Statistica Sinica Preprint No: SS-11-251R3	
Title	Variable selection in robust joint mean and covariance model for longitudinal data analysis
Manuscript ID	SS-11-251R3
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.2011.251
Complete List of Authors	Xueying Zheng Wing Kam Fung and Zhongyi Zhu
Corresponding Author	Zhongyi Zhu
E-mail	zyzhu62@yahoo.com.cn
Notice: Accepted version subject to English editing.	

VARIABLE SELECTION IN ROBUST JOINT MEAN AND COVARIANCE MODEL FOR LONGITUDINAL DATA ANALYSIS

Xueying Zheng[1], Wing Kam Fung[1] and Zhongyi Zhu[2]

[1]*Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong*
[2]*Department of Statistics, Fudan University, Shanghai, China*

Abstract: In longitudinal data analysis, a correct specification of the within-subject covariance matrix cultivates an efficient estimation for mean regression coefficients. In this article, we consider robust variable selection method in a joint mean and covariance model. We propose a set of penalized robust generalized estimating equations to estimate simultaneously the mean regression coefficients, the generalized autoregressive coefficients and innovation variances introduced by the modified Cholesky decomposition. The set of estimating equations select important covariate variables in both mean and covariance models together with the estimating procedure. Under some regularity conditions, we develop the oracle property of the proposed robust variable selection method. Finally, a simulation study and a detailed real data analysis are carried out to assess and illustrate the small sample performance, which show that the proposed method performs favorably by combining the robustifying and penalized estimating techniques together in the joint mean and covariance model.

Key words and phrases: Covariance matrix, Penalized generalized estimating equation, Longitudinal data, Modified Cholesky decomposition, Robustness, Variable selection.

1. Introduction

Longitudinal data arise more and more frequently in a vast of scientific domains, which seek for insightful and comprehensive research in a branch of statistical modeling. Different from other types of data, we often assume independence among distinct subjects but dependence within each subject. Consequently, the within-subject correlation raises a fundamental challenge for the analysis of longitudinal data. The work of Liang and Zeger (1986) is a milestone in the development of methodology for longitudinal data analysis that they proposed the generalized estimating equations (GEE) for estimation of generalized linear re-

gression coefficients. The main advantage of their method is well-known that even the within-subject correlation is treated as a nuisance parameter with an assumed parsimonious structure, GEE still brings about a consistent estimator for the mean regression model. Following the direction of GEE, Qu et al. (2000) raised the quadratic inference function (QIF) to enhance the efficiency by moving a step forward into considering the structure of the covariance matrix. Taking robustification into account, He et al. (2005) proposed the robust GEE method to prevent the unexpected influence from outliers in a longitudinal data set.

Ignoring the within-subject correlation is not appropriate since it will probably result in an inefficient estimator of a regression model. In practice, the within-subject covariance structure itself may be of scientific interest. Most frequently discussed relevant topics include component analysis and factor analysis in multivariate statistical problems. Recent important research on estimation of covariance matrix include, but not limited to, Rothman et al. (2009), El Karoui (2008) and Bickel and Levina (2008), most of which directly deal with individual elements of a covariance matrix.

Similar to the mean regression, covariances may be dependent on various explanatory variables. Following this idea, Pourahmadi (1999, 2000) proposed a joint mean and covariance regression model by decomposing the covariance matrix which employed two sets of new parameters, generalized autoregressive coefficients and innovation variances. Ye and Pan (2006) extended the joint model under the framework of generalized estimating equations which required no assumptions on the distribution of the data. By introducing the generalized autoregressive coefficients and innovation variances which have proper statistical interpretation, their joint model released the constraint of positive definiteness in estimation for the covariance matrix. Instead, since the covariance matrix was treated as crucial as the mean, three generalized estimating equations were proposed to estimate the covariance matrix and the mean simultaneously.

A number of developments have been observed after the publication of the joint mean and covariance model, see Fan et al. (2007), Fan and Wu (2008) and Xu and Mackenzie (2012). Leng et al. (2010) generalized Ye and Pan's model to the semiparametric joint mean and covariance model. Mao et al. (2011) extended Leng et al. (2010)'s work further into the generalized partially linear

varying coefficient model. Zheng et al. (2013) extended the robust estimating equation in He et al. (2005) to the joint mean and covariance model by creating three robust estimating equations to hinder the effect of outliers in both mean and covariance estimation.

Variable selection is a technique of selecting a subset of relevant covariates for constructing reliable statistical models. Many popular variable selection methods are based on the penalized likelihood or penalized estimating equations. Commonly used penalties include, but not limited to, LASSO, ALASSO (adaptive lasso in Zou, 2006), SCAD and Hard penalties. In the scope of longitudinal data analysis, Fu (2003) proposed the penalized generalized estimating equation with LASSO penalty. Other variable selection criteria include AIC and C_p, which have been extended by Pan (2001) and Cantoni et al. (2005) respectively, to the case of longitudinal data under the framework of GEE.

In contrast to the prosperous research on variable selection in the mean model, limited research work has been found for covariance variable selection or identification. Jeng and Daye (2011) noticed that to curve the sparsity in the covariance matrix can improve the efficiency on mean estimation, especially in high dimensional problems. They said their method was a marriage between covariance regularization and variable selection. However, their main objective still focused on the mean estimation while the positive definiteness of the covariance matrix cannot be guaranteed. Within the framework of the joint mean and covariance model, Huang et al. (2006) proposed covariance selection and estimation via the penalized normal likelihood, because they were aware that imposing a penalty would reduce the risk of using too many parameters to capture the dependence. Kou and Pan (2011) proposed a penalized maximum likelihood method for the joint model that they employed variable selection in both models simultaneously, in which the covariance matrix was treated as equal importance as the mean.

In this paper, we aim to develop a penalized robust estimating equations based method to select important explanatory variables that make a remarkable contribution to the joint mean and covariance model for longitudinal data analysis. It has been illustrated by simulation studies that both the classical GEE and the joint mean and covariance model are sensitive to outliers, see details in He et al. (2005) and Zheng et al. (2013). Nevertheless, the discussion on robust variable

selection methods is relatively limited. We consider commonly used SCAD and ALASSO penalty, and show the oracle property of the proposed robust variable selection method. Simulation studies show that even a tiny contamination results in a poor performance in both the estimation and variable selection in the non-robust joint model, which verifies the necessity of robustification consideration. In real data analysis, the robust variable selection procedure has smaller standard errors for the mean coefficients estimation, for which two possible reasons may be attributed: sensible covariance matrix modeling plus accommodation of outliers in both subject and observation levels. Moreover, for outlier detection, we have found some outliers which have not been identified before.

The remainder of this article is organized as follows. Section 2 describes the main model and asymptotic properties. Simulation results with illustrations are given in section 3. In section 4, we implement the proposed method to a hormone data set.

2. Robust variable selection in joint mean and covariance model

2.1 Joint mean and covariance model

Suppose that we have a sample of m subjects. Let $y_i = (y_{i1}, \ldots, y_{in_i})^T$ be the n_i repeated observations at time point $t_i = (t_{i1}, \ldots, t_{in_i})^T$ of the ith subject. Denote $E(y_i) = \mu_i = (\mu_{i1}, \ldots, \mu_{in_i})^T$ and $Cov(y_i) = \Sigma_i$ as the $n_i \times 1$ mean vector and $n_i \times n_i$ covariance matrix of y_i respectively.

To release the constraint of positive definiteness in estimation of the covariance matrix Σ_i, we implement a modified Cholesky decomposition by introducing a unique lower triangular matrix Φ_i with 1's being the diagonal entries and a unique diagonal matrix D_i with positive diagonals such that

$$\Phi_i \Sigma_i \Phi_i^T = D_i.$$

A valuable feature of this decomposition is that both Φ_i and D_i have clear statistical interpretation. The lower-diagonal entries of Φ_i are the negatives of the autoregressive coefficients ϕ_{ijk} defined in

$$\hat{y}_{ij} = \mu_{ij} + \sum_{k=1}^{j-1} \phi_{ijk}(y_{ik} - \mu_{ik}),$$

which is the linear least squares predictor of y_{ij} based on its predecessors $y_{i(j-1)}, \ldots, y_{i1}$. The diagonal entries σ_{ij}^2 of D_i can be seen as the innovation variance $\sigma_{ij}^2 = var(\varepsilon_{ij})$, where $\varepsilon_{ij} = y_{ij} - \hat{y}_{ij}$.

We adopt three linear models for the mean, generalized autoregressive parameters and innovation variances that developed by Ye and Pan (2006):

$$\mu_{ij} = x_{ij}^T\beta,\ \phi_{ijk} = g_{ijk}^T\gamma,\ \log \sigma_{ij}^2 = z_{ij}^T\lambda, \tag{2.1}$$

where x_{ij}, g_{ijk} and z_{ij} are $p \times 1$, $q \times 1$ and $d \times 1$ vectors of covariates, and β, γ and λ are associated parameters. The covariates g_{ijk} and z_{ij} may contain baseline covariates, time and associated interactions. The log-linear model of the innovation variance follows the idea of Cook and Weisberg's (1983) model for the variance. They proposed that the variance often depended on the values of one or more of the explanatory variables or on additional relevant quantities such as time or spatial ordering. Cook and Weisberg therefore developed the log-linear model that allows for dependence of the variance on an arbitrary set of variables. In practice, the whole set of the covariates is difficult to define. An orthogonal form for the polynomial of the time is recommended as the covariate for the autoregressive component by Ye and Pan (2006):

$$g_{ijk} = (1,\ (t_{ij} - t_{ik}),\ (t_{ij} - t_{ik})^2, \cdots,\ (t_{ij} - t_{ik})^{q-1})^T.$$

The linear assumption in (2.1) is not the only choice for the joint modeling. For this concern, we can introduce quadratic assumptions or use nonparametric model and semiparametric model after the decomposition, see Mao et al. (2011) and Leng et al. (2010). Our robust method can also be adopted in those semiparametric model behind the covariance matrix parameters. However, this may further complicate the model that we prefer starting from the linear assumption.

In the models above, estimation for generalized autoregressive coefficients and innovation variances are treated as important as the estimation for the mean. Denote $\theta = (\theta_1, \cdots, \theta_s)^T = (\beta_1, \cdots, \beta_p; \gamma_1, \cdots, \gamma_1; \lambda_q, \cdots, \lambda_d)^T$, where $s = p + q + d$. To select important subsets of the covariates, we need the assumption that all interested explanatory variables, together with their interactions are involved in the initial model. By using the same λ, the proposed method is applicable

for correlated data as long as the correlated (or longitudinal) measurements have similar correlation structure between clusters.

2.2 Penalized robust generalized estimating equations

We propose the following penalized robustified generalized estimating equations

$$U(\theta) = ([U_1(\beta)]^T,\ [U_2(\gamma)]^T,\ [U_3(\lambda)]^T)^T,$$

for the mean, generalized autoregressive parameters and innovation variances, respectively:

$$U_1(\beta) = \sum_{i=1}^{m} X_i^T (V_i^\beta)^{-1} h_i^\beta(\mu_i(\beta)) - m q_{\tau^{(1)}}(|\beta|)\mathrm{sgn}(\beta) = 0, \quad (2.2)$$

$$U_2(\gamma) = \sum_{i=1}^{m} T_i^T (V_i^\gamma)^{-1} h_i^\gamma(\hat{r}_i(\gamma)) - m q_{\tau^{(2)}}(|\gamma|)\mathrm{sgn}(\gamma) = 0, \quad (2.3)$$

$$U_3(\lambda) = \sum_{i=1}^{m} Z_i^T D_i (V_i^\lambda)^{-1} h_i^\lambda(\sigma_i^2(\lambda)) - m q_{\tau^{(3)}}(|\lambda|)\mathrm{sgn}(\lambda) = 0, \quad (2.4)$$

where $h_i^\beta(\mu_i(\beta)) = W_i^\beta[\psi^\beta(\mu_i(\beta)) - C_i^\beta(\mu_i(\beta))]$, $h_i^\gamma(\hat{r}_i(\gamma)) = W_i^\gamma[\psi^\gamma(\hat{r}_i(\gamma)) - C_i^\gamma(\hat{r}_i(\gamma))]$ and $h_i^\lambda(\sigma_i^2(\lambda)) = W_i^\lambda[\psi^\lambda(\sigma_i^2(\lambda)) - C_i^\lambda(\sigma_i^2(\lambda))]$ act as the core of the estimating equations with $X_i = (x_{i1}^T, \ldots, x_{in_i}^T)^T$, $Z_i = (z_{i1}^T, \ldots, z_{in_i}^T)^T$, $g_{ij} = (g_{ij1}^T, \ldots, g_{ij(j-1)}^T)^T$ and $T_i = (g_{i1}^T, \ldots, g_{in_i}^T)^T$.

To be clear, we adopt exactly the same style of penalized estimating equations for mean parameter and the generalized autoregressive coefficients and innovation variances generated from the covariance decomposition. In other words, equations (2.3) and (2.4) are in agreement with that in equation (2.2), in which r_i in $U_2(\gamma)$ and ε_i^2 in $U_3(\lambda)$ play the role similar to that of y_i in $U_1(\beta)$ and can be viewed as working responses. Different from the other joint mean and covariance modeling procedure, the covariance estimation is treated as crucial as the mean estimation in the proposed model, which has the same core spirit as the joint model in Ye and Pan (2006).

We specify items in (2.2)–(2.4) one by one. In $U_2(\gamma)$, r_i and $\hat{r}_i$ are $n_i \times 1$ vectors with jth components $r_{ij} = y_{ij} - \mu_{ij}$ and $\hat{r}_{ij} = E(r_{ij}|r_{i1}, \ldots, r_{i(j-1)}) = \sum_{k=1}^{j-1} \phi_{ijk} r_{ik}$, where $\sum_{k=1}^{0}$ as zero when $j = 1$. In $U_3(\lambda)$, $\varepsilon_{ij} = y_{ij} - \hat{y}_{ij}$ and ε_i^2 and σ_i^2 are $n_i \times 1$ vectors with jth components ε_{ij}^2 and σ_{ij}^2 respectively, where

in fact $E(\varepsilon_i^2) = \sigma_i^2$. Moreover, $T_i^T = \partial \hat{r}_i^T / \partial \gamma$ is the $q \times n_i$ matrix with the jth column $\partial \hat{r}_{ij}/\partial\gamma = \sum_{k=1}^{j-1} r_{ik} g_{ijk}$ and $D_i = \text{diag}\{\sigma_{i1}^2, \ldots, \sigma_{in_i}^2\}$.

Then we define V_i's in (2.2)–(2.4). $V_i^\beta = A_i^{-1/2}\Sigma_i$, A_i is the diagonal elements of Σ_i, $V_i^\gamma = D_i^{1/2}$, $V_i^\lambda = \widetilde{A}_i^{-1/2}\widetilde{\Sigma}_i$ and $\widetilde{A}_i$ is the diagonal elements of $\widetilde{\Sigma}_i$. The sandwich working covariance structure $\widetilde{\Sigma}_i = B_i^{1/2} R_i(\delta) B_i^{1/2}$ can be used to model the true $\widetilde{\Sigma}_i = Cov(\varepsilon_i^2)$ with $B_i = 2\text{diag}\{\sigma_{i1}^4, \ldots, \sigma_{in_i}^4\}$ and $R_i(\delta)$ mimics the correlation between ε_{ij}^2 and ε_{ik}^2 by introducing a new parameter δ. This idea was proposed by Ye and Pan (2006), although they did not provide particular suggestion on how to choose the structure or the value of δ.

Experience tells us that the parameter δ has little effect on the estimation in practice. Considering AR(1) structure, we can estimate δ by the slope from the regression of $\log(\hat{\varepsilon}_{ij}^2, \hat{\varepsilon}_{ik}^2)$ on $\log(|t_j - t_k|)$. Details can be found in Example 4 in Liang and Zegar (1986). In our simulation and real data analysis, the estimate of δ always falls in the interval $[0, 0.3]$. Moreover, results in Table 3 for the simulation study 2 also imply that we can ignore the difference between the independent and AR(1) structure for $R_i(\delta)$.

Penalized robust generalized estimating equations are distinguished from conventional generalized estimating equations in two aspects. First, the undesirable influence of outliers is controlled. In the core of the estimating equations, $\psi^\beta(\mu_i) = \psi(A_i^{-1/2}(y_i - \mu_i))$, $\psi^\gamma(\hat{r}_i) = \psi(D_i^{-1/2}(r_i - \hat{r}_i))$ and $\psi^\lambda(\sigma_i^2) = \psi(\widetilde{A}_i^{-1/2}(\varepsilon_i^2 - \sigma_i^2))$. The function $\psi(\cdot)$ is chosen to limit the influence of outliers in the response variable, and a common choice is Huber's score function $\psi_c(x) = \min\{c, \max\{-c, x\}\}$ for some constant c, say $c = 2$ in our implementation. To ensure Fisher consistency, we use $C_i^\beta(\mu_i) = E[\psi(A_i^{-1/2}(y_i - \mu_i))]$, $C_i^\gamma(\hat{r}_i) = E[\psi(D_i^{-1/2}(r_i - \hat{r}_i))]$ and $C_i^\lambda(\sigma_i^2) = E[\psi(\widetilde{A}_i^{-1/2}(\varepsilon_i^2 - \sigma_i^2))]$. Once assumed that y_i's are under the normal distribution, the three expectations depend only on the choice of constant c in Huber's score function. Another important robust way is through assigning the weights to each subject by diagonal weighting matrices $W_i^\beta = \text{diag}(w_{i1}^\beta, \ldots, w_{in_i}^\beta)$, $W_i^\gamma = \text{diag}(w_{i1}^\gamma, \ldots, w_{in_i}^\gamma)$ and $W_i^\lambda = \text{diag}(w_{i1}^\lambda, \ldots, w_{in_i}^\lambda)$. Similar to Qin et al. (2009), the weight function w_{ij} is chosen to be the Mahalanobis distance in form of

$$w_{ij} = w(p_{ij}) = \min\Big\{1, \ [\frac{b_0}{(p_{ij} - m_p)^T S_p^{-1} (p_{ij} - m_p)}]^{\rho/2}\Big\},$$

with $\rho \geq 1$, where m_p and S_p are some robust estimates of the location and scale of p_{ij} such as the minimum covariance determinant estimators. We introduce the weight function to bound the influence of leverage points, covariate space only. As indicated in He et al. (2005), we can include certain covariates that are likely to contribute to the leverage. In the following simulation study, b_0 is chosen as the 95th percentile of the chi-squared distribution with degrees of freedom equal to the dimension of p_{ij} and ρ is fixed as 1. For simplicity and consistency, we choose $p_{ij} = x_{ij}$ for all three weighting matrices and denote them as $W_i = \text{diag}(w_{i1}, \ldots, w_{in_i})$.

The second central function of penalized robustified generalized estimating equations is capable of selecting variables, which is achieved by adding a penalty term on each estimating equation. Usually, $q_{\tau^{(l)}}(\cdot)$ is the first derivative for some penalty $p_{\tau^{(l)}}(\cdot)$, where $l = 1, 2, 3$. For brevity, we replace $p_{\tau^{(1)}}$, $p_{\tau^{(2)}}$ and $p_{\tau^{(3)}}$ by p_τ and $q_{\tau^{(1)}}$, $q_{\tau^{(2)}}$ and $q_{\tau^{(3)}}$ by q_τ when no misunderstanding arises. In the following simulation and real data analysis, we only consider SCAD and ALASSO penalties to show the asymptotic properties we raise. Fan and Li (2001) defined the smoothly clipped absolute deviation (SCAD) penalty function:

$$p_\tau(|\theta|) = \tau(|\theta|)\{I((|\theta|) < \tau)\} + \frac{(a - |\theta|/2\tau)}{a-1} I(\tau < (|\theta|) < a) + \frac{a^2\tau}{2(a-1)(|\theta|)} I\{(|\theta|) \geq a\tau\},$$

in which $a = 3.7$ was suggested by the authors. As it is well-known that being a compromise between LASSO and Hard penalties, SCAD itself enjoys unbiasedness, sparsity and continuity properties simultaneously, while based on which many oracle procedures have been proved.

As a consistent version of the L_1 penalty, ALASSO penalty is defined as $p_\tau = \tau|\theta|w$, for a known data–driven weight w. In this paper, we employ the weight $w = 1/|\tilde{\theta}|$, where $\tilde{\theta}$ stands for the regression coefficient estimates obtained from solving (2.2)–(2.4) without penalty.

2.3 Asymptotic properties

We denote the m subjects based penalized estimator $\hat{\theta}_m = ((\hat{\theta}_m^{s_1})^T,\ (\hat{\theta}_m^{s_2})^T)^T$ for the true value $\theta_0 = ((\theta_0^{s_1})^T,\ (\theta_0^{s_2})^T)^T$, where $\theta_0^{s_1}$ and $\theta_0^{s_2}$ are the nonzero and zero components of θ_0 respectively. Denote the dimension of $\theta_0^{s_1}$ by s_1 and $s = s_1 + s_2$. The parameter space is assumed compact and the true value θ_0 is in the interior

of the parameter space Θ.

In what follows we first show that the penalized estimator $\hat{\theta}_m$ exists and converges to θ_0 at the rate $O_p(m^{-1/2})$, implying that it has the same consistency rate as the ordinary estimator. We then prove that the $\sqrt{m}$–consistent estimator $\hat{\theta}_m^{s_1}$ has the asymptotic normal distribution and possesses the oracle property under certain regularity conditions.

Theorem 1. Under the assumptions in S1 of the supplementary material, the following results hold:

(a). There exists an approximate zero-crossing solution $\hat{\theta}$ of $U(\theta) = 0$, such that $\hat{\theta} = \theta_0 + O_p(m^{-1/2})$.

(b). For any $\sqrt{m}$ consistent approximate zero-crossing solution of $U(\theta) = 0$, we have

$$\lim_{m\to\infty} P\{\hat{\theta}_j = 0,\ j > s_1\} = 1.$$

The definition of zero-crossing estimator $\hat{\theta}$, which is introduced in Johnson et al. (2008), is given in S1 of the supplementary material. Theorem 1 implies that when we choose proper τ_m, our robust penalized GEE approach can simultaneously achieve the $\sqrt{m}$ consistency of the regularized regression coefficient estimation and the consistency of variable selection.

To obtain the asymptotic distribution of $\hat{\theta}$, we denote $B = \lim_{m\to\infty} \frac{1}{m}\mathrm{Cov}[U^R(\theta_0)]$ and assume it to be positive definite. The definition of $U^R(\theta_0)$ is given in supplement S1. We assume $\kappa_m(\theta) = \mathrm{E}[\frac{1}{m}U^R(\theta)]$, $\kappa_m(\theta_0) = 0$, $\kappa_m(\theta)$ is continuous on Θ and $\kappa_m(\theta)$ is differentiable at θ_0 with nonsingular derivative matrix G. Define $c_m = (q_{\tau_m}(|\theta_{01}^{s_1}|)\mathrm{sgn}(\theta_{01}^{s_1}), \ldots, q_{\tau_m}(|\theta_{0s_1}^{s_1}|)\mathrm{sgn}(\theta_{0s_1}^{s_1}))^T$ and $\Omega = \mathrm{diag}\{-q'_{\tau_m}(|\theta_0|)\mathrm{sgn}(\theta_0)\}$, where τ_m is equal to either $\tau_m^{(1)}$, $\tau_m^{(2)}$ or $\tau_m^{(3)}$, depending on whether θ_{0j} is a component of β_0, γ_0 or λ_0 $(1 \le j \le s)$. θ_{0j} is the jth component of θ_0, and $\theta_{0j}^{s_1}$ is the jth component of $\theta_0^{s_1}$ $(1 \le j \le s_1)$.

Theorem 2. Under the mild conditions as those given in S1 of the supplementary material, for SCAD penalty we have

$$\sqrt{m}(G_m^{s_1} + \Omega_m^{s_1})\{\hat{\theta}_m^{s_1} - \theta_0^{s_1} + (G_m^{s_1} + \Omega_m^{s_1})^{-1}c_m\} \to N_{s_1}(0, B^{s_1})$$

in distribution, where B^{s_1}, $G_m^{s_1}$ and $\Omega_m^{s_1}$ are the $s_1 \times s_1$ submatrix of B, G and Ω corresponding to the nonzero components $\theta_0^{s_1}$.

As a result, the asymptotic covariance matrix $\mathrm{Cov}(\hat{\theta}_m^{s_1})$ of $\hat{\theta}_m^{s_1}$ is

$$\frac{1}{m}(G_m^{s_1}+\Omega_m^{s_1})^{-1}\hat{B}_m^{s_1}(G_m^{s_1}+\Omega_m^{s_1})^{-1}.$$

That is to say, our proposed robust penalized joint model based on SCAD penalty possesses the oracle property that the true model can be correctly identified if it has been known in advance. Proofs of Theorems are sketched in S1 of the supplementary material.

2.4 Implementation

An iterative MM algorithm for estimating β, γ and λ is described in detail in S2 of the supplementary material. Meanwhile, the choice of τ is critical. In practice, we select $\tau^{(1)}$ by minimizing the robustified generalized cross-validation (GCV) criterion:

$$GCV_\beta(\tau)=\frac{RSS_\beta(\tau)/m}{\{1-d(\tau)/m\}^2},$$

where $RSS_\beta(\tau)$ is the robustified residual sum of squares

$$\sum_{i=1}^{m}[W_i\psi(A_i^{-1/2}(y_i-x_i\hat{\beta}_\tau))]^2,$$

and $d(\tau)$ is the effective number of parameters, that is $d_\beta(\tau)=\mathrm{tr}([\hat{G}(\beta)+\Delta_\tau(\hat{\beta}_\tau)]^{-1}[\hat{G}(\beta)]^T)$, here $\hat{\beta}_\tau$ is the solution of the penalized robust GEE when τ is fixed. We select $\hat{\tau}=\mathrm{argmin}_\tau \mathrm{GCV}_\beta$.

Similar to the choice of $\tau^{(1)}$ we select tuning parameters $\tau^{(2)}$ and $\tau^{(3)}$ by minimizing the robustified GCV statistics

$$GCV_\gamma(\tau)=\frac{RSS_\gamma(\tau)/m}{\{1-d_\gamma(\tau)/m\}^2},\ GCV_\lambda(\tau)=\frac{RSS_\lambda(\tau)/m}{\{1-d_\lambda(\tau)/m\}^2},$$

where RSS_γ and RSS_λ are the corresponding robustified residuals, respectively. Specifically, $RSS_\gamma(\tau)=\sum_{i=1}^m[W_i\psi(D_i^{-1/2}(r_i-\hat{r}_i))]^2$ and $RSS_\lambda(\tau)=\sum_{i=1}^m[W_i\psi(\tilde{A}_i^{-1/2}(\hat{\varepsilon}_i^2-\hat{\sigma}_i^2))]^2$. $d_\gamma(\tau)$ and $d_\lambda(\tau)$ are the effective numbers of covariance parameters.

To avoid the computational burden, we recommend selecting parameters sequentially. To be specific, we choose tuning parameters following the steps:

(1) Fix $\tau^{(2)}=\tau^{(3)}=0$, choose $\hat{\tau}^{(1)}=\mathrm{argmin}_{\tau^{(1)}}\mathrm{GCV}_\beta(\tau^{(1)})$;

(2) Fix $\tau^{(1)}=\hat{\tau}^{(1)}$ and $\tau^{(3)}=0$, choose $\hat{\tau}^{(2)}=\mathrm{argmin}_{\tau^{(2)}}\mathrm{GCV}_\gamma(\tau^{(2)})$;

(3) Fix $\tau^{(1)} = \hat{\tau}^{(1)}$ and $\tau^{(2)} = \hat{\tau}^{(2)}$, choose $\hat{\tau}^{(3)} = \text{argmin}_{\tau^{(3)}} \text{GCV}_\lambda(\tau^{(3)})$;
(4) The final choice is $(\hat{\tau}^{(1)}, \hat{\tau}^{(2)}, \hat{\tau}^{(3)})$.

The flexible procedure of tuning parameter selection has several merits. First, it largely reduces the computational burden comparing to minimizing GCVs in a three-dimensional parameter space. Second, even two similar estimating equations of covariance parameters as that of the mean parameter have been proposed, we have to admit that in some cases, pursuing an accurate estimation on covariance structure is not of the same priority as the mean estimation. From a numerical point of view, minimizing GCV_γ and GCV_λ do not always benefit for mean estimation. Moreover, simplifying the procedure of tuning parameters selection is helpful to stabilize the algorithm under contaminations. As a result, we do not recommend minimizing GCV_β, GCV_γ and GCV_λ simultaneously. Sequentially selecting tuning parameters suggests us to adjust and stabilize selection process while balancing the importance of estimation for mean and covariance structure.

3. Simulation

In this section, we conduct a simulation study to assess the performance of the proposed estimators mainly from three aspects: (1) efficiency of the proposed robust model compared with the corresponding non-robust version; (2) necessity of the proposed robust method with the existence of outliers and (3) comparison with classical GEE method under covariance matrix misspecification.

We compare model errors of different variable selection procedures using median of model error (MME), where model errors are evaluated following Fan and Li (2001) as:

$$ME_\beta = (\hat{\beta} - \beta_0)^T XX^T (\hat{\beta} - \beta_0),$$

$$ME_\gamma = (\hat{\gamma} - \gamma_0)^T TT^T (\hat{\gamma} - \gamma_0),\ ME_\lambda = (\hat{\lambda} - \lambda_0)^T ZZ^T (\hat{\lambda} - \lambda_0),$$

where $X = (X_1^T, \ldots, X_m^T)^T$, $T = (T_1^T, \ldots, T_m^T)^T$, and $Z = (Z_1^T, \ldots, Z_m^T)^T$. We employ average correctly fit percentage (CF%) to measure the accuracy of the model selection procedure, where correctly fit means that the procedure select the exact subset model. Moreover, we compare the average numbers of regression coefficients that are correctly shrunk (CS) to zeros, which measure the complexity of the selected model. In sum, MME , CF% and CS are supplementary to each other for measurement of model accuracy. The replications of each scenario are

200 times.

In simulation studies, we generate balanced data sets (i.e. $n_i = n$) for convenience. In practice, our method also works well for the nearly balanced data set (Zhou and Qu, 2012). For example, if we simulate the data set with missing probability at 0.2 (similar to simulation study 1 in Fan and Li, 2001), the estimates for both mean and covariance model are still consistent. These results are omitted for brevity. In the real data example, we can also handle unbalanced data when the observation time information is available. This is a reasonable assumption for longitudinal data, as the subjects' measurements are recorded along with the observation time.

Study 1. We simulate 100 (or 200) subjects, each of which has 5 observations drawn from the multivariate normal distribution $N_5(\mu_i, \Sigma_i)$. The true values of the mean parameter and log-innovation variances are chosen to be $\beta = (3, 0, 0, -2, 1, 0, 0, 0, 0, -4)^T$ and $\lambda = (0, 1, 0, 0, 0, -2, 0)^T$, respectively. Two specifications are designed for generalized autoregressive parameters: (1) $\gamma = (0, 0, 0, 0, 0, 0, 0)^T$ and (2) $\gamma = (0.2, 0, 0, 0, 0, 0, 0)^T$. The mean covariates $x_{ij} = (x_{ijt})_{t=1}^{10}$ are random samples drawn from the multivariate normal distribution with mean 0 and covariance matrix of AR(1) structure with variance $\sigma^2 = 1$ and correlation parameter $\rho = 0.5$ ($i = 1, \ldots, 100$; $j = 1, \ldots, 5$). Then $g_{ijk} = (x_{ijt} - x_{ikt})_{t=1}^{7}$ and $z_{ij} = (x_{ijt})_{t=1}^{7}$ are covariates 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 from (2.2) – (2.4). The responses y_i's are then drawn from the multivariate normal distribution $N(\mu_i, \Sigma_i)$ ($i = 1, \ldots, 100$).

To investigate the robustness of the proposed robust variable selection method against outliers, we consider two contaminations:

C0: randomly choose 1% of x_{ij1} to be $x_{ij1} - 1$;

C1: randomly choose 2% of x_{ij1} to be $x_{ij1} - 3$ and 2% of y_{ij} to be $y_{ij} + 10$.

NC represents no contamination situation hereafter. C1 is commonly used contamination setting in previous research on robust method. For the purpose of comparing the performance of robust model and non-robust model, we consider a tiny contamination in C0. The initial value of the mean estimation is obtained from the robust GEE estimation in He et al. (2005) with independent working

correlation, which in return guarantees consistency of the autoregressive parameters and innovative parameters after the first iteration. If the covariance matrix falls into the spanned space of the covariates, the proposed method converges quickly under no contamination, usually in a few steps of iterations. However, in non-robust modeling, traditional MM algorithm has a large probability of non-convergence under contaminations. Typically, in C1, 90% of the 200 (180) times of replications cannot obtain a final estimation due to divergence of iteration in covariance estimation, even when the initial value of estimation for the mean parameter β is close to the true value. Moreover, $10\% - 20\%$ of the non-robust algorithm fails to converge in C0, where the perturbation is almost negligible. This fact supports the necessity of developing a robust variable selection procedure over its non-robust version in joint modeling of longitudinal data.

To check the asymptotic properties, we simulate with sample size of 100 and 200 subjects. In Table 1, we list the median of model error (MME), average correctly fit percentage (CF%) and average numbers of regression coefficients that are correctly shrunk (CS) for both robust and non-robust methods, under NC (no contamination) and C0 (tiny contamination). Notice that the results for C0 were obtained based on convergence cases only. R_{scad} and R_{alasso} represents the proposed robust method employing SCAD and ALASSO respectively. NR is the non-robust method.

First, as the number of subjects m increases, MME of both robust and non-robust methods decreases, while CF% and CS approach to 1 and the true number of zero parameters respectively. These are consistent with the oracle property. Second, R_{scad} and NR method perform equally well in variable selection under NC, although acceptable loss of efficiency in robust method can be detected from slightly larger MME (in NC R_{scad} comparing to NC NR). However, under C0, robust method apparently outperforms non-robust method in both estimation efficiency and variable selection. Especially in covariance model identification, non-robust method fails to correctly identify innovation variance model under such a tiny contamination in most replications, which indicates that the non-robust joint model is extremely sensitive to perturbations.

Next we compare the performance of R_{scad} and R_{alasso}. We find that R_{scad} outperforms R_{alasso} in β and λ estimation and R_{alasso} performs better in γ es-

timation. In fact, SCAD allows almost no penalty if the true parameter is far from 0 and ALASSO penalizes all parameters, which will increase the MME in its performance, especially in λ. In sum, although both robust methods can resist the contamination according to the simulation, R_{scad} is preferred.

Standard errors for SCAD estimators in Study 1 for non-zero parameters (β_1, β_4, β_5, β_{10}, γ_1, λ_2 and λ_6) are attached in Table S3.1 of the supplementary material. In fact, the standard errors of the robust method are close to those of the non-robust method under no contamination and are much smaller under contamination $C0$. To investigate the influence of outliers in covariance estimation, we list entropy losses and quadratic losses (defined in S3 of the supplementary material) in Table 2. Again, we find poor performance of the non-robust method in C0 comparing to the robust approach in loss comparison. Consequently, we conclude that our robust joint model is necessary in research of joint modeling using the modified Cholesky decomposition in longitudinal data analysis.

Study 2. In this study, we aim to look into the effect of covariance misspecification on mean estimation. We compare the performance of the robust penalized joint model (RPJ, denoted as rpj in Table 3) method with that of robust and non-robust penalized generalized estimating equations (RPGEE and PGEE, denoted as rpgee and pgee) methods that assume a fixed working correlation matrix and solve (2.2) for the mean. The most salient difference of the three methods is that our joint model builds regression models after decomposing the covariance matrix, while RPGEE and PGEE treat the covariance matrix as nuisance where the marginal variance of y_i is estimated by the sample. In PGEE, we set the tuning parameter of Huber function c as 1000 and the weight $W = I$ when solving (2.2) for the mean.

Under the same mean establishment in Study 1, we consider three normal covariance structures: working independence (IN), auto-regressive (AR) and exchangeable (EX) with correlation parameter 0.5. We compare the performance of eight estimators: rpj_{ar} and rpj_{in} are robust joint models with independent and AR(1) correlation structure assumptions for $Cov(\varepsilon_i^2)$ in (2.4) respectively; rpgeee_{in}, rpgeee_{ar} and rpgeee_{ex} (or pgeee_{in}, pgeee_{ar} and pgeee_{ex}) are robust (or non-robust) penalized GEE estimations with IN, AR and EX as working correlation matrices. We adopt SCAD penalty in the whole study.

Table 3 lists the results under NC and C1. In the table, we employ MRME (the median of relative model error) to compare the performance of eight estimators, where the relative model error of the estimator is defined as

$$\frac{\text{ME (model error) of the estimator}}{\text{ME of PGEE estimator with true covariance matrix}}.$$

On one side, in the absence of outliers (NC), in general, pgee estimators perform slightly better than rpgee estimators which in turn are better than the rpj estimators. Besides, rpj_{in} and rpj_{ar} always have similar performance, which suggests that the choice of the δ in (2.4) would not have much effect on the mean and covariance estimation. As a result, we fix $\delta = 0$ in a later application.

On the other side, under C1, rpj estimators improve substantially over the rpgee and pgee estimators. Without the bounded score on the mean estimator, pgee mean estimators collapse in any of IN, AR or EX covariance structure, as all MRME's for rpj and rpgee are less than 1 in C1. By adopting the robust estimator for the mean, rpgee estimators has reasonable performance on the mean estimation. However, rpj estimators further improve the performance of the mean estimator (variable selector). Standard errors for the mean estimators can be found in supplementary Table S3.2, which reveals the influence of outliers on the mean estimation again.

Furthermore, Table S3.3 of the supplementary material lists entropy losses and quadratic losses on covariance matrix estimation. When there is no contamination, rpgee and pgee have comparable performance and are better than rpj. However, under C1, the estimation of the covariance matrix using GEE (both robust and non-robust, i.e. rpgee and pgee) can be seriously affected by the contamination.

4. Real data analysis

In this section, we illustrate our method for estimating the robust penalized joint mean and covariance by analyzing the hormone data, which has been analyzed be Fung et al. (2002), He et al. (2002), Fan et al. (2012) and Qin et al. (2009). The data set contains 492 observations of progesterone level within a menstrual cycle, collecting from 34 women clinical participants. In our model, the response variable y_{ij} is the log-transform of progesterone level and the observation time is t_{ij}, apart from which patient's age and body mass index (BMI) are

recorded. Two objectives are considered when we implement the robust variable selection method in joint mean and covariance model to this hormone data set: (1) to accommodate the influence of outliers and leverage points, and to detect outliers in the data set; (2) to identify statistically significant covariates in linear models of both mean and covariance matrix.

The starting mean model we proposed is given as follows:

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 Age_i + \beta_2 BMI_i + \beta_3 t_{ij} + \beta_4 t_{ij}^2 \\ &\quad + \beta_5 Age_i \times BMI_i + \beta_6 Age_i \times t_{ij} + \beta_7 BMI_i \times t_{ij} + e_{ij} \\ &= x_{ij}^T \beta + e_{ij}. \end{aligned}$$

For the covariance model, we follow the model in (1) and choose the corresponding covariates as

$$g_{ijk} = (1, (t_{ij} - t_{ik}), (t_{ij} - t_{ik})^2, (t_{ij} - t_{ik})^3)^T, \ z_{ij} = x_{ij}.$$

Three estimators are under consideration: rpj represents the robust penalized joint model; pj represents the penalized joint model proposed by Kou and Pan (2011); gee is the widely-used GEE estimators. Table 4 summarizes estimators of the mean parameters with standard errors. We notice that the joint models (rpj and pj) are parsimony than gee that they choose the time as the only significant variable. The observation is consistent with the previous research that both Age, BMI, their interaction and interactions with time are not statistically significant in the model. Only time is significant. We also observe that the regression coefficients for Time obtained by rpj and pj are rather different. It is due to the fact that the non-robust pj estimator is affected by outliers.

Estimates with standard errors for the generalized autoregressive coefficients and innovation variances are summarized in Table 5. We find that the cubic polynomial of time is statistically significant for autoregressive coefficients γ in all fits. Standard errors in non-robust method are larger than those in the robust method, which support the estimation of the robust model again. Unlike estimators of generalized autoregressive coefficients, significant covariates for innovation variances are not found in our analysis. Due to the existence of outliers, the non-robust method fails to select significant covariates for innovation variances.

Outlier detection has been done carefully through the procedure. By investigating into the standardized residuals s_{ij} (i.e. the jth component of $\hat{\Sigma}_i^{-1/2}(y_i - \hat{\mu}_i)$) and the weight function w_{ij}, we find one observation-level outlier (observation 10) and one subject-level leverage point (subject 18). $p_{ij} = (AGE_i, BMI_i)$ is contributed to the weight functions w_{ij} in our robust method. Subject 18 is a leverage point which has not been identified before. It has an extremely high BMI of 38 that heavily downweights the cluster of observations from the patient. A careful inspection on the standardized residual s_{ij} tells us that case 10 is the most extreme point with $s_{ij} = -6.09$. The progesterone level of the 10th observation for subject 1 (case 10) is 2.46, which is very different from its neighborhood observations 9 and 11 measured one day before and one day after, with the progesterone level being 12.8 and 13.4 respectively. This inconsistency has not been noticed in the literature. In fact, all other thirteen observations of subject 1 range from 8.5 to 13.4 except case 10. In particular, this observation is the lowest progesterone level in the whole data set. Therefore, we conclude that case 10 of subject 1 is a clear outlier. The influence of this outlier on the parameter estimates can be referred to Table S4.1 of the supplementary material. Besides, we also notice that some observations have large standardized residuals, such as cases 117, 334 and 372, due to the fact that they are extreme values of the progesterone level within a subject. The effects on these potential outliers are downweighted by our robust method in the estimation of mean and covariance parameters.

Acknowledgement

The authors are grateful to two reviewers, the Associate Editor, and the Co-Editor for their insightful comments and suggestions which have improved the manuscript significantly. This work was supported by the National Natural Science Foundation of China (10931002, 11271180).

References

Bickel, P. J. and Levina, E. (2008). Regularized Estimation of Large Covariance Matrices,*The Annals of Statistics.* **36**, 199-227.

Cantoni, E., Flemming, J. M., and Ronchetti, E. (2005). Variable Selection for Marginal Longitudinal Generalized Linea Models, *Biometrics.* **61**, 507-514.

Cook, R. D and Weisberg, S. (1983). Diagnostics for Heteroscedasticity in Regression, *Biometrika.* **70**, 1-10.

El Karoui, N. (2008). Operator Norm Consistent Estimation of Large Dimensional Sparse Covariance Matrices, *The Annals of Statistics.* **36**, 2717-2756.

Fan, J., Huang, T., and Li, R. (2007). Analysis of Longitudinal Data With Semiparametric Estimation of Covariance Function, *Journal of the American Statistical Association.* **35**, 632-641.

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

Fan, J. and Wu, Y. (2008). Semiparametric Estimation of Covariance Matrices for Longitudinal Data, *Journal of the American Statistical Association.* **103**, 1520-1533.

Fan, Y. L., Qin, G. Y. and Zhu, Z. Y. (2012). Variable Selection in Robust Regression Models for Longitudinal Data, *Journal of Multivariate Analysis.* **109**, 156-167.

Fu, W. J. (2003). Penalized Estimating Equations, *Biometrics.* **59**, 126-132.

Fung, W.K., Zhu, Z. Y., Wei, B., and He, X. (2002). Influence Diagnostics and Outlier Tests for Semiparametric Mixed Models. *Journal of Royal Statistical Society B.* **64**, 565-579.

He, X., Fung, W. K., and Zhu, Z. Y. (2005). Robust Estimation in Generalized Partial Linear Models for Clustered Data, *Journal of the American Statistical Association.* **472**, 1176-1184.

He, X., Zhu, Z. Y., and Fung, W. K. (2002). Estimation in a Semiparametric Model for Longitudinal Data with Unspecified Dependence Structure, *Biometrika.* **89**, 579-590.

Huang, J., Liu, N., Pourahmadi, M., and Liu, L. (2006). Covariance Matrix Selection and Estimation via Penalised Normal Likelihood, *Biometrika.* **93**, 85-98.

Jeng, X. J. and Daye, Z. J. (2011). Sparse Covariance Thresholding for High-dimensional Variable Selection, *Statistica Sinica.* **21**, 625-657.

Johnson, B., Lin, D. Y., and Zeng, D. (2008). Penalized Estimating Functions and Variable Selection in Semiparametric Regression Models, *Journal of the American Statistical Assoication.* **103**, 672-680.

Kou, C. and Pan, J. (2011). Variable Selection for Joint Mean and Covariance Models via Penalised Likelihood, Technical report, Probability and Statistics Group School of Mathematics, The University of Manchester.

Leng, C., Zhang, W., and Pan, J. (2010). Semiparametric Mean-Covariance Regression Analysis for Longitudinal Data, *Journal of the American Statistical Association.* **105**, 181-193.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika.* **73**, 13-22.

Mao, J., Fung W. K. and Zhu, Z. Y. (2011). Joint Estimation of Mean–covariance Model for Longitudinal Data with Basis Function Approximations. *Computational Statistics and Data Analysis.* **55**, 983–992.

Pan, W. (2001). Akaike's Information Criterion in Generalized Estimating Equations. *Biometrics.* **73**, 13-22.

Pourahmadi, M. (1999). Joint Mean-Covariance Models With Applications to Longitudinal Data: Unconstrained Parameterisation. *Biometrika.* **86**, 677-690.

—(2000). Maximum Likelihood Estimation of Generalized Linear Models for Multivariate Normal Covariance Matrix. *Biometrika.* **87**, 425-435.

Qin, G. Y., Zhu, Z. Y., and Fung, W. K. (2009). Robust Estimation of Covariance Parameters in Partial Linear Model for Longitudinal Data. *Journal of Statistical Planning and Inference.* **139**, 558-570.

Qu, A., Lindsay, B., and Li, B. (2000). Improving Generalised Estimating Equations Using Quadratic Inference Functions. *Biometrika*. **87**, 823-836.

Rothman, A. J., Levina, E., and Zhu, J. (2009). Generalized Thresholding of Large Covariance Matrices. *Journal of the American Statistical Association.* **104**, 177-186.

Xu, J. and Mackenzie, G. (2012). Modelling Covariance Structure in Bivariate Marginal Models for Longitudinal Data. *Biometrika*. **99**, 649-662.

Ye, H. and Pan, J. (2006). Modelling Covariance Structures in Generalized Estimating Equations for Longitudinal Data. *Biometrika*. **93**, 927-941.

Zheng, X. Y., Fung, W. K., and Zhu, Z. Y. (2013). Robust Estimation in Joint Mean-Covariance Regression Model for Longitudinal Data. *Annals of Institution of Statistical Mathematics*. To appear.

Zhou, J., and Qu, A. (2012). Informative Estimation and Selection of Correlation Structure for Longitudinal Data. *Journal of the American Statistical Association.* **107**, 701-710.

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

Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong

E-mail: xyzheng@hku.hk

Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong

E-mail: wingfung@hku.hk

Department of Statistics, Fudan University, Shanghai, China

E-mail: zhuzy@fudan.edu.cn

Table 1: Parameter estimation and selection in Study 1

	$\gamma = 0$						$\gamma \neq 0$					
	n=100			n=200			n=100			n=200		
	MME	CF%	CS	MME	CF%	CS	MME	CF%	CS	MME	CF%	CS
NC NR_{scad}												
β	0.040	66.0	5.61	0.030	80.5	5.79	0.032	56.5	5.47	0.021	73.5	5.70
γ	0.016	36.5	5.86	0.000	76.0	6.62	0.021	45.0	5.09	0.004	86.0	5.82
λ	0.057	73.5	4.71	0.032	93.5	4.93	0.043	77.5	4.71	0.026	96.0	4.96
NC R_{scad}												
β	0.050	72.0	5.68	0.040	84.0	5.82	0.039	53.0	5.38	0.026	73.0	5.70
γ	0.018	35.5	5.77	0.000	69.5	6.49	0.025	37.5	5.00	0.006	76.0	5.67
λ	0.078	92.5	4.92	0.048	100	5.00	0.060	93.0	4.92	0.040	99.5	5.00
NC R_{alasso}												
β	0.047	54.5	5.41	0.033	75.5	5.73	0.049	52.0	5.38	0.030	69.0	5.65
γ	0.012	44.5	6.01	0.000	72.0	6.50	0.021	47.0	5.00	0.003	85.5	5.85
λ	0.242	51.0	4.28	0.193	73.5	4.71	0.318	67.0	4.57	0.265	86.0	4.84
C0 NR_{scad}												
β	0.068	65.0	5.57	0.067	81.5	5.79	0.068	57.5	5.42	0.066	71.5	5.63
γ	0.021	33.0	5.66	0.000	59.5	6.37	0.061	28.0	4.45	0.039	52.5	5.14
λ	0.331	17.5	3.12	0.520	15.0	3.02	0.408	19.0	3.06	0.575	12.0	2.83
C0 R_{scad}												
β	0.053	72.0	5.67	0.043	85.0	5.84	0.052	52.0	5.34	0.065	87.5	5.87
γ	0.018	35.0	5.82	0.000	72.0	6.55	0.040	31.5	4.77	0.000	71.0	6.53
λ	0.068	91.0	4.91	0.052	99.5	5.00	0.088	85.0	4.83	0.069	99.0	4.99
C0 R_{alasso}												
β	0.055	52.5	5.41	0.038	74.0	5.70	0.091	50.5	5.29	0.061	64.5	5.54
γ	0.013	41.5	5.93	0.000	74.5	6.55	0.049	28.5	4.54	0.017	71.5	5.61
λ	0.280	43.5	4.22	0.224	67.5	4.63	0.550	60.0	4.49	0.420	84.0	4.82

Simulation results of median of model error (MME), average correctly fit percentage (CF%) and average numbers of regression coefficients that are correctly shrunk (CS) for both robust (R_{scad}, R_{alasso}) and non-robust (NR) method, under NC (no contamination) and C0 (tiny contamination) with 200 replications.

Table 2: Entropy loss (L_1) and quadratic loss (L_2) in estimating Σ in Study 1

		$\gamma = 0$				$\gamma \neq 0$			
		n=100		n=200		n=100		n=200	
		L_1	L_2	L_1	L_2	L_1	L_2	L_1	L_2
NC	NR	0.097	0.253	0.024	0.091	0.104	0.249	0.038	0.097
	R_{scad}	0.117	0.318	0.036	0.129	0.135	0.361	0.054	0.144
	R_{alasso}	0.189	0.756	0.122	0.528	0.251	1.112	0.151	0.698
C0	NR	0.255	1.070	0.289	1.454	0.463	2.318	0.479	2.911
	R_{scad}	0.107	0.272	0.038	0.139	0.190	0.601	0.049	0.187
	R_{alasso}	0.215	0.900	0.134	0.610	0.436	2.289	0.286	1.473

Table 3: Simulation results for Study 2

		IN			AR			EX		
		MRME	CF%	CS	MRME	CF%	CS	MRME	CF%	CS
NC	rpj_{ar}	1.09	92.0	5.92	1.33	76.0	5.74	1.87	67.5	5.63
	rpj_{in}	1.09	92.0	5.92	1.34	75.5	5.73	1.87	67.5	5.63
	rpgee_{ar}	1.04	94.5	5.94	1.12	96.0	5.96	1.56	92.0	5.92
	rpgee_{ex}	1.04	93.5	5.93	1.20	92.0	5.92	1.15	91.0	5.91
	rpgee_{in}	1.07	93.0	5.93	1.33	80.0	5.79	1.92	69.0	5.65
	pgee_{ar}	0.99	91.5	5.91	1.01	94.5	5.95	1.41	89.5	5.90
	pgee_{ex}	0.99	90.5	5.90	1.11	86.5	5.87	1.01	83.5	5.83
	pgee_{in}	1.00	91.5	5.92	1.26	72.0	5.70	1.83	63.0	5.57
C1	rpj_{ar}	0.08	92.0	5.92	0.08	95.5	5.96	0.10	89.0	5.89
	rpj_{in}	0.08	92.0	5.92	0.08	95.5	5.96	0.10	89.0	5.89
	rpgee_{ar}	0.16	90.0	5.90	0.11	88.5	5.88	0.13	83.0	5.83
	rpgee_{ex}	0.16	89.5	5.90	0.10	88.5	5.88	0.13	88.0	5.88
	rpgee_{in}	0.15	89.0	5.89	0.09	76.5	5.75	0.11	71.0	5.68
	pgee_{ar}	0.95	30.0	4.94	0.62	67.0	5.63	0.76	62.5	5.59
	pgee_{ex}	0.96	30.5	4.93	0.58	65.0	5.58	0.75	67.5	5.62
	pgee_{in}	0.92	28.0	4.88	0.52	54.0	5.43	0.63	48.0	5.34

Simulation results of median of relative model error (MRME), average correctly fit percentage (CF%) and average numbers of regression coefficients that are correctly shrunk (CS) under NC (no contamination) and C1 (contamination) with 200 replications. rpj_{ar} and rpj_{in} are robust joint models with independent and AR(1) correlation structure in (2.4) respectively; rpgeee_{in}, rpgeee_{ar} and rpgeee_{ex} (or pgeee_{in}, pgeee_{ar} and pgeee_{ex}) are robust (or non-robust) penalized GEE estimations with IN, AR and EX as working correlation matrices.

Table 4: Estimators of the mean parameters β and standard errors (inside brackets) for progesterone data

	Intercept	Age	BMI	Time	Time^2	Age× BMI	Age× Time	BMI× Time
rpj	0.837	0	0	0.691	0	0	0	0
	(0.074)	(-)	(-)	(0.053)	(-)	(-)	(-)	(-)
pj	0.892	0	0	0.562	0	0	0	0
	(0.078)	(-)	(-)	(0.056)	(-)	(-)	(-)	(-)
gee	0.870	1.684	-2.671	0.709	0.186	-4.829	1.493	0.701
	(0.126)	(2.180)	(2.928)	(0.049)	(0.085)	(50.09)	(0.827)	(0.857)

In the table, rpj represents the robust penalized joint model; pj represents penalized joint model proposed by Kou and Pan (2011); gee is the traditional GEE estimator. Working independence is considered.

Table 5: Estimates of the generalized autoregressive parameters γ and innovation variance parameters λ for progesterone data

	γ_1	γ_2	γ_3	γ_4
rpj	0.902	-2.726	2.339	-0.579
	(0.056)	(0.203)	(0.190)	(0.050)
pj	0.882	-2.623	2.185	-0.523
	(0.063)	(0.234)	(0.223)	(0.060)

	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8
rpj	-1.111	0	0	0	0.180	45.30	0	-1.317
	(0.103)	(-)	(-)	(-)	(0.120)	(34.50)	(-)	(2.435)
pj	-0.882	-1.527	-0.283	0.186	0.029	56.98	0.175	1.313
	(0.099)	(1.398)	(1.761)	(0.086)	(0.115)	(29.03)	(1.699)	(2.044)