Variable selection via the grouped weighted lasso for factor analysis models

Hirose, Kei
Graduate School of Mathematics, Kyushu University

Konishi, Sadanori
Faculty of Mathematics, Kyushu University

出版情報：MI Preprint Series. 2010-10, 2010-03-03. 九州大学大学院数理学研究院
バージョン：
権利関係：

MI Preprint Series
Kyushu University
The Global COE Program
Math-for-Industry Education & Research Hub

Variable selection via the grouped weighted lasso for factor analysis models

Kei Hirose & Sadanori Konishi

MI 2010-10

(Received March 3, 2010)

Faculty of Mathematics
Kyushu University
Fukuoka, JAPAN

Variable selection via the grouped weighted lasso for factor analysis models

Kei Hirose[1] and Sadanori Konishi[2]

[1] *Graduate School of Mathematics, Kyushu University,*

744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan.

[2] *Faculty of Mathematics, Kyushu University,*

744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan.

E-mail: k-hirose@math.kyushu-u.ac.jp, konishi@math.kyushu-u.ac.jp.

Abstract

The L_1 regularization such as the lasso has been widely used in regression analysis since it tends to produce some coefficients that are exactly zero, which leads to variable selection. We consider the problem of variable selection for factor analysis models via the L_1 regularization procedure. In order to select variables each of which is controlled by multiple parameters, we treat parameters as grouped parameters and then apply the grouped lasso. Crucial issues in this modeling procedure include the selection of the number of factors and regularization parameters. Choosing these parameters can be viewed as a model selection and evaluation problem. We derive a model selection criterion for evaluating a factor analysis model via the grouped lasso. The proposed procedure produces estimates that lead to variable selection and also selects the number of factors objectively. Monte Carlo simulations are conducted to investigate the effectiveness of the proposed procedure. A real data example is also given to illustrate our procedure.

Key Words: Factor analysis, Grouped weighted lasso, L_1 regularization, Model selection criterion, Number of factors, Variable selection

1 Introduction

Variable selection is an important topic in statistical analysis. In regression analysis, traditional model selection procedures follow the stepwise deletion and subset selection. However, as analyzed by Breiman (1996), the best subset variable selection often exhibits high variance. To overcome this drawback of subset selection, Tibshirani (1996) proposed a regularization procedure which imposes a L_1 norm penalty on parameters, called the lasso. The L_1 regularization procedure reduces the variance of estimates by sacrificing a little bit of bias, and also produces some coefficients that are exactly zero, which leads to variable selection. Thus, the lasso performs the model selection and estimation simultaneously. Yuan & Lin (2006) proposed the grouped lasso which selects the grouped variables by extending the basic idea of the lasso. Other various kinds of the lasso-type penalty have been proposed (see, e.g., Fan & Li, 2001; Zou & Hastie, 2005; Zou, 2006).

In this paper, we consider the problem of variable selection for factor analysis models. Some authors have discussed this problem and proposed various variable selection procedures (see, e.g., Yanai, 1980; Tanaka, 1983; Gorsuch, 1988; Kano & Ihara, 1994; Ichikawa & Konishi, 1999; Kano & Harada, 2000). Their selection processes are based on a subset selection. Instead of using the subset selection, we focus on the L_1 regularization method. It is well-known that the maximum likelihood factor analysis often yields unstable estimates because of the overparametrization (see, e.g., Akaike, 1987). The regularization method may be useful for factor analysis models because it produces estimates that have small variances. Since each observed variable is controlled by multiple parameters, the ordinary lasso does not work for variable selection. We treat these parameters as grouped parameters and then propose a regularization method via the grouped lasso. Furthermore, we adjust the weight of the grouped lasso penalty so that the proper penalties are imposed on each variable.

Crucial issues in this modeling procedure include the choice of the number of factors and regularization parameters. Regarding the selection of the number of factors, the AIC (Akaike, 1973) and the BIC (Schwarz, 1978) have been widely used. However, these procedures cannot provide suitable values of regularization parameters since they only evaluate models estimated by maximum likelihood procedure. In regression analysis, the selection procedures of the regularization parameter for the lasso has been proposed by Efron *et al.* (2004) and Zou *et al.* (2007), whereas their methods cannot be directly applied to the factor analysis model. We treat a selection of parameters, that include the number of factors and the regularization parameter, as a model selection and evaluation problem, and derive a model selection criterion from a Bayesian viewpoint (Konishi *et al.*, 2004). The proposed method can produce estimates that lead to variable selection and select the number of factors simultaneously.

The remainder of this paper is organized as follows: Section 2 describes the maximum likelihood factor analysis. In Section 3, we introduce a variable selection procedure for factor analysis models via the grouped lasso, and provide a model estimation using the EM algorithms with quadratic approximation. Section 4 derives a model selection criterion for evaluating a factor analysis model via the grouped lasso. Section 5 presents numerical results for both artificial and real datasets. Some concluding remarks are given in Section 6.

2 Maximum likelihood factor analysis

Let $\boldsymbol{X}=(X_1,\cdots,X_p)^T$ be a p-dimensional observable random vector with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$. The factor analysis model is

$$\boldsymbol{X}=\boldsymbol{\mu}+\boldsymbol{\Lambda}\boldsymbol{F}+\boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{\Lambda} = (\lambda_{ij})$ is a $p \times k$ matrix of factor loadings, and $\boldsymbol{F} = (F_1, \cdots, F_k)^T$ 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 $E(\boldsymbol{F}) = \mathbf{0}$, $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, $E(\boldsymbol{F}\boldsymbol{F}^T) = \mathbf{I}_k$, $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) = \mathbf{\Psi}$ and $E(\boldsymbol{F}\boldsymbol{\varepsilon}^T) = \mathbf{0}$, where $\mathbf{I}_k$ is the identity matrix of order k and $\mathbf{\Psi}$ is a $p \times p$ diagonal matrix with i-th diagonal element ψ_i which is called unique variance. Under these assumptions, the variance-covariance matrix of $\boldsymbol{X}$ can be expressed as

$$\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi}.$$

The i-th diagonal element of $\mathbf{\Lambda}\mathbf{\Lambda}^T$ is called communality, which measures the percent of variance in x_i explained by all the factors. It is well-known that factor loadings have a rotational indeterminacy since both $\mathbf{\Lambda}$ and $\mathbf{\Lambda T}$ generate the same covariance matrix $\mathbf{\Sigma}$, where $\mathbf{T}$ is an arbitrary orthogonal matrix.

Assume that the common factors $\boldsymbol{F}$ and the unique factors $\boldsymbol{\varepsilon}$ are, respectively, distributed according to multivariate normal distributions

$$\boldsymbol{F} \sim N_k(\mathbf{0}, \mathbf{I}_k) \quad \text{and} \quad \boldsymbol{\varepsilon} \sim N_p(\mathbf{0}, \mathbf{\Psi}).$$

Suppose that we have a random sample of N observations $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N$ from the p-dimensional normal population $N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$ with $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi}$. Then the log-likelihood function is given by

$$\log f(\boldsymbol{X}_N | \mathbf{\Lambda}, \mathbf{\Psi}) = -\frac{N}{2}\Big\{ p \log(2\pi) + \log|\mathbf{\Sigma}| + \mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{S}) \Big\}, \tag{2}$$

where $\mathbf{X}_N = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N)^T$, $f(\mathbf{X}_N | \mathbf{\Lambda}, \mathbf{\Psi})$ is the likelihood function and $\mathbf{S} = (s_{ij})$ is the sample variance-covariance matrix

$$\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N} (\boldsymbol{x}_n - \bar{\boldsymbol{x}})(\boldsymbol{x}_n - \bar{\boldsymbol{x}})^T,$$

with $\bar{\boldsymbol{x}} = (\bar{x}_1, \ldots, \bar{x}_p)^T$ being the sample mean vector.

The maximum likelihood estimates $\hat{\boldsymbol{\Lambda}}_{\mathrm{ML}}$ and $\hat{\boldsymbol{\Psi}}_{\mathrm{ML}}$ are given as the solutions of $\partial q(\boldsymbol{\Lambda}, \boldsymbol{\Psi})/\partial \boldsymbol{\Lambda} = \mathbf{0}$ and $\partial q(\boldsymbol{\Lambda}, \boldsymbol{\Psi})/\partial \boldsymbol{\Psi} = \mathbf{0}$. Since the solutions cannot be expressed in a closed form, we require some iterative procedure. Some numerical algorithms have been proposed by earlier authors (see, e.g., Jöreskog, 1967; Jennrich & Robinson, 1969; Clarke, 1970; Rubin & Thayer, 1982).

3 Variable selection via the grouped lasso for factor analysis models

In this section we consider the problem of variable selection for factor analysis models via the grouped lasso. First, we provide an insight how a selection procedure works in factor analysis models. From Equation (1), the i-th element of each variable can be written as

$$X_i = \lambda_{i1} F_1 + \cdots + \lambda_{ik} F_k + \varepsilon_i = \boldsymbol{\lambda}_i^T \boldsymbol{F} + \varepsilon_i, \quad i = 1, \ldots, p,$$

where $\boldsymbol{\lambda}_i$ denotes the i-th row of factor loadings $\boldsymbol{\Lambda}$. When all of coefficients on each factor for r-th variable are zero, i.e. $\boldsymbol{\lambda}_r = \mathbf{0}$, we have

$$X_r = \varepsilon_r. \tag{3}$$

The Equation (3) indicates the r-th variable is no more explained by any common factors. Figure 1 presents the path diagram of this situation, and it suggests that the r-th variable could be removed for this model.

In order to implement the variable selection for factor analysis models, we construct grouped parameters each of which consists of

$$\{\lambda_{i1}, \ldots, \lambda_{ik}\}, \quad i = 1, \ldots, p, \tag{4}$$

Table 1: The estimates of parameters based on the maximum likelihood procedures (MLE), the regularization procedure with grouped lasso penalty given by (6) (PMLE) and the regularization procedure with grouped weighted lasso penalty in (8) (PMLE$_w$) when $\rho = 0.1$, 0.15, 0.2 and 0.25.

		$\hat{\lambda}_{11}$	$\hat{\lambda}_{21}$	$\hat{\lambda}_{31}$	$\hat{\lambda}_{41}$
	True value	0.80	0.80	0.80	0.00
	MLE	0.81	0.91	0.84	0.20
$\rho = 0.1$	PMLE	0.70	0.81	0.73	0.09
	PMLE$_w$	0.83	0.85	0.83	0.00
$\rho = 0.15$	PMLE	0.67	0.77	0.70	0.05
	PMLE$_w$	0.80	0.82	0.80	0.00
$\rho = 0.2$	PMLE	0.63	0.73	0.67	0.01
	PMLE$_w$	0.77	0.80	0.77	0.00
$\rho = 0.25$	PMLE	0.60	0.71	0.64	0.00
	PMLE$_w$	0.75	0.78	0.75	0.00

and propose a modeling procedure which tends to produce some of the grouped parameters in (4) that all of its components are exactly zero. The grouped lasso (Yuan & Lin, 2006) is one way to achieve this.

3.1 Grouped lasso penalty

In order to select variables in factor analysis models, we employ the L_1 regularization procedure. The basic idea of regularization method is to add a penalty term to the log-likelihood, and estimate parameters by maximizing the following penalized log-likelihood function:

$$l_\rho(\boldsymbol{\Lambda}, \boldsymbol{\Psi}) = \log f(\boldsymbol{X}_N|\boldsymbol{\Lambda}, \boldsymbol{\Psi}) - p_\rho(\boldsymbol{\Lambda}), \tag{5}$$

where $p_\rho(\boldsymbol{\Lambda})$ is a penalty term with regularization parameter ρ.

In the regularization procedure, it is important to choose the penalty term $p_\rho(\mathbf{\Lambda})$. The lasso penalty (Tibshirani, 1996) is based on the sum of the absolute values of each coefficients. Owing to the property of the lasso penalty it tends to produce some coefficient that are exactly zero. However, the ordinary lasso does not implement the variable selection in factor analysis models since each observed variable consists of grouped parameters in (4) and the lasso cannot produce some of the grouped variables in (4) that all of its components are exactly zero.

Therefore, we apply the grouped lasso (Yuan & Lin, 2006) which selects the members of a group that consists of multiple parameters by extending the basic idea of the lasso. The penalty term based on the grouped lasso is given by

$$p_\rho(\mathbf{\Lambda}) = N\rho \sum_{i=1}^{p} \|\boldsymbol{\lambda}_i\|, \tag{6}$$

where $\|\boldsymbol{\lambda}_i\| = \sqrt{\boldsymbol{\lambda}_i^T \boldsymbol{\lambda}_i}$. However, we observed that the penalty term (6) does not often work well since it selects variable only when the regularization parameter ρ is large.

Here is an example of this phenomena. Assume that $\mathbf{\Lambda} = (0.8, 0.8, 0.8, 0.0)^T$ and $\mathbf{\Psi} = \mathrm{diag}(0.36, 0.36, 0.36, 1.00)$, and we have 50-observations from $N_4(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is given by

$$\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi} = \begin{pmatrix} 1.00 & 0.64 & 0.64 & 0.00 \\ 0.64 & 1.00 & 0.64 & 0.00 \\ 0.64 & 0.64 & 1.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 1.00 \end{pmatrix}. \tag{7}$$

Once the regularization-parameter ρ is determined we can obtain maximum penalized likelihood estimates by EM algorithms (The algorithm will be described in the subsection 3.3). In this example, we investigated 4 variants of ρ: $\rho = 0.1$, 0.15, 0.2 and 0.25. Table 1 shows the estimates of parameters based on the maximum likelihood procedures (MLE), the regularization procedure with grouped lasso penalty given by (6) (PMLE) and the regularization procedure with grouped weighted lasso penalty in (8) (PMLE$_w$).

The grouped weighted lasso penalty will be described in the next subsection.

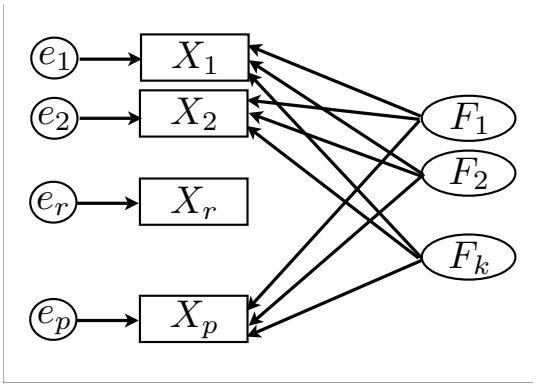

Figure 1: The path diagram of factor analysis model when $\boldsymbol{\lambda}_r = \mathbf{0}$. e_i, $i = 1, \ldots, p$, indicates the error for i-th variable.

When we use the maximum penalized likelihood procedure with grouped lasso penalty in (6) (PMLE), $\hat{\lambda}_{41}$ is exactly zero only when $\rho = 0.25$, whereas $\hat{\lambda}_{11}$, $\hat{\lambda}_{21}$ and $\hat{\lambda}_{31}$ are too small compared with true values because the regularization parameter is too large.

3.2 Grouped weighted lasso penalty

In order to overcome the drawback of ordinary grouped lasso method described in the previous subsection, we propose a weighted grouped L_1 penalty

$$p_\rho(\boldsymbol{\Lambda}) = N\rho \sum_{i=1}^{p} \hat{w}_i \|\boldsymbol{\lambda}_i\|, \tag{8}$$

where $\hat{w}_i$ is the weight of each group. The weighted lasso penalty has been proposed by Zou (2006), Shimamura *et al.* (2007) and Tateishi *et al.* (2010) to improve the performance of the ordinary lasso.

It is important to choose the weight $\hat{w}_i$. We propose weights given as follows

$$\hat{w}_i = \left(\frac{1}{\sum_{h=1}^{p} 1/s^{hh}} \right) \frac{1}{s^{ii}}, \tag{9}$$

where s^{ii} is the ith diagonal elements of $\boldsymbol{S}^{-1}$.

It is known that $\hat{w}_i$ is large when the i-th variable has small correlation with other variables. This can be confirmed by the example of the previous subsection. The correlations between X_4 and X_1, X_4 and X_2, and X_4 and X_3 are zero (see the variance-covariance matrix in (7)), and the weight based on (9) is given by

$$(\hat{w}_1, \hat{w}_2, \hat{w}_3, \hat{w}_4)^T = (0.61, 0.57, 0.61, 2.21)^T.$$

It can be seen that $\hat{w}_4$ is much larger than other weights. This means the grouped weighted estimate $\hat{\lambda}_{41}$ tends to become zero even when ρ is small.

We also obtain estimates of factor loadings based on the grouped weighted lasso (PMLE$_w$) for that example, which is given in Table 1. We observe that $\hat{\lambda}_{41}$ is exactly zero even when $\rho = 0.1$ with the grouped weighted lasso. Since the grouped weighted lasso encourages some coefficients zero even if ρ is small, the estimates of non-zero parameters are close to true values compared with the grouped lasso estimates with $\rho = 0.25$.

In section 5, we compare the performance of the ordinary grouped lasso penalty with the grouped weighted lasso penalty, and we observed that the proposed grouped weighted lasso performs well.

3.3 Estimation

In order to obtain the maximum penalized likelihood estimates in factor analysis models, we employ an EM algorithm. Rubin & Thayer (1982) suggested an estimation procedure via an EM algorithm in maximum likelihood factor analysis. The advantage of the EM algorithms is that even if the likelihood function is not concave with respect to the parameters, the algorithm leads to a maximization of the function.

We provide the expectation and maximization steps for estimating the factor analysis model via the grouped lasso within a general framework of EM algorithms. We regard the common factors as missing variables, and maximize the complete-data log-likelihood

using a posterior distribution for the missing variables. For the L_1 type regularization method, it is difficult to obtain the analytical form of the updated $\boldsymbol{\lambda}_i$ $(i = 1, \ldots, p)$ since the lasso estimate is non-differentiable when $\boldsymbol{\lambda}_i = \mathbf{0}$. Hence we apply to the quadratic approximation given by Fan & Li (2001). Then the iterative procedure is

$$\hat{\boldsymbol{\lambda}}_i = \left\{(\boldsymbol{B} + \boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})/\psi_i^2 + (\rho\hat{w}_i/\|\boldsymbol{\lambda}_i\|)\mathbf{I}_k\right\}^{-1}\left\{\frac{1}{\psi_i^2}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{S}_{i.}\right\}, \tag{10}$$

$$\hat{\boldsymbol{\Psi}} = \mathrm{Diag}\left[\mathbf{S} - 2\mathbf{S}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\hat{\boldsymbol{\Lambda}}^T + \hat{\boldsymbol{\Lambda}}\mathbf{B}\hat{\boldsymbol{\Lambda}}^T + \hat{\boldsymbol{\Lambda}}\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\hat{\boldsymbol{\Lambda}}^T\right], \tag{11}$$

where $\mathbf{S}_{i.}$ is the i-th column of the matrix $\boldsymbol{S}$ and $\boldsymbol{B} = \mathbf{I}_k - \boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}$. The derivation of the procedure is detailed in Appendix A.

In order to eliminate the rotational indeterminacy from $\boldsymbol{\Lambda}$, we impose restrictions that $\lambda_{ij} = 0$ $(i > j)$ (see, for example, Anderson & Rubin; 1956).

4 Model selection criterion

As described in the example in the previous section, the maximum penalized estimates depend on the regularization parameter ρ. Hence, it is important to choose the appropriate value of ρ. Moreover, the selection of the number of factors k is also an essential problem. The selection of parameters ρ and k can be viewed as a model selection and evaluation problem. In this section we derive a model selection criterion for evaluating a factor analysis model via the grouped weighted lasso.

The generalized Bayesian information criterion (GBIC), proposed by Konishi *et al.* (2004), enables us to choose adjusted parameters including the regularization-parameter ρ and the number of factors k simultaneously by extending the Bayesian information criterion (BIC) proposed by Schwarz (1978). BIC only deals with models estimated by the maximum likelihood method, whereas the model selection criterion GBIC also applies to models estimated by the maximum penalized likelihood method. For model selection criteria we refer to Konishi & Kitagawa (2008) and references given therein.

Suppose that $\boldsymbol{\theta}$ is a parameter vector given by

$$\boldsymbol{\theta} = (\boldsymbol{\lambda}_{\cdot 1}^T, \boldsymbol{\lambda}_{\cdot 2}^T, \cdots, \boldsymbol{\lambda}_{\cdot k}^T, \mathrm{Diag}(\boldsymbol{\Psi})^T)^T,$$

where $\boldsymbol{\lambda}_{\cdot i} = (\lambda_{i,i}, \lambda_{i+1,i}, \cdots, \lambda_{p,i})^T$. We used the definition of $\boldsymbol{\lambda}_{\cdot i}$ which consists of only the lower elements of $\boldsymbol{\Lambda}$ because it eliminates the rotational indeterminacy as described in the previous section. Let $f(\mathbf{X}_N|\hat{\boldsymbol{\theta}})$ be the estimated model by maximum penalized likelihood methods. Then we have a statistical model

$$f(\mathbf{X}_N|\hat{\boldsymbol{\theta}}) = (2\pi)^{-\frac{Np}{2}} |\hat{\boldsymbol{\Sigma}}|^{-\frac{N}{2}} \exp\left\{-\frac{N}{2}\mathrm{tr}\left(\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{S}\right)\right\}, \tag{12}$$

where $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}^T + \hat{\boldsymbol{\Psi}}$.

It should be noted that the model selection criterion GBIC cannot be directly derived for the L_1 type regularization method, since we need a second order differential $\partial^2\|\hat{\boldsymbol{\lambda}}_i\|/\partial\lambda_{ij}\partial\lambda_{ij'}$ and it is difficult to derive it when $\hat{\boldsymbol{\lambda}}_i = \mathbf{0}$. In order to overcome this difficulty, we define an Active set

$$\mathcal{A} = \{j : \hat{\theta}_j \neq 0\} \tag{13}$$

and derive a second order differential of the penalized log-likelihood of $\boldsymbol{\theta}_{\mathcal{A}}$ instead of $\boldsymbol{\theta}$, where

$$\boldsymbol{\theta}_{\mathcal{A}}^T = (\cdots \theta_j \cdots)_{j\in\mathcal{A}}.$$

The basic idea of the elimination of the non-zero parameter is given by Efron *et al.* (2004) and Zou *et al.* (2007).

The model selection criterion GBIC for evaluating the factor analysis model via the grouped wighted lasso is given by

$$\begin{aligned}\mathrm{GBIC} = &-p^* \log(2\pi) + p^* \log N + \log|J_\rho(\hat{\boldsymbol{\theta}}_{\mathcal{A}})| + N\left\{p\log(2\pi) + \log|\hat{\boldsymbol{\Sigma}}| + \mathrm{tr}(\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{S})\right\} \\ &-2\log C + 2N\rho\sum_{i=1}^{p}\hat{w}_i\|\hat{\boldsymbol{\lambda}}_i\|,\end{aligned} \tag{14}$$

where C is given in (B2) in Appendix B, p^* is the number of non-zero parameters and $J_\rho(\hat{\boldsymbol{\theta}}_{\mathcal{A}})$ is a second order differential of the penalized log-likelihood function for the Active set parameters. We choose optimum values of the hyper-parameter ρ and the number of factors k which simultaneously minimize the value of the model selection criterion in (14). The derivation of GBIC is detailed in Appendix B.

Other traditional model selection criteria include the AIC (Akaike, 1973) and BIC (Schwarz, 1978). These model selection criteria are given by

$$\text{AIC} = -2\log f(\mathbf{X}_N|\hat{\boldsymbol{\Lambda}}_{\text{ML}}, \hat{\boldsymbol{\Psi}}_{\text{ML}}) + 2p^*,$$
$$\text{BIC} = -2\log f(\mathbf{X}_N|\hat{\boldsymbol{\Lambda}}_{\text{ML}}, \hat{\boldsymbol{\Psi}}_{\text{ML}}) + p^*\log N.$$

However, these procedures cannot provide suitable values of regularization parameter ρ since these approaches cannot evaluate models estimated by the maximum penalized likelihood method including the L_1 regularization procedure.

5 Numerical Examples

In this section, we present Monte Carlo simulations to investigate the effectiveness of our modeling strategies. The proposed procedure is also applied to a job application dataset.

5.1 Numerical comparison

In the simulation study, the following two models are used:

Model (A):

$$\boldsymbol{\Lambda} = \begin{pmatrix} & 0 & \\ \boldsymbol{\Lambda}_0^T & 0 & \\ & 0 & \end{pmatrix}^T, \ \boldsymbol{\Psi} = \begin{pmatrix} 0.0975\mathbf{I}_{12} & \mathbf{0}_{12} \\ \mathbf{0}_{12}^T & 1 \end{pmatrix},$$

Model (B):

$$\mathbf{\Lambda} = \begin{pmatrix} & 0 & 0 \\ \mathbf{\Lambda}_0^T & 0 & 0 \\ & 0 & 0 \end{pmatrix}^T, \ \mathbf{\Psi} = \begin{pmatrix} 0.0975\mathbf{I}_{12} & \boldsymbol{O}_{12,2} \\ \boldsymbol{O}_{2,12} & \mathbf{I}_2 \end{pmatrix},$$

where $\mathbf{\Lambda}_0$ is the 12×3 matrix

$$\mathbf{\Lambda}_0^T = \begin{pmatrix} 0.95 & 0.00 & 0.00 & 0.95 & 0.00 & 0.00 & 0.95 & 0.00 & 0.00 & 0.95 & 0.00 & 0.00 \\ 0.00 & 0.95 & 0.00 & 0.00 & 0.95 & 0.00 & 0.00 & 0.95 & 0.00 & 0.00 & 0.95 & 0.00 \\ 0.00 & 0.00 & 0.95 & 0.00 & 0.00 & 0.95 & 0.00 & 0.00 & 0.95 & 0.00 & 0.00 & 0.95 \end{pmatrix}^T$$

and $\boldsymbol{O}_{a,b}$ is $a \times b$ 0-matrix. The number of observations were $N = 50$ and $N = 100$ for each Model, and 100 datasets were generated for each setting. For Model (A), X_{13} is not explained by any common factors, whereas X_{13} and X_{14} are unimportant variables for Model (B).

To investiate the performance of the proposed method, the following three points were examined:

(a) the grouped weighted lasso often yields estimates that can lead to variable selection,

(b) penalized maximum likelihood estimates have smaller mean squared error than maximum likelihood estimates,

(c) the model selection criterion selects the true number of factors frequently.

Firstly, (a) and (b) are investigated. We fixed $k = 3$ and selected the best model by varying the regularization parameter ρ. To investigate (a), we show that how many times the proposed procedures estimated the grouped parameters correctly zero out of 100 datasets. Regarding (b), the sum of squared error of $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ were calculated for each dataset, and these values were averaged over the 100 simulations, which are given by

$$\mathrm{MSE}_{\mathbf{\Lambda}} = \frac{1}{100}\sum_{t=1}^{100} \|\hat{\mathbf{\Lambda}}(t) - \mathbf{\Lambda}\|^2, \ \mathrm{MSE}_{\mathbf{\Psi}} = \frac{1}{100}\sum_{t=1}^{100} \|\hat{\mathbf{\Psi}}(t) - \mathbf{\Psi}\|^2,$$

Table 2: The MSE for parameters $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ and the number of correctly selected models when the number of factors is fixed.

			ML	PML	PML_w
Model(A)	$N = 50$	$\text{MSE}_{\mathbf{\Lambda}}$	0.388	0.376	0.331
		$\text{MSE}_{\mathbf{\Psi}}$	0.021	0.021	0.014
		Correct	0	0	80
	$N = 100$	$\text{MSE}_{\mathbf{\Lambda}}$	0.208	0.198	0.177
		$\text{MSE}_{\mathbf{\Psi}}$	0.008	0.008	0.007
		Correct	0	1	78
Model(B)	$N = 50$	$\text{MSE}_{\mathbf{\Lambda}}$	0.443	0.414	0.323
		$\text{MSE}_{\mathbf{\Psi}}$	0.027	0.026	0.014
		Correct	0	0	73
	$N = 100$	$\text{MSE}_{\mathbf{\Lambda}}$	0.238	0.221	0.180
		$\text{MSE}_{\mathbf{\Psi}}$	0.009	0.009	0.007
		Correct	0	0	71

where $\hat{\mathbf{\Lambda}}(t)$ and $\hat{\mathbf{\Psi}}(t)$ are t-th estimates and $\|\boldsymbol{A}\|$ is the square-root of sum of squares of each element of $\boldsymbol{A}$.

Table 2 shows the MSE for parameters $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ and the number of correctly selected models, in which the column labeled "Correct" presents the number of correctly selected models. The procedures "ML", "PML" and "PML_w" are as follows:

ML: Maximum likelihood procedure

PML: Penalized maximum likelihood procedure with ordinary grouped lasso

PML_w: Penalized maximum likelihood procedure with weighted grouped lasso

For example, for Model (A), the ML procedure selected the correct variables 0 times because it cannot shrink communality exactly zero. In other words, the estimates of $\boldsymbol{\lambda}_{13}$ did not become zero with ML procedure. The PML also selected the correct variables 0 times when $N = 50$ since the hyper-parameter selected by the GBIC was too small. On

Table 3: The number of correctly selected the true number of factors.

		AIC	BIC	GBIC (PML)	GBIC (PML_w)
Model (A)	$N = 50$	73	100	100	100
	$N = 100$	83	100	100	100
Model (B)	$N = 50$	80	100	100	100
	$N = 100$	72	100	100	100

the other hand, the PML_w procedure selected the correct variable 80 times when $N = 50$. When $N = 100$, the PML_w procedure also often selected correct variables, whereas the ordinary lasso selected them only once. Similarly, for Model (B), the PML_w selected the correct variables more than 70 times but the ML and PML never selected correctly. Moreover, the MSE of the PML_w procedure was smaller than that of the ML and PML for both Model (A) and Model (B).

Secondly, we examine (c): selection of the number of factors. For investigating (a) and (b) we fixed the number of factors. On the other hand, for investigating (c), the number of factors is not fixed and we choose k and ρ using the model selection criterion GBIC given by (14). We also selected the number of factors using AIC and BIC, which only deal with the models estimated by the maximum likelihood method, to compare the performance of AIC and BIC with that of GBIC.

Table 3 shows that how many times the model selection criteria selected the true number number of factors out of 100 datasets. For example, the AIC selected the three factor model 73 times out of 100 datasets in model (A) when $N = 50$. For models (A) and (B), the AIC did not always select the true number of factors, whereas the BIC and GBIC chose the true number of factors 100 times, which means BIC-type criteria perform well.

We observed that the proposed procedure of PML_w performed well in the sense that it most often selects the correct variables, and mean squared errors of PML_w was smaller than that of ML and PML. Furthermore, the PML_w selected the number of factors correctly.

5.2 Job application dataset

We illustrate our modeling procedure through a job application dataset in Kendall (1980). This dataset contains 48 applicants for a certain job, who have been scored on $p = 15$ variables regarding their acceptability. The variables are

(1) Form of letter application,	(2) Appearance,	(3) Academic ability,
(4) Likeability,	(5) Self confidence,	(6) Lucidity,
(7) Honesty,	(8) Salesmanship,	(9) Experience,
(10) Drive,	(11) Ambition,	(12) Grasp,
(13) Potential,	(14) Keenness to join,	(15) Suitability.

First, we focus on the selection of the number of factors. The result of AIC, BIC and GBIC are shown in Table 4. The minimum GBIC was selected for varying values of k and ρ.

The model selection criterion AIC and BIC selected 4 factor model and 7 factor model, respectively. Note that the maximum likelihood estimates of some unique variances turned out to be zero or negative for $k \geq 4$. This problem is called improper solutions (see, e.g., van Driel, 1978). The AIC and the BIC selected models that resulted in improper solutions.

Table 5 shows the estimates of $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ obtained by maximum likelihood procedure for 4 factor model. The estimates of factor loadings $\mathbf{\Lambda}$ are rotated by varimax method (Kaiser, 1958). We obtained the improper solutions since the estimates of ψ_{14} turned out to be zero.

Table 4: The number of factors selected by the AIC, BIC and the GBIC, and the variables not selected for each procedure.

	AIC	BIC	GBIC (PML)	GBIC (PML_w)
k	7	4	4	3
variables not selected	–	–	–	X_3

Table 5: The estimates of $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ obtained by maximum likelihood procedure for 4 factor model.

	Factor1	Factor2	Factor3	Factor4	uniqueness
Form of letter application	0.13	0.72	0.11	−0.12	0.44
Appearance	0.45	0.14	0.24	0.16	0.69
Academic ability	0.07	0.12	0.00	0.68	0.52
Likeability	0.23	0.24	0.83	−0.05	0.20
Self confidence	0.92	−0.10	0.15	−0.09	0.11
Lucidity	0.84	0.12	0.30	0.06	0.19
Honesty	0.25	−0.22	0.74	−0.02	0.34
Salesmanship	0.89	0.24	0.08	−0.07	0.13
Experience	0.09	0.78	−0.05	0.17	0.36
Drive	0.77	0.39	0.18	−0.06	0.22
Ambition	0.90	0.18	0.11	−0.06	0.14
Grasp	0.78	0.28	0.36	0.16	0.15
Potential	0.73	0.35	0.44	0.25	0.09
Keenness to join	0.42	0.39	0.56	−0.59	0.00
Suitability	0.36	0.77	0.05	0.14	0.25

Table 6: The estimates of $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ obtained by PMLE$_w$.

	Factor1	Factor2	Factor3	uniqueness
Form of letter application	0.08	0.45	0.10	0.57
Appearance	0.27	0.12	0.17	0.72
Academic ability	0.00	0.00	0.00	1.00
Likeability	0.10	0.17	0.86	0.01
Self confidence	0.79	−0.12	0.11	0.11
Lucidity	0.69	0.08	0.25	0.19
Honesty	0.18	−0.17	0.56	0.45
Salesmanship	0.74	0.18	0.08	0.14
Experience	0.05	0.60	−0.05	0.41
Drive	0.63	0.30	0.12	0.23
Ambition	0.75	0.10	0.10	0.16
Grasp	0.65	0.21	0.24	0.21
Potential	0.58	0.28	0.34	0.19
Keenness to join	0.32	0.19	0.48	0.42
Suitability	0.26	0.70	0.05	0.16

On the other hand, the GBIC for the grouped weighted lasso selected 3 factor model. The estimates of $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ obtained by PMLE$_w$ are given in Table 6. The estimates of factor loadings are rotated by varimax method. We observe that $\hat{\boldsymbol{\lambda}}_3 = \mathbf{0}$ which means we can interpret that the variable X_3 (Academic Ability) is unimportant in constructing the 3 factor model. Moreover, we can obtain the interpretable common factors in the following: *Motivation and Ability*, *Career and Adequacy* and *Character*.

6 Concluding Remarks

We proposed a procedure for variable selection via the L_1 regularization for factor analysis models. Since there are multiple parameters in each variable we treated them

as grouped parameters, then applied the group weighted lasso regularization. In order to select regularization parameters we derived a model selection criterion for evaluating models estimated by the maximum penalized likelihood procedure. The proposed modeling strategy is applied to the analysis of a simulation example, and the proposed procedure selects appropriate variables, produces estimates that have small mean squared error and selects the true number of factors simultaneously. The modeling process is also applied to the Kendall's dataset, and obtained a different interpretation which cannot be obtained by maximum likelihood procedure.

Crucial issues in this modeling procedure include the selection of the weight of the penalty. The proposed weighted penalty performed better than the ordinary lasso, but derived heuristically. As a future research topic, it is interest to propose a new weighted lasso penalty which is derived theoretically.

Appendix A: The derivation of EM algorithm for factor analysis model via the grouped lasso

To apply the EM algorithm into factor analysis model, we consider the common factors $\boldsymbol{f}_n$ to be missing data and maximize the complete-data penalized log-likelihood given by

$$l_\rho^C(\boldsymbol{\theta}) = \sum_{n=1}^{N} \log f(\boldsymbol{x}_n, \boldsymbol{f}_n) - p_\rho(\boldsymbol{\Lambda}),$$

where $f(\boldsymbol{x}_n, \boldsymbol{f}_n)$ is the density of the complete-data distribution and $p_\rho(\boldsymbol{\Lambda})$ is the penalty term in (8).

To derive the posterior mean of the log-likelihood, we use the conditional distribution of common factors $\boldsymbol{F}_n$ given the observed $\boldsymbol{x}_n$. It is well-known that the conditional distribution is given (see, e.g., Anderson, 2003) by

$$\boldsymbol{F}_n | \boldsymbol{x}_n \sim N_k(\boldsymbol{\Lambda}^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \bar{\boldsymbol{x}}), \mathbf{B}), \tag{A1}$$

where $\mathbf{B} = \mathbf{I}_k - \mathbf{\Lambda}^T\mathbf{\Sigma}^{-1}\mathbf{\Lambda}$. Then the values for $E[\boldsymbol{F}_n|\boldsymbol{x}_n]$ and $E[\boldsymbol{F}_n\boldsymbol{F}_n^T|\boldsymbol{x}_n]$ in the E-step are

$$E[\boldsymbol{F}_n|\boldsymbol{x}_n] = \mathbf{\Lambda}^T\mathbf{\Sigma}^{-1}(\boldsymbol{x}_n - \bar{\boldsymbol{x}}), \tag{A2}$$

$$E[\boldsymbol{F}_n\boldsymbol{F}_n^T|\boldsymbol{x}_n] = \mathbf{B} + E[\boldsymbol{F}_n|\boldsymbol{x}_n]E[\boldsymbol{F}_n|\boldsymbol{x}_n]^T. \tag{A3}$$

The expectation of the complete log-likelihood with respect to the distributions of (A1) is

$$E[l_\rho^C(\boldsymbol{\theta})] = -\sum_{n=1}^{N}\Bigg[\frac{p}{2}\log(2\pi) + \frac{p}{2}\log|\mathbf{\Psi}| + \frac{1}{2}\mathrm{tr}\left(E[\boldsymbol{F}_n\boldsymbol{F}_n^T]\right) + \frac{1}{2}\mathrm{tr}\left\{\mathbf{\Psi}^{-1}(\boldsymbol{x}_n - \bar{\boldsymbol{x}})(\boldsymbol{x}_n - \bar{\boldsymbol{x}})^T\right\}$$
$$-E[\boldsymbol{F}_n]^T\mathbf{\Lambda}^T\mathbf{\Psi}^{-1}(\boldsymbol{x}_n - \bar{\boldsymbol{x}}) + \frac{1}{2}\mathrm{tr}(\mathbf{\Lambda}^T\mathbf{\Psi}^{-1}\mathbf{\Lambda}E[\boldsymbol{F}_n\boldsymbol{F}_n^T]) + \rho\sum_{i=1}^{p}\hat{w}_i\|\boldsymbol{\lambda}_i\|\Bigg],$$

where $E[\boldsymbol{F}_n] = E[\boldsymbol{F}_n|\boldsymbol{x}_n]$ and $E[\boldsymbol{F}_n\boldsymbol{F}_n^T] = E[\boldsymbol{F}_n\boldsymbol{F}_n^T|\boldsymbol{x}_n]$.

Since we use the L_1 type penalty, it is difficult to update parameters in the M-step analytically. Therefore, we use the quadratic approximation proposed by Fan & Li (2001). Suppose that $\boldsymbol{\lambda}_{i_0}$ is the current step of $\boldsymbol{\lambda}_i$. If $\boldsymbol{\lambda}_{i_0}$ is very close to $\mathbf{0}$, then set $\boldsymbol{\lambda}_{i_0} = \mathbf{0}$. Otherwise they can be locally approximated by a quadratic functions given by

$$\|\boldsymbol{\lambda}_i\| \approx \|\boldsymbol{\lambda}_{i_0}\| + \frac{1}{2\|\boldsymbol{\lambda}_{i_0}\|}(\boldsymbol{\lambda}_i^T\boldsymbol{\lambda}_i - \boldsymbol{\lambda}_{i_0}^T\boldsymbol{\lambda}_{i_0}) \qquad \text{for } \boldsymbol{\lambda}_{i_0} \neq \mathbf{0}. \tag{A4}$$

The new parameter estimates in the M-step are obtained by maximizing $E[l_\rho^C(\boldsymbol{\theta})]$ with the approximation in (A4) with respect to $\mathbf{\Lambda}$ and $\mathbf{\Psi}$, resulting in

$$\hat{\boldsymbol{\lambda}}_i = \left\{\sum_{n=1}^{N}\left\{E[\boldsymbol{F}_n\boldsymbol{F}_n^T] + (\rho\hat{w}_i/\|\boldsymbol{\lambda}_i\|)\mathbf{I}_k\right\}\right\}^{-1}\left\{\sum_{n=1}^{N}(x_{nj} - \bar{x}_j)E[\boldsymbol{F}_n]^T\right\}, \tag{A5}$$

$$\hat{\mathbf{\Psi}} = \frac{1}{N}\mathrm{Diag}\left[\sum_{n=1}^{N}\left\{(\boldsymbol{x}_n - \bar{\boldsymbol{x}})(\boldsymbol{x}_n - \bar{\boldsymbol{x}})^T - 2(\boldsymbol{x}_n - \bar{\boldsymbol{x}})E[\boldsymbol{F}_n]^T\hat{\mathbf{\Lambda}}^T + \hat{\mathbf{\Lambda}}E[\boldsymbol{F}_n\boldsymbol{F}_n^T]\hat{\mathbf{\Lambda}}^T\right\}\right]. \tag{A6}$$

The updated parameters given by (10) and (11) can be obtained by substituting (A2) and (A3) into (A5) and (A6).

Appendix B: The derivation of the GBIC

Let us consider the problem of selecting a model from a set of candidate models $M_1, \cdots, M_r$. The model M_t $(t = 1, \cdots, r)$ has a probability density $f_t(\boldsymbol{x}|\boldsymbol{\theta}_t)$, and $\boldsymbol{\theta}_t$ has a prior density $\pi_t(\boldsymbol{\theta}_t|\rho_t)$, where ρ_t is a hyper-parameter. The Bayesian procedure for selecting a model is to choose the model with the largest posterior probability given by

$$\begin{aligned}\Pr(M_t|\boldsymbol{x}) &\propto \Pr(M_t) \int f_t(\boldsymbol{x}|\boldsymbol{\theta}_t)\pi_t(\boldsymbol{\theta}_t|\rho_t)d\boldsymbol{\theta}_t \\ &=: \Pr(M_t)p_t(\boldsymbol{x}|\rho_t),\end{aligned}$$

where $\Pr(M_t)$ is the prior probability for model M_t and $p_t(\boldsymbol{x}|\rho_t)$ is the marginal likelihood. If it is assumed that the prior probability $\Pr(M_t)$ is the same for all models, it follows that the model that maximizes the marginal likelihood $p_t(\boldsymbol{x}|\rho_t)$ of the data must be selected.

In the factor analysis model via the grouped weighted lasso, the prior distribution $\pi(\boldsymbol{\theta}|\rho)$ is given by

$$\pi_\rho(\boldsymbol{\Lambda}) = C \prod_{i=1}^{p} \exp\left(-N\rho\hat{w}_i\|\boldsymbol{\lambda}_i\|\right), \tag{B1}$$

where C is the normalizing constant given by

$$C = (C_k)^{p-k} \prod_{j=1}^{k-1} C_j, \quad C_j = \frac{(N\rho\hat{w}_j)^j}{2^j \pi^{\frac{j-1}{2}} \Gamma\left(\frac{j+1}{2}\right)} \tag{B2}$$

with $\Gamma(\cdot)$ being the Gamma function.

The posterior distribution is then given by

$$\begin{aligned}\pi(\boldsymbol{\Lambda}, \boldsymbol{\Psi}|\boldsymbol{X}_N) &= \frac{f(\boldsymbol{X}_N|\boldsymbol{\Lambda}, \boldsymbol{\Psi})\pi_\rho(\boldsymbol{\Lambda})}{\int\int f(\boldsymbol{X}_N|\boldsymbol{\Lambda}, \boldsymbol{\Psi})\pi_\rho(\boldsymbol{\Lambda})d\boldsymbol{\Lambda}d\boldsymbol{\Psi}} \\ &\propto f(\boldsymbol{X}_N|\boldsymbol{\Lambda}, \boldsymbol{\Psi})\pi_\rho(\boldsymbol{\Lambda}).\end{aligned} \tag{B3}$$

In a Bayesian framework the parameters $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ are estimated through mode of the posterior distribution. It is equivalent to obtain estimates by maximizing the penalized log-likelihood function in (5) with the hyper-parameter ρ which can be considered as a regularization parameter.

The model selection criterion GBIC (Konishi *et al.*, 2004) is obtained by minimizing $-2\log p_t(\boldsymbol{x}|\rho_t)$ with the use of Laplace approximation (Tierney & Kadane, 1986). Since the model selection criterion GBIC cannot be directly derived for the L_1 type regularization method, we define an Active set in (13) and derive a second order differential of the penalized log-likelihood of the Active set $\boldsymbol{\theta}_{\mathcal{A}}$ instead of $\boldsymbol{\theta}$. Then the GBIC is written as follows (see Equation (10) in Konishi *et al.*, 2004):

$$\text{GBIC} = -p^*\log(2\pi) + p^*\log N + \log|J_\rho(\hat{\boldsymbol{\theta}}_{\mathcal{A}})| - 2\left\{\log f(\mathbf{X}_N|\hat{\boldsymbol{\theta}}) + \log\pi_\rho(\hat{\boldsymbol{\Lambda}})\right\}. \quad \text{(B4)}$$

By substituting (12) and (B1) into (B4), we obtain a model selection criterion for evaluating the factor analysis model via the grouped lasso given by (14). The matrix $J_\rho(\boldsymbol{\theta})$ consists of the elements of $\frac{\partial^2 l_\rho(\boldsymbol{\theta})}{\partial\lambda_{ab}\partial\lambda_{cd}}$, $\frac{\partial^2 l_\rho(\boldsymbol{\theta})}{\partial\psi_i\partial\lambda_{cd}}$ and $\frac{\partial^2 l_\rho(\boldsymbol{\theta})}{\partial\psi_i\partial\psi_j}$, which are given by

$$\begin{aligned}
\frac{\partial^2 l_\rho(\boldsymbol{\theta})}{\partial\lambda_{ab}\partial\lambda_{cd}} = N\Big\{ & (\boldsymbol{\Sigma}^{-1})_{ac}(\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})_{bd} + (\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})_{ad}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})_{cb} \\
& -(\boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1})_{ac}(\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})_{bd} - (\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})_{ad}(\boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})_{cb} \\
& -(\boldsymbol{\Sigma}^{-1})_{ac}(\boldsymbol{\Lambda}^T\boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})_{bd} - (\boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})_{ad}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})_{cb} \\
& -(\boldsymbol{\Sigma}^{-1})_{ac}(\mathbf{I}_k)_{bd} + (\boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1})_{ac}(\mathbf{I}_k)_{bd}\Big\} \\
& +\rho\hat{w}_c\frac{(\boldsymbol{\Lambda})_{(a,b)}(\boldsymbol{\Lambda})_{(c,d)}(\mathbf{I}_p)_{(a,c)}}{\sqrt{(\boldsymbol{\Lambda}\boldsymbol{\Lambda}')_{(c,c)}}^3} - \rho\hat{w}_c\frac{(\mathbf{I}_p)_{(a,c)}(\mathbf{I}_k)_{(b,d)}}{\sqrt{(\boldsymbol{\Lambda}\boldsymbol{\Lambda}')_{(c,c)}}}\Big\}, \\
\frac{\partial^2 l_\rho(\boldsymbol{\theta})}{\partial\psi_i\partial\lambda_{cd}} = N\Big\{ & (\boldsymbol{\Sigma}^{-1})_{ci}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})_{id} - (\boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1})_{ci}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})_{id} \\
& -(\boldsymbol{\Sigma}^{-1})_{ci}(\boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})_{id}\Big\}, \\
\frac{\partial^2 l_\rho(\boldsymbol{\theta})}{\partial\psi_i\partial\psi_j} = \frac{N}{2}\Big\{ & (\boldsymbol{\Sigma}^{-1})_{ij}^2 - 2(\boldsymbol{\Sigma}^{-1})_{ij}(\boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1})_{ij}\Big\},
\end{aligned}$$

where $(\mathbf{A})_{\alpha\beta}$ is a (α,β)-th element of a matrix $\mathbf{A}$.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *2nd International Symposium on Information Theory* (Petrov and, B. N.

and Csaki, F., eds.), Akademiai Kiado, 267–281. (Reproduced in *Breakthroughs in Statistics* **1**, S. Kotz and N. L. Johnson eds., Foundations and Basic Theory, Springer–Verlag, (1992) 610–624.)

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

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis* (3rd ed.). Hoboken, N.J.: Wiley.

Anderson, T. W., & Rubin, H. (1956). Statistical inference in factor analysis. In J. Neyman (Ed.), *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, **5**, (pp. 111–150). Berkeley: University of California Press.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Annals of Statistics*, **24**, 2350–2383.

Clarke, M. R. B. (1970). A rapidly convergent method for maximum-likelihood factor analysis. *British Journal of Mathematical and Statistical Psychology*, **23**, 43–52.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression (with discussion). *Annals of Statistics*, **32**, 407–499.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**, 1348–1360.

Gorsuch, R. L. (1988). Exploratory factor analysis. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (2nd ed., pp. 231-258.) New York and London: Plenum Press.

Ichikawa, M., & Konishi, S. (1999). Model evaluation and information criteria in covariance structure analysis. *British Journal of Mathematical and Statistical Psychology*, **52**, 285–302.

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

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psy-*

chometrika, **32**, 443–482.

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

Kano, Y., & Ihara, M. (1994). Identification of inconsistent variates in factor analysis. *Psychometrika*, **59**, 5–20.

Kano, Y., & Harada, A. (2000). Stepwise variable selection in factor analysis. *Psychometrika*, **65**, 7–22.

Kendall, M. G. (1980). *Multivariate Analysis* (2nd.ed.). London: Charles Griffin.

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

Konishi, S., & Kitagawa, G. (2008). *Information Criteria and Statistical Modeling*. New York: Springer.

Lawley, D. N., & Maxwell, A. E. (1971). *Factor Analysis as a Statistical Method* (2nd ed.). London: Butterworths.

Rubin, D. B., & Thayer, D. T. (1982). EM algorithms for ML factor analysis. *Psychometrika*, **47**, 69–76.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Shimamura, T., Imoto, S., Yamaguchi, R., & Miyano, S. (2007). Weighted lasso in graphical Gaussian modeling for large gene network estimation based on microarray data. *Genome Informatics*, **19**, 142–153.

Tanaka, Y. (1983). Some criteria for variable selection in factor analysis. *Behaviormetrika*, **13**, 31–45.

Tateishi, S., Matsui, H., & Konishi, S. (2010). Nonlinear regression modeling via the lasso-type regularization. *Journal of Statistical Planning and Inference*, **140**, 1125–1134.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journals of the Royal Statistical Society Series B*, **58**, 267–288.

Tierny, L., & Kanade, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, **81**, 82–86.

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

Yanai, H. (1980). A proposition of generalized method for forward selection of variables. *Behaviormetrika*, **7**, 95–107.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journals of the Royal Statistical Society Series B*, **68**, 49–67.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journals of the Royal Statistical Society Series B*, **67**, 301–320.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, **101**, 1418–1429.

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

List of MI Preprint Series, Kyushu University

The Global COE Program
Math-for-Industry Education & Research Hub

MI

MI2008-1 Takahiro ITO, Shuichi INOKUCHI & Yoshihiro MIZOGUCHI
Abstract collision systems simulated by cellular automata

MI2008-2 Eiji ONODERA
The intial value problem for a third-order dispersive flow into compact almost Hermitian manifolds

MI2008-3 Hiroaki KIDO
On isosceles sets in the 4-dimensional Euclidean space

MI2008-4 Hirofumi NOTSU
Numerical computations of cavity flow problems by a pressure stabilized characteristic-curve finite element scheme

MI2008-5 Yoshiyasu OZEKI
Torsion points of abelian varieties with values in nfinite extensions over a p-adic field

MI2008-6 Yoshiyuki TOMIYAMA
Lifting Galois representations over arbitrary number fields

MI2008-7 Takehiro HIROTSU & Setsuo TANIGUCHI
The random walk model revisited

MI2008-8 Silvia GANDY, Masaaki KANNO, Hirokazu ANAI & Kazuhiro YOKOYAMA
Optimizing a particular real root of a polynomial by a special cylindrical algebraic decomposition

MI2008-9 Kazufumi KIMOTO, Sho MATSUMOTO & Masato WAKAYAMA
Alpha-determinant cyclic modules and Jacobi polynomials

MI2008-10 Sangyeol LEE & Hiroki MASUDA
Jarque-Bera Normality Test for the Driving Lévy Process of a Discretely Observed Univariate SDE

MI2008-11 Hiroyuki CHIHARA & Eiji ONODERA
A third order dispersive flow for closed curves into almost Hermitian manifolds

MI2008-12 Takehiko KINOSHITA, Kouji HASHIMOTO and Mitsuhiro T. NAKAO
On the L^2 a priori error estimates to the finite element solution of elliptic problems with singular adjoint operator

MI2008-13 Jacques FARAUT and Masato WAKAYAMA
Hermitian symmetric spaces of tube type and multivariate Meixner-Pollaczek polynomials

MI2008-14 Takashi NAKAMURA
Riemann zeta-values, Euler polynomials and the best constant of Sobolev inequality

MI2008-15 Takashi NAKAMURA
Some topics related to Hurwitz-Lerch zeta functions

MI2009-1 Yasuhide FUKUMOTO
Global time evolution of viscous vortex rings

MI2009-2 Hidetoshi MATSUI & Sadanori KONISHI
Regularized functional regression modeling for functional response and predictors

MI2009-3 Hidetoshi MATSUI & Sadanori KONISHI
Variable selection for functional regression model via the L_1 regularization

MI2009-4 Shuichi KAWANO & Sadanori KONISHI
Nonlinear logistic discrimination via regularized Gaussian basis expansions

MI2009-5 Toshiro HIRANOUCHI & Yuichiro TAGUCHII
Flat modules and Groebner bases over truncated discrete valuation rings

MI2009-6 Kenji KAJIWARA & Yasuhiro OHTA
Bilinearization and Casorati determinant solutions to non-autonomous 1+1 dimensional discrete soliton equations

MI2009-7 Yoshiyuki KAGEI
Asymptotic behavior of solutions of the compressible Navier-Stokes equation around the plane Couette flow

MI2009-8 Shohei TATEISHI, Hidetoshi MATSUI & Sadanori KONISHI
Nonlinear regression modeling via the lasso-type regularization

MI2009-9 Takeshi TAKAISHI & Masato KIMURA
Phase field model for mode III crack growth in two dimensional elasticity

MI2009-10 Shingo SAITO
Generalisation of Mack's formula for claims reserving with arbitrary exponents for the variance assumption

MI2009-11 Kenji KAJIWARA, Masanobu KANEKO, Atsushi NOBE & Teruhisa TSUDA
Ultradiscretization of a solvable two-dimensional chaotic map associated with the Hesse cubic curve

MI2009-12 Tetsu MASUDA
Hypergeometric τ-functions of the q-Painlevé system of type $E_8^{(1)}$

MI2009-13 Hidenao IWANE, Hitoshi YANAMI, Hirokazu ANAI & Kazuhiro YOKOYAMA
A Practical Implementation of a Symbolic-Numeric Cylindrical Algebraic Decomposition for Quantifier Elimination

MI2009-14 Yasunori MAEKAWA
On Gaussian decay estimates of solutions to some linear elliptic equations and its applications

MI2009-15 Yuya ISHIHARA & Yoshiyuki KAGEI
Large time behavior of the semigroup on L^p spaces associated with the linearized compressible Navier-Stokes equation in a cylindrical domain

MI2009-16 Chikashi ARITA, Atsuo KUNIBA, Kazumitsu SAKAI & Tsuyoshi SAWABE
Spectrum in multi-species asymmetric simple exclusion process on a ring

MI2009-17 Masato WAKAYAMA & Keitaro YAMAMOTO
Non-linear algebraic differential equations satisfied by certain family of elliptic functions

MI2009-18 Me Me NAING & Yasuhide FUKUMOTO
Local Instability of an Elliptical Flow Subjected to a Coriolis Force

MI2009-19 Mitsunori KAYANO & Sadanori KONISHI
Sparse functional principal component analysis via regularized basis expansions and its application

MI2009-20 Shuichi KAWANO & Sadanori KONISHI
Semi-supervised logistic discrimination via regularized Gaussian basis expansions

MI2009-21 Hiroshi YOSHIDA, Yoshihiro MIWA & Masanobu KANEKO
Elliptic curves and Fibonacci numbers arising from Lindenmayer system with symbolic computations

MI2009-22 Eiji ONODERA
A remark on the global existence of a third order dispersive flow into locally Hermitian symmetric spaces

MI2009-23 Stjepan LUGOMER & Yasuhide FUKUMOTO
Generation of ribbons, helicoids and complex scherk surface in laser-matter Interactions

MI2009-24 Yu KAWAKAMI
Recent progress in value distribution of the hyperbolic Gauss map

MI2009-25 Takehiko KINOSHITA & Mitsuhiro T. NAKAO
On very accurate enclosure of the optimal constant in the a priori error estimates for H_0^2-projection

MI2009-26 Manabu YOSHIDA
Ramification of local fields and Fontaine's property (Pm)

MI2009-27 Yu KAWAKAMI
Value distribution of the hyperbolic Gauss maps for flat fronts in hyperbolic three-space

MI2009-28 Masahisa TABATA
Numerical simulation of fluid movement in an hourglass by an energy-stable finite element scheme

MI2009-29 Yoshiyuki KAGEI & Yasunori MAEKAWA
Asymptotic behaviors of solutions to evolution equations in the presence of translation and scaling invariance

MI2009-30 Yoshiyuki KAGEI & Yasunori MAEKAWA
On asymptotic behaviors of solutions to parabolic systems modelling chemotaxis

MI2009-31 Masato WAKAYAMA & Yoshinori YAMASAKI
Hecke's zeros and higher depth determinants

MI2009-32 Olivier PIRONNEAU & Masahisa TABATA
Stability and convergence of a Galerkin-characteristics finite element scheme of lumped mass type

MI2009-33 Chikashi ARITA
Queueing process with excluded-volume effect

MI2009-34 Kenji KAJIWARA, Nobutaka NAKAZONO & Teruhisa TSUDA
Projective reduction of the discrete Painlevé system of type$(A_2 + A_1)^{(1)}$

MI2009-35 Yosuke MIZUYAMA, Takamasa SHINDE, Masahisa TABATA & Daisuke TAGAMI
Finite element computation for scattering problems of micro-hologram using DtN map

MI2009-36 Reiichiro KAWAI & Hiroki MASUDA
Exact simulation of finite variation tempered stable Ornstein-Uhlenbeck processes

MI2009-37 Hiroki MASUDA
On statistical aspects in calibrating a geometric skewed stable asset price model

MI2010-1 Hiroki MASUDA
Approximate self-weighted LAD estimation of discretely observed ergodic Ornstein-Uhlenbeck processes

MI2010-2 Reiichiro KAWAI & Hiroki MASUDA
Infinite variation tempered stable Ornstein-Uhlenbeck processes with discrete observations

MI2010-3 Kei HIROSE, Shuichi KAWANO, Daisuke MIIKE & Sadanori KONISHI
Hyper-parameter selection in Bayesian structural equation models

MI2010-4 Nobuyuki IKEDA & Setsuo TANIGUCHI
The Itô-Nisio theorem, quadratic Wiener functionals, and 1-solitons

MI2010-5 Shohei TATEISHI & Sadanori KONISHI
Nonlinear regression modeling and detecting change point via the relevance vector machine

MI2010-6 Shuichi KAWANO, Toshihiro MISUMI & Sadanori KONISHI
Semi-supervised logistic discrimination via graph-based regularization

MI2010-7 Teruhisa TSUDA
UC hierarchy and monodromy preserving deformation

MI2010-8 Takahiro ITO
Abstract collision systems on groups

MI2010-9 Hiroshi YOSHIDA, Kinji KIMURA, Naoki YOSHIDA, Junko TANAKA & Yoshihiro MIWA
An algebraic approach to underdetermined experiments

MI2010-10 Kei HIROSE & Sadanori KONISHI
Variable selection via the grouped weighted lasso for factor analysis models