Sparsest Factor Analysis for Clustering Variables:
A Matrix Decomposition Approach

Abstract

We propose a new procedure for sparse factor analysis (FA) such that each variable loads only one common factor. Thus, the loading matrix has a single nonzero element in each row and zeros elsewhere. Such a loading matrix is the sparsest possible for certain number of variables and common factors. For this reason, the proposed method is named sparsest FA (SSFA). It may also be called FA-based variable clustering, since the variables loading the same common factor can be classified into a cluster. In SSFA, all model parts of FA (common factors, their correlations, loadings, unique factors, and unique variances) are treated as fixed unknown parameter matrices and their least squares function is minimized through specific data matrix decomposition. A useful feature of the algorithm is that the matrix of common factor scores is re-parameterized using QR decomposition in order to efficiently estimate factor correlations. A simulation study shows that the proposed procedure can exactly identify the true sparsest models. Real data examples demonstrate the usefulness of the variable clustering performed by SSFA.

Key words: Exploratory Factor analysis, Sparsest Loadings, Matrix decomposition factor analysis, Variable clustering, QR re-parameterization

1. Introduction

Factor analysis (FA) aims to explain the interrelationships among p observed variables by m ($<< p$) 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 them accounts for the unique variance associated with only one observed variable. FA can be formulated in several ways. Among them, maximum likelihood FA (MLFA) and matrix decomposition FA (MDFA) need to be outlined before introducing our study.

In MLFA, a $p \times 1$ observed vector $\mathbf{x}$, whose expectation $E[\mathbf{x}]$ equals to the $p \times 1$ zero vector $\mathbf{0}_p$, is modeled as

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

Here, random vectors $\mathbf{f}$ ($m \times 1$) and $\mathbf{u}$ ($p \times 1$) contain common and unique factor scores respectively, while $\mathbf{\Lambda}$ is the $p \times m$ matrix of factor loadings and $\mathbf{\Psi}$ is a $p \times p$ diagonal matrix, the squares of whose diagonal elements are called unique variances (e.g., Mulaik, 2010). It is assumed that $\mathbf{f} \sim N_m(\mathbf{0}_m, \mathbf{\Phi})$, i.e, $\mathbf{f}$ follows the m-variate normal distribution whose average vector is $\mathbf{0}_m$ and covariance matrix is $\mathbf{\Phi}$, where the diagonal elements of $\mathbf{\Phi}$ are ones, implying that $\mathbf{\Phi}$ is a correlation matrix. Further, $\mathbf{u} \sim N_p\, (\mathbf{0}_m, \mathbf{I}_p)$ and $E[\mathbf{fu}'] = \mathbf{O}_{m\times p}$ are assumed with $\mathbf{I}_p$ the $p \times p$ identity matrix and $\mathbf{O}_{m\times p}$ the $m\times p$ matrix of zeros. Those assumptions lead to $\mathbf{x} \sim N_p(\mathbf{0}_m, \mathbf{\Lambda\Phi\Lambda}'+\mathbf{\Psi})$, which gives the negative log likelihood

$$l(\mathbf{\Lambda},\mathbf{\Psi},\mathbf{\Phi}) = \log|\mathbf{\Lambda\Phi\Lambda}'+\mathbf{\Psi}| + \mathrm{tr}\mathbf{S}(\mathbf{\Lambda\Phi\Lambda}'+\mathbf{\Psi})^{-1} \tag{2}$$

for sample covariance matrix $\mathbf{S} = n^{-1}\mathbf{X}'\mathbf{X}$. Here, $|\mathbf{\Sigma}|$ denotes the determinant of matrix $\mathbf{\Sigma}$, and $\mathbf{X}$ is the n-observation × p-variables data matrix column-centered as $\mathbf{1}_n'\mathbf{X} = \mathbf{0}_p'$ with $\mathbf{1}_n$ the $n \times 1$ vector of ones. Since of the rotational freedom of $\mathbf{\Lambda}$, the attained value of (2) remains unchanged even if $\mathbf{\Phi}$ is set at $\mathbf{I}_m$. Thus, (2) is minimized over $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ with $\mathbf{\Phi} = \mathbf{I}_m$.

In MDFA, factor scores are also treated as fixed parameters, and $\mathbf{X}$ is modeled as

$$\mathbf{X} = \mathbf{F\Lambda}' + \mathbf{U\Psi} + \mathbf{E} \,. \tag{3}$$

Here, matrices $\mathbf{F}$ ($n \times m$) and $\mathbf{U}$ ($n \times p$) contain common and unique factor scores, respectively, with $\mathbf{E}$ ($n \times p$) containing errors. Its squared norm $||\mathbf{E}||^2$, i.e.,

$$f(\mathbf{F}, \mathbf{U}, \mathbf{\Lambda}, \mathbf{\Psi}) = ||\mathbf{X} - (\mathbf{F\Lambda}'+\mathbf{U\Psi})||^2 , \tag{4}$$

is minimized over $\mathbf{F}$, $\mathbf{U}$, $\mathbf{\Lambda}$, and $\mathbf{\Psi}$, subject to the constraints

$$n^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_m, \quad n^{-1}\mathbf{U}'\mathbf{U} = \mathbf{I}_p, \quad \text{and} \quad n^{-1}\mathbf{F}'\mathbf{U} = \mathbf{O}_{m\times p} \,. \tag{5}$$

Here, the common factors are assumed to be mutually uncorrelated as $n^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_m$ for the same reason as in MLFA. This approach was recently introduced independently by Henk A. L. Kiers in Socǎn (2003) and de Leeuw (2004), and later studied and elaborated by Unkel and Trendafilov (2010), Trendafilov and Unkel (2011), and Stegeman (2016). It is empirically known that MDFA and MLFA provide almost equivalent solutions (Adachi, 2012, 2014).

A main purpose of FA is to explore the relationships of observed variables to the underlying common factors. Those relationships are captured by interpreting the loading matrix $\mathbf{\Lambda}$. The interpretation of $\mathbf{\Lambda}$ is facilitated, if $\mathbf{\Lambda}$ is sparse, i.e., includes a large number of zero loadings, as zero ones clearly show no relationship between the corresponding factors and variables, implying that only nonzero elements are considered for interpreting factors. A classic approach to such facilitation of interpretation is to rotate the loading matrix $\mathbf{\Lambda}$ so that it approximates a sparse matrix. However, the rotated loadings cannot be exactly zeros. Thus, users must neglect loadings below certain magnitudes and make them zero without further adjustment of the remaining (nonzero) loadings.

New FA procedures which do not show this drawback when rotating loadings have recently been proposed, and are generally called sparse FA (Adachi & Trendafilov, 2014; Hirose & Yamamoto, 2014a,b). In sparse FA, the goal is to obtain a loading matrix $\mathbf{\Lambda}$ which has a large proportion of exactly zero elements (loadings). Here, it should be noted that the optimal locations of the zero elements in $\mathbf{\Lambda}$ are unknown and have to be estimated. That is, the parameter estimation in sparse FA includes the location of the zero loadings. A great number of sparse principal component analysis (PCA) procedures have been successfully developed in the last decade (e.g. Jollife, Uddin & Trendafilov, 2003; Zou, Hastie, & Tibshirani, 2006). In sparse PCA, a sparse weight matrix is produced that contain the weights for the linear combinations of variables that will form components. In analogy, a sparse factor loading matrix is estimated in sparse FA together with the other parameters.

In sparse PCA, a major approach to achieve sparseness is by using penalty function (Trendafilov, 2014). This has also been employed in Hirose and Yamamoto's sparse FA (2014a,b). The authors have developed the R-package "FANC" which implements their procedure. For short, it is referred to as FANC. It is based on MLFA and the negative log likelihood (2) is combined with a penalty function $P_\gamma(\mathbf{\Lambda})$. Thus, FANC is formulated as

$$\min_{\mathbf{\Lambda},\mathbf{\Psi},\mathbf{\Phi}} l(\mathbf{\Lambda}, \mathbf{\Psi}, \mathbf{\Phi}) + \rho P_\gamma(\mathbf{\Lambda}) \quad . \tag{6}$$

Here $P_\gamma(\mathbf{\Lambda})$ is penalizing $\mathbf{\Lambda}$ to have nonzero elements, with ρ and γ being tuning parameters. The latter γ specifies the form of function $P_\gamma(\mathbf{\Lambda})$, while the former ρ controls the relative importance of $P(\mathbf{\Lambda})$: larger ρ values promote sparser loadings. When $\mathbf{\Lambda}$ is sparse, it does not have rotational freedom. Note, that $\mathbf{\Phi}$ is also estimated by (6), and thus, setting $\mathbf{\Phi} = \mathbf{I}_m$ is restrictive.

On the other hand, the level of sparseness is directly constrained to be a specified integer in Adachi and Trendafilov's procedure (2014). SOFA is based on MDFA and formulated using a least squares function (4) and solves the following type of problem

$$\min_{\mathbf{F}, \mathbf{\Lambda}, \mathbf{U}, \mathbf{\Psi}} f(\mathbf{F}, \mathbf{\Lambda}, \mathbf{U}, \mathbf{\Psi}) \text{ subject to } \mathrm{Card}(\mathbf{\Lambda}) = c \text{ and } (5)\ , \tag{7}$$

where Card($\mathbf{\Lambda}$) expresses the cardinality of $\mathbf{\Lambda}$, i.e. the number of its nonzero elements, and c is a pre-specified integer. As already stated, setting factor correlation matrix $\mathbf{\Phi}$ to $\mathbf{I}_m$ is restrictive in sparse cases. But, $n^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_m$ in (5) remains valid, as it is technically necessary for achieving (7) (Adachi & Trendafilov, 2014, p. 229). Thus, it is called sparse orthogonal FA (SOFA), as factors are supposed to be uncorrelated, i.e., orthogonal.

Bayesian sparse FA procedures have also been developed in which sparseness is induced by special prior probabilities (Knowles & Ghahramani, 2011; Rattray, Stegle, Sharp, & Winn, 2009). We are interested in procedures for sparse parameter FA estimation and thus, they are beyond the scope of the paper.

It is considered that a sparser loading matrix can be interpreted easier. In this respect, the interpretation of *the sparsest* matrix is the easiest possible. In this paper, we propose a FA procedure for finding the sparsest loading matrix $\mathbf{\Lambda}$. Such a matrix looks as follows:

$$\underset{p \times m}{\mathbf{\Lambda}} = \begin{bmatrix} 0 \cdots 0 & \# & 0 & \cdots 0 \\ 0 \cdots \cdots & \cdots & \cdots & 0 \quad \# \\ & & \vdots & \\ 0 \quad 0 & \# & 0 & \cdots \cdots 0 \end{bmatrix} \tag{8}$$

where # indicates the nonzero elements. That is, each of the p variables loads only one of the m factors, and thus Card($\mathbf{\Lambda}$) = p. This is the lowest limit of sparseness of $\mathbf{\Lambda}$, because if Card($\mathbf{\Lambda}$) < p, then at least one of the variables does not load any common factor. The procedure for obtaining such loadings as in (8) can be referred to as the sparsest FA. The idea of constraining loadings to be sparsest is already presented in the area of principal component analysis (PCA): Vichi and Saporta (2009) have proposed a procedure called disjoint PCA in which a component loading matrix is constrained as (8). In some sense, sparsest FA extends the same idea to a more complicated model involving unique variances, namely FA (Trendafilov, Unkel, and Krzanowski, 2011).

For sparsest FA, the penalty approach is not convenient, since it is unknown beforehand what value of the tuning parameter gives the solution with a particular value of Card($\mathbf{\Lambda}$). As shown in (6), two tuning parameters must be specified in FANC. Thus, one must take efforts with trial-and-errors, in order to find the appropriate ρ and γ values that give the sparsest solution. On the other hand, Card($\mathbf{\Lambda}$) = p is pre-specified in SOFA (7). However, such a constraint cannot enforce every variable to load only one factor: it might happen that a row has two nonzero elements while another one is filled with zeros. Moreover, the SOFA procedure is restricted to the solutions with uncorrelated common factors, i.e. $n^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_m$.

In sparsest FA, the loadings are restricted to the sparsest, i.e. look like in (8). This restrictive property may be disadvantageous. However, in addition to easing interpretation, it allows us to classify variables into a few groups: we can regard the variables that load the same common factor as a group in the sparsest FA solution. Such classification of entities into groups is one of the major subjects in multivariate analysis (Everitt, 1993; Gan, Ma, & Wu, 2007) and, more recently, in data mining (Aggarwal, 2015, Zaki & Meira, 2014). It is well known that most clustering procedures are designed to cluster observations rather than variables (see Gan, et al., 2007). However, clustering variable is an important problem for which very few methods exist. For example, in the field of the psychological testing for which FA was originally proposed (Spearman, 1904), it is desired to cluster test items, as the items classified into the same group can be found to measure the same psychological trait (Goldberg, 1992). Such needs would also be encountered in other fields. For that reason, it is worth considering the sparsest FA, which may also be called FA-based variable clustering.

In this paper, we propose a sparsest FA procedure in which common factors are not restricted to uncorrelated ones. The proposed procedure is referred to as SSFA by abbreviating sparsest FA. The remaining part of the paper is organized as follows: in the next section SSFA is formulated, and its algorithm is detailed in Section 3. SSFA is numerically illustrated and compared with SOFA and FANC in Section 4.

2. Formulation

Our proposed SSFA is based on MDFA. The reason for choosing the approach is that its objective function (4) is a quadratic function of $\mathbf{\Lambda}$ which can be easily minimized over $\mathbf{\Lambda}$ subject to the constraint that $\mathbf{\Lambda}$ is the sparsest, as show in Section 3.4. Thus, the proposed method can be viewed as a modification of MDFA-based SOFA formulated as (7). In SSFA, Card($\mathbf{\Lambda}$) = c in (7) is replaced by the sparsest constraint on $\mathbf{\Lambda}$, and the orthogonality constraint $n^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_m$ in (5) is relaxed to $n^{-1}\text{diag}(\mathbf{F}'\mathbf{F}) = \mathbf{I}_m$: the factor correlations in $\mathbf{\Phi} = n^{-1}\mathbf{F}'\mathbf{F}$ are also estimated in SSFA, with diag(•) denoting the diagonal matrix whose diagonal elements are those of a parenthesized one. The relaxation is attained by re-parameterizing $\mathbf{F}$ using its constrained QR decomposition as $\mathbf{F} = \mathbf{QR}$ (e.g., Seber, 2008). Here,

$$\frac{1}{n}\mathbf{Q}'\mathbf{Q} = \mathbf{I}_m \ , \tag{9}$$

and $\mathbf{R} = (r_{jk}) = [\mathbf{r}_1, \ldots , \mathbf{r}_m]$ is an upper triangular matrix satisfying diag($\mathbf{R}'\mathbf{R}$) = $\mathbf{I}_m$. This implies that a part of the jth column of $\mathbf{R}$ and r_{11} are constrained as

$$\mathbf{r}_j = [\mathbf{r}_{j1}',\mathbf{0}_{m-j}']', \ ||\mathbf{r}_{j1}|| = 1 \ (j = 1, \ldots, m), \text{ and } r_{11} = 1 \tag{10}$$

with $\mathbf{r}_{j1}$ the j × 1 unknown vector and $\mathbf{0}_{m-j}$ the $(m-j) \times 1$ vector of zeros.

The two constraints in (5)

$$\frac{1}{n}\mathbf{U}'\mathbf{U} = \mathbf{I}_p \tag{11}$$

and $n^{-1}\mathbf{F}'\mathbf{U} = \mathbf{O}_{m\times p}$ are also used in SSFA, where the latter constraint is rewritten as

$$\frac{1}{n}\mathbf{Q}'\mathbf{U} = \mathbf{O}_{m\times p} \tag{12}$$

The remaining constraint is loading matrix $\mathbf{\Lambda} = [\boldsymbol{\lambda}_1, \ldots , \boldsymbol{\lambda}_p]' = (\lambda_{ij})$ being the sparsest, i.e.,

$$\boldsymbol{\lambda}_i = [\lambda_{i1}, \ldots , \lambda_{im}]' \text{ being filled with zero except for an element ,} \tag{13}$$

which implies

$$\mathbf{\Lambda}'\mathbf{\Lambda} \text{ being a diagonal matrix .} \tag{14}$$

Substituting $\mathbf{F} = \mathbf{QR}$ into (4), it can be rewritten as

$$f(\mathbf{Q}, \mathbf{U}, \mathbf{R}, \mathbf{\Lambda}, \mathbf{\Psi}) = ||\mathbf{X} - (\mathbf{QR\Lambda}'+\mathbf{U\Psi})||^2 \ . \tag{15}$$

SSFA is thus formulated as minimizing (15) with respect to **Q**, **U**, **R**, **Λ**, and **Ψ** over subject to the constraints of (9) to (13). Here, we should note that $\mathbf{F} = \mathbf{QR}$, (9), and (10) imply that

$$\mathbf{\Phi} = \frac{1}{n}\mathbf{F}'\mathbf{F} = n^{-1}\mathbf{R}'\mathbf{Q}'\mathbf{QR} = \mathbf{R}'\mathbf{R} \tag{16}$$

is a factor correlation matrix.

3. Algorithm

The minimization of (15) subject to the constraints of (9) to (13) can be attained by alternately iterating the steps, in each of which [**Q**, **U**], **Ψ**, **R**, or **Λ** is updated so that (15) is decreased. We refer to those steps as *factor score*, *unique variance*, *correlation*, and *loading steps*, respectively. As found in the next subsection, it is unnecessary, indeed, to update **Q** and **U** in the factor score step for estimating the optimal **Ψ**, **R**, and **Λ**. Furthermore, they can be obtained without the original data matrix **X**, if only the sample correlation matrix $\mathbf{S} = n^{-1}\mathbf{X}'\mathbf{X}$ is available. After detailing the steps, the algorithm for SSFA is summarized in the final subsection.

3.1. Factor score step

We collect parameter matrices into $n \times (m+p)$ and $p \times (m+p)$ block matrices as $\mathbf{Z} = [\mathbf{Q}, \mathbf{U}]$ and $\mathbf{B} = [\mathbf{\Lambda R}', \mathbf{\Psi}]$, respectively. Then, loss function (15) may be rewritten as $f(\mathbf{Z}, \mathbf{B}) = ||\mathbf{X} - \mathbf{ZB}'||^2$ and the constraints (9), (11), and (12) are summarized into

$$\frac{1}{n}\mathbf{Z}'\mathbf{Z} = \mathbf{I}_{m+p} \quad . \tag{17}$$

In this step, we consider minimizing $f(\mathbf{Z}, \mathbf{B})$ over $\mathbf{Z}$ subject to (17) with $\mathbf{B}$ kept fixed. Since $f(\mathbf{Z}, \mathbf{B}) = ||\mathbf{X} - \mathbf{ZB}'||^2$ is expanded as $c - 2\mathrm{tr}\mathbf{B}'\mathbf{X}'\mathbf{Z}$ with $c = \mathrm{tr}\mathbf{X}'\mathbf{X} + n\mathrm{tr}\mathbf{BB}'$ independent of $\mathbf{Z}$, the minimization amounts to maximizing the linear form

$$\eta(\mathbf{Z}) = \mathrm{tr}\mathbf{B}'\mathbf{X}'\mathbf{Z} \tag{18}$$

over $\mathbf{Z}$ subject to (17). As found in ten Berge's (1983, 1993) theorems, this maximization is attained using the singular value decomposition (SVD) of the $n \times (m+p)$ matrix $n^{-1/2}\mathbf{XB}$:

$$\frac{1}{\sqrt{n}}\mathbf{XB} = \mathbf{K\Delta L}' = [\mathbf{K}_1, \mathbf{K}_2]\begin{bmatrix} \mathbf{\Delta}_1 & \\ & {}_m\mathbf{O}_m \end{bmatrix}\begin{bmatrix} \mathbf{L}_1' \\ \mathbf{L}_2' \end{bmatrix} = \mathbf{K}_1\mathbf{\Delta}_1\mathbf{L}_1' \ . \tag{19}$$

Here, $\mathbf{\Delta}$ is an $(p+m) \times (p+m)$ diagonal matrix with its first $p \times p$ diagonal block being positive definite matrix $\mathbf{\Delta}_1$, and the block matrices $\mathbf{K} = [\mathbf{K}_1, \mathbf{K}_2]$ and $\mathbf{L} = [\mathbf{L}_1, \mathbf{L}_2]$ satisfy $\mathbf{K}'\mathbf{K} = \mathbf{L}'\mathbf{L} = \mathbf{LL}' = \mathbf{I}_{p+m}$ with $\mathbf{K}_1$ and $\mathbf{L}_1$ being $n \times p$ and $(p+m) \times p$ matrices, respectively. In (19), we have assumed

$$\mathrm{rank}(\mathbf{XB}) = p. \tag{20}$$

The linear form (18) satisfies $\mathrm{tr}\mathbf{B}'\mathbf{X}'\mathbf{Z} \leq n\mathrm{tr}\mathbf{\Delta} = n\mathrm{tr}\mathbf{\Delta}_1$ under (17) and the upper bound $n\mathrm{tr}\mathbf{\Delta}_1$ is achieved for

$$\mathbf{Z} = \sqrt{n}\,\mathbf{KL}' = \sqrt{n}\,\mathbf{K}_1\mathbf{L}_1' + \sqrt{n}\,\mathbf{K}_2\mathbf{TL}_2', \tag{21}$$

where $\mathbf{T}$ is an arbitrary $m \times m$ orthogonal matrix (Trendafilov & Unkel, 2011).

It follows from (21), that the optimal factor scores cannot be uniquely determined (e.g., Eldén, 2007). But, the p-variables × $(m+p)$-factors covariance matrix $n^{-1}\mathbf{X}'\mathbf{Z} = n^{-1}[\mathbf{X}'\mathbf{Q}, \mathbf{X}'\mathbf{U}]$ is uniquely determined as shown next. Assumption (20) implies that $\mathbf{B}$ has full-row rank and thus its Moore-Penrose inverse is given by $\mathbf{B}^+ = \mathbf{B}'(\mathbf{BB}')^{-1}$. Making use of $\mathbf{BB}^+ = \mathbf{I}_p$ and (19), one finds that $n^{-1/2}\mathbf{X} = n^{-1/2}\mathbf{XBB}^+ = \mathbf{K}_1\mathbf{\Delta}_1\mathbf{L}_1'\mathbf{B}^+$. This equation and (21) give that:

$$\frac{1}{n}\mathbf{X}'\mathbf{Z} = (n^{-1/2}\mathbf{X})'(n^{-1/2}\mathbf{Z}) = (\mathbf{B}'^+\mathbf{L}_1\mathbf{\Delta}_1\mathbf{K}_1')(\mathbf{KL}') = \mathbf{B}'^+\mathbf{L}_1\mathbf{\Delta}_1\mathbf{L}_1'. \tag{22}$$

From (19) it follows that the eigenvalue decomposition (EVD) of $\mathbf{B}'(n^{-1}\mathbf{X}'\mathbf{X})\mathbf{B} = \mathbf{B}'\mathbf{S}\mathbf{B}$ is defined as

$$\mathbf{B}'\mathbf{S}\mathbf{B} = \mathbf{L}_1\mathbf{\Delta}_1^2\mathbf{L}_1' \,. \tag{23}$$

One can show that (22) follows from (23). Moreover, if (20) holds and the diagonal elements of $\mathbf{\Delta}_1$ take distinct values, then $\mathbf{L}_1$ and $\mathbf{\Delta}_1$ in (23) are uniquely determined. This fact and the uniqueness of $\mathbf{B}^+$ imply that (22) is also unique.

It is shown in the next subsections that $\mathbf{B} = [\mathbf{\Lambda}\mathbf{R}', \mathbf{\Psi}]$ can be updated if only (22) is available. This implies that the update of the factor score matrix $\mathbf{Z} = [\mathbf{Q}, \mathbf{U}]$ can be skipped. Further, the availability of the original data matrix $\mathbf{X}$ is not essential: if only a sample covariance matrix $\mathbf{S}$ is available, we can use it, and given $\mathbf{B}$ in (23), to obtain $\mathbf{L}_1$ and $\mathbf{\Delta}_1$, which gives (22).

3.2. Unique variances step

Loss function (15) can be expanded as

$$f(\mathbf{Q}, \mathbf{U}, \mathbf{R}, \mathbf{\Lambda}, \mathbf{\Psi}) = n\text{tr}\mathbf{S} + n\text{tr}\mathbf{\Lambda}\mathbf{R}'\mathbf{R}\mathbf{\Lambda}' + n\text{tr}\mathbf{\Psi}^2 - 2\text{tr}\mathbf{X}'\mathbf{Q}\mathbf{R}\mathbf{\Lambda}' - 2\text{tr}\mathbf{X}'\mathbf{U}\mathbf{\Psi} \tag{15$'$}$$

using (9), (11), and (12). The purpose of this step is to minimize (15′) over diagonal matrix $\mathbf{\Psi}$ with the other parameters fixed. Since (15′) can be rewritten as $||n^{1/2}\mathbf{\Psi} - n^{-1/2}\text{diag}(\mathbf{X}'\mathbf{U})||^2 + c^\#$ with $c^\#$ a constant independent of $\mathbf{\Psi}$, the minimizer is found to be given by

$$\mathbf{\Psi} = \frac{1}{n}\text{diag}(\mathbf{X}'\mathbf{U}) \,. \tag{24}$$

One can compare this with (22) and take into account $\mathbf{X}'\mathbf{Z} = \mathbf{X}'[\mathbf{Q}, \mathbf{U}] = [\mathbf{X}'\mathbf{Q}, \mathbf{X}'\mathbf{U}]$ to find that (24) is rewritten as

$$\mathbf{\Psi} = \text{diag}(\mathbf{B}'^{+}\mathbf{L}_1\mathbf{\Delta}_1\mathbf{L}_1'\mathbf{H}^p), \tag{25}$$

where $\mathbf{H}^p = [\mathbf{O}_{p\times m}, \mathbf{I}_p]'$ is a block matrix of size $(p+m) \times p$. Here, we should distinguish between $\mathbf{\Psi}$ on the left hand side of (25) and its counterpart in $\mathbf{B} = [\mathbf{\Lambda}\mathbf{R}',\mathbf{\Psi}]$ on the right hind side. The former $\mathbf{\Psi}$ is the updated one, while the latter one is $\mathbf{\Psi}$ from the previous iteration: (25) can be expressed as $\mathbf{\Psi}_{\text{new}} = \text{diag}([\mathbf{\Lambda}\mathbf{R}',\mathbf{\Psi}_{\text{old}}]'^{+}\mathbf{L}_1\mathbf{\Delta}_1\mathbf{L}_1'\mathbf{H}^p)$. Formula (25) shows that unique variances can be updated without the original data in $\mathbf{X}$ and the scores in $[\mathbf{Q}, \mathbf{U}]$.

3.3. Correlation step

The purpose of this step is to minimize (15′) over $\mathbf{R}$ subject to (6) with the other parameters. The name of the step follows from that $\mathbf{R}$ forms a factor correlation matrix as (16). We can use (24) in (15′) and have $f(\mathbf{Q}, \mathbf{R}, \mathbf{\Lambda}, \mathbf{\Psi}) = n\text{tr}\mathbf{S} + n\text{tr}\mathbf{\Lambda}\mathbf{R}'\mathbf{R}\mathbf{\Lambda}' - 2\text{tr}\mathbf{X}'\mathbf{Q}\mathbf{R}\mathbf{\Lambda}' - n\text{tr}\mathbf{\Psi}^2$, which is further rewritten as

$$f(\mathbf{R}, \mathbf{\Lambda}, \mathbf{\Psi}) = n(\text{tr}\mathbf{S} + \text{tr}\mathbf{\Lambda}'\mathbf{\Lambda} - 2\text{tr}\mathbf{YR\Lambda}' - \text{tr}\mathbf{\Psi}^2) \,, \qquad (15'')$$

where we have defined $\mathbf{Y}$ as $\mathbf{Y} = n^{-1}\mathbf{X}'\mathbf{Q}$ and used the fact that (10) and (14) imply $\text{tr}\mathbf{\Lambda R}'\mathbf{R\Lambda}' = \text{tr}\mathbf{R}'\mathbf{R\Lambda}'\mathbf{\Lambda} = \text{tr}\mathbf{\Lambda}'\mathbf{\Lambda}$. Function (15″) shows that our task is to maximize $\text{tr}\mathbf{YR\Lambda}' = \text{tr}\mathbf{\Lambda}'\mathbf{YR} = \text{tr}(\mathbf{Y}'\mathbf{\Lambda})'\mathbf{R} = \sum_{j=1}^{m}(\mathbf{Y}'\mathbf{\Lambda})_j'\mathbf{r}_j$ subject to (10), with $(\mathbf{Y}'\mathbf{\Lambda})_j$ denoting the jth column of $\mathbf{Y}'\mathbf{\Lambda}$. The maximization is easily attained by setting the $\mathbf{r}_{j1}$ in $\mathbf{r}_j = [\mathbf{r}_{j1}', \mathbf{0}_{m-j}']'$ at

$$\mathbf{r}_{j1} = \frac{(\mathbf{Y}'\mathbf{\Lambda})_{j1}}{\|(\mathbf{Y}'\mathbf{\Lambda})_{j1}\|} \qquad (26)$$

for $j \geq 2$, with $(\mathbf{Y}'\mathbf{\Lambda})_{j1}$ the $j \times 1$ vector containing the first j elements of $(\mathbf{Y}'\mathbf{\Lambda})_j = [(\mathbf{Y}'\mathbf{\Lambda})_{j1}', (\mathbf{Y}'\mathbf{\Lambda})_{j2}']'$ and r_{11} fixed at 1. Comparing $\mathbf{Y} = n^{-1}\mathbf{X}'\mathbf{Q}$ with (22) and considering $n^{-1}\mathbf{X}'\mathbf{Z} = n^{-1}[\mathbf{X}'\mathbf{Q}, \mathbf{X}'\mathbf{U}]$, we find that $\mathbf{Y}$ is given by

$$\mathbf{Y} = \mathbf{B}'^{+}\mathbf{L}_1\mathbf{\Delta}_1\mathbf{L}_1'\mathbf{H}_m \qquad (27)$$

where $\mathbf{H}_m = [\mathbf{I}_m, \mathbf{O}_{m\times p}]'$ is a block matrix of size $(p+m) \times m$. In (26) and (27), we can find $\mathbf{R}$ to be updated without $\mathbf{X}$ and $[\mathbf{Q}, \mathbf{U}]$.

3.4. Loadings step

The purpose of this step is minimizing (15″) over $\mathbf{\Lambda}$ subject to constraint (13) with the other parameters fixed. Using $\text{tr}\mathbf{\Lambda}'\mathbf{\Lambda} = \sum_{i=1}^{p}\sum_{j=1}^{m}\lambda_{ij}^2$ and $\text{tr}\mathbf{YR\Lambda}' = \sum_{i=1}^{p}\sum_{j=1}^{m}(\mathbf{y}_i'\mathbf{r}_j)\lambda_{ij}$ with $\mathbf{y}_i'$ the i-th row of $\mathbf{Y}$, we can rewrite (15″) into the form $f = n\sum_{i=1}^{p}\sum_{j=1}^{m}g_{ij}(\lambda_{ij}) + c^*$, where $c^* = n\text{tr}(\mathbf{S} - \mathbf{\Psi}^2)$ is irrelevant to $\mathbf{\Lambda}$ and

$$g_{ij}(\lambda_{ij}) = \lambda_{ij}^2 - 2(\mathbf{y}_i'\mathbf{r}_j)\lambda_{ij} \quad . \qquad (28)$$

Using $\lambda_{i,J(i)}$ for the element in $\boldsymbol{\lambda}_i'$ (the ith row of $\mathbf{\Lambda}$) to be given a nonzero value, the above purpose is found to be attained for

$$J(i) = \arg\min_{1\leq j\leq m} \min_{\lambda_{ij}} g_{ij}(\lambda_{ij}) \quad . \qquad (29)$$

Here, it follows from (28), that $\min_{\lambda_{ij}} g_{ij}(\lambda_{ij})$ is achieved for $\lambda_{ij} = \mathbf{y}_i'\mathbf{r}_j$. Taking this into account, (29) can be rewritten as $J(i) = \text{argmin}_{1\leq j\leq m}\, g_{ij}(\mathbf{y}_i'\mathbf{r}_j)$. Therefore, the update formula in this step is given by

$$\lambda_{ij} = \begin{cases} \mathbf{y}_i'\mathbf{r}_j & iff \;\; j = J(i) \\ 0 & \text{otherwise} \end{cases} \quad . \qquad (30)$$

The resulting loading matrix $\mathbf{\Lambda} = [\boldsymbol{\lambda}_1, \dots , \boldsymbol{\lambda}_p]'$ containing (30) satisfies

$$\mathrm{tr}\mathbf{YR\Lambda}' = \sum_{i=1}^{p}\sum_{j=1}^{m}(\mathbf{y}_i'\mathbf{r}_j)\lambda_{ij} = \sum_{i=1}^{p}(\mathbf{y}'_{i,J(i)}\mathbf{r}_{i,J(i)})\lambda_{i,J(i)} = \sum_{i=1}^{p}\lambda_{i,J(i)}^2 = \mathrm{tr}\mathbf{\Lambda\Lambda}', \tag{31}$$

which is used in the next subsection.

The update by (30) can produce empty columns $\mathbf{0}_p$ in $\mathbf{\Lambda}$, which violates assumption (20). We can deal with this problem, by restarting the algorithm with different initial values if such empty columns arose. As described in the next subsection, the initial $\mathbf{\Lambda}$ is chosen randomly, which implies that the initial values in the restart differ from the previous ones. Here, we must consider the two cases following the restart:

[1] Empty columns occur again, resulting that the algorithm cannot be terminated.

[2] The value of loss function (15) when empty columns arise is lower than the value for the solution obtained finally, which implies the possibility of the global minimizer including empty columns.

Cases [1] and [2] suggest that the data set is not suitable for SSFA and/or imply that m (the number of factors) is set to an improper one. In the simulation studies to be reported in Section 4, an empty column was never found. In the real data examples in Section 5, such columns rarely arose, but the restart always lead to a solution without an empty column and Case [2] was not observed.

3.5. Iterative Algorithm

The above results show that we can obtain the SSFA solution of $\mathbf{\Lambda}$, $\mathbf{\Psi}$, and $\mathbf{R}$ without the update of [$\mathbf{Q}$, $\mathbf{U}$] and without $n \times p$ data matrix $\mathbf{X}$, only if a $p \times p$ sample correlation matrix $\mathbf{S}$ is available which is more compact and thus easier to treat than $\mathbf{X}$. The algorithm for SSFA is formally stated as the following list of steps:

Step 1. Initialize $\mathbf{\Lambda}$, $\mathbf{\Psi}$, and $\mathbf{R}$.
Step 2. Perform EVD (23)
Step 3. Update $\mathbf{\Psi}$ with (25)
Step 4. Obtain $\mathbf{Y}$ with (27)
Step 5. Update of the columns of $\mathbf{R}$ with (26)
Step 6. Update the elements of $\mathbf{\Lambda}$ with (30)
Step 7. Go back to Step 1 if $\mathbf{\Lambda}$ has an empty column.
Step 8. Finish if convergence is reached; otherwise, go back to Step 2.

For checking convergence in the above Step 8, we can use the standardized loss function value

$$f_{\mathrm{S}}(\mathbf{\Theta}) = 1 - \frac{\mathrm{tr}\,\mathbf{\Lambda\Lambda}' + \mathrm{tr}\,\mathbf{\Psi}^2}{\mathrm{tr}\,\mathbf{S}} \quad . \tag{32}$$

with $\mathbf{\Theta} = \{\mathbf{\Lambda}, \mathbf{\Psi}, \mathbf{\Phi}\}$ a set of parameter matrices. Function (32) is derived as follows: we can use (31) in (15″) to have $f(\mathbf{\Theta}) = n\mathrm{tr}\mathbf{S} - n\mathrm{tr}\mathbf{\Lambda\Lambda}' - n\mathrm{tr}\mathbf{\Psi}^2$, whose division by $n\mathrm{tr}\mathbf{S}$ gives (32) with $0 \le f_{\mathrm{S}}(\mathbf{\Theta}) \le 1$. Although $\mathbf{\Phi}$ is not included in the right hand side of (32), it depends on $\mathbf{\Phi}$, since the optimal $\mathbf{\Lambda}$ is a function of $\mathbf{R}$ forming $\mathbf{\Phi}$ as found in (30). In this paper, the convergence is defined as the decrease of $f_{\mathrm{S}}(\mathbf{\Theta})$ from the previous round being less than 0.1^5. As the range of $f_{\mathrm{S}}(\mathbf{\Theta})$ is [0, 1], such a decrease is small enough to be neglected.

In Step 1, we initialize $\mathbf{R}$ at $\mathbf{I}_m$ and $\mathbf{\Psi}$ at $\mathrm{diag}(\mathbf{I}_p - \mathbf{\Lambda\Lambda}')^{1/2}$ using $\mathbf{\Lambda}$ chosen as follows. The nonzero-elements of $\mathbf{\Lambda}$ are randomly chosen subject to that it is the sparsest and each column has at least three nonzero loadings. Each nonzero loading was randomly drawn from $U(0.5, 0.98)$ or $U(-0.98, -0.5)$ with $U(\alpha, \beta)$ denoting the uniform distribution defined for range $[\alpha, \beta]$. The above α and β values were chosen, supposing that SSFA is performed for standardized data and the resulting loadings tend to take the values within $[-1, 1]$.

3.6. Multiple Run Procedure

It should be noted that the presented iterative algorithm is not guaranteed to converge to the global minima, since the sets of parameters are alternately updated. To increase the possibility of obtaining the global minimizer, we start the algorithm with multiple times with different initial $\mathbf{\Lambda}$. Among the resulting solutions, the one attaining the lowest loss function value is selected as the optimal solution. This procedure for selecting the optimal solution is detailed in the next paragraph. The issue of local minima and empty columns is also a problem in other clustering procedures, e.g. the popular *k*-means clustering (Gan, et al., 2007).

Let us use $\mathbf{\Theta}_l = \{\mathbf{\Lambda}_l, \mathbf{\Psi}_l, \mathbf{\Phi}_l\}$ for the solution resulting from the *l*th run of the SSFA and use $f_{\mathrm{s}}(\mathbf{\Theta}_l)$ for the corresponding loss function value (32) with $l = 1, \dots, L$. We regard $\mathbf{\Theta}_{l^*}$ with $l^* = \mathrm{argmin}_{1\le l\le L}\, f(\mathbf{\Theta}_l)$ as the optimal solution, and define $\mathbf{\Theta}_l$ being a local minimizer as $\Delta(\mathbf{\Theta}_l, \mathbf{\Theta}_{l^*}) = p^{-1}\Sigma_i | \lambda_{i,\#}^{[l]} - \lambda_{i,\#}^{[l^*]} | + p^{-1}\Sigma_i | \psi_i^{[l]2} - \psi_i^{[l^*]2} | + M^{-1}\Sigma_{j<k} | \phi_{jk}^{[l]} - \phi_{jk}^{[l^*]} | > 3 \times 0.1^3$ with $\lambda_{i,\#}^{[l]}$ being the nonzero element of the *i*th row of $\mathbf{\Lambda}_l$, $\psi_i^{[l]}$ the *i*th diagonal element of $\mathbf{\Psi}_l$, and $\mathbf{\Phi}_l = (\phi_{jk}^{[l]})$. Here, the suitable L (number of runs) is unknown beforehand. We thus employ a strategy in which L is initialized at an integer and increased until L is considered to be sufficient. We define the sufficiency as that the solutions $\mathbf{\Theta}_1, \dots, \mathbf{\Theta}_L$ resulting from L runs include the two equivalently optimal solutions $\mathbf{\Theta}_{l^*}$ and $\mathbf{\Theta}_{l^\#}$ satisfying $\Delta(\mathbf{\Theta}_{l^*}, \mathbf{\Theta}_{l^\#}) \le 0.1^3$ and $l^* = \mathrm{argmin}_{1\le l\le L}\, f(\mathbf{\Theta}_l)$ with $l^\# \ne l^*$. Our strategy can thus be called a *two-optimal-solutions stopping* procedure, which is formally stated as follows:

[1] Set L = 50 and obtain $l^*= \text{argmin}_{1\leq l\leq L}\, f(\mathbf{\Theta}_l)$.

[2] Go to [6] if $\Delta(\mathbf{\Theta}_{l^*}, \mathbf{\Theta}_{l^\#}) \leq 3 \times 0.1^3$ with $l^\# \neq l^*$ and $1 \leq l^\# \leq L$; otherwise, go to [3].

[3] Set $L := L + 1$, and let $\mathbf{\Theta}_{l^\#}$ be the output from another run.

[4] Exchange $\mathbf{\Theta}_{l^*}$ for $\mathbf{\Theta}_{l^\#}$ if $f(\mathbf{\Theta}_{l^\#}) < f(\mathbf{\Theta}_{l^*})$.

[5] Go to [6] if $\Delta(\mathbf{\Theta}_{l^*}, \mathbf{\Theta}_{l^\#}) \leq 3 \times 0.1^3$ or L = 200; otherwise, go back to [3].

[6] Finish with choosing $\mathbf{\Theta}_{l^*}$ as the optimal solution.

4. Simulation Studies

In this section, we report the simulation studies for assessing how well the true variable clusters and parameter values can be recovered by SSFA. For the purpose of variable clustering, it is more persuasive if the SSFA usefulness is demonstrated with comparisons to competitors. Thus, SSFA is compared to the existing sparse FA procedures, SOFA and FANC. Their purposes are not variable clustering itself, but they can be enforced to eventually produce the sparsest loadings with (13). Indeed, SOFA may produce such loadings without its factor orthogonality restriction seriously influencing loading estimation. FANC may also yield the sparsest loadings if tuning parameters are suitably chosen.

4.1. True Parameters and Data Synthesis

We performed a small simulation study in order to see how well the parameters with the sparsest loadings are recovered by SSFA, SOFA, and FANC. The Panel (A) in Tables 1, 2, and 3 shows our considered sets of the true $\mathbf{\Lambda}$, $\mathbf{\Psi}^2$, and $\mathbf{\Phi}$ with the numbers of variables and factors $\{p, m\}$ set to {15, 3}, {20, 4}, and {24, 5}. From each set of $\mathbf{\Lambda}$, $\mathbf{\Psi}^2$, and $\mathbf{\Phi}$, we synthesized a data set as follows: an $n \times p$ data matrix $\mathbf{X}$ with $n = 10 \times p$ is generated, each of whose row was sampled from the p-variate normal distribution $N_p(\mathbf{0}_p, \mathbf{\Sigma})$ with mean vector $\mathbf{0}_p$ and covariance matrix $\mathbf{\Sigma} = \mathbf{\Lambda\Phi\Lambda}' + \mathbf{\Psi}^2$, respectively. Every set of $\mathbf{\Lambda}$, $\mathbf{\Psi}^2$, and $\mathbf{\Phi}$ in Tables 1, 2, and 3 satisfies $\text{diag}(\mathbf{\Sigma}) = \mathbf{I}_p$ so that the true $\mathbf{\Sigma}$ was a correlation matrix. Thus, a $p \times p$ sample correlation matrix obtained from $\mathbf{X}$ was to be analyzed. The reasons why we choose $N_p(\mathbf{0}_p, \mathbf{\Sigma})$ to generate $\mathbf{X}$ and why we analyze the correlations (rather than covariances) are described in the following two paragraphs in turn.

The model part $\mathbf{X}^* = \mathbf{QR\Lambda}' + \mathbf{U\Psi}$ in SSFA loss function (15) leads to the covariance matrix $n^{-1}\mathbf{X}^{*\prime}\mathbf{X}^* = \mathbf{\Sigma} = \mathbf{\Lambda\Phi\Lambda}' + \mathbf{\Psi}^2$, since of (9), (11), (12), and (16). In FANC, the covariance matrix has also the same form $\mathbf{\Sigma}$. Further, normality $\mathbf{x} \sim N_p(\mathbf{0}_p, \mathbf{\Sigma})$ is assumed under the formulation $\mathbf{x} = \mathbf{\Lambda f} + \mathbf{\Psi u}$ introduced in Section 1. On the other hand, SOFA and SSFA are based on the least squares formulation (4) and have no normality assumption. We choose the FANC assumption $\mathbf{x} \sim N_p(\mathbf{0}_p, \mathbf{\Sigma})$ for the simulation study in which a data distribution must be specified. In a

Table 1. True parameters with m = 3 and their estimated counterparts

(A)True				(B) SSFA				(C) SOFA				(D) FANC[*]			
	$\mathbf{\Lambda}$		$\mathbf{\Psi}^2$		$\mathbf{\Lambda}$		$\mathbf{\Psi}^2$		$\mathbf{\Lambda}$		$\mathbf{\Psi}^2$		$\mathbf{\Lambda}$		$\mathbf{\Psi}^2$
0.9	.	.	0.19	0.87	.	.	0.23	0.86	.	.	0.24	0.87	.	.	0.24
-0.8	.	.	0.36	-0.83	.	.	0.30	-0.83	.	.	0.29	-0.83	.	.	0.32
0.7	.	.	0.51	0.73	.	.	0.46	0.71	.	.	0.47	0.72	.	.	0.48
-0.6	.	.	0.64	-0.56	.	.	0.66	-0.54	.	.	0.68	-0.54	.	.	0.71
0.5	.	.	0.75	0.44	.	.	0.80	0.43	.	.	0.80	0.45	.	.	0.80
-0.4	.	.	0.84	-0.36	.	.	0.86	-0.36	.	.	0.86	-0.37	.	.	0.86
.	0.8	.	0.36	.	0.78	.	0.37	0.39	.	-0.45	0.57	.	0.78	.	0.39
.	-0.7	.	0.51	.	-0.72	.	0.46	-0.48	.	0.46	0.52	.	-0.71	.	0.49
.	0.6	.	0.64	.	0.51	.	0.72	.	.	.	0.88	.	0.52	.	0.73
.	-0.5	.	0.75	.	-0.41	.	0.82	.	.	.	0.92	.	-0.41	.	0.83
.	0.4	.	0.84	.	0.32	.	0.89	.	0.34	.	0.83	.	0.31	.	0.90
.	.	0.7	0.51	.	.	0.74	0.45	.	.	0.70	0.48	.	.	0.72	0.48
.	.	-0.6	0.64	.	.	-0.45	0.79	.	.	-0.47	0.77	.	.	-0.46	0.79
.	.	0.5	0.75	.	.	0.60	0.63	.	.	0.56	0.66	.	.	0.60	0.64
.	.	-0.4	0.84	.	.	-0.50	0.74	.	.	-0.44	0.77	.	.	-0.50	0.75
	$\mathbf{\Phi}$				$\mathbf{\Phi}$				$\mathbf{\Phi}$				$\mathbf{\Phi}$		
1	0.4	0.3		1.00	0.48	0.33		1.00	.	.		1.00	0.47	0.35	
0.4	1	-0.4		0.48	1.00	-0.41		.	1.00	.		0.47	1.00	-0.41	
0.3	-0.4	1		0.33	-0.41	1.00		.	.	1.00		0.35	-0.41	1.00	

[*] ρ = 0.26; γ = 1.01

sense, this choice favors FANC and is conservative for testing SSFA.

As detailed in Adachi (2012), the MDFA loss function (4) is not scale-free as MLFA in (2) (e.g., Harman, 1976). Thus, the MDFA and MDFA-based SSFA solutions for covariances are essentially different from those for the correlations obtained from the same data set. Thus, when SSFA is performed for covariances, the inter-variable differences in variances (i.e., scales) influence solutions. In the simulation studies, and also in the next section, we consider cases where such influences are avoided. Then, FA is performed on correlations with the variances homogeneous among the variables.

4.2. Illustrative Results

For each of the resulting three data sets, we carried out SSFA, SOFA, and FANC with m set to the true number. In SOFA and FANC, obtaining a solution with Card($\mathbf{\Lambda}$) = p is required for allowing the resulting $\mathbf{\Lambda}$ to be the sparsest. Thus, Card($\mathbf{\Lambda}$) is set to p in SOFA formulated as (7). In FANC (6), the correspondence of the values of tuning parameters ρ and γ to Card($\mathbf{\Lambda}$) cannot be known beforehand, i.e., it can be found by only looking at the solution resulting in the trial with ρ and γ set to specific values, although FANC provides the best combination of ρ and γ giving the solution with the least value of BIC = $2l(\mathbf{\Lambda},\mathbf{\Psi},\mathbf{\Phi}) + df(\rho,\ \gamma)\log n$. Here, $l(\mathbf{\Lambda},\mathbf{\Psi},\mathbf{\Phi})$ is the resulting value of the negative log likelihood (2), and the degree of freedom $df(\rho,\ \gamma)$, which is a function of ρ and γ, is obtained according to Mazumder, Friedman, & Hastie (2011) (Hirose & Yamamoto, 2014). However, the resulting combination is not guaranteed to give a solution with the desired Card($\mathbf{\Lambda}$). Indeed, this BIC-based selection gave the solution of Card($\mathbf{\Lambda}$) > p for every data set. Thus, we increased the value of ρ (penalty weight) to find a solution with a smaller Card($\mathbf{\Lambda}$) while γ was kept fixed. It gave the solutions with Card($\mathbf{\Lambda}$) = p, as found in Tables 1, 2, and 3, where SSFA and FANC solutions are also presented.

First, let us note the results in Table 1 for m = 3. We can find that SSFA and FANC

Table 2. True parameters with m = 4 and their estimated counterparts

(A)True					(B) SSFA					(C) SOFA					(D) FANC*				
Λ				**Ψ**^{2}	**Λ**				**Ψ**^{2}	**Λ**				**Ψ**^{2}	**Λ**				**Ψ**^{2}
0.9	.	.	.	0.19	0.89	.	.	.	0.20	0.86	.	.	.	0.20	0.89	.	.	.	0.21
-0.8	.	.	.	0.36	-0.81	.	.	.	0.33	-0.77	.	.	.	0.34	-0.81	.	.	.	0.35
0.7	.	.	.	0.51	0.67	.	.	.	0.54	0.67	.	0.39	.	0.39	0.68	.	.	.	0.54
-0.6	.	.	.	0.64	-0.56	.	.	.	0.67	-0.54	.	.	.	0.68	-0.56	.	.	.	0.69
0.5	.	.	.	0.75	0.55	.	.	.	0.68	0.51	.	.	.	0.70	0.55	.	.	.	0.70
-0.4	.	.	.	0.84	-0.39	.	.	.	0.82	-0.40	.	.	.	0.81	-0.39	.	.	.	0.85
.	0.8	.	.	0.36	.	0.77	.	.	0.40	.	0.72	.	.	0.42	.	0.76	.	.	0.42
.	-0.7	.	.	0.51	.	-0.72	.	.	0.48	.	-0.73	.	.	0.44	.	-0.72	.	.	0.48
.	0.6	.	.	0.64	.	0.56	.	.	0.68	.	0.54	.	.	0.68	.	0.57	.	.	0.68
.	-0.5	.	.	0.75	.	-0.53	.	.	0.71	.	-0.48	.	.	0.74	.	-0.52	.	.	0.73
.	0.4	.	.	0.84	.	0.42	.	.	0.81	.	0.39	.	.	0.82	.	0.42	.	.	0.82
.	.	0.7	.	0.51	.	.	0.63	.	0.59	0.43	.	.	.	0.75	0.35	.	.	.	0.87
.	.	-0.6	.	0.64	.	.	-0.64	.	0.57	-0.41	.	.	.	0.77	-0.36	.	.	.	0.87
.	.	0.5	.	0.75	.	.	0.45	.	0.78	0.39	.	.	.	0.84	0.35	.	.	.	0.88
.	.	-0.4	.	0.84	.	.	-0.25	.	0.92	.	.	.	.	0.96	.	.	-0.99	.	0.02
.	.	.	0.8	0.36	.	.	.	0.84	0.27	.	.	.	0.83	0.27	.	.	.	0.85	0.28
.	.	.	-0.7	0.51	.	.	.	-0.74	0.44	.	.	.	-0.74	0.44	.	.	.	-0.74	0.45
.	.	.	0.6	0.64	.	.	.	0.64	0.58	.	.	.	0.62	0.59	.	.	.	0.63	0.60
.	.	.	-0.5	0.75	.	.	.	-0.46	0.78	.	.	.	-0.47	0.77	.	.	.	-0.46	0.79
.	.	.	0.4	0.84		.	.	0.46	0.77	.	.	.	0.46	0.77	.	.	.	0.46	0.79
Φ					**Φ**					**Φ**					**Φ**				
1.0	0.4	0.3	-0.2		1.00	0.49	0.58	-0.15		1.00	.	.	.		1.00	0.45	0.11	-0.18	
0.4	1.0	-0.4	0.3		0.49	1.00	-0.13	0.21		.	1.00	.	.		0.45	1.00	-0.03	0.19	
0.3	-0.4	1.0	-0.3		0.58	-0.13	1.00	-0.19		.	.	1.00	.		0.11	-0.03	1.00	-0.06	
-0.2	0.3	-0.3	1.0		-0.15	0.21	-0.19	1.00		.	.	.	1.00		-0.18	0.19	-0.06	1.00	

*ρ = 0.15; γ = 1.01

successfully provided the sparsest **Λ**, i.e., identified the locations of the nonzero loadings. But, SOFA failed to give the sparsest **Λ**, with two variables not loading any factor. It implies that the variations in the two variables are not explained by the common factors, which is contradictory to the model underlying the data set. For comparing SSFA and FANC in the goodness of recovering nonzero parameter values, we obtained the averaged absolute differences AAD(**Λ**) = $p^{-1}\Sigma_i|\hat{\lambda}_{i,\#}-\lambda_{i,\#}|$, AAD(**Ψ**^{2}) = $p^{-1}\Sigma_i|\hat{\psi}_i^2-\psi_i^2|$, and AAD(**Φ**) = $M^{-1}\Sigma_{j<k}|\hat{\phi}_{jk}-\phi_{jk}|$. Here, $\lambda_{i,\#}$ denotes the non-zero element in the ith row of the true **Λ**, ψ_i^2 is the true unique variance for variable i, and ϕ_{jk} is the (j, k) element of the true **Φ**, with $\hat{\lambda}_{i,\#}$, $\hat{\psi}_i^2$, and $\hat{\phi}_{jk}$ the estimated counterparts and M = m(m–1)/2. The resulting values were AAD(**Λ**) = 0.02, AAD(**Ψ**^{2}) = 0.06, and AAD(**Φ**) = 0.04 for both SSFA and FANC: they presented the equivalent AAD values which are small enough to show good recovery of parameters.

Next, we note the results in Table 2 for m = 4. We can find that only SSFA succeeded in recovering the sparsest **Λ**, while SOFA and FANC failed it with only one variable loading the third factor in their solutions. For the SSFA solution, AAD(**Λ**) = 0.01, AAD(**Ψ**^{2}) = 0.05, and AAD(**Φ**) = 0.15: loadings and unique variances were recovered very well, though two estimated factor correlations (0.58 and –0.13) were far from the true values (0.3 and –0.4), which implies that the recovery of each parameter cannot be indicated by AAD values.

Finally, let us note the solutions in Table 3 for m = 5. It is found that SSFA and FANC successfully recovered the sparsest **Λ**, while SOFA failed. For the SSFA solution, AAD(**Λ**) = 0.01, AAD(**Ψ**^{2}) = 0.05, and AAD(**Φ**) = 0.05, while AAD(**Λ**) = 0.01, AAD(**Ψ**^{2}) = 0.04, and

Table 3. True parameters with m = 5 and their estimated counterparts

(A)True						(B) SSFA						(C) SOFA						(D) FANC*					
		Λ			**Ψ**^{2}			**Λ**			**Ψ**^{2}			**Λ**			**Ψ**^{2}			**Λ**			**Ψ**^{2}
0.9	.	.	.	.	0.19	0.88	.	.	.	.	0.22	0.84	.	.	.	.	0.23	0.88	.	.	.	.	0.22
-0.8	.	.	.	.	0.36	-0.82	.	.	.	.	0.32	-0.82	.	.	.	.	0.30	-0.82	.	.	.	.	0.33
0.7	.	.	.	.	0.51	0.68	.	.	.	.	0.53	0.68	.	.	.	.	0.52	0.68	.	.	.	.	0.54
-0.6	.	.	.	.	0.64	-0.64	.	.	.	.	0.58	-0.62	.	.	.	.	0.59	-0.63	.	.	.	.	0.61
0.5	.	.	.	.	0.75	0.46	.	.	.	.	0.78	0.43	.	.	.	.	0.80	0.46	.	.	.	.	0.79
-0.4	.	.	.	.	0.84	-0.47	.	.	.	.	0.76	-0.46	.	.	.	.	0.76	-0.47	.	.	.	.	0.78
.	0.8	.	.	.	0.36	.	0.83	.	.	.	0.31	.	0.83	.	.	.	0.26	.	0.83	.	.	.	0.32
.	-0.7	.	.	.	0.51	.	-0.68	.	.	.	0.53	.	-0.61	.	.	.	0.56	.	-0.67	.	.	.	0.55
.	0.6	.	.	.	0.64	.	0.60	.	.	.	0.63	.	0.56	.	.	.	0.64	.	0.60	.	.	.	0.64
.	-0.5	.	.	.	0.75	.	-0.52	.	.	.	0.72	.	-0.51	.	.	.	0.71	.	-0.52	.	.	.	0.73
.	0.4	.	.	.	0.84	.	0.41	.	.	.	0.82	.	0.38	.	.	.	0.82	.	0.41	.	.	.	0.83
.	.	0.7	.	.	0.51	.	.	0.72	.	.	0.47	.	-0.43	.	.	.	0.72	.	.	0.70	.	.	0.51
.	.	-0.6	.	.	0.64	.	.	-0.60	.	.	0.63	.	.	.	.	.	0.87	.	.	-0.59	.	.	0.66
.	.	0.5	.	.	0.75	.	.	0.47	.	.	0.77	.	.	.	.	.	0.91	.	.	0.48	.	.	0.77
.	.	-0.4	.	.	0.84	.	.	-0.52	.	.	0.71	.	.	.	.	.	0.89	.	.	-0.53	.	.	0.72
.	.	.	0.8	.	0.36	.	.	.	0.77	.	0.39	.	0.31	.	0.72	.	0.36	.	.	.	0.78	.	0.40
.	.	.	-0.7	.	0.51	.	.	.	-0.65	.	0.56	.	.	.	-0.63	.	0.56	.	.	.	-0.65	.	0.58
.	.	.	0.6	.	0.64	.	.	.	0.60	.	0.62	.	.	.	0.57	.	0.62	.	.	.	0.60	.	0.64
.	.	.	-0.5	.	0.75	.	.	.	-0.48	.	0.76	.	.	.	-0.49	.	0.74	.	.	.	-0.49	.	0.76
.	.	.	0.4	.	0.84	.	.	.	0.47	.	0.77	.	.	.	0.40	.	0.79	.	.	.	0.45	.	0.80
.	.	.	.	0.7	0.51	.	.	.	.	0.76	0.42	.	-0.34	.	.	0.62	0.45	.	.	.	.	0.75	0.43
.	.	.	.	-0.6	0.64	.	.	.	.	-0.64	0.58	.	.	-0.56	.	-0.64	0.22	.	.	.	.	-0.64	0.59
.	.	.	.	0.5	0.75	.	.	.	.	0.52	0.71	.	.	.	.	0.53	0.68	.	.	.	.	0.53	0.71
.	.	.	.	-0.4	0.84	.	.	.	.	-0.53	0.71	.	.	.	.	-0.49	0.71	.	.	.	.	-0.51	0.74
		Φ						**Φ**						**Φ**						**Φ**			
1.0	0.4	0.3	-0.2	0.3		1.00	0.29	0.32	-0.25	0.33		1.00	.	.	.	.		1.00	0.27	0.32	-0.24	0.34	
0.4	1.0	-0.4	0.3	-0.4		0.29	1.00	-0.46	0.40	-0.47		.	1.00	.	.	.		0.27	1.00	-0.49	0.40	-0.48	
0.3	-0.4	1.0	-0.3	0.2		0.32	-0.46	1.00	-0.37	0.27		.	.	1.00	.	.		0.32	-0.49	1.00	-0.38	0.26	
-0.2	0.3	-0.3	1.0	-0.3		-0.25	0.40	-0.37	1.00	-0.37		.	.	.	1.00	.		-0.24	0.40	-0.38	1.00	-0.35	
0.3	-0.4	0.2	-0.3	1.0		0.33	-0.47	0.27	-0.37	1.00		.	.	.	.	1.00		0.34	-0.48	0.26	-0.35	1.00	

*ρ = 0.26; γ = 1.01

AAD(**Φ**) = 0.06 for the FANC solution. Those results do not show substantial differences between the two methods, with the AAD values beings small enough and showing the good recovery of parameter values.

In summary, only SSFA exactly recovered the sparsest loadings for the three data sets, while SOFA failed for all of them and FANC failed for m = 4. We can also consider that the performances of FANC was good in that it succeeded the recovery of the sparsest loadings for the two of the data sets. However, it needs the task for searching out the tuning parameters that provide sparsest loadings. On the other hand, they are given straightforwardly in SSFA. This easiness and the good performances demonstrated that SSFA is the most suitable for finding the sparsest loadings.

4.3. Results for Multiple Data Sets

We replicated the data generation in Section 4.1 to have 15 data sets for each m. The resulting data were analyzed by the three procedures. For FANC, it is too laborious to find the tuning parameters that provide sparsest loadings for each of the multiple data sets. Thus, FANC was performed with ρ = 0.26 and γ = 1.01 chosen for m = 3 and 5 in Section 4.1. Those tuning parameters values were also used for m = 4, as they were found to give better results than ρ = 0.14 and γ = 1.01 in Table 2. FANC was also performed with BIC-based selection of tuning parameters.

Table 4 shows the resulting averages and standard deviations of MIS$_0$(**Λ**) and AAD values over 15 data sets. Here, MIS$_0$(**Λ**) stands for the misidentification rates of zero/nonzero loadings: MIS$_0$(**Λ**) = $w/(pm)$ with w the number of the resulting λ_{ij} that are zero in spite of the

Table 4. Averages (Ave) and standard deviations (Std) of the differences of solutions from the true counterparts.

m	Index	SSFA		SOFA		FANC*		FANC$_{\text{BIC}}$	
		Ave	Std	**Ave**	Std	**Ave**	Std	**Ave**	Std
3	Mis$_0$($\boldsymbol{\Lambda}$)	**0.006**	0.015	**0.172**	0.074	**0.028**	0.045	**0.043**	0.050
	AAD($\boldsymbol{\Lambda}$)	**0.019**	0.006	**0.198**	0.010	**0.028**	0.024	**0.034**	0.025
	AAD($\boldsymbol{\Psi}^2$)	**0.057**	0.009	**0.171**	0.012	**0.062**	0.016	**0.062**	0.016
	AAD($\boldsymbol{\Phi}$)	**0.077**	0.029		0.000	**0.100**	0.060	**0.164**	0.173
4	Mis$_0$($\boldsymbol{\Lambda}$)	**0.000**	0.000	**0.100**	0.032	**0.013**	0.037	**0.024**	0.045
	AAD($\boldsymbol{\Lambda}$)	**0.011**	0.002	**0.146**	0.007	**0.016**	0.019	**0.019**	0.020
	AAD($\boldsymbol{\Psi}^2$)	**0.050**	0.009	**0.162**	0.010	**0.053**	0.018	**0.054**	0.016
	AAD($\boldsymbol{\Phi}$)	**0.063**	0.020		0.000	**0.067**	0.038	**0.072**	0.037
5	Mis$_0$($\boldsymbol{\Lambda}$)	**0.011**	0.014	**0.118**	0.039	**0.021**	0.041	**0.048**	0.060
	AAD($\boldsymbol{\Lambda}$)	**0.014**	0.008	**0.120**	0.007	**0.019**	0.021	**0.031**	0.039
	AAD($\boldsymbol{\Psi}^2$)	**0.055**	0.011	**0.168**	0.011	**0.056**	0.020	**0.054**	0.015
	AAD($\boldsymbol{\Phi}$)	**0.068**	0.029		0.000	**0.085**	0.055	**0.091**	0.048

$^*\rho = 0.26$; $\gamma = 1.01$

true λ_{ij} being nonzero or vice versa. In this paper, the standard deviations are used which are not unbiased. In Table 4, we can find that SSFA shows the smallest average among the procedures, which indicates the best recovery, for every case, except for AAD($\boldsymbol{\Psi}^2$) when m = 5. Further, it can be considered that the averages of SSFA are satisfactorily small enough and the corresponding standard deviations are also small, which implies that a solution with exceptionally bad recovery was not obtained.

In order to compare SSFA more directly with the others, we subtracted an SSFA index value from another procedure's one for each data set: d_i = Index$_i$ – Index$_i^{[\text{SS}]}$ was obtained, where Index$_i^{[\text{SS}]}$ is the value of an index obtained with SSFA for data set i (=1, … , 15), while Index$_i$ is the corresponding value for another procedure. The resulting average $\bar{d}$ and standard deviation SD_d for $d_1,\dots, d_{15}$ gave the t-statistic t = $\bar{d}$ /{SD_d/(15–1)$^{1/2}$}, which are presented in Table 5. Although it is not obvious whether this statistic follows the t-distribution with the degree-of-freedom 15–1, the 95 percentile 1.76 for that distribution can be used as a benchmark. Thus, the t-values exceeding 1.76 have been bold-faced in Table 5. It shows that SOFA is substantially inferior to SSFA, and FANC with the BIC-based selection is so in the

Table 5. t-statistics indicating the inferiority from SSFA solutions

m	Index	SOFA	FANC*	FANC$_{\text{BIC}}$
3	MIS$_0$($\boldsymbol{\Lambda}$)	**7.74**	**1.84**	**2.78**
	AAD($\boldsymbol{\Lambda}$)	**54.19**	1.41	**2.27**
	AAD($\boldsymbol{\Psi}^2$)	**24.64**	1.65	1.46
	AAD($\boldsymbol{\Phi}$)		1.27	1.81
4	MIS$_0$($\boldsymbol{\Lambda}$)	**11.83**	1.36	**2.05**
	AAD($\boldsymbol{\Lambda}$)	**77.31**	1.20	1.64
	AAD($\boldsymbol{\Psi}^2$)	**33.30**	0.73	1.03
	AAD($\boldsymbol{\Phi}$)		0.59	1.28
5	MIS$_0$($\boldsymbol{\Lambda}$)	**10.66**	0.95	**2.36**
	AAD($\boldsymbol{\Lambda}$)	**43.15**	0.98	1.65
	AAD($\boldsymbol{\Psi}^2$)	**38.05**	0.37	-0.17
	AAD($\boldsymbol{\Phi}$)		1.45	**2.17**

$^*\rho = 0.26$; $\gamma = 1.01$

identification of zero/non-zero loadings. Whether FANC with $\rho = 0.26$ and $\gamma = 1.01$ is inferior to SSFA is inconclusive, but it is superior in that SSFA does not involve the cumbersome procedures for tuning parameters as in FANC, as described often so far.

5. Real Data Examples

For illustration, we use three real data sets. SSFA, SOFA, and FANC are carried out for the first of the three sets, as its variables are known to have "population" sparsest loadings. For the remaining two data sets, such prior knowledge does not exist. Thus, we performed only SSFA to illustrate how useful its variable clustering is. In SSFA, the empty columns (i.e., $\mathbf{0}_p$ included in $\mathbf{\Lambda}$) described in Section 3.4 occurred in the 1%, 8%, and, 10% of all runs, for the data sets in Sections 5.1, 5.2, and 5.3, respectively, but the restart always lead to a solution without such a column: the SSFA algorithm never needed cancellation.

5.1. Big-five Data

In this section, we use the 25 × 25 correlation matrix obtained from the 190-participants × 25-items data matrix collected by the first author and available as "Big Five Personality Test Data" from http://bstat.jp/en_material/. This data set contains the self-ratings of the persons (university students) to what extent they are characterized by the personalities described by the 25 items. According to a theory in personality psychology, the items can be exclusively classified into the five groups shown in the first column of Table 6 (Costa & McCrae, 1992; Goldberg, 1992). Obviously, for this data set, the sparsest $\mathbf{\Lambda}$ should be estimated, in which the five items within a group load one of $m = 5$ factors.

We set $m = 5$ to perform SSFA, SOFA, and FANC, with Card($\mathbf{\Lambda}$) set to 25 in SOFA. In FANC, we took the same procedure as in the last subsection for finding the tuning parameters ρ and γ that provide the loadings Card($\mathbf{\Lambda}$) = 25: the BIC-based suggested the solution with $\rho = 0.17$, $\gamma = 1.01$, and Card($\mathbf{\Lambda}$) = 49 > 25. Thus, we increased the value of penalty weight ρ with keeping γ fixed. But, the solution with Card($\mathbf{\Lambda}$) = 25 was not found, though the ones with Card($\mathbf{\Lambda}$) = 25±1 were found, since ρ cannot take arbitrary real numbers, but distinct values, in FANC: ρ could not take a value which exactly corresponds to Card($\mathbf{\Lambda}$) = 25. Therefore, we change both ρ and γ values, in order to heuristically find a solution with Card($\mathbf{\Lambda}$) = 25. The resulting solution with Card($\mathbf{\Lambda}$) = 25 is presented in Table 6 together with the SSFA and SOFA solutions.

It is found in Table 6(A), that SSFA gives the sparsest $\mathbf{\Lambda}$ predicted by the theory in personality psychology. On the other hand, SOFA and FANC are found to fail providing such loadings, see Table 6(B) and (C). In the SOFA solution, only two items load the third factor,

Table 6. Solutions for big-five data

Variable	(A) SSFA						(B) SOFA						(C) FANC					
			Λ			Ψ²			Λ			Ψ²			Λ			Ψ²
worry	0.75	.	.	.	.	0.40	0.71	.		.	.	0.41	0.41	.	.	.	.	0.57
sensitive	0.65	.	.	.	.	0.52	0.70	.	0.35	.	.	0.37	0.23	.	.	.	.	0.78
pessimistic	0.79	.	.	.	.	0.34	0.72	.		.	.	0.37	0.52	.	.	.	.	0.41
unrest	0.42	.	.	.	.	0.71	0.48	.	0.26	.	-0.31	0.57	.	.	.	.	.	1.00
careful	0.71	.	.	.	.	0.45	0.66	.	.	.	.	0.47	0.44	.	.	.	.	0.53
sociable	.	0.85	.	.	.	0.26	.	0.83	.	.	.	0.26	.	0.58	.	.	.	0.29
talkative	.	0.75	.	.	.	0.38	.	0.80	.	.	.	0.34	.	0.45	.	.	.	0.49
voluntary	.	0.76	.	.	.	0.39	.	0.73	.	.	.	0.41	.	0.47	.	.	.	0.46
cheerful	.	0.83	.	.	.	0.27	.	0.79	.	.	.	0.29	.	0.55	.	.	.	0.33
showy	.	0.64	.	.	.	0.55	.	0.64	.	.	.	0.55	.	0.33	.	.	.	0.66
creative	.		0.73	.	.	0.41	.		.	.	.	0.78	.		0.41	.	.	0.59
adventurous	.		0.78	.	.	0.36	.		.	.	.	0.72	.		0.65	.	.	0.25
progressive	.		0.68	.	.	0.50	.		.	.	.	0.78	.		0.29	.	.	0.73
flexible	.		0.54	.	.	0.65	.	0.36	.	.	.	0.76	.	0.11		.	.	0.90
imaginative	.		0.41	.	.	0.75	.	.	.	.	.	0.86	.	.	0.02	.	.	0.98
mild	.	.	.	0.51	.	0.68	.	.	.	0.58	.	0.62	.	.	.	0.04	.	0.96
tenderhearted	.	.	.	0.59	.	0.60	.	.	.	0.57	.	0.60	.	.	.	0.11	.	0.90
altruistic	.	.	.	0.70	.	0.48	.	.	.	0.68	.	0.48	.	.	.	0.32	.	0.70
cooperative	.	.	.	0.68	.	0.50	.	.	.	0.64	.	0.50	.	.	.	0.41	.	0.59
sympathetic	.	.	.	0.79	.	0.35	.	.	.	0.74	.	0.37	.	.	.	0.61	.	0.32
deliberate	.	.	.	.	0.61	0.59	.	.	.	.	0.59	0.60	.	.	.	.	0.26	0.75
reliable	.	.	.	.	0.60	0.53	.	0.34	.	.	0.56	0.50	.	0.13	.	.	0.21	0.72
diligent	.	.	.	.	0.77	0.38	.	.	.	.	0.78	0.35	.	.	.	.	0.46	0.50
systematic	.	.	.	.	0.64	0.55	.	.	.	.	0.66	0.52	.	.	.	.	0.39	0.60
methodical	.	.	.	.	0.77	0.35	.	.	.	.	0.73	0.39	.	.	.	.	0.56	0.37
			Φ						Φ						Φ			
Factor 1	1	-0.29	-0.43	0.15	0.24		1.00	.	.	.	.		1.00	-0.33	-0.40	0.23	0.39	
Factor 2	-0.29	1	0.41	0.25	0.13		.	1.00	.	.	.		-0.33	1.00	0.29	0.26	0.08	
Factor 3	-0.43	0.41	1	0.1	-0.18		.	.	1.00	.	.		-0.40	0.29	1.00	-0.02	-0.28	
Factor 4	0.15	0.25	0.1	1	0.36		.	.	.	1.00	.		0.23	0.26	-0.02	1.00	0.30	
Factor 5	0.24	0.13	-0.18	0.36	1		.	.	.	.	1.00		0.39	0.08	-0.28	0.30	1.00	

$^*\rho = 0.41$; $\gamma = 3.75$

and four variables do not load any factor, which implies that their variations are not explained by any factor. In the FANC solution, one variable does not load any factor and two variables load the factors, which are not predicted by the theory. Here, the possibility remains that another combination of ρ and γ values may lead to the predicted sparsest $\mathbf{\Lambda}$ in FANC. However, it is unknown how to choose the ρ and γ values that give such a solution: we must resort to a heuristic procedure to select those values. This is disadvantageous for FANC. Therefore, we can conclude that this real data example demonstrated the superiority of SSFA in finding the sparsest loadings that cluster variables.

The comparisons of nonzero loadings among (A), (B), and (C) show that the values in (A) and (B) are mutually similar and larger than the corresponding ones in (C), with some nonzero loadings in (C) close to zero. It suggests that the use of penalty functions might shrink non-zero values in FANC. However, this problem is out of the scope and remains for the studies of penalized sparse FA.

5.2. Intelligence Test Data

The second example is Holzinger and Swineford's (1939) students × items matrix of the test scores which resulted in the intelligence test with 24 items administered for the students. Their raw scores are available at the website for Izenman (2008, p. 587). This data set has often been used for illustrating FA, as found in Harman (1976) and Izenman (2008), for example. Following the authors, we set m = 4 to carry out SSFA for the 24 × 24 inter-variable

Table 7. SSFA solution for intelligence test data

Variable	Λ				Ψ^2
Visual Perception	0.64	.	.	.	0.55
Cubes	0.38	.	.	.	0.83
Paper Form Board	0.38	.	.	.	0.83
Flags	0.54	.	.	.	0.67
General Information	.	0.83	.	.	0.30
Paragraph Comprehension	.	0.82	.	.	0.32
Sentence Completion	.	0.86	.	.	0.23
Word Classification	.	0.75	.	.	0.42
Word Meaning	.	0.85	.	.	0.26
Addition	.	.	0.57	.	0.61
Code	.	.	0.72	.	0.45
Counting Dots	.	.	0.63	.	0.58
Straight-Curved Capitals	.	.	0.69	.	0.50
Word Recognition	0.40	.	.	.	0.79
Number Recognition	0.36	.	.	.	0.81
Figure Recognition	0.60	.	.	.	0.62
Object-Number	0.38	.	.	.	0.78
Number-Figure	0.43	.	.	.	0.78
Figure-Word	.	.	.	0.47	0.75
Deduction	.	.	.	0.59	0.62
Numerical Puzzles	.	.	.	0.65	0.55
Problem Reasoning	.	.	.	0.67	0.53
Series Completion	0.73	.	.		0.42
Arithmetic Problems	.	.	.	0.66	0.53
	Φ				
Factor 1	1.00	0.53	0.59	0.84	
Factor 2	0.53	1.00	0.39	0.69	
Factor 3	0.59	0.39	1.00	0.60	
Factor 4	0.84	0.69	0.60	1.00	

correlation matrix obtained form the scores.

In intelligence tests, the items (variables) need to be clustered so that those in a cluster can be associated with a specific intellectual ability. Such needs are satisfied by the resulting SSFA solution in Table 7 as follows: a number of the ten items that form a cluster with loading the first factor (the first column of $\mathbf{\Lambda}$) involve visual and recognition tasks. It implies that those items can be regarded as measuring the ability for the visual and recognition task performances. The cluster of items loading the second factor is found related to the verbal ability. The item cluster loading the third factor involves the simple tasks and how fast they are completed. The items for which problem solving with thinking is needed form the fourth cluster. The nonzero loadings for the second factor are found to be larger than the others, which implies that the verbal ability can be measured better by the corresponding test items as compared with the other abilities.

5.3. Geochemical Data

The third example is Sampson's (1968) 11 × 11 inter-variable correlation matrix, which was obtained from the measurements on 11 geochemical variables for 122 brine samples. For that matrix, Reyment and Jöreskog (1996) have performed the standard orthogonal FA with m = 4. We also set m = 4 to carry out SSFA.

The resulting SSFA solution is presented in Table 8, together with the original one obtained by Reyment and Jöreskog (1996). Considering their loadings of large magnitudes,

Table 8. SSFA and the standard orthogonal FA solutions for geochemical data, where bold font is used for the latter loadings whose absolute values exceed 0.5.

Variable	SSFA					Reyment & Jöreskog (1996)				
	Λ				Ψ^2	Λ				Ψ^2
calcium	0.99	.	.	.	0.01	**0.81**	0.10	0.36	0.45	0.00
magnesium	0.79	.	.	.	0.37	**0.81**	0.17	0.10	0.19	0.28
sodium	.	0.29	.	.	0.91	0.04	0.26	0.05	0.12	0.91
bicarbonate	-0.18	.	.	.	0.96	-0.10	-0.09	-0.32	-0.03	0.88
sulfate	.	.	0.79	.	0.37	-0.01	0.41	-0.06	0.01	0.83
chloride	.	.	.	0.98	0.03	0.48	0.34	0.03	**0.78**	0.04
salts	.	0.89	.	.	0.21	0.38	**0.77**	-0.05	0.27	0.19
gravity	.	.	.	0.88	0.22	0.42	0.30	0.02	**0.72**	0.21
temperature	.	.	-0.20	.	0.95	-0.06	-0.05	0.24	-0.01	0.94
resistance	.	-0.79		.	0.33	-0.17	**-0.77**	**-0.57**	-0.24	0.00
pH	.	.		-0.34	0.87	-0.43	-0.04	0.16	-0.14	0.77
	Φ					Φ				
Factor 1	1.00	0.61	0.03	0.81		1.00	.	.	.	
Factor 2	0.61	1.00	0.44	0.72		.	1.00	.	.	
Factor 3	0.03	0.44	1.00	0.22		.	.	1.00	.	
Factor 4	0.81	0.72	0.22	1.00		.	.	.	1.00	

we see that the corresponding loadings in SSFA's first, second, and fourth factors also show non-zero values, except the third factor. For this third factor Reyment and Jöreskog write that it is not easy to interpret. However, in the SSFA solution, this is straightforward: the richness in sulfate and low temperature load the third factor to form a cluster, where the adjective "low" is attached to "temperature", since its loading is negative. The other factors can be interpreted in a similar way: the lack in bicarbonate and the richness in calcium and magnesium are classified into the same cluster; the resistance property and the richness in sodium and salts forms a cluster, and the remaining one consists of pH, chloride, and gravity. It can also be found that the variables loading any particular factor are either high or low in magnitude. For example, calcium and magnesium load the first factor heavily, while bicarbonate does not.

6. Concluding Remarks

We proposed the sparsest factor analysis (SSFA) for finding the optimal sparsest loading matrix, which has a single nonzero entry in every row and zeros elsewhere. For obtaining the SSFA solution, we presented an alternating least squares algorithm in which the matrix of common factor scores is re-parameterized using QR decomposition. In the simulation study, SSFA was shown to recover the true sparsest loadings very well.

In the existing sparse FA procedures, the number of nonzero loadings can take any value. In SSFA their number is restricted to *p*, the number of the input variables. This may also be a weakness of SSFA. However, sparsest loadings are usually required, when variables are anticipated clustered, as illustrated by the real data examples in Section 5. There, it was demonstrated that SSFA is genuinely useful for clustering variables.

In this paper, model selection problems were out of the scope. The problems include assessing whether the sparsest constraint on loadings is appropriate for the particular data and the selected number of factors. It remains for future studies to consider indices useful for such model selection.

References

Adachi, K. (2012). Some contributions to data-fitting factor analysis with empirical comparisons to covariance-fitting factor analysis. *J. Jpn. Soc. Comp. Statist.*, **25**, 25-38.

Adachi, K. (2014). A matrix-intensive approach to factor analysis. *Jpn. J. Statist.*, **44**, 363-382 (in Japanese).

Adachi, K., & Trendafilov, N. T. (2014). Sparse orthogonal factor analysis. In M. Carpita, E. Brentari, & E. M. Qannari, (Eds.), *Advances in latent variables: Studies in theoretical and applied statistics*, pp. 227 - 239, Heidelberg, Springer.

Aggarwal, C. C. (2015). *Data mining: The textbook*. New York: Springer.

Costa, P. T., & McCrae, R. R. (1992). NEO PI-R Professional Manual: Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI). Odessa, FL: Psychological Assessment Resources.

de Leeuw, J. (2004). Least squares optimal scaling of partially observed linear systems. In K. van Montfort, J. Oud and A. Satorra (Eds.), *Recent developments of structural equation models: Theory and applications*. Pp. 121-134. Dordrecht: Kluwer Academic Publishers.

Eldén, L. (2007). *Matrix methods in data mining and pattern recognition*. Philadelphia: Siam.

Everitt, B. S. (1993). *Cluster analysis* (3rd edition). London: Edward Arnold.

Gan, G., Ma, C., and Wu, J. (2007). *Data clustering: Theory, algorithms, and Applications*. Philadelphia, PA, Society of Industrial and Applied Mathematics (SIAM).

Goldberg, L. R. (1992). The development of markers for the Big-five factor structure. *Psychological Assessment, 4*, 26-42.

Harman, H.H. (1976). *Modern factor analysis* (Third Edition). Chicago: The University of Chicago Press.

Hirose, K., & Yamamoto, M. (2014a). Estimation of an oblique structure via penalized likelihood factor analysis. *Comput. Statist. Data Anal.*, **79**, 120-132.

Hirose, K., & Yamamoto, M. (2014b). Sparse estimation via nonconcave penalized likelihood in factor analysis model. *Statist. Comput.*, http://dx.doi.org/10.1007/ s11222-014-9475-z (to appear).

Holzinger, K. J., & Swineford, F. (1939). *A study in factor analysis: The stability of a bi-factor solution*. University of Chicago: Supplementary Educational Monographs, No. 48.

Izenman, A. J (2008). *Modern multivariate statistical techniques*: *Regression, classification, and manifold learning*. New York, NY. Springer.

Jolliffe, I. T., Trendafilov, N. T., & Uddin, M. (2003). A modified principal component technique based on the LASSO. *J. Comput. Graph. Statist.*, **12**, 531-547.

Knowles, D. & Ghahramani, Z. (2011). Nonparametric Bayesian sparse factor models with applications to gene expression modeling. *Annals of Applied Statistics,* **5**, 1534-1552.

Mazumder, R., Friedman, J., & Hastie, T. (2011). SparseNet: Coordinate descent with nonconvex penalties. *J. Amer. Statis. Assoc., 106*, 1125-1138.

Mulaik, S. A. (2010). *Foundations of factor analysis*. *Second Edition*. Boca Raton: CRC Press.

Rattray, M., Stegle, O., Sharp, K., & Winn, J. (2009). Inference algorithms and learning theory for Bayesian sparse factor analysis. *Journal of Physics: Conference Series*, **197**, 1-10, doi:10.1088/1742-6596/197/1/012002.

Reyment, R., & Jöreskog, K. G., (1996). *Applied factor analysis in the natural sciences*. Cambridge, Cambridge University Press.

Sampson, R. J. (1968). *R*-mode factor analysis program in FORTRAN II for the IBM 1620 computer. *Kansas geological survey computer contributions, 20*.

Seber, G. A. F., (2008). *A matrix handbook for statisticians*. Hoboken, NJ: Wiley.

Sočan, G. (2003). *The incremental value of minimum rank factor analysis*. PhD Thesis, University of Groningen: Groningen.

Spearman, C. (1904). ‘General intelligence’ objectively determined and measured. *American Journal of Psychology*, **15**, 201-293.

Stegeman, A. (2016). A new method for simultaneous estimation of the factor model parameters, factor scores, and unique parts. *Comput. Statist. Data Anal.*, **99**, 189-203.

ten Berge, J. M. F. (1983). A generalization of Kristof’s theorem on the trace of certain matrix products. *Psychometrika* **48**, 519-523.

ten Berge, J. M. F. (1993). *Least squares optimization in multivariate analysis*. Leiden, The Netherlands: DSWO Press.

Trendafilov, N. T. (2014). From simple structure to sparse components: a review. *Comput. Statist.*, **29**, 431-454.

Trendafilov, N. T. & Unkel, S. (2011). Exploratory factor analysis of data matrices with more variables than observations. *Journal of Computational and Graphical Statistics*, **20**, 874-891.

Trendafilov, N. T. Unkel, S., & Krzanowski, W. (2011). Exploratory factor and principal analyses: some new aspects. *Statistics and Computing*, **23**, 209-220.

Unkel, S. and Trendafilov, N. T. (2010). Simultaneous parameter estimation in exploratory factor analysis: An expository review. *International Statistical Review* **78**, 363-382.

Vichi, M. and Saporta, G. (2009). Clustering and disjoint principal component analysis. *Comput. Statist. Data Anal.*, **53**, 3194-3208.

Zaki, M. J., & Meira, W. (2014) *Data Mining and Analysis: Fundamental Concepts and Algorithms*, Cambridge University Press, Cambridge, UK.

Zou, D. M., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *J. Comput. Graph. Statist.*, **15**, 265-286.