Chapter 13
Variable Selection in Joint Mean and Covariance Models

Chaofeng Kou and Jianxin Pan

Abstract In this paper, we propose a penalized maximum likelihood method for variable selection in joint mean and covariance models for longitudinal data. Under certain regularity conditions, we establish the consistency and asymptotic normality of the penalized maximum likelihood estimators of parameters in the models. We further show that the proposed estimation method can correctly identify the true models, as if the true models would be known in advance. We also carry out real data analysis and simulation studies to assess the small sample performance of the new procedure, showing that the proposed variable selection method works satisfactorily.

13.1 Introduction

In longitudinal studies, one of the main objectives is to find out how the average value of the response varies over time and how the average response profile is affected by different treatments or various explanatory variables of interest. Traditionally the within-subject covariance matrices are treated as nuisance parameters or assumed to have a very simple parsimonious structure, which inevitably leads to a misspecification of the covariance structure. Although the misspecification need not affect the consistency of the estimators of the parameters in the mean, it can lead to a great loss of efficiency of the estimators. In some circumstances, for example, when missing data are present, the estimators of the mean parameters can be severely biased if the covariance structure is misspecified. Therefore, correct specification of the covariance structure is really important.

On the other hand, the within-subject covariance structure itself may be of scientific interest, for example, in prediction problems arising in econometrics and finance. Moreover, like the mean, the covariances may be dependent on various explanatory variables. A natural constraint for modelling of covariance structures

C. Kou · J. Pan (✉)
Department of Mathematics, University of Manchester, Manchester, UK
e-mail: ckou@maths.man.ac.uk; jianxin.pan@manchester.ac.uk

T. Holgersson, M. Singull (eds.), *Recent Developments in Multivariate and Random Matrix Analysis*, https://doi.org/10.1007/978-3-030-56773-6_13

is that the estimated covariance matrices must be positive definite, making the covariance modelling rather challenging. Chiu et al. [2] proposed to solve this problem by using a matrix logarithmic transformation, defined as the inverse of the matrix exponential transformation by taking the spectral decomposition of the covariance matrix. Since there are no constraints on the upper triangular elements of the matrix logarithm, any structures of interest may be imposed on the elements of the matrix logarithm. But the limitation of this approach is that the matrix logarithm is lack a clear statistical interpretation. An alternative method to deal with the positive definite constraint of covariance matrices is to work on the modified Cholesky decomposition advocated by Pourahmadi [9, 10], and use regression formulations to model the unconstrained elements in the decomposition. The key idea is that any covariance matrix can be diagonalized by a unique lower triangular matrix with 1's as its diagonal elements. The elements of the lower triangular matrix and the diagonal matrix enjoy a very clear statistical interpretation in terms of autoregressive coefficients and innovation variances, see, e.g., Pan and MacKenzie [8]. Ye and Pan [13] proposed an approach for joint modelling of mean and covariance structures for longitudinal data within the framework of generalized estimation equations, which does not require any distribution assumptions and only assumes the existence of the first four moments of the responses. However, a challenging issue for modelling joint mean and covariance structures is the high-dimensional problem, which arises frequently in many fields such as genomics, gene expression, signal processing, image analysis and finance. For example, the number of explanatory variables may be very large. Intuitively, all these variables should be included in the initial model in order to reduce the modelling bias. But it is very likely that only a small number of these explanatory variables contribute to the model fitting and the majority of them do not. Accordingly, these insignificant variables should be excluded from the initial model to increase prediction accuracy and avoid overfitting problem. Variable selection thus can improve estimation accuracy by effectively identifying the important subset of the explanatory variables, which may be just tens out of several thousands of predictors with a sample size being in tens or hundreds.

There are many variable selection criteria existing in the literature. Traditional variable selection criteria such as Mallow's C_p criteria, Akaike Information Criterion (AIC) and Bayes Information Criterion (BIC) all involve a combinatorial optimization problem, with computational loads increasing exponentially with the number of explanatory variables. This intensive computation problem hampers the use of traditional procedures. Fan and Li [4] discussed a class of penalized likelihood based methods for variable selection, including the bridge regression by Frank and Friedman [6], Lasso by Tibshirani [11] and smoothly clipped absolute deviation by Fan and Li [4]. In the setting of finite parameters, [4] further studied oracle properties for non-concave penalized likelihood estimators in the sense that the penalized maximum likelihood estimator can correctly identify the true model as if we would know it in advance. Fan and Peng [5] extended the results by letting the number of parameters have the order $o(n^{1/3})$ and showed that the oracle properties still hold in this case. Zou [14] proposed an adaptive Lasso in a finite parameter

setting and showed that the Lasso does not have oracle properties as conjectured by Fan and Li [4], but the adaptive Lasso does.

In this paper 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. We show that the proposed approach produces good estimation results and can correctly identify zero regression coefficients for the joint mean and covariance models, simultaneously. The rest of the paper is organized as follows. In Sect. 13.2, we first describe a reparameterisation of covariance matrix 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. The standard error formula of the parameter estimators and the choice of the tuning parameters are provided. In Sect. 13.3, we study the variable selection method and its sample properties when the number of explanatory variables tends to infinity with the sample size. In Sect. 13.4, we illustrate the proposed method via a real data analysis. In Sect. 13.5, we carry out simulation studies to assess the small sample performance of the method. In Sect. 13.6, we give a further discussion on the proposed variable selection method. Technical details on calculating the penalized likelihood estimators of parameters are given in Appendix A, and theoretical proofs of the theorems that summarize the asymptotic results are presented in Appendix B.

13.2 Variable Selection via Penalized Maximum Likelihood

13.2.1 Joint Mean and Covariance Models

Suppose that there are n independent subjects and the ith subject has m_i repeated measurements. Let y_{ij} be the jth measurement of the ith subject and t_{ij} be the time at which the measurement y_{ij} is made. Throughout this paper we assume that $\mathbf{y}_i = (y_{i1}, \ldots, y_{im_i})^T$ is a random sample of the ith subject from the multivariate normal distribution with the mean μ_i and covariance matrix Σ_i, where $\mu_i = (\mu_{i1}, \ldots, \mu_{im_i})^T$ is an $(m_i \times 1)$ vector and Σ_i is an $(m_i \times m_i)$ positive definite matrix $(i = 1, \ldots, n)$. We consider the simultaneous variable selection procedure for the mean and covariance structures using penalized maximum likelihood estimation methods.

To deal with the positive definite constraint of the covariance matrices, we design an effective regularization approach to gain statistical efficiency and overcome the high dimensionality problem in the covariance matrices. We actually use a statistically meaningful representation that reparameterizes the covariance matrices by the modified Cholesky decomposition advocated by Pourahmadi [9, 10]. Specifically,

any covariance matrix Σ_i $(1 \leq i \leq n)$ can be diagonalized by a unique lower triangular matrix T_i with 1's as its diagonal elements. In other words,

$$T_i \Sigma_i T_i^T = D_i, \tag{13.1}$$

where D_i is a unique diagonal matrix with positive diagonal elements. The elements of T_i and D_i have a very clear statistical interpretation in terms of autoregressive least square regressions. More precisely, the below-diagonal entries of $T_i = (-\phi_{ijk})$ are the negatives of the regression coefficients of $\widehat{y}_{ij} = \mu_{ij} + \sum_{k=1}^{j-1} \phi_{ijk}(y_{ik} - \mu_{ik})$, the linear least square predictor of y_{ij} based on its predecessors $y_{i1}, \ldots, y_{i(j-1)}$, and the diagonal entries of $D_i = \text{diag}(\sigma_{i1}^2, \ldots, \sigma_{im_i}^2)$ are the prediction error variances $\sigma_{ij}^2 = \text{var}(y_{ij} - \widehat{y}_{ij})$ $(1 \leq i \leq n, 1 \leq j \leq m_i)$. The new parameters ϕ_{ijk}'s and σ_{ij}^2's are called generalized autoregressive parameters and innovation variances, respectively. By taking log transformation to the innovation variances, the decomposition (13.1) converts the constrained entries of $\{\Sigma_i : i = 1, \ldots, n\}$ into two groups of unconstrained autoregressive regression parameters and innovation variances, given by $\{\phi_{ijk} : i = 1, \ldots, n; j = 2, \ldots, m_i; k = 1, \ldots, (j-1)\}$ and $\{\log \sigma_{ij}^2 : i = 1, \ldots, n, j = 1, \ldots, m_i\}$, respectively.

Based on the modified Cholesky decomposition, the unconstrained parameters μ_{ij}, ϕ_{ijk} and $\log \sigma_{ij}^2$ are modelled in terms of the linear regression models

$$\mu_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}, \quad \phi_{ijk} = \mathbf{z}_{ijk}^T \boldsymbol{\gamma} \quad \text{and} \quad \log \sigma_{ij}^2 = \mathbf{h}_{ij}^T \boldsymbol{\lambda}, \tag{13.2}$$

where $\mathbf{x}_{ij}$, $\mathbf{z}_{ijk}$ and $\mathbf{h}_{ij}$ are $(p \times 1)$, $(q \times 1)$ and $(d \times 1)$ covariates vectors, and $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$ are the associated regression coefficients. The covariates $\mathbf{x}_{ij}$, $\mathbf{z}_{ijk}$ and $\mathbf{h}_{ij}$ may contain baseline covariates, polynomials in time and their interactions, etc. For example, when modelling stationary growth curve data using polynomials in time, the explanatory variables may take the forms $\mathbf{x}_{ij} = (1, t_{ij}, t_{ij}^2, \ldots, t_{ij}^{p-1})^T$, $\mathbf{z}_{ijk} = (1, (t_{ij} - t_{ik}), (t_{ij} - t_{ik})^2, \ldots, (t_{ij} - t_{ik})^{q-1})^T$ and $\mathbf{h}_{ij} = (1, t_{ij}, t_{ij}^2, \ldots, t_{ij}^{d-1})^T$. An advantage of the model (13.2) is that the resulting estimators of the covariance matrices can be guaranteed to be positive definite. In this paper we assume that the covariates $\mathbf{x}_{ij}$, $\mathbf{z}_{ijk}$ and $\mathbf{h}_{ij}$ may be of high dimension and we would select the important subsets of the covariates $\mathbf{x}_{ij}$, $\mathbf{z}_{ijk}$ and $\mathbf{h}_{ij}$, simultaneously. We first assume all the explanatory variables of interest, and perhaps their interactions as well, are already included into the initial models. We then aim to remove the unnecessary explanatory variables from the models.

13.2.2 Penalized Maximum Likelihood

Many traditional variable selection criteria can be considered as a penalized likelihood which balances modelling biases and estimation variances [4]. Let $\ell(\boldsymbol{\theta})$

denote the log-likelihood function. For the joint mean and covariance models (13.2), we propose the penalized likelihood function

$$Q(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) - n\sum_{i=1}^{p} p_{\tau^{(1)}}(|\beta_i|) - n\sum_{j=1}^{q} p_{\tau^{(2)}}(|\gamma_j|) - n\sum_{k=1}^{d} p_{\tau^{(3)}}(|\lambda_k|), \quad (13.3)$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_s)^T = (\beta_1, \ldots, \beta_p; \gamma_1, \ldots, \gamma_q; \lambda_1, \ldots, \lambda_d)^T$ with $s = p+q+d$ and $p_{\tau^{(l)}}(\cdot)$ is a given penalty function with the tuning parameter $\tau^{(l)}$ $(l = 1, 2, 3)$. 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, generalized autoregressive parameters and log-innovation variances, respectively. The function form of $p_\tau(\cdot)$ determines the general behavior of the estimators. Antoniadis [1] defined the hard thresholding rule for variable selection by taking the hard thresholding penalty function as $P_\tau(|t|) = \tau^2 - (|t| - \tau)^2 I(|t| < \tau)$, where $I(.)$ is the indicator function. The penality function $p_\tau(\cdot)$ may also be chosen as L_p penalty. Especially, the use of L_1 penalty, defined by $p_\tau(t) = \tau|t|$, leads to the least absolute shrinkage and selection operator (Lasso) proposed by Tibshirani [11]. Fan and Li [4] suggested using the smoothly clipped absolute deviation (SCAD) penalty function, which is defined by

$$p_\tau(|t|) = \begin{cases} \tau|t| & \text{if } 0 \le |t| < \tau \\ -(|t|^2 - 2a\tau|t| + \tau^2)/\{2(a-1)\} & \text{if } \tau \le |t| < a\tau \\ (a+1)\tau^2/2 & \text{if } |t| \ge a\tau \end{cases} \quad (13.4)$$

for some $a > 2$. This penalty function is continuous, symmetric and convex on $(0, \infty)$ but singular at the origin. It improves the Lasso by avoiding excessive estimation biases. Details of penalty functions can be found in [4].

The penalized maximum likelihood estimator of $\boldsymbol{\theta}$, denoted by $\widehat{\boldsymbol{\theta}}$, maximizes the function $Q(\boldsymbol{\theta})$ in (13.3). With appropriate penalty functions, maximizing $Q(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ leads to certain parameter estimators vanishing from the initial models so that the corresponding explanatory variables are automatically removed. Hence, through maximizing $Q(\boldsymbol{\theta})$ we achieve the goal of selecting important variables and obtaining the parameter estimators, simultaneously. In Appendix A, we provide the technical details and an algorithm for calculating the penalized maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$.

13.2.3 Asymptotic Properties

In this subsection we consider the consistency and asymptotic normality of the penalized maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$. To emphasize its dependence on the subject number n, we also denote it by $\widehat{\boldsymbol{\theta}}_n$. We assume that the number of the

parameters, $s = p + q + d$, is fixed in the first instance. In the next section we will consider the case when s is a variable tending to infinity with n. Denote the true value of $\boldsymbol{\theta}$ by $\boldsymbol{\theta}_0$. Without loss of generality, we assume that $\boldsymbol{\theta}_0 = ((\boldsymbol{\theta}_0^{(1)})^T, (\boldsymbol{\theta}_0^{(2)})^T)^T$ where $\boldsymbol{\theta}_0^{(1)}$ and $\boldsymbol{\theta}_0^{(2)}$ are the nonzero and zero components of $\boldsymbol{\theta}_0$, respectively. Otherwise the components of $\boldsymbol{\theta}_0$ can be reordered. Denote the dimension of $\boldsymbol{\theta}_0^{(1)}$ by s_1. In what follows we first show that the penalized maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_n$ exists and converges to $\boldsymbol{\theta}_0$ at the rate $O_p(n^{-1/2})$, implying that it has the same consistency rate as the ordinary maximum likelihood estimator. We then prove that the $\sqrt{n}$-consistent estimator $\widehat{\boldsymbol{\theta}}_n$ has the asymptotic normal distribution and possesses the oracle property under certain regularity conditions. The results are summarized in the following two theorems and the detailed proofs are provided in Appendix B. To prove the theorems in this paper, we require the following regularity conditions:

(A1) The covariates $\mathbf{x}_{ij}$, $\mathbf{z}_{ijk}$ and $\mathbf{h}_{ij}$ are fixed. Also, for each subject the number of repeated measurements, m_i, is fixed ($i = 1, \ldots, n$; $j = 1, \ldots, m_i$; $k = 1, \ldots, j-1$).

(A2) The parameter space is compact and the true value $\boldsymbol{\theta}_0$ is in the interior of the parameter space.

(A3) The design matrices $\mathbf{x}_i$, $\mathbf{z}_i$ and $\mathbf{h}_i$ in the joint models are all bounded, meaning that all the elements of the matrices are bounded by a single finite real number.

(A4) The dimensions of the parameter vectors $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\lambda}$, that is, p_n, q_n and d_n, have the same order as s_n.

(A5) The nonzero components of the true parameters $\theta_{01}^{(1)}, \ldots, \theta_{0s_1}^{(1)}$ satisfy

$$\min_{1 \le j \le s_1} \left\{ \frac{|\theta_{0j}^{(1)}|}{\tau_n} \right\} \to \infty \ \ (\text{as } n \to \infty),$$

where τ_n is equal to either $\tau_n^{(1)}$, $\tau_n^{(2)}$ or $\tau_n^{(3)}$, depending on whether $\theta_{0j}^{(1)}$ is a component of $\boldsymbol{\beta}_0$, $\boldsymbol{\gamma}_0$, and $\boldsymbol{\lambda}_0$ ($j = 1, \ldots, s_1$).

Theorem 13.1 *Let*

$$a_n = \max_{1 \le j \le s} \{p'_{\tau_n}(|\theta_{0j}|) : \theta_{0j} \neq 0\} \text{ and } b_n = \max_{1 \le j \le s} \{|p''_{\tau_n}(|\theta_{0j}|)| : \theta_{0j} \neq 0\},$$

where $\boldsymbol{\theta}_0 = (\theta_{01}, \ldots, \theta_{0s})^T$ *is the true value of* $\boldsymbol{\theta}$, *and* τ_n *is equal to either* $\tau_n^{(1)}$, $\tau_n^{(2)}$ *or* $\tau_n^{(3)}$, *depending on whether* θ_{0j} *is a component of* $\boldsymbol{\beta}_0$, $\boldsymbol{\gamma}_0$ *or* $\boldsymbol{\lambda}_0$ ($1 \le j \le s$). *Assume* $a_n = O_p(n^{-1/2})$, $b_n \to 0$ *and* $\tau_n \to 0$ *as* $n \to \infty$. *Under the conditions (A1)–(A3) above, with probability tending to* 1 *there must exist a local maximizer* $\widehat{\boldsymbol{\theta}}_n$ *of the penalized likelihood function* $Q(\boldsymbol{\theta})$ *in (13.3) such that* $\widehat{\boldsymbol{\theta}}_n$ *is a* $\sqrt{n}$*-consistent estimator of* $\boldsymbol{\theta}_0$.

We now consider the asymptotic normality property of $\widehat{\theta}_n$. Let

$$A_n = \text{diag}(p''_{\tau_n}(|\theta_{01}^{(1)}|), \ldots, p''_{\tau_n}(|\theta_{0s_1}^{(1)}|)),$$
$$\boldsymbol{c}_n = (p'_{\tau_n}(|\theta_{01}^{(1)}|)\text{sgn}(\theta_{01}^{(1)}), \ldots, p'_{\tau_n}(|\theta_{0s_1}^{(1)}|)\text{sgn}(\theta_{0s_1}^{(1)}))^T,$$

where τ_n has the same definition as that in Theorem 13.1, and $\theta_{0j}^{(1)}$ is the jth component of $\boldsymbol{\theta}_0^{(1)}$ $(1 \le j \le s_1)$. Denote the Fisher information matrix of $\boldsymbol{\theta}$ by $\mathscr{I}_n(\boldsymbol{\theta})$.

Theorem 13.2 *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 $\bar{\mathscr{I}}_n = \mathscr{I}_n(\boldsymbol{\theta}_0)/n$ converges to a finite and positive definite matrix $\mathscr{I}(\boldsymbol{\theta}_0)$ as $n \to \infty$. Under the same mild conditions as these given in Theorem 13.1, if $\tau_n \to 0$ and $\sqrt{n}\tau_n \to \infty$ as $n \to \infty$, then the $\sqrt{n}$-consistent estimator $\widehat{\boldsymbol{\theta}}_n = (\widehat{\boldsymbol{\theta}}_n^{(1)T}, \widehat{\boldsymbol{\theta}}_n^{(2)T})^T$ in Theorem 13.1 must satisfy $\widehat{\boldsymbol{\theta}}_n^{(2)} = 0$ and

$$\sqrt{n}(\bar{\mathscr{I}}_n^{(1)})^{-1/2}(\bar{\mathscr{I}}_n^{(1)} + A_n)\left\{(\widehat{\boldsymbol{\theta}}_n^{(1)} - \boldsymbol{\theta}_0^{(1)}) + (\bar{\mathscr{I}}_n^{(1)} + A_n)^{-1}\boldsymbol{c}_n\right\} \to \mathscr{N}_{s_1}(\boldsymbol{0}, I_{s_1})$$

in distribution, where $\bar{\mathscr{I}}_n^{(1)}$ is the $(s_1 \times s_1)$ submatrix of $\bar{\mathscr{I}}_n$ corresponding to the nonzero components $\boldsymbol{\theta}_0^{(1)}$ and I_{s_1} is the $(s_1 \times s_1)$ identity matrix.

Note for the SCAD penalty we can show

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

for $t > 0$, where $a > 2$ and $(x)_+ = xI(x > 0)$. Since $\tau_n \to 0$ as $n \to \infty$, we then have $a_n = 0$ and $b_n = 0$ so that $\boldsymbol{c}_n = \boldsymbol{0}$ and $A_n = \boldsymbol{0}$ when the sample size n is large enough. It can be verified that in this case the conditions in Theorems 13.1 and 13.2 are all satisfied. Accordingly, we must have

$$\sqrt{n}(\bar{\mathscr{I}}_n^{(1)})^{1/2}(\widehat{\boldsymbol{\theta}}_n^{(1)} - \boldsymbol{\theta}_0^{(1)}) \to \mathscr{N}_{s_1}(\boldsymbol{0}, I_{s_1})$$

in distribution. This means that the estimator $\widehat{\boldsymbol{\theta}}_n^{(1)}$ shares the same sampling property as if we would know $\boldsymbol{\theta}_0^{(2)} = \boldsymbol{0}$ in advance. In other words, the penalized maximum likelihood estimator of $\boldsymbol{\theta}$ based on the SCAD penalty can correctly identify the true model as if we would know it in advance. This property is the so-called

oracle property by Fan and Li [4]. Similarly, the parameter estimator based on the hard thresholding penalty also possesses the oracle property. For the Lasso penalty, however, the parameter estimator does not have the oracle property. A brief explanation for this is given as follows. Since $p_{\tau_n}(t) = \tau_n t$ for $t > 0$ and then $p'_{\tau_n}(t) = \tau_n$, the assumption of $a_n = O_p(n^{-1/2})$ in Theorem 13.1 implies $\tau_n = O_p(n^{-1/2})$, leading to $\sqrt{n}\tau_n = O_p(1)$. On the other hand, one of the conditions in Theorem 13.2 is $\sqrt{n}\tau_n \to \infty$ as $n \to \infty$, which conflicts the assumption of $\sqrt{n}\tau_n = O_p(1)$. Hence the oracle property cannot be guaranteed in this case.

13.2.4 Standard Error Formula

As a consequence of Theorem 13.2, the asymptotic covariance matrix of $\widehat{\boldsymbol{\theta}}_n^{(1)}$ is

$$\mathrm{Cov}(\widehat{\boldsymbol{\theta}}_n^{(1)}) = \frac{1}{n}(\bar{\mathscr{I}}_n^{(1)} + A_n)^{-1}\bar{\mathscr{I}}_n^{(1)}(\bar{\mathscr{I}}_n^{(1)} + A_n)^{-1} \tag{13.5}$$

so that the asymptotic standard error for $\widehat{\boldsymbol{\theta}}_n^{(1)}$ is straightforward. However, $\bar{\mathscr{I}}_n^{(1)}$ and A_n are evaluated at the true value $\boldsymbol{\theta}_0^{(1)}$, which is unknown. A natural choice is to evaluate $\bar{\mathscr{I}}_n^{(1)}$ and A_n at the estimator $\widehat{\boldsymbol{\theta}}_n^{(1)}$ so that the estimator of the asymptotic covariance matrix of $\widehat{\boldsymbol{\theta}}_n^{(1)}$ is obtained through (13.5).

Corresponding to the partition of $\boldsymbol{\theta}_0$, we assume $\boldsymbol{\theta} = (\boldsymbol{\theta}^{(1)T}, \boldsymbol{\theta}^{(2)T})^T$. Denote

$$\ell'(\boldsymbol{\theta}_0^{(1)}) = \left[\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{(1)}}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \quad \text{and} \quad \ell''(\boldsymbol{\theta}_0^{(1)}) = \left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{(1)}\partial \boldsymbol{\theta}^{(1)T}}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0},$$

where $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_0^{(1)T}, \mathbf{0})^T$. Also, let

$$\Sigma_{\tau_n}(\boldsymbol{\theta}_0^{(1)}) = \mathrm{diag}\left\{\frac{p'_{\tau_n^{(1)}}(|\theta_{01}^{(1)}|)}{|\theta_{01}^{(1)}|}, \ldots, \frac{p'_{\tau_n^{(s_1)}}(|\theta_{0s_1}^{(1)}|)}{|\theta_{0s_1}^{(1)}|}\right\}.$$

Using the observed information matrix to approximate the Fisher information matrix, the covariance matrix of $\widehat{\boldsymbol{\theta}}_n^{(1)}$ can be estimated through

$$\widehat{\mathrm{Cov}}(\widehat{\boldsymbol{\theta}}_n^{(1)}) = \left\{\ell''(\widehat{\boldsymbol{\theta}}_n^{(1)}) - n\Sigma_{\tau_n}(\widehat{\boldsymbol{\theta}}_n^{(1)})\right\}^{-1} \widehat{\mathrm{Cov}}\left\{\ell'(\widehat{\boldsymbol{\theta}}_n^{(1)})\right\}\left\{\ell''(\widehat{\boldsymbol{\theta}}_n^{(1)}) - n\Sigma_{\tau_n}(\widehat{\boldsymbol{\theta}}_n^{(1)})\right\}^{-1},$$

where $\widehat{\mathrm{Cov}}\{\ell'(\widehat{\boldsymbol{\theta}}_n^{(1)})\}$ is the covariance of $\ell'(\boldsymbol{\theta}^{(1)})$ evaluated at $\boldsymbol{\theta}^{(1)} = \widehat{\boldsymbol{\theta}}_n^{(1)}$.

13.2.5 Choosing the Tuning Parameters

The penalty function $p_{\tau^{(l)}}(\cdot)$ involves the tuning parameter $\tau^{(l)}$ $(l = 1, 2, 3)$ that controls the amount of penalty. We may use K-fold cross-validation or generalized cross-validation [4, 11] to choose the most appropriate tuning parameters τ's. For the purpose of fast computation, we prefer the K-fold cross-validation approach, which is described briefly as follows. First, we randomly split the full dataset $\mathscr{D}$ into K subsets which are of about the same sample size, denoted by $\mathscr{D}^v$ $(v = 1, \ldots, K)$. For each v, we use the data in $\mathscr{D} - \mathscr{D}^v$ to estimate the parameters and $\mathscr{D}^v$ to validate the model. We also use the log-likelihood function to measure the performance of the cross-validation method. For each $\boldsymbol{\tau} = (\tau^{(1)}, \tau^{(2)}, \tau^{(3)})^T$, the K-fold likelihood based cross-validation criterion is defined by

$$\mathrm{CV}(\boldsymbol{\tau}) = \frac{1}{K}\sum_{v=1}^{K}\left\{\sum_{i\in I_v}\log(|\widehat{\Sigma}_i^{-v}|) + \sum_{i\in I_v}(\mathbf{y}_i - \mathbf{x}_i\widehat{\boldsymbol{\beta}}^{-v})^T(\widehat{\Sigma}_i^{-v})^{-1}(\mathbf{y}_i - \mathbf{x}_i\widehat{\boldsymbol{\beta}}^{-v})\right\},$$

where I_v is the index set of the data in $\mathscr{D}^v$, and $\widehat{\boldsymbol{\beta}}^{-v}$ and $\widehat{\Sigma}_i^{-v}$ are the estimators of the mean parameter $\boldsymbol{\beta}$ and the covariance matrix Σ_i obtained by using the training dataset $\mathscr{D} - \mathscr{D}^v$. We then choose the most appropriate tuning parameter τ by minimizing $\mathrm{CV}(\tau)$. In general, we may choose the number of data subsets as $K = 5$ or $K = 10$.

13.3 Variable Selection when the Number of Parameters $s = s_n \to \infty$

In the previous section, we assume that the numbers of the parameters $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\lambda}$, i.e., p, q and d and therefore s, are fixed. In some circumstances, it is not uncommon that the number of explanatory variables increase with the sample size. In this section we consider the case where the number of parameters s_n is a variable, which goes to infinity as the sample size n tends to infinity. In what follows, we study the asymptotic properties of the penalized maximum likelihood estimator in this case.

As before, we assume that $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_0^{(1)T}, \boldsymbol{\theta}_0^{(2)T})^T$ is the true value of $\boldsymbol{\theta}$ where $\boldsymbol{\theta}_0^{(1)}$ and $\boldsymbol{\theta}_0^{(2)}$ are the nonzero and zero components of $\boldsymbol{\theta}_0$, respectively. Also, we denote the dimension of $\boldsymbol{\theta}_0$ by s_n, which increases with the sample size n this time. Similar to the previous section, we first show that there exists a consistent penalized maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_n$ that converges to $\boldsymbol{\theta}_0$ at the rate $O_p(\sqrt{s_n/n})$. We then show that the $\sqrt{n/s_n}$-consistent estimator $\widehat{\boldsymbol{\theta}}_n$ has an asymptotic normal distribution and possesses the oracle property.

Theorem 13.3 *Let*

$$a_n^* = \max_{1\le j\le s_n}\{p'_{\tau_n}(|\theta_{0j}|) : \theta_{0j}\neq 0\} \text{ and } b_n^* = \max_{1\le j\le s_n}\{|p''_{\tau_n}(|\theta_{0j}|)| : \theta_{0j}\neq 0\},$$

where $\boldsymbol{\theta}_0 = (\theta_{01},\ldots,\theta_{0s_n})^T$ *is the true value of* $\boldsymbol{\theta}$*, and* τ_n *is equal to either* $\tau_n^{(1)}$*,* $\tau_n^{(2)}$ *or* $\tau_n^{(3)}$*, depending on whether* θ_{0j} *is a component of* β_0*,* γ_0 *or* λ_0 $(1\le j\le s)$*. Assume* $a_n^* = O_p(n^{-1/2})$*,* $b_n^*\to 0$*,* $\tau_n\to 0$ *and* $s_n^4/n\to 0$ *as* $n\to\infty$*. Under the conditions (A1)–(A5) above, with probability tending to one there exists a local maximizer* $\widehat{\boldsymbol{\theta}}_n$ *of the penalized likelihood function* $Q(\boldsymbol{\theta})$ *in (13.3) such that* $\widehat{\boldsymbol{\theta}}_n$ *is a* $\sqrt{n/s_n}$*-consistent estimator of* $\boldsymbol{\theta}_0$*.*

In what follows we consider the asymptotic normality property of the estimator $\widehat{\boldsymbol{\theta}}_n$. Denote the number of nonzero components of $\boldsymbol{\theta}_0$ by $s_{1n}(\le s_n)$. Let

$$A_n^* = \mathrm{diag}(p''_{\tau_n}(|\theta_{01}^{(1)}|),\ldots,p''_{\tau_n}(|\theta_{0s_{1n}}^{(1)}|)),$$
$$\boldsymbol{c}_n^* = (p'_{\tau_n}(|\theta_{01}^{(1)}|)\mathrm{sgn}(\theta_{01}^{(1)}),\ldots,p'_{\tau_n}(|\theta_{0s_{1n}}^{(1)}|)\mathrm{sgn}(\theta_{0s_{1n}}^{(1)}))^T,$$

where τ_n is equal to either $\tau_n^{(1)}$, $\tau_n^{(2)}$ or $\tau_n^{(3)}$, depending on whether θ_{0j} is a component of $\boldsymbol{\beta}_0$, $\boldsymbol{\gamma}_0$ or $\boldsymbol{\lambda}_0$ $(1\le j\le s)$, and $\theta_{0j}^{(1)}$ is the jth component of $\boldsymbol{\theta}_0^{(1)}$ $(1\le j\le s_{1n})$. Denote the Fisher information matrix of $\boldsymbol{\theta}$ by $\mathscr{I}_n(\boldsymbol{\theta})$.

Theorem 13.4 *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 $\bar{\mathscr{I}}_n = \mathscr{I}_n(\boldsymbol{\theta}_0)/n$ *converges to a finite and positive definite matrix* $\mathscr{I}(\boldsymbol{\theta}_0)$ *as* $n\to\infty$*. Under the same mild conditions as these in Theorem 13.3, if* $\tau_n\to 0$*,* $s_n^5/n\to 0$ *and* $\tau_n\sqrt{n/s_n}\to\infty$ *as* $n\to\infty$*, then the* $\sqrt{n/s_n}$*-consistent estimator* $\widehat{\boldsymbol{\theta}}_n = (\widehat{\boldsymbol{\theta}}_n^{(1)T},\widehat{\boldsymbol{\theta}}_n^{(2)T})^T$ *in Theorem 13.3 must satisfy* $\widehat{\boldsymbol{\theta}}_n^{(2)} = 0$ *and*

$$\sqrt{n}M_n(\bar{\mathscr{I}}_n^{(1)})^{-1/2}(\bar{\mathscr{I}}_n^{(1)}+A_n^*)\left\{(\widehat{\boldsymbol{\theta}}_n^{(1)}-\boldsymbol{\theta}_0^{(1)})+(\bar{\mathscr{I}}_n^{(1)}+A_n^*)^{-1}\boldsymbol{c}_n^*\right\}\to\mathscr{N}_k(\mathbf{0},\boldsymbol{G})$$

in distribution, where $\bar{\mathscr{I}}_n^{(1)}$ *is the* $(s_{1n}\times s_{1n})$ *submatrix of* $\bar{\mathscr{I}}_n$ *corresponding to the nonzero components* $\boldsymbol{\theta}_0^{(1)}$*,* M_n *is an* $(k\times s_{1n})$ *matrix satisfying* $M_nM_n^T\to\boldsymbol{G}$ *as* $n\to\infty$*,* $\boldsymbol{G}$ *is an* $(k\times k)$ *positive definite matrix and* $k(\le s_{1n})$ *is a constant.*

The technical proofs of Theorems 13.3 and 13.4 are provided in Appendix B. Similar to the finite parameters setting, for the SCAD penalty and hard thresholding

penalty functions, it can be verified that the conditions in Theorems 13.3 and 13.4 are all satisfied. In this case, we have

$$\sqrt{n}M_n(\bar{\mathscr{I}}_n^{(1)})^{1/2}(\widehat{\boldsymbol{\theta}}_n^{(1)} - \boldsymbol{\theta}_0^{(1)}) \to \mathscr{N}_k(\mathbf{0}, \boldsymbol{G})$$

in distribution. That means the estimator $\widehat{\boldsymbol{\theta}}_n^{(1)}$ shares the same sampling property as if we would know $\boldsymbol{\theta}_0^{(2)} = 0$ in advance. In other words, the estimation procedures based on the SCAD and hard thresholding penalty have the oracle property. However, the L_1-based penalized maximum likelihood estimator like Lasso does not have this property. Based on Theorem 13.4, similar to the finite parameters case the asymptotic covariance estimator of $\widehat{\boldsymbol{\theta}}^{(1)}$ can also be constructed but the details are omitted.

13.4 Real Data Analysis

In this section, we apply the proposed procedure to the well known CD4+ cell data analysis, of which the data details can be found in [3]. The human immune deficiency virus (HIV) causes AIDS by reducing a person's ability to fight infection. The HIV attacks an immune cell called the CD4+ cell which orchestrates the body's immunoresponse to infectious agents. An uninfected individual usually has around 1100 cells per millilitre of blood. When infected, the CD4+ cells decrease in number with time and an infected person's CD4+ cell number can be used to monitor the disease progression. The data set we analyzed consists of 369 HIV-infected men. Altogether there are 2376 values of CD4+ cell numbers, with several repeated measurements being made for each individual at different times covering a period of approximately eight and a half years.

For this unbalanced longitudinal data set, information from several explanatory variables is recorded, including X_1 =time, X_2 =age, X_3 =smoking habit (the number of packs of cigarettes smoked per day), X_4 =recreational drug use (1, yes; 0, no), X_5 =number of sexual partners, and X_6 =score on center for epidemiological studies of depression scale. The objectives of our analysis are: (a) to identify covariates that really affect the CD4+ cell numbers in the sense that they are statistically significant in either the mean or covariance models, and (b) to estimate the average time course for the HIV-infected men by taking account of measurement errors in the CD4+ cell collection. Ye and Pan [13] analyzed the CD4+ count data with a focus on the second objective and did not include the explanatory variables except the time. Following [13], we propose to use three polynomials in time, one of degree 6 and two cubics, to model the mean μ_{ij}, the generalized autoregressive parameters ϕ_{ijk} and the log-innovation variances $\log\sigma_{ij}^2$. In the meantime, the explanatory variables $X_2, \ldots, X_6$ above and the intercept X_0 are also included in the initial models for the selection purpose. The ordinary maximum likelihood estimation and the penalized maximum likelihood estimation methods using the

Table 13.1 Estimated tuning parameters

Parameters	SCAD	LASSO	Hard-thresholding
$\tau^{(1)}$	0.42	0.01	0.79
$\tau^{(2)}$	0.21	0.01	0.46
$\tau^{(3)}$	0.84	0.04	0.88

Table 13.2 Estimators of the mean parameters $\boldsymbol{\beta}$

Coefficient	MLE	SCAD	LASSO	Hard-thresholding
β_1 (X_0)	776.60(20.96)	776.68(20.31)	775.35 (20.96)	776.60(20.96)
β_2 (X_1)	−209.05(14.24)	−209.10(9.40)	−209.04(14.25)	−209.05(14.24)
β_3 (X_1^2)	−14.47(8.36)	−14.49(8.04)	−14.51(8.37)	−14.47(8.36)
β_4 (X_1^3)	32.68(5.93)	32.74(2.17)	32.72 (5.93)	32.68(5.93)
β_5 (X_1^4)	−1.97(1.05)	−1.97(1.02)	−1.96(1.05)	−1.97(1.05)
β_6 (X_1^5)	−1.84(0.57)	−1.84(0.21)	−1.85(0.55)	−1.84(0.57)
β_7 (X_1^6)	0.25(0.08)	0.26(0.02)	0.26 (0.08)	0.25(0.08)
β_8 (X_2)	0.88(1.34)	0.88(0.007)	0.88 (1.35)	0.88(1.34)
β_9 (X_3)	61.27(5.36)	61.32(6.35)	61.04 (6.30)	61.27(6.36)
β_{10} (X_4)	45.70(18.84)	45.71(18.71)	45.61 (18.84)	45.70(18.84)
β_{11} (X_5)	−3.61(2.09)	−3.60(2.09)	−3.64(2.09)	−3.61(2.09)
β_{12} (X_6)	−2.24(0.80)	−2.30(0.82)	0(−)	−2.24(0.80)

SCAD, Lasso and Hard-thresholding penalty functions are all considered. The unknown tuning parameters $\tau^{(l)}$ $(l = 1, 2, 3)$ of the penalty functions are estimated through using the 5-fold cross-validation principle described in Sect. 13.2.5, and the resulting estimators are summarized in Table 13.1. It is noted that the SCAD penalty function given in (13.4) also involves another parameter a. Here we choose $a = 3.7$ as suggested by Fan and Li [4].

For the mean, generalized autoregressive parameters and log-innovation variances, the estimated regression coefficients and their associated standard errors, in parentheses, by different penalty estimation methods, are presented in Tables 13.2, 13.3, and 13.4. It is noted that in Table 13.3 $\gamma_1, \ldots, \gamma_4$ correspond to the coefficients of the cubic polynomial in time lag, γ_5 is associated with the time-independent covariate X_2, and the other coefficients $\gamma_6, \ldots, \gamma_{13}$ correspond to the time-dependent covariates $X_3, \ldots, X_6$ measured at two different time points, denoted by $(X_{31}, X_{32}),\ldots,(X_{61}, X_{62})$.

From Tables 13.2, 13.3, and 13.4, it is clear that for the mean structure the estimated regression coefficients of the sixth power polynomial in time are statistically significant. For the generalized autoregressive parameters and the innovation variances, the estimated regression coefficients of cubic polynomials in time are significant. This confirms the conclusion drawn by Ye and Pan [13]. Furthermore, Table 13.2 shows that there is little evidence for the association between age and immune response, but the smoking habit and the use of recreational drug have significant positive effects on the CD4+ numbers. In addition, the number of sexual partners seems to have little effect on the immune response, although it shows

Table 13.3 Estimators of the generalized autoregressive parameters $\boldsymbol{\gamma}$

Coefficient	MLE	SCAD	LASSO	Hard-thresholding
γ_1 (X_0)	0.29(0.06)	0.29 (0.02)	0.29 (0.06)	0.29 (0.06)
γ_2 (X_1)	−0.33(0.09)	−0.33(0.02)	−0.33(0.09)	−0.33(0.09)
γ_3 (X_1^2)	0.20(0.04)	0.20 (0.01)	0.20 (0.04)	0.20 (0.04)
γ_4 (X_1^3)	−0.03(0.004)	−0.03(0.002)	−0.03(0.003)	−0.03(0.004)
γ_5 (X_2)	−0.001(0.0008)	0(−)	0(−)	0(−)
γ_6 (X_{31})	−0.01(0.008)	−0.01(0.005)	−0.01(0.007)	−0.01(0.007)
γ_7 (X_{32})	0.007(0.008)	0(−)	0(−)	0(−)
γ_8 (X_{41})	−0.01(0.02)	0.01 (0.06)	0.01 (0.01)	0.01 (0.02)
γ_9 (X_{42})	0.02(0.02)	0.02 (0.07)	0.02 (0.01)	0.02 (0.02)
γ_{10} (X_{51})	0.001(0.002)	0(−)	0(−)	0(−)
γ_{11} (X_{52})	−0.005(0.003)	0(−)	0(−)	0(−)
γ_{12} (X_{61})	0.004(0.0009)	0(−)	0(−)	0(−)
γ_{13} (X_{62})	0.006(0.001)	0(−)	0(−)	0(−)

Table 13.4 Estimators of the log-innovation variance parameters $\boldsymbol{\lambda}$

Coefficient	MLE	SCAD	LASSO	Hard-thresholding
λ_1 (X_0)	11.64(0.07)	11.63 (0.04)	11.63 (0.08)	11.64 (0.07)
λ_2 (X_1)	−0.22(0.03)	−0.22(0.01)	−0.22(0.03)	−0.22(0.03)
λ_3 (X_1^2)	−0.03(0.01)	−0.03(0.04)	−0.03(0.01)	−0.03(0.01)
λ_4 (X_1^3)	−0.02(0.003)	−0.02(0.001)	−0.02(0.004)	−0.02(0.003)
λ_5 (X_2)	−0.005(0.004)	0(−)	0(−)	0(−)
λ_6 (X_3)	0.21(0.02)	0.21 (0.01)	0.21 (0.02)	0.21 (0.02)
λ_7 (X_4)	−0.12(0.07)	−0.12(0.005)	−0.12(0.06)	−0.12(0.07)
λ_8 (X_5)	−0.02(0.008)	−0.02(0.004)	−0.02(0.008)	−0.02(0.009)
λ_9 (X_6)	−0.006(0.003)	0(−)	0(−)	0(−)

some evidence of negative association. Also, there is a negative association between depression symptoms (score) and immune response.

Interestingly, Table 13.3 clearly indicates that except the cubic polynomial in time lag all other covariates do not have significant influences to the generalized autoregressive parameters, implying that the generalized autoregressive parameters are characterized only by the cubic polynomial in time lag. For the log-innovation variances, however, Table 13.4 shows that in addition to the cubic polynomial in time, the smoking habit, the use of recreational drug, and the number of sexual partners do have significant effects, implying that the innovation variances and therefore the within-subject covariances are not homogeneous and are actually dependent on the covariates of interests. Finally, we notice that in this data example the SCAD, Lasso and Hard thresholding penalty based methods perform very similarly in terms of the selected variables.

13.5 Simulation Study

In this section we conduct a simulation study to assess the small sample performance of the proposed procedures. We simulate 100 subjects, each of which has five observations drawn from the multivariate normal distribution $\mathcal{N}_5(\mu_i, \Sigma_i)$, where the mean μ_i and the within-subject covariance matrix Σ_i are formed by the joint models (13.2) in the framework of the modified Cholesky decomposition. We choose the true values of the parameters in the mean, generalized autoregressive parameters and log-innovation variances to be $\boldsymbol{\beta} = (3, 0, 0, -2, 1, 0, 0, 0, 0, -4)^T$, $\boldsymbol{\gamma} = (-4, 0, 0, 2, 0, 0, 0)^T$ and $\boldsymbol{\lambda} = (0, 1, 0, 0, 0, -2, 0)^T$, respectively. We form the mean covariates $\mathbf{x}_{ij} = (x_{ijt})_{t=1}^{10}$ by drawing random samples from the multivariate normal distribution with mean 0 and covariance matrix of AR(1) structure with $\sigma^2 = 1$ and $\rho = 0.5$ ($i = 1, 2, \ldots, 100$; $j = 1, 2, \ldots, 5$). We then form the covariates $\mathbf{z}_{ijk} = (x_{ijt} - x_{ikt})_{t=1}^{7}$ and $\mathbf{h}_{ij} = (x_{ijt})_{t=1}^{7}$ for the generalized autoregressive parameters and the log-innovation variances. Using these values, the mean μ_i and covariance matrix Σ_i are constructed through the modified Cholesky decomposition. The responses $\mathbf{y}_i$ are then drawn from the multivariate normal distribution $\mathcal{N}(\mu_i, \Sigma_i)$ ($i = 1, 2, \ldots, 100$).

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 estimation procedures for finding out the ordinary maximum likelihood estimators and penalized maximum likelihood estimators with SCAD, Lasso and Hard-thresholding penalty functions are considered. The unknown tuning parameters $\tau^{(l)}, l = 1, 2, 3$ for the penalty functions are chosen by a 5-fold cross-validation criterion in the simulation. For each of these methods, the average of zero coefficients over the 1000 simulated data sets is reported in Table 13.5. Note that 'True' in Table 13.5 means the average number of zero regression coefficients that are correctly estimated as zero, and 'Wrong' depicts the average number of non-zero regression coefficients that are erroneously set to zero. In addition, the non-zero parameter estimators, and their associated standard errors as well, are provided in Table 13.6. From those simulation results, it is clear that the SCAD penalty method outperforms the Lasso and Hard thresholding penalty approaches in the sense of correct variable selection rate, which significantly reduces the model uncertainty and complexity.

Table 13.5 Average number of zero regression coefficients

	SCAD		LASSO		Hard-thresholding	
Parameter	True	Wrong	True	Wrong	True	Wrong
$\boldsymbol{\beta}$	5.42	0.00	4.76	0.00	4.92	0.00
$\boldsymbol{\gamma}$	4.18	0.06	3.28	0.08	3.55	0.21
$\boldsymbol{\lambda}$	4.53	0.00	3.70	0.00	4.06	0.00

Table 13.6 Estimators of non-zero regression coefficients

Coefficient	True value	SCAD	LASSO	Hard-thresholding
β_1	3	3.08(0.95)	3.08(0.95)	3.09(0.93)
β_4	−2	−1.94(0.68)	−1.93(0.63)	−1.95(0.65)
β_5	1	0.95(0.32)	0.96(0.39)	0.97(0.39)
β_{10}	−4	−4.12(1.65)	−4.13(1.74)	−4.14(1.75)
γ_1	−4	−4.13(1.88)	−4.07(2.14)	−4.10(2.14)
γ_4	2	1.77(0.79)	1.71(0.85)	1.75(0.85)
λ_2	1	1.05(0.05)	1.03(0.06)	1.03(0.06)
λ_6	−2	−2.20(0.83)	−2.11(0.81)	−2.11(0.82)

13.6 Discussion

Within the framework of joint modelling of mean and covariance structures for longitudinal data, we proposed a variable selection method based on penalized likelihood approaches. Like the mean, the covariance structures may be dependent on various explanatory variables of interest so that simultaneous variable selection to the mean and covariance structures becomes fundamental to avoid the modelling biases and reduce the model complexities.

We have shown that under mild conditions the proposed penalized maximum likelihood estimators of the parameters in the mean and covariance models are asymptotically consistent and normally distributed. Also, we have shown that the SCAD and Hard thresholding penalty based estimation approaches have the oracle property. In other words, they can correctly identify the true models as if the true models would be known in advance. In contrast, the Lasso penalty based estimation method does not share the oracle property. We also considered the case when the number of explanatory variables goes to infinity with the sample size and obtained similar results to the case with finite number of variables.

The proposed method differs from [7] where they only addressed the issue of variable selection in the mean model without modelling the generalized autoregressive parameters and innovation variances. It is also different from [12] where a different decomposition of the covariance matrix, namely moving average coefficient based model, was employed, and the variable selection issue was discussed under the decomposition but with the number of explanatory variables fixed. In contrast, the proposed models and methods in this paper are more flexible, interpretable and practicable.

Appendix A: Penalized Maximum Likelihood Estimation

Firstly, note the first two derivatives of the log-likelihood function $\ell(\boldsymbol{\theta})$ are continuous. Around a given point $\boldsymbol{\theta}_0$, the log-likelihood function can be approximated by

$$\ell(\boldsymbol{\theta}) \approx \ell(\boldsymbol{\theta}_0) + \left[\frac{\partial \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right]^T (\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \left[\frac{\partial^2 \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right] (\boldsymbol{\theta} - \boldsymbol{\theta}_0).$$

Also, given an initial value t_0 we can approximate the penalty function $p'_\tau(t)$ by a quadratic function [4]

$$[p_\tau(|t|)]' = p'_\tau(|t|)\text{sgn}(t) \approx \frac{p'_\tau(|t_0|)t}{t_0}, \quad \text{for } t \approx t_0.$$

In other words,

$$p_\tau(|t|) \approx p_\tau(|t_0|) + \frac{1}{2}p'_\tau(|t_0|)\frac{t^2 - t_0^2}{|t_0|}, \quad \text{for } t \approx t_0.$$

Therefore, the penalized likelihood function (13.3) can be locally approximated, apart from a constant term, by

$$\begin{aligned} Q(\boldsymbol{\theta}) \approx \ell(\boldsymbol{\theta}_0) &+ \left[\frac{\partial \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right]^T (\boldsymbol{\theta} - \boldsymbol{\theta}_0) \\ &+ \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \left[\frac{\partial^2 \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right] (\boldsymbol{\theta} - \boldsymbol{\theta}_0) - \frac{n}{2}\boldsymbol{\theta}^T \Sigma_\tau(\boldsymbol{\theta}_0)\boldsymbol{\theta}, \end{aligned}$$

where

$$\begin{aligned} \Sigma_\tau(\boldsymbol{\theta}_0) = \text{diag}\Big\{ &\frac{p'_{\tau^{(1)}}(|\beta_{01}|)}{|\beta_{01}|}, \ldots, \frac{p'_{\tau^{(1)}}(|\beta_{0p}|)}{|\beta_{0p}|}, \frac{p'_{\tau^{(2)}}(|\gamma_{01}|)}{|\gamma_{01}|}, \ldots, \frac{p'_{\tau^{(2)}}(|\gamma_{0q}|)}{|\gamma_{0q}|}, \\ &\frac{p'_{\tau^{(3)}}(|\lambda_{01}|)}{|\lambda_{01}|}, \ldots, \frac{p'_{\tau^{(3)}}(|\lambda_{0d}|)}{|\lambda_{0d}|} \Big\}, \end{aligned}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_s)^T = (\beta_1, \ldots, \beta_p, \gamma_1, \ldots, \gamma_q, \lambda_1, \ldots, \lambda_d)^T$ and $\boldsymbol{\theta}_0 = (\theta_{01}, \ldots, \theta_{0s})^T = (\beta_{01}, \ldots, \beta_{0p}, \gamma_{01}, \ldots, \gamma_{0q}, \lambda_{01}, \ldots, \lambda_{0d})^T$. Accordingly, the quadratic maximization problem for $Q(\boldsymbol{\theta})$ leads to a solution iterated by

$$\boldsymbol{\theta}_1 \approx \boldsymbol{\theta}_0 + \left\{\frac{\partial^2 \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} - n\Sigma_\tau(\boldsymbol{\theta}_0)\right\}^{-1} \left\{n\Sigma_\tau(\boldsymbol{\theta}_0)\boldsymbol{\theta}_0 - \frac{\partial \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right\}.$$

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

$$\begin{aligned}
-2\ell(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \log|\Sigma_i| + \sum_{i=1}^{n} (\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta})^T \Sigma_i^{-1} (\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}), \\
&= \sum_{i=1}^{n} \log|D_i| + \sum_{i=1}^{n} (\boldsymbol{r}_i - \mathbf{z}_i\boldsymbol{\gamma})^T D_i^{-1} (\boldsymbol{r}_i - \mathbf{z}_i\boldsymbol{\gamma}), \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m_i} \log\sigma_{ij}^2 + \sum_{i=1}^{n}\sum_{j=1}^{m_i} \frac{(r_{ij} - \widehat{r}_{ij})^2}{\sigma_{ij}^2},
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{r}_i &= \mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta} = (r_{i1}, \ldots, r_{im_i})^T, \\
\widehat{r}_{ij} &= \sum_{k=1}^{j-1} \phi_{ijk} r_{ik}, \quad (j = 2, \ldots, m_i) \\
\mathbf{z}_i &= (\mathbf{z}_{i1}, \ldots, \mathbf{z}_{im_i})^T, \\
\mathbf{z}_{ij} &= \sum_{k=1}^{j-1} r_{ik}\mathbf{z}_{ijk}, \quad (j = 2, \ldots, m_i) \\
\mathbf{x}_i &= (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{im_i})^T, \quad (i = 1, \ldots, n).
\end{aligned}$$

Therefore, the resulting score functions are

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

where

$$\begin{aligned}
U_1(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \mathbf{x}_i^T \Sigma_i^{-1} (\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}), \\
U_2(\boldsymbol{\gamma}) &= \sum_{i=1}^{n} \mathbf{z}_i^T D_i^{-1} (\boldsymbol{r}_i - \mathbf{z}_i\boldsymbol{\gamma}), \\
U_3(\boldsymbol{\lambda}) &= \frac{1}{2}\sum_{i=1}^{n} \mathbf{h}_i^T D_i^{-1} (\boldsymbol{\varepsilon}_i^2 - \Sigma_i^2),
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{h}_i &= (\mathbf{h}_{i1}, \ldots, \mathbf{h}_{im_i})^T, \\
\boldsymbol{\varepsilon}_i^2 &= (\varepsilon_{i1}^2, \ldots, \varepsilon_{im_i}^2)^T, \\
\varepsilon_{ij}^2 &= (r_{ij} - \widehat{r}_{ij})^2, \quad (j = 1, \ldots, m_i) \\
\Sigma_i^2 &= (\sigma_{i1}^2, \ldots, \sigma_{im_i}^2)^T.
\end{aligned}$$

According to [13], the Fisher information matrix $\mathscr{I}_n(\boldsymbol{\theta})$ must be block diagonal. In other words, $\mathscr{I}_n(\boldsymbol{\theta}) = \text{diag}(\mathscr{I}_{11}, \mathscr{I}_{22}, \mathscr{I}_{33})$, where

$$\begin{aligned}
\mathscr{I}_{11} &= \sum_{i=1}^{n} \mathbf{x}_i^T \Sigma_i^{-1} \mathbf{x}_i, \\
\mathscr{I}_{22} &= \sum_{i=1}^{n} E(\mathbf{z}_i^T D_i^{-1} \mathbf{z}_i), \\
\mathscr{I}_{33} &= \frac{1}{2} \sum_{i=1}^{n} \mathbf{h}_i^T \mathbf{h}_i.
\end{aligned}$$

By using the Fisher information matrix to approximate the observed information matrix, we obtain the following iteration solution

$$\begin{aligned}
\boldsymbol{\theta}_1 &\approx \boldsymbol{\theta}_0 + \left\{ \frac{\partial^2 \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} - n \Sigma_\tau(\boldsymbol{\theta}_0) \right\}^{-1} \left\{ n \Sigma_\tau(\boldsymbol{\theta}_0) \boldsymbol{\theta}_0 - \frac{\partial \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \right\} \\
&\approx \boldsymbol{\theta}_0 + \{ \mathscr{I}_n(\boldsymbol{\theta}_0) + n \Sigma_\tau(\boldsymbol{\theta}_0) \}^{-1} \{ U(\boldsymbol{\theta}_0) - n \Sigma_\tau(\boldsymbol{\theta}_0) \boldsymbol{\theta}_0 \} \\
&= \{ \mathscr{I}_n(\boldsymbol{\theta}_0) + n \Sigma_\tau(\boldsymbol{\theta}_0) \}^{-1} \{ U(\boldsymbol{\theta}_0) + \mathscr{I}_n(\boldsymbol{\theta}_0) \boldsymbol{\theta}_0 \}.
\end{aligned}$$

Since $\mathscr{I}_n(\boldsymbol{\theta})$ is block diagonal, the above iteration solution is equivalent to

$$\begin{aligned}
\boldsymbol{\beta}_1 &= \left\{ \sum_{i=1}^{n} \mathbf{x}_i^T \Sigma_i^{-1} \mathbf{x}_i + n \Sigma_{\tau^{(1)}}(\boldsymbol{\beta}_0) \right\}^{-1} \left\{ \sum_{i=1}^{n} \mathbf{x}_i^T \Sigma_i^{-1} \mathbf{y}_i \right\}, \\
\boldsymbol{\gamma}_1 &= \left\{ \sum_{i=1}^{n} \mathbf{z}_i^T D_i^{-1} \mathbf{z}_i + n \Sigma_{\tau^{(2)}}(\boldsymbol{\gamma}_0) \right\}^{-1} \left\{ \sum_{i=1}^{n} \mathbf{z}_i^T D_i^{-1} \boldsymbol{r}_i \right\}, \\
\boldsymbol{\lambda}_1 &= \left\{ \sum_{i=1}^{n} \mathbf{h}_i^T \mathbf{h}_i + 2n \Sigma_{\tau^{(3)}}(\boldsymbol{\lambda}_0) \right\}^{-1} \left\{ \sum_{i=1}^{n} \mathbf{h}_i^T D_i^{-1} (\boldsymbol{\varepsilon}_i^2 - \Sigma_i^2 + D_i \log \Sigma_i^2) \right\},
\end{aligned}$$

where all the relevant quantities on the right hand side are evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, and

$$\Sigma_{\tau^{(1)}}(\boldsymbol{\beta_0}) = \text{diag}\{\frac{p'_{\tau^{(1)}}(|\beta_{01}|)}{|\beta_{01}|}, \ldots, \frac{p'_{\tau^{(1)}}(|\beta_{0p}|)}{|\beta_{0p}|}\},$$

$$\Sigma_{\tau^{(2)}}(\boldsymbol{\gamma_0}) = \text{diag}\{\frac{p'_{\tau^{(2)}}(|\gamma_{01}|)}{|\gamma_{01}|}, \ldots, \frac{p'_{\tau^{(2)}}(|\gamma_{0q}|)}{|\gamma_{0q}|}\},$$

$$\Sigma_{\tau^{(3)}}(\boldsymbol{\lambda_0}) = \text{diag}\{\frac{p'_{\tau^{(3)}}(|\lambda_{01}|)}{|\lambda_{01}|}, \ldots, \frac{p'_{\tau^{(3)}}(|\lambda_{0d}|)}{|\lambda_{0d}|}\}.$$

Finally, the following algorithm summarizes the computation of the penalized maximum likelihood estimators of the parameters in the joint mean and covariance models.

Algorithm

0. Take the ordinary least squares estimators (without penalty) $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{\gamma}^{(0)}$ and $\boldsymbol{\lambda}^{(0)}$ of $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$ as their initial values.
1. Given the current values $\{\boldsymbol{\beta}^{(s)}, \boldsymbol{\gamma}^{(s)}, \boldsymbol{\lambda}^{(s)}\}$, update

$$\boldsymbol{r}_i^{(s)} = \mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}^{(s)},\ \phi_{ijk}^{(s)} = \mathbf{z}_{ijk}^T\boldsymbol{\gamma}^{(s)},\ \log[(\sigma_{ij}^2)^{(s)}] = \mathbf{h}_{ij}^T\boldsymbol{\lambda}^{(s)},$$

and then use the above iteration solutions to update $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$ until convergence. Denote the updated results by $\boldsymbol{\gamma}^{(s+1)}$ and $\boldsymbol{\lambda}^{(s+1)}$.

2. For the updated values $\boldsymbol{\gamma}^{(s+1)}$ and $\boldsymbol{\lambda}^{(s+1)}$, form

$$\phi_{ijk}^{(s+1)} = \mathbf{z}_{ijk}^T\boldsymbol{\gamma}^{(s+1)},\ \text{and}\ \log[(\sigma_{ij}^2)^{(s+1)}] = \mathbf{h}_{ij}^T\boldsymbol{\lambda}^{(s+1)},$$

and construct

$$\Sigma_i^{(s+1)} = (T_i^{(s+1)})^{-1}D_i^{(s+1)}[(T_i^{(s+1)})^T]^{-1}.$$

Then update $\boldsymbol{\beta}$ according to

$$\boldsymbol{\beta}^{(s+1)} = \left\{\sum_{i=1}^n \mathbf{x}_i^T(\Sigma_i^{(s+1)})^{-1}\mathbf{x}_i + n\Sigma_{\tau^{(1)}}(\boldsymbol{\beta}^{(s)})\right\}^{-1}\left\{\sum_{i=1}^n \mathbf{x}_i^T(\Sigma_i^{(s+1)})^{-1}\mathbf{y}_i\right\}.$$

3. Repeat Step 1 and Step 2 above until certain convergence criteria are satisfied. For example, it can be considered as convergence if the L_2-norm of the difference of the parameter vectors between two adjacent iterations is sufficiently small.

Appendix B: Proofs of Theorems

Proof of Theorem 13.1 Note that $p_{\tau_n}(0) = 0$ and $p_{\tau_n}(\cdot) > 0$. Obviously, we have

$$\begin{aligned} Q(\boldsymbol{\theta}_0 + n^{-1/2}\mathbf{u}) - Q(\boldsymbol{\theta}_0) \\ &\leq [\ell(\boldsymbol{\theta}_0 + n^{-1/2}\mathbf{u}) - \ell(\boldsymbol{\theta}_0)] - n\sum_{j=1}^{s_1}[p_{\tau_n}(|\theta_{0j} + n^{-1/2}u_j|) - p_{\tau_n}(|\theta_{0j}|)] \\ &= K_1 + K_2. \end{aligned}$$

We consider K_1 first. By using Taylor expansion, we know

$$\begin{aligned} K_1 &= \ell(\boldsymbol{\theta}_0 + n^{-1/2}\mathbf{u}) - \ell(\boldsymbol{\theta}_0) \\ &= n^{-1/2}\mathbf{u}^T\ell'(\boldsymbol{\theta}_0) + \frac{1}{2}n^{-1}\mathbf{u}^T\ell''(\boldsymbol{\theta}^*)\mathbf{u} \\ &= K_{11} + K_{12}, \end{aligned}$$

where $\boldsymbol{\theta}^*$ lies between $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_0 + n^{-1/2}u$. Note the fact that $n^{-1/2}\|\ell'(\boldsymbol{\theta}_0)\| = O_p(1)$. By applying Cauchy-Schwartz inequality, we obtain

$$K_{11} = n^{-1/2}\mathbf{u}^T\ell'(\boldsymbol{\theta}_0) \leq n^{-1/2}\|\ell'(\boldsymbol{\theta}_0)\|\|\mathbf{u}\| = O_p(1).$$

According to Chebyshev's inequality, we know that for any $\varepsilon > 0$,

$$\begin{aligned} P\left\{\frac{1}{n}\|\ell''(\boldsymbol{\theta}_0) - E\ell''(\boldsymbol{\theta}_0)\| \geq \varepsilon\right\} &\leq \frac{1}{n^2\varepsilon^2}E\left\{\sum_{j=1}^{s}\sum_{l=1}^{s}\left(\frac{\partial^2\ell(\boldsymbol{\theta}_0)}{\partial\theta_j\partial\theta_l} - E\frac{\partial^2\ell(\boldsymbol{\theta}_0)}{\partial\theta_j\partial\theta_l}\right)^2\right\} \\ &\leq \frac{Cs^2}{n\varepsilon^2} = o(1) \end{aligned}$$

so that $n^{-1}\|\ell''(\boldsymbol{\theta}_0) - E\ell''(\boldsymbol{\theta}_0)\| = o_p(1)$. It then follows directly that

$$\begin{aligned} K_{12} &= \frac{1}{2}n^{-1}\mathbf{u}^T\ell''(\boldsymbol{\theta}^*)\mathbf{u} = \frac{1}{2}\mathbf{u}^T\{n^{-1}[\ell''(\boldsymbol{\theta}_0) - E\ell''(\boldsymbol{\theta}_0) - \mathscr{I}_n(\boldsymbol{\theta}_0)]\}\mathbf{u}[1 + o_p(1)] \\ &= -\frac{1}{2}\mathbf{u}^T\mathscr{I}(\boldsymbol{\theta}_0)\mathbf{u}[1 + o_p(1)]. \end{aligned}$$

Therefore we conclude that K_{12} dominates K_{11} uniformly in $\|\mathbf{u}\| = C$ if the constant C is sufficiently large.

We then study the term K_2. It follows from Taylor expansion and Cauchy-Schwartz inequality that

$$\begin{aligned} K_2 &= -n\sum_{j=1}^{s_1}[p_{\tau_n}(|\theta_{0j}+n^{-1/2}u_j|)-p_{\tau_n}(|\theta_{0j}|)] \\ &= -\sum_{j=1}^{s_1}\{n^{1/2}p'_{\tau_n}(|\theta_{0j}|)\mathrm{sgn}(\theta_{0j})u_j+\frac{1}{2}p''_{\tau_n}(|\boldsymbol{\theta}_{0j}|)u_j^2[1+o_p(1)]\} \\ &\le \sqrt{s_1}n^{1/2}\|\mathbf{u}\|\max_{1\le j\le s}\{p'_{\tau_n}(|\theta_{0j}|):\theta_{0j}\neq 0\}+2\|\mathbf{u}\|^2\max_{1\le j\le s}\{|p''_{\tau_n}(|\theta_{0j}|)|:\theta_{0j}\neq 0\} \\ &= \sqrt{s_1}n^{1/2}\|\mathbf{u}\|a_n+2\|\mathbf{u}\|^2b_n. \end{aligned}$$

Since it is assumed that $a_n = O_p(n^{-1/2})$ and $b_n \to 0$, we conclude that K_{12} dominates K_2 if we choose a sufficiently large C. Therefore for any given $\varepsilon > 0$, there exists a large constant C such that

$$P\left\{\sup_{\|\mathbf{u}\|=C}Q(\boldsymbol{\theta}_0+n^{-1/2}\mathbf{u})<Q(\boldsymbol{\theta}_0)\right\}\ge 1-\varepsilon,$$

implying that there exists a local maximizer $\widehat{\boldsymbol{\theta}}_n$ such that $\widehat{\boldsymbol{\theta}}_n$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\theta}_0$. The proof of Theorem 13.1 is completed. □

Proof of Theorem 13.2 First, we prove that under the conditions of Theorem 13.2, for any given $\boldsymbol{\theta}^{(1)}$ satisfying $\boldsymbol{\theta}^{(1)}-\boldsymbol{\theta}_0^{(1)} = O_p(n^{-1/2})$ and any constant $C > 0$, we have

$$Q\{((\boldsymbol{\theta}^{(1)})^T,\mathbf{0}^T)^T\} = \max_{\|\boldsymbol{\theta}^{(2)}\|\le Cn^{-1/2}}Q\{((\boldsymbol{\theta}^{(1)})^T,(\boldsymbol{\theta}^{(2)})^T)^T\}.$$

In fact, for any θ_j $(j = s_1+1,\ldots,s)$, using Taylor's expansion we obtain

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\theta})}{\partial\theta_j} &= \frac{\partial\ell(\boldsymbol{\theta})}{\partial\theta_j}-np'_{\tau_n}(|\theta_j|)\mathrm{sgn}(\theta_j) \\ &= \frac{\partial\ell(\boldsymbol{\theta}_0)}{\partial\theta_j}+\sum_{l=1}^{s}\frac{\partial^2\ell(\boldsymbol{\theta}^*)}{\partial\theta_j\partial\theta_l}(\theta_l-\theta_{0l})-np'_{\tau_n}(|\theta_j|)\mathrm{sgn}(\theta_j) \end{aligned}$$

where $\boldsymbol{\theta}^*$ lies between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$. By using the standard argument, we know

$$\frac{1}{n}\frac{\partial\ell(\boldsymbol{\theta}_0)}{\partial\theta_j} = O_p(n^{-1/2}) \text{ and } \frac{1}{n}\left\{\frac{\partial^2\ell(\boldsymbol{\theta}_0)}{\partial\theta_j\partial\theta_l}-E\left(\frac{\partial^2\ell(\boldsymbol{\theta}_0)}{\partial\theta_j\partial\theta_l}\right)\right\} = o_p(1).$$

Note $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| = O_p(n^{-1/2})$. We then have

$$\frac{\partial Q(\boldsymbol{\theta})}{\partial \theta_j} = n\tau_n\{-\tau_n^{-1} p'_{\tau_n}(|\theta_j|)\mathrm{sgn}(\theta_j) + O_p(n^{-1/2}\tau_n^{-1})\}.$$

According to the assumption in Theorem 13.2, we obtain

$$\liminf_{n\to\infty} \liminf_{t\to 0+} \frac{p'_{\tau_n}(t)}{\tau_n} > 0 \ \text{ and } \ n^{-1/2}\tau_n^{-1} = (\sqrt{n}\tau_n)^{-1} \to 0,$$

so that

$$\frac{\partial Q(\boldsymbol{\theta})}{\partial \theta_j} \begin{cases} < 0, \text{ for } \ 0 < \theta_j < Cn^{-1/2}; \\ > 0, \text{ for } \ -Cn^{-1/2} < \theta_j < 0. \end{cases}$$

Therefore $Q(\boldsymbol{\theta})$ achieves its maximum at $\boldsymbol{\theta} = ((\boldsymbol{\theta}^{(1)})^T, \mathbf{0}^T)^T$ and the first part of Theorem 13.2 has been proved. □

Second, we discuss the asymptotic normality of $\widehat{\boldsymbol{\theta}}_n^{(1)}$. From Theorem 13.1 and the first part of Theorem 13.2, there exists a penalized maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_n^{(1)}$ that is the $\sqrt{n}$-consistent local maximizer of the function $Q\{((\boldsymbol{\theta}^{(1)})^T, \mathbf{0}^T)^T\}$. The estimator $\widehat{\boldsymbol{\theta}}_n^{(1)}$ must satisfy

$$\begin{aligned} 0 &= \frac{\partial Q(\boldsymbol{\theta})}{\partial \theta_j}\bigg|_{\boldsymbol{\theta}=\binom{\widehat{\boldsymbol{\theta}}_n^{(1)}}{\mathbf{0}}} = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_j}\bigg|_{\boldsymbol{\theta}=\binom{\widehat{\boldsymbol{\theta}}_n^{(1)}}{\mathbf{0}}} - np'_{\tau_n}(|\widehat{\theta}_{nj}^{(1)}|)\mathrm{sgn}(\widehat{\theta}_{nj}^{(1)}) \\ &= \frac{\partial \ell(\boldsymbol{\theta}_0)}{\partial \theta_j} + \sum_{l=1}^{s_1}\left\{\frac{\partial^2 \ell(\boldsymbol{\theta}_0)}{\partial \theta_j \partial \theta_l} + o_p(1)\right\}(\widehat{\theta}_{nl}^{(1)} - \theta_{0l}^{(1)}) \\ &\quad -np'_{\tau_n}(|\theta_{0j}^{(1)}|)\mathrm{sgn}(\theta_{0j}^{(1)}) - n\{p''_{\tau_n}(|\theta_{0j}^{(1)}|) + o_p(1)\}(\widehat{\theta}_{nj}^{(1)} - \theta_{0j}^{(1)}). \end{aligned}$$

In other words, we have

$$\left\{-\frac{\partial^2 \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^{(1)} \partial (\boldsymbol{\theta}^{(1)})^T} + nA_n + o_p(1)\right\}(\widehat{\boldsymbol{\theta}}_n^{(1)} - \boldsymbol{\theta}_0^{(1)}) + \boldsymbol{c}_n = \frac{\partial \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^{(1)}}.$$

Using the Liapounov form of the multivariate central limit theorem, we obtain

$$\frac{1}{\sqrt{n}}\frac{\partial \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^{(1)}} \to \mathscr{N}_{s_1}(\mathbf{0}, \mathscr{I}^{(1)})$$

in distribution. Note that

$$\frac{1}{n}\left\{\frac{\partial^2 \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^{(1)} \partial (\boldsymbol{\theta}^{(1)})^T} - E\left(\frac{\partial^2 \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^{(1)} \partial (\boldsymbol{\theta}^{(1)})^T}\right)\right\} = o_p(1),$$

it follows immediately by using Slustsky's theorem that

$$\sqrt{n}(\bar{\mathscr{I}}_n^{(1)})^{-1/2}(\bar{\mathscr{I}}_n^{(1)} + A_n)\left\{(\widehat{\boldsymbol{\theta}}_n^{(1)} - \boldsymbol{\theta}_0^{(1)}) + (\bar{\mathscr{I}}_n^{(1)} + A_n)^{-1}\boldsymbol{c}_n\right\} \to \mathscr{N}_{s_1}(\mathbf{0}, I_{s_1})$$

in distribution. The proof of Theorem 13.2 is complete. □

Proof of Theorem 13.3 Let $\alpha_n = (n/s_n)^{-1/2}$. Note $p_{\tau_n}(0) = 0$ and $p_{\tau_n}(\cdot) > 0$. We then have

$$\begin{aligned} Q(\boldsymbol{\theta}_0 + \alpha_n \mathbf{u}) - Q(\boldsymbol{\theta}_0) &\le [\ell(\boldsymbol{\theta}_0 + \alpha_n \mathbf{u}) - \ell(\boldsymbol{\theta}_0)] - n\sum_{j=1}^{s_{1n}}[p_{\tau_n}(|\theta_{0j} + \alpha_n u_j|) - p_{\tau_n}(|\theta_{0j}|)] \\ &= K_1 + K_2. \end{aligned}$$

Using Taylor's expansion, we obtain

$$\begin{aligned} K_1 &= \ell(\boldsymbol{\theta}_0 + \alpha_n \mathbf{u}) - \ell(\boldsymbol{\theta}_0) \\ &= \alpha_n \mathbf{u}^T \ell'(\boldsymbol{\theta}_0) + \frac{1}{2}\alpha_n^2 \mathbf{u}^T \ell''(\boldsymbol{\theta}_0^*)\mathbf{u} \\ &= K_{11} + K_{12}, \end{aligned}$$

where $\boldsymbol{\theta}_0^*$ lies between $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_0 + \alpha_n \mathbf{u}$. Note that $\|\ell'(\boldsymbol{\theta}_0)\| = O_p(\sqrt{ns_n})$. By using Cauchy-Schwartz inequality, we conclude that

$$|K_{11}| = |\alpha_n \mathbf{u}^T \ell'(\boldsymbol{\theta}_0)| \le \alpha_n \|\ell'(\boldsymbol{\theta}_0)\|\|\mathbf{u}\| = O_p(\alpha_n (ns_n)^{1/2})\|\mathbf{u}\| = O_p(n\alpha_n^2)\|\mathbf{u}\|.$$

According to Chebyshev's inequality, for any $\varepsilon > 0$ we have

$$\begin{aligned} P\{\left\|\frac{s_n}{n}\left(\ell''(\boldsymbol{\theta}_0) - E\ell''(\boldsymbol{\theta}_0)\right)\right\| \ge \varepsilon\} &\le \frac{1}{\varepsilon^2} E\left(\left\|\frac{s_n}{n}\left(\ell''(\boldsymbol{\theta}_0) - E\ell''(\boldsymbol{\theta}_0)\right)\right\|^2\right) \\ &= \frac{s_n^2}{n^2\varepsilon^2} E\left\{\sum_{j=1}^{s_n}\sum_{l=1}^{s_n}\left(\frac{\partial^2 \ell(\boldsymbol{\theta}_0)}{\partial \theta_j \partial \theta_l} - E\frac{\partial^2 \ell(\boldsymbol{\theta}_0)}{\partial \theta_j \partial \theta_l}\right)^2\right\} \\ &\le \frac{C s_n^4}{n\varepsilon^2} = o(1), \end{aligned}$$

which implies that $\frac{s_n}{n}\left\|\ell''(\boldsymbol{\theta}_0) - E\ell''(\boldsymbol{\theta}_0)\right\| = o_p(1)$. It then follows that

$$K_{12} = \frac{1}{2}\alpha_n^2 \mathbf{u}^T \ell''(\boldsymbol{\theta}_0^*)\mathbf{u} = \frac{1}{2}n\alpha_n^2 \mathbf{u}^T \left\{[\frac{1}{n}(\ell''(\boldsymbol{\theta}_0) - E\ell''(\boldsymbol{\theta}_0)) - \mathscr{I}_n(\boldsymbol{\theta}_0)]\right\}\mathbf{u}[1 + o_p(1)]$$
$$= -\frac{1}{2}n\alpha_n^2 \mathbf{u}^T \mathscr{I}(\boldsymbol{\theta}_0)\mathbf{u}[1 + o_p(1)].$$

Therefore we know that K_{12} dominates K_{11} uniformly in $\|\mathbf{u}\| = C$ for a sufficiently large constant C.

We now turn to K_2. It follows from Taylor's expansion that

$$K_2 = -n\sum_{j=1}^{s_{1n}}[p_{\tau_n}(|\theta_{0j} + \alpha_n u_j|) - p_{\tau_n}(|\theta_{0j}|)]$$
$$= -\sum_{j=1}^{s_{1n}}\{n\alpha_n p'_{\tau_n}(|\theta_{0j}|)\mathrm{sgn}(\theta_{0j})u_j + \frac{1}{2}n\alpha_n^2 p''_{\tau_n}(|\theta_{0j}|)u_j^2[1 + o_p(1)]\}$$
$$\leq \sqrt{s_{1n}}n\alpha_n\|\mathbf{u}\| \max_{1\leq j\leq s_n}\{p'_{\tau_n}(|\theta_{0j}|) : \theta_{0j} \neq 0\}$$
$$+2n\alpha_n^2\|\mathbf{u}\|^2 \max_{1\leq j\leq s_n}\{|p''_{\tau_n^{(j)}}(|\theta_{0j}|)| : \theta_{0j} \neq 0\}$$
$$\leq \sqrt{s_n}n\alpha_n\|\mathbf{u}\|a_n^* + 2n\alpha_n^2\|\mathbf{u}\|^2 b_n^*$$
$$= n\alpha_n^2\|\mathbf{u}\|O_p(1) + 2n\alpha_n^2\|\mathbf{u}\|^2 b_n^*.$$

Since $b_n^* \to 0$ as $n \longrightarrow 0$, it is clear that K_{12} dominates K_2 if a sufficiently large constant C is chosen. In other words, for any given $\varepsilon > 0$ there exists a large constant C such that

$$P\left\{\sup_{\|\mathbf{u}\|=C} Q(\boldsymbol{\theta}_0 + \alpha_n\mathbf{u}) < Q(\boldsymbol{\theta}_0)\right\} \geq 1 - \varepsilon$$

as long as n is large enough. This implies that there exists a local maximizer $\widehat{\boldsymbol{\theta}}_n$ in the ball $\{\boldsymbol{\theta}_0 + \alpha_n\mathbf{u} : \|\mathbf{u}\| \leq C\}$ such that $\widehat{\boldsymbol{\theta}}_n$ is a $\sqrt{n/s_n}$-consistent estimator of $\boldsymbol{\theta}_0$. The proof of Theorem 13.3 is completed. □

Proof of Theorem 13.4 The proof of Theorem 13.4 is similar to that of Theorem 13.2. In what follows we only give a very brief proof. First, it is easy to show that under the conditions of Theorem 13.4, for any given $\boldsymbol{\theta}^{(1)}$ satisfying $\|\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}_0^{(1)}\| = O_p((n/s_n)^{-1/2})$ and any constant C, the following equality holds

$$Q\{((\boldsymbol{\theta}^{(1)})^T, \mathbf{0}^T)^T\} = \max_{\|\boldsymbol{\theta}^{(2)}\|\leq C(n/s_n)^{-1/2}} Q\{((\boldsymbol{\theta}^{(1)})^T, (\boldsymbol{\theta}^{(2)})^T)^T\}.$$

Based on this fact and Theorem 13.3, there exists an $\sqrt{n/s_n}$-consistent estimator $\widehat{\boldsymbol{\theta}}_n^{(1)}$ that is the local maximizer of $Q\{((\boldsymbol{\theta}^{(1)})^T, \mathbf{0}^T)^T\}$. Let $\bar{\mathscr{I}}_n^{(1)} = \mathscr{I}_n^{(1)}/n$. Similar to the proof of Theorem 13.2, we can show that

$$(\bar{\mathscr{I}}_n^{(1)} + A_n^*)(\widehat{\boldsymbol{\theta}}_n^{(1)} - \boldsymbol{\theta}_0^{(1)}) + \boldsymbol{c}_n^* = \frac{1}{n}\frac{\partial \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^{(1)}} + o_p(\frac{1}{\sqrt{n}}),$$

so that

$$\begin{aligned}&\sqrt{n}M_n(\bar{\mathscr{I}}_n^{(1)})^{-1/2}(\bar{\mathscr{I}}_n^{(1)} + A_n^*)\{(\widehat{\boldsymbol{\theta}}_n^{(1)} - \boldsymbol{\theta}_0^{(1)}) + (\bar{\mathscr{I}}_n^{(1)} + A_n^*)^{-1}\boldsymbol{c}_n^*\}\\ &= \frac{1}{\sqrt{n}}M_n(\bar{\mathscr{I}}_n^{(1)})^{-1/2}\frac{\partial \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^{(1)}} + o_p(M_n(\bar{\mathscr{I}}_n^{(1)})^{-1/2}).\end{aligned}$$

By using Lindeberg-Feller central limit theorem, we can show that

$$\frac{1}{\sqrt{n}}M_n(\bar{\mathscr{I}}_n^{(1)})^{-1/2}\frac{\partial \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^{(1)}}$$

has an asymptotic multivariate normal distribution. The result in Theorem 13.4 follows immediately according to Slustsky's theorem. The proof of Theorem 13.4 is complete. □

Acknowledgments We would like to thank the editors and an anonymous referee for their very constructive comments and suggestions, which makes the paper significantly improved. This research is supported by a research grant from the Royal Society of the UK (R124683).

References

1. Antoniadis, A.: Wavelets in statistics: a review (with discussion). J. Ital. Stat. Assoc. **6**, 97–144 (1997)
2. Chiu, T.Y.M., Leonard, T., Tsui, K.W.: The matrix-logarithm covariance model. J. Am. Stat. Assoc. **91**, 198–210 (1996)
3. Diggle, P.J., Heagerty, P.J., Liang, K.Y., Zeger, S.L.: Analysis of Longitudinal Data, 2nd edn. Oxford University, Oxford (2002)
4. Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. **96**, 1348–60 (2001)
5. Fan, J., Peng, H.: Nonconcave Penalized likelihood with a diverging number of parameters. Ann. Stat. **32**, 928–61 (2004)
6. Frank, I.E., Friedman, J.H.: A statistical view of some chemometrics regression tools. Technometrics **35**, 109–148 (1993)
7. Lee, J., Kim, S., Jhong, J., Koo, J.: Variable selection and joint estimation of mean and covariance models with an application to eQTL data. Comput. Math. Methods Med. **2018**, 13 (2018)
8. Pan, J., MacKenzie, G.: Model selection for joint mean-covariance structures in longitudinal studies. Biometrika **90**, 239–44 (2003)

9. Pourahmadi, M.: Joint mean-covariance models with applications to lontidinal data: unconstrained parameterisation. Biometrika **86**, 677–90 (1999)
10. Pourahmadi, M.: Maximum likelihood estimation fo generalised linear models for multivariate normal covariance matrix. Biometrika **87**, 425–35 (2000)
11. Tibshirani, R.: Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. B **58**, 267–88 (1996)
12. Xu, D., Zhang, Z., Wu, L.: Joint variable selection of mean-covariance model for longitudinal data. Open J. Stat. **3**, 27–35 (2013)
13. Ye, H.J., Pan, J.: Modelling of covariance structures in generalized estimating equations for longitudinal data. Biometrika **93**, 927–41 (2006)
14. Zou, H.: The adaptive Lasso and its oracle properties. J. Am. Stat. Assoc. **101**, 1418–29 (2006)