Open Journal of Statistics, **2013, 3, 27-35**
http://dx.doi.org/10.4236/ojs.2013.31004 Published Online February 2013 (http://www.scirp.org/journal/ojs)

Joint Variable Selection of Mean-Covariance Model for Longitudinal Data

Dengke Xu[1], Zhongzhan Zhang[1], Liucang Wu[1,2]
[1]College of Applied Sciences, Beijing University of Technology, Beijing, China
[2]Faculty of Science, Kunming University of Science and Technology, Kunming, China
Email: zzhang@bjut.edu.cn

Received October 23, 2012; revised November 24, 2012; accepted December 9, 2012

ABSTRACT

In this paper we reparameterize covariance structures in longitudinal data analysis through the modified Cholesky decomposition of itself. Based on this modified Cholesky decomposition, the within-subject covariance matrix is decomposed into a unit lower triangular matrix involving moving average coefficients and a diagonal matrix involving innovation variances, which are modeled as linear functions of covariates. Then, we propose a penalized maximum likelihood method for variable selection in joint mean and covariance models based on this decomposition. Under certain regularity conditions, we establish the consistency and asymptotic normality of the penalized maximum likelihood estimators of parameters in the models. Simulation studies are undertaken to assess the finite sample performance of the proposed variable selection procedure.

Keywords: Joint Mean and Covariance Models; Variable Selection; Cholesky Decomposition; Longitudinal Data; Penalized Maximum Likelihood Method

1. Introduction

In recent years, the method of joint modeling of mean and covariance on the general linear model with multivariate normal errors, was heuristically introduced by Pourahmadi [1,2]. The key advantages of such models include the convenience in statistical interpretation and computational ease in parameter estimation, which is described in Section 2. On the other hand, the estimation of the covariance matrix is important in a longitudinal study. A good estimator for the covariance can improve the efficiency of the regression coefficients. Furthermore, the covariance estimation itself is also of interest [3]. A number of authors have studied the problem of estimating the covariance matrix. Pourahmadi [1,2] considered generalized linear models for the components of the modified Cholesky decomposition of the covariance matrix. Fan *et al.* [4] and Fan and Wu [5] proposed to use a semiparametric model for the covariance function. Recently, Rothman *et al.* [6] proposed a new regression interpretation of the Cholesky factor of the covariance matrix by parameterizing itself and guaranteed the positive-definiteness of the estimated covariance at no additional computational cost. Furthermore, based on this decomposition [6], Zhang and Leng [7] proposed efficient maximum likelihood estimates for joint mean-covariance analysis.

As is well known, as a part of modeling strategy, variable selection is an important topic in most statistical analysis, and has been extensively explored over the last three decades. In a traditional linear regression setting, many selection criteria (e.g., AIC and BIC) have been extensively used in practice. Nevertheless, those selection methods suffer from expensive computational costs. As computational efficiency is more desirable in many situations, various shrinkage methods have been developed, which include but are not limited to: the nonnegative garrotte [8], the LASSO [9], the bridge regression [10], the SCAD [11], and the one-step sparse estimator [12]. Recently, Zhang and Wang [13] proposed a new criterion, named PICa, to simultaneously select explanatory variables in the mean model and variance model in heteroscedastic linear models based on the model structure. Zhao and Xue [14] presented a variable selection procedure by using basis function approximations and a partial group SCAD penalty for semiparametric varying coefficient partially linear models with longitudinal data.

In this paper we show that the modified Cholesky decomposition of the covariance matrix, rather than its inverse, also has a natural regression interpretation, and therefore all Cholesky-based regularization methods can be applied to the covariance matrix itself instead of its inverse to obtain a sparse estimator with guaranteed positive definiteness. Furthermore, we aim to develop an

efficient penalized likelihood based method to select important explanatory variables that make a significant contribution to the joint modelling of mean and covariance structures for longitudinal data. With proper choices of the penalty functions and the tuning parameters, we establish the consistency and asymptotic normality of the resulting estimator. Simulation studies are used to illustrate the proposed methodologies. Compared with existing methods, our procedure offers the following differences and improvements. Firstly, Zhang and Leng [7] discussed efficient maximum likelihood estimates and model selection for joint mean-covariance analysis based BIC. As is well known, BIC selection method would suffer from expensive computational costs. However, our method can select significant variables and obtain the parameter estimators simultaneously in the joint modelling of mean and covariance structures for longitudinal data, that implies that our method can avoid the heavy computational burden. Secondly, in this paper we assume the covariates may be of high dimension, which become increasingly common in many health studies, and our method also can select the important subsets of the covatiates. Thirdly, we reparameterize covariance structures in longitudinal data analysis through the modified Cholesky decomposition of itself, which is brought closer to time series analysis, for which the moving average model may provide an alternative, equally powerful and parsimonious representation.

The rest of this paper is organized as follows. In Section 2 we first describe a reparameterization of covariance matrix itself through the modified Cholesky decomposition and introduce the joint mean and covariance models for longitudinal data. We then propose a variable selection method for the joint models via penalized likelihood function. Asymptotic properties of the resulting estimators are considered in Section 3. In Section 4 we give the computation of the penalized likelihood estimator as well as the choice of the tuning parameters. In Section 5 we carry out simulation studies to assess the finite sample performance of the method.

2. Variable Selection for Joint Mean-Covariance Model

2.1. Modified Cholesky Decomposition of the Covariance Matrix

Suppose that there are *n* independent subjects and the *i*th subject has m_i repeated measurements. Specifically, denote the response vector $y_i = \left(y_{i1}, \cdots, y_{im_i}\right)^{\mathrm{T}}$ for the *i*th subject, $i = 1, \cdots, n$, which are observed at time $t_i = \left(t_{i1}, \cdots, t_{im_i}\right)^{\mathrm{T}}$. We assume that the response vector is normally distributed as $y_i \sim N\left(\mu_i, \Sigma_i\right)$, where $\mu_i = \left(\mu_{i1}, \cdots, \mu_{im_i}\right)^{\mathrm{T}}$ is an $\left(m_i \times 1\right)$ vector and Σ_i is an $\left(m_i \times m_i\right)$ positive definite matrix $\left(i = 1, \cdots, n\right)$. As a tool for regularizing the inverse covariance matrix, Pourahmadi [1] suggested using the modified Cholesky factorization of Σ_i^{-1}. To parametrize Σ_i, Pourahmadi [1] first proposed to decompose it as $T_i \Sigma_i T_i^{\mathrm{T}} = D_i$. The lower triangular matrix T_i is unique with 1's on its diagonal and the below diagonal entries of T_i are the negative autoregressive parameters ϕ_{ijk} in the model

$$y_{ij} - \mu_{ij} = \sum_{k=1}^{j-1} \phi_{ijk}\left(y_{ik} - \mu_{ik}\right) + \varepsilon_{ij}.$$

The diagonal entries of D_i are the innovation variances as $\sigma_{ij}^2 = \mathrm{Var}\left(\varepsilon_{ij}\right)$.

According to the idea of the proposed decomposition in Rothman *et al.* [6], we let $L_i = T_i^{-1}$, a lower triangular matrix with 1's on its diagonal, we can write $\Sigma_i = T_i D_i T_i^{\mathrm{T}}$. We actually use a new statistically meaningful representation that reparameterizes the covariance matrices by the modified Cholesky decomposition advocated by Rothman *et al*. [6]. The entries l_{ijk} in L_i can be interpreted as the moving average coefficients in

$$y_{ij} - \mu_{ij} = \sum_{k=1}^{j-1} l_{ijk}\varepsilon_{ik} + \varepsilon_{ij},\ j = 2, \cdots, m_i,$$

where $\varepsilon_{i1} = y_{i1} - \mu_{i1}$ and $\varepsilon_i \sim N\left(0, D_i\right)$ for $\varepsilon_i = \left(\varepsilon_{i1}, \cdots, \varepsilon_{im_i}\right)^{\mathrm{T}}$. Note that the parameters l_{ijk} and $\log\left(\sigma_{ij}^2\right)$ are unconstrained.

Based on the modified Cholesky decomposition and motivated by [1,2] and Ye and Pan [15], the unconstrained parameters μ_{ij}, l_{ijk} and $\log\left(\sigma_{ij}^2\right)$ are modeled in terms of the generalized linear regression models (JMVGLRM)

$$g\left(\mu_{ij}\right) = x_{ij}^{\mathrm{T}}\beta, l_{ijk} = z_{ijk}^{\mathrm{T}}\gamma, \log\left(\sigma_{ij}^2\right) = h_{ij}^{\mathrm{T}}\lambda. \quad (1)$$

Here $g(\cdot)$ is a monotone and differentiable known link function, and x_{ij}, z_{ijk} and h_{ij} are the $p \times 1$, $q \times 1$ and $d \times 1$ vectors of covariates, respectively. The covariate x_{ij} and h_{ij} are the usual covariates used in regression analysis, while z_{ijk} is usually taken as a polynomial of time difference $t_{ik} - t_{ij}$. In addition, denote $X_i = \left(x_{i1}, \cdots, x_{im_i}\right)^{\mathrm{T}}$ and $H_i = \left(h_{i1}, \cdots, h_{im_i}\right)^{\mathrm{T}}$. We further refer to γ as moving average coefficients and λ as innovation coefficients. In this paper we assume that the covariates x_{ij}, z_{ijk} and h_{ij} may be of high dimension and we would select the important subsets of the covariates x_{ij}, z_{ijk} and h_{ij}, simultaneously. We first assume all the explanatory variables of interest, and perhaps their interactions as well, are already included into the initial models. Then, we aim to remove the unnecessary explanatory variables from the models.

2.2. Penalized Maximum Likelihood for JMVGLRM

Many traditional variable selection criteria can be considered as a penalized likelihood which balances modelling biases and estimation variances [11]. Let $\ell(\theta)$ denote the log-likelihood function. For the JMVGLRM, we propose the penalized likelihood function

$$L(\theta)=\ell(\theta)-\sum_{i=1}^{p}p_{\tau^{(1)}}\left(|\beta_i|\right)-\sum_{j=1}^{q}p_{\tau^{(2)}}\left(|\gamma_j|\right)-\sum_{k=1}^{d}p_{\tau^{(3)}}\left(|\lambda_k|\right) \tag{2}$$

where $\theta=(\theta_1,\cdots,\theta_s)^{\mathrm{T}}=(\beta_1,\cdots,\beta_p;\gamma_1,\cdots,\gamma_q;\lambda_1,\cdots,\lambda_d)^{\mathrm{T}}$ with $s=p+q+d$ and $p_{\tau^{(l)}}(\cdot)$ is a given penalty function with the tuning parameter $\tau^{(l)}(l=1,2,3)$. The tuning parameters can be chosen by a data-driven criterion such as cross validation (CV), generalized cross-validation (GCV) [9], or the BIC-type tuning parameter selector [16] which is described in Section 4. Here we use the same penalty function $p(\cdot)$ for all the regression coefficients but with different tuning parameters $\tau^{(1)}$, $\tau^{(2)}$ and $\tau^{(3)}$ for the mean parameters, moving average parameters and log-innovation variances, respectively. Note that the penalty functions and tuning parameters are not necessarily the same for all the parameters. For example, we wish to keep some important variables in the final model and therefore do not want to penalize their coefficients. In this paper, we use the smoothly clipped absolute deviation (SCAD) penalty whose first derivative satisfies

$$p'_{\lambda}(\cdot)=\lambda\left\{I(t\le\lambda)+\frac{(a\lambda-t)_{+}}{(a-1)\lambda}I(t>\lambda)\right\}$$

for some $a>2$ [11]. Following the convention in [11], we set $a=3.7$ in our work. The SCAD penalty is a spline function on an interval near zero and constant outside, so that it can shrink small value of an estimate to zero while having no impact on a large one.

The penalized maximum likelihood estimator of θ, denoted by $\hat{\theta}$, maximizes the function $L(\theta)$ in (2). With appropriate penalty functions, maximizing $L(\theta)$ with respect to θ leads to certain parameter estimators vanishing from the initial models so that the corresponding explanatory variables are automatically removed. Hence, through maximizing $L(\theta)$ we achieve the goal of selecting important variables and obtaining the parameter estimators, simultaneously. In Section 4, we provide the technical details and an algorithm for calculating the penalized maximum likelihood estimator $\hat{\theta}$.

3. Asymptotic Properties

We next study the asymptotic properties of the resulting penalized likelihood estimate. We first introduce some notations. Let θ_0 denote the true values of θ. Furthermore, let

$$\theta_0=(\theta_{01},\cdots,\theta_{0s})^{\mathrm{T}}=\left(\left(\theta_0^{(1)}\right)^{\mathrm{T}},\left(\theta_0^{(2)}\right)^{\mathrm{T}}\right)^{\mathrm{T}}.$$

For ease of presentation and without loss of generality, it is assumed that $\theta_0^{(1)}$ consists of all nonzero components of θ_0 and that $\theta_0^{(2)}=0$. Denote the dimension of $\theta_0^{(1)}$ by s_1. Let

$$a_n=\max_{1\le j\le s}\left\{\left|p'_{\tau_n}\left(|\theta_{0j}|\right)\right|:\theta_{0j}\ne 0\right\}$$

and

$$b_n=\max_{1\le j\le s}\left\{p''_{\tau_n}\left(|\theta_{0j}|\right):\theta_{0j}\ne 0\right\}.$$

Here we denote $\tau^{(l)}$ as τ_n to emphasize its dependence on sample size n. τ_n is equal to either $\tau^{(1)}$, $\tau^{(2)}$ or $\tau^{(3)}$, depending on whether θ_{0j} is a component of β_0, γ_0 or $\lambda_0(1\le j\le s)$.

To obtain the asymptotic properties in the paper, we require the following regularity conditions:

(C1): The covariate vectors x_{ij}, z_{ijk} and h_{ij} are fixed. Also, for each subject the number of repeated measurements, m_i, is fixed

$$(i=1,\cdots,n,\ j=1,\cdots,m_i,\ k=1,\cdots,j-1).$$

(C2): The parameter space is compact and the true value θ_0 is in the interior of the parameter space.

(C3): The design matrices X_i and H_i in the joint models are all bounded, meaning that all the elements of the matrices are bounded by a single finite real number.

Theorem 1 Assume $a_n=O_p\left(n^{-1/2}\right)$, $b_n\to 0$ and $\tau_n\to 0$ as $n\to\infty$. Under the conditions (C1)-(C3), with probability tending to 1 there must exist a local maximizer $\hat{\theta}_n$ of the penalized likelihood function $L(\theta)$ in (2) such that $\hat{\theta}_n$ is a $\sqrt{n}$-consistent estimator of θ_0.

The following theorem gives the asymptotic normality property of $\hat{\theta}_n$. Let

$$A_n=\mathrm{diag}\left(p''_{\tau_n}\left(\left|\theta_{01}^{(1)}\right|\right),\cdots,p''_{\tau_n}\left(\left|\theta_{0s_1}^{(1)}\right|\right)\right)$$

$$c_n=\left(p'_{\tau_n}\left(\left|\theta_{01}^{(1)}\right|\right)\mathrm{sgn}\left(\theta_{01}^{(1)}\right),\cdots,p'_{\tau_n}\left(\left|\theta_{0s_1}^{(1)}\right|\right)\mathrm{sgn}\left(\theta_{0s_1}^{(1)}\right)\right)^{\mathrm{T}},$$

where $\theta_{0j}^{(1)}$ is the jth component of $\theta_0^{(1)}(1\le j\le s_1)$.

Denote the Fisher information matrix of θ by $I_n(\theta)$.

Theorem 2 Assume that the penalty function $p_{\tau_n}(t)$ satisfies

$$\liminf_{n\to\infty}\liminf_{t\to 0^{+}}\frac{p'_{\tau_n}(t)}{\tau_n}>0$$

and $\bar{I}_n=I_n(\theta_0)/n$ converges to a finite and positive

definite matrix $I(\theta_0)$ as $n \to \infty$. Under the same mild conditions as these given in Theorem 1, if $\tau_n \to 0$ and $\sqrt{n}\tau_n \to \infty$ as $n \to \infty$, then the $\sqrt{n}$-consistent estimator $\hat{\theta}_n = \left(\left(\hat{\theta}_n^{(1)}\right)^{\mathrm{T}}, \left(\hat{\theta}_n^{(2)}\right)^{\mathrm{T}} \right)^{\mathrm{T}}$ in Theorem 1 must satisfy

1) $\hat{\theta}_n^{(2)} = 0$ with probability tending to 1.

2) $\sqrt{n}\left(\bar{I}_n^{(1)}\right)^{-1/2}\left(\bar{I}_n^{(1)} + A_n\right)\left\{\left(\hat{\theta}_n^{(1)} - \theta_0^{(1)}\right) + \left(\bar{I}_n^{(1)} + A_n\right)^{-1} c_n\right\} \xrightarrow{L} N_{s_1}\left(0, I_{s_1}\right)$,

where "$\xrightarrow{L}$" stands for the convergence in distribution; $\bar{I}_n^{(1)}$ is the $(s_1 \times s_1)$ submatrix of $\bar{I}_n$ corresponding to the nonzero components $\theta_0^{(1)}$ and I_{s_1} is the $(s_1 \times s_1)$ identity matrix.

Remark: The proofs of the Theorems 1 and 2 are similar to [11]. To save space, the proofs are omitted.

4. Computation

4.1. Algorithm

Because $L(\theta)$ is irregular at the origin, the commonly used gradient method is not applicable. Now, we develop an iterative algorithm based on the local quadratic approximation of the penalty function $p_\tau(\cdot)$ as in [11].

Firstly, note the first two derivatives of the log-likelihood function $\ell(\theta)$ are continuous. Around a given point θ_0, the log-likelihood function can be approximated by

$$\ell(\theta) \approx \ell(\theta_0) + \left[\frac{\partial \ell(\theta_0)}{\partial \theta}\right]^{\mathrm{T}}(\theta - \theta_0) + \frac{1}{2}(\theta - \theta_0)^{\mathrm{T}}\left[\frac{\partial^2 \ell(\theta_0)}{\partial \theta \partial \theta^{\mathrm{T}}}\right](\theta - \theta_0).$$

Also, given an initial value t_0 we can approximate the penalty function $p_\tau'(t)$ by a quadratic function [11]

$$p_\tau(|t|) \approx p_\tau(|t_0|) + \frac{1}{2}\frac{p_\tau'(|t_0|)}{|t_0|}\left(t^2 - t_0^2\right),$$

for $t \approx t_0$.

Therefore, the penalized likelihood function (2) can be local approximated, apart from a constant term, by

$$L(\theta) \approx \ell(\theta_0) + \left[\frac{\partial \ell(\theta_0)}{\partial \theta}\right]^{\mathrm{T}}(\theta - \theta_0) + \frac{1}{2}(\theta - \theta_0)^{\mathrm{T}}\left[\frac{\partial^2 \ell(\theta_0)}{\partial \theta \partial \theta^{\mathrm{T}}}\right](\theta - \theta_0) - \frac{n}{2}\theta^{\mathrm{T}}\Sigma_\tau(\theta_0)\theta,$$

where

$$\Sigma_\tau(\theta_0) = \operatorname{diag}\left\{\frac{p'_{\tau_1^{(1)}}(|\beta_{01}|)}{|\beta_{01}|}, \cdots, \frac{p'_{\tau_p^{(1)}}(|\beta_{0p}|)}{|\beta_{0p}|}, \frac{p'_{\tau_1^{(2)}}(|\gamma_{01}|)}{|\gamma_{01}|}, \cdots, \frac{p'_{\tau_q^{(2)}}(|\gamma_{0q}|)}{|\gamma_{0q}|}, \frac{p'_{\tau_1^{(3)}}(|\lambda_{01}|)}{|\lambda_{01}|}, \cdots, \frac{p'_{\tau_d^{(3)}}(|\lambda_{0d}|)}{|\lambda_{0d}|}\right\},$$

where

$$\theta = (\theta_1, \cdots, \theta_s)^{\mathrm{T}} = (\beta_1, \cdots, \beta_p, \gamma_1, \cdots, \gamma_q, \lambda_1, \cdots, \lambda_d)^{\mathrm{T}}$$

and

$$\theta_0 = (\theta_{01}, \cdots, \theta_{0s})^{\mathrm{T}} = (\beta_{01}, \cdots, \beta_{0p}, \gamma_{01}, \cdots, \gamma_{0q}, \lambda_{01}, \cdots, \lambda_{0d})^{\mathrm{T}}.$$

Accordingly, the quadratic maximization problem for $L(\theta)$ leads to a solution iterated by

$$\theta_1 \approx \theta_0 + \left\{\frac{\partial^2 \ell(\theta_0)}{\partial \theta \partial \theta^{\mathrm{T}}} - n\Sigma_\tau(\theta_0)\right\}^{-1}\left\{n\Sigma_\tau(\theta_0)\theta_0 - \frac{\partial \ell(\theta_0)}{\partial \theta}\right\}.$$

Secondly, as the data are normally distributed the log-likelihood function $\ell(\theta)$ can be written as

$$\ell(\theta) = -\frac{1}{2}\sum_{i=1}^{n}\log(|\Sigma_i|) - \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu_i)^{\mathrm{T}}\Sigma_i^{-1}(y_i - \mu_i) = -\frac{1}{2}\sum_{i=1}^{n}\log(|D_i|) - \frac{1}{2}\sum_{i=1}^{n}\varepsilon_i^{\mathrm{T}}D_i^{-1}\varepsilon_i.$$

Therefore, the resulting score functions are

$$U(\theta) = \frac{\partial \ell(\theta)}{\partial \theta} = \left(U_1^{\mathrm{T}}(\beta), U_2^{\mathrm{T}}(\gamma), U_3^{\mathrm{T}}(\lambda)\right)^{\mathrm{T}},$$

where

$$U_1(\beta) = \frac{\partial \ell(\theta)}{\partial \beta} = \sum_{i=1}^{n} X_i^{\mathrm{T}}\Delta_i\Sigma_i^{-1}\left(y_i - \mu_i(X_i\beta)\right);$$

$$U_2(\gamma) = \frac{\partial \ell(\theta)}{\partial \gamma} = -\sum_{i=1}^{n}\frac{\partial \varepsilon_i^{\mathrm{T}}}{\partial \gamma}D_i^{-1}\varepsilon_i;$$

$$U_3(\lambda) = \frac{\partial \ell(\theta)}{\partial \lambda} = \frac{1}{2}\sum_{i=1}^{n}H_i^{\mathrm{T}}\left(D_i^{-1}f_i - 1_{m_i}\right).$$

Here $\Delta_i = \Delta_i(X_i\beta) = \operatorname{diag}\left\{\dot{g}^{-1}\left(x_{i1}^{\mathrm{T}}\beta\right), \cdots, \dot{g}^{-1}\left(x_{im_i}^{\mathrm{T}}\beta\right)\right\}$, $\dot{g}^{-1}(\cdot)$ is the derivative of the inverse of the link function $g^{-1}(\cdot)$, and $\varepsilon_i = (\varepsilon_{i1}, \cdots, \varepsilon_{im_i})^{\mathrm{T}}$ with

$$\varepsilon_{ij} = r_{ij} - \sum_{k=1}^{j-1} l_{ijk}\varepsilon_{ik};\ f_i = (f_{i1}, \cdots, f_{im_i})^{\mathrm{T}}$$

with $f_{ij} = \varepsilon_{ij}^2$ and 1_{m_i} is a vector of 1's. Denote

$$I_n(\theta) = -E\frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta^{\mathrm{T}}} = \begin{pmatrix} -E\frac{\partial^2 \ell(\theta)}{\partial\beta\partial\beta^{\mathrm{T}}} & -E\frac{\partial^2 \ell(\theta)}{\partial\beta\partial\gamma^{\mathrm{T}}} & -E\frac{\partial^2 \ell(\theta)}{\partial\beta\partial\lambda^{\mathrm{T}}} \\ -E\frac{\partial^2 \ell(\theta)}{\partial\gamma\partial\beta^{\mathrm{T}}} & -E\frac{\partial^2 \ell(\theta)}{\partial\gamma\partial\gamma^{\mathrm{T}}} & -E\frac{\partial^2 \ell(\theta)}{\partial\gamma\partial\lambda^{\mathrm{T}}} \\ -E\frac{\partial^2 \ell(\theta)}{\partial\lambda\partial\beta^{\mathrm{T}}} & -E\frac{\partial^2 \ell(\theta)}{\partial\lambda\partial\gamma^{\mathrm{T}}} & -E\frac{\partial^2 \ell(\theta)}{\partial\lambda\partial\lambda^{\mathrm{T}}} \end{pmatrix} = \begin{pmatrix} I_{11} & I_{12} & I_{13} \\ I_{21} & I_{22} & I_{23} \\ I_{31} & I_{32} & I_{33} \end{pmatrix},$$

where

$$I_{11} = \sum_{i=1}^{n} X_i^{\mathrm{T}} \Delta_i \Sigma_i^{-1} \Delta_i X_i;$$

$$I_{22} = \sum_{i=1}^{n} E\frac{\partial \varepsilon_i^{\mathrm{T}}}{\partial\gamma} D_i^{-1} \frac{\partial \varepsilon_i}{\partial\gamma};$$

$$I_{33} = \frac{1}{2}\sum_{i=1}^{n} H_i^{\mathrm{T}} H_i$$

and $I_{12} = I_{21} = I_{13} = I_{31} = I_{23} = I_{32} = 0$.

Finally, by using the Fisher information matrix to approximate the observed information matrix, the following algorithm summarizes the computation of penalized maximum likelihood estimators of the parameters in JMVGLRM.

Algorithm:

Step 1. Take the ordinary maximum likelihood estimators (without penalty) β_0, γ_0, λ_0 of β, γ, λ as their initial values.

Step 2. Given the current values

$$\theta^{(m)} = \left\{\beta^{(m)\mathrm{T}}, \gamma^{(m)\mathrm{T}}, \lambda^{(m)\mathrm{T}}\right\}^{\mathrm{T}}$$

update it by

$$\theta^{(m+1)} = \theta^{(m)} + \left\{I_n\left(\theta^{(m)}\right) + n\Sigma_\tau\left(\theta^{(m)}\right)\right\}^{-1} \cdot \left\{U\left(\theta^{(m)}\right) - n\Sigma_\tau\left(\theta^{(m)}\right)\theta^{(m)}\right\}.$$

Step 3. Repeat Step 2 above until certain convergence criteria are satisfied.

4.2. Choosing the Tuning Parameters

The penalty function $p_{\tau^{(l)}}(\cdot)$ involves the tuning parameters $\tau^{(l)}\left(l = 1,2,3\right)$ that controls the amount of penalty. Many selection criteria, such as CV, GCV, AIC and BIC selection can be used to select the tuning parameters. Wang *et al.* [16] suggested using the BIC for the SCAD estimator in linear models and partially linear models, and proved its model selection consistency property, *i.e.*, the optimal parameter chosen by BIC can identify the true model with probability tending to one. Hence, we use their suggestion throughout this paper. So the BIC will be used to choose the optimal $\{\tau_i, i = 1,\cdots,s\}$ which is equal to either $\left\{\tau_i^{(1)}, i = 1,\cdots,p\right\}$, $\left\{\tau_j^{(2)}, j = 1,\cdots,q\right\}$ or $\left\{\tau_k^{(3)}, k = 1,\cdots,d\right\}$. Nevertheless, in real application, how to simultaneously select a total of *s* shrinkage parameters $\{\tau_i, i = 1,\cdots,s\}$ is challenging. To bypass this difficulty, we follow the idea of [12,16,17], and simplify the tuning parameters as

1) $\tau_{1i} = \dfrac{\tau_1}{\left|\tilde{\beta}_i^{(0)}\right|}, i = 1,\cdots,p,$

2) $\tau_{2j} = \dfrac{\tau_2}{\left|\tilde{\gamma}_j^{(0)}\right|}, j = 1,\cdots,q,$

3) $\tau_{3k} = \dfrac{\tau_3}{\left|\tilde{\lambda}_k^{(0)}\right|}, k = 1,\cdots,d$

in the numerical studies followed, where $\tilde{\beta}_i^{(0)}$, $\tilde{\gamma}_j^{(0)}$ and $\tilde{\lambda}_k^{(0)}$ are respectively the *i*th element, *j*th element and *k*th element of the unpenalized estimate $\tilde{\beta}^{(0)}$, $\tilde{\gamma}^{(0)}$ and $\tilde{\lambda}^{(0)}$. Consequently, the original *s* dimensional problem about τ_i becomes a three dimensional problem about $\tau = \left(\tau_1, \tau_2, \tau_3\right)^{\mathrm{T}}$. τ can be selected according to the following BIC-type criterion

$$\mathrm{BIC}_\tau = -\frac{2}{n}\ell\left(\hat{\beta}, \hat{\gamma}, \hat{\lambda}\right) + df_\tau \times \frac{\log(n)}{n}.$$

where $0 \le df_\tau \le s$ is simply the number of nonzero coefficients of $\hat{\theta}$.

From our simulation study, we found that this method works well.

5. Simulation Studies

In this section we conduct simulation studies to assess the small sample performance of the proposed procedures. We consider the sample size *n* = 100, 200, and 400 respectively. Each subject is supposed to be measured by m_i times with $m_i - 1 \sim \mathrm{Binomial}(11, 0.8)$. In the simulation study, 1000 repetitions of random samples are generated by using the above data generation procedure. For each simulated data set, the proposed variable selection procedures for finding out penalized maximum likelihood estimators with SCAD and adaptive lasso (ALASSO) penalty functions [17] are considered. The unknown tuning parameters $\tau^{(l)}$, $\left(l = 1,2,3\right)$ for the penalty functions are chosen by BIC criterion in the simulation. The performance of estimator $\hat{\beta}$, $\hat{\gamma}$ and $\hat{\lambda}$ will be assessed by the mean square error (MSE), defined as

$$\mathrm{MSE}\left(\hat{\beta}\right)=E\left(\hat{\beta}-\beta_0\right)^{\mathrm{T}}\left(\hat{\beta}-\beta_0\right),$$

$$\mathrm{MSE}(\hat{\gamma})=E\left(\hat{\gamma}-\gamma_0\right)^{\mathrm{T}}\left(\hat{\gamma}-\gamma_0\right),$$

$$\mathrm{MSE}(\hat{\lambda})=E\left(\hat{\lambda}-\lambda_0\right)^{\mathrm{T}}\left(\hat{\lambda}-\lambda_0\right).$$

5.1. Example 1: Linear Mean Model for JMVGLRM

In this example, we first consider the linear model for mean parameters as a special JMVGLRM. We choose the true values of the mean parameters, moving average parameters and log-innovation variances to be $\beta=\left(\beta_1,\beta_2,\cdots,\beta_{10}\right)^{\mathrm{T}}$ with $\beta_1=1$, $\beta_2=-0.5$, $\beta_4=0.5$, $\lambda=\left(\lambda_1,\lambda_2,\cdots,\lambda_7\right)^{\mathrm{T}}$ with $\lambda_2=0.5$, $\lambda_3=0.4$, and $\gamma=\left(\gamma_1,\gamma_2,\cdots,\gamma_7\right)^{\mathrm{T}}$ with $\gamma_1=-0.3$, $\gamma_2=0.3$, respectively, while the remaining coefficients, corresponding to the irrelevant variables, are given by zeros. In the models $x_{ij}=\left(1,x_{1ij}^{\mathrm{T}}\right)^{\mathrm{T}}$ with x_{1ij} is generated from a multivariate normal distribution with mean zero, marginal variance 1 and all correlations 0.5. We take $h_{ij}=\left(x_{ijt}\right)_{t=1}^{7}$ and

$$z_{ijk}=\left(1,t_{ij}-t_{ik},\left(t_{ij}-t_{ik}\right)^2,\cdots,\left(t_{ij}-t_{ik}\right)^6\right)^{\mathrm{T}}$$

and the measurement times t_{ij} are generated from the uniform distribution $U[0,2]$. Using these values, the mean μ_i and covariance matrix Σ_i are constructed through the modified Cholesky decomposition described in Section 2. The responses y_i are then drawn from the multivariate normal distribution $N\left(\mu_i,\Sigma_i\right), i=1,\cdots,n$.

The average number of the estimated zero coefficients for the parametric components, with 1000 simulation runs, is reported in **Table 1**. Note that "Correct" in **Table 1** means the average number of zero regression coefficients that are correctly estimated as zero, and "Incorrect" depicts the average number of non-zero regression coefficients that are erroneously set to zero.

From **Table 1**, we can make the following observations. Firstly, the performances of variable selection procedures with different penalty functions become better and better as *n* increases. For example, the values in the column labeled "Correct" become more and more closer to the true number of zero regression coefficients in the models. Secondly, the SCAD and ALASSO penalty methods perform similarly in the sense of correct variable selection rate, which significantly reduces the model uncertainty and complexity. Thirdly, for the designed settings, the overall performance of the variable selection procedure is satisfactory.

Next, we compare the two decomposition methods under two data generating processes, autoregressive (AR) decomposition [1] and moving average (MA) decomposition [6]. The main measurements for comparison are differences between the fitted mean $\hat{\mu}_i$ and the true mean μ_i, and the fitted covariance matrix $\hat{\Sigma}_i$ to the true Σ_i. In particular, we define two relative errors as

$$\mathrm{RERR}\left(\hat{\mu}\right)=\frac{1}{n}\sum_{i=1}^{n}\frac{\left\|\hat{\mu}_i-\mu_i\right\|}{\left\|\mu_i\right\|},\ \mathrm{RERR}\left(\hat{\Sigma}\right)=\frac{1}{n}\sum_{i=1}^{n}\frac{\left\|\hat{\Sigma}_i-\Sigma_i\right\|}{\left\|\Sigma_i\right\|}.$$

Here $\|A\|$ denotes the largest singular value of *A*. We compute the averages of these two relative errors for 1000 replications with *n* = 100 and 200. **Table 2** gives the averages of relative errors for the MA decomposition and AR decomposition, when the data are generated from our model under different true covariance matrix. In **Table 2**, "MA.data" ("AR.data") means that the true covariance matrix follows the moving average structure (autoregressive structure). "MA.fit" ("AR.fit") means we

Table 1. Variable selection for JMVGLRM (linear mean model) using different penalties and sample size.

Model	n	SCAD			ALASSO		
		MSE	Correct	Incorrect	MSE	Correct	Incorrect
	100	0.0012	6.9340	0	0.0012	7.0000	0
β	200	7.8107e−004	6.9870	0	8.3486e−004	7.0000	0
	400	0.0005	6.9990	0	0.0006	7.0000	0
	100	1.3369e−004	4.9080	0	1.2259e−004	4.9880	0
γ	200	6.5800e−005	4.9850	0	7.4626e−005	5.0000	0
	400	4.6587e−005	4.9980	0	5.3295e−005	5.0000	0
	100	0.0417	4.8700	0.0010	0.0356	4.9750	0.0010
λ	200	0.0254	4.9380	0	0.0246	4.9970	0
	400	0.0218	4.9940	0	0.0190	5.0000	0

Table 2. Average of relative errors using different methods and sample size.

True	Method	n	MA.fit		AR.fit	
			RERR($\hat{\mu}$)	RERR($\hat{\Sigma}$)	RERR($\hat{\mu}$)	RERR($\hat{\Sigma}$)
MA.data	SCAD	100	0.0159	0.1999	0.0609	0.8528
		200	0.0127	0.1719	0.0554	0.8245
	ALASSO	100	0.0161	0.1548	0.0696	0.8158
		200	0.0131	0.1442	0.0646	0.8047
AR.data	SCAD	100	0.0495	0.6636	0.0404	0.3061
		200	0.0427	0.6527	0.0356	0.2639
	ALASSO	100	0.0541	0.6960	0.0400	0.2370
		200	0.0436	0.6473	0.0362	0.2253

decompose the covariance matrix by MA decomposition (AR decomposition) to fit data. We see that when the true covariance matrix follows the moving average structure, the errors in estimating μ and Σ both increase when incorrectly decomposing the covariance matrix using the autoregressive structure, and vice versa. However, for this simulation study, model misspecification seems to affect the MA decomposition less than AR decomposition.

5.2. Example 2: Generalized Linear Mean Model for JMVGLRM

Consider the following logistic link function to model the mean component in the JMVGLRM, then we have

$$\log it\left(\mu_{ij}\right) = x_{ij}^{\mathrm{T}}\beta .$$

We use the settings in example 1 to assess the performance of the proposed variable selection procedures, and the simulation results are reported in **Table 3**.

The results in **Table 3** show that under different sample size, the proposed variable selection methods have the desired performance, which is substantively similar to the previous example.

5.3. Example 3: High-Dimensional Setup for JMVGLRM

In this example, we discuss how the proposed variable selection procedures can be applied to the "large *n*, diverging *s*" setup for JMVGLRM. We consider the following high-dimensional logistic mean model in JMVGLRM:

$$\log it\left(\mu_{ij}\right) = x_{ij}^{\mathrm{T}}\beta_0, l_{ijk} = z_{ijk}^{\mathrm{T}}\gamma_0, \log\left(\sigma_{ij}^2\right) = h_{ij}^{\mathrm{T}}\lambda_0,$$

where β_0 is a *p*-dimensional vector of parameters with $p = \left\lfloor 4n^{1/3} \right\rfloor - 4$ for n = 100, 200 and 400, and $\lfloor u \rfloor$ denotes the largest integer not greater than *u*. In addition, γ_0 is a *q*-dimensional vector of parameters with $q = \left\lfloor 2n^{1/3} \right\rfloor - 2$ and λ_0 is a *d*-dimensional vector of parameters with $d = \left\lfloor 3n^{1/3} \right\rfloor - 3 \cdot x_{ij} = \left(1, x_{1ij}^{\mathrm{T}}\right)^{\mathrm{T}}$ with x_{1ij} is generated from a multivariate normal distribution with mean zero, marginal variance 1 and all correlations 0.5. We take

$$h_{ij} = \left(x_{ijt}\right)_{t=1}^{d}$$

$$z_{ijk} = \left(1, t_{ij} - t_{ik}, \left(t_{ij} - t_{ik}\right)^2, \cdots, \left(t_{ij} - t_{ik}\right)^{q-1}\right)^{\mathrm{T}},$$

where the measurement times t_{ij} are generated from the uniform distribution $U[0,2]$.

The true coefficient vectors are

$$\beta_0 = \left(1, -0.5, 0.5, 0_{p-3}\right)^{\mathrm{T}}$$

$$\lambda_0 = \left(-0.4, -0.4, 0_{d-2}\right)^{\mathrm{T}}$$

$$\gamma_0 = \left(-0.6, 0.6, 0_{q-2}\right)^{\mathrm{T}},$$

and, where 0_m denotes a *m*-vector of 0's. Using these values, the mean μ_i and covariance matrix Σ_i are constructed through the modified Cholesky decomposition described in Section 2. Then, the responses y_i are then drawn from the multivariate normal distribution $N\left(\mu_i, \Sigma_i\right), i = 1, \cdots, n.$ The summary of simulation results are reported in **Table 4**.

It is easy to see from **Table 4** that, the proposed variable selection method is able to correctly identify the true submodel, and works remarkably well, even if it is the "large *n*, diverging *s*" setup for JMVGLRM.

6. Acknowledgements

This work is supported by grants from the National Natural Science Foundation of China (10971007, 11271039, 11261025); Funding Project of Science and Technology

Table 3. Variable selection for JMVGLRM (generalized linear mean model) using different penalties and sample size.

Model	n	SCAD			ALASSO		
		MSE	Correct	Incorrect	MSE	Correct	Incorrect
	100	0.1346	6.8820	0.0300	0.1591	6.9580	0.0700
β	200	0.1028	6.9920	0	0.0886	6.9980	0.0010
	400	0.0838	7.0000	0	0.0727	7.0000	0
	100	1.2997e−004	4.8480	0	1.4948e−004	4.9900	0
γ	200	7.2503e−005	4.9720	0	8.3386e−005	5.0000	0
	400	2.5737e−005	4.9820	0	5.9863e−005	5.0000	0
	100	0.0149	4.9270	0	0.0297	4.9980	0.0030
λ	200	0.0086	4.9940	0	0.0178	5.0000	0
	400	0.0059	5.0000	0	0.0135	5.0000	0

Table 4. Variable selection for high-dimensional JMVGLRM (generalized linear mean model) using different penalties and sample size.

Model	(n, p/q/d)	SCAD			ALASSO		
		MSE	Correct	Incorrect	MSE	Correct	Incorrect
	(100, 14)	0.0053	10.7840	0.0090	0.0063	11.0000	0.0090
β	(200, 18)	0.0004	14.9900	0	0.0011	15.0000	0
	(400, 24)	0.0002	20.9980	0	0.0005	21.0000	0
	(100, 7)	0.0022	4.8690	0.0060	0.0022	4.9990	0.0060
γ	(200, 9)	6.2065e−006	6.8200	0	5.4613e−006	6.9820	0
	(400, 12)	1.8671e−005	9.7170	0	3.1547e−006	9.8960	0
	(100, 10)	0.0151	7.8060	0.0060	0.0276	7.9910	0.0060
λ	(200, 14)	0.0117	11.9260	0	0.0225	12.0000	0
	(400, 18)	0.0071	15.9880	0	0.0105	16.0000	0

Research Plan of Beijing Education Committee (JC-006790201001); Beijing municipal key disciplines (No. 006000541212010).

REFERENCES

[1] M. Pourahmadi, "Joint Mean-Covariance Models with Applications to Lontidinal Data: Unconstrained Parameterisation," *Biometrika*, Vol. 86, No. 3, 1999, pp. 677-690. doi:10.1093/biomet/86.3.677

[2] M. Pourahmadi, "Maximum Likelihood Estimation for Generalised Linear Models for Multivariate Normal Covariance Matrix," *Biometrika*, Vol. 87, No. 2, 2000, pp. 425-435. doi:10.1093/biomet/87.2.425

[3] P. T. Diggle and A. Verbyla, "Nonparametric Estimation of Covariance Structure in Longitudinal Data," *Biometrics*, Vol. 54, No. 2, 1998, pp. 401-415. doi:10.2307/3109751

[4] J. Q. Fan, T. Huang and R. Li, "Analysis of Longitudinal Data with Semiparametric Estimation of Covariance Function," *Journal of the American Statistical Association*, Vol. 102, No. 478, 2007, pp. 632-641. doi:10.1198/016214507000000095

[5] J. Q. Fan and Y. Wu, "Semiparametric Estimation of Covariance Matrices for Longitudinal Data," *Journal of the American Statistical Association*, Vol. 103, No. 484, 2008, pp. 1520-1533. doi:10.1198/016214508000000742

[6] A. J. Rothman, E. Levina and J. Zhu, "A New Approach to Cholesky-Based Covariance Regularization in High Dimensions," *Biometrika*, Vol. 97, No. 3, 2010, pp. 539-550. doi:10.1093/biomet/asq022

[7] W. P. Zhang and C. L. Leng, "A Moving Average Cholesky Factor Model in Covariance Modeling for Longitudinal Data," *Biometrika*, Vol. 99, No. 1, 2012, pp. 141-150. doi:10.1093/biomet/asr068

[8] L. Breiman, "Better Subset Selection Using Nonnegative Garrote," *Techonometrics*, Vol. 37, No. 4, 1995, pp. 373-

384. doi:10.1080/00401706.1995.10484371

[9] R. Tibshirani, "Regression Shrinkage and Selection via the LASSO," *Journal of Royal Statistical Society*, *Series B*, Vol. 58, No. 1, 1996, pp. 267-288.

[10] W. J. Fu, "Penalized Regression: The Bridge versus the LASSO," *Journal of Computational and Graphical Statistics*, Vol. 7, No. 3, 1998, pp.397-416.

[11] J. Q. Fan and R. Li, "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of American Statistical Association*, Vol. 96, No. 456, 2001, pp. 1348-1360. doi:10.1198/016214501753382273

[12] H. Zou and R. Li, "One-Step Sparse Estimates in Nonconcave Penalized Likelihood Models," *The Annals of Statistics*, Vol. 36, No. 4, 2008, pp. 1509-1533. doi:10.1214/009053607000000802

[13] Z. Z. Zhang and D. R. Wang, "Simultaneous Variable Selection for Heteroscedastic Regression Models," *Science China Mathematic*, Vol. 54, No. 3, 2011, pp. 515-530. doi:10.1007/s11425-010-4147-8

[14] P. X. Zhao and L. G. Xue, "Variable Selection in Semiparametric Regression Analysis for Longitudinal Data," *Annals of the Institute of Statistical Mathematics*, Vol. 64, No. 1, 2012, pp. 213-231. doi:10.1007/s10463-010-0312-7

[15] H. J. Ye and J. X. Pan, "Modelling of Covariance Structures in Generalized Estimating Equations for Longitudinal Data," *Biometrika*, Vol. 93, No. 4, 2006, pp. 927-941. doi:10.1093/biomet/93.4.927

[16] H. Wang, G. Li and C. L. Tsai, "Tuning Parameter Selectors for the Smoothly Clipped Absolute Deviation Method," *Biometrika*, Vol. 94, No. 3, 2007, pp. 553-568. doi:10.1093/biomet/asm053

[17] H. Zou, "The Adaptive Lasso and Its Oracle Properties," *Journal of American Statistical Association*, Vol. 101, No. 476, 2006, pp. 1418-1429. doi:10.1198/016214506000000735

