Exploratory factor analysis of data matrices with more variables than observations

Nickolay T. Trendafilov * and Steffen Unkel

Department of Mathematics and Statistics

The Open University

Milton Keynes, UK

July 3, 2010

Abstract

A new approach for exploratory factor analysis (EFA) of data matrices with more variables p than observations n is presented. First, the classic EFA model ($n > p$) is considered as a specific data matrix decomposition with fixed unknown matrix parameters. Then, it is generalized to a new model, called for short GEFA, which covers both cases of data, with either $n > p$ or $p \geq n$. An alternating least squares algorithm **GEFALS** is proposed for simultaneous estimation of all GEFA model parameters. As principal component analysis (PCA), **GEFALS** is based on singular value decomposition, which makes GEFA an attractive alternative to PCA for descriptive data analysis and dimensionality reduction. The existence and uniqueness of the GEFA parameters estimations is studied and the convergence properties of **GEFALS** are established. Finally, the new approach is illustrated with Thurstone's 26-variable box data and real high-dimensional data, while the performance of **GEFALS** – with simulation experiment.

Key words: Factor analysis, $p > n$, Procrustes problem, Singular value decomposition.

*Corresponding author (e-mail:N.Trendafilov@open.ac.uk)

1 Introduction

Exploratory factor analysis (EFA) is a model that 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 (Bartholomew and Knott 1999; Harman 1976; Mulaik 1972; Thurstone 1947).

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 covariance or 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.

Unlike the EFA factorization of the correlation matrix, fitting the EFA model directly to the data yields factor loadings and common factor scores *simultaneously* (Horst 1965; Jöreskog 1962; Lawley 1942; McDonald 1979; Whittle 1952). A brief summary of this class of EFA methods is given in De Leeuw (2004), and their weaknesses are reviewed. The early works had difficulties to define properly the EFA data fitting problem: they ended up with an unbounded log-likelihood function. The least squares (LS) approach adopted by Horst (1965) did principal component analysis (PCA), instead of EFA. The most recent and promising approach was proposed by McDonald (1979), however its "computational and statistical properties...are quite complicated..." (De Leeuw 2004). Particularly, McDonald (1979) employed the Guttman's approach to find common and unique factors, however the latter ones require strictly positive unique variances. As a remedy, De Leeuw (2004) proposed simultaneous estimation of *all* EFA model unknowns by optimizing a LS loss function, designed for the classical case of 'vertical' data with $n > p$.

However, in many modern applications, the number of observations is less than the number of variables. For example, data arising from experiments in genome research are usually in the form of large horizontal matrices 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 paper extends the approach of De Leeuw (2004) for EFA of 'horizontal' data matrices with $p \geq n$. The EFA model is considered as a specific data matrix decomposition with fixed unknown matrix parameters. New assumptions are imposed on the EFA model parameters which necessarily require the acceptance of unique factors with zero variance.

As in PCA (Jolliffe 2002) , the new EFA parameters estimation is based on the singular value decomposition (SVD) (Golub and Van Loan 1996). Thus, the new approach to EFA makes it computationally competitive descriptive technique for data analysis along with PCA.

The paper is organized as follows. The next Section briefly reviews the standard EFA models with random and fixed common factors. Section 3 outlines approaches for fitting the EFA model to vertical data. In Section 4, a generalization of the EFA model is introduced, called GEFA hereafter, to cope with both vertical and horizontal data. Then, new numerical procedure **GEFALS** for simultaneous estimation of all GEFA unknowns is proposed. Section 5 deals with the existence and uniqueness of the GEFA estimators and the **GEFALS** convergence properties. In Section 6 **GEFALS** is applied to the Thurstone's 26-variable box data (Thurstone 1947) and to a real high-dimensional data from atmospheric science. Section 7 concludes the paper by summarizing the main findings.

2 The EFA models

Let $\mathbf{z} \in \mathbb{R}^{p\times 1}$ be a random vector of standardized manifest variables. Suppose that the EFA model holds which states that $\mathbf{z}$ can be written in the form (e.g., Mulaik 1972, p. 97):

$$\mathbf{z} = \mathbf{\Lambda f} + \mathbf{\Psi u} \ , \tag{1}$$

where $\mathbf{f} \in \mathbb{R}^{k\times 1}$ is a random vector of k ($k \ll p$) common factors, $\mathbf{\Lambda} \in \mathbb{R}^{p\times k}$ is a matrix of fixed factor loadings, $\mathbf{u} \in \mathbb{R}^{p\times 1}$ is a random vector of unique factors and $\mathbf{\Psi}$ is a $p \times p$ diagonal matrix of fixed coefficients called uniquenesses. The choice of k in EFA is subject to some limitations (e.g., Mulaik 1972, p.138), which will not be discussed here.

Assume that $\mathrm{E}(\mathbf{f}) = \mathbf{O}_{k\times 1}$, $\mathrm{E}(\mathbf{u}) = \mathbf{O}_{p\times 1}$, $\mathrm{E}(\mathbf{uu}^\top) = \mathbf{I}_p$ and $\mathrm{E}(\mathbf{fu}^\top) = \mathbf{O}_{k\times p}$, where $\mathbf{I}_p$ is an identity matrix of order p and $\mathbf{O}_{p\times k}$ is a $p \times k$ matrix of zeros. Furthermore, let $\mathrm{E}(\mathbf{zz}^\top) = \mathbf{\Sigma}$

and $\mathrm{E}(\mathbf{f}\mathbf{f}^\top) = \mathbf{\Phi}$ be correlation matrices, i.e. positive definite (p.d.) matrices with unit main diagonals. Then, the k-model (1) and the assumptions made imply that

$$\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}^\top + \mathbf{\Psi}^2 \ . \tag{2}$$

The correlated common factors are called oblique. In the sequel, it is assumed that the common factors are uncorrelated (orthogonal), that is, $\mathbf{\Phi} = \mathbf{I}_k$. Thus, (2) reduces to

$$\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^\top + \mathbf{\Psi}^2 \ . \tag{3}$$

Unlike the random EFA model (1), the fixed EFA model considers $\mathbf{f}$ to be a vector of non-random parameters which vary from one case to another (Lawley 1942).

3 Fitting the EFA model in the classical case ($n > p$)

Let $\mathbf{Z}$ be the $n \times p$ data matrix collecting n independent centered observations on $p(< n)$ variables with unit length. According to (1), let $\mathbf{F}$ and $\mathbf{U}$ denote the matrices of common and unique factors, respectively. Then, the k-factor model (1) and the related assumptions from Section 2 imply that EFA represents the data $\mathbf{Z}$ as follows:

$$\mathbf{Z} = \mathbf{F}\mathbf{\Lambda}^\top + \mathbf{U}\mathbf{\Psi} \ , \tag{4}$$

$$\text{subject to} \quad \mathbf{F}^\top\mathbf{F} = \mathbf{I}_k, \ \mathbf{U}^\top\mathbf{U} = \mathbf{I}_p, \ \mathbf{U}^\top\mathbf{F} = \mathbf{O}_{p\times k} \text{ and } \mathbf{\Psi} \text{ diagonal.} \tag{5}$$

The EFA data representation (4)–(5) implies the following representation of the sample correlation matrix:

$$\mathbf{Z}^\top\mathbf{Z} = \mathbf{\Lambda}\mathbf{\Lambda}^\top + \mathbf{\Psi}^2 \ . \tag{6}$$

The equation (6) is fundamental for the standard EFA (with random common factors), where a pair $\{\mathbf{\Lambda}, \mathbf{\Psi}\}$ is sought which gives the best fit to $\mathbf{Z}^\top\mathbf{Z}$ with respect to some discrepancy measure. The process of finding $\{\mathbf{\Lambda}, \mathbf{\Psi}\}$, for some specified value of k, 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 certain log-likelihood function equivalent to the following data fitting problem (Mulaik 1972, p. 163):

$$\min_{\mathbf{\Lambda},\mathbf{\Psi}} \log(\det(\mathbf{\Lambda}\mathbf{\Lambda}^\top + \mathbf{\Psi}^2)) + \text{trace}((\mathbf{\Lambda}\mathbf{\Lambda}^\top + \mathbf{\Psi}^2)^{-1}(\mathbf{Z}^\top\mathbf{Z})) \ , \tag{7}$$

referred to as maximum likelihood (ML) factor analysis. If nothing is assumed about the distribution of the data, (7) 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 (7). A natural choice is the LS approach for fitting the EFA model. It can be formulated as the following general class of weighted LS problems (Bartholomew and Knott 1999, pp. 53–56):

$$\min_{\mathbf{\Lambda},\mathbf{\Psi}} ||(\mathbf{Z}^\top\mathbf{Z} - \mathbf{\Lambda}\mathbf{\Lambda}^\top - \mathbf{\Psi}^2)\mathbf{\Gamma}||_F^2 \ , \tag{8}$$

where$||\mathbf{X}||_F = \sqrt{\text{trace}(\mathbf{X}^\top\mathbf{X})}$ denotes the Frobenius norm of a matrix $\mathbf{X}$ and $\mathbf{\Gamma}$ is a matrix of weights. The case $\mathbf{\Gamma} = \mathbf{\Sigma}^{-1}$ is known as generalized least squares (GLS) factor analysis. If $\mathbf{\Gamma} = \mathbf{I}_p$, (8) reduces to an unweighted least squares (ULS) factor analysis. The standard numerical solutions of the ML, ULS and GLS factor analysis are iterative, usually based on a gradient or Newton-Raphson procedure.

Suppose that a pair $\{\mathbf{\Lambda}, \mathbf{\Psi}\}$ is obtained by solving the factor extraction problem. Then, common factor scores can be computed as a function of $\mathbf{Z}$, $\mathbf{\Lambda}$ and possibly $\mathbf{\Psi}$ in a number of ways (Harman 1976; Mulaik 1972).

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. Instead, De Leeuw (2004) formulated the EFA model (4) as a specific data matrix decomposition. Then, the EFA problem is to minimize the following least squares goodness-of-fit criterion:

$$f(\mathbf{F}, \mathbf{\Lambda}, \mathbf{U}, \mathbf{\Psi}) \ = \ ||\mathbf{Z} - \mathbf{F}\mathbf{\Lambda}^\top - \mathbf{U}\mathbf{\Psi}||_F^2 \ , \tag{9}$$

$$\text{subject to} \qquad \mathbf{F}^\top\mathbf{F} = \mathbf{I}_k, \ \mathbf{U}^\top\mathbf{U} = \mathbf{I}_p, \ \mathbf{U}^\top\mathbf{F} = \mathbf{O}_{p\times k} \text{ and } \mathbf{\Psi} \text{ diagonal.} \tag{10}$$

De Leeuw (2004) proposed an alternating least squares (ALS) algorithm to minimize (9) – (10). 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 simultaneously by solving a Procrustes problem. Indeed, let $\mathbf{B} = [\mathbf{F} \;\; \mathbf{U}]$ and $\mathbf{A} = [\mathbf{\Lambda} \;\; \mathbf{\Psi}]$ be block matrices with dimensions $n \times (k+p)$ and $p \times (k+p)$. Note that (10) implies $\mathbf{B}^\top \mathbf{B} = \mathbf{I}_{p+k}$. Then (9) can be rewritten as:

$$f = \left|\left|\mathbf{Z} - \mathbf{B}\mathbf{A}^\top\right|\right|_F^2 = ||\mathbf{Z}||_F^2 + \text{trace}(\mathbf{A}\mathbf{B}^\top\mathbf{B}\mathbf{A}^\top) - 2\ \text{trace}(\mathbf{B}^\top\mathbf{Z}\mathbf{A})\ . \tag{11}$$

Minimizing (11) subject to $\mathbf{B}$ orthonormal is a standard Procrustes problem. The minimization of f in (11) is equivalent to the maximization of trace($\mathbf{B}^\top\mathbf{Z}\mathbf{A}$) (for given or estimated $\mathbf{A}$). Solution of this maximization problem is given by $\mathbf{B} = \mathbf{Q}\mathbf{P}^\top$, where $\mathbf{Z}\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{P}^\top$ is the 'economy' SVD (33) of $\mathbf{Z}\mathbf{A}$. As shown in the Appendix, this solution is not unique for $n > p$, as $\text{rank}(\mathbf{Z}\mathbf{A}) \leq \min\{\text{rank}(\mathbf{Z}), \text{rank}(\mathbf{A})\} = p < k + p$.

Once an orthonormal $\mathbf{B}$ is found for given or estimated $\mathbf{A}(= [\mathbf{\Lambda} \;\; \mathbf{\Psi}])$, i.e. new $\mathbf{F}$ and $\mathbf{U}$ are available, update $\mathbf{\Lambda} = \mathbf{Z}^\top\mathbf{F}$ and $\mathbf{\Psi} = \text{diag}(\mathbf{U}^\top\mathbf{Z})$, making use of the identities:

$$\mathbf{F}^\top\mathbf{Z} = \mathbf{F}^\top\mathbf{F}\mathbf{\Lambda}^\top + \mathbf{F}^\top\mathbf{U}\mathbf{\Psi} = \mathbf{\Lambda}^\top, \tag{12}$$

$$\mathbf{U}^\top\mathbf{Z} = \mathbf{U}^\top\mathbf{F}\mathbf{\Lambda}^\top + \mathbf{U}^\top\mathbf{U}\mathbf{\Psi} = \mathbf{\Psi} \text{ (and thus diagonal)}\ , \tag{13}$$

which follow from the EFA model (4) and the imposed constraints. The ALS procedure of finding $\{\mathbf{F}, \mathbf{U}\}$ and $\{\mathbf{\Lambda}, \mathbf{\Psi}\}$ continues until certain convergence criterion is met. It can be summarized in the following algorithm **EFALS**:

$\mathbf{F} \leftarrow \texttt{rand}(n, k) - .5\ ,\ \mathbf{U} \leftarrow \texttt{rand}(n, p) - .5$;

$\mathbf{\Lambda} \leftarrow \mathbf{Z}^\top\mathbf{F}\ ,\ \mathbf{\Psi} \leftarrow \text{diag}(\mathbf{U}^\top\mathbf{Z})\ ,\ \mathbf{A} \leftarrow [\mathbf{\Lambda}\ \mathbf{\Psi}]$;

$f_{old} = ||\mathbf{Z}||_F^2, f = ||\mathbf{Z} - \mathbf{B}\mathbf{A}^\top||_F^2$;

while $|f_{old} - f| > 10^{-6}$

$\mathbf{B} \leftarrow \mathbf{Q}\mathbf{P}^\top$, where $\mathbf{Z}\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{P}^\top$ is the economy SVD of $\mathbf{Z}\mathbf{A}$;

$\mathbf{F} \leftarrow \mathbf{B}(:, 1:k)$;

$\mathbf{U} \leftarrow \mathbf{B}(:, k+1:k+p)$;

$\mathbf{\Lambda} \leftarrow \mathbf{Z}^\top\mathbf{F}\ ,\ \mathbf{\Psi} \leftarrow \text{diag}(\mathbf{U}^\top\mathbf{Z})\ ,\ \mathbf{A} \leftarrow [\mathbf{\Lambda}\ \mathbf{\Psi}]$;

$f_{old} = f, f = ||\mathbf{Z} - \mathbf{B}\mathbf{A}^\top||_F^2$;

end while

Assuming that $\text{rank}(\mathbf{Z}) \geq k$, the **EFALS** update of $\mathbf{\Lambda}$ $(= \mathbf{Z}^\top \mathbf{F})$ gives in general $\text{rank}(\mathbf{\Lambda}) \leq \min\{\text{rank}(\mathbf{Z}), \text{rank}(\mathbf{F})\} = k$. It will be shown in Section 5 that, in fact, the updating formula $\mathbf{\Lambda} = \mathbf{Z}^\top \mathbf{F}$ secures that $\text{rank}(\mathbf{\Lambda}) = k$.

4 Generalized EFA for both $p < n$ and $p \geq n$

If $p \geq n$, 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 rank-deficient sample correlation matrix by the EFA correlation structure (6). They try to approximate a singular symmetric matrix $\mathbf{Z}^\top \mathbf{Z}$ by a positive definite one having the specific form $\mathbf{\Lambda}\mathbf{\Lambda}^\top + \mathbf{\Psi}^2$ imposed by EFA and assuming $\mathbf{\Psi}^2$ p.d. Alternatively, one can employ ULS factor analysis (8).

However, there is a conceptual difficulty to adopt these two approaches to EFA. When $p > n$, the rank of $\mathbf{U}$ can be at most n and the constraint $\mathbf{U}^\top \mathbf{U} = \mathbf{I}_p$ cannot be fulfilled any more. The rest of the classic EFA constraints $\mathbf{F}^\top \mathbf{F} = \mathbf{I}_k$ and $\mathbf{U}^\top \mathbf{F} = \mathbf{O}_{p \times k}$ remain valid when $p \geq n$. Moreover, they imply that $\text{rank}(\mathbf{U}) \leq n - k$ (Horn and Johnson 1986, 0.4.5 (c)). Thus, the classical EFA correlation structure (6) turns into

$$\mathbf{Z}^\top \mathbf{Z} = \mathbf{\Lambda}\mathbf{\Lambda}^\top + \mathbf{\Psi}\mathbf{U}^\top \mathbf{U}\mathbf{\Psi} \; . \tag{14}$$

For $p > n$, the correlation structure (14) coincides with the classical one (6), if the more general constraint $\mathbf{U}^\top \mathbf{U}\mathbf{\Psi} = \mathbf{\Psi}$ is introduced in place of $\mathbf{U}^\top \mathbf{U} = \mathbf{I}_p$.

Lemma 4.1. *If $p > n$, $\boldsymbol{U}^\top \boldsymbol{U}\mathbf{\Psi} = \mathbf{\Psi}$ implies that $\mathbf{\Psi}^2$ is positive semi-definite (p.s.d.)*

Proof : A product of symmetric and diagonal matrix with non-zero entries can be diagonal only if the symmetric matrix is diagonal too. If it is assumed that $\mathbf{\Psi}^2 > \mathbf{O}$, then $\mathbf{U}^\top \mathbf{U}\mathbf{\Psi} = \mathbf{\Psi}$ implies $\mathbf{U}^\top \mathbf{U} = \mathbf{I}_p$, which contradicts to $\text{rank}(\mathbf{U}^\top \mathbf{U}) \leq n - k$. Thus, $\mathbf{\Psi}$ should have zero entries, i.e. $\mathbf{\Psi}^2 \geq \mathbf{O}$. ■

Lemma 4.1 shows that if $p > n$, then variables with zero unique variances necessarily exist, i.e. unique factors with zero variances should be accepted in the EFA model. This proves that the EFA modification by Robertson and Symons (2007) requiring $\mathbf{\Psi}^2$ p.d. for the case $p \geq n$, is in fact not consistent with the EFA model (4).

Let r denote the number of the zero entries in $\mathbf{\Psi}$, i.e. $\mathrm{rank}(\mathbf{\Psi}) = p - r$. Assume for simplicity, that the corresponding r variables with zero unique variances are the first r variables. Then $\mathbf{U}$ can be partitioned as $\mathbf{U} = [\mathbf{U}_1\ \mathbf{U}_2]$, where $\mathbf{U}_1$ and $\mathbf{U}_2$ are block-matrices of sizes $n \times r$ and $n \times (p-r)$, respectively. Similarly $\mathbf{\Psi} = \begin{bmatrix} \mathbf{\Psi}_1 & \mathbf{O}_{r\times(p-r)} \\ \mathbf{O}_{(p-r)\times r} & \mathbf{\Psi}_2 \end{bmatrix}$, where $\mathbf{\Psi}_1 \equiv \mathbf{O}_r$. Then, the condition $\mathbf{U}^\top\mathbf{U}\mathbf{\Psi} = \mathbf{\Psi}$ can be rewritten as:

$$\begin{aligned}
\mathbf{U}^\top\mathbf{U}\mathbf{\Psi} &= \begin{bmatrix} \mathbf{U}_1^\top \\ \mathbf{U}_2^\top \end{bmatrix} \begin{bmatrix} \mathbf{U}_1 & \mathbf{U}_2 \end{bmatrix} \begin{bmatrix} \mathbf{\Psi}_1 & \mathbf{O}_{r\times(p-r)} \\ \mathbf{O}_{(p-r)\times r} & \mathbf{\Psi}_2 \end{bmatrix}, \\
&= \begin{bmatrix} \mathbf{U}_1^\top\mathbf{U}_1 & \mathbf{U}_1^\top\mathbf{U}_2 \\ \mathbf{U}_2^\top\mathbf{U}_1 & \mathbf{U}_2^\top\mathbf{U}_2 \end{bmatrix} \begin{bmatrix} \mathbf{\Psi}_1 & \mathbf{O}_{r\times(p-r)} \\ \mathbf{O}_{(p-r)\times r} & \mathbf{\Psi}_2 \end{bmatrix}, \\
&= \begin{bmatrix} \mathbf{U}_1^\top\mathbf{U}_1\mathbf{\Psi}_1 & \mathbf{U}_1^\top\mathbf{U}_2\mathbf{\Psi}_2 \\ \mathbf{U}_2^\top\mathbf{U}_1\mathbf{\Psi}_1 & \mathbf{U}_2^\top\mathbf{U}_2\mathbf{\Psi}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{\Psi}_1 & \mathbf{O}_{r\times(p-r)} \\ \mathbf{O}_{(p-r)\times r} & \mathbf{\Psi}_2 \end{bmatrix} = \mathbf{\Psi}. \qquad (15)
\end{aligned}$$

By construction $\mathbf{\Psi}_1 \equiv \mathbf{O}_r$, and the constraint (15) reduces to: $\mathbf{U}_1^\top\mathbf{U}_2\mathbf{\Psi}_2 = \mathbf{O}_{r\times(p-r)}$ and $\mathbf{U}_2^\top\mathbf{U}_2\mathbf{\Psi}_2 = \mathbf{\Psi}_2$. As $\mathbf{\Psi}_2^2 > 0$, they imply that $\mathbf{U}_1^\top\mathbf{U}_2 = \mathbf{O}_{r\times(p-r)}$ and $\mathbf{U}_2^\top\mathbf{U}_2 = \mathbf{I}_{p-r}$. As $\mathbf{U}_2$ is a submatrix of $\mathbf{U}$ and $\mathrm{rank}(\mathbf{U}) \leq n - k$, it also follows that $p - r \leq n - k$. Then one finds that $r \geq p - n + k > k$, i.e. the number of the variables with zero unique variances will always be greater than the chosen common factors number, if $p > n$. Moreover, from the factor analysis equation (4) and assuming that $\mathrm{rank}(\mathbf{Z}) = n - 1$ as $\mathbf{Z}$ is centered, one finds that (Horn and Johnson 1986, 0.4.5 (d)):

$$n - 1 \leq \mathrm{rank}(\mathbf{F}\mathbf{\Lambda}^\top + \mathbf{U}\mathbf{\Psi}) \leq \mathrm{rank}(\mathbf{F}\mathbf{\Lambda}^\top) + \mathrm{rank}(\mathbf{U}\mathbf{\Psi}) \leq k + p - r\ ,$$

i.e. $r \leq p - n + k + 1$. This shows that for given n, p and k, the number of the variables with zero unique variances may take only two values, as $r \in [p - n + k, p - n + k + 1]$. Of course, if $\mathrm{rank}(\mathbf{Z}) < n - 1$, then the number of zeros r may increase.

There is a long standing debate in classical EFA ($n > p$) about the acceptance of zero entries in $\mathbf{\Psi}^2$ commonly referred to as 'Heywood' cases. Some authors argue that in such situations the Heywood case variable is explained entirely by the corresponding common factor (e.g., Bartholomew and Knott 1999, p. 61), while others find it unrealistic (e.g., Anderson 1984, pp. 561–562), and require $\mathbf{\Psi}^2$ to be p.d. It turns out, that an EFA model covering both cases $p \geq n$ and $n > p$ should accept $\mathbf{\Psi}^2$ being p.s.d. The following generalization of

the classical EFA model (4) – (5):

$$\mathbf{Z} = \mathbf{F}\boldsymbol{\Lambda}^\top + \mathbf{U}\boldsymbol{\Psi} \ , \tag{16}$$

$$\text{subject to} \quad \mathbf{F}^\top\mathbf{F} = \mathbf{I}_k,\ \mathbf{U}^\top\mathbf{U}\boldsymbol{\Psi} = \boldsymbol{\Psi},\ \mathbf{U}^\top\mathbf{F} = \mathbf{O}_{p\times k} \text{ and } \boldsymbol{\Psi} \text{ diagonal} \ , \tag{17}$$

covers both cases $p \geq n$ and $n > p$. From now on, it is called for short the GEFA model. Note, that GEFA implies the same identities as for the classical case ($n > p$):

$$\mathbf{F}^\top\mathbf{Z} = \mathbf{F}^\top\mathbf{F}\boldsymbol{\Lambda}^\top + \mathbf{F}^\top\mathbf{U}\boldsymbol{\Psi} = \boldsymbol{\Lambda}^\top, \tag{18}$$

$$\mathbf{U}^\top\mathbf{Z} = \mathbf{U}^\top\mathbf{F}\boldsymbol{\Lambda}^\top + \mathbf{U}^\top\mathbf{U}\boldsymbol{\Psi} = \boldsymbol{\Psi} \text{ (and thus diagonal)} \ . \tag{19}$$

Lemma 4.2. *If $n < p+k$, the $k-$factor GEFA constraints $\boldsymbol{F}^\top\boldsymbol{F} = \boldsymbol{I}_k$ and $\boldsymbol{U}^\top\boldsymbol{F} = \boldsymbol{O}_{p\times k}$ are equivalent to the constraint $\boldsymbol{F}\boldsymbol{F}^\top + \boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{I}_n$ and rank$(\boldsymbol{F}) = k$.*

Proof : According to $\mathbf{U}^\top\mathbf{F} = \mathbf{O}_{p\times k}$, the columns of $\mathbf{U}$ should be in the nullspace of $\mathbf{F}$ in $\mathbb{R}^n$. Denote by $\mathbf{F}_\perp$ the $n \times (n-k)$ matrix containing an orthonormal basis of the nullspace of $\mathbf{F}$ in $\mathbb{R}^n$, where $\mathbf{F}_\perp$ can be found by the QR factorization of $\mathbf{F}$:

$$\mathbf{F} = \mathbf{Q}\mathbf{R} = \underbrace{\begin{bmatrix} \mathbf{F} & \mathbf{F}_\perp \end{bmatrix}}_{\mathbf{Q}} \underbrace{\begin{bmatrix} \mathbf{I}_k \\ \mathbf{O}_{(n-k)\times k} \end{bmatrix}}_{\mathbf{R}} \ , \tag{20}$$

where $\mathbf{Q}$ is orthogonal and $\mathbf{R}$ is an upper triangular matrix. Then, there exists a full row-rank $(n-k)\times p$ matrix $\widetilde{\mathbf{U}}$ such that $\mathbf{U} = \mathbf{F}_\perp\widetilde{\mathbf{U}}$. As $\widetilde{\mathbf{U}}$ can be chosen orthonormal, i.e. $\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top = \mathbf{I}_{n-k}$, then it follows that $\mathbf{U}\mathbf{U}^\top = \mathbf{F}_\perp\mathbf{F}_\perp^\top$, and $\mathbf{F}\mathbf{F}^\top + \mathbf{F}_\perp\mathbf{F}_\perp^\top = \mathbf{I}_n$ implies $\mathbf{F}\mathbf{F}^\top + \mathbf{U}\mathbf{U}^\top = \mathbf{I}_n$. In other words, the constraints $\mathbf{F}^\top\mathbf{F} = \mathbf{I}_k$ and $\mathbf{U}^\top\mathbf{F} = \mathbf{O}_{p\times k}$ can be combined into a new constraint $\mathbf{F}\mathbf{F}^\top + \mathbf{U}\mathbf{U}^\top = \mathbf{I}_n$.

The converse is also true. Indeed, $\mathbf{U}\mathbf{U}^\top = \mathbf{I}_n - \mathbf{F}\mathbf{F}^\top = \mathbf{F}_\perp\mathbf{F}_\perp^\top$ gives that

$$\mathbf{F}^\top - \mathbf{F}^\top\mathbf{F}\mathbf{F}^\top = (\mathbf{I}_k - \mathbf{F}^\top\mathbf{F})\mathbf{F}^\top = \mathbf{F}^\top\mathbf{F}_\perp\mathbf{F}_\perp^\top = \mathbf{O}_{k\times n} \ ,$$

and

$$\mathbf{U}^\top - \mathbf{U}^\top\mathbf{F}\mathbf{F}^\top = \mathbf{U}^\top\mathbf{F}_\perp\mathbf{F}_\perp^\top = \widetilde{\mathbf{U}}^\top\mathbf{F}_\perp^\top\mathbf{F}_\perp\mathbf{F}_\perp^\top = \widetilde{\mathbf{U}}^\top\mathbf{F}_\perp^\top = \mathbf{U}^\top \ ,$$

i.e. $(\mathbf{I}_k - \mathbf{F}^\top\mathbf{F})\mathbf{F}^\top = \mathbf{O}_{k\times n}$ and $\mathbf{U}^\top\mathbf{F}\mathbf{F}^\top = \mathbf{O}_{p\times n}$, which imply that $\mathbf{I}_n - \mathbf{F}^\top\mathbf{F} = \mathbf{O}_k$ and $\mathbf{U}^\top\mathbf{F} = \mathbf{O}_{p\times k}$ as $\mathbf{F}^\top$ has full row rank. ■

Lemma 4.3. ***FF***$^\top$ + ***UU***$^\top$ = ***I***$_n$ *does not imply* ***U***$^\top$***UΨ*** = **Ψ**.

PROOF : The new constraint $\mathbf{F}\mathbf{F}^\top + \mathbf{U}\mathbf{U}^\top = \mathbf{I}_n$ simply gives the identity $\mathbf{U} = \mathbf{U}\mathbf{U}^\top\mathbf{U}$, which multiplied by $\mathbf{\Psi}$ leads to

$$\mathbf{U}\mathbf{\Psi} = \mathbf{U}\mathbf{U}^\top\mathbf{U}\mathbf{\Psi} \Rightarrow \mathbf{U}(\mathbf{\Psi} - \mathbf{U}^\top\mathbf{U}\mathbf{\Psi}) = \mathbf{O}_{n\times p} \ . \tag{21}$$

Since $\mathbf{U}$ has not full column rank, (21) demonstrates that the constraint $\mathbf{U}^\top\mathbf{U}\mathbf{\Psi} = \mathbf{\Psi}$ does not necessarily follow from $\mathbf{F}\mathbf{F}^\top + \mathbf{U}\mathbf{U}^\top = \mathbf{I}_n$ and must be imposed *separately.* ■

Thus, the GEFA model (16) – (17) requires the minimization of:

$$f(\mathbf{F}, \mathbf{\Lambda}, \mathbf{U}, \mathbf{\Psi}) \ = \ ||\mathbf{Z} - \mathbf{F}\mathbf{\Lambda}^\top - \mathbf{U}\mathbf{\Psi}||_F^2 \ , \tag{22}$$

$$\text{subject to} \qquad \text{rank}(\mathbf{F}) = k, \ \mathbf{F}\mathbf{F}^\top + \mathbf{U}\mathbf{U}^\top = \mathbf{I}_n \ , \mathbf{U}^\top\mathbf{U}\mathbf{\Psi} = \mathbf{\Psi}(\text{diagonal}). \tag{23}$$

Let $\mathbf{B} = [\mathbf{F} \ \ \mathbf{U}]$ and $\mathbf{A} = [\mathbf{\Lambda} \ \ \mathbf{\Psi}]$ be block matrices with dimensions $n \times (k + p)$ and $p \times (k + p)$. Then, for given $\mathbf{A}$ the following problem:

$$\min_{\mathbf{B}} \left|\left|\mathbf{Z} - \mathbf{B}\mathbf{A}^\top\right|\right|_F^2 \ , \text{ subject to } \mathbf{B}\mathbf{B}^\top = \mathbf{I}_n \ , \tag{24}$$

is a standard Procrustes problem, i.e. its solution is found by maximizing trace($\mathbf{B}^\top\mathbf{Z}\mathbf{A}$). The loss function $\left|\left|\mathbf{Z} - \mathbf{B}\mathbf{A}^\top\right|\right|_F^2 = ||\mathbf{Z}||_F^2 + \text{trace}(\mathbf{B}^\top\mathbf{B}\mathbf{A}^\top\mathbf{A}) - 2\ \text{trace}(\mathbf{B}^\top\mathbf{Z}\mathbf{A})$ contains two terms depending on $\mathbf{B}$, because $\mathbf{B}^\top\mathbf{B}$ is not an identity matrix as in (11). Nevertheless, making use of the constraint $\mathbf{U}^\top\mathbf{U}\mathbf{\Psi} = \mathbf{\Psi}$ (diagonal), one can see that

$$\begin{aligned}
\text{trace}(\mathbf{B}^\top\mathbf{B}\mathbf{A}^\top\mathbf{A}) \ &= \ \text{trace}\left\{ \begin{bmatrix} \mathbf{F}^\top \\ \mathbf{U}^\top \end{bmatrix} \begin{bmatrix} \mathbf{F} & \mathbf{U} \end{bmatrix} \begin{bmatrix} \mathbf{\Lambda}^\top \\ \mathbf{\Psi} \end{bmatrix} \begin{bmatrix} \mathbf{\Lambda} & \mathbf{\Psi} \end{bmatrix} \right\} \ , \\
&= \ \text{trace}\left\{ \begin{bmatrix} \mathbf{I}_k & \mathbf{O}_{k\times p} \\ \mathbf{O}_{p\times k} & \mathbf{U}^\top\mathbf{U} \end{bmatrix} \begin{bmatrix} \mathbf{\Lambda}^\top\mathbf{\Lambda} & \mathbf{\Lambda}^\top\mathbf{\Psi} \\ \mathbf{\Psi}\mathbf{\Lambda} & \mathbf{\Psi}^2 \end{bmatrix} \right\} \ , \\
&= \ \text{trace}\left\{ \begin{bmatrix} \mathbf{\Lambda}^\top\mathbf{\Lambda} & \mathbf{\Lambda}^\top\mathbf{\Psi} \\ \mathbf{U}^\top\mathbf{U}\mathbf{\Psi}\mathbf{\Lambda} & \mathbf{U}^\top\mathbf{U}\mathbf{\Psi}^2 \end{bmatrix} \right\} \ , \\
&= \ \text{trace}(\mathbf{\Lambda}^\top\mathbf{\Lambda}) + \text{trace}(\mathbf{\Psi}^2) \ ,
\end{aligned} \tag{25}$$

i.e. trace($\mathbf{B}^\top\mathbf{B}\mathbf{A}^\top\mathbf{A}$) does not depend on $\mathbf{F}$ and $\mathbf{U}$. Thus, the solution of (24) is equivalent to the maximization of trace($\mathbf{B}^\top\mathbf{Z}\mathbf{A}$) and simply requires the 'economy' SVD of $\mathbf{A}^\top\mathbf{Z}^\top$. After solving (24) for $\mathbf{B} = [\mathbf{F} \ \ \mathbf{U}]$, the values of $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ are updated making use of the identities

(18) and (19). The ALS procedure of finding $\{\mathbf{F}, \mathbf{U}\}$ and $\{\mathbf{\Lambda}, \mathbf{\Psi}\}$ continues until certain convergence criterion is met. It can be summarized in the following algorithm **GEFALS** containing **EFALS** as a subset:

$\mathbf{F} \leftarrow \texttt{rand}(n, k) - .5$, $\mathbf{U} \leftarrow \texttt{rand}(n, p) - .5$;

$\mathbf{\Lambda} \leftarrow \mathbf{Z}^\top \mathbf{F}$, $\mathbf{\Psi} \leftarrow \text{diag}(\mathbf{U}^\top \mathbf{Z})$, $\mathbf{A} \leftarrow [\mathbf{\Lambda}\ \mathbf{\Psi}]$;

$f_{old} = ||\mathbf{Z}||_F^2, f = ||\mathbf{Z} - \mathbf{B}\mathbf{A}^\top||_F^2$;

while $|f_{old} - f| > 10^{-6}$

if $n \geq p + k$

$\mathbf{B} \leftarrow \mathbf{Q}\mathbf{P}^\top$, where $\mathbf{Z}\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{P}^\top$ is the economy SVD of $\mathbf{Z}\mathbf{A}$;

else

$\mathbf{B} \leftarrow \mathbf{P}\mathbf{Q}^\top$, where $\mathbf{A}^\top\mathbf{Z}^\top = \mathbf{Q}\mathbf{D}\mathbf{P}^\top$ is the economy SVD of $\mathbf{A}^\top\mathbf{Z}^\top$;

endif

$\mathbf{F} \leftarrow \mathbf{B}(:, 1 : k)$;

$\mathbf{U} \leftarrow \mathbf{B}(:, k + 1 : k + p)$;

$\mathbf{\Lambda} \leftarrow \mathbf{Z}^\top \mathbf{F}$, $\mathbf{\Psi} \leftarrow \text{diag}(\mathbf{U}^\top \mathbf{Z})$, $\mathbf{A} \leftarrow [\mathbf{\Lambda}\ \mathbf{\Psi}]$;

$f_{old} = f, f = ||\mathbf{Z} - \mathbf{B}\mathbf{A}^\top||_F^2$;

end while

If the k-factor GEFA model holds then it also holds if the loadings are rotated. Let $\mathbf{T}$ be an arbitrary orthogonal $k \times k$ matrix. Then, (4) may be rewritten as

$$\mathbf{Z} = \mathbf{F}\mathbf{T}\mathbf{T}^\top\mathbf{\Lambda}^\top + \mathbf{U}\mathbf{\Psi} \ , \tag{26}$$

which is a model with loading matrix $\mathbf{\Lambda}\mathbf{T}$ and common factors $\mathbf{F}\mathbf{T}$. The assumptions on the EFA parameters are not violated by this transformation. Thus, any matrix $\mathbf{\Lambda}\mathbf{T}$ 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. Such kind of reparametrization is mentioned by Anderson and Rubin (1956) and is first employed for EFA by Trendafilov (2005). Clearly, if lower triangular loadings matrix $\mathbf{L}$ is adopted in the GEFA model, the only indeterminacy left is simultaneous changes of the signs of $\mathbf{L}$ and $\mathbf{F}$, which do not alter their interpretation and the model fit. Then, the original GEFA loss function (22) modifies to:

$$f(\mathbf{L}, \mathbf{\Psi}, \mathbf{F}, \mathbf{U}) = \left|\left| \mathbf{Z} - [\mathbf{F} \quad \mathbf{U}] \begin{bmatrix} \mathbf{L}^\top \\ \mathbf{\Psi} \end{bmatrix} \right|\right|_F^2 , \tag{27}$$

and is minimized subject to the GEFA constraints (23). **GEFALS** solves this modified problem by simply replacing the updating formula $\mathbf{\Lambda} \leftarrow \mathbf{Z}^\top\mathbf{F}$ with $\mathbf{L} \leftarrow \texttt{tril}(\mathbf{Z}^\top\mathbf{F})$, where `tril()` is the operator taking the lower triangular part of its argument.

5 Convergence properties of GEFALS

The GEFA objective function decreases at each **GEFALS** step. Thus **GEFALS** is globally convergent algorithm, that is, it converges from any starting value. Let the GEFA parameters indexed by $_0$ and $_+$ denote their old and updated values. Then, the solution $\mathbf{B}_+$ of the Procrustes problem (24), reduces the GEFA objective function as follows:

$$f(\mathbf{A}_0, \mathbf{B}_0) \;\; = \;\; ||\mathbf{Z} - \mathbf{B}_0\mathbf{A}_0^\top||_F^2 \geq ||\mathbf{Z} - \mathbf{B}_+\mathbf{A}_0^\top||_F^2 = f(\mathbf{A}_0, \mathbf{B}_+) \; , \tag{28}$$

where $\mathbf{B}_+\mathbf{B}_+^\top = \mathbf{I}_n$. Note, that the formulation of the Procrustes problem (24) is based on the constraint $\mathbf{U}^\top\mathbf{U}\mathbf{\Psi} = \mathbf{\Psi}$ (diagonal) as shown in (25). The solution (update) $\mathbf{B}_+$ is global (at each step) and is given by the SVD of a rank deficient matrix. Unfortunately, such a global solution is not unique as demonstrated in the Appendix.

With available update $\mathbf{B}_+$, the GEFA objective function (28) turns into

$$f(\mathbf{A}_0, \mathbf{B}_+) = f(\mathbf{F}_+, \mathbf{\Lambda}_0, \mathbf{U}_+, \mathbf{\Psi}_0) = ||(\mathbf{Z} - \mathbf{F}_+\mathbf{\Lambda}_0^\top) - \mathbf{U}_+\mathbf{\Psi}_0||_F^2 \; , \tag{29}$$

which is a function of $\mathbf{\Lambda}_0$ and $\mathbf{\Psi}_0$ only. For fixed $\mathbf{\Psi}_0$, the minimization of (29) is equivalent to the minimization of

$$\phi(\mathbf{\Lambda}) = \text{trace}\mathbf{\Lambda}^\top\mathbf{\Lambda} - 2\text{trace}\mathbf{\Lambda}^\top\mathbf{Z}^\top\mathbf{F}_+ \; .$$

Lemma 5.1. *$\phi(\boldsymbol{X}) = trace\boldsymbol{X}^\top\boldsymbol{X} + trace\boldsymbol{A}^\top\boldsymbol{X}$ is strictly convex on $\mathbb{R}^{p\times k}$ for any $\boldsymbol{A} \in \mathbb{R}^{p\times k}$.*

PROOF : The Hessian of $\phi(\mathbf{X})$ is $\mathbf{I}_{pk}$. ■

As the gradient of $\phi(\mathbf{\Lambda})$ is $\nabla = \mathbf{\Lambda} - \mathbf{Z}^\top\mathbf{F}_+$, Lemma 5.1 implies that $\phi(\mathbf{\Lambda})$ has global minimum in $\mathbb{R}^{p\times k}$ attained at $\mathbf{\Lambda}_+ = \mathbf{Z}^\top\mathbf{F}_+$. As $\mathbb{R}^{p\times k}$ is convex one is tempted to say that $\mathbf{\Lambda}_+ = \mathbf{Z}^\top\mathbf{F}_+$ is the unique minimizer of $\phi(\mathbf{\Lambda})$. However, it is easy to see that $\mathbf{\Lambda}_+\mathbf{Q}^\top = \mathbf{Z}^\top\mathbf{F}_+\mathbf{Q}^\top$ is a minimizer too, for any orthogonal $\mathbf{Q}$. To clarify this, note that $\mathbf{\Lambda}_+ = \mathbf{Z}^\top\mathbf{F}_+$ is also a solution of the first order optimality conditions:

$$\mathbf{F}_+^\top\mathbf{Z}\mathbf{\Lambda} \text{ be symmetric, and } (\mathbf{I}_p - \mathbf{\Lambda}(\mathbf{\Lambda}^\top\mathbf{\Lambda})^{-1}\mathbf{\Lambda}^\top)\mathbf{Z}^\top\mathbf{F}_+ = \mathbf{O}_{p\times k} \ , \tag{30}$$

for the minimizers $\mathbf{\Lambda}$ of $\phi(\mathbf{\Lambda})$ over the *nonconvex* noncompact Stiefel manifold $\mathbb{ST}(p,k) \subset \mathbb{R}^{p\times k}$ of all $p \times k$ matrices of rank k. This explains why $\mathbf{\Lambda}_+ = \mathbf{Z}^\top\mathbf{F}_+$ is not the unique minimizer of $\phi(\mathbf{\Lambda})$, and why it has full column rank k. Hence, (29) can be reduced to

$$\begin{aligned} f(\mathbf{F}_+, \mathbf{\Lambda}_0, \mathbf{U}_+, \mathbf{\Psi}_0) &= ||(\mathbf{Z} - \mathbf{F}_+\mathbf{\Lambda}_0^\top) - \mathbf{U}_+\mathbf{\Psi}_0||_F^2 \geq ||(\mathbf{Z} - \mathbf{F}_+\mathbf{F}_+^\top\mathbf{Z}) - \mathbf{U}_+\mathbf{\Psi}_0||_F^2 \\ &= f(\mathbf{F}_+, \mathbf{\Lambda}_+, \mathbf{U}_+, \mathbf{\Psi}_0) \ . \end{aligned} \tag{31}$$

Finally, an update of $\mathbf{\Psi}_0$ is needed which further reduces (31). The minimization of (31) is equivalent to the minimization of

$$\begin{aligned} \varphi(\mathbf{\Psi}) &= \text{trace}(\mathbf{\Psi}\mathbf{U}_+^\top\mathbf{U}_+\mathbf{\Psi}) - 2\text{trace}((\mathbf{Z} - \mathbf{F}_+\mathbf{F}_+^\top\mathbf{Z})^\top\mathbf{U}_+\mathbf{\Psi}) \\ &= \text{trace}(\mathbf{\Psi}\text{diag}(\mathbf{U}_+^\top\mathbf{U}_+\mathbf{\Psi})) - 2\text{trace}(\text{diag}((\mathbf{Z}^\top - \mathbf{Z}^\top\mathbf{F}_+\mathbf{F}_+^\top)\mathbf{U}_+)\mathbf{\Psi}) \\ &= \text{trace}\mathbf{\Psi}^2 - 2\text{trace}(\text{diag}(\mathbf{Z}^\top\mathbf{U}_+)\mathbf{\Psi}) \ , \end{aligned}$$

and, by Lemma 5.1, its unique minimum is attained for $\mathbf{\Psi}_+ = \text{diag}(\mathbf{Z}^\top\mathbf{U}_+)$. Then:

$$\begin{aligned} f(\mathbf{F}_+, \mathbf{\Lambda}_+, \mathbf{U}_+, \mathbf{\Psi}_0) \geq\ & ||(\mathbf{Z} - \mathbf{F}_+\mathbf{F}_+^\top\mathbf{Z}) - \mathbf{U}_+\text{diag}(\mathbf{U}_+^\top\mathbf{Z})||_F^2 \ = \\ & ||\mathbf{Z} - \mathbf{F}_+\mathbf{\Lambda}_+^\top - \mathbf{U}_+\mathbf{\Psi}_+||_F^2 \ = f(\mathbf{F}_+, \mathbf{\Lambda}_+, \mathbf{U}_+, \mathbf{\Psi}_+) \ . \end{aligned}$$

The updates $\mathbf{\Lambda}_+$ and $\mathbf{\Psi}_+$ are global minimizers at each step. Thus, **GEFALS** globally minimizes the GEFA objective function, i.e. starting from any initial point. Note that the above derivations make use of $\mathbf{F}_+^\top\mathbf{F}_+ = \mathbf{I}_k$ and $\mathbf{F}_+^\top\mathbf{U}_+ = \mathbf{O}_{k\times p}$, and impose the constraint $\mathbf{U}_+^\top\mathbf{U}_+\mathbf{\Psi} = \mathbf{\Psi}$ (diagonal).

In general, the convergence of the objective function does not guarantee convergence of the parameters. The convergence of the updates $\mathbf{B}$ is guaranteed as the set of all $\mathbf{B}$ for which $\mathbf{B}\mathbf{B}^\top = \mathbf{I}_n$ is compact, and the objective function f is continuous. The convergence of the updates $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ is less obvious as $\mathbf{\Lambda}$ stays on the noncompact Stiefel manifold $\mathbb{ST}(p,k)$ and $\mathbf{\Psi}$ – on $\mathbb{R}^p$. To see this, consider the following general inequality $||\mathbf{x}-\mathbf{y}|| \geq |\ ||\mathbf{x}|| - ||\mathbf{y}||\ |$, which is true for any $\mathbf{x}, \mathbf{y}$, and any norm. Its application gives:

$$||(\mathbf{Z}-\mathbf{U}\mathbf{\Psi})-\mathbf{F}\mathbf{\Lambda}^\top||_F^2 \geq (||\mathbf{Z}-\mathbf{U}\mathbf{\Psi}||_F - ||\mathbf{F}\mathbf{\Lambda}^\top||_F)^2 \ = (||\mathbf{Z}-\mathbf{U}\mathbf{\Psi}||_F - ||\mathbf{\Lambda}||_F)^2 \ ,$$

and

$$||(\mathbf{Z}-\mathbf{F}\mathbf{\Lambda}^\top)-\mathbf{U}\mathbf{\Psi}||_F^2 \geq (||\mathbf{Z}-\mathbf{F}\mathbf{\Lambda}^\top||_F - ||\mathbf{U}\mathbf{\Psi}||_F)^2 \ = (||\mathbf{Z}-\mathbf{F}\mathbf{\Lambda}^\top||_F - ||\mathbf{\Psi}||_F)^2 \ ,$$

which show that $f \to \infty$, when $||\mathbf{\Lambda}||_F \to \infty$ and $||\mathbf{\Psi}||_F \to \infty$, respectively, and the rest of the parameters are fixed. Then, it follows that all level sets of $f(\mathbf{\Lambda})$ and $f(\mathbf{\Psi})$ are bounded (Ortega and Rheinboldt 1970, 4.3.2). By definition, they are also closed, as f is continuous, and thus, compact. Hence, global minimizers of $f(\mathbf{\Lambda})$ and $f(\mathbf{\Psi})$ exist. As discussed before, the global minimizer of $f(\mathbf{\Psi})$ is unique, and there is no unique global minimizer of $f(\mathbf{\Lambda})$. However, if $\mathbf{\Lambda}$ is assumed to be lower triangular, then $f(\mathbf{L})$ has a unique minimizer as the subspace of all lower triangular matrices $\mathcal{L}(p,k)$ with rank k is convex.

Finally, it can be shown that **GEFALS** is a specific gradient descent method. Thus, the ALS process has linear rate of convergence, as any other gradient method.

Indeed, it has been shown already that the **GEFALS** updating formulas $\mathbf{\Lambda} = \mathbf{Z}^\top\mathbf{F}$ and $\mathbf{\Psi} = \mathrm{diag}(\mathbf{Z}^\top\mathbf{U})$ are solutions of the corresponding first order optimality conditions, obtained from the gradients of the GEFA objective function with respect to $\mathbf{\Lambda}$ and $\mathbf{\Psi}$. Now, consider the Procrustes problem (24). The gradient of its objective function is $\mathbf{ZA}$. The first order optimality conditions for the minimizers $\mathbf{B}$ of (24) are

$$\mathbf{ZAB}^\top \text{be symmetric, and } \mathbf{ZA}(\mathbf{I}_{p+k} - \mathbf{B}^\top\mathbf{B}) = \mathbf{O}_{n\times(p+k)} \ . \tag{32}$$

It is easy to check that $\mathbf{B} = (\mathbf{ZAA}^\top\mathbf{Z}^\top)^{-1/2}\mathbf{ZA}$ solves (32) and satisfies $\mathbf{B}\mathbf{B}^\top = \mathbf{I}_n$. Let $\mathbf{ZA} = \mathbf{QDP}^\top$ be the SVD of $\mathbf{ZA}$. Then, after substitution, one finds $\mathbf{B} = \mathbf{QP}^\top$, which is exactly the standard solution of the Procrustes problem (24).

6 Applications

All computations are carried out using MATLAB R2009a (Matlab 2009) on a PC under Windows XP. The codes used in the numerical experiments are available upon request.

6.1 Thurstone's 26-variable box data

Thurstone collected a random sample of 20 boxes and measured their three dimensions x (length), y (width) and z (height) (Harman 1976, p.157). In this data set, the boxes constitute the observational units. The variables of the example are twenty-six functions of these dimensions listed in Table 1. Most 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 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. Note, that the correlation matrix given on (Thurstone 1947, p.371) and heavily used in classic EFA is based on 30 boxes, which dimensions are not available.

The observed variables are mean-centered and scaled to have unit norm. They are collected in a 20×26 data matrix $\mathbf{Z}$. The first three eigenvalues of $\mathbf{Z}^\top \mathbf{Z}$ (12.4217, 7.1807, 5.5386, .2963,...) are considerably greater than one, and than the rest ones. Thus, according to the Kaiser's criterion three common factors will be sought, i.e. $k = 3$.

First, standard EFA least squares solutions $\{\mathbf{\Lambda}, \mathbf{\Psi}\}$ are obtained, i.e. by solving (8) with $\mathbf{\Gamma} = \mathbf{I}_p$. To make these solutions comparable to the ones obtained by the **GEFALS**, the following reparameterizations of the classical EFA model are employed.

The eigenvalue decomposition (EVD) reparameterization of the EFA correlation structure (6) is given as follows (Trendafilov 2003):

$$\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^\top + \mathbf{\Psi}^2 = \mathbf{Q}\mathbf{D}^2\mathbf{Q}^\top + \mathbf{\Psi}^2 \ ,$$

where $\mathbf{Q}\mathbf{D}^2\mathbf{Q}^\top$ is the EVD of the positive semi-definite matrix $\mathbf{\Lambda}\mathbf{\Lambda}^\top$ of rank at most k. Thus, the modified EFA problem is to find a triple $\{\mathbf{Q}, \mathbf{D}, \mathbf{\Psi}\}$, instead of $\{\mathbf{\Lambda}, \mathbf{\Psi}\}$ as in the classical case, and is related to the canonical form factor solution (Harman 1976, 8.7).

Similarly, the lower triangular (LT) reparameterization of the EFA correlation structure

(6) is given as (Trendafilov 2005)

$$\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^\top + \mathbf{\Psi}^2 = \mathbf{L}\mathbf{L}^\top + \mathbf{\Psi}^2 \ ,$$

where $\mathbf{L}\mathbf{L}^\top$ is the Cholesky decomposition of $\mathbf{\Lambda}\mathbf{\Lambda}^\top$. Thus, the modified EFA problem is to find a pair $\{\mathbf{L}, \mathbf{\Psi}\}$, instead of $\{\mathbf{\Lambda}, \mathbf{\Psi}\}$ as in the classical case.

For both reparameterizations, the corresponding LS fitting problems are solved by making use of the projected gradient approach (Trendafilov 2003, 2005).

Table 1: Standard LS solutions for Thurstone's 26-variable box data.

	EVD reparameterization				LT reparameterization			
Variable	$\mathbf{\Lambda}$			$\mathbf{\Psi}^2$	$\mathbf{L}$			$\mathbf{\Psi}^2$
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

Then, LS estimations of all GEFA model unknowns are obtained simultaneously, making use of the new iterative algorithm **GEFALS**. To reduce the chance of mistaking local with

global solutions, the algorithm was run twenty times, each with different random starts for the orthonormal block matrix $\mathbf{B} = [\mathbf{F} \;\; \mathbf{U}]$. The algorithm was stopped when successive function values differed by less than $\epsilon = 10^{-6}$. The corresponding results for $\{\mathbf{\Lambda}, \mathbf{\Psi}^2\}$ applying two parameterizations for the loadings are provided in Table 2.

Table 2: **GEFALS** solutions for the Thurstone's 26-variable box data.

	$\mathbf{\Lambda} = \mathbf{Z}^\top\mathbf{F}$ error of fit = .175174				$\mathbf{L} = \text{tril}(\mathbf{Z}^\top\mathbf{F})$ error of fit = .175184			
Variable	$\mathbf{\Lambda}$			$\mathbf{\Psi}^2$	$\mathbf{L}$			$\mathbf{\Psi}^2$
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

For both algorithms the twenty runs led to the same minimum of the loss function, up to the third decimal place. Numerical experiments showed that **GEFALS** with $\mathbf{L} \leftarrow \texttt{tril}(\mathbf{Z}^\top\mathbf{F})$ is slower but yields pretty stable loadings. In contrast, **GEFALS** with $\mathbf{\Lambda} \leftarrow \mathbf{Z}^\top\mathbf{F}$ is faster, but converges to quite different $\mathbf{\Lambda}$. The **GEFALS** solution with updating $\mathbf{\Lambda} \leftarrow \mathbf{Z}^\top\mathbf{F}$ reported in Table 2 is the one that resembles most the **GEFALS** solution with updating $\mathbf{L} \leftarrow \texttt{tril}(\mathbf{Z}^\top\mathbf{F})$.

GEFALS gives the same $\boldsymbol{\Psi}^2$ and goodness-of-fit for both types of loadings. As expected, **GEFALS** allows for unique factors with zero variance. In contrast, the classical EFA solutions have positive $\boldsymbol{\Psi}^2$, because the projected gradient algorithm for both EVD and LT reparameterization is designed to yield p.d. $\boldsymbol{\Psi}^2$ (Trendafilov 2003, 2005).

The algorithms employing 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 loadings matrices in Table 1 and Table 2, the remaining loadings perfectly identify the box dimensions used to generate each of the variables.

6.2 Numerical performance and simulation illustrations

Here, numerical aspects of the **GEFALS** algorithm are discussed. The simulation experiments provided give some insights on its performance.

In Section 5 it has been shown that the **GEFALS** decreases the objective function (22) at each iteration step. Thus, it is a globally convergent algorithm, that is, convergence to a minimizer is reached independently of the initial value. However, there is no guarantee that the minimizer found is the global minimum of the problem. The standard remedy in such situations is to try several runs with different starting values.

Another issue related to the **GEFALS** performance is that its convergence is linear. This means that the first few **GEFALS** steps reduce (22) sharply, which is then followed by a number of steps with little descent. This is a common weakness of all gradient methods, which may result in slow convergence. The alternating nature of **GEFALS** may additionally worsen the situation. The standard escape is the development of methods, employing second order derivative information. Their convergence is typically quadratic, i.e. they are faster than the gradient methods. The weakness of the second order methods is that they are *locally* convergent, that is, only a "good" starting value ensures convergence to a local minimum. In practice, one frequently combines a gradient method – to locate a good starting value, with a second order method to obtain the solution fast.

While second order methods are not available, a way to speed up the convergence of **GEFALS** is to imitate the above strategy: run it first with low accuracy, and use the result

as a starting value for a second run with high accuracy.

The following simulation experiments illustrate the **GEFALS** performance. First, the Thurstone's 26-variable box data problem is solved by **GEFALS** using both updating formulas and 100 random starts. The results reported in Table 3 are the mean fit for the 100 runs, its standard deviation (std. dev.), and the minimal and the maximal fit obtained. Practically, no local minima are found for these data. The updating formula $\mathbf{L} = \texttt{tril}(\mathbf{Z}^\top\mathbf{F})$ gives slightly better results. The average CPU times in seconds required by the two updating formulas are .014 and .0667, respectively. Using rational starts reduces the CPU times to .0134 and .0169, respectively. When **GEFALS** makes use of the updating $\mathbf{\Lambda} = \mathbf{Z}^\top\mathbf{F}$, rational starts for $\mathbf{\Lambda}$ and $\mathbf{F}$ are obtained by the SVD of the data $\mathbf{Z}$. Alternatively, when **GEFALS** makes use of the updating $\mathbf{L} = \texttt{tril}(\mathbf{Z}^\top\mathbf{F})$, rational starts for $\mathbf{L}$ and $\mathbf{F}$ are obtained by the QR decomposition of $\mathbf{Z}$. In both cases, random starts are used for $\mathbf{U}$ and $\mathbf{\Psi}$.

Table 3: Results for 100 random starts of **GEFALS** for Thurstone's 26-variable box data.

Parameterization	mean	std. dev.	min	max
$\mathbf{\Lambda} = \mathbf{Z}^\top\mathbf{F}$	.1752	$2.3056\ 10^{-5}$	.1751	.1753
$\mathbf{L} = \texttt{tril}(\mathbf{Z}^\top\mathbf{F})$	.1752	$4.4121\ 10^{-7}$	.1752	.1752

Next, the **GEFALS** performance is illustrated on random data with $n = 25$, $p = 30$ and four extracted common factors ($k = 4$). The data are generated as follows. In a first step, a random $n \times p$ matrix $\mathbf{X}_s$ is generated using uniform random numbers in $[-.5, .5]$. Then, **GEFALS** is applied on $\mathbf{X}_s$. Let the resulting GEFA parameters be $\mathbf{\Lambda}_s, \mathbf{F}_s, \mathbf{U}_s$ and $\mathbf{\Psi}_s$. The data for the simulation experiment are formed as $\mathbf{X} = \mathbf{F}_s\mathbf{\Lambda}_s^\top + \mathbf{U}_s\mathbf{\Psi}_s$. This matrix $\mathbf{X}$ will be analyzed by **GEFALS** using either random or rational starts (with random starts for $\mathbf{U}$ and $\mathbf{\Psi}$), and seven different termination criteria listed in Table 4, using different accuracy and number of iterations. The seventh criterion makes at most 100 iterations requiring the magnitude of the difference between the consecutive function values to be $< 10^{-4}$, and then this accuracy requirement is switched to $< 10^{-6}$.

For each termination criterion 100 **GEFALS** runs are made with updating $\mathbf{\Lambda} = \mathbf{Z}^\top\mathbf{F}$, and using random (rand) and rational (ratio) starts. The reported values in Table 4 are the mean fit over the 100 runs, its standard deviation, the mean CPU time and the median CPU time in seconds.

Table 4: Results for **GEFALS** (with $\mathbf{\Lambda} = \mathbf{Z}^\top\mathbf{F}$) applied to a random data matrix with $n = 25, p = 30$ and $k = 4$. For each of the listed seven termination criteria 100 rational (ratio) and random (rand) starts are used.

#	Termination criterion	Start	mean	std. dev.	CPU mean/median
1	`(abs(f0 - f)<= 1e-6)|(max_iter>100)`	ratio	.2532	.0086	.1261/.1302
		rand	.2532	.0080	.1256/.1202
2	`(abs(f0 - f)<= 1e-5)|(max_iter>1000)`	ratio	.2493	.0079	.9257/1.1116
		rand	.2488	.0074	.8717/.9964
3	`(abs(f0 - f)<= 1e-6)|(max_iter>1000)`	ratio	.2521	.0085	1.2195/1.3119
		rand	.2532	.0085	1.2179/1.2919
4	`(abs(f0 - f)<= 1e-5)|(max_iter>4000)`	ratio	.2470	.0060	1.3746/.9113
		rand	.2456	.0045	1.2281/1.0165
5	`(abs(f0 - f)<= 1e-6)|(max_iter>4000)`	ratio	.2501	.0076	4.0400/5.1825
		rand	.2501	.0080	4.0079/5.1825
6	`(abs(f0 - f)<= 1e-6)`	ratio	.2461	.0051	13.6353/11.3313
		rand	.2456	.0039	12.6581/7.5108
7	`(abs(f0 - f)<= 1e-4)|(max_iter>100)`	ratio	.2448	.0036	11.8760/8.5072
	and followed by `(abs(f0 - f)<= 1e-6)`	rand	.2464	.0059	9.8972/6.0136

Table 4 reveals that for the first five criteria there is little difference between the CPU times when using rational or random starts. However, for the last two criteria the random starts require less CPU time. The greater median CPU time (than the mean) indicates that there are fewer runs requiring very little CPU times to converge. Conversely, the smaller median CPU time indicates that there are fewer runs requiring considerable CPU times to converge. According to Table 4, very similar (average) fit is obtained by **GEFALS** when using different termination criteria. More information about the **GEFALS** performance is obtained from the histograms of the fits using different starting values and termination criteria (see Figure 1).

The histograms in Figure 1 have 20 bins which means that the difference between the fits allocated into two neighboring bins is between .002 and .004. The histograms show that the rational starts ensure hitting lower minima more frequently. Moreover, the seventh criterion gives very good results for about 10–14% less CPU time than the sixth criterion.

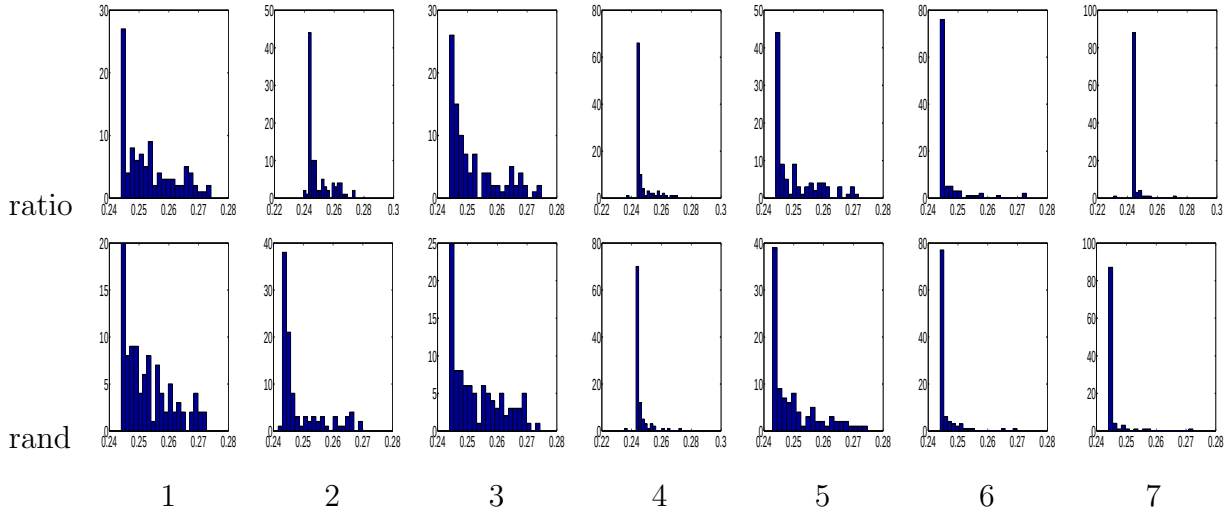

Figure 1: Histograms (with 20 bins) of the minima obtained by **GEFALS** with the seven termination criteria listed in Table 4.

6.3 Atmospheric science data

Climate is a natural system that is characterized by complex and high-dimensional phenomena. To improve the understanding of the physical behaviour of the system, it is often useful to reduce the dimensionality of the data.

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

Unlike PCA, the EFA application in atmospheric science is quite uncommon. Here, GEFA is applied to data from the National Center for Environmental Prediction/National Center for Atmospheric Research (NCEP/NCAR) re-analysis project. The data consist of winter monthly sea-level pressures (SLP) over the Northern Hemisphere north of 20°N. The winter season is defined by the months December, January and February. The data set spans the period from December 1948 to February 2006 ($n = 174$ observations) and is available on a regular grid with a 2.5° latitude $\times$ 2.5° longitude resolution ($p = 29 \times 144 = 4176$ variables

representing gridpoints). The data set was kindly provided by Dr Abdel Hannachi.

Prior to GEFA, 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 analyzed by GEFA.

The new GEFA model is applied with $p > n$ and $k = 5$, where the five factors account for 60.2% of the total variance of the data. This choice is made to 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 sample covariance matrix and finding numerical solution of the LS problem (8) with $\mathbf{\Gamma} = \mathbf{I}_p$ by means of an iterative Newton-Raphson procedure (implemented in **SAS**®) takes about 2.5 hours.

GEFA 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. This spatial map representation greatly aids interpretation, as is illustrated in Figure 2.

For the winter SLP data, the plots represent the first (i) and second (ii) column of the 4176×5 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 et al. 2007). The second EFA pattern (ii) yields the North Pacific Oscillation (NPO) or Pacific pattern, a monopolar structure sitting over the North Pacific (Hannachi et al. 2007). For the twenty different random starts, the obtained GEFA loadings look similar.

It is of interest to compare the spatial patterns obtained by GEFA and by PCA/EOF analysis. Figure 3 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. This tripolar structure corresponds to the familiar Annular Oscillation (AO) (Hannachi et al. 2007). Like GEFA 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.

7 Discussion

The EFA model is reconsidered as a specific data matrix decomposition with fixed unknown matrix parameters. They are found simultaneously by **EFALS** algorithm, which fits the EFA model directly to the data. This is in contrast to the classical EFA, where only estimates of the factor loadings and unique variances are found by fitting the EFA correlation structure to the sample correlation matrix.

The classical EFA model is generalized to the GEFA model which allows positive semi-definite $\mathbf{\Psi}^2$. GEFA can be applied to data matrices with either $p \geq n$ or $p < n$. All GEFA parameters are found simultaneously by the **GEFALS** algorithm, based on SVD, which makes GEFA a reasonable alternative to PCA.

Since the GEFA model postulates the existence of k common and p unique factors such

that the p observed variables can be represented as their linear combinations, the scores of the n observations on the common and unique factors are not uniquely identifiable. In other words, an infinite set of scores for the common and unique factors can be constructed satisfying the GEFA model. This form of indeterminacy is known as 'factor indeterminacy' (e.g., Mulaik 2005). However, the non-uniqueness of the factor scores in GEFA is not a problem for **GEFALS** which finds estimates for all matrix parameters. From this point of view, the numerical procedure developed in this paper avoids the conceptual problem of factor score indeterminacy and facilitates the estimation of both **F** and **U**.

Acknowledgements

The authors are grateful to the Associate Editor and three anonymous reviewers for their helpful comments on the first draft of this paper.

Appendix

The following is an amended version of the proposition considered by De Leeuw (2004).

Lemma 7.1. *Let* $\boldsymbol{X}$ *be a* $m \times n$ $(m \geq n)$ *matrix of rank* r $(\leq n)$ *and its 'economy' SVD* $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$ *be written in details as follows:*

$$\boldsymbol{X} = \begin{bmatrix} \underset{m \times r}{\boldsymbol{U}_1} & \underset{m \times (n-r)}{\boldsymbol{U}_2} \end{bmatrix} \begin{bmatrix} \underset{r \times r}{\boldsymbol{D}} & \underset{r \times (n-r)}{\boldsymbol{O}} \\ \underset{(n-r) \times r}{\boldsymbol{O}} & \underset{(n-r) \times (n-r)}{\boldsymbol{O}} \end{bmatrix} \begin{bmatrix} \underset{r \times n}{\boldsymbol{V}_1^\top} \\ \underset{(n-r) \times n}{\boldsymbol{V}_2^\top} \end{bmatrix}, \tag{33}$$

where $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}_n$ *and* $\boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{V}\boldsymbol{V}^\top = \boldsymbol{I}_n$. *Then*

$$\max_{\boldsymbol{Q}^\top\boldsymbol{Q}=\boldsymbol{I}_n} \mathit{trace}\boldsymbol{Q}^\top\boldsymbol{X} = \mathit{trace}\boldsymbol{D}$$

and it is attained for any $\boldsymbol{Q} = \boldsymbol{U}_1\boldsymbol{V}_1^\top + \boldsymbol{U}_2\boldsymbol{P}\boldsymbol{V}_2^\top$, *where* $\boldsymbol{P}^\top\boldsymbol{P} = \boldsymbol{P}\boldsymbol{P}^\top = \boldsymbol{I}_{n-r}$.

PROOF : By projecting the gradient, $\mathbf{X}$, of the objective function trace$\mathbf{Q}^\top\mathbf{X}$ onto $\mathbf{Q}^\top\mathbf{Q} = \mathbf{I}_n$, one finds the following first order optimality conditions for its maximizer (Trendafilov 2003):

- $\mathbf{Q}^\top\mathbf{X}$ be symmetric
- $\mathbf{X} = \mathbf{Q}\mathbf{Q}^\top\mathbf{X}$.

According to the first optimality condition, let $\mathbf{S} = \mathbf{Q}^\top\mathbf{X}$ be some $n \times n$ symmetric matrix. Then, $\mathbf{X} = \mathbf{Q}\mathbf{S}$ and $\mathbf{S}^2 = \mathbf{X}^\top\mathbf{X} = \mathbf{V}\mathbf{D}^2\mathbf{V}^\top$, and $\mathbf{S}$ can be expressed as:

$$\mathbf{S} = \begin{bmatrix} \mathbf{V}_1 & \mathbf{V}_2 \\ n \times r & n \times (n-r) \end{bmatrix} \begin{bmatrix} \mathbf{D} & \mathbf{O} \\ r \times r & r \times (n-r) \\ \mathbf{O} & \mathbf{O} \\ (n-r) \times r & (n-r) \times (n-r) \end{bmatrix} \begin{bmatrix} \mathbf{V}_1^\top \\ r \times n \\ \mathbf{V}_2^\top \\ (n-r) \times n \end{bmatrix} . \tag{34}$$

Let the unknown $\mathbf{Q}$ be written as:

$$\mathbf{Q} = \begin{bmatrix} \mathbf{U}_1 & \mathbf{U}_2 \\ m \times r & m \times (n-r) \end{bmatrix} \begin{bmatrix} \mathbf{Q}_1 \\ r \times n \\ \mathbf{Q}_2 \\ (n-r) \times n \end{bmatrix} . \tag{35}$$

By making use of (33), (34) and (35), one can see that the identity $\mathbf{X} = \mathbf{Q}\mathbf{S}$ implies $\mathbf{Q}_1\mathbf{V}_1 = \mathbf{I}_r$ and $\mathbf{Q}_2\mathbf{V}_1 = \mathbf{O}_{(n-r)\times r}$. Then, it follows that $\mathbf{Q}_1 = \mathbf{V}_1^\top$ and $\mathbf{Q}_2 = \mathbf{P}\mathbf{V}_2^\top$ for any full rank $(n-r) \times (n-r)$ matrix $\mathbf{P}$. Moreover, the identities $\mathbf{I}_n = \mathbf{Q}^\top\mathbf{Q} = \mathbf{Q}_1^\top\mathbf{Q}_1 + \mathbf{Q}_2^\top\mathbf{Q}_2$ and $\mathbf{I}_n = \mathbf{V}_1\mathbf{V}_1^\top + \mathbf{V}_2\mathbf{V}_2^\top$ show that $\mathbf{P}^\top\mathbf{P} = \mathbf{I}_{n-r}$. Finally, $\mathbf{P}\mathbf{P}^\top = \mathbf{I}_{n-r}$ follows from the second optimality condition $\mathbf{X} = \mathbf{Q}\mathbf{Q}^\top\mathbf{X}$, by making use of (33) and (35). ■

References

Anderson, T. W., 1984. An Introduction to Multivariate Statistical Analysis, 2nd Edition. John Wiley & Sons: New York.

Anderson, T. W., Rubin, H., 1956. Statistical inference in factor analysis. In: Neyman, J. (Ed.), Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability, Vol. V. University of California Press: Berkeley, pp. 111–150.

Bartholomew, D. J., Knott, M., 1999. Latent Variable Models and Factor Analysis, 2nd Edition. Edward Arnold: London.

De Leeuw, J., 2004. Least squares optimal scaling of partially observed linear systems. In: van Montfort, K., Oud, J., Satorra, A. (Eds.), Recent Developments on Structural Equation Models: Theory and Applications. Kluwer Academic Publishers: Dordrecht, NL, pp. 121–134.

Golub, G. H., Van Loan, C. F., 1996. Matrix Computations, 3rd Edition. The John Hopkins University Press: Baltimore, MD.

Hannachi, A., Jolliffe, I. T., Stephenson, D. B., 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, 3rd Edition. University of Chicago Press: Chicago, IL.

Horn, R., Johnson, C., 1986. Matrix Analysis. Cambridge University Press: Cambridge, UK.

Horst, P., 1965. Factor Analysis of Data Matrices. Holt, Rinehart and Winston: New York, NY.

Jennrich, R. I., Trendafilov, N. T., 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, 2nd Edition. Springer: New York, NY.

Jöreskog, K. G., 1962. On the statistical treatment of residuals in factor analysis. Psychometrika 27, 335–354.

Lawley, D. N., 1942. Further investigations in factor estimation. Proceedings of the Royal Society of Edinburgh: Section A 61, 176–185.

Matlab, 2009. MATLAB R2009a. The MathWorks, Inc.

McDonald, R. P., 1979. The simultaneous estimation of factor loadings and scores. British Journal of Mathematical and Statistical Psychology 32, 212–228.

Mulaik, S. A., 1972. The foundations of factor analysis. McGraw-Hill: New York.

Mulaik, S. A., 2005. Looking back on the indeterminacy controversies in factor analysis. In: Maydeu-Olivares, A., McArdle, J. J. (Eds.), Contemporary Psychometrics: A Festschift for Roderick P. McDonald. Lawrence Erlbaum: Mahwah, pp. 173–206,.

Ortega, J. M., Rheinboldt, W. C., 1970. Iterative Solution of Nonlinear Equations in Several Variables. Academic Press: New York, NY.

Robertson, D., Symons, J., 2007. Maximum likelihood factor analysis with rank-deficient sample covariance matrices. Journal of Multivariate Analysis 98, 813–828.

Thurstone, L. L., 1947. Multiple Factor Analysis. University of Chicago Press: Chicago, IL.

Trendafilov, N. T., 2003. Dynamical system approach to factor analysis parameter estimation. British Journal of Mathematical and Statistical Psychology 56, 27–46.

Trendafilov, N. T., 2005. Fitting the factor analysis model in ℓ_1 norm. British Journal of Mathematical and Statistical Psychology 58, 19–31.

Whittle, P., 1952. On principal components and least squares methods in factor analysis. Skandinavisk Aktuarietidskrift 35, 223–239.

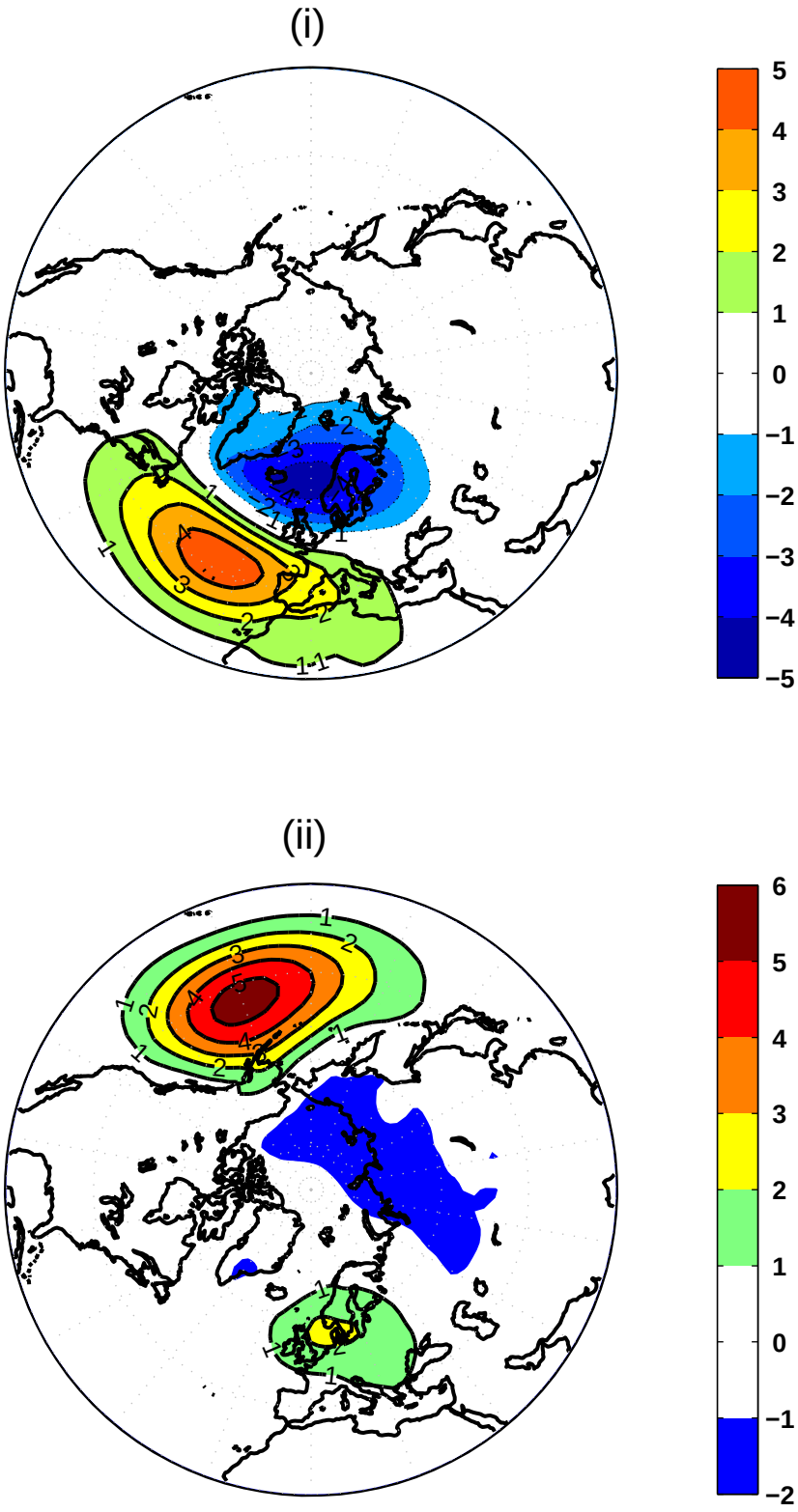

Figure 2: Spatial map representations of the first (i) and second (ii) column of the EFA loading matrix for winter SLP data ($k = 5$).

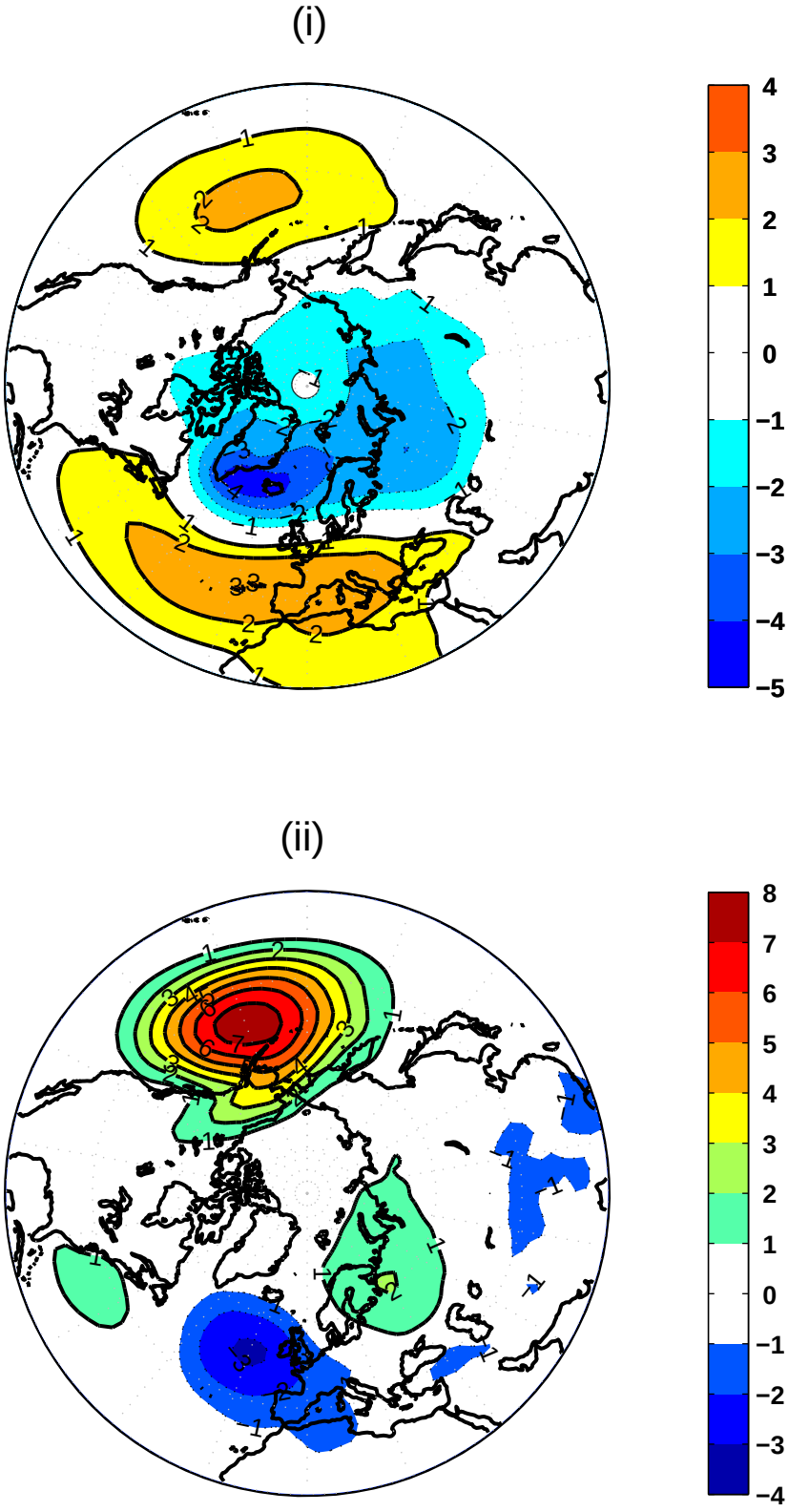

Figure 3: Spatial map representations of the two leading EOFs one (i) and two (ii) for winter SLP data ($k = 5$). The EOFs have been multiplied by 100.