Sparse Orthogonal Factor Analysis

Kohei Adachi and Nickolay T. Trendafilov

Abstract

We propose a sparse orthogonal factor analysis (SOFA) procedure in which the optimal loadings and unique variances are estimated subject to additional constraint which directly requires some factor loadings to be exact zeros. More precisely, the constraint specifies the required number of zero factor loadings without any restriction on their locations. Such loadings are called sparse which gives the name of the method. SOFA solutions are obtained by minimizing a FA loss function under the sparseness constraint making use of an alternate least squares algorithm. We further present a sparseness selection procedure in which SOFA is performed repeatedly by setting the sparseness at each of a set of feasible integers. Then, the SOFA solution with the optimal sparseness can be chosen using an index for model selection. This procedure allows us to find the *optimal* orthogonal confirmatory factor analysis model among all possible models. SOFA and the sparseness selection procedure are assessed by simulation and illustrated with well known data sets.

Keywords

Confirmatory factor analysis • Factor analysis • Optimal sparseness selection • Sparse loadings • Sparse principal component analysis

K. Adachi (✉)
Graduate School of Human Sciences, Osaka University, 1-2 Yamadaoka, Suita, Osaka 565-0871, Japan
e-mail: adachi@hus.osaka-u.ac.jp

N.T. Trendafilov
Department of Mathematics and Statistics, The Open University, Buckinghamshire, UK

M. Carpita et al. (eds.), *Advances in Latent Variables*, Studies in Theoretical and Applied Statistics, DOI 10.1007/10104_2014_2, Published online: 28 October 2014

1 Introduction

Factor analysis (FA) is a model that aims to explain the interrelationships among observed variables by a small number of latent variables called common factors. The relationships of the factors to observed variables are described by a factor loading matrix. FA is classified as exploratory (EFA) or confirmatory (CFA). In EFA, the loading matrix is unconstrained and has rotational freedom which is exploited to rotate the matrix so that some of its elements approximate zero. In CFA, some loadings are constrained to be zero and the loading matrix has no rotational freedom [9].

One refers to a loading matrix with a number of exactly zero elements as being *sparse*, which is an indispensable property for loadings to be interpretable. In EFA, a loading matrix is rotated toward a sparse matrix, but the literal sparseness is not attained, since rotated loadings cannot exactly be equal to zero. Thus, the user must decide which of them can be viewed as approximately zeros. On the other hand, some loadings are fixed exactly to zero in CFA. However, the problem in CFA is that the number of zero loadings and their locations must be chosen by users. That is, the users' subjective decision is needed in both EFA and CFA.

In order to overcome these difficulties, we propose a new FA procedure, which is neither EFA nor CFA. The optimal orthogonal factor solution is estimated such that it has a sparse loading matrix with a suitable number of zero elements. Note that, their locations are also estimated in an optimal way. The procedure to be proposed consists of the following two stages:

(a) The optimal solution is obtained for a specified number of zero loadings.
(b) The optimal number of zero loadings is selected among possible numbers.

Stages (a) and (b) would be described in Sects. 2–4, respectively.

In the area of principal component analysis (PCA), many procedures, called sparse PCA, have been proposed in the last decade (e.g. [8, 13, 16]). As in our FA procedure, they obtain sparse loadings. However, besides the difference between PCA and FA, our approach does not rely on penalty functions, which is the standard way to induce sparseness in the existing sparse PCA.

2 Sparse Factor Problem

The main goal of FA is to estimate the p-variables $\times m$-factors matrix $\mathbf{\Lambda}$ containing loadings and the $p \times p$ diagonal matrix $\mathbf{\Psi}^2$ including unique variances from the n-observation $\times p$-variables ($n > p$) column-centered data matrix $\mathbf{X}$. For this goal, FA can be formulated by a number of different objective functions, among which we choose the least squares function

$$f = ||\mathbf{X} - \mathbf{F}\mathbf{\Lambda}' - \mathbf{U}\mathbf{\Psi}||^2 \tag{1}$$

recently utilized in several works [1,4,14,15]. Here, $\mathbf{F}$ is the $n \times m$ matrix containing common factor scores and $\mathbf{U}$ is the $n \times p$ matrix of unique factor scores. The factor score matrices are constrained to satisfy

$$n^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_m, n^{-1}\mathbf{U}'\mathbf{U} = \mathbf{I}_p, \text{ and } \mathbf{F}'\mathbf{U} = {}_m\mathbf{O}_p \tag{2}$$

with $\mathbf{I}_m$ the $m \times m$ identity matrix and ${}_m\mathbf{O}_p$ the $m \times p$ matrix of zeros.

We propose to minimize (1) over $\mathbf{F}$, $\mathbf{U}$, $\mathbf{\Lambda}$, and Ψ subject to (2) and

$$SP(\mathbf{\Lambda}) = q, \tag{3}$$

where $SP(\mathbf{\Lambda})$ expresses the sparseness of $\mathbf{\Lambda}$, i.e., the number of its elements being zero, and q is a specified integer.

The reason for our choosing loss function (1) is that we can define

$$\mathbf{A} = n^{-1}\mathbf{X}'\mathbf{F} \tag{4}$$

to decompose (1) as

$$f = ||\mathbf{X} - \mathbf{F}\mathbf{A}' - \mathbf{U}\mathbf{\Psi} - (\mathbf{F}\mathbf{\Lambda}' - \mathbf{F}\mathbf{A}')||^2 = ||\mathbf{X} - \mathbf{F}\mathbf{A}' - \mathbf{U}\Psi||^2 + n||\mathbf{\Lambda} - \mathbf{A}||^2. \tag{1'}$$

This equality is derived from the fact that $(\mathbf{X} - \mathbf{F}\mathbf{A}' - \mathbf{U}\Psi)'(\mathbf{F}\mathbf{\Lambda}' - \mathbf{F}\mathbf{A}') = n\mathbf{A}\mathbf{\Lambda}' - n\mathbf{A}\mathbf{A}' - n\mathbf{A}\mathbf{\Lambda}' + n\mathbf{A}\mathbf{A}' = {}_p\mathbf{O}_p$ is given using (2) and (4). In (1′) only a simple function $||\mathbf{\Lambda} - \mathbf{A}||^2$ is relevant to $\mathbf{\Lambda}$ and thus can be easily minimized over $\mathbf{\Lambda}$ subject to (3) as seen in the next section. It is difficult for other objective functions of FA to be rewritten into simple forms as (1′). For example, the likelihood function for FA includes the determinant of a function of $\mathbf{\Lambda}$ which is difficult to handle.

3 Algorithm

For minimizing (1) subject to (2) and (3), we consider alternately iterating the update of each parameter matrix.

First, let us consider updating $\mathbf{\Lambda}$ so that (1) or (1′) is minimized subject to (3) while $\mathbf{F}$, $\mathbf{U}$, and Ψ are kept fixed. This amounts to minimizing $g(\mathbf{\Lambda}) = ||\mathbf{\Lambda} - \mathbf{A}||^2$ under (3), since of (1′). Using $\mathbf{\Lambda} = (\lambda_{ij})$ and $\mathbf{A} = (a_{ij})$, we can rewrite $g(\mathbf{\Lambda})$ as

$$g(\mathbf{\Lambda}) = \sum_{(i,j)\in\mathbf{N}} a_{ij}^2 + \sum_{(i,j)\in\mathbf{N}^\perp} (\lambda_{ij} - a_{ij})^2 \geq \sum_{(i,j)\in\mathbf{N}} a_{ij}^2, \tag{5}$$

where $\mathbf{N}$ denotes the set of the q pairs of (i, j) for the loadings λ_{ij} to be zero and $\mathbf{N}^\perp$ is the complement to $\mathbf{N}$. The inequality in (5) shows that $g(\mathbf{\Lambda})$ attains its lower limit $\Sigma_{(i,j)\in N} a_{ij}^2$ when the loading λ_{ij} with $(i, j) \in \mathbf{N}^\perp$ is set equal to a_{ij}. Further,

the limit $\Sigma_{(i,j)\in N} a_{ij}^2$ is minimal when $\mathbf{N}$ contains the (i, j) for the q smallest a_{ij}^2 among all squared elements of $\mathbf{A}$. The optimal $\mathbf{\Lambda} = (\lambda_{ij})$ is thus given by

$$\lambda_{ij} = \begin{cases} 0 \ \text{ iff} \, a_{ij}^2 \leq a_{[q]}^2 \\ a_{ij} \ \ \text{otherwise} \end{cases} \tag{6}$$

with $a_{[q]}^2$ the q-th smallest value among all a_{ij}^2.

Next, let us consider the update of the diagonal matrix Ψ. Substituting (2) in (1) simplifies the objective function to

$$f = n\text{tr}\mathbf{S} + n\text{tr}\mathbf{\Lambda}\mathbf{\Lambda}' + n\text{tr}\mathbf{\Psi}^2 - 2n\text{tr}\mathbf{X}'\mathbf{F}\mathbf{\Lambda}' - 2\text{tr}\mathbf{X}'\mathbf{U}\mathbf{\Psi} \tag{1''}$$

with $\mathbf{S} = n^{-1/2}\mathbf{X}'\mathbf{X}$ the sample covariance matrix. Since (1″) can further be rewritten as $||n^{1/2}\mathbf{\Psi} - n^{-1/2}\text{diag}(\mathbf{X}'\mathbf{U})||^2 + c$ with c a constant irrelevant to $\mathbf{\Psi}$, the minimizer is found to be given by

$$\mathbf{\Psi} = \text{diag}(n^{-1}\mathbf{X}'\mathbf{U}), \tag{7}$$

when $\mathbf{F}, \mathbf{U}$, and $\mathbf{\Lambda}$ are considered fixed.

Finally, let us consider minimizing (1) over $n \times (m + p)$ block matrix $[\mathbf{F}, \mathbf{U}]$ subject to (2) with $\mathbf{\Psi}$ and $\mathbf{\Lambda}$ kept fixed. We note that (1″) is rewritten as $f = c^* - 2n\text{tr}\mathbf{B}'\mathbf{X}'[\mathbf{F}, \mathbf{U}]$ with $\mathbf{B} = [\mathbf{\Lambda}, \mathbf{U}]$ an $p \times (m + p)$ matrix and c^* a constant irrelevant to $[\mathbf{F}, \mathbf{U}]$. As proved in Appendix 1, f is minimized for

$$n^{-1}\mathbf{X}'[\mathbf{F}, \mathbf{U}] = \mathbf{B}'^{+}\mathbf{Q}\mathbf{\Delta}\mathbf{Q}', \tag{8}$$

where $\mathbf{B}'^{+}$ is the Moore-Penrose inverse of $\mathbf{B}'$ and $\mathbf{Q}\mathbf{\Delta}\mathbf{Q}'$ is obtained through the eigenvalue decomposition (EVD) of $\mathbf{B}'\mathbf{S}\mathbf{B}$:

$$\mathbf{B}'\mathbf{S}\mathbf{B} = \mathbf{Q}\mathbf{\Delta}^2\mathbf{Q}', \tag{9}$$

with $\mathbf{Q}'\mathbf{Q} = \mathbf{I}_p$ and Δ^2 the positive definite diagonal matrix. Rewriting (8) as $[n^{-1}\mathbf{X}'\mathbf{F},\ n^{-1}\mathbf{X}'\mathbf{U}] = \mathbf{B}'^{+}\mathbf{Q}\mathbf{\Delta}\,Q'$ and comparing it with (4) and (7), one finds:

$$\mathbf{A} = \mathbf{B}'^{+}\mathbf{Q}\mathbf{\Delta}\mathbf{Q}'\mathbf{H}_m \tag{4'}$$

$$\mathbf{\Psi} = \text{diag}(\mathbf{B}'^{+}\mathbf{Q}\mathbf{\Delta}\mathbf{Q}'\mathbf{H}^p) \tag{7'}$$

where $\mathbf{H}_m = [\mathbf{I}_m,\ {}_m\mathbf{O}_p]'$ and $\mathbf{H}^p = [{}_p\mathbf{O}_m,\ \mathbf{I}_p]'$.

The above equations show that $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ can be updated if only the sample covariance matrix $\mathbf{S}(= n^{-1}\mathbf{X}'\mathbf{X})$ is available. In other words, the updating of $[\mathbf{F}, \mathbf{U}]$ can be avoided when the original data matrix $\mathbf{X}$ is not given, That is, the decomposition (9) gives the matrices $\mathbf{Q}$ and $\mathbf{\Delta}$ needed in (4′) and (7′), with (4′) being used for (6). Further, the resulting loss function value can be computed without the

use of $\mathbf{X}$: (6) implies $\mathbf{\Lambda}'\mathbf{A} = \mathbf{\Lambda}'\mathbf{\Lambda}$, and substituting this, (4), and $\mathbf{B} = [\mathbf{\Lambda}, \mathbf{U}]$ into (1″) leads to $f = n\mathrm{tr}\mathbf{S} + n\mathrm{tr}\mathbf{\Lambda}\mathbf{\Lambda}' - 2n\mathrm{tr}\mathbf{\Lambda}'\mathbf{A} - n\mathrm{tr}\mathbf{\Psi}^2 = n\{\mathrm{tr}\mathbf{S} - \mathrm{tr}(\mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}^2)\} = n(\mathrm{tr}\mathbf{S} - \mathrm{tr}\mathbf{B}\mathbf{B}')$. Then, the standardized loss function value

$$f_S(\mathbf{B}) = 1 - \mathrm{tr}\mathbf{B}\mathbf{B}'/\mathrm{tr}\mathbf{S}, \tag{10}$$

which takes a value within [0,1], can be used for convenience instead of f.

The optimal $\mathbf{B} = [\mathbf{\Lambda}, \mathbf{\Psi}]$ is thus given by the following algorithm:

Step 1. Initialize $\mathbf{\Lambda}$ and $\mathbf{\Psi}$.
Step 2. Set $\mathbf{B} = [\mathbf{\Lambda}, \mathbf{\Psi}]$ to perform EVD (9).
Step 3. Obtain $\mathbf{A}$ by (4′) to update $\mathbf{\Lambda}$ with (6).
Step 4. Update $\mathbf{\Psi}$ with (7′).
Step 5. Finish if convergence is reached; otherwise, go back to Step 2.

The convergence of the updated parameters in Step 5 is defined as the decrease of (10) being less than 0.1^7. To avoid missing the global minimum, we run the algorithm multiple times with random start in Step 1. The procedure for selection of the optimal solution is described in Appendix 2. We denote the resulting solution of $\mathbf{B}$ as $\hat{\mathbf{B}}_q = [\hat{\mathbf{\Lambda}}_q, \hat{\mathbf{\Psi}}_q]$, where the subscript q indicates the particular number of zeros used in (3).

4 Sparseness Selection

Sparseness can be restated as parsimony: the greater $SP(\mathbf{\Lambda})$ implies that fewer parameters are to be estimated and the resulting loss function value is greater. Thus, the sparseness selection means to choose a FA model with the optimal combination of the attained loss function value and parsimony. For such model selection, we can use the information criteria [10] which are defined using maximum likelihood (ML) estimates. Although ML method is not used in our algorithm, we assume that $\hat{\mathbf{B}}_q = [\hat{\mathbf{\Lambda}}_q, \hat{\mathbf{\Psi}}_q]$ is equivalent to the ML-CFA solution which maximizes log likelihood $L(\mathbf{\Lambda}, \mathbf{\Psi}) = -0.5n\{\log|\mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}^2| + \mathrm{tr}\mathbf{S}(\mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}^2)^{-1}\}$ with the locations of the zero loadings constrained to be those of $\hat{\mathbf{\Lambda}}_q$. This assumption would be validated empirically in the next section. Under this assumption, we propose to use an information criterion BIC [10] for choosing the optimal q. BIC can be expressed as

$$BIC(q) = -2L(\hat{\mathbf{\Lambda}}_q, \hat{\mathbf{\Psi}}_q) - q\log n + c^{\#} \tag{11}$$

for $\hat{\mathbf{B}}_q$ with $c^{\#}$ a constant irrelevant to q. The optimal sparseness is thus defined as

$$\hat{q} = \underset{q_{\min} \le q \le q_{\max}}{\arg\min} \; BIC(q) \tag{12}$$

and $\hat{\mathbf{B}}_{\hat{q}}$ is chosen as the final solution $\hat{\mathbf{B}}$. Here, we set $q_{\min} = m(m-1)/2$, since it prevents $\mathbf{\Lambda}$ from rotational indeterminacy if q goes below it. On the other hand, we set $q_{\max} = p(m-1)$, since it prevents $\mathbf{\Lambda}$ from having an empty column if q were greater than the limit.

5 Simulation Study

We performed a simulation study to assess the proposed procedure with respect to (1) identifying the true sparseness and the locations of zero loadings; (2) goodness of the recovery of parameter values; (3) sensitivity to local minima; (4) whether SOFA solutions are equivalent to the solutions of the ML-CFA procedure with the locations of the zero elements in $\mathbf{\Lambda}$ set to those obtained by SOFA.

We used the five types of the true $\mathbf{\Lambda}$ shown in Table 1, which are desired to be possessed by FA solutions. The first three types have simple structure, while the remaining two have bi-factor simple structure as defined by [6]. For each type, we generated 40 sets of $\{\mathbf{\Lambda}, \mathbf{\Psi}, \mathbf{S}\}$ by the following steps: (1) each diagonal element of $\mathbf{\Psi}$ was set to $u(0.1^{1/2}, 0.7^{1/2})$, where $u(\alpha, \beta)$ denotes a value drawn from the uniform distribution of the range $[\alpha, \beta]$. (2) A nonzero value in $\mathbf{\Lambda}$ was set to $u(0.4, 1)$, while an element denoted by "r" in Table 1 was randomly set to zero or $u(0.4, 1)$. (3) $\mathbf{\Lambda}$ was normalized so as to satisfy $\mathrm{diag}(\mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}^2) = \mathbf{I}_p$. (4) Setting $n = 200p$, we sampled each row of $\mathbf{X}$ from the centered p-variate normal distribution with its covariance matrix $\mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}^2$. (5) Inter-variable correlation matrix $\mathbf{S}$ was obtained from $\mathbf{X}$. For the resulting data sets, we carried out SOFA: its algorithm was run multiple times for each of $q = q_{\min}, \ldots, q_{\max}$ until the two equivalently optimal

Table 1 Three loading matrices of simple structure (left) and two ones of bi-factor structure (right), with nonzero and zero elements denoted by # and blank cells, respectively

#	r		#	r			#		r	r		#	#			#	#		
#	r		#				#		r	r		#	#			#	#		
#			#				#		r	r		#	#			#	#		
#			#				#		r	r		#	r			#	r		
r	#		r	#			r	#		r		#	r			#	r	r	
r	#		r	#			r	#		r		#	r			#		#	
	#		r	#			r	#		r		#		#		#		#	
	#		r	#			r	#		r		#		#		#		#	
			r	r	#		r	r	#			#		#		#		r	
				r	#		r	r	#			#		r		#		r	r
				r	#		r	r	#			#		r		#			#
				r	#		r	r	#			#		r		#			#
								r	r	#						#			#
								r	r	#						#			r
								r	r	#						#	r		r
								r	r	#									

Table 2 Percentiles of index values for assessing the SOFA solutions

		(B) Rate		(C) Difference to the true value		(D) Difference to ML-CFA	
Percentile	(A) BES	R_{00}	$R_{\#\#}$	$\mathbf{\Lambda}$	$\mathbf{\Psi}^2$	$\mathbf{\Lambda}$	$\mathbf{\Psi}^2$
5	−0.133	0.843	0.972	0.013	0.023	0.002	0.004
25	−0.031	0.968	1.000	0.017	0.032	0.003	0.005
50	0.000	1.000	1.000	0.021	0.038	0.004	0.006
75	0.000	1.000	1.000	0.026	0.046	0.006	0.008
95	0.000	1.000	1.000	0.040	0.056	0.009	0.011

solutions are found by the procedure in Appendix 2. As done there, we use L_q for the number of runs necessitated.

To asses the sensitivity of SOFA to local minima, we counted L_q and averaged it over q for each data set. The sensitivity is indicated by L_q as described in Appendix 2. The quartiles of the averaged L_q values over 200 data sets were 89, 120, and 155: the second quartile 120 implies that the $120 - 2 = 118$ solutions (except two equivalently optimal solutions) are local minimizers among 120 solutions for a half of data sets. This demonstrates high sensitivity to local minima. Nevertheless, good performances of the proposed procedure are shown next.

For each of 200 data sets, we obtained some index values to assess the correctness of the $\hat{q}$ selected by BIC and the corresponding optimal solution $\hat{\mathbf{B}}_{\hat{q}} = [\hat{\mathbf{\Lambda}}_{\hat{q}}, \hat{\mathbf{\Psi}}_{\hat{q}}]$. The percentiles of the index values over the 200 cases are shown in Panels (A), (B), and (C) of Table 2. The first index is BES $= (\hat{q} - q)/q$ which assesses the relative bias of the estimated sparseness from the true q. The percentiles in Panel (A) show that sparseness was satisfactorily estimated, though it tended to be underestimated. The indices R_{00} and $R_{\#\#}$in Panel (B) are the rates of the zero and non-zero elements in the true $\mathbf{\Lambda}$ correctly identified by $\hat{\mathbf{\Lambda}}$. Panel (B) shows that non-zero elements have been exactly identified. The indices in Panel (C) are mean absolute differences $||\hat{\mathbf{\Lambda}}_{\hat{q}} - \mathbf{\Lambda}||_1/(pm)$ and $||\hat{\mathbf{\Psi}}^2_{\hat{q}} - \mathbf{\Psi}^2||_1/p$, where $||\cdot||_1$ denotes the sum of the absolute values of the elements of the argument. The percentiles of the differences show that the parameter values were recovered very well.

For each data set, ML-CFA was also performed with the locations of the zero loadings fixed at those in $\hat{\Lambda}_{\hat{q}}$. For ML-CFA, we used the EM algorithm with [2] formulas. Let $\mathbf{\Lambda}_{\mathrm{ML}}$ and $\mathbf{\Psi}_{\mathrm{ML}}$ denote the resulting $\mathbf{\Lambda}$ and $\mathbf{\Psi}$. Panel (D) in Table 2 shows the percentiles of $||\hat{\mathbf{\Lambda}}_{\hat{q}} - \mathbf{\Lambda}_{\mathrm{ML}}||_1/(pm)$ and $||\hat{\mathbf{\Psi}}^2_{\hat{q}} - \mathbf{\Psi}^2_{\mathrm{ML}}||_1/p$. There, we find that the differences were small enough to be ignored, which validate the use of ML-based BIC in SOFA.

6 Examples

We illustrate SOFA with two famous examples. The first one is a real data set known as [6] twenty four psychological test data, which contain the scores of $n = 145$ students for $p = 24$ problems. The correlation matrix is available in [5], p. 124.

Table 3 Solution for 24 psychological test data with empty cells standing for zero

Abilities	Variables (problems)	$\mathbf{\Lambda}$				ψ_i^2
		1	2	3	4	
Spatial perception	Visual perception	0.67				0.52
	Cubes	0.43				0.79
	Paper form board	0.52		−0.19		0.66
	Flags	0.54				0.68
Verbal processing	General information	0.56	0.59			0.31
	Paragraph comprehension	0.58	0.58			0.31
	Sentence completion	0.55	0.64			0.26
	Word classification	0.62	0.35			0.47
	Word meaning	0.59	0.60			0.26
Speed of performances	Addition	0.26	0.16	0.80		0.25
	Code	0.42		0.47	0.26	0.50
	Counting dots	0.37		0.62		0.45
	Straight-curved capitals	0.56		0.38		0.51
Memory	Word recognition	0.36			0.46	0.64
	Number recognition	0.34			0.45	0.67
	Figure recognition	0.54	−0.15		0.35	0.55
	Object-number	0.36		0.20	0.52	0.54
	Number-figure	0.45		0.27	0.33	0.59
	Figure-word	0.43			0.22	0.74
Mathematics	Deduction	0.66				0.54
	Numerical puzzles	0.58		0.30		0.55
	Problem reasoning	0.65				0.56
	Series completion	0.74				0.43
	Arithmetic problems	0.54	0.21	0.40		0.49

From the EFA solution for the matrix, [7] found bi-factor structure using their proposed bi-factor rotation with $m = 4$. We analyzed the correlation matrix by SOFA with the same number of factors. The optimal $SP(\mathbf{\Lambda}) = 48$ was found by BIC. The solution is shown in Table 3. Its first column shows the abilities made up by [5], p. 125, which are considered necessary for solving the corresponding groups of problems. This grouping can be used to give clear interpretation of $\hat{\mathbf{\Lambda}}$: the first, second, third, and fourth factors stand in turn for the general ability related to all problems, the skill of verbal processing, the speed of performances, and the accuracy of memory, respectively. It matches the bi-factor structure found by [7]. However, our result allows us to interpret the factors simply by observing the nonzero loadings, while [7] obtain reasonable interpretation only after considering the loadings greater than or equal to 0.3 in magnitude. This choice is subjective and likely to lead to suboptimal and misleading solutions.

Table 4 Solution for the box problem with empty cells standing for zero

	$\mathbf{\Lambda}$			
Variables	x	y	z	ψ_i^2
x^2	0.95			0.08
y^2		0.96		0.08
z^2			0.94	0.09
xy	0.67	0.61		0.17
xz	0.64		0.64	0.17
yz		0.66	0.63	0.15
$(x^2 + y^2)^{1/2}$	0.69	0.64		0.10
$(x^2 + z^2)^{1/2}$	0.68		0.64	0.12
$(y^2 + z^2)^{1/2}$		0.66	0.67	0.11
$2x + 2y$	0.68	0.67		0.08
$2x + 2z$	0.67		0.68	0.08
$2y + 2z$		0.66	0.68	0.09
$\log x$	0.89			0.19
$\log y$		0.87		0.23
$\log z$			0.88	0.21
xyz	0.47	0.49	0.54	0.22
$(x^2 + y^2 + z^2)^{1/2}$	0.57	0.52	0.54	0.10
e^x	0.71			0.48
e^y		0.68		0.52
e^z			0.71	0.49

The second example considers [12] box problem which gives simulated data traditionally used as a benchmark for testing FA procedures. As described in Appendix 3, we followed [12] to generate 20 variables using functions of 3×1 common factor vector $[x, y, z]'$, with the functions defined as in the first column of Table 4. Those procedures gave the correlation matrix (Table 5) to be analyzed. The ideal solution for this problem is the one such that variables load the factor(s) used for defining the variables: for example, the fourth variable should ideally load x and y. The SOFA solution with $SP(\mathbf{\Lambda}) = 27$ selected by BIC is shown in Table 4, where we find that the ideal loadings were obtained.

Table 5 Correlation coefficients multiplied by 100 for the box problem

Var.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	100																			
2	−5	100																		
3	−2	0	100																	
4	62	57	−2	100																
5	59	−3	60	42	100															
6	−4	64	59	40	41	100														
7	63	61	−2	82	42	40	100													
8	65	1	61	47	81	40	46	100												
9	−3	62	63	40	42	83	41	44	100											
10	62	63	2	87	44	45	90	48	45	100										
11	63	−2	63	44	86	42	45	89	45	46	100									
12	−4	62	63	41	40	86	40	44	90	45	45	100								
13	82	−6	1	57	60	−3	59	62	−2	59	64	−4	100							
14	−5	80	2	57	0	58	57	2	60	59	0	63	−5	100						
15	−2	1	79	1	58	60	−2	55	62	2	61	64	1	2	100					
16	41	47	50	66	72	73	61	64	65	65	68	67	43	44	50	100				
17	52	50	51	68	67	64	74	76	72	75	74	71	48	46	48	77	100			
18	74	1	−6	48	41	−5	53	47	−4	50	42	−1	52	3	−7	28	43	100		
19	−7	72	6	35	−2	51	43	0	50	41	−3	45	−7	49	5	34	37	−5	100	
20	1	4	76	2	41	48	1	48	50	4	46	50	−1	7	51	40	43	−3	8	100

7 Discussions

In this paper, we proposed a new FA procedure named SOFA (sparse orthogonal factor analysis), which is neither EFA nor CFA. In SOFA, FA loss function (1) is minimized over loadings and unique variances subject to the direct sparseness constraint on loadings. This minimization algorithm alternately estimates the locations of the zero loadings and the values of the nonzero ones. Further, the best sparseness is selected using BIC. The simulation study demonstrated that the true sparseness and parameter values are recovered well by SOFA, and the examples illustrated that SOFA produces reasonable sparse solutions.

As stated already, a weakness of the rotation methods in EFA is that the user must decide which rotated loadings can be viewed as potential zeros. Another weakness of the rotation methods is that they do not involve the original data, because the rotation criteria are functions of the loading matrix only [3]. Thus, the rotated loadings may possess structure which is not relevant to the true loadings of the underlying data. In contrast, SOFA minimizes (1) so that the FA model is optimally fitted to the data set under the sparseness constraint, and thus can find the loadings underlying the data set with a suitable sparseness.

Our proposed procedure of SOFA with the sparseness selection by BIC allows us to find an optimal orthogonal solution with the best sparseness. If one tries to find that optimal solution by CFA without any prior knowledge about the solution, CFA must be performed over all possible models, i.e., over all possible locations of q zero loadings with changing q from $q_{\min}$ to $q_{\max}$. That is, the number of the models to be tested is so enormous that it is unfeasible. An optimal model can, however, be found by our procedure. It is thus regarded as an automatic finder of an optimal orthogonal CFA model.

A drawback of SOFA is that its solutions are restricted to the orthogonal ones without inter-factor correlations. It thus remains for future studies to develop a sparse oblique FA procedure with the correlations included in parameters.

Acknowledgements The works were partially supported by Grant #4387 by The Great Britain Sasakawa Foundation.

Appendix 1: Update of $n^{-1}X'$ [F,U]

We prove that $c^* - \mathrm{tr}\mathbf{B}'\mathbf{X}'[\mathbf{F},\mathbf{U}]$ is minimized, or equivalently, $\mathrm{tr}\mathbf{B}'\mathbf{X}'[\mathbf{F},\mathbf{U}]$ is maximized, for (8) subject to (2), supposing that the rank of $\mathbf{XB}$ is p. First, let us consider maximizing $\mathrm{tr}\mathbf{B}'\mathbf{X}'[\mathbf{F},\mathbf{U}]$ under the constrains in (2) summarized in $n^{-1}[\mathbf{F},\mathbf{U}]'[\mathbf{F},\mathbf{U}] = \mathbf{I}_{m+p}$. The maximizer is given by

$$[\mathbf{F},\mathbf{U}] = n^{1/2}\mathbf{P}\mathbf{Q}' + n^{1/2}\mathbf{P}_\perp\mathbf{Q}'_\perp \tag{13}$$

through the singular value decomposition of $n \times (m + p)$ matrix $n^{-1/2}\mathbf{XB}$;

$$n^{-1/2}\mathbf{XB} = [\mathbf{P}, \mathbf{P}_\perp] \begin{bmatrix} \mathbf{\Delta} & \\ & {}_m\mathbf{O}_m \end{bmatrix} \begin{bmatrix} \mathbf{Q}' \\ \mathbf{Q}'_\perp \end{bmatrix} = \mathbf{P}\mathbf{\Delta}\mathbf{Q}'. \tag{14}$$

Here, $[\mathbf{P}, \mathbf{P}_\perp]$ and $[\mathbf{Q}, \mathbf{Q}_\perp]$ are $n \times (p + m)$ and $p \times (p + m)$ orthonormal matrices, respectively, whose blocks $\mathbf{P}$ and $\mathbf{Q}$ consist of p columns, and $\mathbf{\Delta}$ is a $p \times p$ diagonal matrix [11]. Next, let us note that the rank of $\mathbf{XB}$ being p implies $\mathbf{B}$ being of full-row rank, which leads to $\mathbf{B}\mathbf{B}^+ = \mathbf{I}_p$. Using this fact in (14) we have $n^{-1}\mathbf{X} = n^{-1/2}\mathbf{P}\mathbf{\Delta}\mathbf{Q}'\mathbf{B}^+$, which is transposed and post-multiplied by (13) to give (8), since of $\mathbf{P}'\mathbf{P}_\perp = {}_p\mathbf{O}_{p-m}$. Further, (8) is obtained with (9) followed from (14).

Appendix 2: Multiple Runs Procedure

The initial $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ in the SOFA algorithm (Sect. 3) are chosen randomly. Each diagonal element of $\mathbf{\Psi}$ is initialized at $u(0.1^{1/2}, 0.7^{1/2})$ with $u(\alpha, \beta)$ a value drawn from the uniform distribution of the range $[\alpha, \beta]$. Each loading of $\mathbf{\Lambda} = (\lambda_{ij})$ is set to $u(0.3, 1)$, and the value $\lambda^2_{[q]}$ is obtained that is the q-th smallest among all λ^2_{ij}, which is followed by transforming the loadings with $\lambda^2_{ij} \leq \lambda^2_{[q]}$ into zeros. Further, the initial $\mathbf{\Lambda}$ is normalized so as to satisfy $\mathrm{diag}(\mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}^2) = \mathbf{I}_p$.

Let $\mathbf{B}_{ql} = [\mathbf{\Lambda}_{ql}, \mathbf{\Psi}_{ql}]$ denote the solution of $\mathbf{B}$ resulting from the l-th run of the SOFA algorithm for $SP(\mathbf{\Lambda})$ set at a specified q, with $l = 1, \ldots, L_q$. We regard $\mathbf{B}_{ql*} = [\mathbf{\Lambda}_{ql*}, \mathbf{\Psi}_{ql*}]$ with $l^* = \arg\min_{1 \leq l \leq L_q} f_S(\mathbf{B}_{ql})$ as the optimal solution $\hat{\mathbf{B}}_q$, and define $\mathbf{B}_{ql}$ being a local minimizer as $\Delta(\mathbf{B}_{ql}, \mathbf{B}_{ql*}) = 0.5(||\mathbf{\Lambda}_{ql} - \mathbf{\Lambda}_{ql*}||_1/mp + ||\mathbf{\Psi}_{ql} - \mathbf{\Psi}_{ql*}||_1/p) > 0.1^3$, with $||\cdot||_1$ denoting the sum of the absolute values of the elements of the argument. Here, the suitable L_q (number of runs) is unknown beforehand. We thus employ a strategy in which L_q is initialized at an integer and increased until $\{\mathbf{B}_{ql};\ l = 1, \ldots, L_q\}$ include the two equivalently optimal solutions $\mathbf{B}_{ql*}$ and $\mathbf{B}_{ql^\#}$ satisfying $\Delta(\mathbf{B}_{ql*}, \mathbf{B}_{ql^\#}) \leq 0.1^3$ and $l^* = \mathrm{argmin}_{1 \leq l \leq L} f(\mathbf{\Theta}_l)$ with $l^\# \neq l^*$. This procedure is formally stated as follows:

1. Set $L_q = 50$ and obtain $l^* = \arg\min_{1 \leq l \leq L_q} f_S(\mathbf{B}_{ql})$
2. Go to 6, if $\Delta(\mathbf{B}_{ql*}, \mathbf{B}_{ql^\#}) \leq 0.1^3$ is satisfied for $l^* \neq l^\#$; otherwise, go to 3.
3. Set $L_q := L_q + 1$, and let $\mathbf{B}_{ql^\#}$ be the output from another run.
4. Exchange $\mathbf{B}_{ql*}$ for $\mathbf{B}_{ql^\#}$ if $f_S(\mathbf{B}_{ql^\#}) < f_S(\mathbf{B}_{ql*})$.
5. Go to 6 if $\Delta(\mathbf{B}_{ql*}, \mathbf{B}_{ql^\#}) \leq 0.1^3$ or $L_q = 200$; otherwise, go back to 3.
6. Finish with $\hat{\mathbf{B}}_q$ set at $\mathbf{B}_{ql*}$.

In this procedure, except $\mathbf{B}_{ql*}$ and $\mathbf{B}_{ql^\#}$, the rest $L_q - 2$ solutions are local minimizers, thus the L_q value clearly indicates the sensitivity to local minima.

Appendix 3: Box Problem Data

In the box problem, the 3×1 common factor score vector $\mathbf{f} = [x,\ y,\ z]'$ is supposed to yield 20×1 observation vector $\mathbf{x}$ with $\mathbf{x} = \boldsymbol{\phi}(x,\ y,\ z) + \boldsymbol{\Psi}\mathbf{u}$, where $\boldsymbol{\phi}(x, y, z)$ is the vector function with its 20 elements defined as in the first column of Table 4. The original [12] box data matrix is 20×20, whose rows are 20 realizations of $\mathbf{x}' = \boldsymbol{\phi}'(x, y, z)$ without unique factor $\boldsymbol{\Psi}\mathbf{u}$. Here, x, y, z was set to the lengths, widths, and heights of boxes, from which the name of the problem originates. However, the 20×20 data matrix does not suit the cases of $n > p$ considered in this paper. We thus simulated the 400×20 $\mathbf{X}$ based on $\mathbf{x} = \boldsymbol{\phi}(x, y, z) + \boldsymbol{\Psi}\mathbf{u}$ with the following steps: First, we set x, y, and z at $u(1, 10)$ to have $400 \times 20\boldsymbol{\Phi}$ whose rows are the realizations of $\boldsymbol{\phi}'(x, y, z)$. Second, we sampled each element of $\mathbf{u}$ from the standard normal distribution to have 400×20 $\mathbf{U}$ with its rows $\mathbf{u}'$ and set the diagonal elements of $\boldsymbol{\Psi}$ to $0.1^{1/2}$. Third, we standardized the columns of $\boldsymbol{\Phi}$ so that their average and variance were zero and one, and had $\mathbf{X} = \boldsymbol{\Phi} + \mathbf{U}\boldsymbol{\Psi}$ whose inter-column correlations are shown in Table 5.

Bibliography

1. Adachi, K.: Some contributions to data-fitting factor analysis with empirical comparisons to covariance-fitting factor analysis. J. Jpn. Soc. Comput. Stat. **25**, 25–38 (2012)
2. Adachi, K.: Factor analysis with EM algorithm never gives improper solutions when sample covariance and initial parameter matrices are proper. Psychometrika **78**, 380–394 (2013)
3. Browne, M.W.: An overview of analytic rotation in exploratory factor analysis. Multivariate Behav. Res. **36**, 111–150 (2001)
4. de Leeuw, J.: Least squares optimal scaling of partially observed linear systems. In: van Montfort, K., Oud, J., Satorra, A. (eds.) Recent Developments of Structural Equation Models: Theory and Applications, pp. 121–134. Kluwer Academic, Dordrecht (2004)
5. Harman, H.H.: Modern Factor Analysis, 3rd edn. University of Chicago Press, Chicago (1976)
6. Holzinger, K.J., Swineford, F.: A study in factor analysis: the stability of a bi-factor solution. University of Chicago: Supplementary Educational Monographs No. 48 (1939)
7. Jennrich, R.I., Bentler, P.M.: Exploratory bi-factor analysis. Psychometrika **76**, 537–549 (2011)
8. Jolliffe, I.T., Trendafilov, N.T., Uddin, M.: A modified principal component technique based on the LASSO. J. Comput. Graphical Stat. **12**, 531–547 (2003)
9. Mulaik, S.A.: Foundations of Factor Analysis, 2nd edn. CRC, Boca Raton (2010)
10. Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**, 461–464 (1978)
11. ten Berge, J.M.F.: Least Square Optimization in Multivariate Analysis. DSWO Press, Leiden (1993)
12. Thurstone, L.L.: Multiple Factor Analysis. University of Chicago Press, Chicago (1947)
13. Trendafilov, N.T.: From simple structure to sparse components: a review. Comput. Stat. **29**, 431–454 (2014)
14. Trendafilov, N.T., Unkel, S.: Exploratory factor analysis of data matrices with more variables than observations. J. Comput. Graphical Stat. **20**, 874–891 (2011)
15. Unkel, S., Trendafilov, N.T.: Simultaneous parameter estimation in exploratory factor analysis: an expository review. Int. Stat. Rev. **78**, 363–382 (2010)
16. Zou, H., Hastie, T., Tibshirani, R.: Sparse principal component analysis. J. Comput. Graphical Stat. **15**, 265–286 (2006)