——— (2002): "Setting up alternating least squares and iterative majorization algorithms for solving various matrix optimization problems," *Computational Statistics & Data Analysis*, 41, 157–170.

Kiers, H. A. L., P. M. Kroonenberg, and J. M. F. ten Berge (1992): "An efficient algorithm for TUCKALS3 on data with large number of observation units," *Psychometrika*, 57, 415–422.

Kiers, H. A. L., and I. Van Mechelen (2001): "Three-way component analysis: Principles and illustrative application," *Psychological Methods*, 6, 84–110.

Kosfeld, R. (1996): "Robust exploratory factor analysis," *Statistical papers*, 37, 105–122.

Kroonenberg, P. M. (1983): *Three-Mode Principal Component Analysis: Theory and Applications.* DSWO Press: Leiden.

——— (2003): "Three-mode analysis of multimode covariance matrices," *British Journal of Mathematical and Statistical Psychology*, 356, 305–335.

——— (2005): "Model selection procedures in three-mode component models," in *New Developments in Classification and Data Analysis*, ed. by M. Vichi, P. Monari, S. Mignani, and A. Montanari, pp. 167–172. Springer: Berlin, Heidelberg.

——— (2008): *Applied Multiway Data Analysis.* John Wiley & Sons: Hoboken, New Jersey.

Kroonenberg, P. M., and J. De Leeuw (1980): "Principal component analysis of three-mode data by means of alternating least squares algorithms," *Psychometrika*, 45, 69–97.

Kroonenberg, P. M., J. M. F. ten Berge, P. Brouwer, and H. A. L. Kiers (1989): "Gram-Schmidt versus Bauer-Rutishauser in alternating least squares algorithms for three-mode principal component analysis," *Computational Statistics Quarterly*, 5, 81–87.

Kruskal, J. B. (1977): "Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics," *Linear Algebra and its Applications*, 18, 95–138.

Krzanowski, W. J. (1988): *Principles of Multivariate Analysis: A User's Perspective.* Oxford University Press: Oxford.

Lathauwer, L. D., B. D. Moor, and J. Vanderwalle (1996): "Independent component analysis based on higher-order statistics only," *Proceedings of the 8th IEEE Signal Processing Workshop on Statistical Signal and Array Processing (SSAP 1996)*, pp. 356–359.

Lawley, D. N. (1942): "Further investigations in factor estimation," *Proceedings of the Royal Society of Edinburgh: Section A*, 61, 176–185.

Lawley, D. N., and A. E. Maxwell (1971): *Factor Analysis as a Statistical Method.* Butterworth: London, 2nd edn.

Ledermann, W. (1937): "On the rank of the reduced correlation matrix in multiple factor analysis," *Psychometrika*, 2, 85–93.

——— (1939): "On a shortened method of estimation of mental factors by regression," *Psychometrika*, 4, 109–115.

LEE, T.-W. (1998): *Independent Component Analysis: Theory and Applications.* Kluwer Academic Publishers: Boston.

LÜTKEPOHL, H. (1996): *Handbook of Matrices.* John Wiley & Sons: Chichester.

MAGNUS, J. R., AND H. NEUDECKER (1988): *Matrix Differential Calculus with Applications in Statistics and Econometrics.* John Wiley & Sons: Chichester.

MARDIA, K. V., J. T. KENT, AND J. M. BIBBY (1979): *Multivariate Analysis.* Academic Press: London.

MARYON, R. H. (1979): "Eigenanalysis of the Northern Hemispherical 15-day mean surface pressure field and its application to long-range forecasting," *Met O 13 Branch Memorandum No. 82 (unpublished).* UK Meteorological Office: Bracknell.

MAVRIDIS, D., AND I. MOUSTAKI (2008): "Detecting outliers in factor analysis using the forward search algorithm," *Multivariate Behavioral Research*, 43, 453–475.

MCCAMMON, R. B. (1966): "Principal components analysis and its application in large-scale correlation studies," *Journal of Geology*, 74, 721–733.

MCCULLAGH, P. (1987): *Tensor Methods in Statistics.* Chapman & Hall: London.

MCDONALD, R. P. (1979): "The simultaneous estimation of factor loadings and scores," *British Journal of Mathematical and Statistical Psychology*, 32, 212–228.

MOOD, A. M., F. A. GRAYBILL, AND D. C. BOES (1974): *Introduction to the Theory of Statistics.* McGraw-Hill: Tokyo, 3rd edn.

MOOIJAART, A. (1985): "Factor analysis for non-normal variables," *Psychometrika*, 50, 323–342.

MOSTELLER, F., AND J. W. TUKEY (1977): *Data Analysis and Regression.* Addison-Wesley: Reading, Massachusetts.

MOULINES, E., J.-F. CARDOSO, AND E. GASSIAT (1997): "Maximum likelihood for blind source seperation and deconvolution of noisy signals using mixture models," Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'97), Munich, 3617-3620.

MULAIK, S. A. (1972): *The Foundations of Factor Analysis.* McGraw-Hill: New York.

——— (2005): "Looking back on the indeterminacy controversies in factor analysis," in *Contemporary Psychometrics: A Festschrift for Roderick P. McDonald*, ed. by A. Maydeu-Olivares, and J. J. McArdle, pp. 173–206,. Lawrence Erlbaum: Mahwah.

NASON, G. P. (1992): "Design and choice of projection indices," PhD Thesis, University of Bath: Bath.

PAWITAN, Y. (2001): *In All Likelihood: Statistical Modelling and Inference Using Likelihood.* Oxford University Press: Oxford.

PENNY, W. D., S. J. ROBERTS, AND R. M. EVERSON (2001): "ICA: Model order selection and dynamic source models," in *Independent Component Analysis: Principles and Practice*, ed. by S. Roberts, and R. Everson, pp. 299–314. Cambridge University Press: Cambridge.

PISON, G., P. J. ROUSSEEUW, P. FILZMOSER, AND C. CROUX (2003): "Robust factor analysis," *Journal of Multivariate Analysis*, 84, 145–172.

RAO, C. R. (1996): “Principal component and factor analyses,” in *Handbook of Statistics, Vol. 14*, ed. by G. S. Maddala, and C. R. Rao, pp. 489–505. Elsevier: Amsterdam.

RICHMAN, M. B. (1986): “Rotation of principal components,” *International Journal of Climatology*, 6, 293–335.

ROBERTSON, D., AND J. SYMONS (2007): “Maximum likelihood factor analysis with rank-deficient sample covariance matrices,” *Journal of Multivariate Analysis*, 98, 813–828.

ROBITZSCH, A. (2003): *Kontraste und M-Schätzer in der Independent Component Analysis.* Diplomarbeit: Technische Universität Dresden, Institut für Mathematische Stochastik.

ROUSSEEUW, P. J. (1985): “Multivariate estimation with high breakdown point,” in *Mathematical Statistics and Applications, vol. B*, ed. by W. Grossmann, pp. 283–297. Reidel: Dordrecht.

ROUSSEEUW, P. J., AND A. M. LEROY (1987): *Robust Regression and Outlier Detection.* John Wiley & Sons: Hoboken, New Jersey.

ROZEBOOM, W. W. (1991): “Theory and practice of analytic hyperplane optimization,” *Multivariate Behavioral Research*, 26, 179–197.

RUBIN, D. B., AND D. T. THAYER (1982): “EM algorithms for ML factor analysis,” *Psychometrika*, 47, 69–76.

RUTISHAUSER, H. (1969): “Computational aspects of F. L. Bauer’s simultaneous iteration method,” *Numerische Mathematik*, 13, 4–13.

SCHNEEWEISS, H., AND H. MATHES (1995): "Factor analysis and principal components," *Journal of Multivariate Analysis*, 55, 105–124.

SHAMPINE, L. F., AND M. W. REICHELT (1997): "The MATLAB ODE suite," *SIAM Journal on Scientific Computing*, 18, 1–22.

SOČAN, G. (2003): "The Incremental Value of Minimum Rank Factor Analysis," PhD Thesis, University of Groningen: Groningen.

STEGEMAN, A. (2007): "Comparing independent component analysis and the Parafac model for artificial multi-subject fMRI data," Technical Report: Heymans Insititute, University of Groningen.

STEGEMAN, A., AND A. MOOIJAART (2008): "Independent component analysis with errors by least squares covariance fitting," Technical report: Heymans Insititute, University of Groningen.

STEWART, G. W. (1998): *Matrix Algorithms I: Basic Decompositions.* SIAM: Philadelphia.

STIEFEL, E. (1935-1936): "Richtungsfelder und Fernparallelismus in n-dimensionalen Mannigfaltigkeiten," *Commentarii Mathematici Helvetici*, 8, 305–335.

STONE, J. V. (2004): *Independent Component Analysis: A Tutorial Introduction.* MIT Press: Cambridge, Massachusetts.

STUART, A., AND J. K. ORD (1994): *Kendall's Advanced Theory of Statistics, Vol. 1: Distribution Theory.* John Wiley & Sons: New York, 6th edn.

TEN BERGE, J. M. F., J. DE LEEUW, AND P. M. KROONENBERG (1987): "Some additional results on principal components analysis of three-mode data by means of alternating least squares algorithms," *Psychometrika*, 52, 183–191.

THE MATHWORKS (2008): *Using MATLAB, Version 7.7 (R2008b).* The MathWorks Inc.

THURSTONE, L. L. (1935): *The Vectors of Mind.* University of Chicago Press: Chicago.

——— (1947): *Multiple Factor Analysis.* University of Chicago Press: Chicago.

TOMASI, G., AND R. BRO (2006): "A comparison of algorithms for fitting the Parafac model," *Computational Statistics & Data Analysis*, 50, 1700–1734.

TRENDAFILOV, N. T. (2003): "Dynamical system approach to factor analysis parameter estimation," *British Journal of Mathematical and Statistical Psychology*, 56, 27–46.

——— (2005): "Fitting the factor analysis model in $\ell_1$ norm," *British Journal of Mathematical and Statistical Psychology*, 58, 19–31.

——— (2006): "Dynamical system approach to multivariate data analysis," *Journal of Computational and Graphical Statistics*, 15, 628–650.

TUCKER, L. R. (1966): "Some mathematical notes on three-mode factor analysis," *Psychometrika*, 31, 279–311.

TUKEY, P. A., AND J. W. TUKEY (1981): "Preparation; prechosen sequences of views," in *Interpreting Multivariate Data*, ed. by V. Barnett, pp. 189–213. John Wiley & Sons: Chichester.

UNKEL, S., AND N. T. TRENDAFILOV (2007): "Noisy independent component analysis as a method of rotating the factor scores," in *Proceedings of the 7th International Conference on Independent Component Analysis and Signal Separation (ICA 2007)*, ed. by M. E. Davies, C. J. James, S. A. Abdallah, and M. D. Plumbley, pp. 810–817. Springer: Berlin, Heidelberg.

——— (2009a): "Exploratory factor analysis of data matrices with more variables than observations," Submitted.

——— (2009b): "Factor analysis as data matrix decomposition: A new approach for quasi-sphering in noisy ICA," in *Proceedings of the 8th International Conference on Independent Component Analysis and Signal Separation (ICA 2009)*, ed. by T. Adali, C. Jutten, J. M. T. Romano, and A. K. Barros, pp. 163–170. Springer: Berlin, Heidelberg.

——— (2009c): "A majorization algorithm for simultaneous parameter estimation in robust exploratory factor analysis," *Computational Statistics & Data Analysis*, under revision.

UNKEL, S., N. T. TRENDAFILOV, A. HANNACHI, AND I. T. JOLLIFFE (2009): "Independent exploratory factor analysis with application to atmospheric science data," *Journal of Applied Statistics*, to appear.

VERBOON, P. (1994): *A Robust Approach to Nonlinear Multivariate Analysis.* DSWO Press: Leiden.

VERBOON, P., AND W. J. HEISER (1992): "Resistant orthogonal procrustes analysis," *Journal of Classification*, 9, 237–256.

WALLACE, J. M., AND D. S. GUTZLER (1981): "Teleconnections in the geopotential height field during the northern hemisphere winter," *Monthly Weather Review*, 109, 784–812.

WALSH, J. E., AND M. B. RICHMAN (1981): "Seasonality in the associations between surface temperature over the United States and the North Pacific Ocean," *Monthly Weather Review*, 109, 767–783.

WALSH, J. E., M. B. RICHMAN, AND D. W. ALLEN (1982): "Spatial coherence of monthly precipitation in the United States," *Monthly Weather Review*, 110, 272–286.

WHITTLE, P. (1952): "On principal components and least squares methods in factor analysis," *Skandinavisk Aktuarietidskrift*, 35, 223–239.

WOLD, H. (1966): "Nonlinear estimation by iterative least squares procedures," in *Research papers in statistics: Festschrift for Jerzy Newman*, ed. by F. N. David, pp. 411–444. John Wiley & Sons: New York.

YATES, A. (1987): *Multivariate Exploratory Data Analysis: A Perspective on Exploratory Factor Analysis.* State University of New York: Albany.

YOUNG, G. (1941): "Maximum likelihood estimation and factor analysis," *Psychometrika*, 6, 49–53.