Estimation of oblique structure via penalized likelihood factor analysis

Kei Hirose and Michio Yamamoto

Division of Mathematical Science, Graduate School of Engineering Science, Osaka University,

1-3, Machikaneyama-cho, Toyonaka, Osaka, 560-8531, Japan

E-mail: hirose@sigmath.es.osaka-u.ac.jp, myamamoto@sigmath.es.osaka-u.ac.jp.

Abstract

We consider the problem of sparse estimation via a lasso-type penalized likelihood procedure in a factor analysis model. Typically, the model estimation is done under the assumption that the common factors are orthogonal (uncorrelated). However, the lasso-type penalization method based on the orthogonal model can often estimate a completely different model from that with the true factor structure when the common factors are correlated. In order to overcome this problem, we propose to incorporate a factor correlation into the model, and estimate the factor correlation along with parameters included in the orthogonal model by maximum penalized likelihood procedure. An entire solution path is computed by the EM algorithm with coordinate descent, which permits the application to a wide variety of convex and nonconvex penalties. The proposed method can provide sufficiently sparse solutions, and be applied to the data where the number of variables is larger than the number of observations. Monte Carlo simulations are conducted to investigate the effectiveness of our modeling strategies. The results show that the lasso-type penalization based on the orthogonal model cannot often approximate the true factor structure, whereas our approach performs well in various situations. The usefulness of the proposed procedure is also illustrated through the analysis of real data.

Key Words: Nonconvex penalty, Oblique structure, Rotation technique, Penalized likelihood factor analysis

1   Introduction

Factor analysis provides a practical tool for exploring the covariance structure among a set of observed random variables by construction of a smaller number of random variables called common factors. In exploratory factor analysis, a traditional estimation procedure in use is the following two-step approach: the model is estimated by the maximum likelihood method under the assumption that the common factors are uncorrelated (orthogonal), and then rotation techniques, such as the varimax method (Kaiser 1958)

and the promax method (Hendrickson and White 1964), are utilized to find sparse factor loadings. However, it is well known that the maximum likelihood method often yields unstable estimates because of overparametrization (e.g., Akaike 1987). In particular, the commonly-used algorithms for maximum likelihood factor analysis (e.g., Jöreskog 1967; Jennrich and Robinson 1969; Clarke 1970; Lawley and Maxwell 1971) cannot often be applied when the number of variables is larger than the number of observations. Furthermore, the rotation techniques cannot often produce a sufficiently sparse solution. In order to overcome these difficulties, we apply a penalized likelihood procedure that produces the sparse solutions, such as the lasso (Tibshirani 1996).

The lasso-type penalized likelihood factor analysis has been recently studied by several researchers. Ning and Georgiou (2011) and Choi et al. (2011) applied the weighted lasso to obtain sparse factor loadings, and numerically demonstrated that the penalization method often outperformed the rotation technique with maximum likelihood procedure. Hirose and Yamamoto (2012) showed that the penalization method is a generalization of the rotation technique with maximum likelihood method, and applied the nonconvex penalties such as minimax concave penalty (MC+, Zhang 2010) and smoothly clipped absolute deviation (SCAD, Fan and Li 2001) to achieve sparser solutions than the lasso.

In these studies, the common factors are assumed to be uncorrelated (orthogonal) as is the case with the maximum likelihood exploratory factor analysis. In some cases, however, analysts may prefer to relax the requirement that the common factors are orthogonal (e.g., Mulaik 1972). Moreover, we found that the lasso-type penalization technique based on the orthogonal model can often estimate a completely different model from that with the true factor structure when the common factors are correlated (oblique). Empirically, the estimated factor loadings in the first column often become dense (i.e., all elements are non-zero), even if the first column of true loading matrix is sparse.

In order to handle this fundamental problem, we propose to incorporate a factor correlation into the model, and estimate the factor correlation along with parameters included in the orthogonal model by maximum penalized likelihood procedure. A pathwise algorithm via the EM algorithm (Rubin and Thayer 1982) with coordinate descent for nonconvex penalties (Mazumder et al. 2011) is introduced according to the basic idea given by Hirose and Yamamoto (2012). Our algorithm produces the entire solution path for a wide variety of convex and nonconvex penalties including the lasso, SCAD, and MC+ family. Furthermore, the proposed methodology can provide sparser solutions than the rotation technique with maximum likelihood method, and be applied to the data where the number of variables is larger than the number of observations.

The remainder of this paper is organized as follows: Section 2 shows that the lasso-type penalized likelihood factor analysis based on the orthogonal model cannot often approximate the oblique structure. In Section 3, we introduce a penalized factor analysis

via the oblique model, and provide a computational algorithm based on the EM algorithm and coordinate descent to obtain the entire solution path. Section 4 presents numerical results for both artificial and real datasets. Some concluding remarks are given in Section 5.

2 Penalized likelihood factor analysis based on the orthogonal model may not approximate the oblique structure

2.1 Model and Estimation

We briefly describe a lasso-type penalized likelihood factor analysis based on the orthogonal model (Choi et al. 2011; Ning and Georgiou 2011; Hirose and Yamamoto 2012). Suppose that $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ is a p-dimensional observable random vector with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$. The factor analysis model (e.g., Mulaik 1972) is

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{F} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\Lambda} = (\lambda_{ij})$ is a $p \times m$ matrix of factor loadings, and $\boldsymbol{F} = (F_1, \cdots, F_m)^T$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_p)^T$ are unobservable random vectors. The elements of $\boldsymbol{F}$ and $\boldsymbol{\varepsilon}$ are called common factors and unique factors, respectively. It is assumed that the common factors $\boldsymbol{F}$ and the unique factors $\boldsymbol{\varepsilon}$ are multivariate-normally distributed with $E(\boldsymbol{F}) = \boldsymbol{0}$, $E(\boldsymbol{\varepsilon}) = \boldsymbol{0}$, $E(\boldsymbol{F}\boldsymbol{F}^T) = \mathbf{I}_m$, $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) = \boldsymbol{\Psi}$, and are independent (i.e., $E(\boldsymbol{F}\boldsymbol{\varepsilon}^T) = \mathbf{O}$). Here $\mathbf{I}_m$ is the $m \times m$ identity matrix, and $\boldsymbol{\Psi}$ is a $p \times p$ diagonal matrix with the i-th diagonal element ψ_i, which is called unique variance. Under these assumptions, the observable random vector $\boldsymbol{X}$ is multivariate-normally distributed with variance-covariance matrix $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi}$.

Let $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N$ be a random sample of N observations from the p-dimensional normal population $N_p(\boldsymbol{\mu}, \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi})$. The estimates of factor loadings and unique variances, say, $\hat{\boldsymbol{\Lambda}}_{\text{ort}}$ and $\hat{\boldsymbol{\Psi}}_{\text{ort}}$ ("ort" is an abbreviation for orthogonal), are obtained by maximizing the penalized log-likelihood function

$$(\hat{\boldsymbol{\Lambda}}_{\text{ort}}, \hat{\boldsymbol{\Psi}}_{\text{ort}}) = \arg\max_{\boldsymbol{\Lambda}, \boldsymbol{\Psi}} \ell_\rho^{\text{ort}}(\boldsymbol{\Lambda}, \boldsymbol{\Psi}),$$

where $\ell_\rho^{\text{ort}}(\boldsymbol{\Lambda}, \boldsymbol{\Psi})$ is the penalized log-likelihood function

$$\ell_\rho^{\text{ort}}(\boldsymbol{\Lambda}, \boldsymbol{\Psi}) = \ell^{\text{ort}}(\boldsymbol{\Lambda}, \boldsymbol{\Psi}) - N \sum_{i=1}^{p} \sum_{j=1}^{m} \rho P(|\lambda_{ij}|).$$

Here $\ell^{\text{ort}}(\boldsymbol{\Lambda}, \boldsymbol{\Psi})$ is the log-likelihood function

$$\ell^{\text{ort}}(\boldsymbol{\Lambda}, \boldsymbol{\Psi}) = -\frac{N}{2} \left[p \log(2\pi) + \log|\boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi}| + \text{tr}\{(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi})^{-1}\mathbf{S}\} \right],$$

$P(\cdot)$ is a penalty function, and ρ is a regularization parameter. The matrix $\mathbf{S} = (s_{ij})$ is the sample variance-covariance matrix.

The lasso-type penalty function $P(\cdot)$ produces sparse solutions for some ρ, i.e., some of the factor loadings can be estimated by exactly zero. The lasso is continuous and fast, but biased and then estimates an overly dense model (Zou 2006; Zhao and Yu 2007; Zhang 2010). Typically, a nonconcave penalization procedure such as MC+ (Zhang 2010) and SCAD (Fan and Li 2001) can achieve sparser models than the lasso. For example, the MC+ (Zhang 2010) is given by

$$\rho P(|\theta|; \rho; \gamma) = \rho \int_0^{|\theta|} \left(1 - \frac{x}{\rho\gamma}\right)_+ dx$$
$$= \rho \left(|\theta| - \frac{\theta^2}{2\rho\gamma}\right) I(|\theta| < \rho\gamma) + \frac{\rho^2\gamma}{2} I(|\theta| \geq \rho\gamma).$$

For each value of $\rho > 0$, $\gamma \to \infty$ yields soft threshold operator (i.e., lasso penalty) and $\gamma \to 1+$ produces hard threshold operator.

2.2   Problem of the lasso via orthogonal model

The lasso-type penalization based on the orthogonal model can perform well when the true common factors are uncorrelated. In practical situations, however, the true common factors may often be correlated: $E[\boldsymbol{F}\boldsymbol{F}^T] = \boldsymbol{\Phi}$ with $\boldsymbol{\Phi}$ being the factor correlation. In this case, the covariance matrix of the observed variables $\boldsymbol{X}$ is expressed as $\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi}$. If each factor is highly correlated to each other, i.e., the absolute values non-diagonal elements of $\boldsymbol{\Phi}$ are large, the lasso based on the orthogonal model can often estimate a completely different model from that with the true factor structure. A typical example of this phenomena is given as follows:

Example 2.1. *Suppose that the true factor loadings, unique variances and factor correlation are given by*

$$\boldsymbol{\Lambda} = \begin{pmatrix} 0.9 & 0.9 & 0.9 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.9 & 0.9 & 0.9 \end{pmatrix}^T, \quad \boldsymbol{\Psi} = 0.19\mathbf{I}_6, \quad \boldsymbol{\Phi} = \begin{pmatrix} 1.0 & 0.6 \\ 0.6 & 1.0 \end{pmatrix}.$$

We generated 50 observations from $\boldsymbol{X} \sim N_6(\mathbf{0}, \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi})$. The model was estimated by the penalized likelihood method with $P(\theta) = \theta$ (i.e., the lasso) and $\rho = 0.01$. The estimated factor landings were

$$\hat{\boldsymbol{\Lambda}}_{\mathrm{ort}} = \begin{pmatrix} 0.87 & 0.83 & 0.89 & 0.59 & 0.53 & 0.54 \\ 0.00 & 0.05 & -0.02 & 0.71 & 0.62 & 0.62 \end{pmatrix}^T. \tag{1}$$

Because the lasso tends to produce some of the loadings being zero, $\hat{\lambda}_{12}$, $\hat{\lambda}_{22}$ and $\hat{\lambda}_{32}$ were close to or exactly zero. However, $\hat{\lambda}_{41}$, $\hat{\lambda}_{51}$ and $\hat{\lambda}_{61}$ were far from zero although the true

parameters are zero. The lasso based on the orthogonal model was not able to approximate the true factor structure.

The problem that the lasso based on orthogonal model cannot approximate the true factor structure is closely related to the rotation problem. In orthogonal model, the true covariance matrix $\mathbf{\Lambda\Phi\Lambda}^T + \mathbf{\Psi}$ is estimated by $\hat{\mathbf{\Lambda}}_{\text{ort}}\hat{\mathbf{\Lambda}}_{\text{ort}}^T + \hat{\mathbf{\Psi}}_{\text{ort}}$, which means $\hat{\mathbf{\Lambda}}_{\text{ort}}$ approximates $\mathbf{\Lambda G}$, where $\mathbf{G} = (g_{ii'})$ is an $m \times m$ matrix that satisfies $\mathbf{GG}^T = \mathbf{\Phi}$. The matrix $\mathbf{G}$ is not an identity matrix unless the factor correlation $\mathbf{\Phi}$ is an identity matrix, so that $\hat{\mathbf{\Lambda}}_{\text{ort}}$ is not always close to $\mathbf{\Lambda}$ even if $\hat{\mathbf{\Lambda}}_{\text{ort}} \approx \mathbf{\Lambda G}$.

Furthermore, the matrix $\mathbf{G}$ can have a rotational indeterminacy, since $\mathbf{GG}^T = \mathbf{G}^*(\mathbf{G}^*)^T = \mathbf{\Phi}$, where $\mathbf{G}^* = \mathbf{GT}$ with $\mathbf{T}$ being an arbitrary orthogonal matrix. The rotational indeterminacy of $\mathbf{G}$ leads to $\ell(\mathbf{\Lambda G}, \mathbf{\Psi}) = \ell(\mathbf{\Lambda G}^*, \mathbf{\Psi})$. Thus, if $\hat{\mathbf{\Lambda}}_{\text{ort}}$ and $\hat{\mathbf{\Psi}}_{\text{ort}}$ are expressed as $\hat{\mathbf{\Lambda}}_{\text{ort}} = \mathbf{\Lambda G}$ and $\hat{\mathbf{\Psi}}_{\text{ort}} = \mathbf{\Psi}$, the matrix $\mathbf{G}$ is obtained by solving the following problem:

$$\max_{\mathbf{G}} \ell_\rho^{\text{ort}}(\mathbf{\Lambda G}, \mathbf{\Psi}) \quad \text{s.t.,} \quad \mathbf{GG}^T = \mathbf{\Phi},$$

which is equivalent to

$$\min_{\mathbf{G}} \sum_{i=1}^{p} \sum_{j=1}^{m} P(|\breve{\lambda}_{ij}|) \quad \text{s.t.,} \quad \mathbf{GG}^T = \mathbf{\Phi}, \tag{2}$$

where $\breve{\lambda}_{ij}$ is the (i, j)-th element of $\mathbf{\Lambda G}$.

How is the matrix $\mathbf{G}$ estimated? To explain this, we assume that the true factor loadings $\mathbf{\Lambda}$ possess perfect simple structure, that is, each row has at most one nonzero element. The problem in (2) is then written as

$$\min_{\mathbf{G}} \sum_{i=1}^{m} \sum_{i'=1}^{m} P(w_{ii'}|g_{ii'}|) \quad \text{s.t.,} \quad \mathbf{GG}^T = \mathbf{\Phi}, \tag{3}$$

where $w_{ii'}$ are positive values. Because the objective function is based on the L_1 loss, some of the elements of $\mathbf{G}$ become exactly zero. Empirically, one of the elements of $\mathbf{G}$ often becomes 1. When $g_{qr} = 1$, we have $g_{qi'} = 0$ $(i' \neq r)$ and $g_{ir} = \phi_{ir}$ $(i \neq q)$, so that all elements of r-th column of $\mathbf{\Lambda G}$ become non-zero unless $\phi_{ir} = 0$, which does not approximate the perfect simple structure. In this way, the lasso-type penalization via orthogonal model can often estimate a completely different model from that with the true factor structure when the common factors are highly correlated.

Example 2.2. *In Example 2.1, the problem in (3) is written as*

$$\min_{\mathbf{G}}(|g_{11}| + |g_{12}| + |g_{21}| + |g_{22}|) \quad \text{s.t.,} \quad \mathbf{GG}^T = \begin{pmatrix} 1.0 & 0.6 \\ 0.6 & 1.0 \end{pmatrix}.$$

The solution of $\mathbf{G}$ is given by

$$\mathbf{G} = \begin{pmatrix} 1.0 & 0.0 \\ 0.6 & 0.8 \end{pmatrix}.$$

In this case, we have

$$\mathbf{\Lambda G} = \begin{pmatrix} 0.90 & 0.90 & 0.90 & 0.54 & 0.54 & 0.54 \\ 0.00 & 0.00 & 0.00 & 0.72 & 0.72 & 0.72 \end{pmatrix}^T,$$

which is quite similar to the maximum penalized likelihood estimates of factor loadings based on orthogonal model in (1).

3 Estimation of oblique structure via penalized likelihood factor analysis

The lasso-type penalized likelihood factor analysis based on the orthogonal model cannot often approximate the oblique structure as described in the Section 2.2. In this Section, we propose to incorporate a factor correlation into the model, and estimate the oblique model by maximum penalized likelihood procedure.

3.1 Model Estimation

Let $\ell(\mathbf{\Lambda}, \mathbf{\Psi}, \mathbf{\Phi})$ be the log-likelihood function based on the oblique model

$$\ell(\mathbf{\Lambda}, \mathbf{\Psi}, \mathbf{\Phi}) = -\frac{N}{2} \left[p \log(2\pi) + \log|\mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}^T + \mathbf{\Psi}| + \text{tr}\{(\mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}^T + \mathbf{\Psi})^{-1}\mathbf{S}\} \right], \qquad (4)$$

and $\ell_\rho(\mathbf{\Lambda}, \mathbf{\Psi}, \mathbf{\Phi})$ be the penalized log-likelihood function

$$\ell_\rho(\mathbf{\Lambda}, \mathbf{\Psi}, \mathbf{\Phi}) = \ell(\mathbf{\Lambda}, \mathbf{\Psi}, \mathbf{\Phi}) - N \sum_{i=1}^{p} \sum_{j=1}^{m} \rho P(|\lambda_{ij}|). \qquad (5)$$

We estimate the factor loadings, unique variance, and factor correlation, say, $\hat{\mathbf{\Lambda}}_{\text{obl}}$, $\hat{\mathbf{\Psi}}_{\text{obl}}$, and $\hat{\mathbf{\Phi}}_{\text{obl}}$ ("obl" is an abbreviation for oblique), by maximum penalized likelihood procedure simultaneously:

$$(\hat{\mathbf{\Lambda}}_{\text{obl}}, \hat{\mathbf{\Psi}}_{\text{obl}}, \hat{\mathbf{\Phi}}_{\text{obl}}) = \arg \max_{\mathbf{\Lambda}, \mathbf{\Psi}, \mathbf{\Phi}} \ell_\rho(\mathbf{\Lambda}, \mathbf{\Psi}, \mathbf{\Phi}).$$

Example 3.1. *The lasso based on oblique model was applied to the dataset used in the Example 2.1. When $\rho = 0.01$, the estimates of factor loadings were*

$$\hat{\mathbf{\Lambda}}_{\text{obl}} = \begin{pmatrix} 0.85 & 0.78 & 0.89 & 0.00 & 0.00 & 0.02 \\ 0.03 & 0.09 & 0.00 & 0.93 & 0.82 & 0.81 \end{pmatrix}^T,$$

which closely approximate the true factor loadings compared with orthogonal factor loadings $\hat{\mathbf{\Lambda}}_{\text{ort}}$ in (1).

3.2 Algorithm

It is well known that the solutions estimated by the lasso-type regularization methods are not usually expressed in a closed form mainly because the penalty term includes a non-differentiable function. In regression analysis, a number of researchers have proposed fast algorithms to obtain the entire solutions (e.g., Least angle regression, Efron et al. 2004; Coordinate descent algorithm, Friedman et al. 2007; Generalized path seeking, Friedman 2012). In particular, the coordinate descent algorithm is known as a remarkably fast algorithm (Friedman et al. 2010) and can also be applied to a wide variety of convex and nonconvex penalties (Breheny and Huang 2011; Mazumder et al. 2011). Thus, we employ the coordinate descent algorithm to obtain the entire solution.

In the coordinate descent algorithm, each step is fast if an explicit formula for each coordinate-wise maximization is given, whereas the log-likelihood function in (4) may not lead to the explicit formula. In order to derive the explicit formula, we apply the EM algorithm (Rubin and Thayer 1982) to the penalized likelihood factor analysis. The coordinate descent algorithm is utilized to maximize the nonconcave function in the maximization step of the EM algorithm. Because the complete-data log-likelihood function takes the quadratic form, the explicit formula for each coordinate-wise maximization is available.

3.2.1 Update Equation for Fixed Regularization Parameter

First, we provide the update equations of factor loadings, unique variances, and factor correlation when ρ and γ are fixed. Suppose that $\boldsymbol{\Lambda}_{\text{old}}$, $\boldsymbol{\Psi}_{\text{old}}$ and $\boldsymbol{\Phi}_{\text{old}}$ are the current values of factor loadings, unique variances, and factor correlation. The model can be estimated by maximizing the expectation of the complete-data penalized log-likelihood function with respect to $\boldsymbol{\Lambda}$, $\boldsymbol{\Psi}$ and $\boldsymbol{\Phi}$:

$$
\begin{aligned}
E[l_\rho^C(\boldsymbol{\Lambda}, \boldsymbol{\Psi}, \boldsymbol{\Phi})] = & -\frac{N}{2} \sum_{i=1}^p \log \psi_i - \frac{N}{2} \sum_{i=1}^p \frac{s_{ii} - 2\boldsymbol{\lambda}_i^T \mathbf{b}_i + \boldsymbol{\lambda}_i^T \mathbf{A} \boldsymbol{\lambda}_i}{\psi_i} \\
& - \frac{N}{2} \log |\boldsymbol{\Phi}| - \frac{N}{2} \text{tr}(\boldsymbol{\Phi}^{-1} \mathbf{A}) - N \sum_{i=1}^p \sum_{j=1}^m \rho P(|\lambda_{ij}|) + \text{const.},
\end{aligned}
$$

(6)

where $\mathbf{b}_i = \mathbf{M}^{-1} \boldsymbol{\Lambda}_{\text{old}}^T \boldsymbol{\Psi}_{\text{old}}^{-1} \mathbf{s}_i$ and $\mathbf{A} = \mathbf{M}^{-1} + \mathbf{M}^{-1} \boldsymbol{\Lambda}_{\text{old}}^T \boldsymbol{\Psi}_{\text{old}}^{-1} \mathbf{S} \boldsymbol{\Psi}_{\text{old}}^{-1} \boldsymbol{\Lambda}_{\text{old}} \mathbf{M}^{-1}$. Here $\mathbf{M} = \boldsymbol{\Lambda}_{\text{old}}^T \boldsymbol{\Psi}_{\text{old}}^{-1} \boldsymbol{\Lambda}_{\text{old}} + \boldsymbol{\Phi}_{\text{old}}^{-1}$, and $\mathbf{s}_i$ is the i-th column vector of $\mathbf{S}$. The derivation of the complete-data penalized log-likelihood function is described in Appendix.

The new parameter $(\boldsymbol{\Lambda}_{\text{new}}, \boldsymbol{\Psi}_{\text{new}}, \boldsymbol{\Phi}_{\text{new}})$ can be computed by maximizing the complete-data penalized log-likelihood function, i.e.,

$$
(\boldsymbol{\Lambda}_{\text{new}}, \boldsymbol{\Psi}_{\text{new}}, \boldsymbol{\Phi}_{\text{new}}) = \arg \max_{\boldsymbol{\Lambda}, \boldsymbol{\Psi}, \boldsymbol{\Phi}} E[l_\rho^C(\boldsymbol{\Lambda}, \boldsymbol{\Psi}, \boldsymbol{\Phi})].
$$

(7)

7

The solution in (7) is not usually expressed in a closed form because the penalty term includes a nondifferentiable function, so that the coordinate descent algorithm is utilized.

Let $\tilde{\boldsymbol{\lambda}}_i^{(j)}$ be an $(m-1)$-dimensional vector $(\tilde{\lambda}_{i1}, \tilde{\lambda}_{i2}, \ldots, \tilde{\lambda}_{i(j-1)}, \tilde{\lambda}_{i(j+1)}, \ldots, \tilde{\lambda}_{im})^T$. The parameter λ_{ij} can be updated by maximizing (6) with the other parameters $\tilde{\boldsymbol{\lambda}}_i^{(j)}$, $\boldsymbol{\Psi}$ and $\boldsymbol{\Phi}$ being fixed, i.e., we solve the following problem:

$$
\begin{aligned}
\tilde{\lambda}_{ij} &= \arg\min_{\lambda_{ij}} \frac{1}{2\psi_i} \left\{ a_{jj}\lambda_{ij}^2 - 2\left(b_{ij} - \sum_{k\neq j} a_{kj}\tilde{\lambda}_{ik}\right)\lambda_{ij} \right\} + \rho P(|\lambda_{ij}|) \\
&= \arg\min_{\lambda_{ij}} \frac{1}{2}\left(\lambda_{ij} - \frac{b_{ij} - \sum_{k\neq j} a_{kj}\tilde{\lambda}_{ik}}{a_{jj}}\right)^2 + \frac{\psi_i \rho}{a_{jj}} P(|\lambda_{ij}|).
\end{aligned}
$$

This is equivalent to minimizing the following penalized squared-error loss function

$$
S(\tilde{\theta}) = \arg\min_{\theta} \left\{ \frac{1}{2}(\theta - \tilde{\theta})^2 + \rho^* P(|\theta|) \right\}.
$$

The solution $S(\tilde{\theta})$ can be expressed in a closed form for a variety of convex and nonconvex penalties. For example, the update equation for MC+ penalty is given by

$$
S(\tilde{\theta}) = \begin{cases} \dfrac{\mathrm{sgn}(\tilde{\theta})(|\tilde{\theta}| - \rho^*)_+}{1 - 1/\gamma} & \text{if } |\tilde{\theta}| \leq \rho^*\gamma \\ \tilde{\theta} & \text{if } |\tilde{\theta}| > \rho^*\gamma. \end{cases}
$$

After updating $\boldsymbol{\Lambda}$ by the coordinate descent algorithm, the new values of $\boldsymbol{\Psi}_{\text{new}}$ and $\boldsymbol{\Phi}_{\text{new}}$ are obtained by maximizing the expected penalized log-likelihood function in (6) as follows:

$$
\begin{aligned}
(\psi_i)_{\text{new}} &= s_{ii} - 2(\hat{\boldsymbol{\lambda}}_i^T)_{\text{new}}\mathbf{b}_i + (\hat{\boldsymbol{\lambda}}_i)_{\text{new}}^T \mathbf{A}(\hat{\boldsymbol{\lambda}}_i)_{\text{new}} \quad \text{for } i = 1, \ldots, p, \\
\boldsymbol{\Phi}_{\text{new}} &= \arg\min_{\boldsymbol{\Phi}}\{\log|\boldsymbol{\Phi}| + \mathrm{tr}(\boldsymbol{\Phi}^{-1}\mathbf{A})\},
\end{aligned}
$$

where $(\psi_i)_{\text{new}}$ is the i-th diagonal element of $\boldsymbol{\Psi}_{\text{new}}$ and $(\hat{\boldsymbol{\lambda}}_i)_{\text{new}}$ is the i-th column of $\hat{\boldsymbol{\Lambda}}_{\text{new}}$. The explicit formula of $\boldsymbol{\Phi}_{\text{new}}$ may not be easily derived, because the diagonal elements of $\boldsymbol{\Phi}$ are fixed by 1. Therefore, the non-diagonal elements of $\boldsymbol{\Phi}_{\text{new}}$ are estimated by Broyden-Fletcher-Goldfarb-Shanno (BFGS) optimization procedure.

3.2.2 Pathwise Algorithm

A pathwise algorithm for orthogonal case has been proposed by Hirose and Yamamoto (2012), and we apply their algorithm to the oblique case. The pathwise algorithm can produce the solution for the grid of increasing ρ values $P = \{\rho_1, \ldots, \rho_K\}$ and a grid of increasing values $\Gamma = \{\gamma_1, \ldots, \gamma_T\}$ efficiently, where γ_T gives the lasso penalty (e.g., $\gamma_T = \infty$ for MC+ family). First, we compute the lasso solution path for $P = \{\rho_1, \ldots, \rho_K\}$ by decreasing the sequence of values for ρ, starting with the largest value $\rho = \rho_K$ for

which the estimates of factor loadings $\hat{\mathbf{\Lambda}} = \mathbf{O}$. Next, the value of γ_{T-1} is selected, and the solutions are produced for the sequence of $P = \{\rho_1, \ldots, \rho_K\}$. The solution at (γ_{T-1}, ρ_k) can be computed by using the solution at (γ_T, ρ_k), which leads to improved and smoother objective value surfaces (Mazumder et al. 2011). In the same way, for $t = T - 2, \ldots, 1$, the solution at (γ_t, ρ_k) can be computed by using the solution at (γ_{t+1}, ρ_k).

3.3 Selection of the Regularization Parameter

In this modeling procedure, it is important to select the appropriate value of the regularization parameter ρ. The following two selection procedures are introduced.

3.3.1 Model Selection Criteria

The selection of the regularization parameter can be viewed as a model selection and evaluation problem. In regression analysis, the degrees of freedom of the lasso (Zou et al. 2007) may be used for selecting the regularization parameter. With the use of the degrees of freedom, the following model selection criteria are introduced:

$$\text{AIC} = -2\ell(\hat{\mathbf{\Lambda}}, \hat{\mathbf{\Psi}}, \hat{\mathbf{\Phi}}) + 2p^*$$
$$\text{BIC} = -2\ell(\hat{\mathbf{\Lambda}}, \hat{\mathbf{\Psi}}, \hat{\mathbf{\Phi}}) + p^* \log N,$$
$$\text{CAIC} = -2\ell(\hat{\mathbf{\Lambda}}, \hat{\mathbf{\Psi}}, \hat{\mathbf{\Phi}}) + p^*(\log N + 1),$$

where the number of parameters is given by $p^* = df(\rho_k) + m_0(m_0 - 1)/2 + p$. Here $df(\rho_k)$ is the number of nonzero parameters for the lasso penalty at $\rho = \rho_k$, $m_0(m_0 - 1)/2$ is the number of parameters in factor correlation matrix and p is the number of parameters in unique variances. Note that this formula can be applied to any value of γ if the reparameterization of the penalty function (Mazumder et al. 2011) is carried out, because the reparameterization constrains the degrees of freedom to be constant as γ varies.

3.3.2 Goodness-of-Fit Index

It may be easy to interpret the estimated model when the factor loadings are sufficiently sparse. However, a model that is too sparse does not fit the data. Therefore, it is reasonable to select a regularization parameter that produces sparse solutions and also yields large values for the following goodness-of-fit index (GFI) and the adjusted GFI (AGFI):

$$\text{GFI} = 1 - \frac{\text{tr}[\{\hat{\mathbf{\Sigma}}^{-1}(\mathbf{S} - \hat{\mathbf{\Sigma}})\}^2]}{\text{tr}[\{\hat{\mathbf{\Sigma}}^{-1}\mathbf{S}\}^2]},$$
$$\text{AGFI} = 1 - \frac{p(p+1)(1 - \text{GFI})}{p(p+1) - 2df},$$

9

where $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Phi}}\hat{\boldsymbol{\Lambda}}^T + \hat{\boldsymbol{\Psi}}$. The GFI and AGFI take values from 0 through 1. In our experience, the model is fitted well if the value of the GFI is greater than 0.9.

3.4 Treatment for Improper Solutions

It is well-known that the maximum likelihood estimates of unique variances can turn out to be zero or negative, which is referred as the improper solutions, and many researchers have studied this problem (e.g., Van Driel 1978; Anderson and Gerbing 1984; Kano 1998). In general, the occurrence of improper solutions makes converge of the algorithm slow and unstable. In order to handle this issue, we add a penalty with respect to $\boldsymbol{\Psi}$ to (5) according to the basic idea given by Martin and McDonald (1975) and Hirose et al. (2011):

$$\ell_\rho^*(\boldsymbol{\Lambda}, \boldsymbol{\Psi}, \boldsymbol{\Phi}) = \ell_\rho(\boldsymbol{\Lambda}, \boldsymbol{\Psi}, \boldsymbol{\Phi}) - \frac{N}{2}\eta \mathrm{tr}(\boldsymbol{\Psi}^{-1/2}\mathbf{S}\boldsymbol{\Psi}^{-1/2}),$$

where η is a tuning parameter. Note that when $\psi_i \to 0$, $\mathrm{tr}(\boldsymbol{\Psi}^{-1/2}\mathbf{S}\boldsymbol{\Psi}^{-1/2}) \to \infty$. Thus, the penalty term $\mathrm{tr}(\boldsymbol{\Psi}^{-1/2}\mathbf{S}\boldsymbol{\Psi}^{-1/2})$ prevents the occurrence of improper solutions. Hirose et al. (2011) derived a generalized Bayesian information criterion (Konishi et al. 2004) for selecting the appropriate value of η, whereas it is difficult to derive generalized Bayesian model criterion in lasso-type penalization procedure. In practice, the penalty term can prevent the occurrence of improper solution even when η is very small such as 0.001.

We provide a package `fanc` in `R` (R Development Core Team 2010), which implements our algorithm to produce the entire solution path. The package `fanc` is available from Comprehensive R Archive Network (CRAN) at `http://cran.r-project.org/web/packages/fanc/index.`

4 Numerical Examples

4.1 Monte Carlo Simulations

In the simulation study, we used three models according to the following factor loadings:

Model (A):
$$\boldsymbol{\Lambda} = \begin{pmatrix} 0.9 & 0.9 & 0.9 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.8 & 0.8 & 0.8 \end{pmatrix}^T,$$

Model (B):
$$\boldsymbol{\Lambda} = \begin{pmatrix} 0.9 & 0.9 & 0.9 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.8 & 0.8 & 0.8 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.7 & 0.7 & 0.7 \end{pmatrix}^T,$$

Model (C):

10

$$\mathbf{\Lambda} = \begin{pmatrix} 0.9 \cdot \mathbf{1}_{25} & \mathbf{0}_{25} & \mathbf{0}_{25} & \mathbf{0}_{25} \\ \mathbf{0}_{25} & 0.8 \cdot \mathbf{1}_{25} & \mathbf{0}_{25} & \mathbf{0}_{25} \\ \mathbf{0}_{25} & \mathbf{0}_{25} & 0.7 \cdot \mathbf{1}_{25} & \mathbf{0}_{25} \\ \mathbf{0}_{25} & \mathbf{0}_{25} & \mathbf{0}_{25} & 0.6 \cdot \mathbf{1}_{25} \end{pmatrix},$$

where $\mathbf{1}_q$ is a q-dimensional vector with each element being 1, and $\mathbf{0}_q$ is a q-dimensional zero vector. For all models, we set $\mathbf{\Phi} = 0.4 \cdot \mathbf{I}_m + 0.6 \cdot \mathbf{1}_m \mathbf{1}_m^T$, and $\mathbf{\Psi} = \mathrm{diag}(\mathbf{I}_m - \mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}^T)$. The Model (C) is a relatively large model compared with Models (A) and (B).

For each model, 1000 data sets were generated with $\mathbf{x} \sim N_p(\mathbf{0}, \mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}^T + \mathbf{\Psi})$. The number of observations was $N = 50, 100$, and 200. The model was estimated by the maximum penalized likelihood method based on both orthogonal model and oblique model. The penalty functions were the MC+ family with $\gamma = 2.10$ and the lasso, and the regularization parameter was selected by the AIC, BIC and CAIC. For comparison, we also estimated the model by the maximum likelihood method, and employed the rotation techniques based on the following criteria: the lasso loss criterion (both orthogonal and oblique models), the varimax criterion (orthogonal model) and promax criterion (oblique model). For example, the lasso loss criterion for oblique model is formulated as follows:

$$\min_{\mathbf{T}} \sum_{i=1}^{p} \sum_{j=1}^{m} |\hat{\lambda}_{ij}^*|, \quad \text{s.t.} \quad \hat{\mathbf{\Lambda}}^* = \hat{\mathbf{\Lambda}}_{\mathrm{MLE}} \mathbf{T}, \ \mathrm{diag}(\mathbf{T}^T \mathbf{T}) = \mathbf{I},$$

where $\hat{\mathbf{\Lambda}}^* = (\hat{\lambda}_{ij}^*)$ and $\hat{\mathbf{\Lambda}}_{\mathrm{MLE}}$ is the maximum likelihood estimates of factor loadings. Note that the lasso loss function is included in the class of component loss function (Jennrich 2004, 2006).

Tables 1, 2 and 3 show the mean squared error of factor loadings, the true positive rate (TPR), and true negative rate (TNR). The mean squared error is defined by MSE $= \sum_{s=1}^{1000} \|\mathbf{\Lambda} - \hat{\mathbf{\Lambda}}^{(s)}\|^2 / (1000pm)$, where $\hat{\mathbf{\Lambda}}^{(s)}$ is the estimated factor loading for the s-th dataset. The TPR (TNR) indicates the proportion of cases where non-zero (zero) factor loadings correctly set to non-zero (zero). Note that the maximum likelihood estimates are not available when $N \leq p$, so that the results of rotation techniques based on the maximum likelihood estimates for $N = 50$ and $N = 100$ in Model (C) were not displayed. We can see that

- The lasso-type regularization with orthogonal model yielded large MSE and small TNR even when the number of observations N was sufficiently large, which suggests the orthogonal model may produce different factor structure from the true one.

- For Model (C), although the maximum likelihood estimates were not available when $N = 50$ and $N = 100$, the penalized likelihood procedure via BIC and CAIC with MC+ relatively selected correct models.

Table 1: Mean squared error, the true positive rate (TPR), and true negative rate (TNR) for Model (A). In the second column, "obl" and "ort" indicate the oblique model and orthogonal model, respectively. In the last column, "P/V" indicates the promax rotation for oblique case, and the varimax rotation for orthogonal case.

			Penalization						Rotation	
			AIC		BIC		CAIC		—	—
N			MC+	lasso	MC+	lasso	MC+	lasso	lasso	P/V
50	obl	MSE	0.14	0.18	0.14	0.20	0.16	0.24	0.17	0.14
		TPR	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00
		TNR	0.75	0.47	0.84	0.55	0.86	0.57	0.05	0.00
	ort	MSE	1.21	1.13	1.21	1.08	1.21	1.02	1.27	0.53
		TPR	0.97	0.98	0.97	0.98	0.96	0.98	0.98	1.00
		TNR	0.30	0.20	0.33	0.23	0.35	0.26	0.14	0.00
100	obl	MSE	0.06	0.09	0.04	0.10	0.03	0.11	0.08	0.06
		TPR	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		TNR	0.81	0.46	0.91	0.56	0.93	0.58	0.06	0.00
	ort	MSE	1.03	0.95	1.05	0.90	1.05	0.87	1.04	0.48
		TPR	0.99	0.99	0.98	0.99	0.98	0.99	0.99	1.00
		TNR	0.34	0.21	0.35	0.24	0.36	0.25	0.14	0.00
200	obl	MSE	0.02	0.05	0.01	0.06	0.01	0.06	0.04	0.03
		TPR	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		TNR	0.87	0.47	0.97	0.56	0.97	0.60	0.07	0.00
	ort	MSE	0.89	0.84	0.89	0.78	0.89	0.77	0.88	0.46
		TPR	0.99	1.00	0.99	1.00	0.99	1.00	1.00	1.00
		TNR	0.41	0.21	0.42	0.26	0.43	0.26	0.14	0.00

- The MC+ family often performed better than the lasso in terms of both the MSE and model consistency.

- The BIC and CAIC may often select the correct model compared with the AIC.

4.2   Analysis of Harman's psychological tests data

We illustrate the proposed procedure by Harman's psychological tests data (Harman 1976). This data represents scores of $N = 145$ subjects on the 24 psychological tests. The dataset is available from the `datasets` in the software `R` (R Development Core Team 2010). Table 4 shows the factor loadings estimated by MC+ based on both orthogonal and oblique models at $\gamma = 2.10$. The value of ρ was selected by the BIC. With the MC+ based on the orthogonal model, all elements of the first column of the estimated factor

Table 2: Mean squared error, the true positive rate (TPR), and true negative rate (TNR) for Model (B). In the second column, "obl" and "ort" indicate the oblique model and orthogonal model, respectively. In the last column, "P/V" indicates the promax rotation for oblique case, and the varimax rotation for orthogonal case.

			Penalization						Rotation	
			AIC		BIC		CAIC		—	—
N			MC+	lasso	MC+	lasso	MC+	lasso	lasso	P/V
50	obl	MSE	0.84	0.81	1.00	1.21	1.08	1.41	0.81	0.72
		TPR	0.97	0.96	0.91	0.86	0.88	0.82	1.00	1.00
		TNR	0.69	0.47	0.81	0.56	0.85	0.59	0.02	0.00
	ort	MSE	2.56	2.03	2.66	1.95	2.66	2.05	2.33	1.33
		TPR	0.94	0.97	0.91	0.90	0.89	0.84	0.99	1.00
		TNR	0.35	0.25	0.40	0.35	0.45	0.40	0.14	0.00
100	obl	MSE	0.29	0.34	0.44	0.62	0.53	0.83	0.41	0.31
		TPR	1.00	1.00	0.96	0.94	0.94	0.90	1.00	1.00
		TNR	0.77	0.47	0.88	0.57	0.90	0.59	0.03	0.00
	ort	MSE	2.35	2.02	2.43	1.80	2.43	1.76	2.24	1.13
		TPR	0.96	0.98	0.94	0.96	0.94	0.94	0.99	1.00
		TNR	0.38	0.23	0.42	0.30	0.44	0.34	0.16	0.00
200	obl	MSE	0.10	0.16	0.07	0.18	0.08	0.22	0.19	0.12
		TPR	1.00	1.00	1.00	1.00	0.99	0.99	1.00	1.00
		TNR	0.86	0.48	0.95	0.56	0.96	0.58	0.03	0.00
	ort	MSE	2.05	1.85	2.11	1.64	2.13	1.59	1.95	1.02
		TPR	0.98	0.99	0.97	0.99	0.97	0.99	0.99	1.00
		TNR	0.41	0.21	0.44	0.27	0.45	0.28	0.16	0.00

loadings were relatively large. This phenomena has been described in Section 2.2. On the other hand, the MC+ based on the oblique model estimated a loading matrix where the first column was sparse. The AGFI and GFI of the oblique model were 0.78 and 0.87, respectively, which might be large enough to conclude that the estimated model fit the observed data.

5   Concluding remarks

In exploratory factor analysis, the lasso based on the orthogonal model often fails in approximating the oblique structure. We have shown that this disadvantage comes from the rotation problem of factor loadings. Then, a maximum penalized likelihood factor analysis based on the oblique model has been proposed to handle this problem. Our

Table 3: Mean squared error, the true positive rate (TPR), and true negative rate (TNR) for Model (C). In the second column, "obl" and "ort" indicate the oblique model and orthogonal model, respectively. In the last column, "P/V" indicates the promax rotation for oblique case, and the varimax rotation for orthogonal case.

			Penalization						Rotation	
			AIC		BIC		CAIC		—	—
N			MC+	lasso	MC+	lasso	MC+	lasso	lasso	P/V
50	obl	MSE	8.54	10.0	7.68	17.2	8.84	20.6	—	—
		TPR	0.99	1.00	0.92	0.97	0.86	0.92	—	—
		TNR	0.43	0.14	0.85	0.36	0.96	0.44	—	—
	ort	MSE	28.0	15.3	27.4	15.5	22.9	20.3	—	—
		TPR	0.98	1.00	0.97	0.99	0.94	0.92	—	—
		TNR	0.26	0.06	0.32	0.20	0.51	0.36	—	—
100	obl	MSE	3.51	5.73	1.79	12.4	2.08	13.2	—	—
		TPR	1.00	1.00	0.99	1.00	0.98	0.99	—	—
		TNR	0.58	0.14	0.99	0.41	0.99	0.43	—	—
	ort	MSE	29.5	15.5	23.0	11.9	13.8	12.6	—	—
		TPR	0.98	1.00	0.98	1.00	0.97	0.99	—	—
		TNR	0.32	0.04	0.51	0.13	0.74	0.22	—	—
200	obl	MSE	0.97	3.39	0.67	9.24	0.67	10.2	2.18	1.68
		TPR	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		TNR	0.91	0.13	1.00	0.44	1.00	0.48	0.00	0.00
	ort	MSE	30.2	16.1	19.2	12.1	9.12	11.1	18.0	13.0
		TPR	0.99	1.00	0.99	1.00	0.99	1.00	1.00	1.00
		TNR	0.41	0.03	0.65	0.07	0.86	0.09	0.02	0.00

modeling strategy has been investigated through Monte Carlo simulations and the analysis of a real data. Simulation results show that the proposed procedure can yield much smaller mean squared error and true negative rate compared with the penalized likelihood factor analysis via the orthogonal model. Furthermore, the MC+ often produced sparser solutions than the lasso, so that the true factor structure can often be reconstructed. In the Harman's psychological data example, the orthogonal model estimated factor loadings where first column was dense, whereas our procedure produced sparse factor loadings.

As a future research topic, it would be interesting to construct a penalization procedure via nonconvex penalties for structural equation modeling, such as LISREL (Jöreskog and Sörbom 1996), which is able to express much more complex covariance structure between observable variables and common factors. In this paper, the tuning parameter was selected by the information criteria based on the degrees of freedom of the lasso. The degrees of free-

Table 4: Loading matrices estimated by MC+ based on both orthogonal and oblique models at $\gamma = 2.10$ for 24 psychological tests data. The value of ρ was selected by the BIC.

MC+ (Orthogonal)				MC+ (Oblique)			
0.75	−0.08	0.00	0.00	0.73	0.00	0.00	0.00
0.46	0.00	0.00	0.00	0.47	0.00	0.00	0.00
0.56	0.00	0.16	0.00	0.66	0.00	−0.18	0.00
0.59	0.00	0.00	0.00	0.58	0.00	0.00	0.00
0.47	0.63	−0.17	0.00	0.00	0.72	0.22	0.00
0.49	0.67	0.00	0.00	0.00	0.75	0.00	0.17
0.48	0.67	−0.12	−0.12	0.00	0.79	0.13	0.00
0.56	0.41	−0.16	0.00	0.23	0.48	0.19	0.00
0.49	0.71	0.00	0.00	0.00	0.81	0.00	0.13
0.17	0.14	−0.81	0.24	−0.31	0.00	0.93	0.11
0.36	0.14	−0.42	0.34	0.00	0.00	0.44	0.35
0.37	−0.10	−0.63	0.13	0.14	−0.19	0.71	0.00
0.59	0.00	−0.41	0.00	0.37	0.00	0.44	0.00
0.23	0.25	0.00	0.48	0.00	0.00	0.00	0.58
0.26	0.15	0.00	0.45	0.00	0.00	0.00	0.53
0.51	0.00	0.10	0.42	0.44	−0.14	−0.15	0.49
0.26	0.19	−0.11	0.54	0.00	0.00	0.00	0.65
0.43	0.00	−0.19	0.43	0.28	−0.20	0.18	0.43
0.38	0.06	0.00	0.30	0.22	0.00	0.00	0.35
0.56	0.25	0.00	0.16	0.37	0.24	0.00	0.17
0.55	0.00	−0.31	0.16	0.38	0.00	0.39	0.00
0.56	0.24	0.00	0.16	0.37	0.23	0.00	0.18
0.67	0.19	−0.10	0.10	0.50	0.22	0.15	0.00
0.43	0.29	−0.41	0.24	0.00	0.22	0.47	0.22

dom of the lasso are usually applied to the regression model, whereas we have not given a mathematical support for the degrees of freedom of the lasso in factor analysis model yet. Another interesting topic is to provide a theoretical justification of the information criteria given by Section 3.3.1.

Appendix: Derivation of complete-data penalized log-likelihood function in EM algorithm

In order to apply the EM algorithm, first, the common factors $\boldsymbol{f}_n$ can be regarded as missing data and maximize the complete-data penalized log-likelihood function

$$l_\rho^C(\boldsymbol{\Lambda}, \boldsymbol{\Psi}, \boldsymbol{\Phi}) = \sum_{n=1}^{N} \log f(\boldsymbol{x}_n, \boldsymbol{f}_n) - N \sum_{i=1}^{p} \sum_{j=1}^{m} \rho P(|\lambda_{ij}|),$$

where the density function $f(\boldsymbol{x}_n, \boldsymbol{f}_n)$ is defined by

$$f(\boldsymbol{x}_n, \boldsymbol{f}_n) = (2\pi)^{-p/2} |\boldsymbol{\Psi}|^{-1/2} \exp\left\{ -\frac{(\boldsymbol{x}_n - \boldsymbol{\Lambda} \boldsymbol{f}_n)^T \boldsymbol{\Psi}^{-1} (\boldsymbol{x}_n - \boldsymbol{\Lambda} \boldsymbol{f}_n)}{2} \right\}$$

$$\cdot (2\pi)^{-m/2} |\boldsymbol{\Phi}|^{-1/2} \exp\left(-\frac{\boldsymbol{f}_n^T \boldsymbol{\Phi}^{-1} \boldsymbol{f}_n}{2} \right)$$

$$= \prod_{i=1}^{p} \left\{ (2\pi\psi_i)^{-1/2} \exp\left(-\frac{(x_{ni} - \boldsymbol{\lambda}_i^T \boldsymbol{f}_n)^2}{2\psi_i} \right) \right\}$$

$$\cdot (2\pi)^{-m/2} |\boldsymbol{\Phi}|^{-1/2} \exp\left(-\frac{\boldsymbol{f}_n^T \boldsymbol{\Phi}^{-1} \boldsymbol{f}_n}{2} \right)$$

Then, the expectation of l_ρ^C can be taken with respect to the distributions $f(\boldsymbol{f}_n | \boldsymbol{x}_n, \boldsymbol{\Lambda}, \boldsymbol{\Psi}, \boldsymbol{\Phi})$,

$$E[l_\rho^C(\boldsymbol{\Lambda}, \boldsymbol{\Psi}, \boldsymbol{\Phi})] = -\frac{N(p+m)}{2} \log(2\pi) - \frac{N}{2} \sum_{i=1}^{p} \log \psi_i$$

$$-\frac{1}{2} \sum_{n=1}^{N} \sum_{i=1}^{p} \frac{x_{ni}^2 - 2x_{ni} \boldsymbol{\lambda}_i^T E[\boldsymbol{F}_n | \boldsymbol{x}_n] + \boldsymbol{\lambda}_i^T E[\boldsymbol{F}_n \boldsymbol{F}_n^T | \boldsymbol{x}_n] \boldsymbol{\lambda}_i}{\psi_i}$$

$$-\frac{N}{2} \log |\boldsymbol{\Phi}| - \frac{1}{2} \mathrm{tr}\left\{ \sum_{n=1}^{N} \boldsymbol{\Phi}^{-1} E[\boldsymbol{F}_n \boldsymbol{F}_n^T | \boldsymbol{x}_n] \right\} - N \sum_{i=1}^{p} \sum_{j=1}^{m} \rho P(|\lambda_{ij}|)$$

For given $\boldsymbol{\Lambda}_{\mathrm{old}}$, $\boldsymbol{\Psi}_{\mathrm{old}}$ and $\boldsymbol{\Phi}_{\mathrm{old}}$, the posterior $f(\boldsymbol{f}_n | \boldsymbol{x}_n, \boldsymbol{\Lambda}_{\mathrm{old}}, \boldsymbol{\Psi}_{\mathrm{old}}, \boldsymbol{\Phi}_{\mathrm{old}})$ is normally distributed with $E[\boldsymbol{F}_n | \boldsymbol{x}_n] = \mathbf{M}^{-1} \boldsymbol{\Lambda}_{\mathrm{old}}^T \boldsymbol{\Psi}_{\mathrm{old}}^{-1} \boldsymbol{x}_n$ and $E[\boldsymbol{F}_n \boldsymbol{F}_n^T | \boldsymbol{x}_n] = \mathbf{M}^{-1} + E[\boldsymbol{F}_n | \boldsymbol{x}_n] E[\boldsymbol{F}_n | \boldsymbol{x}_n]^T$, where $\mathbf{M} = \boldsymbol{\Lambda}_{\mathrm{old}}^T \boldsymbol{\Psi}_{\mathrm{old}}^{-1} \boldsymbol{\Lambda}_{\mathrm{old}} + \boldsymbol{\Phi}_{\mathrm{old}}^{-1}$. Then, we have

$$\sum_{n=1}^{N} E[\boldsymbol{F}_n] x_{ni} = \sum_{n=1}^{N} \mathbf{M}^{-1} \boldsymbol{\Lambda}_{\mathrm{old}}^T \boldsymbol{\Psi}_{\mathrm{old}}^{-1} \boldsymbol{x}_n x_{ni} = N \mathbf{M}^{-1} \boldsymbol{\Lambda}_{\mathrm{old}}^T \boldsymbol{\Psi}_{\mathrm{old}}^{-1} \mathbf{s}_i,$$

$$\sum_{n=1}^{N} E[\boldsymbol{F}_n \boldsymbol{F}_n^T] = \sum_{n=1}^{N} (\mathbf{M}^{-1} + \mathbf{M}^{-1} \boldsymbol{\Lambda}_{\mathrm{old}}^T \boldsymbol{\Psi}_{\mathrm{old}}^{-1} \boldsymbol{x}_n \boldsymbol{x}_n^T \boldsymbol{\Psi}_{\mathrm{old}}^{-1} \boldsymbol{\Lambda}_{\mathrm{old}} \mathbf{M}^{-1})$$

$$= N(\mathbf{M}^{-1} + \mathbf{M}^{-1} \boldsymbol{\Lambda}_{\mathrm{old}}^T \boldsymbol{\Psi}_{\mathrm{old}}^{-1} \mathbf{S} \boldsymbol{\Psi}_{\mathrm{old}}^{-1} \boldsymbol{\Lambda}_{\mathrm{old}} \mathbf{M}^{-1}),$$

Let $\mathbf{M}^{-1} \boldsymbol{\Lambda}_{\mathrm{old}}^T \boldsymbol{\Psi}_{\mathrm{old}}^{-1} \mathbf{s}_i$ and $\mathbf{M}^{-1} + \mathbf{M}^{-1} \boldsymbol{\Lambda}_{\mathrm{old}}^T \boldsymbol{\Psi}_{\mathrm{old}}^{-1} \mathbf{S} \boldsymbol{\Psi}_{\mathrm{old}}^{-1} \boldsymbol{\Lambda}_{\mathrm{old}} \mathbf{M}^{-1}$ be $\mathbf{b}_i$ and $\mathbf{A}$, respectively. Then, the expectation of l_ρ^C in (6) can be derived.

References

Akaike, H. (1987), "Factor analysis and AIC," *Psychometrika*, 52(3), 317–332.

Anderson, J., and Gerbing, D. (1984), "The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis," *Psychometrika*, 49(2), 155–173.

Breheny, P., and Huang, J. (2011), "Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection," *The annals of applied statistics*, 5(1), 232.

Choi, J., Zou, H., and Oehlert, G. (2011), "A Penalized Maximum Likelihood Approach to Sparse Factor Analysis," *Statistics and Its Interface*, 3(4), 429–436.

Clarke, M. (1970), "A Rapidly Convergent Method for Maximum-Likelihood Factor Analysis," *British Journal of Mathematical and Statistical Psychology*, 23(1), 43–52.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression (with discussion)," *The Annals of Statistics*, 32, 407–499.

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360.

Friedman, J. H. (2012), "Fast sparse regression and classification," *International Journal of Forecasting*, 28(3), 722–738.

Friedman, J., Hastie, H., Höfling, H., and Tibshirani, R. (2007), "Pathwise Coordinate Optimization," *The Annals of Applied Statistics*, 1, 302–332.

Friedman, J., Hastie, T., and Tibshirani, R. (2010), "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, 33.

Harman, H. (1976), *Modern factor analysis* University of Chicago Press.

Hendrickson, A., and White, P. (1964), "Promax: A quick method for rotation to oblique simple structure," *British Journal of Statistical Psychology*, 17(1), 65–70.

Hirose, K., Kawano, S., Konishi, S., and Ichikawa, M. (2011), "Bayesian information criterion and selection of the number of factors in factor analysis models," *Journal of Data Science*, 9(1), 243–259.

Hirose, K., and Yamamoto, M. (2012), "Sparse estimation via nonconcave penalized likelihood in a factor analysis model," *arXiv preprint arXiv:1205.5868*, .

Jennrich, R. (2004), "Rotation to simple loadings using component loss functions: The orthogonal case," *Psychometrika*, 69(2), 257–273.

Jennrich, R. (2006), "Rotation to simple loadings using component loss functions: The oblique case," *Psychometrika*, 71(1), 173–191.

Jennrich, R., and Robinson, S. (1969), "A Newton-Raphson algorithm for maximum likelihood factor analysis," *Psychometrika*, 34(1), 111–123.

Jöreskog, K. (1967), "Some contributions to maximum likelihood factor analysis," *Psychometrika*, 32(4), 443–482.

Jöreskog, K. G., and Sörbom, D. (1996), *LISREL 8 user's reference guide* Scientific Software.

Kaiser, H. (1958), "The varimax criterion for analytic rotation in factor analysis," *Psychometrika*, 23(3), 187–200.

Kano, Y. (1998), Improper Solutions in Exploratory Factor Analysis: Causes and Treatments,, in *Advances in data science and classification: proceedings of the 6th Conference of the International Federation of Classification Societies (IFCS-98), Università" La Sapienza", Rome, 21-24 July, 1998*, Springer Verlag, p. 375.

Konishi, S., Ando, T., and Imoto, S. (2004), "Bayesian information criteria and smoothing parameter selection in radial basis function networks," *Biometrika*, 91(1), 27–43.

Lawley, D., and Maxwell, A. (1971), *Factor analysis as a statistical method*, Vol. 18 Butterworths London.

Martin, J., and McDonald, R. (1975), "Bayesian estimation in unrestricted factor analysis: A treatment for Heywood cases," *Psychometrika*, 40(4), 505–517.

Mazumder, R., Friedman, J., and Hastie, T. (2011), "SparseNet: Coordinate Descent with Nonconvex Penalties," *Journal of the American Statistical Association*, 106, 1125–1138.

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

Ning, L., and Georgiou, T. T. (2011), Sparse factor analysis via likelihood and ℓ_1 regularization,, in *50th IEEE Conference on Decision and Control and European Control Conference*, pp. 5188–5192.

R Development Core Team (2010), "R: A Language and Environment for Statistical Computing," *R Foundation for Statistical Computing Vienna Austria*, . Available at http://www.R-project.org

Rubin, D., and Thayer, D. (1982), "EM algorithms for ML factor analysis," *Psychometrika*, 47(1), 69–76.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Ser. B*, 58, 267–288.

Van Driel, O. (1978), "On various causes of improper solutions in maximum likelihood factor analysis," *Psychometrika*, 43(2), 225–243.

Zhang, C. (2010), "Nearly Unbiased Variable Selection Under Minimax Concave Penalty," *The Annals of Statistics*, 38, 894–942.

Zhao, P., and Yu, B. (2007), "On model selection consistency of Lasso," *Journal of Machine Learning Research*, 7(2), 2541.

Zou, H. (2006), "The Adaptive Lasso and its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429.

Zou, H., Hastie, T., and Tibshirani, R. (2007), "On the Degrees of Freedom of the Lasso," *The Annals of Statistics*, 35, 2173–2192.