



# Variable selection in robust regression models for longitudinal data

Yali Fan<sup>a,b</sup>, Guoyou Qin<sup>c,d</sup>, Zhongyi Zhu<sup>a,\*</sup>

<sup>a</sup> Department of Statistics, Fudan University, Shanghai 200433, China

<sup>b</sup> College of Science, University of Shanghai for Science and Technology, Shanghai 200093, China

<sup>c</sup> Department of Biostatistics, School of Public Health, Fudan University, Shanghai 200032, China

<sup>d</sup> Key Laboratory of Public Health Safety, Ministry of Education of China, Fudan University, China

## ARTICLE INFO

### Article history:

Received 22 March 2011

Available online 23 March 2012

### AMS subject classification:

62J99

### Keywords:

Longitudinal data

Penalized estimating equation

Robust method

Variable selection

## ABSTRACT

In this article, we consider variable selection in robust regression models for longitudinal data. We propose a penalized robust estimating equation to estimate the regression parameters and to select the important covariate variables simultaneously. Under some regularity conditions, we show the oracle properties of the proposed robust variable selection methods. A simulation study shows the robustness of the proposed methods against outliers. Moreover, it is found by the simulation study that incorporating the correlation structure into the procedure of variable selection will lead to better performance than ignoring the correlation structure for longitudinal data. In the end, the proposed methods are illustrated in the analysis of a real data set.

© 2012 Elsevier Inc. All rights reserved.

## 1. Introduction

Longitudinal data is ubiquitous across medicine, epidemiology, economics and sociology. In longitudinal studies, often many variables are measured, and the inclusion of redundant variables in the predictive regression model can reduce the accuracy and efficiency for estimation. Thus, variable selection is an essential part of regression modeling for longitudinal data. Longitudinal data is characterized by the correlation among the observations in the same subjects and the independence between the observations from different subjects. Selecting important covariates using regression models for longitudinal data is challenging, because the likelihood function is usually unavailable due to the correlation. The generalized estimating equations approach, proposed by Liang and Zeger [11], provide us with an efficient way to analyze longitudinal data with a marginal mean model. Some literatures, e.g., [12,2,4] extended the traditional variable selection criteria, such as AIC, BIC and  $C_p$ , to the case of longitudinal data under the framework of generalized estimating equations. Fu [6] applied the bridge penalty to the generalized estimating equations model to deal with collinearity in longitudinal studies. Dziak [4] also generalized LASSO [16] and SCAD [5] to the longitudinal case by solving a penalized generalized estimating equation.

However, it is well known that the classical estimating equation approach is very sensitive to outliers. Hence, it is expected that the variable selection methods could be distorted by the presence of data contamination or outliers. Robustness against outliers is a very important issue in longitudinal studies. Because for longitudinal data, the observations from the same subject often share the same values for the covariates, e.g. age, height, weight, then an outlier in a subject-level measurement can generate multiple outliers in the sample. So, for longitudinal data, outliers could be outlying observations or outlying subjects [7]. Many authors studied the influence of outliers to the estimate and have developed outlier diagnostic methods, such as Preisser and Qaqish [13] and Fung et al. [7]. Regarding estimation methods, robust methods have attracted

\* Corresponding author.

E-mail address: [zhuzy@fudan.edu.cn](mailto:zhuzy@fudan.edu.cn) (Z. Zhu).

much attention and have been discussed in many literatures, see [8,14]. However, the study on robust variable selection methods is relatively limited. Cantoni [1] used weighted estimating equations to build robust quasi-likelihood functions and to construct a class of test statistics for variable selection. Cantoni et al. [2] proposed generalized  $C_p$  criterion using a weighted quadratic predictive risk, and the weights allow data analysts to incorporate robustness, but their methods require Monte Carlo simulation to estimate the effective degrees of freedom of the model fitting and this is very computationally expensive.

In this paper, we propose robust variable selection approach based on a penalized robust estimating equation that incorporated the correlation structure for longitudinal data. Specifically, in the robust estimating equation, we make use of the covariate-dependent weights to downweight the effect of leverage points and a bounded score function on the Pearson residuals to dampen the effect of outliers in the response. Then we combine the estimation and selection by using continuous penalties. We consider several commonly used penalty functions, including SCAD [5], ALASSO [18], [3], LASSO [16] and elastic net (EN, [19]). We show the oracle properties of the proposed robust variable selection methods based on SCAD, ALASSO and Hard penalties. A simulation study shows that, when data is contaminated by outliers, the proposed robust methods can maintain good performance in terms of model error and model accuracy. It is also found by simulation that incorporating correlation can improve estimation efficiency and variable selection power for longitudinal data.

The remainder of this article is organized as follows. The main results are described in Section 2, including the penalized robust estimating equation and the asymptotic properties of the proposed methods. An efficient algorithm is given in Section 3. We report the simulation results in Section 4 and apply the proposed methods to a real data analysis in Section 5. The proof of the main results is given in the Appendix.

## 2. Main results

### 2.1. Penalized robust estimating equation

Consider a longitudinal study with  $n$  subjects and  $m_i$  observations over time for the  $i$ th subject ( $i = 1, 2, \dots, n$ ). Each observation consists of a response variable  $y_{ij}$  and a covariate vector  $x_{ij} \in \mathbb{R}^d$ . We specify a marginal linear regression model on  $(x_{ij}, y_{ij})$

$$Y_i = X_i \beta + \epsilon_i, \quad i = 1, 2, \dots, n, \quad (2.1)$$

where  $Y_i = (y_{i1}, y_{i2}, \dots, y_{im_i})^T$ ,  $X_i = (x_{i1}, x_{i2}, \dots, x_{im_i})^T$ , and  $\beta$  is a  $d \times 1$  regression coefficient,  $\epsilon_i = (\epsilon_{i1}, \epsilon_{i2}, \dots, \epsilon_{im_i})^T$  denote the random error vector,  $\epsilon_{ij}$  has mean zero and variance  $\sigma^2$ . We assume that the observations from the different subjects are independent.

We consider the following robust estimating equation, which is similar to Cantoni [1] and He et al. [8],

$$U_n^R(\beta) = \sum_{i=1}^n U_\beta(X_i \beta) = \sum_{i=1}^n X_i^T V_i^{-1} h_i(X_i \beta), \quad (2.2)$$

where  $V_i = R_i(\alpha) A_i^{-\frac{1}{2}}$ ,  $R_i(\alpha)$  is the working correlation matrix depends on unknown parameter vector  $\alpha$ ,  $A_i = \sigma^2 I_{m_i}$ . The core of the estimating equation is  $h_i(X_i \beta) = W_i(\psi(A_i^{-\frac{1}{2}}(Y_i - X_i \beta) - C_i))$ , where  $C_i = E(\psi(A_i^{-\frac{1}{2}}(Y_i - X_i \beta)))$  is used to ensure Fisher consistency of the estimator, and the expectation is taken under the conditional distribution of  $Y_i$  given  $X_i$  in the model (2.1). The function  $\psi(\cdot)$  is chosen to downweight the influence of outliers in the response variable. A common choice is Huber's score function  $\psi_c(x) = \min\{c, \max(-c, x)\}$ , the tuning constant  $c$  here is typically chosen to give a certain level of asymptotic efficiency at the underlying distribution. For the Gaussian distribution and symmetric Huber's function, which is the case this paper concerned with, the correction term  $C_i$  is exactly equal to zero. The weight matrix  $W_i = \text{diag}(w_{i1}, w_{i2}, \dots, w_{im_i})$  is used to downweight the effect of leverage points. Similar to Sinha [14], we choose the weight function as a function of the Mahalanobis distance in the form

$$w_{ij} = w(x_{ij}) = \min \left\{ 1, \left\{ \frac{b_0}{(x_{ij} - m_x)^T S_x^{-1} (x_{ij} - m_x)} \right\}^{\frac{r}{2}} \right\},$$

where  $r \geq 1$ ,  $b_0$  is the 0.95 quantile of the chi-square distribution with the degree of freedom equal to the dimension of  $x_{ij}$ ,  $m_x$  and  $S_x$  are some robust estimates of location and scatter of  $x_{ij}$ , like the median of  $x_{ij}$  and median absolute deviance of  $x_{ij}$ . For data without outlying points, we can set  $\psi(x) = x$  and  $w_{ij} = 1$  in the robust estimating Eq. (2.2), and this will lead to the classical nonrobust estimating equation.

In order to select important covariate variables and estimate them simultaneously, we use the penalized robust estimating equation

$$U_n^p(\beta) = U_n^R(\beta) - nq_\lambda(|\beta|) \text{sgn}(\beta), \quad (2.3)$$

where  $q_\lambda(|\beta|) = (q_{\lambda,1}(|\beta_1|), \dots, q_{\lambda,d}(|\beta_d|))^T$ . In most cases,  $q_{\lambda,j}(\cdot) = p'_{\lambda,j}(\cdot)$ ,  $j = 1, \dots, d$ , for some penalty function  $p_{\lambda,j}(\cdot)$ . As demonstrated by He et al. [8] and Sinha [14], robust estimating equation like (2.2) can produce robust estimates against

outliers or influential observations. However, their methods had not involved the variable selection issue. By penalized estimating equation, the proposed methods are able to shrink small components of coefficient to zero, thus performing variable selection in addition to produce robust estimators of the nonzero components. Note that the proposed penalization strategy can safely handle data containing outlying points. Cantoni [1] proposed robust variable selection methods based on a class of test statistics rather than penalty methods. Cantoni et al. [2] suggested to perform robust variable selection by using a generalized  $C_p$  criterion, which is a discrete penalties based method. The penalized robust estimation Eq. (2.3) focuses on robust variable selection using continuous penalties rather than older discrete penalties. This equation is an extension of those considered by Johnson et al. [10] to the case of longitudinal data with robust setting.

In this paper, we consider several commonly used penalty functions. The LASSO penalty function [16],  $p_{\lambda,j}(|\beta_j|) = \lambda|\beta_j|$  (for all  $j$ ), is one of the most popular shrinkage estimators. The Hard penalty function  $p_{\lambda,j}(|\beta_j|) = \lambda^2 - (|\beta_j| - \lambda)^2 I(|\beta_j| < \lambda)$  (for all  $j$ ), which is generated from wavelet thresholding rules, corresponds to the best subset selection and stepwise deletion in certain cases. The SCAD penalty function is presented as a compromise between the Hard penalty and LASSO and is defined as

$$p_{\lambda,j}(|\beta_j|) = \lambda|\beta_j|I(|\beta_j| < \lambda) + \frac{(a - |\beta_j|/2\lambda)}{a - 1}I(\lambda < |\beta_j| \leq a\lambda) + \frac{a^2\lambda}{(a - 1)2|\beta_j|}I(|\beta_j| \geq a\lambda), \quad j = 1, \dots, d,$$

where  $a > 2$ . Elastic net(EN) penalty [19],

$$p_{\lambda,j}(|\beta_j|) = \lambda_1|\beta_j| + \lambda_2\beta_j^2, \quad j = 1, \dots, d$$

is introduced as a mixing penalty to effectively select grouped variables. ALASSO penalty is a consistent version of the  $\ell_1$  penalty and is defined as  $p_{\lambda,j}(|\beta_j|) = \lambda|\beta_j|w_j$ , for a known data-driven weight  $w_j$ . In this paper, we use the weight  $w_j = 1/|\tilde{\beta}_j|$ ,  $j = 1, \dots, d$ , where  $\tilde{\beta} = (\tilde{\beta}_1, \dots, \tilde{\beta}_d)^T$  refers to the regression coefficient estimates obtained from solving  $U_n^R(\beta) = 0$ .

Our proposed estimator for  $\beta$  is solution of  $U_n^p(\beta) = 0$ . Following [10], we use a general definition of solution to the penalized robust estimating equation, which is called zero crossing. An estimator  $\beta^* = (\beta_1^*, \dots, \beta_d^*)^T$  is called a zero crossing of  $U_n^p(\beta)$  if

$$\lim_{n \rightarrow \infty} \lim_{\epsilon \rightarrow 0+} n^{-1} U_j^p(\beta^* + \epsilon e_j) U_j^p(\beta^* - \epsilon e_j) \leq 0, \quad j = 1, \dots, d, \quad (2.4)$$

where  $U_j^p(\cdot)$  is the  $j$ th component of  $U_n^p(\cdot)$ ,  $e_j$  is the  $j$ th canonical unit vector. This definition means that the element-wise product of  $U_n^p(\beta^*)$  goes to zero or changes the sign at zero. In this paper, when we say the solution of  $U_n^p(\beta) = 0$ , we mean the zero crossing solution.

## 2.2. Asymptotic Properties

In this subsection, we will establish the consistency of our proposed robust variable selection method and the asymptotic normality of the estimator for regression coefficient. Furthermore we will show that the SCAD, Hard and ALASSO solved by (2.3) have the so-called *oracle* properties [5], that is, under appropriate conditions, the solution of (2.3) behaves asymptotically as if the true model were known *a priori*.

We assume that  $\beta \in \Theta$ ,  $\Theta \subseteq R^d$ , where  $\Theta$  is bounded subset in  $R^d$ . Let  $\beta_0 = (\beta_{01}, \dots, \beta_{0d})^T$  denote the true value of  $\beta$ . Without loss of generality, we suppose that  $\beta_{0j} \neq 0$  for  $j \leq s$  and  $\beta_{0j} = 0$  for  $j > s$ . Denote  $\kappa_n(\beta) = E(\frac{1}{n} U_n^R(\beta))$ , where the expectation is taken with the underlying distribution of  $Y_i$  conditional on  $X_i$  under the assumed model (2.1). Throughout the paper, we use  $\|\cdot\|$  to denote the  $L_2$  norm of a vector. The following assumptions are needed to establish the asymptotic properties.

- (C.1) The random function  $U_\beta(\cdot)$  in (2.2) satisfied the Lipschitz condition, that is, there exists a random function  $m(\cdot)$  with bounded second moment such that,  $\|U_{\beta_1}(\cdot) - U_{\beta_2}(\cdot)\| \leq m(\cdot) \|\beta_1 - \beta_2\|$ , for every  $\beta_1, \beta_2$  in the neighborhood of  $\beta_0$ .
- (C.2)  $\kappa_n(\beta_0) = 0$ ,  $\kappa_n(\beta)$  is continuous on  $\Theta$  and  $\kappa_n(\beta)$  is differentiable at  $\beta_0$  with nonsingular derivative matrix  $D$ .
- (C.3)  $E \|U_n^R(\beta_0)\|^2 < \infty$ , and exist  $\delta > 0$ , such that,

$$\lim_{n \rightarrow \infty} \frac{\sum_{i=1}^n E \|U_\beta(X_i \beta_0)\|^{2+\delta}}{(E \|U_n^R(\beta_0)\|^2)^{1/2})^{2+\delta}} = 0.$$

- (C.4)  $B = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n X_i^T V_i^{-1} \text{Cov}(h_i(X_i \beta_0)) (V_i^{-1})^T X_i$  is positive definite.
- (C.5) For any nonzero fixed  $\beta \in \Theta$ ,  $\lim_{n \rightarrow \infty} \sqrt{n} q_{\lambda_n}(|\beta|) = 0$ , and  $\lim_{n \rightarrow \infty} q'_{\lambda_n}(|\beta|) = 0$ .
- (C.6) For any  $M > 0$ ,  $\lim_{n \rightarrow \infty} \sqrt{n} \inf_{|\beta| \leq Mn^{-1/2}} q_{\lambda_n}(|\beta|) \rightarrow \infty$ .

**Remark.** The assumption (C.1) is the smoothness condition imposed on the score function  $U_\beta(\cdot)$ , and usually it is easy to check. When  $\psi(\cdot)$  is chosen as Huber's function  $\psi_c(x) = \min\{c, \max(-c, x)\}$ , condition (C.1) holds automatically. Under

condition  $\kappa_n(\beta_0) = 0$ , the estimation Eq. (2.2) is identifiable, that is we assume the model is properly specified. To obtain an asymptotic linear expansion of (2.2), we need (C.2). The assumptions (C.3), (C.4) are the usual conditions for central limit theory and they are expected to hold under general design. The conditions (C.5), (C.6) are imposed on the penalty function and are key to obtaining the oracle property. For  $\beta_{0j} \neq 0$ , condition (C.5) prevents the  $j$ th component of  $U_n^R(\beta)$  from being dominated by the penalty term, but for  $\beta_{0j} = 0$ , the condition (C.6) implies that the  $j$ th component of  $U_n^R(\beta)$  is dominated by the penalty term, such that any consistent solution of  $U_n^R(\beta) = 0$ , say  $\beta^*$ , must satisfy  $\beta_j^* = 0$ .

If we choose proper regularization parameter  $\lambda_n$ , several commonly used penalties will satisfy conditions (C.5)–(C.6). For the SCAD penalty and Hard penalty, we choose  $\lambda_n$  such that

$$\lambda_n \rightarrow 0, \quad \sqrt{n}\lambda_n \rightarrow \infty. \quad (2.5)$$

For the SCAD penalty, we have

$$q_{\lambda_n}(|\beta|) = \lambda_n \{I(|\beta| < \lambda_n) + \frac{(a\lambda_n - |\beta|)_+}{(a-1)\lambda_n} I(|\beta| \geq \lambda_n)\},$$

where  $c_+ = cI(c \geq 0)$ ,  $a > 2$ . Under condition (2.5), we have  $\lim_{n \rightarrow \infty} \sqrt{n}q_{\lambda_n}(|\beta|) = \lim_{n \rightarrow \infty} q'_{\lambda_n}(|\beta|) = 0$  and  $\sqrt{n} \inf_{|\beta| \leq Mn^{-1/2}} q_{\lambda_n}(|\beta|) = \sqrt{n}\lambda_n \rightarrow \infty$ . For the Hard penalty, that is,

$$q_{\lambda_n}(|\beta|) = 2(\lambda_n - |\beta|)I(|\beta| < \lambda_n),$$

it is again straightforward to verify that conditions (C.5)–(C.6) hold.

For the ALASSO penalty, we assume that

$$\sqrt{n}\lambda_n \rightarrow 0, \quad n\lambda_n \rightarrow \infty. \quad (2.6)$$

If we take  $q_{\lambda_n}(|\beta|) = \lambda_n/|\tilde{\beta}|$ , then  $\sqrt{n}q_{\lambda_n}(|\beta|) = \sqrt{n}\lambda_n/|\tilde{\beta}| \rightarrow 0$  for  $1/|\tilde{\beta}| < \infty$ , and  $q'_{\lambda_n}(|\beta|) = 0$ . Because we have  $\sqrt{n}(\tilde{\beta} - \beta_0) = O_p(1)$  under some regularity conditions (see [8]), then  $\sqrt{n} \inf_{|\beta| \leq Mn^{-1/2}} q_{\lambda_n}(|\beta|) = \sqrt{n} \inf_{|\beta| \leq Mn^{-1/2}} \lambda_n/|\tilde{\beta}| = M^{-1}n\lambda_n \rightarrow \infty$ .

Since  $q_{\lambda_n}(|\beta|) = \lambda_n$  for LASSO penalty, and  $q_{\lambda_n}(|\beta|) = \lambda_{1n} + 2\lambda_{2n}|\beta|$  for EN penalty, so conditions (C.5)–(C.6) do not hold for these two penalty functions.

The following lemma states the asymptotic linear expression of the robust estimating equation in (2.2).

**Lemma.** Under (C.1)–(C.3), for any  $M > 0$ , we have

$$\sup_{\|\beta - \beta_0\| \leq Mn^{-1/2}} \|n^{-1/2} U_n^R(\beta) - n^{-1/2} U_n^R(\beta_0) - n^{1/2} D(\beta - \beta_0)\| = o_p(1).$$

The following theorem states the main theoretical results regarding the proposed penalized robust estimating equation based estimators, including the existence of the  $\sqrt{n}$  consistent estimator, and the sparsity and the asymptotic normality of the resulting estimators.

**Theorem.** Under the assumption of (C.1)–(C.6), the following results hold:

- There exists a solution  $\hat{\beta}$  of  $U_n^P(\beta) = 0$ , such that  $\hat{\beta} = \beta_0 + O_p(n^{-1/2})$ .
- For any  $\sqrt{n}$  consistent solution of  $U_n^P(\beta) = 0$ , denoted by  $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_d)^T$ , we have

$$\lim_n P\{\hat{\beta}_j = 0, j > s\} = 1.$$

- Denote  $\hat{\beta}_1 = (\hat{\beta}_1, \dots, \hat{\beta}_s)^T$ , and  $\beta_{01} = (\beta_{01}, \dots, \beta_{0s})^T$ , then  $n^{1/2}(D_{11} + \Sigma_{11})\{\hat{\beta}_1 - \beta_{01} + (D_{11} + \Sigma_{11})^{-1}b_n\} \rightarrow N(0, B_{11})$ , in distribution,

where  $D_{11}$ ,  $\Sigma_{11}$ ,  $B_{11}$  are the first  $s \times s$  submatrices of  $D$ ,  $\text{diag}\{-q'_{\lambda_n}(|\beta_0|)\text{sgn}(\beta_0)\}$ ,  $B$ , and  $b_n = -(q_{\lambda_n}(|\beta_{0s}|)\text{sgn}(\beta_{0s}))^T$ .

The theorem implies that when we choose a proper  $\lambda_n$ , our robust penalized estimating equation approach can simultaneously achieve the  $\sqrt{n}$  consistency of the regularized regression coefficient estimation and the consistency of variable selection. At the same time, the asymptotic variance-covariance matrix of  $\hat{\beta}_1$  is given by

$$n^{-1}(D_{11} + \Sigma_{11})^{-1}B_{11}(D_{11} + \Sigma_{11})^{-1},$$

which is approximately equal to the usual sandwich-covariance estimate of  $\hat{\beta}_1$  by condition (C.5). The proof of the theorem and lemma are presented in Appendix.

### 3. Implementation

In this paper, we solve the penalized robust estimating equation using a majorize–minorize(MM) algorithm [9]. Let  $\beta^{(k)}$  denote the  $k$ th iterate value for fixed  $\lambda$  and  $\beta^{(0)}$  be the solution to  $U_n^R(\beta) = 0$ , then the iterative MM algorithm is written:

$$\beta^{(k+1)} = \beta^{(k)} + [D\beta^{(k)} + \Delta_\lambda(\beta^{(k)})]^{-1}U_n^P(\beta^{(k)}), \quad k \geq 0,$$

where  $\Delta_\lambda(\beta) = \text{diag}(q_\lambda(|\beta_1|)/(\epsilon + |\beta_1|), \dots, q_\lambda(|\beta_d|)/(\epsilon + |\beta_d|))$ , here  $\epsilon$  is a small number, and is chosen to be  $10^{-6}$  in our simulation. The MM algorithm continues until successive iterates values are less than a user-defined threshold. In our case, we continue until  $\|\beta^{(k+1)} - \beta^{(k)}\| < 10^{-4}$ .

For fixed  $\lambda$ , the solution of  $U_n^P(\beta) = 0$  also involves the iteration alternate between  $\beta$  and the nuisance parameters  $\sigma$  and  $\alpha$ . If we choose the special case of working independence  $R_i(\alpha) = I$  and  $\psi(x) = x$ , Eq. (2.2) no longer depends on the nuisance parameters, an initial estimate  $\beta^{(0)}$  is taken to be the solution to  $U_n^R(\beta) = 0$  in this special case and it can be obtained from standard statistics software. Similar to He et al. [8], a robust estimate of  $\sigma$  is obtained through the median absolute deviation

$$\hat{\sigma} = \{1.483 \text{ median}\{|y_{ij} - x_{ij}^T \beta^*| - \text{median}(y_{ij} - x_{ij}^T \beta^*)|\}\},$$

where  $\beta^*$  is the current estimate of  $\beta$ . The specific robust estimate of the correlation parameter  $\alpha$  depends on the choice of correlation structure. In this paper, we use the same estimator of  $\alpha$  as Liang and Zeger [11], except that the Pearson residual we used here is Huber's robust version.

We need to choose  $\lambda$  for LASSO, ALASSO and Hard penalty functions,  $(a, \lambda)$  for the SCAD penalty and  $\lambda = (\lambda_1, \lambda_2)$  for the EN penalty. Fan and Li [5] indicated that the choice of  $a = 3.7$  performs well in a variety of situations, thus we use their suggestion throughout our numerical analysis. Here, we propose a robust GCV to choose  $\lambda$  for reducing the impact of outliers. We select the penalty parameter  $\lambda$  by minimizing the robustified GCV statistic

$$\text{GCV}_R(\lambda) = \frac{\text{RSS}_R(\lambda)/n}{\{1 - d(\lambda)/n\}^2}, \quad (3.1)$$

where  $\text{RSS}_R(\lambda)$  is the robustified residual sum of squares  $\|W(\psi(A^{-1/2}(Y - X\hat{\beta}_\lambda)))\|^2$ ,  $Y = (Y_1, \dots, Y_n)^T$ ,  $X = X = (X_1^T, \dots, X_n^T)^T$ ,  $W = \text{diag}\{W_1, \dots, W_n\}$ , and  $d(\lambda)$  is the effective number of parameters, that is  $d(\lambda) = \text{tr}([\hat{D} + \Delta_\lambda(\hat{\beta}_\lambda)]^{-1}\hat{D}^T)$ , here  $\hat{\beta}_\lambda$  is the solution of the penalized robust estimating equation when  $\lambda$  is fixed, and  $\hat{D} = \partial(\frac{1}{n}U_n^R(\hat{\beta}_\lambda))/\partial(\hat{\beta})$  is an estimation of  $D$ . This definition for an effective number of parameters is similar to Fan and Li [5] and Tibshirani [16] and is an extension of those defined by Johnson et al. [10] to the longitudinal data case. We select  $\hat{\lambda} = \text{argmin}_\lambda \text{GCV}_R$ .

### 4. Simulation results

We simulated 500 data sets from the following marginal regression model for longitudinal data:

$$y_{ij} = x_{ij}^T \beta_0 + \epsilon_{ij}, \quad i = 1, 2, \dots, n, \quad j = 1, 2, \dots, m. \quad (4.1)$$

We consider  $\beta_0 = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$ ,  $m = 5$ ,  $x_{ij}$  are drawn from a standard normal distribution with the correlation between the  $k$ th and  $l$ th component of  $x_{ij}$  equal to  $0.5^{|l-k|}$ ,  $\epsilon_i = (\epsilon_{i1}, \dots, \epsilon_{im})^T$  independently follow multivariate normal distribution  $N(0, \sigma^2 R_i(\alpha))$ . This model is an extension of those considered by Tibshirani [16] and Fan and Li [5] to the case of longitudinal data. In our simulations, the  $r$  in the weight function  $w_{ij}$  is chosen to be 1 and the constant  $c$  in Huber's function  $\psi_c(\cdot)$  is chosen to be 2.

We assess the performance of our proposed variable selection procedures from three aspects: model error, model accuracy and model complexity. Model error can measure the predictive error of the selected model, and the accuracy and complexity of the selected model can evaluate the performance of estimation methods for variable selection. For linear model (2.1), model error is  $\text{ME} \equiv (\hat{\beta} - \beta_0)^T E(XX^T)(\hat{\beta} - \beta_0)$ , see [5]. Since the mean of  $X$  is zero vector, we estimate  $E(XX^T)$  by its sample covariance matrix. We compare the model error of different variable selection procedures using the median of relative model error(MRME), where the ratio of model error (RME) is defined as the model error of the selected model over that of the oracle model for each Monte Carlo data set. The average correctly fit percentage (C-f) measures the accuracy of the model selection procedure, where “correctly fit” means that the procedure selects the exact subset model. We also compared the average numbers of regression coefficients that are correctly shrunk(C-s) to zero, which can measure the complexity of the selected model.

To illustrate the oracle properties of our proposed robust variable selection methods, we use the true correlation structure as a working correlation, and let  $n = 50, 100, 200$ . The results over 500 simulated datasets are reported in Table 1, which included the results for both robust methods and non-robust methods. The non-robust methods compared here are defined through the same penalized estimating equation as (2.3), except that  $W_i = I$  and  $\psi(x) = x$ . It is observed that the methods with oracle properties, which include SCAD, Hard and ALASSO, outperform LASSO and EN in terms of model error, correctly fit percentage and the numbers of regression coefficients that are correctly shrunk(C-s) to zero. As  $n$  increases, the MRME

**Table 1**

Simulation results on variable selection with normal errors, where the response is correlated as an exchangeable structure with  $\alpha = 0.5$ , and “-R” “-NR” represent robust and non-robust respectively.

	n=50			n=100			n=200		
	MRME C-f C-s			MRME C-f C-s			MRME C-f C-s		
SCAD-R	1.98	0.66	4.59	1.89	0.84	4.83	1.64	0.96	4.96
SCAD-NR	1.74	0.61	4.49	1.54	0.77	4.73	1.29	0.91	4.92
Hard-R	1.54	0.97	4.97	1.42	0.99	4.98	1.41	0.99	4.99
Hard-NR	1.31	0.98	4.98	1.19	0.98	4.98	1.14	0.99	4.99
ALASSO-R	1.86	0.97	4.97	1.58	0.97	4.98	1.57	0.99	4.99
ALASSO-NR	1.35	0.96	4.96	1.28	0.96	4.96	1.21	0.98	4.98
LASSO-R	3.10	0.26	3.98	3.57	0.31	4.03	3.21	0.34	4.14
LASSO-NR	2.84	0.17	3.80	3.38	0.21	3.86	2.71	0.21	3.92
EN-R	3.22	0.27	3.98	3.48	0.30	4.03	3.50	0.38	4.20
EN-NR	2.91	0.16	3.80	3.42	0.22	3.86	2.71	0.27	3.95

**Table 2**

Simulation results on penalized robust estimating equation versus the penalized non-robust estimating equation, where the working correlation is taken as the true correlation (exchangeable,  $\alpha = 0.5$ ).

Methods	Contamination 1			Contamination 2			Contamination 3		
	MRME	C-f	C-s	MRME	C-f	C-s	MRME	C-f	C-s
SCAD-R	2.12	0.52	4.18	2.52	0.60	4.45	2.58	0.70	4.53
SCAD-NR	5.00	0.42	3.34	9.05	0.54	4.28	21.2	0.65	4.53
Hard-R	2.12	0.95	4.95	2.08	0.96	4.96	2.10	0.96	4.96
Hard-NR	7.86	0.94	4.83	7.22	0.95	4.98	11.5	0.94	4.98
ALASSO-R	1.92	0.96	4.96	2.20	0.97	4.97	1.95	0.96	4.97
ALASSO-NR	5.61	0.90	4.80	12.6	0.96	4.97	27.0	0.95	4.88
LASSO-R	4.32	0.35	3.82	4.42	0.31	3.99	3.52	0.41	3.99
LASSO-NR	5.26	0.25	3.80	9.96	0.22	3.86	22.2	0.40	4.00
EN-R	4.59	0.28	3.29	4.88	0.23	3.39	3.94	0.21	3.64
EN-NR	21.2	0.20	3.15	23.6	0.17	2.86	23.4	0.16	3.55

of SCAD, Hard and Alasso decrease, at the same time, their C-f approach to 1 and their C-s goes to 5. Furthermore, it is also found that when there are no outliers in the data, robust methods and non-robust methods performed similarly in terms of C-f and C-s, while the non-robust methods have less MRME than that of robust methods. This is not a surprise since the robust method will lose efficiency in some degree when the data contains no outliers.

To investigate the robustness of our proposed robust variable selection method against outliers, we consider three methods to create outliers:

Contamination 1. We perturb the covariate  $x_{ij}$  for one randomly chosen subject to  $x_{ij} + 3.5$ .

Contamination 2. We perturb the covariate  $x_{ij}$  for one randomly chosen subject to  $x_{ij} + 3.5$  and perturb the response  $y_{ij}$  for another randomly chosen subject to  $y_{ij} + 10$ .

Contamination 3. We perturb the covariate  $x_{ij}$  for three randomly chosen subjects to  $x_{ij} + 3.5$  and perturb the response  $y_{ij}$  for another three randomly chosen subjects to  $y_{ij} + 10$ .

We compare the performance of the non-robust variable selection method with that of the corresponding proposed robust method when data is contaminated with outliers, where  $n = 50$  and  $\sigma = 1$ . It is observed from Table 2, that apparently the proposed robust variable selection method can decrease the model error noticeably, especially when the contamination percentage is high. At the same time, the proposed robust methods can increase the correctly fit percentage and reduce the model complexity. We think the underlying reason for the good performance of robust methods is that the bounded score function and the covariate dependent weights can reduce the MSE of the resulting estimators in the presence of outliers. Results for  $n = 100, 200$  are similar and thus are omitted.

To study the sensitivity of the choice of working correlation structure on the variable selection, we compare every procedure under three working correlation structures: exchangeable (EX), AR(1) and working independence (Inde), the results are reported in Tables 3 and 4. Table 3 summarizes the results for non-contaminated data, and then we perturb the data by Contamination 1 to compare the robust methods and corresponding non-robust methods in Table 4. We set  $n = 50$  in this simulation.

It is shown from Tables 3 and 4 that the estimation procedures with a correct AR(1) working correlation have the smallest MRME and highest correctly fitted percentage (C-f) as well as C-s, and thus the estimators perform much better than their counterparts with independent working correlation that ignores within-subject correlation. For example, when  $\alpha = 0.8$ , the C-f gained by incorporating correlation as much as 22% in the SCAD approach. Estimation based on a misspecified exchangeable correlation structure will lead to some efficiency loss compared with using the correct AR(1) structure, but in general, it still performs better than using a working independent structure, especially in the  $\alpha = 0.8$  case. In intuition, the reason for the working independence tended to be associated with the highest MRME is it led to a poor variance estimate.



**Table 3**

Simulation study for the sensitivity of our propose robust variable selection procedure about the working correlation structure, where the true correlation structure is AR(1) and  $\alpha = 0.2, 0.5, 0.8$ .

Method		$\alpha = 0.2$			$\alpha = 0.5$			$\alpha = 0.8$		
		MRME C-f C-s			MRME C-f C-s			MRME C-f C-s		
Scad	EX	2.08	0.65	4.80	2.07	0.61	4.55	1.38	0.62	4.57
	AR(1)	2.21	0.66	4.79	1.60	0.74	4.64	1.15	0.83	4.88
	Inde	2.35	0.63	4.61	2.36	0.63	4.58	2.38	0.63	4.57
Hard	EX	2.17	0.89	4.89	1.60	0.90	4.88	1.14	0.94	4.96
	AR(1)	2.07	0.89	4.88	1.57	0.90	4.87	1.06	0.96	4.96
	Inde	2.14	0.85	4.88	2.04	0.87	4.84	2.03	0.86	4.88
ALASSO	EX	2.17	0.60	4.54	1.83	0.61	4.57	1.27	0.65	4.70
	AR(1)	2.10	0.89	4.86	1.72	0.93	4.90	1.19	0.98	4.98
	Inde	2.21	0.87	4.88	2.11	0.88	4.89	2.31	0.89	4.84
LASSO	EX	4.17	0.21	3.80	3.74	0.27	3.89	2.44	0.37	4.26
	AR(1)	3.97	0.22	3.86	3.31	0.28	3.93	2.10	0.45	4.30
	Inde	4.38	0.20	3.77	4.09	0.25	3.80	4.04	0.27	3.79
EN	EX	4.63	0.20	3.70	4.02	0.21	3.74	2.70	0.29	3.99
	AR(1)	4.48	0.20	3.82	3.37	0.25	3.80	2.14	0.37	4.61
	Inde	5.02	0.18	3.56	4.39	0.20	3.77	4.59	0.22	3.98

**Table 4**

Simulation results for various selection methods in the presence of outliers with AR(1) true correlation, where the correlation parameter  $\alpha = 0.8$ . The data was contaminated by randomly selected one subject and plus 3.5 to the covariate.

Methods	MRME			C-f			C-s		
	AR(1)	EX	Ind	AR(1)	EX	Ind	AR(1)	EX	Ind
SCAD-NR	5.64	11.8	20.5	0.76	0.65	0.39	4.98	4.92	3.78
SCAD-R	1.50	1.94	4.44	0.88	0.67	0.56	4.99	4.87	4.44
Hard-NR	3.92	15.1	21.3	0.98	0.98	0.68	4.94	4.96	4.65
Hard-R	1.15	1.30	3.22	0.99	0.99	0.76	4.99	4.98	4.74
ALASSO-NR	3.99	11.2	20.5	0.97	0.99	0.66	4.98	4.97	4.64
ALASSO-R	1.92	1.88	3.97	1.00	0.98	0.85	5.00	4.98	4.73
LASSO-NR	5.00	5.51	19.9	0.52	0.60	0.38	4.38	4.21	3.64
LASSO-R	2.06	2.84	4.36	0.56	0.61	0.53	4.43	4.32	3.89
EN-NR	6.29	6.85	18.5	0.37	0.36	0.29	4.16	4.29	3.46
EN-R	2.54	2.65	4.68	0.46	0.53	0.30	4.34	4.28	3.68

In general, estimation procedures with oracle properties are more efficient than the other methods. Furthermore, we can see from Table 4, that when data contain outliers, nonrobust methods can impact greatly in terms of MRME, despite which working correlation structure was used, while the corresponding robust methods perform much better. The results of a simulation experiment with exchangeable true correlation structure are similar and thus omitted.

As we mentioned in the Introduction, several methods exist for variable selection in the framework of a generalized estimating equation, such as [12,2,6,4]. The methods of Pan [12] and Cantoniet al. [2] are based on discrete penalized criterion, and the method developed in [6] is a penalized generalized estimating equation using continuous bridge penalties, which include LASSO as a special case. Dziak [4] extended Fu's work by replacing the bridge penalties with SCAD. Since we focus on the classical linear model in a Gaussian family for longitudinal data in this paper, we compare our results with a part of the results computed by Dziak [4], which is also a linear model with Gaussian distribution in a classical non-robust setting when no outliers were considered in the data.

Dziak [4] (Page 37) simulated 100 data sets from the following marginal regression model for longitudinal data:

$$y_{ij} = 20 + x_{ij}^T \beta_0 + 3\epsilon_{ij}, \quad i = 1, 2, \dots, 50, j = 1, 2, \dots, 7. \quad (4.2)$$

He set  $\beta_0 = (3, 1, 0, 0, 2, 0, 0, 0, 0.5, 0)^T$ ,  $x_{ij}$  are drawn from standard normal distribution with the correlation between the  $k$ th and  $l$ th component of  $x_{ij}$  equal to  $0.6^{|l-k|}$ , with  $\epsilon_i = (\epsilon_{i1}, \dots, \epsilon_{im})^T$  multivariate normal with either AR(1) or exchangeable true correlation, correlation parameter  $\alpha = 0.7$ . Under this setting, he also computed the results for estimation and variable selection using the methods of Pan [12], Fu [6], Cantoniet al. [2]. We compute our results by setting  $\psi(x) = x$  and  $\omega_{ij} = 1$ , and summarize the comparison in Tables 5 and 6 for true exchangeable correlation.

It is found from Tables 5 and 6 that three penalized estimating equation methods, including [6,4] and our methods, tend to perform best when using a correct exchangeable correlation structure. In general, using a misspecified AR(1) working correlation also performs better than ignoring the within-cluster correlation because these methods are based on a penalized estimating equation which can incorporate working correlation to increase the efficiency. However, Pan [12] and Cantoni et al. [2] used best subset selection methods and these methods seem not very sensitive to misspecified correlation. The advantage of these two subset selection methods is that they enjoy lowest model error and have a comparable variable

**Table 5**

Comparison results with Exchangeable true correlation. The columns show the mean model error, mean number of correct deletions (best is 6) and mean number of incorrect deletions (worst is 4) of the selected model under different working correlation structures. Pan-AIC and Cantoni- $C_p$  represent the methods of Pan [12] and Cantoni et al. [2]. Fu-LASSO<sub>1</sub> and Fu-LASSO<sub>2</sub> represent [6]'s methods tuning by a Quasi-GCV criterion and BIC respectively. D-SCAD<sub>1</sub> and D-SCAD<sub>2</sub> represent Dzak's results with SCAD penalties and tuning by QGCV and BIC. These results were computed by ([4], page 57–60). M-LASSO and M-SCAD represent the result given by our nonrobust methods.

Methods	Mean model error			Correct deletion			Wrong deletion		
	Ind	AR(1)	EX	Ind	AR(1)	EX	Ind	AR(1)	EX
Pan-AIC	0.23	0.24	0.20	5.35	4.82	4.72	0.35	0.24	0.24
Cantoni- $C_p$	0.22	0.24	0.20	5.46	4.98	4.86	0.36	0.28	0.25
Fu-LASSO <sub>1</sub>	0.22	0.35	0.29	3.22	2.32	1.32	0.04	0.00	0.00
Fu-LASSO <sub>2</sub>	0.23	0.31	0.25	3.78	1.97	1.01	0.05	0.00	0.00
D-SCAD <sub>1</sub>	0.24	0.24	0.21	4.87	5.65	5.76	0.26	0.22	0.14
D-SCAD <sub>2</sub>	0.19	0.22	0.18	5.16	4.65	4.36	0.24	0.03	0.02
M-LASSO	0.31	0.27	0.25	4.52	4.77	4.99	0.25	0.05	0.02
M-SCAD	0.34	0.23	0.29	5.69	5.78	5.69	0.27	0.15	0.05
M-ALASSO	0.28	0.26	0.25	5.84	5.97	5.99	0.38	0.40	0.39
M-Hard	0.39	0.28	0.25	5.96	5.98	5.99	0.97	0.97	0.66
M-EN	0.36	0.26	0.23	4.48	4.46	4.90	0.06	0.00	0.01

**Table 6**

Comparison for variable selection results with Exchangeable true correlation. The columns labeled C, O, U and M show the proportion of the correct model, an overfit model (at least one erroneous inclusion but no erroneous deletion), an underfit model (no erroneous inclusion but at least one erroneous deletion), or a misfit model (both erroneous inclusion and erroneous deletion) that was selected by each method.

Methods	Working Ind				Working AR(1)				Working EX			
	C	O	U	M	C	O	U	M	C	O	U	M
Pan-AIC	.35	.31	.14	.21	.24	.52	.06	.19	.21	.56	.07	.17
Cantoni- $C_p$	.35	.31	.14	.21	.26	.47	.09	.19	.22	.53	.08	.18
Fu-LASSO <sub>1</sub>	.03	.94	.01	.04	.01	.99	.00	.00	.00	1	.00	.00
Fu-LASSO <sub>2</sub>	.05	.90	.01	.04	.01	.99	.00	.00	.00	1	.00	.00
D-SCAD <sub>1</sub>	.27	.48	.15	.11	.50	.29	.20	.02	.67	.20	.13	.01
D-SCAD <sub>2</sub>	.32	.20	.41	.08	.50	.27	.20	.03	.65	.27	.08	.01
M-LASSO	.16	.75	.01	.08	.25	.71	.00	.04	.32	.66	.00	.02
M-SCAD	.56	.21	.17	.07	.68	.21	.10	.01	.69	.25	.04	.02
M-ALASSO	.55	.08	.31	.06	.60	.01	.37	.02	.60	.01	.39	.00
M-Hard	.19	.00	.77	.04	.20	.00	.78	.02	.52	.00	.47	.01
M-EN	.13	.81	.01	.05	.19	.81	.00	.00	.26	.73	.00	.01

selection performance. Fu's methods based on LASSO penalty are preservative in that they always have lowest wrong deletion and highest over-fit. Dzak's SCAD based penalized methods perform much better than the counter part of Fu's methods. The superiority of our methods is that our SCAD and ALASSO penalized solution always get a high correct-fit proportion and get a comparable model error. The main difference between our methods and that of Fu's and Dzak's is that we use different tuning criterions. The tuning criteria used by Fu and Dzak include quasi-GCV (QGCV) and BIC, while we use GCV defined in (3.1). It seems that the tuning criterion does effect the estimation and variable selection. A careful comparison between different tuning criteria is of interest to the authors but beyond the scope of the current paper. For true AR(1) correlation, the comparison obtains a similar conclusion and is thus omitted.

In conclusion, our proposed robust variable selection methods perform much more efficiently than the classical non-robust methods when data are contaminated with outliers. Furthermore, for longitudinal data, when the observations in the same subjects are highly correlated, incorporation of the correlation structure into the procedure of variable selection will lead to better performance than ignoring the correlation and using a working independent correlation structure. Compared to the existing variable selection methods for longitudinal data in a linear regression model, our methods (non-robust setting) based on SCAD and ALASSO penalties perform better in terms of selection and obtain a comparable model error.

## 5. Real data analysis

In this section, we applied the proposed robust variable selection methods to analyze the longitudinal progesterone data of [15]. This longitudinal hormone study on progesterone has collected urine samples from 34 women in a menstrual cycle. Among the 34 subjects, a total of 492 observations were obtained, with each woman contributing from 11 to 28 observations over time. Urine samples were collected from each woman in a menstrual cycle and urinary progesterone was assayed on alternate days. Each woman's menstrual cycle lengths were standardized uniformly to a reference 28-day cycle. The log-transformed progesterone level is taken to be the response. In addition to the time effect, two covariates are available. They are age and body mass index(BMI).



**Table 7**

Penalized robust estimating equation estimates for progesterone data.

	SCAD-R	Hard-R	ALASSO-R	LASSO-R	EN-R
Age	0 (–)	0 (–)	0 (–)	0 (–)	0 (–)
BMI	0 (–)	0 (–)	0 (–)	0 (–)	0 (–)
Time	0.4804 (0.0361)	0.5605 (0.1269)	0.5685 (0.0412)	0.4977 (0.0374)	0.4936 (0.0374)
Age*BMI	0 (–)	0 (–)	0 (–)	0 (–)	0 (–)
Age*time	0 (–)	0 (–)	0 (–)	0 (–)	0 (–)
BMI*time	0 (–)	0 (–)	0 (–)	0 (–)	0 (–)
Time*time	0.0217 (0.0060)	0 (–)	0 (–)	0.0305 (0.0085)	0.0286 (0.0078)
Intercept	0.8768 (0.0748)	0.9864 (0.0900)	0.9431 (0.0818)	0.8718 (0.0761)	0.8702 (0.0754)
$MSE_{CV}$	3.0094	3.1901	3.0176	3.1104	3.0832

**Table 8**

Penalized non-robust estimating equation estimates for progesterone data.

	SCAD-NR	Hard-NR	ALASSO-NR	LASSO-NR	EN-NR
Age	0 (–)	1.7269 (1.9972)	0 (–)	0 (–)	0 (–)
BMI	0 (–)	–2.4345 (1.6674)	0 (–)	0 (–)	0 (–)
Time	0.5301 (0.0374)	0.5458 (0.0346)	0.6143 (0.0436)	0.5685 (0.0400)	0.5788 (0.0400)
Age*BMI	0 (–)	0 (–)	0 (–)	0 (–)	0 (–)
Age*time	0 (–)	0 (–)	0 (–)	0 (–)	0 (–)
BMI*time	0 (–)	0 (–)	0 (–)	0 (–)	0 (–)
Time*time	0.0522 (0.0161)	0 (–)	0 (–)	0.0785 (0.0291)	0.0834 (0.0332)
Intercept	0.8681 (0.0819)	0.9412 (0.0800)	0.9631 (0.0837)	0.8717 (0.0889)	0.8756 (0.0917)
$MSE_{CV}$	3.0515	3.3771	3.0259	3.1058	3.1107

Let  $Y_{ij}$  denote the  $j$ th log-transformed progesterone value measured standardized day  $t_{ij}$  for the  $i$ th woman, and by  $Age_i$  and  $BMI_i$  her age and body mass index. Zhang et al. [17] fitted the semiparametric mixed model to the data, where the time effect is non-linear. Fung et al. [7] detected the outliers of this data using semiparametric mixed models. We consider the following linear model:

$$Y_{ij} = \beta_0 + \beta_1 Age_i + \beta_2 BMI_i + \beta_3 t_{ij} + \beta_4 Age_i BMI_i + \beta_5 Age_i t_{ij} + \beta_6 BMI_i t_{ij} + \beta_7 t_{ij}^2 + \epsilon_{ij}$$

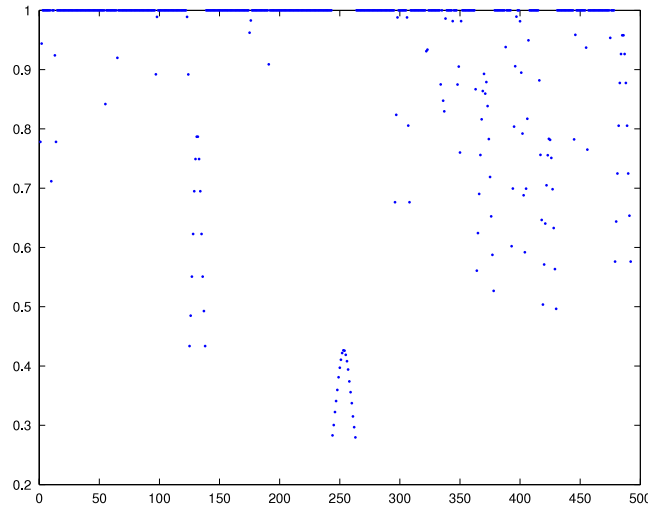
where  $\epsilon_{ij}$  are independent measurement errors following a normal  $(0, \sigma^2)$  distribution. We assume the serial correlation between consecutive days on the transformed time scale is the same for different women, and we use the AR(1) correlation structure as the working correlation.

We also implement a leave-one-out cross validation procedure to compare the performance among different penalize methods. We assess the goodness of fit using the mean squared error for cross validation procedures ( $MSE_{CV}$ )

$$MSE_{CV} = \frac{1}{n} \sum_{i=1}^n \|Y_i - X_i \hat{\beta}_{(-i)}\|,$$

where  $n = 34$ . The  $\hat{\beta}_{(-i)}$  is the estimate obtained based on the data of the other 33 subjects except the  $i$ th subject. We summarize the results both for robust and non-robust settings in Tables 7 and 8.

It is found that robust methods based on SCAD, LASSO and EN penalty estimators yield very similar models, they choose Time and Time\*Time covariates. ALASSO and the Hard robust estimator only kept Time covariate in the models. These result seem to be consistent with the findings in [7,17], where Age and BMI were found to have no significant effect on progesterone level. Compared to the non-robust estimates with the corresponding robust ones, we found that the point estimates are similar, however, the standard errors for the robust estimates coefficients are uniformly smaller than their corresponding errors of non-robust estimates coefficients counterparts. Furthermore, from these tables, we see that  $MSE_{CV}$



**Fig. 1.** Observation weights. 492 observation labeled on horizontal axis and their weights in robust estimating Eq. (2.2) labeled on vertical axis.

of the predictive model is smallest when we use the robust SCAD method, and largest when we use the nonrobust Hard method, and generally, robust methods perform better than the corresponding non-robust one in terms of  $MSE_{CV}$ . It can be seen that the Hard-nonrobust method generated an over-fitted model and gave a large standard deviation for term Age and BMI. This may be heuristically understood as the fact that progesterone data may contain outlier points.

We plot the weight of each observation in Fig. 1, the weight is defined as

$$s_{ij} = w_{ij} \frac{\psi((y_{ij} - x_{ij}\beta^{(0)})/\hat{\sigma})}{(y_{ij} - x_{ij}\beta^{(0)})/\hat{\sigma}},$$

where  $\beta^{(0)}$  is the solution to  $U_n^R(\beta) = 0$ . These weights allow one to identify the outlying observations. The 18th subject corresponding to the 244th observation to 263th observation is heavily downweighted (weight less than 0.4), and the weights of the 244th observation and 263th observation less than 0.3. These means the observation for the 18th subject, especially the 244th observation and 263th observation, are outliers of progesterone data under the linear regression model. These findings seem to be consistent with [7] in some degree. The 18th subject was one of the influential subjects according to Funget al. [7].

## 6. Conclusion

We have developed a robust estimator for longitudinal data in the linear regression model. Our estimator extends the usual robust estimation for longitudinal data [8,14] by solving a penalized robust estimating equation, the resulting estimators simultaneously select variables and estimate their regression coefficients. Our proposed estimate and variable selection methods can automatically downweight the outliers which commonly arise in longitudinal data. Compared to some variable selection methods for longitudinal data that consider robustness in literature, our proposed methods can easily implemented using standard software. We also find that for longitudinal data, the performance of our proposed variable selection procedure does depend on proper specification of the working correlation structure. Further work includes extending the proposed robust variable selection methods to generalized linear models and studying the effect of different tuning criteria on estimation and selection procedures.

## Acknowledgment

This work was supported by the National Natural Science Foundation of China (10931002, 1091120386, 10801039). The authors thank two anonymous referees for their insightful comments and helpful suggestions.

## Appendix. Sketch of proofs

**Proof of Lemma.** Let  $\mathcal{F} = \{U_\beta : \beta \in \Theta, \Theta \subseteq \mathbb{R}^d\}$ , where  $\Theta$  is bounded. Under condition (C.1),  $\mathcal{F}$  is Donsker class, see Vaart, A.W. van der, 1998, (Page 270, Theorem 19.5). Let  $\mathbb{G}_n U_\beta$  denote  $\sqrt{n}(\mathbb{P}_n - P)U_\beta$ , where  $\mathbb{P}_n$  is the empirical measure,

that is,  $\mathbb{P}_n U_\beta = \frac{1}{n} U_n^R(\beta)$ . By (C.3) and the properties of Donsker class, for every  $\beta$  satisfied  $\|\beta - \beta_0\| \leq Mn^{-1/2}$ , we have  $\mathbb{G}_n U_\beta - \mathbb{G}_n U_{\beta_0} \rightarrow 0$ , in probability. By (C.2), the convergence is uniform about  $\beta$  in the neighborhood of  $\beta_0$ , so we have

$$\sup_{\|\beta - \beta_0\| \leq Mn^{-1/2}} \|\mathbb{G}_n U_\beta - \mathbb{G}_n U_{\beta_0}\| \rightarrow 0, \quad \text{in probability.} \quad (\text{A.1})$$

Notice that  $\kappa_n(\beta_0) = 0$ , so for every  $\beta$  satisfied  $\|\beta - \beta_0\| \leq Mn^{-1/2}$ ,

$$\begin{aligned} \mathbb{G}_n U_\beta - \mathbb{G}_n U_{\beta_0} &= \sqrt{n} \frac{1}{n} U_n^R(\beta) - \sqrt{n} \kappa_n(\beta) - \sqrt{n} \frac{1}{n} U_n^R(\beta_0) + \sqrt{n} \kappa_n(\beta_0) \\ &= n^{-1/2} U_n^R(\beta) - n^{-1/2} U_n^R(\beta_0) - n^{1/2} \kappa_n(\beta) \\ &= n^{-1/2} U_n^R(\beta) - n^{-1/2} U_n^R(\beta_0) - n^{1/2} D(\beta - \beta_0) - n^{1/2} o_p(n^{-1/2}) \\ &= n^{-1/2} U_n^R(\beta) - n^{-1/2} U_n^R(\beta_0) - n^{1/2} D(\beta - \beta_0) - o_p(1) \\ &= o_p(1). \end{aligned}$$

So, (A.1) can be rewritten as

$$\sup_{\|\beta - \beta_0\| \leq Mn^{-1/2}} \|n^{-1/2} U_n^R(\beta) - n^{-1/2} U_n^R(\beta_0) - n^{1/2} D(\beta - \beta_0)\| = o_p(1).$$

This completes the Lemma.  $\square$

**Proof of Theorem.** We consider  $\hat{\beta} = (\hat{\beta}_1^T, \mathbf{0}^T)^T$ , where  $\hat{\beta}_1^T = \beta_{01} - n^{-1} D_{11}^{-1} U_1^R(\beta_0)$ , here  $\beta_{01}$  and  $U_1^R(\cdot)$  denote the first  $s$ -components of  $\beta_0$  and  $U_n^R(\cdot)$ . By condition (C.3), (C.4) and the central limit theorem, we have

$$n^{-1/2} D_{11}^{-1} U_1^R(\beta_0) \rightarrow N(0, D_{11}^{-1} B_{11} (D_{11}^{-1})^T), \quad \text{in distribution.}$$

Therefore,  $n^{-1/2} D_{11}^{-1} U_1^R(\beta_0) = O_p(1)$ , that is  $\hat{\beta} = \beta_0 + O_p(n^{-1/2})$ . To prove (a), we show  $\hat{\beta}$  is a solution of  $U_n^p(\beta) = 0$  next.

Let  $U_1^p(\cdot)$  denote the first  $s$ -components of  $U_n^p(\cdot)$ , by the conclusion of the lemma, we have

$$\begin{aligned} n^{-1/2} U_1^p(\hat{\beta}) &= n^{-1/2} U_1^R(\hat{\beta}) - n^{1/2} q_{\lambda_n}(|\hat{\beta}_1|) \text{sgn}(\hat{\beta}_1) \\ &= n^{-1/2} U_1^R(\beta_0) + n^{1/2} D_{11}(\hat{\beta}_1 - \beta_{01}) + o_p(1) - n^{1/2} q_{\lambda_n}(|\hat{\beta}_1|) \text{sgn}(\hat{\beta}_1) \\ &= o_p(1) - n^{1/2} q_{\lambda_n}(|\hat{\beta}_1|) \text{sgn}(\hat{\beta}_1). \end{aligned}$$

Because for  $j = 1, \dots, s$ ,  $\sqrt{n} q_{\lambda_n}(|\beta_{0j}|) \rightarrow 0$ , under condition (C.5), we have that, for any  $\epsilon \rightarrow 0^+$

$$\begin{aligned} n^{-1/2} U_j^p(\hat{\beta} \pm \epsilon e_j) &= o_p(1) - n^{1/2} q_{\lambda_n}(|\hat{\beta}_j \pm \epsilon|) \text{sgn}(\hat{\beta}_j \pm \epsilon) \\ &= o_p(1), \end{aligned}$$

where  $U_j^p(\cdot)$  is the  $j$ th component of  $U_n^p(\cdot)$ .

Under condition (C.6), for  $j = s+1, \dots, d$ ,  $\hat{\beta}_j = 0$ ,  $n^{-1/2} U_j^p(\hat{\beta} + \epsilon e_j)$  and  $n^{-1/2} U_j^p(\hat{\beta} - \epsilon e_j)$  are dominated by  $-n^{1/2} q_{\lambda_n}(\epsilon)$  and  $n^{1/2} q_{\lambda_n}(\epsilon)$ , so they have opposite signs when  $\epsilon$  goes to zero, this implies  $\hat{\beta}$  is a solution of  $U_j^p(\beta) = 0$ .

To proof part (b), we will show for any  $\epsilon > 0$ , for any  $\sqrt{n}$  consistent solution of  $U_n^p(\beta) = 0$ , denoted by  $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_d)^T$ , when  $n$  is sufficiently large, for any  $\epsilon > 0$ ,

$$P\{\hat{\beta}_j \neq 0, j = s+1, \dots, d\} < \epsilon.$$

Because  $\hat{\beta}_j = O_p(n^{-1/2})$ , for  $j = s+1, \dots, d$ , there exist some  $M > 0$  such that when  $n$  is large enough,

$$P\{\hat{\beta}_j \neq 0, j = s+1, \dots, d\} < \epsilon/2 + P\{\hat{\beta}_j \neq 0, j = s+1, \dots, d, |\hat{\beta}_j| < Mn^{-1/2}\} \quad (\text{A.2})$$

Next, we need to show  $P\{\hat{\beta}_j \neq 0, j = s+1, \dots, d, |\hat{\beta}_j| < Mn^{-1/2}\} < \epsilon/2$  when  $n$  is large enough.

Using the  $j$ th component of  $U_n^p(\hat{\beta})$  and the definition of the solution, we obtain that on the set of  $\{\hat{\beta}_j \neq 0, j = s+1, \dots, d, |\hat{\beta}_j| < Mn^{-1/2}\}$ ,  $[n^{-1/2} U_j^p(\hat{\beta})]^2 = o_p(1)$ , combined with the conclusion of Lemma 1, we have

$$\begin{aligned} [n^{-1/2} U_j^p(\hat{\beta})]^2 &= [n^{-1/2} U_j^R(\beta_0) + n^{1/2} D_j(\hat{\beta} - \beta_0) + o_p(1) - n^{1/2} q_{\lambda_n}(|\hat{\beta}_j|) \text{sgn}(\hat{\beta}_j)]^2 \\ &= [O_p(1) - n^{1/2} q_{\lambda_n}(|\hat{\beta}_j|) \text{sgn}(\hat{\beta}_j)]^2 \\ &= o_p(1), \end{aligned}$$

where  $D_j$  is the  $j$ th row of  $D$ . As a result, there exists some  $M' > 0$  such that for large  $n$ ,

$$P\{\hat{\beta}_j \neq 0, j = s + 1, \dots, d, |\hat{\beta}_j| < Mn^{-1/2}, n^{1/2}q_{\lambda_n}(|\hat{\beta}_j|) > M'\} < \epsilon/2. \quad (\text{A.3})$$

Because for any  $M > 0$ ,  $\lim_n \sqrt{n} \inf_{|\beta| \leq Mn^{-1/2}} q_{\lambda_n}(|\beta|) \rightarrow \infty$ ,  $\{\hat{\beta}_j \neq 0, j = s + 1, \dots, d, |\hat{\beta}_j| < Mn^{-1/2}\}$  implies that  $n^{1/2}q_{\lambda_n}(|\hat{\beta}_j|) > M'$  for large  $n$ . Thus

$$\begin{aligned} &P\{\hat{\beta}_j \neq 0, j = s + 1, \dots, d, |\hat{\beta}_j| < Mn^{-1/2}\} \\ &= P\{\hat{\beta}_j \neq 0, j = s + 1, \dots, d, |\hat{\beta}_j| < Mn^{-1/2}, n^{1/2}q_{\lambda_n}(|\hat{\beta}_j|) > M'\} < \epsilon/2. \end{aligned} \quad (\text{A.4})$$

Therefore, by (A.2), (A.3), (A.4), we have for any  $\epsilon > 0$  that there exist  $M > 0$ ,  $M' > 0$ , when  $n$  is large enough,

$$\begin{aligned} P\{\hat{\beta}_j \neq 0, j = s + 1, \dots, d\} &< \epsilon/2 + P\{\hat{\beta}_j \neq 0, j = s + 1, \dots, d, |\hat{\beta}_j| < Mn^{-1/2}\} \\ &= \epsilon/2 + P\{\hat{\beta}_j \neq 0, j = s + 1, \dots, d, |\hat{\beta}_j| < Mn^{-1/2}, n^{1/2}q_{\lambda_n}(|\hat{\beta}_j|) > M'\} \\ &< \epsilon/2 + \epsilon/2 = \epsilon. \end{aligned}$$

To prove part (c), using  $n^{-1/2}U_1^R(\beta_0) + n^{1/2}D_{11}(\hat{\beta}_1 - \beta_{01}) - n^{1/2}q_{\lambda_n}(|\hat{\beta}_1|)\text{sgn}(\hat{\beta}_1) = o_p(1)$ , and the Taylor expansion of the last term, we have that

$$\begin{aligned} &n^{1/2}D_{11}(\hat{\beta}_1 - \beta_{01}) - n^{1/2}q_{\lambda_n}(|\beta_{01}|)\text{sgn}(\beta_{01}) - n^{1/2}q'_{\lambda_n}(|\beta_{01}|)\text{sgn}(\beta_{01})(\hat{\beta}_1 - \beta_{01}) - n^{1/2}\frac{1}{2}o_p(\|\hat{\beta}_1 - \beta_{01}\|^2)\text{sgn}(\beta_{01}) \\ &= -n^{1/2}U_1^R(\beta_0) + o_p(1). \end{aligned}$$

Using  $b_n$ ,  $\Sigma_{11}$ , we can rewrite the above equation as

$$n^{1/2}D_{11}(\hat{\beta}_1 - \beta_{01}) + n^{1/2}\Sigma_{11}(\hat{\beta}_1 - \beta_{01}) + n^{1/2}b_n = -n^{1/2}U_1^R(\beta_0) + o_p(1).$$

So, we have

$$n^{1/2}(D_{11} + \Sigma_{11})\{\hat{\beta}_1 - \beta_{01} + (D_{11} + \Sigma_{11})^{-1}b_n\} = -n^{1/2}U_1^R(\beta_0) + o_p(1) \rightarrow_d N(0, B_{11}). \quad \square$$

## References

- [1] E. Cantoni, A robust approach to longitudinal data analysis, *Canadian Journal of Statistics* 32 (2004) 169–180.
- [2] E. Cantoni, J.M. Flemming, E. Ronchetti, Variable selection for marginal longitudinal generalized linear models, *Biometrics* 61 (2005) 507–514.
- [3] D.L. Donoho, I.M. Johnstone, Ideal spatial adaptation by wavelet shrinkage, *Biometrika* 81 (1994) 425–455.
- [4] J.J. Dziak, Variable selection for longitudinal data by penalized quadratic inference functions. Ph.D. Dissertation, Department of Statistics, The Pennsylvania State University, 2006.
- [5] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association* 96 (2001) 1348–1360.
- [6] W.J. Fu, Penalized estimating equations, *Biometrics* 59 (2003) 126–132.
- [7] W.K. Fung, Z.Y. Zhu, B.C. Wei, X. He, Influence diagnostics and outlier tests for semiparametric mixed models, *Journal of Royal Statistical Society, Ser. B* 64 (2002) 565–579.
- [8] X. He, W.K. Fung, Z.Y. Zhu, Robust estimation in generalized partial linear models for clustered data, *Journal of American Statistical Association* 100 (2005) 1176–1184.
- [9] D. Hunter, R. Li, Variable selection using MM algorithms, *Annals of statistics* 33 (2005) 1617–1642.
- [10] B.A. Johnson, D.Y. Lin, D.L. Zeng, Penalized estimating functions and variable selection in semiparametric regression models, *Journal of the American Statistical Association* 103 (2008) 672–680.
- [11] K.Y. Liang, S.L. Zeger, Longitudinal data analysis using generalized linear models, *Biometrika* 73 (1986) 13–22.
- [12] W. Pan, Akaike's information criterion in generalized estimating equations, *Biometrics* 57 (2001) 120–125.
- [13] J.S. Preisser, B.F. Qaqish, Deletion diagnostics for generalized estimating equations, *Biometrika* 83 (1996) 551–562.
- [14] S.K. Sinha, Robust inference in generalized linear models for longitudinal data, *Canadian Journal of Statistics* 34 (2006) 261–278.
- [15] M.F. Sowers, M. Crutchfield, J.F. Randolph, B. Shapiro, B. Zhang, M.L. Pietra, M.A. Schork, Urinary ovarian and gonadotrophin hormone levels in premenopausal women with low bone mass, *Journal of Bone Mining Research* 13 (1998) 1191–1202.
- [16] R. Tibshirani, Regression shrinkage and selection via the lasso, *Journal of Royal Statistical Society* 58 (1996) 267–288.
- [17] D.W. Zhang, X.H. Lin, J. Raz, M. Sowers, Semiparametric stochastic mixed models for longitudinal data, *Journal of American Statistical Association* 93 (1998) 710–719.
- [18] H. Zou, The adaptive lasso and its oracle properties, *Journal of American Statistical Association* 101 (2006) 1418–1429.
- [19] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, *Journal of Royal Statistical Society, Ser. B* 67 (2005) 301–320.