Sparse exploratory factor analysis *

Nickolay T. Trendafilov†
School of Mathematics and Statistics, Open University, UK
Sara Fontanella
Department of Medicine, Imperial College London, UK
Kohei Adachi
Graduate School of Human Sciences, Osaka University, Japan

June 13, 2017

Abstract

Sparse principal component analysis is a very active research area in the last decade. It produces component loadings with many zero entries which facilitates their interpretation and helps avoid redundant variables. The classic factor analysis is another popular dimension reduction technique which shares similar interpretation problems and could greatly benefit from sparse solutions. Unfortunately, there are very few works considering sparse versions of the classic factor analysis. Our goal is to contribute further in this direction.

We revisit the most popular procedures for exploratory factor analysis, maximum likelihood and least squares. Sparse factor loadings are obtained for them by, first, adopting a special re-parameterization and, second, by introducing additional ℓ_1-norm penalties into the standard factor analysis problems. As a result we propose sparse versions of the major factor analysis procedures. We illustrate the developed algorithms on well known psychometric problems. Our sparse solutions are critically compared to ones obtained by other existing methods.

Key words: eigenvalue re-parameterization, penalties inducing sparseness, optimization on matrix manifolds.

*This work is supported by a grant RPG-2013-211 from The Leverhulme Trust, UK.
†Corresponding author: Nickolay.Trendafilov@open.ac.uk

1 Introduction

Sparse principal component analysis (PCA) is a very active research area in the last decade. The origins of this new concept are considered in (Jolliffe, 2002). The usual practice to interpret either component or factor loadings is to ignore the ones with small magnitude, or set to zero loadings smaller than certain threshold value. This makes the loadings matrix *sparse* artificially and subjectively. The initial idea behind the sparse PCA was to eliminate this subjective thresholding of the component loadings and facilitate their interpretation, especially when the number of the original variables is large. It was additionally realized that sparse PCA helps avoiding redundant variables.

A great number of papers appeared (and still continue to appear) to solve this difficult but very important for the modern applications problem (Trendafilov, 2014). Exploratory factor analysis (EFA) is another popular dimension reduction technique (Mulaik, 2010). However, there exist very few works dealing with the modernization of the classic EFA parameter estimation in order to produce sparse factor loadings, e.g. Choi *et al.* (2011); Hirose and Yamamoto (2015); Ning and Georgiou (2011). The recently proposed alternative to the classic rotation approach by Trendafilov and Adachi (2015) is primarily designed for interpreting component loadings.

A vector or a matrix is called sparse when the number of the non-zero entries, called cardinality, is much smaller than the number of all elements. Clearly, this definition is not precise and leaves room for discussion for the level of spareness. Compared to the classic Thurstone's concept for simple structure, the sparseness concept does not make any requirements for the pattern of the zero entries. Some comparative discussion of these two concepts is available in (Trendafilov, 2014).

Currently, there are two ways to impose sparseness: by penalties, or by explicit requirement for the cardinality of the solution. The LASSO (**L**east **A**bsolute **S**hrinkage and **S**election Operator) is the most popular sparseness inducing penalty. For a vector x of fixed length, say $\|x\|_2 = 1$, LASSO penalizes the sum of the absolute values of its elements, i.e. $\|x\|_1 < \tau$. The reduction of τ produces more and more zero entries in x. In the available sparse techniques, adopting penalties, the level of sparseness is controlled by such tuning parameter(s). However, the sparser solutions give worsen fit to the data. Thus, the big problem of any sparse technique is to compromise between sparseness and goodness-of-fit.

In this work, we modify the classic EFA by introducing sparse-inducing constraints on the factor loadings. The main goal is to obtain easily interpretable factor loadings which are sparse in an optimal way. Some preliminary results were announced in (Fontanella *et al.*, 2014).

The EFA model is considerably more complicated than PCA. Particularly, obtaining sparse factor loadings is considerably more complicated by the presence of other parameters affecting the overall fit of the EFA model. For example, the nice feature of the sparse PCA that sparse component loadings can be obtained

column by column by deflation is not available for EFA. Also, penalizing the ℓ_1 matrix norm of the factor loadings can lead to solutions with large unique variances. Thus, imposing sparseness inducing constraints on the factor loadings in EFA is less straightforward than with the component loadings in PCA.

The currently existing works on sparse EFA (Choi *et al.*, 2011; Hirose and Yamamoto, 2015; Ning and Georgiou, 2011) use penalties to obtain sparse factor loadings. Choi *et al.* (2011) and Hirose and Yamamoto (2015) enhance the classic EM algorithm with penalties, while Ning and Georgiou (2011) solve the classic maximum likelihood (ML) EFA with additional LASSO-type penalty.

We also adopt LASSO-like penalties to achieve sparse factor loadings. They can be readily incorporated in the EFA reparameterization proposed by Trendafilov (2003) for arbitrary EFA formulation(ML or other). This reparameterization considers the matrix of factor loadings Λ as a product of an ortonormal matrix Q, and a diagonal matrix D. Then, Q is sparsified with LASSO-type penalties, and the sparseness is preserved after multiplication by D. Thus, Q takes care for the pattern of sparseness of Λ, i.e. the locations of the zero loadings, while D adjusts the magnitudes of Λ for better fit.

The paper is organized as follows. Section 2 briefly revisits the EFA model and the assumptions imposed on its parameters, as well as the EFA reformulation as a parameter estimation (optimization) problem under several most popular goodness-of-fit measures (cost functions). New parameterization of the EFA models is considered in Section 3, which is then utilized to define the corresponding sparse EFA problems in Section 4. The sparse EFA problems are solved as optimization problems on matrix manifolds. The performance of the proposed algorithms is demonstrated on several artificial and real well-know data sets in Section 5. The Appendix contains the derivation of the gradient of the sparse inducing penalty based on the ℓ_1-norm.

2 The classic EFA model and its estimation

EFA is a model-based multivariate technique that aims to explain the relationships among p manifest random variables by r ($\ll p$) latent random variables called *common* factors F. The EFA model assumes that some portion of the variation of each observed variable remains unaccounted for by the common factors. Thus, p additional latent variables called *unique* factors U are introduced, each of which accounts for this portion of variance of the corresponding manifest variable (Mulaik, 2010). In formal terms, the EFA model represents/approximates a given $n \times p$ data matrix Z of p observed (standardized) variables on n observations as a linear combination of r common and p unique factors F and U

$$Z \approx F\Lambda^\top + U\Psi, \tag{1}$$

where Λ and Ψ are parameter matrices with sizes $p \times r$ and $p \times p$ respectively. Λ contains the *factor loadings*, and Ψ is diagonal and contains the standard de-

viations for the unique factors U. The choice of r is either subjective or based on preliminary validation. In both cases its value is subject to some limitations (Mulaik, 2010). The r-factor model (1) assumes that all involved random variables (Z, F and U) have zero means and unit variances, and that both common and unique factors are uncorrelated. Most importantly, F and U are also assumed *mutually* uncorrelated, and all diagonal entries of Ψ are assumed *non-zero*. Following the r-model defined above and the assumptions made, it can be found that the sample correlation matrix R is presented/approximated by EFA as:

$$R \approx R_{ZZ} = \Lambda\Lambda^T + \Psi^2 \ . \tag{2}$$

Thus, the main problem of EFA is to find the pair $\{\Lambda, \Psi\}$ which gives the best fit in some sense to the sample correlation matrix R (for certain r). If the data are assumed normally distributed the maximum likelihood principle can be applied (Mulaik, 2010). Then, finding $\{\Lambda, \Psi\}$ can be formulated as minimizing the following negative loglikelihood function (Jöreskog, 1977; Mulaik, 2010):

$$\min_{\Lambda,\Psi} \quad \log(\det(\Lambda\Lambda^T + \Psi^2)) + \mathrm{trace}((\Lambda\Lambda^T + \Psi^2)^{-1}R) \ , \tag{3}$$

which for short is called ML-EFA.

If nothing is assumed about the distribution of the data, the loglikelihood function (3) can still be used as a measure of the discrepancy between the model and the sample correlation matrices, R_{ZZ} and R. There are a number of other discrepancy measures (Jöreskog, 1977) which are used in place of (3). A natural choice is the least squares approach for fitting the factor analysis model (2), which can be formulated as the following general class of weighted least squares problems:

$$\min_{\Lambda,\Psi} \|(R - \Lambda\Lambda^T - \Psi^2)V\|^2 \ , \tag{4}$$

where V is a matrix of weights, and $\|\|$ denotes the Frobenius matrix norm $\|A\|^2 = \mathrm{trace}A^TA$. The case of $V = I_p$, an identity $p \times p$ matrix, is known as the least squares factor analysis, LS-EFA. The second special case $V = R^{-1}$, is known as the generalized least squares problem, GLS-EFA.

The solutions of the minimization problems ML, LS and GLS are *not* unique. To eliminate the rotational indeterminacy, the unknowns Λ and Ψ are sought subject to the following constraints (Jöreskog, 1977): for ML and GLS,

$$\Lambda^T\Psi^{-2}\Lambda \text{ to be diagonal} \ , \tag{5}$$

and for LS,

$$\Lambda^T\Lambda \text{ to be diagonal} \ . \tag{6}$$

The constraint (5) explains why Ψ is required by EFA to have non-zero diagonal entries. This assumption is equivalent to the assertion that no observable random

variable can ever be explained entirely by a common factor. This assumption and several other features, e.g. factor scores indeterminacy (Mulaik, 2010), make the EFA model highly controversial, which probably explains why EFA is far less popular dimension reduction technique than PCA.

For any orthogonal $r \times r$ matrix P we have:

$$R_{ZZ} = \Lambda\Lambda^T + \Psi^2 = \Lambda P P^T \Lambda^T + \Psi^2 = \Lambda P (\Lambda P)^T + \Psi^2 \ , \tag{7}$$

which is known as the rotation indeterminacy in EFA. Indeed, the constraint (5) eliminates the indeterminacy (7), however such solutions are usually difficult for interpretation. Instead, the common practice is to make use of (7): rotate the initially found factor loadings Λ by some kind of "simple structure" rotation (Mulaik, 2010) to make them more interpretable. By "interpretable" it is meant that each factor has only few large loadings. The rule is to ignore, effectively make *zero*, the remaining rather small ones. In fact, the factor loadings interpretation relies on artificially constructed *sparse* loadings Λ, many of which are neglected, and thus considered zeros.

We propose to modify the EFA fitting problems (3) and (4) by introducing sparse-inducing constraints. Then, the resulting factor loadings Λ will be sparse in an optimal way. This strategy is not new. The same interpretation problem occurs in PCA. Its solution led in the last decade to developing a great number of new procedures directly producing sparse component loadings, which considerably simplifies their interpretation. In contrast, there are very few works on sparse EFA, e.g. (Choi *et al.*, 2011; Hirose and Yamamoto, 2015; Ning and Georgiou, 2011). The proposed work makes a further contribution to this new research area.

3 PCA-like reparameterization of EFA

It has been argued in Trendafilov (2003), that, in fact, the constraints (5) and (6) facilitate the algorithms for numerical solution of the different EFA definitions (3) and (4), see for details e.g. (Jöreskog, 1977; Mulaik, 2010). As we mentioned, occasionally (5) and (6) may facilitate the interpretation of Λ, but in general this is not the case. The alternative traditional approach to rotate the initial factor loadings Λ to "simple structure" gives, in turn, rotated factor loading violating (5) and (6).

In this work we adopt the new formulation of the EFA estimation problems (3) and (4) proposed in (Trendafilov, 2003). The constraints (5) and (6) will not be needed any more. The only *natural* constraints inferred from the r-factor analysis model (2) are that the $p \times r$ matrix Λ should have full column rank, and that the $p \times p$ diagonal matrix Ψ^2 should be positive definite. Additionally, we relax the second condition and assume positive *semi*-definite diagonal Ψ^2. There are two reasons for this. From EFA model point of view this constraint seems too restrictive. From numerical point of view the algorithms developed in (Trendafilov,

2003) do not relay on $\Psi^2 > 0$. Moreover, maintaining $\Psi^2 > 0$ may contradict achieving high level of sparseness (Section 5 and Section 6).

Consider the eigenvalue decomposition of the positive semi definite $\Lambda\Lambda^T$ of rank at most r in (2), i.e. let $\Lambda\Lambda^T = QD^2Q^T$, where D^2 is an $r \times r$ diagonal matrix composed by the largest (nonnegative) r eigenvalues of $\Lambda\Lambda^T$ arranged in descending order and Q is a $p \times r$ orthonormal matrix containing the corresponding eigenvectors. Note that for this reparameterization $\Lambda^T\Lambda$ is diagonal, i.e. the condition (6) is fulfilled automatically. Then (2) can be rewritten as:

$$R_{ZZ} = QD^2Q^T + \Psi^2 \ . \tag{8}$$

It may look that the reformulated EFA model (8) is more restrictive. In fact, it is more general. To see this, one can write (8) in more abstract terms as

$$R_{ZZ} = S + \Psi^2 \ ,$$

where S denotes a semidefinite symmetric matrix of low rank r. The classic EFA considers two separate cases with either uncorrelated (orthogonal) factors and $S = \Lambda\Lambda^T$, or with correlated factors and $S = \Lambda\Phi\Lambda^T$, where Φ is the correlation matrix collecting the correlations among the factors. No matter which of these two forms is used for S, it can be rewritten as $S = QD^2Q^T$ making use of the truncated eigenvalue decomposition of S. In other words, (8) absorbs both cases of either uncorrelated or correlated factors, and thus, it is more general that the original EFA formulation.

Thus, instead of the pair $\{\Lambda, \Psi\}$, a triple $\{Q, D, \Psi\}$ is sought in (Trendafilov, 2003). Thus, the new factor loadings Λ are given by QD. Clearly, when Q is sparse, Λ will have the same sparseness. Note, that the model (8) with sparse Q (and Λ) does not permit rotations, only permutations are possible. In order to maintain the factor analysis constraints, the triple $\{Q, D, \Psi\}$ should be sought such that Q be a $p \times r$ orthonormal matrix, and D and Ψ – diagonal. Note, that we do not insist for non-singular Ψ, however the singularity of D implies failing of the r-factor analysis model.

The new formulation of the factor analysis estimation problems is straightforward. Indeed, for a given sample correlation matrix R, the ML-EFA is reformulated as follows:

$$\min_{Q,D,\Psi} \log(\det(QD^2Q^T + \Psi^2)) + \mathrm{trace}((QD^2Q^T + \Psi^2)^{-1}R) \ , \tag{9}$$

and the LS- and the GLS-EFA estimation problems are rewritten as:

$$\min_{Q,D,\Psi} \|(R - QD^2Q^T - \Psi^2)V\|^2 \ . \tag{10}$$

4 Sparse factor loadings with penalized EFA

In the new EFA formulation (8), the factor loadings Λ are parameterized as QD. This implies that Λ and Q have the same patterns of zero entries, i.e. they are

equally sparse. To see this, consider parameterization of the following hypothetical 5×2 sparse factor loadings matrix Λ:

$$\Lambda = \begin{pmatrix} \lambda_{11} & 0 \\ 0 & \lambda_{22} \\ \lambda_{31} & 0 \\ \lambda_{41} & 0 \\ 0 & \lambda_{52} \end{pmatrix} = QD = \begin{pmatrix} q_{11} & 0 \\ 0 & q_{22} \\ q_{31} & 0 \\ q_{41} & 0 \\ 0 & q_{52} \end{pmatrix} \begin{pmatrix} d_1 & 0 \\ 0 & d_2 \end{pmatrix} = \begin{pmatrix} q_{11}d_1 & 0 \\ 0 & q_{22}d_2 \\ q_{31}d_1 & 0 \\ q_{41}d_1 & 0 \\ 0 & q_{52}d_2 \end{pmatrix} ,$$

which demonstrates that Q is solely responsible for the locations of the zeros in Λ, and D adjusts the magnitudes of the nonzero loadings.

Thus, to achieve sparse factor loadings Λ, one simply needs sparse orthonormal Q, which is a problem resembling the well-known sparse PCA, e.g. (Trendafilov and Jolliffe, 2006).

Let $\mathbf{q}_i$ denote the ith column of Q, i.e. $Q = (\mathbf{q}_1, \mathbf{q}_2, ..., \mathbf{q}_r)$, and $\boldsymbol{\tau} = (\tau_1, \tau_2, ..., \tau_r)$ be a vector of tuning parameters, one for each column of Q. We consider a penalized version of EFA, where the ℓ_1 norm of each of the columns of Q is penalized, i.e. $\|\mathbf{q}_i\|_1 \leq \tau_i$ for all $i = 1, 2, ..., r$. Introduce the following discrepancy vector $\mathbf{q}_\tau = (\|\mathbf{q}_1\|_1, \|\mathbf{q}_2\|_1, ..., \|\mathbf{q}_r\|_1) - \boldsymbol{\tau}$, which can also be expressed as $\mathbf{q}_\tau = 1_p^\top [Q \odot \text{sign}(Q)] - \boldsymbol{\tau}$, where $\text{sign}(Q)$ is a matrix containing the signs of the elements of Q, and 1_p is a vector with p unit elements. We adapt the scalar penalty function $\max\{x, 0\}$ used by Trendafilov and Jolliffe (2006) to introduce the following vector penalty function $P_\tau(Q) = [\mathbf{q}_\tau \odot (1_r + \text{sign}(\mathbf{q}_\tau)]/2$. Then, the penalized versions of (9) and (10) can be defined, for the ML-EFA as:

$$\min_{Q,D,\Psi} \log(\det(R_{ZZ})) + \text{trace}((R_{ZZ})^{-1}R) + \mu P_\tau(Q)^\top P_\tau(Q) \ , \tag{11}$$

and for the LS- and the GLS-EFA as:

$$\min_{Q,D,\Psi} \|(R - R_{ZZ})V\|^2 + \mu P_\tau(Q)^\top P_\tau(Q) \ . \tag{12}$$

Note, that $P_\tau(Q)^\top P_\tau(Q)$ penalizes the sum of squares of $\|\mathbf{q}_i\|_1 - \tau_i$ for all $i = 1, 2, ..., r$, i.e. precise fit of $\|\mathbf{q}_i\|_1$ to each tuning parameter τ_i cannot be achieved.

5 Numerical examples

In this Section we first explore the behavior of the proposed sparse EFA on simulated data considered by Choi *et al.* (2011). Then, in contrast to Choi *et al.* (2011); Hirose and Yamamoto (2015); Ning and Georgiou (2011), we consider two examples from the classic EFA. Our goal is to demonstrate that the results from the new procedure agree well with the results from the old EFA solutions, but are easier and clearer for interpretation.

5.1 Simulated data (Choi *et al.*, 2011)

We examine the performance of the proposed approach by employing the simulated data constructed by Choi *et al.* (2011). They take a hypothetical 12×4 sparse loadings matrix Λ with the following non-zero entries: $\lambda_{11} = \lambda_{21} = \lambda_{31} = 1.8, \lambda_{42} = \lambda_{52} = \lambda_{62} = 1.7, \lambda_{73} = \lambda_{83} = \lambda_{93} = 1.6$ and $\lambda_{10,4} = \lambda_{11,4} = \lambda_{12,4} = 1.5$, and $\Psi^2 = \text{Diag}(1.27,\ .61,\ .74,\ .88,\ .65,\ .81,\ .74,\ 1.3,\ 1.35,\ .74,\ .92,\ 1.32)$. Thus, the "population" sparse loadings matrix has 36 zeros. The "population" covariance matrix is created by (2), and then we normalize it to obtain a correlation matrix used to generate normally distributed zero mean independent samples.

We generate 100 data matrices each of which is analyzed by sparse ML-EFA. For this reason we solve (11) for six decreasing values $\tau_0 (= \sqrt{12}, 3.0534, 2.6427, 2.2321, 1.8214, 1.4107)$, where each value of τ_0 is applied to all columns, i.e. $\tau = \tau_0 1_r$. In such situations, we use τ for both scalar or vector, depending on the context. The solution for any particular τ is used as a starting value for the next run with the consecutive τ. The starting values for the first $\tau (= \sqrt{12} = 3.4641)$ are chosen randomly.

The goal of this simulation experiment is to demonstrate that, in general, the sparsity level of the factor loadings increases when τ decreases. In order to show also how the number of zero loadings increases within the 100 runs, we provide a more complicated graphical display in Figure 1. For $\tau = \sqrt{12}$, nearly all factor loadings matrices are dense, only 4 of them contain a single zero entry. For $\tau = 2.6427$, there are 22 factor loadings matrices with no zero entry, 49 – with a single zero entry, 22 – with two zero entries, and the rest seven have three zero loadings. For $\tau = 1.4107$, there are 93 factor loadings matrices with 36 zero entries, 6 – with a 35 zeros, and only one – with 34 zero entries. In other words, with $\tau = 1.4107$ the sparse ML-EFA achieves 93% exact recovery of the underlying sparseness. The case $\tau = 1$ is not depicted, as it produces excessive sparseness. Clearly, the correct tuning parameter for this problem is around $\tau = 1.4107$. After the correct sparseness is localized, one can perform further runs to achieve the best corresponding fit.

5.2 Harman's Five Socio-Economic Variables (Harman, 1976, p.14)

First, we illustrate the proposed procedures for sparse EFA on a well known data set from classic EFA, namely the Harman's Five Socio-Economic Variables (Harman, 1976, p.14). This small data set is interesting because the two- and the three-factor solutions from LS- and ML-EFA are 'Heywood cases' (Harman, 1976; Mulaik, 2010), i.e. Ψ^2 contains zero diagonal entries, or $\Psi^2 \geq 0$. One-factor solution is not considered interesting as it explains only 57.47% of the total variance.

Table 1 contains several sparse LS-EFA solutions of (12) starting with $\tau = \sqrt{5} = 2.2361$, which is equivalent to the standard (non sparse) LS-EFA solution.

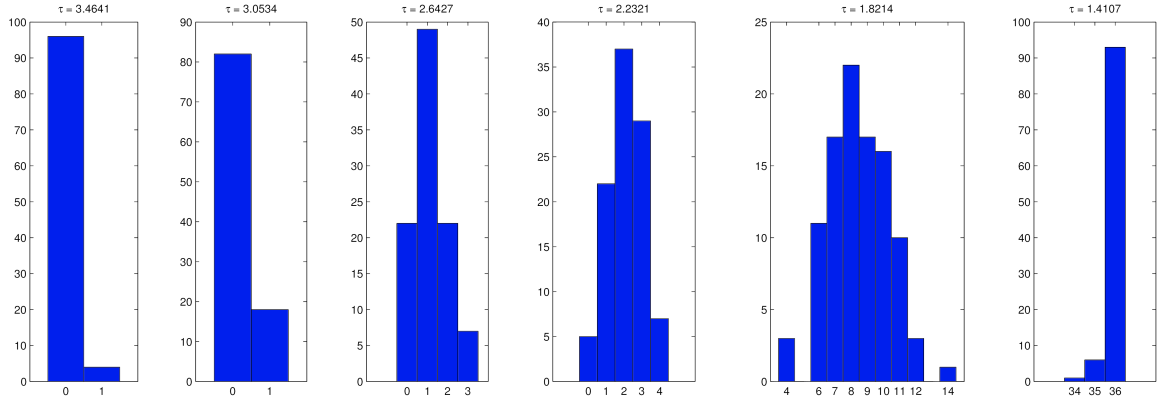

Figure 1: Number of zeros obtained in 100 runs of sparse ML-EFA (11) for different τ. For each plot, the x axis denotes the number of zero entries in a single matrix of factor loadings, e.g none, or one, or 36. The y axis denotes the number of matrices with the corresponding (on x) number of zero entries.

For all of them we have $\Psi^2 \geq 0$. Clearly, POP, EMPLOY and HOUSE tend to be explained by the common factors only, which is already suggested by the non sparse solution ($\tau = \sqrt{5}$). Increasing the sparseness of the factor loadings results in variables entirely explained by either a common or unique factor. The presence of loadings with magnitudes over 1 demonstrates the well known weakness of LS-EFA in fitting the unit diagonal of a correlation matrix. It is well known that ML-EFA does not exhibit this problem which is illustrated by the next example.

VARS	$\tau = \sqrt{5}$			$\tau = 1.824$			$\tau = 1.412$			$\tau = 1$		
	QD		Ψ^2	QD		Ψ^2	QD		Ψ^2	QD		Ψ^2
POP	-.62	-.78	.00	.07	1.0	.00	-.00	1.0	.00	.00	-.99	.00
SCHOOL	-.70	.52	.23	.94	-.20	.07	.85	-.00	.27	-.28	-.00	.92
EMPLOY	-.70	-.68	.04	.19	.87	.21	-.00	1.0	.00	-.00	-.99	.00
SERVICES	-.88	.15	.20	.78	.23	.34	.58	.13	.65	-.18	-.00	.97
HOUSE	-.78	.60	.03	1.0	-.22	.00	1.1	-.07	.00	-1.2	.00	.00

Table 1: LS-EFA solutions for Five Socio-Economic Variables, (Harman, 1976, p.14).

5.3 Holzinger-Harman's Twenty-Four Psychological Tests (Harman, 1976, p.123)

Next, we illustrate the proposed procedures for sparse EFA on another well known data set from classic EFA, namely the Holzinger-Harman' Twenty-Four Psycho-

logical Tests (Harman, 1976, p.123). It is widely used to illustrate different aspects of classic EFA (Harman, 1976; Mulaik, 2010).

The correlation matrix (Harman, 1976, p.124) of these data is non-singular and we apply ML-EFA (11). The first five columns of Table 2 contain the solution (factor loadings QD and unique variances Ψ^2) of (11) with $\tau = \sqrt{24} = 4.899$, i.e. the standard ML-EFA solution, which is nearly identical to the ML solution obtained in (Harman, 1976, p.215). Then, we rotate (with normalization) the factor loadings QD from the first four columns by VARIMAX from `MATLAB` (MATLAB, 2014), and the result is given in the next four columns of Table 2.

	$\tau = \sqrt{24} = 4.899$					Varimax rotated				$\tau = 2.2867$				$\tau = 2.1697$				
	QD				Ψ^2	QD and $T_{.41}$				QD			Ψ^2	QD				Ψ^2
1	.60	.39	-.22	.02	.44	**.69**	.16	.19	.16	-.88			.31	-.83				.41
2	.37	.25	-.13	-.03	.78	**.44**	.12	.08	.10	-.25			.86					1.0
3	.41	.39	-.14	-.12	.64	**.57**	.14	-.02	.11	-.53			.70	-.39				.76
4	.49	.25	-.19	-.10	.65	**.53**	.23	.10	.08	-.55			.69	-.55				.67
5	.69	-.28	-.03	-.30	.35	.19	**.74**	.21	.15		.82		.35		.81			.36
6	.69	-.20	.08	-.41	.31	.20	**.77**	.07	.23		.84		.32		.84			.32
7	.68	-.29	-.08	-.41	.28	.20	**.81**	.15	.07		.86		.29		.86			.29
8	.67	-.10	-.12	-.19	.49	.34	**.57**	.24	.13		.64		.54		.63			.54
9	.70	-.21	.08	-.45	.26	.20	**.81**	.04	.23		.87		.27		.86			.28
10	.48	-.49	-.09	.54	.24	.12	.17	**.83**	.17	.17		.91	.29	.07		.89		.33
11	.56	-.14	.09	.33	.55	.12	.18	**.51**	.37			.63	.59			.61		.59
12	.47	-.14	-.26	.51	.44	.21	.02	**.72**	.09			.72	.50			.75		.48
13	.60	.03	-.30	.24	.49	**.44**	.19	**.53**	.08	-.29		.47	.51	-.13		.47		.58
14	.42	.02	.41	.06	.65	.05	.20	.08	**.55**				.75				.37	.79
15	.39	.10	.36	.09	.70	.12	.12	.07	**.52**				.71				.49	.73
16	.51	.35	.25	.09	.55	.41	.07	.06	**.53**				.68				.50	.72
17	.47	-.00	.38	.20	.60	.06	.14	.22	**.57**				.54				.77	.50
18	.52	.15	.15	.31	.59	.29	.03	.34	**.46**				.61				.68	.59
19	.44	.11	.15	.09	.76	.24	.15	.16	.37				.82				.18	.89
20	.61	.12	.04	-.12	.59	.40	.38	.12	.30		.33		.77		.31			.78
21	.59	.06	-.12	.23	.58	.38	.17	**.44**	.22	-.02		.50	.68			.37		.75
22	.61	.13	.04	-.11	.60	.40	.37	.12	.30		.28		.80	-.60				.64
23	.69	.14	-.10	-.04	.50	**.50**	.37	.24	.24	-.59		.02	.64	-.70				.55
24	.65	-.21	.02	.18	.50	.16	.37	**.50**	.30			.63	.59			.61		.60
(3)			14.28									16.71				17.08		

Table 2: ML-EFA solutions for Twenty-Four Psychological Tests (Harman, 1976, p.123) with their values for the ML objective function (3).

The classic EFA approach to interpret the rotated loadings QD is to set up a threshold which cuts off the loadings with lesser magnitudes. It is helpful to sort the loadings magnitudes and look for jumps indicating for possible cut-off values. The left hand side panel of Figure 2 gives a plot of the sorted rotated ML loadings. Clearly, the largest jump in the loadings magnitudes is at .58. If this is taken as a cut-off value only 7 non-zero loadings will be left for interpretation (out of 96), which may seem too extreme simplification. The next largest jump is at .25, which would leave 38 loadings for interpretation. This is nearly 40% of all loadings and may seem too many. The next largest jumps are at .46 and .31, which would leave for interpretation 18 and 34 loadings respectively. Interpreting 34 loadings still looks difficult, so one can drop .31 as a cut-off point. The other cut-off point .46 looks attractive with only 18 points for interpretation (which is nearly 20% of all loadings), but one may feel unconformable with such a high threshold. Then, consider the next ones being either .34 or .41, leaving respectively 32 and 22 for interpretation. As 32 loadings still look too many for interpretation, for this example we choose to work with .41 as a cut-off point and the corresponding

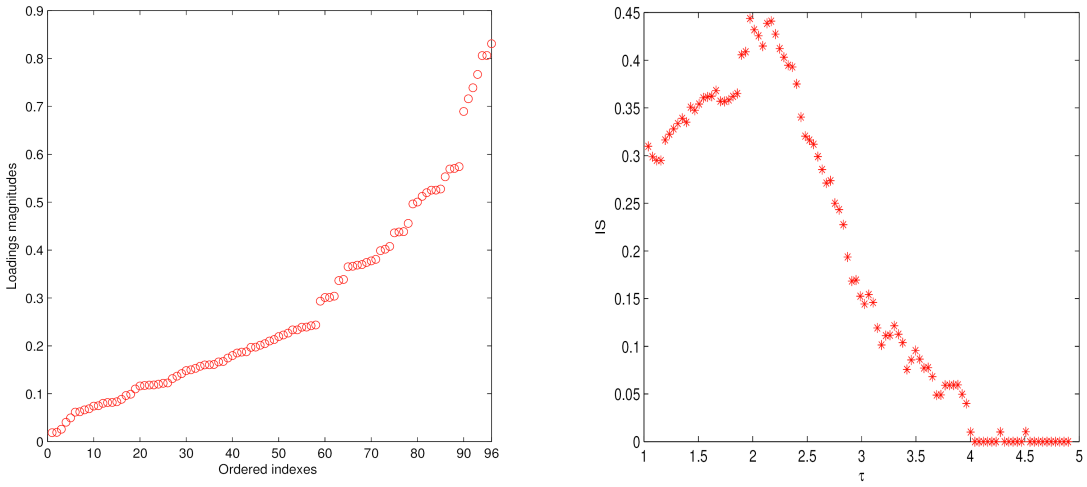

Figure 2: Number of zeros obtained in 100 runs of sparse ML-EFA (11) for different τ.

loadings are given in bold in Table 2. The hypothetical matrix of thresholded loadings (with bold loadings and zeros elsewhere) is denoted by $T_{.41}$.

It is clear that this lengthy way of choosing the interpretation threshold is completely subjective and lacks any optimality. For large loadings matrices such an approach would be simply impossible to apply. The sparse EFA provides a reasonable alternative by directly producing sparse matrix of factor loadings. However, the sparse solutions produce worse fit than the classic ones. In general, the increase of the loadings sparseness worsens the fit. Thus, one needs to find a compromise between fit and sparseness, i.e. to optimize the value of the tuning parameter τ in (11). In sparse PCA, tuning parameters as τ are usually found by cross-validation for large applications, or by employing information criteria (Trendafilov, 2014). For small applications, as those considered in the paper, the optimal tuning parameter τ can be easily located by solving the problem for several values of τ and compromising between sparseness and fit. Another option is to solve (11) for a range of values of τ and choose the most appropriate of them based on some index of sparseness. Here we use the following one:

$$\mathrm{IS}(\tau) = \frac{\text{original fit}}{\text{fit for } \tau} \times \left(\frac{\#_0}{pr}\right)^2 , \tag{13}$$

where $\#_0$ is the number of zeros among all pr loadings of $\Lambda = QD$. IS increases with the sparseness and when the fit of the sparse solution of (11) is close to the original one, i.e. with $\mu = 0$.

We solve (11) for 100 values of τ from $\tau = \sqrt{24} = 4.899$ to 1. The values of the

index of sparseness (13) are depicted in the right hand side panel of Figure 2. The maximum of IS is for $\tau = 2.1697$ and is IS(2.1697) = 0.4409. The corresponding sparse factor loadings QD and unique variances Ψ^2 are depicted in the last five columns of Table 2. This matrix of factor loadings has 25 non-zero loadings (26%). The ML fit is 17.08. It turns out, that it is closer to the hypothetical $T_{.46}$ in least-squares sense than to $T_{.41}$. The sparse matrix obtained with $\tau = 2.2867$ (the second largest IS(2.2867) = 0.4032) is also depicted in Table 2. It may look more like $T_{.41}$, but in fact, it is also closer to $T_{.46}$ in least-squares sense than to $T_{.41}$. Thus, the adopted index of sparseness (13) implies that the choice of the cut-off point .41 is incorrect, and should have been set to .46.

6 Comparison to other methods

In this section we compare the performance of the proposed procedure `SEFA` for sparse EFA with three other approaches by Choi *et al.* (2011); Hirose and Yamamoto (2015) and Ning and Georgiou (2011) developed with the same purpose. However, it turns out that the available software `GEM` for the method by Choi *et al.* (2011) does not work properly. The code realizing the method by Ning and Georgiou (2011) is lost, according to the authors' response.

Thus, our procedure will be compared with the method proposed by Hirose and Yamamoto (2015). Their codes in `R` are available online as the package `fanc`, which also will be used for short reference to their work. We will demonstrate how `fanc` finds sparse factor solutions for the Five Socio-Economic Variables, (Harman, 1976, p.14), and for the Twenty-Four Psychological Tests (Harman, 1976, p.123). The results will be compared with the performance of `SEFA`.

The solution with $\rho = .001$ (and less) is identical with the non-constrained solution ($\rho = 0$) depicted in Table 7. We find several (`fanc`) solutions with increasing values of ρ. They are reproduced in Table 3:

VARS	$\rho = .005$			$\rho = .01$			$\rho = .05$			$\rho = .1$		
	$\Lambda = QD$		Ψ^2	Λ		Ψ^2	QD		Ψ^2	QD		Ψ^2
POP	.991	.001	.005	.983		.005	.938		.005	.916	-.002	.005
SCHOOL		.891	.193		.882	.194		.828	.195		.802	.196
EMPLOY	.966	.117	.036	.959	.114	.036	.913	.107	.036	.890	.101	.036
SERVICES	.424	.783	.185	.418	.774	.185	.388	.727	.186	.374	.706	.186
HOUSE	.006	.955	.074	.004	.947	.074		.895	.073		.870	.073
Value of (3)	-1.072			-1.071			-1.052			-1.032		

Table 3: Four `fanc` solutions with LASSO penalty ($\gamma = \infty$) for Five Socio-Economic Variables, (Harman, 1976, p.14).

By looking at Table 3, one can conclude that the solution with $\rho = .05$ is the best: it has three exact zero loadings, and its fit is better than the next one with $\rho = .1$, which has two zero loadings. Hirose and Yamamoto (2015) provide a number of ways to evaluate the quality of their solutions. Some of them are provided in Table 4 for completeness.

ρ_0	ρ	Goodness-of-fit			Criteria			
		GFI	AGFI	SRMR	AIC	BIC	CAIC	EBIC
.001	.0010784	.7039321	$-\infty$	.0123146	17.13224	24.40584	39.40584	70.45754
.005	.0052776	.7056696	-3.4149556	.0197663	15.13684	21.92553	35.92553	63.37206
.01	.0116752	.7070369	-1.1972236	.0301804	13.14765	19.45144	32.45144	56.29280
.05	.0571375	.7097125	-.4514375	.0983878	11.37472	17.19360	29.19360	49.42979
.1	.0849835	.7081857	-1.1886072	.1327246	13.61488	19.91867	32.91867	56.76003
.15	.1264004	.7033769	-1.2246731	.1759992	14.07468	20.37846	33.37846	57.21982
.2	.1880019	.6916270	-1.3127974	.2264768	14.91420	21.21799	34.21799	58.05935
.7	.6185911	.5735386	-2.1984603	.3974809	22.95663	29.26041	42.26041	66.10178
.8	.9200627	.2063196	-.4881507	.4632527	44.27396	47.66831	54.66831	56.87865

Table 4: Quality measures for several `fanc` solutions with LASSO penalty for Five Socio-Economic Variables, (Harman, 1976, p.14).

Note, that ρ_0 is the input value for `fanc`, while ρ is the actual value of the tuning parameter used by `fanc` to produce the loadings. According to all goodness-of-fit measures and the information criteria collected in Table 4, the solution with $\rho_0 = .05$ seems to be the best one indeed. What seems surprising is that `fanc` is incapable to produce sparser solutions, containing more zeros than three. Instead, the increase of ρ_0 results in loadings containing only two zeros. Such solutions are depicted in Table 3 with $\rho = .1$ and in Table 5 with $\rho = .7$. As ρ controls the importance of the penalty term, it is logical to expect and desirable to have sparser loadings with larger ρ, which is not the case. Moreover, further increase of ρ ($\geq .8$) simply results in invalid solutions containing one zero column. The remedy proposed by Hirose and Yamamoto (2014) is to replace the LASSO constraint by the MC+ one. The best solution we found is with $\rho = 1.6,\ \gamma = 1.5$, and is depicted in Table 5. It has five zero loadings and ML fit 1.159163. In order to get sparser solutions Hirose and Yamamoto (2014) suggest trying correlated factors. We were unable to identify a pair of parameters (ρ, γ) for which the oblique solution provides better sparseness and/or ML fit. For comparison, the `SEFA` solution with $\tau = .9$ is depicted in the last three columns of Table 5. It has six zero loadings and even better (lower) minimum of the objective function (3).

VARS	$\rho = .7$			$\rho = 1.6,\ \gamma = 1.5$			`SEFA` ($\tau = .6$)		
	Λ		Ψ^2	Λ		Ψ^2	QD		Ψ^2
POP	.705	-.015	.005	.690		.005	.971	-.000	.054
SCHOOL		.554	.214		.684	.007	.000	.760	.266
EMPLOY	.677	.053	.037	.663	.077	.030	1.00	.000	.000
SERVICES	.229	.477	.211			1.00	.000	.000	1.00
HOUSE		.645	.076		.449	.299	.000	1.00	.000
Value of (3)		-0.254			1.159			0.758	

-

Table 5: More `fanc` solutions with LASSO penalty for Five Socio-Economic Variables, (Harman, 1976, p.14). The solution for $\rho = .8$ has an empty (zero) second column.

With this simple example we demonstrate that `fanc` has two serious drawbacks to be taken into account for practical use. First, `fanc` is incapable to produce full

range of sparse solutions. Second, the relationship between the sparseness and the parameter ρ for its control is not linear. This considerably complicates the location of an optimal ρ (which provides a reasonable fit with a sensible sparseness) – the main difficulty in any sparse analysis of large data. The presence of another parameter γ for MC+, puts additional difficulty in the `fanc` application.

Now, let us move to the Twenty-Four Psychological Tests (Harman, 1976, p.123). The standard ML solution (`fanc` with $\rho = 0$) has four zeros. The solution with $\rho = 0.1$ seems the sparsest possible with LASSO penalty and has 26 zeros. The further increase of ρ produces invalid solutions. For $\rho = 0.14$, the loadings already have one zero column. Then, let us replace the LASSO constraint by MS+. Hirose and Yamamoto (2014) find solution with $\rho = .02,\ \gamma = 4$ which resembles the PROMAX solution. This solution is not satisfactory, it is not sparse enough as it has 70 non-zero loadings. The best solution they find with MS+ constraint is with $\rho = .14,\ \gamma = 1.1$, which still has plenty of non-zero loadings, 59. Our solution with the same $(\rho,\ \gamma)$ in Table 6 is nearly identical to the one reported in (Hirose and Yamamoto, 2014, Table 4.). In order to get sparser solutions, Hirose and Yamamoto (2014) suggest employing correlated factors. For this data set this strategy pays off. The best solution with oblique factors obtained with $\rho = .21,\ \gamma = 1.1$ is reported in (Hirose and Yamamoto, 2014, Table 4.) and has only 28 non-zero loadings. After several runs of `fanc`, we are unable to repeat this solution. Our solution with oblique factors and same $(\rho,\ \gamma)$ is depicted in Table 6 and has 36 non-zero loadings.

	Orth, $\rho = 0$						Orth, $\rho = .1$						Orth, $\rho = .14,\ \gamma = 1.1$						Oblique $\rho = .21,\ \gamma = 1.1$					
	Λ				Ψ^2		Λ			Ψ^2			Λ			Ψ^2			Λ			Ψ^2	
1	.06	.38	.64	.06	.44		.12	.54		.45			.45	.57	.09	.46	.73			.47				
2	.03	.25	.40	0	.78		.06	.30		.79			.28	.38		.78	.47			.78				
3	.04	.29	.51	-.12	.64		.10	.38	-.07	.68			.29	.52		.65	.55			.69				
4	-.02	.38	.45	-.02	.65		.16	.35		.67			.40	.43		.66	.58			.66				
5	-.07	.80	.01	.07	.35		.61		.09	.36	-.02		.78		-.17	.36				.80	.35			
6	.01	.83		-.08	.31	.03	.64			.32			.77		-.32	.31				.81	.34			
7	-.15	.83	-.01		.28	-.04	.65		.02	.30	-.17		.78		-.25	.29				.83	.32			
8	-.05	.68	.20	.10	.49		.46	.15	.07	.50			.69	.15		.50	.22			.57	.50			
9		.86	-.02	-.11	.26	.00	.68		-.01	.26			.79		-.36	.25		-.08		.90	.24			
10	.07	.31	-.07	.81	.24		.07		.73	.23			.52	-.32	.63	.24	-.38	1.0			.21			
11	.27	.38	.12	.46	.55	.20	.12	.10	.37	.56	.26		.52		.34	.54		.46	-.28		.57			
12	.02	.21	.27	.67	.44			.21	.52	.48			.42		.61	.45		.70			.50			
13	-.03	.39	.43	.42	.49		.11	.34	.30	.54			.52	.25	.37	.52	.31	.47			.52			
14	.47	.35	.00	.05	.65	.42	.16		.01	.64	.45		.37			.66			-.56		.69			
15	.46	.29	.08	.05	.70	.37	.10	.06		.71	.45		.32			.69			-.53		.72			
16	.46	.31	.38		.55	.34	.06	.31		.57	.42		.35	.37		.55	.35		-.37		.61			
17	.50	.34	.04	.20	.60	.42	.11	.03	.13	.61	.50		.40			.59			-.67		.54			
18	.39	.27	.31	.29	.59	.27	.01	.26	.19	.61	.35		.39	.21	.27	.59		.24	-.46		.62			
19	.29	.32	.20	.11	.76	.16	.11	.17	.04	.79	.24		.37	.18		.77			-.49		.76			
20	.16	.54	.30	.00	.59	.07	.31	.25		.60			.56	.30		.60	.41			.33	.58			
21	.12	.38	.36	.35	.58		.12	.33	.26	.58			.51	.26	.32	.57	.34	.42			.57			
22	.17	.53	.30	.01	.60	.05	.30	.26		.61			.55	.30		.61	.40			.33	.59			
23	.09	.56	.41	.11	.50		.30	.36	.06	.49			.61	.36		.50	.51			.30	.49			
24	.16	.54	.10	.41	.50	.05	.28	.11	.36	.51			.65		.26	.51		.51		.29	.51			
(3)			14.28						14.95						14.41				14.66					

Table 6: Several `fanc` solutions with LASSO ($\gamma = \infty$) and MS+ penalty for Twenty-Four Psychological Tests (Harman, 1976, p.123).

The correlations among the oblique factors for our (.21, 1.1)-`fanc` solution are:

$$\Phi = \begin{pmatrix} 1.0 & .51 & -.52 & .50 \\ .51 & 1.0 & -.51 & .47 \\ -.52 & -.51 & 1.0 & -.49 \\ .50 & .47 & -.49 & 1.0 \end{pmatrix}.$$

Clearly, the solution with oblique factors looks better, than the orthogonal one. Indeed, the corresponding values of the index of sparseness (13) are 0.3805 and 0.1472 respectively. Nevertheless, we see again that `fanc` is unable to produce solutions with arbitrary sparseness. The `fanc` problems to achieve reasonable sparseness with LASSO, further continue when applying the MC+ penalty. For some data, `fanc` can achieve better sparseness by employing correlated (oblique) factors. This looks as a way to boost the performance of the numerical method without clear benefit for the EFA solution, e.g. better fit. As illustrated in Table 2, the same problem can be solved by `SEFA` very satisfactory.

Finally, we mention briefly the two other methods proposed by Choi *et al.* (2011) and Ning and Georgiou (2011). Choi *et al.* (2011) find sparse factor loadings Λ by minimizing (3) and adding to it a LASSO type constraint ($\lambda\|\Lambda\|_1$). The minimization is carried out by an EM algorithm and the sparsification of Λ is achieved by the LARSEN algorithm. However, the available code `GEM` (`G`eneralized `E`xpectation-`M`aximization algorithm) realizing the method does not work properly. A `GEM` solution (with switched off sparsification) is depicted in the second three columns of Table 7. This solution is very different from any known ML-EFA solution (Harman, 1976). For example, the ML-EFA solution obtained by `SEFA` (with $\mu = 0$) is depicted in the first three columns of Table 7. It is not surprising (because of the EFA reparameterization) that this solution is nearly identical to the canonical form solution (Harman, 1976, Table 10.5). The `fanc` solution (with $\rho = 0$) is very similar to the standard ML solution (Harman, 1976, Table 10.5). The value of the ML objective function (3) obtained by `GEM` is much higher than both `SEFA` and `fanc`. It looks like `GEM` has some kind of normalization problem.

VARS	`SEFA`			`GEM`			`fanc`		
	$\Lambda = QD$		Ψ^2	Λ		Ψ^2	QD		Ψ^2
POP	.783	.622	.000	.282	.026	.005	.997	.007	.005
SCHOOL	-.552	.711	.190	.033	-.254	.018	.008	.898	.193
EMPLOY	.689	.697	.040	.283	-.006	.005	.974	.124	.036
SERVICES	-.147	.891	.184	.148	-.215	.015	.433	.792	.185
HOUSE	-.579	.766	.078	.031	-.278	.005	.014	.962	.074
Value of (3)	-1.0758			63.0146			-1.0723		

Table 7: Three ML-EFA solutions for Five Socio-Economic Variables, (Harman, 1976, p.14).

To obtain sparse loadings by `GEM` one needs to increase the value of the tuning parameter (λ). However, `GEM` starts to produce very frequently loadings matrices Λ with zero column, or does not converge at all. For this reason, `GEM` is not considered any further.

Ning and Georgiou (2011) consider the same spares ML-EFA formulation as in (Choi *et al.*, 2011), but propose different algorithm for its minimization. Their algorithm is guaranteed to converge for rather small values of $\lambda > 0$. This indicates that the method may not be able to obtain arbitrary level of sparseness. The codes realizing their algorithm are not available for additional checks of its performance.

7 Concluding remarks

The well-established practice to interpret factor loadings by setting to zero loadings smaller in magnitude than certain threshold value, makes the factor loadings matrix *sparse* artificially and subjectively. The idea behind the sparse EFA is to eliminate this subjective thresholding of the factor loadings and obtain their values in an optimal way.

We propose sparse versions of the classic ML- and LS-EFA by adding penalty term which drives some of the factor loadings to zero. To achieve this, we introduce new FA parameterization which replaces the original factor loadings Λ with two new unknowns, an othonormal Q and diagonal D (Section 3). The new model absorbs both cases of uncorrelated and correlated factors, which are otherwise considered separately by classic EFA. The benefit of this PCA-like parameterization is demonstrated in the numerical examples: the sparseness and fit of a FANC solution with *correlated* factors (Table 6) are easily achieved by the proposed method/parameterization (Table 2). The specific features of the new EFA model are additionally utilized to obtain sparse loadings.

The resulting sparse loadings are obtained in an optimal way and are easily interpretable. The level of sparseness of the loadings is effectively controlled by a tuning parameter τ. Its reduction enforces sparser loadings. The extreme value of $\tau = 1$ results in factor loadings matrix with a single non-zero entry per column/factor. The importance of the penalty term can be additionally controlled by another tuning parameter μ. However, involving two parameters (μ and τ) is unnecessary complication of the problem. In practice, μ is only used to switch between sparse and ordinary EFA.

The new EFA procedures readily produce interpretable factor loadings. Unfortunately, this is achieved on the expense of loosing some portion of the fit of the sparse EFA model (2) to the sample correlation matrix R. Further research is needed to quantify this loss, and possibly relate it to the sparseness of the factor loadings in new sparse EFA algorithms. This is a very tricky aspect of the sparse methods and requires adjustment of the involved tuning parameters controlling the balance between the model fit and the sparsity, i.e. the interpretability of the results. Here, we employed specially designed index of sparseness. Other possible options are to adapt standard information criteria, e.g. Hirose and Yamamoto (2015), or rely on cross-validation procedures, e.g. Choi *et al.* (2011). In any case, this additionally complicates the overall application of the sparse methods.

Note, that the columns of the new factor loadings QD are always orthogonal.

In fact, the orthonormality of Λ is not quite new for classic EFA. For example, the classic LS-EFA requires $\Lambda^\top\Lambda$ to be diagonal (6). The rotation methods (e.g. VARIMAX) do not produce orthonormal loadings matrix. They try to generate loadings with either large or small magnitudes. That is the main point, why sparse loadings are preferred for clear interpretation. As the small loadings are not zero, one is inevitably forced to subjectively decide which small is small enough to be dropped off consideration, i.e. be taken effectively as zero. This is the main goal of the proposed sparse EFA: to avoid such subjective choices.

A weakness of the proposed method is that it tends to produce few very small non-zero loadings. It makes sense to consider replacing LASSO by some "harder" thresholding. Our main future interest is to focus on developing sparse EFA algorithms for analyzing large data. The currently available algorithms (the one proposed here and the others by Choi *et al.* (2011); Hirose and Yamamoto (2015); Ning and Georgiou (2011)) are not very appropriate to serve this purpose for a number of reasons. In general, the available methods for sparse EFA are not satisfactory compared to the available methods for sparse PCA. We believe, that the present work will inspire interest in this new and challenging problem.

Acknowledgements

We are grateful to the Reviewers for the careful reading of the manuscript and their helpful comments. We also thank Dr Kei Hirose, Osaka University, for his help with `fanc`.

Appendix 1

Here we find the gradient of the penalty term $P_\tau(Q)^\top P_\tau(Q)$ in (11) and (12), which can be then combined with the gradients of the objective functions of ML-, LS-, or GLS-EFA. Let start with

$$d(P(Q_\tau)^\top P_\tau(Q)) = 2d(P_\tau(Q))^\top P_\tau(Q) = 2d(P_\tau)^\top P_\tau, \tag{14}$$

which requires the calculation of $d(P_\tau)$. At this point we need an approximation of $\text{sign}(x)$, and we employ the one already used in (Trendafilov and Jolliffe, 2006), which is $\text{sign}(x) \approx \tanh(\gamma x)$ for some large $\gamma > 0$, or for short $\text{th}(\gamma x)$. See also (Hage and Kleinsteuber, 2014; Luss and Teboulle, 2013). Then

$$\begin{aligned} 2(dP_\tau) &= (d\mathbf{q}_\tau) \odot [1_r + \text{th}(\gamma\mathbf{q}_\tau)] + \mathbf{q}_\tau \odot [1_r - \text{th}^2(\gamma\mathbf{q}_\tau)] \odot \gamma(d\mathbf{q}_\tau) \ , \\ &= (d\mathbf{q}_\tau) \odot \left\{1_r + \text{th}(\gamma\mathbf{q}_\tau) + \gamma\mathbf{q}_\tau \odot [1_r - \text{th}^2(\gamma\mathbf{q}_\tau)]\right\} \ , \end{aligned} \tag{15}$$

where 1_r is a $r \times 1$ vector with unit entries. The next differential to be found is:

$$\begin{aligned} dq_\tau &= 1_p^\top \left\{(dQ) \odot \text{th}(\gamma Q) + Q \odot [1_{p\times r} - \text{th}^2(\gamma Q)] \odot \gamma(dQ)\right\} \\ &= 1_p^\top \left\{(dQ) \odot \{\text{th}(\gamma Q) + (\gamma Q) \odot [1_{p\times r} - \text{th}^2(\gamma Q)\}\right\} \ , \end{aligned} \tag{16}$$

where $1_{p\times r}$ is a $p \times r$ matrix with unit entries.

Now we are ready to find the gradient ∇_Q of the penalty term with respect to Q. To simplify the notations, let

$$\mathbf{w} = 1_r + \mathrm{th}(\gamma \mathbf{q}_\tau) + (\gamma \mathbf{q}_\tau) \odot [1_r - \mathrm{th}^2(\gamma \mathbf{q}_\tau)] \ , \tag{17}$$

and

$$W = \mathrm{th}(\gamma Q) + (\gamma Q) \odot [1_{p\times r} - \mathrm{th}^2(\gamma Q)] \ . \tag{18}$$

Going back to (14) and (15), we find that:

$$\begin{aligned} 2(dP_\tau)^\top P_\tau &= \mathrm{trace}[(d\mathbf{q}_\tau) \odot \mathbf{w}]^\top P_\tau = \mathrm{trace}(d\mathbf{q}_\tau)^\top (\mathbf{w} \odot P_\tau) \\ &= \mathrm{trace}\{1_p^\top [(dQ) \odot W]\}^\top (\mathbf{w} \odot P_\tau) \\ &= \mathrm{trace}[(dQ)^\top \odot W^\top] 1_p (\mathbf{w} \odot P_\tau) \\ &= \mathrm{trace}(dQ)^\top \{W \odot [1_p (\mathbf{w} \odot P_\tau)]\} \ , \end{aligned} \tag{19}$$

making use of the identity $\mathrm{trace}(A \odot B)C = \mathrm{trace}A(B^\top \odot C)$. Thus, the gradient ∇_Q of the penalty term with respect to Q is:

$$\nabla_Q = W \odot [1_p (\mathbf{w} \odot P_\tau)] \ . \tag{20}$$

Appendix 2

Here we summarize some technical details related to the numerical solutions employed in the work.

The gradients of the ML-, LS- and GLS-EFA objective functions with respect to the unknowns $\{Q, D, \Psi\}$ are given in Trendafilov (2003) as the following block-matrix: $(-YQD^2, -Q^TYQ \odot D, -Y \odot \Psi)$. For ML-EFA, one has $Y = 2R_{ZZ}^{-1}(R - R_{ZZ})R_{ZZ}^{-1}$, and for LS- and GLS-EFA it changes to $Y = 4(R - R_{ZZ})V^2$. Additionally, we need the gradient ∇_Q of the penalty term $P_\tau(Q)^\top P_\tau(Q)$ with respect to Q, which should be added to $-YQD^2$. Its derivation is given in details in the Appendix.

The dynamical system approach employed in (Trendafilov, 2003) can be readily applied for solving (11) and (12). It involves numerical integration of matrix ordinary differential equations (ODE) for $\{Q, D, \Psi\}$ defined by their projected gradients. Particularly, it involves projected gradient dynamical system for Q on the Stiefel manifold of all $p \times r$ orthonormal matrices. There exist a number of specialized numerical methods for solving such problem listed in (Trendafilov, 2003), e.g. Del Buono and Lopez (2001) and etc. In contrast to the standard EFA alternating approaches (Jöreskog, 1977; Mulaik, 2010), the dynamical system approach gives matrix algorithms which produce *simultaneous* solution for $\{Q, D, \Psi\}$ exploiting the geometry of their specific matrix structures. Moreover, such algorithms are *globally* convergent, i.e. the convergence is reached *independently* of the starting (initial) point (Absil *et al.*, 2008; Trendafilov, 2003).

The numerical ODE solvers currently available in `MATLAB` (MATLAB, 2014) are not suitable for solving large optimization problems. They track the whole trajectory defined by the ODE which is time-consuming and undesirable when the asymptotic state is of interest only. This limits the application of the proposed approach to solving (11) and (12) for rather small data sets.

An alternative way is to employ iterative algorithms directly working on matrix manifolds (Absil *et al.*, 2008; Edelman *et al.*, 1998; Wen and Yin, 2013). The listed above gradients can be readily used for solving (11) and (12) by employing `MANOPT`, a free `MATLAB`-based software for optimization on matrix manifolds (Boumal *et al.*, 2014). The `MANOPT` code for solving (11) and (12) can be obtained from the authors upon request, and will be available online. Note that by choosing $\mu = 0$, one can obtain solutions for the standard ML-, LS- and GLS-EFA problems (9) and (10).

References

Absil, P.-A., Mahony, R., and Sepulchre, R. (2008). *Optimization Algorithms on Matrix Manifolds*. Princeton University Press.

Boumal, N., Mishra, B., Absil, P.-A., and Sepulchre, R. (2014). MANOPT: a Matlab toolbox for optimization on manifolds. *Journal of Machine Learning Research*, **15**, 1455–1459.

Choi, J., Zou, H., and Oehlert, G. (2011). A penalized maximum likelihood approach to sparse factor analysis. *Statistics and Its Interface*, **3**, 429–436.

Del Buono, N. and Lopez, L. (2001). Runge-Kutta type methods based on geodesics for systems of ODEs on the Stiefel manifold. *BIT Numerical Mathematics*, **41**(5), 912–923.

Edelman, A., Arias, T. A., and Smith, S. T. (1998). The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, **20**, 303–353.

Fontanella, S., Trendafilov, N., and Adachi, K. (2014). Sparse exploratory factor analysis. In *Proceedings of COMPSTAT 2014*, pages 281–288.

Hage, C. and Kleinsteuber, M. (2014). Robust PCA and subspace tracking from incomplete observations using ℓ_0-surrogates. *Computatinal Statistics*, **29**, 467–487.

Harman, H. H. (1976). *Modern Factor Analysis*. University of Chicago Press, Chicago, IL, 3rd edition.

Hirose, K. and Yamamoto, M. (2014). Estimation of an oblique structure via penalized likelihood factor analysis. *Computational Statistics and Data Analysis*, **79**, 120–132.

Hirose, K. and Yamamoto, M. (2015). Sparse estimation via nonconcave penalized likelihood in a factor analysis model. *Statistics and Computing*, **25**, 863–875.

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer-verlag, New York, NY, 2nd edition.

Jöreskog, K. G. (1977). Factor analysis by least-squares and maximum likelihood methods. In K. Enslein, A. Ralston, and H. S. Wilf, editors, *Mathematical methods for digital computers*, pages 125–153. John Wiley & Sons, New York, NY.

Luss, R. and Teboulle, M. (2013). Conditional gradient algorithms for rank-one matrix approximations with a sparsity constraint. *SIAM Review*, **55**, 65–98.

MATLAB (2014). *MATLAB R2014b*. The MathWorks, Inc, New York, NY.

Mulaik, S. A. (2010). *The Foundations of Factor Analysis*. Chapman and Hall/CRC, Boca Raton, FL, 2nd edition.

Ning, N. and Georgiou, T. T. (2011). Sparse factor analysis via likelihood and ℓ_1-regularization. 50th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC) Orlando, FL, USA, December 12-15, 2011.

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

Trendafilov, N. T. (2014). From simple structure to sparse components: a review. *Computational Statistics*, **29**, 431–454.

Trendafilov, N. T. and Adachi, K. (2015). Sparse versus simple structure loadings. *Psychometrika*, **80**, 776–790.

Trendafilov, N. T. and Jolliffe, I. T. (2006). Projected gradient approach to the numerical solution of the SCoTLASS. *Computational Statistics and Data Analysis*, **50**, 242–253.

Wen, Z. and Yin, W. (2013). A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, **142**, 397–434.