Sparse exploratory factor analysis

Sara Fontanella, *Open University*, Sara.Fontanella@open.ac.uk
Nickolay Trendafilov, *Open University*, Nickolay.Trendafilov@open.ac.uk
Kohei Adachi, *Osaka University*, adachi@hus.osaka-u.ac.jp

Abstract. Sparse principal component analysis is a very active research area in the last decade. In the same time, there are very few works on sparse factor analysis. We propose a new contribution to the area by exploring a procedure for sparse factor analysis where the unknown parameters are found simultaneously.

Keywords. ℓ_1 penalties, Matrix manifolds, Projected gradients.

1 Introduction

Exploratory factor analysis (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. 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 are introduced, each of which accounts for this portion of variance of the corresponding manifest variable [12]. 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

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

where Λ is a $p \times r$ parameter matrix of *factor loadings*. The choice of r is either subjective or based on preliminary validation. In both case its value is subject to some limitations [12]. 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, they are also assumed *mutually* uncorrelated, and the $p \times p$ matrix Ψ is assumed diagonal with *non-zero* diagonal entries. 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 [12]. Then, finding $\{\Lambda, \Psi\}$ can be formulated as minimizing the following negative loglikelihood function [9, 12]:

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

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 [9] 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 $V = I_p$ 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 minimization problems ML, LS and GLS listed above are not *unconstrained*. The unknowns Λ and Ψ are sought subject to the following constraints [9]: 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 [12], make the EFA model highly controversial, which probably explains why EFA is far less popular dimension reduction technique than principal components (PCA).

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

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

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 [12] 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.[3, 13]. The proposed work will be a further contribution to this new research area.

2 New EFA parameters

It has been argued in [15], 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. [9, 12]. 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 [15]. 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 [15] do not relay on $\Psi^2 > 0$. Moreover, maintaining $\Psi^2 > 0$ may contradict to achieving high level of sparseness (Section 5).

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:

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

Thus, instead of the pair $\{\Lambda, \Psi\}$, a triple $\{Q, D, \Psi\}$ is sought in [15]. Note, that the model (8) does not permit rotations, only permutations are possible. Thus, the new factor loadings Λ are given by QD. Clearly, when Q is sparse, Λ will have the same sparseness. In order to maintain the factor analysis constraints, the triple $\{Q, D, \Psi\}$ should be sought such that Q be an $p \times r$ 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:

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

3 Sparse factor loadings

Let q_i denote the ith column of Q, i.e. $Q = (q_1, q_2, ..., q_r)$, and $\tau = (\tau_1, \tau_2, ..., \tau_r)$ 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. $\|q_i\|_1 \leq \tau_i$ for all $i = 1, 2, ..., r$. Introduce the following discrepancy vector $q_\tau = (\|q_1\|_1, \|q_2\|_1, ..., \|q_r\|_1) - \tau$, which can also be expressed as $q_\tau = 1_p^\top[Q \odot \mathrm{sign}(Q)] - \tau$, where $\mathrm{sign}(Q)$ 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 [16] to introduce the following vector penalty function $P_\tau(Q) = [q_\tau \odot (1_p + \text{sign}(q_\tau)]/2$. Then, the penalized versions of (9) and (10) can be defined, for the ML-EFA as:

$$\min_{Q,D,\Psi} \log(\det(QD^2Q^T + \Psi^2)) + \text{trace}((QD^2Q^T + \Psi^2)^{-1}R) + P_\tau(Q)^\top P_\tau(Q) \ , \tag{11}$$

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

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

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

4 Gradients and Stiefel gradients

The gradients of the ML-, LS- and GLS-EFA objective functions with respect to the unknowns $\{Q, D, \Psi\}$ are given in [15] 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 $Y = 4V(R - R_{ZZ})V$. Now we need to find the gradient ∇_Q of the penalty term $P_\tau(Q)^\top P_\tau(Q)$ with respect to Q, which should be added to $-YQD^2$.

Making use of the identity $\text{trace}(A \odot B)C = \text{trace}A(B^\top \odot C)$, we find that:

$$\nabla_Q = \frac{1}{2} W \odot [1_p(w \odot P_\tau)] \ , \tag{13}$$

where 1_p is a $p \times 1$ vector and $1_{p\times r}$ is a $p \times r$ matrix with unit entries, and

$$w = 1_p + \text{th}(\gamma q_\tau) + (\gamma q_\tau) \odot [1_p - \text{th}^2(\gamma q_\tau)] \ , \tag{14}$$

and

$$W = \text{th}(\gamma Q) + (\gamma Q) \odot [1_{p\times r} - \text{th}^2(\gamma Q)] \ . \tag{15}$$

The dynamical system approach employed in [15] 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, e.g. [4] and others listed in [15]. In contrast to the standard EFA alternating approaches [9, 12], 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.

The numerical ODE solvers currently available in `MATLAB` [11] 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 [1, 5, 17]. 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 [2].

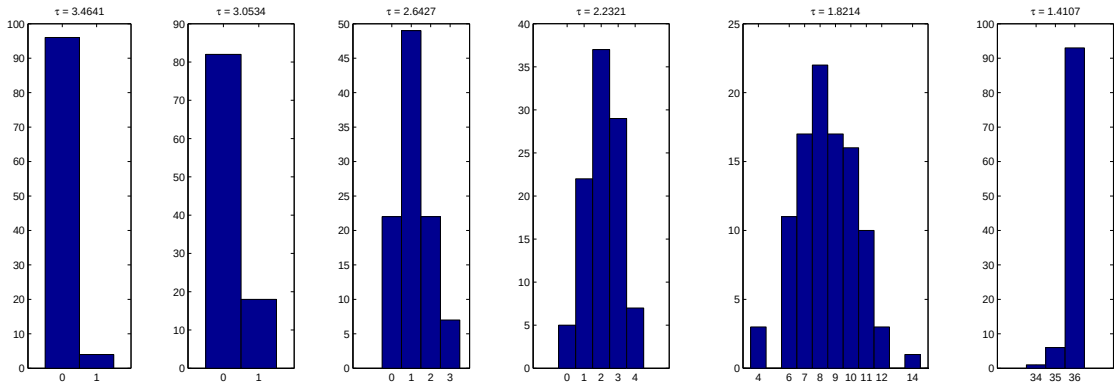

Figure 1. Number of zeros obtained in 100 runs of sparse ML-EFA (11) for different τ.

.

5 Numerical examples

In this Section we first explore the behavior of the proposed sparse EFA on simulated data considered in [3]. Then, in contrast to [3, 13], we consider two examples from the classic EFA.

Simulated data [3]

We examine the performance of the proposed approach by employing the simulated data constructed in [3]. They take a hypothetical 12×4 sparse loadings matrix Λ with the following nonzero 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 $\Psi^2 = \text{Diag}(1.27,\ .61,\ .74,\ .88,\ .65,\ .81,\ .74,\ 1.3,\ 1.35,\ .74,\ .92,\ 1.32)$. 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 of $\tau (= \sqrt{12},\ 3.0534,\ 2.6427,\ 2.2321,\ 1.8214,\ 1.4107)$. 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 number of the zero loadings among all $12 \times 4 = 48$ for each τ are depicted 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.

Harman's Five Socio-Economic Variables [8, 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 [8, p.14]. This small data set is interesting because the two- and the three-factor solutions from LS- and ML-EFA are 'Heywood cases' [8, 12], 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. 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, [8, p.14].

Holzinger-Harman's Twenty-Four Psychological Tests [8, p.123]

Finally, we illustrate the proposed procedures for sparse EFA on another well known data set from classic EFA, namely the Holzinger-Harman' Twenty-Four Psychological Tests [8, p.123]. It is widely used to illustrate different aspects of classic EFA [8, 12].

The correlation matrix [8, 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 [8, p.215]. Then, we rotate (with normalization) the factor loadings QD from the first four columns by VARIMAX from `MATLAB` [11], and the result is given in the next four columns of Table 2. The loadings in bold correspond to non-zero loadings of the sparse ML-EFA solution of (11) obtained with $\tau = 2.2997$ and depicted in the last columns of Table 2. Further decrease of τ results in sparser loadings, but regarded as too simplified. Note, that to interpret the VARIMAX solution, one must subjectively ignore the loadings with small absolute values. The sparse factor loadings are easily interpreted only by focusing on the nonzero loadings.

VARS	$\tau = \sqrt{24} = 4.899$					Varimax				$\tau = 2.2997$				
	QD				Ψ^2	Rotated QD				QD				Ψ^2
1	.60	.39	-.22	.02	.44	**.69**	.16	.16	.19	.88				.32
2	.37	.25	-.13	-.03	.78	**.44**	.12	.10	.08	.28				.85
3	.41	.39	-.14	-.12	.64	**.57**	.14	.11	-.02	.54				.70
4	.49	.25	-.19	-.10	.65	**.53**	.23	.08	.10	.55				.69
5	.69	-.28	-.03	-.30	.35	.19	**.74**	.15	.21		.82			.35
6	.69	-.20	.08	-.41	.31	.20	**.77**	.23	.07		.84			.32
7	.68	-.29	-.08	-.41	.28	.20	**.81**	.07	.15		.86			.29
8	.67	-.10	-.12	-.19	.49	.34	**.57**	.13	.24		.64			.54
9	.70	-.21	.08	-.45	.26	.20	**.81**	.23	.04		.87			.27
10	.48	-.49	-.09	.54	.24	**-.12**	.17	.17	**.83**	-.18			.91	.28
11	.56	-.14	.09	.33	.55	.12	.18	.37	**.51**				.63	.59
12	.47	-.14	-.26	.51	.44	.21	.02	.09	**.72**				.72	.50
13	.60	.03	-.30	.24	.49	**.44**	.19	.08	**.53**	.30			.47	.51
14	.42	.02	.41	.06	.65	.05	.20	**.55**	.08			-.47		.74
15	.39	.10	.36	.09	.70	.12	.12	**.52**	.07			-.53		.70
16	.51	.35	.25	.09	.55	.41	.07	**.53**	.06			-.57		.68
17	.47	-.00	.38	.20	.60	.06	.14	**.57**	.22			-.72		.54
18	.52	.15	.15	.31	.59	.29	.03	**.46**	.34			-.65		.61
19	.44	.11	.15	.09	.76	.24	.15	**.37**	.16			-.35		.82
20	.61	.12	.04	-.12	.59	.40	**.38**	.30	.12		.34			.76
21	.59	.06	-.12	.23	.58	.38	.17	.22	**.44**				.51	.69
22	.61	.13	.04	-.11	.60	.40	**.37**	.30	.12		.30			.79
23	.69	.14	-.10	-.04	.50	**.50**	.37	.24	**.24**	.58			.04	.63
24	.65	-.21	.02	.18	.50	.16	.37	.30	**.50**				.63	.58

Table 2. ML-EFA solutions for Twenty-Four Psychological Tests [8, p.123].

6 Conclusion

We propose a new method to construct sparse factor loadings for the classic EFA. This is, in fact, a new approach to EFA, which readily produces interpretable EFA results. Unfortunately, this can be 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.

There are few methods for sparse PCA, e.g. [6, 14, 16], able to produce either orthonormal component loadings or uncorrelated components. In contrast to PCA, the factor loadings $\Lambda(= QD)$, both original and sparse, are not orthonormal. However, how the sparse factor loadings affect the correlations among the estimated factors remains to be studied.

Acknowledgement

This work is supported by a grant RPG-2013-211 from The Leverhulme Trust, UK.

Bibliography

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

[2] Boumal, N., Mishra, B., Absil, P.-A. and Sepulchre, R., (2014) Manopt: a Matlab toolbox for optimization on manifolds, *The Journal of Machine Learning Research*, to appear.

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

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

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

[6] Farcomeni, A. (2009) An exact approach to sparse principal component analysis, *Computational Statistics*, 24, 583–604.

[7] Hage, C., and Kleinsteuber, M. (2014) Robust PCA and subspace tracking from incomplete observations using ℓ_0-surrogates, *Computational Statistics*, to appear.

[8] Harman, H. H. (1976) *Modern Factor Analysis*, 3rd Ed., Chicago: University of Chicago Press.

[9] Jöreskog, K. G. (1977) Factor analysis by least-squares and maximum likelihood methods, in *Mathematical Methods for Digital Computers*, (K. Enslein, A. Ralston and H.S. Wilf, Eds.) Vol. 3, pp. 125–153, New York: Wiley.

[10] Luss, R., and Teboulle, M. (2012) Conditional gradient algorithms for rank-one matrix approximations with a sparsity constraint, `http://arxiv.org/pdf/1107.1163.pdf`.

[11] MATLAB (2011) *MATLAB R2011a*, The MathWorks, Inc., New York.

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

[13] Ning, L., and Georgiou, 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, 5188–5192.

[14] Qi, X., Luo, R., Zhao, H. (2013) Sparse principal component analysis by choice of norm, *Journal of Multivariate Analysis*, 114, 127–160.

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

[16] 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.

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