Efficient and doubly-robust methods for variable selection and parameter estimation in longitudinal data analysis

Liya Fu · Zhuoran Yang · Fengjing Cai · You-Gan Wang

Received: 1 February 2020 / Accepted: 1 October 2020 / Published online: 12 October 2020
© Springer-Verlag GmbH Germany, part of Springer Nature 2020

Abstract
New technologies have produced increasingly complex and massive datasets, such as next generation sequencing and microarray data in biology, dynamic treatment regimes in clinical trials and long-term wide-scale studies in the social sciences. Each study exhibits its unique data structure within individuals, clusters and possibly across time and space. In order to draw valid conclusion from such large dimensional data, we must account for intracluster correlations, varying cluster sizes, and outliers in response and/or covariate domains to achieve valid and efficient inferences. A weighted rank-based method is proposed for selecting variables and estimating parameters simultaneously. The main contribution of the proposed method is four fold: (1) variable selection using adaptive lasso is extended to robust rank regression so that protection against outliers in both response and predictor variables is obtained; (2) within-subject correlations are incorporated so that efficiency of parameter estimation is improved; (3) the computation is convenient via the existing function in statistical software R. (4) the proposed method is proved to have desirable asymptotic properties for fixed number of covariates \( p \). Simulation studies are carried out to evaluate the proposed method for a number of scenarios including the cases when \( p \) equals to the number of subjects. The simulation results indicate that the proposed method is efficient and robust. A hormone dataset is analyzed for illustration. By adding additional redundant variables as covariates, the penalty approach and weighting schemes are proven to be effective.

Keywords Correlated data · Outliers · Rank-based method · Variable selection

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00180-020-01038-3) contains supplementary material, which is available to authorized users.

Extended author information available on the last page of the article
1 Introduction

Longitudinal data is commonly utilized in economics, medical studies, and environmental research. A large number of covariates are often collected in longitudinal studies. The inclusion of redundant variables can reduce the accuracy and efficiency of parameter estimation. Therefore, it is important to select the appropriate covariates in analyzing longitudinal data. However, it is a challenge to select significant variables in longitudinal data due to underlying correlations and unavailable likelihood. Fan and Li (2004) provided a penalized weighted least-squares approach for variable selection in a semiparametric model in longitudinal data analysis. Ni et al. (2010) proposed a double-penalized Gaussian likelihood approach for simultaneous model selection and parameter estimation in a semiparametric mixed model for longitudinal data. Wang et al. (2012) and Cho and Qu (2013) considered the penalized generalized estimating equations (GEE) (Liang and Zeger 1986) and the penalized quadratic inference functions through smoothly clipped absolute deviation (SCAD) (Fan and Li 2001) with high dimension covariates. All the methods mentioned above are essentially based on the weighted least squares (WLS); thus, they are sensitive to outliers.

In longitudinal studies, the collected data often deviates from normality, and the response variable and/or covariates may contain some potential outliers, which often results in serious problems for variable selection and parameter estimation. Therefore, robust methods have attracted much attention in recent years. However, the literature on variable selection and against outliers in response or/and covariates for longitudinal data is quite limited. Fan et al. (2012) and Guo et al. (2014) constructed a penalized robust GEE approach by applying the Huber’s score function to the standardized residuals in linear regression models and the semiparametric mean-covariance regression models for longitudinal data, respectively. Lv et al. (2015) utilized a bounded exponential score function (Wang et al. 2013) in the GEE framework to choose variables and estimate parameters. However, both the Huber’s score function and the exponential score function require specifying a tuning parameter to control the level of robustness at the cost of efficiency loss.

The well-known rank-based method has many beneficial properties; for example, it is robust and distribution free (Jaeckel 1972). Jung and Ying (2003) and Fu et al. (2010) studied rank regression with longitudinal data under an independence assumption. Wang and Zhao (2008) considered a weighted rank method to take account of correlations and varying cluster sizes. Fan and Wang (2012) constructed a generalized estimating equations based on the rank method under an exchangeable correlation structure assumption. The above methods can’t be used to select variables and can be sensitive to outliers in covariates. As far as we know, studies on variable selection based on ranks are rather limited due to mathematical challenges and computational complexity.

Wang and Li (2009) and Yang et al. (2015) proposed weighted rank-based methods penalized by SCAD for automatic variable selection and parameter estimation, which are applicable for independent data only. Xu et al. (2010) utilized the rank-based method to choose variables in an accelerated failure time (AFT) model. Fu and Wang (2018) considered the rank-based method based on the independence assumption for variable selection, and the method is only robust for response variable. In this paper, we...
extend the rank-based method to longitudinal data and propose a weighted rank-based method for selecting covariates and estimating parameters based on the Wilcoxon dispersion function penalized by SCAD. The new method is robust against outliers in response and heavy-tailed distributions. It is also robust against leverage points in covariates. Furthermore, the proposed method is effective because of incorporating intracluster correlations and varying cluster sizes, and has the oracle properties. The computation of minimizing the penalized weighted dispersion function is convenient via the existing function in the statistical software R.

The rest of the paper is organized as follows: the proposed method is presented in Sect. 2. Simulation studies are carried out to evaluate the performance of the proposed method in Sect. 3. The data from the longitudinal hormone study is used to illustrate the proposed method in Sect. 4. Finally, our conclusions are drawn. The proof of the oracle properties is shown in the Appendix.

2 Methodology

Suppose \((Y_{ik}, X_{ik})\) are the observed response and predictors at the \(k\)th time point from the \(i\)th subject or cluster, where \(k = 1, \ldots, n_i\) and \(i = 1, \ldots, N\). Assume that observations from the different subject or clusters are independent and observations from the same subject or cluster are correlated. Here \(n_i\) is often referred to as cluster size. Consider the following linear regression model:

\[
Y_{ik} = \beta_0 + X_{ik}^T \beta + \epsilon_{ik},
\]

in which \(\beta_0\) is the intercept, \(\beta = (\beta_1, \ldots, \beta_p)^T\) is a \(p\) dimensional unknown parameter vector, and \(\epsilon_{ik}\) is an error term. Suppose that the median of \(\epsilon_{ik} - \epsilon_{jl}\) is zero when \(i \neq j\), and \(\epsilon_{i1}, \ldots, \epsilon_{in_i}\) are correlated. We partition \(\beta\) as \((\beta_{10}^T, \beta_{20}^T)^T\) with \(\beta_{10} \in \mathbb{R}^d\) and \(\beta_{20} \in \mathbb{R}^{p-d}\). Suppose that the true parameter values in model (1) are \(\beta_T = (\beta_{10}^*, \beta_{20}^T)^T\). We aim to identify the covariates with zero coefficients \(\beta_{20}\) consistently and resolutely, and meanwhile estimate the other nonzero coefficients.

2.1 Weighting approach for efficiency and robustness

To seek doubly-robust and efficient parameter estimates and simultaneously select important covariates, we propose minimizing the following penalized dispersion function with two given weights \(b_{ikjl}\) and \(w_{ij}\):

\[
Q_W(\beta) = M^{-2} \sum_{i=1}^{N} \sum_{j<i}^{N} n_i \sum_{k=1}^{n_j} \sum_{l=1}^{n_j} b_{ikjl} w_{ij} |\epsilon_{ik} - \epsilon_{jl}| + \sum_{s=1}^{p} P_\lambda(|\beta_s|),
\]

where \(M = \sum_{i=1}^{N} n_i\) is the total number of observations, and \(P_\lambda(\cdot)\) is a penalty function encouraging sparsity in \(\beta\), and \(\lambda > 0\) is a tuning parameter controlling the complexity of the model. The weight \(w_{ij} = w_i w_j\) is used to capture the effects of
the within-subject correlations and varying cluster sizes, and the weight $b_{ikjl}$ is used to control for effects of possible outliers in the covariates. Moreover, the first term in $Q_W(\beta)$ is based on the Wilcoxon-type dispersion function and hence derive robust estimates when the response variable contains outliers. When $b_{ikjl} = 1$ and $w_{ij} = 1$ for $i \neq j$, the first term in $Q_W(\beta)$ can yield the well-known rank-based Wilcoxon estimate, and the proposed penalized function $Q_W(\beta)$ will lead to the function of Fu and Wang (2018). For $b_{ikjl}$, we consider the generalized rank (GR) weight proposed by Naranjo et al. (1994) and the high-breakdown rank (HBR) weight proposed by Chang et al. (1999). For GR weight, $b_{ikjl} = h_{ikjl}$, in which $h_{ik} = \min \left\{ 1, \left[ \frac{c}{d_i^2(X_{ik})} \right]^{\kappa/2} \right\}$, and $d_i^2(X_{ik}) = (X_{ik} - \hat{\mu}_x)^T S_\kappa^{-1}(X_{ik} - \hat{\mu}_x)$, where $c$ and $\kappa$ are tuning constants and $\hat{\mu}_x$ and $S_\kappa$ are the robust estimates of the location and covariance of $X_{ik}$ (Rousseeuw and Zomeren 1990; Terpstra and McKean 2005). For the tuning parameters $\kappa$ and $c$, we use $\kappa = 2$ and $c = \chi^2_{0.95}(p)$, which is the 0.95 quantile of a $\chi^2(p)$ distribution. For HBR weight, $b_{ikjl} = \psi \left( \frac{c_2}{a_{ikl}} \right)$, in which $a_{ikl} = \hat{\epsilon}_{ik}/\hat{\sigma}(\hat{\chi}^2_{0.95}(p)/d_i^2(X_{ik}))$, where $\psi(t) = 1, t$ or $-1$ according to whether $t \geq 1, -1 < t < 1$ or $t \leq -1$, and the tuning constant $c_2 = \{\text{med}(a_{ikl}) + 3 \text{MAD}(a_{ikl})\}^2$, and $\hat{\sigma} = 1.483 \text{med}(|\hat{\epsilon}_{ik}(\hat{\beta}^0)| - \text{med}(|\hat{\epsilon}_{ik}(\hat{\beta}^0)|))$, where $\hat{\beta}^0 = (\hat{\beta}^0_1, \ldots, \hat{\beta}^0_p)^T$ is a consistent estimate of $\beta$.

For $w_i$, we consider $w_i = 1/(1 + (n_i - 1)\rho)$ proposed by Wang and Zhao (2008), where $\rho$ is the average correlation coefficient. This weight incorporates correlations and different cluster sizes in a simple way and the resulting estimator is efficient and robust (Wang and Zhao 2008). The parameter $\rho$ is estimated via the moment method and is given by the following formula (3), which was proposed by Wang and Carey (2003) and Wang and Zhao (2008).

$$\hat{\rho} = \frac{\sum_{i=1}^{N} \sum_{k=1}^{n_i} \sum_{l \neq k}^{n_i} (r_{ik} - \bar{r})(r_{il} - \bar{r})}{\sum_{i=1}^{N} (n_i - 1) \sum_{k=1}^{n_i} (r_{ik} - \bar{r})^2},$$

(3)

where $r_{ik} = \sum_{j=1}^{N} \sum_{l=1}^{n_j} I(\hat{\epsilon}_{jl} \leq \hat{\epsilon}_{ik})$, and $\bar{r}$ is the average of rank sum of all the residual terms. We use the widely used SCAD penalty function for the penalty function $P_\lambda(\cdot)$ (Fan and Li 2001). To reduce the computational burdens, Zou and Li (2008) proposed a local linear approximation to the SCAD penalty, which retains the same asymptotic properties. Therefore, the SCAD estimates are derived from the following objective function:

$$Q_W(\beta) = L_W(\beta) + \sum_{s=1}^{p} P_\lambda'(|\hat{\beta}^0_s|)|\beta_s|,$$
where

\[ L_W(\beta) = M^{-2} \sum_{i=1}^{N} \sum_{j<i}^{N} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} b_{ikjl} w_{ij} |\epsilon_{ik} - \epsilon_{jl}|, \]

and

\[ P'_\lambda(\theta) = \lambda \left\{ I(\theta \leq \lambda) + \frac{(a\lambda - \theta)^+}{(a-1)\lambda} I(\theta > \lambda) \right\}, \]

for \( a > 2 \) and \( \theta > 0 \). Fan and Li (2001) indicated that \( a = 3.7 \) performs well for a variety of cases; hence, we will use \( a = 3.7 \) throughout this paper.

Besides robustness and effectiveness, another appealing feature of the rank-based method with the SCAD penalty is that its computation can be conveniently carried out using the statistical software R since the penalty term can be easily merged with the first weighted term. The procedures are given as follows. The objective function \( Q_W(\beta) \) can be written as:

\[
Q_W(\beta) = M^{-2} \sum_{i=1}^{N} \sum_{j<i}^{N} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} b_{ikjl} w_{ij} |\epsilon_{ik} - \epsilon_{jl}| + \sum_{s=1}^{p} P'_\lambda(|\hat{\beta}_s^0|) |\beta_s| \\
= M^{-2} \sum_{i=1}^{N} \sum_{j<i}^{N} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} b_{ikjl} w_{ij} |Y_{ik} - Y_{jl}| - (X_{ik} - X_{jl})^T \beta | \\
+ \sum_{s=1}^{p} P'_\lambda(|\hat{\beta}_s^0|) |\beta_s| \\
= M^{-2} \sum_{r=1}^{M(M-1)/2+p} |\tilde{Y}_r - \tilde{X}_r^T \beta|,
\]

where \((\tilde{Y}_r, \tilde{X}_r)\) are pseudo observations. The first \( M(M-1)/2 \) pseudo-observations are \((b_{ikjl} w_{ij} (Y_{ik} - Y_{jl}), b_{ikjl} w_{ij} (X_{ik} - X_{jl}))\) for \( 1 \leq k \leq n_i, 1 \leq l \leq n_j \) and \( 1 \leq j < i \leq N \). The last \( p \) pseudo observations are \((0, M^2 P'_\lambda(|\hat{\beta}_s^0|) E_s), s = 1, \ldots, p\), where \( E_s \) is a \( p \)-dimensional vector, with the \( s \)th element being 1 and all the other elements being zeroes. Therefore, \( Q_W(\beta) \) can be treated as the \( L_1 \) loss function for the pseudo-data \((\tilde{Y}_r, \tilde{X}_r)\) for \( r = 1, \ldots, M(M-1)/2+p \). The penalty estimate can be obtained using the \( rq \) function in the \texttt{quantreg} package in the statistical software R (Koenker 2005). Furthermore, the initial estimate \( \hat{\beta}^0 \) can be obtained by minimizing \( L_W(\beta) \). Because \( Q_W(\beta) \) is invariant to location, the intercept \( \beta_0 \) cannot be simultaneously estimated with \( \beta \). We estimate it by the median of \( \{\epsilon_{ik}(\hat{\beta}), k = 1, \ldots, n_i; \quad i = 1, \ldots, N\} \).

For the tuning parameter \( \lambda \), we select it by a data driven method via minimizing the following objective function over a given interval,
\[
BIC_\lambda = \log \left( M^{-2} \sum_{i=1}^{N} \sum_{j<i}^{N} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} b_{ikjl} w_{ij} |Y_{ik} - Y_{jl} - (X_{ik} - X_{jl})^T \hat{\beta}_\lambda| \right) \\
+ \frac{df_\lambda \log N}{N},
\]

where \( \hat{\beta}_\lambda \) is the penalized estimator with a tuning parameter \( \lambda \), and \( df_\lambda \) is the number of nonzero components in \( \hat{\beta}_\lambda \).

### 2.2 Asymptotic properties

Let \( f_{ik} \) and \( F_{ik} \) be the density and cumulative distribution functions of \( \epsilon_{ik} \), respectively. Define \( D = M^{-2} \sum_{i=1}^{N} \sum_{j<i}^{N} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} b_{ikjl} w_{ij} (X_{ik} - X_{jl}) (X_{ik} - X_{jl})^T \int f_{ik} dF_{ij} \). Let \( D_{11} \) be the first \( d \times d \) submatrix of \( D \). Denote \( \Sigma_{11} = \text{diag}(P_{\lambda}''(|\beta_{10}^*|) \text{sign}(\beta_{10}^*), \ldots, P_{\lambda}''(|\beta_d^*|) \text{sign}(\beta_d^*)) \) and

\[
P_{\lambda}'(|\beta_{10}^*|) \text{sign}(\beta_{10}^*) = (P_{\lambda}'(|\beta_1^*|) \text{sign}(\beta_1^*), \ldots, P_{\lambda}'(|\beta_d^*|) \text{sign}(\beta_d^*))^T.
\]

The theorem below indicates that the proposed estimators have the oracle properties, and the proof of the theorem is given in the Appendix.

**Theorem 1** Under some regularity conditions given in the Appendix, the proposed estimator \( \hat{\beta} = (\hat{\beta}_{10}^T, \hat{\beta}_{20}^T)^T \) has the following properties, as \( \lambda \to 0 \) and \( \sqrt{M} \lambda \to \infty \),

(a) **Sparsity**: \( P(\hat{\beta}_{20} = 0) \to 1 \).
(b) **Asymptotic normality**:

\[
\sqrt{M}(\hat{\beta}_{10} - \beta_{10}^*) - (D_{11} + \Sigma_{11})^{-1} P_{\lambda}'(|\beta_{10}^*|) \text{sign}(\beta_{10}^*) \to N(0, B_{11}),
\]

where \( B_{11} = (D_{11} + \Sigma_{11})^{-1} V_{11} (D_{11} + \Sigma_{11})^{-1} \), in which \( V_{11} \) is the first \( d \times d \) block matrix of \( V \), as given in the Appendix.

### 3 Simulation studies

In this section, we carry out simulation studies to demonstrate the robustness and efficiency of the proposed method. The data is generated from the following model:

\[
Y_{ik} = X_{ik}^T \beta + \epsilon_{ik}, \quad k = 1, \ldots, n_i; \quad i = 1, \ldots, 50,
\]

where \( \beta = (3, 1.5, 2, 0_{p-3})^T \), and \( X_{ik1} \) and \( X_{ik2} \) are subject-level covariates; that is, they do not change within each subject or cluster, but may differ among subjects/clusters. Covariate \( X_{ik3} \) is a within-cluster covariate; it changes within each subject or cluster.

We generate \( X_{ik1} \) and \( X_{ik2} \) independently from the standard normal distribution and generate \( X_{ik3} \) from a uniform distribution \( U(-1, 1) \). Covariates \( (X_{ik4}, \ldots, X_{ikp}) \) are...
drawn from a multivariate normal distribution with a covariance matrix, \( R_i(0.5^{k-l}) \).

Four different cases are considered for error terms \( \epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{in_i})^T \) and covariates, respectively.

Case (1) We consider error terms \( \epsilon_i \) follow a multivariate normal distribution, \( N(0, \sigma^2 R_i(\alpha)) \).

Case (2) Error terms \( \epsilon_i \) follow a multivariate \( t \) distribution with two degrees of freedom, \( T_2(0, \sigma^2 R_i(\alpha)) \), which may contain some underlying outliers.

Case (3) To investigate the effect of outliers in the covariate direction, we randomly choose 3% of \( X_{ik} \) to be \( X_{ik} + 5 \) in Case (1).

Case (4) In Case (1), we randomly choose 3% of \( X_{ik} \) to be \( X_{ik} + 5 \). \( Y_{ik} \) are contaminated by adding an outlier equal to 6 or \(-6\) with a probability of 0.05.

For these four cases, we randomly generated \( n_i \) from integer values between 3 and 10 with an equal probability. We set \( p = 8 \) and \( \sigma^2 = 1, 9 \). For the correlation matrix \( R_i(\alpha) \), we used an exchangeable correlation structure with \( \alpha = 0.5 \) and 0.8. We conducted a simulation study with 100 independent realizations for all cases. The multivariate normal and the multivariate student’s \( t \) random numbers are generated using the \texttt{rmvtnorm} and \texttt{rmvt} functions in the \texttt{mvtnorm} library. We also carried out the simulation studies for \( p = N = 50 \) for each case. The results given as a supplementary material have the same pattern as those for \( p = 8 \). Therefore, we only summarized the results for \( p = 8 \) in Tables 1, 2, 3, 4.

We compare the proposed method with the oracle procedures, which set the zero coefficients to zero and estimated the nonzero coefficients by excluding the covariates of zero coefficients through the GEE method with the true correlation structure (denoted by gee.orcal). In Tables 1, 2, 3, 4, IND denotes the penalized objective function \( Q_W(\beta) \) with weights \( b_{ikjl} = 1 \) and \( w_i = 1 \), which is corresponding to the method of Fu and Wang (2018); WIL denotes \( Q_W(\beta) \) with weights \( b_{ikjl} = 1 \) and \( w_i = [1 + (1 + n_i - 1)\hat{\rho}]^{-1} \). GR and HBR correspond to \( Q_W(\beta) \) with GR and HBR weights (for \( b_{ikjl} \) and \( w_i \)). We evaluate the performance of the proposed method in terms of model errors proposed by Fan and Li (2001). We report biases, the relative efficiencies (Eff) of WIL, GR, HBR and gee.orcal to IND for the first three parameters, the average number of the \( p - 3 \) true zeroes coefficients that are properly estimated to be zero (CN), and the average number of three nonzero coefficients improperly estimated to be zero (IC). We also present the mean of the model errors (MME) and the percentile that correctly identified the true models (CP).

From Table 1 (multivariate normal distribution), we can see that the results of WIL, GR, and HBR are similar. The CNs of the weighted methods, WIL, GR, and HBR, are close to five. All the methods obtain unbiased estimates. Moreover, IC increases as \( \sigma^2 \) increases. When \( \sigma^2 = 1 \), WIL, GR, and HBR perform better than IND in terms of MME, CP, and CN. The efficiency of WIL, GR, and HBR is much higher than that of IND for the subject-level covariates but is slightly lower than that of IND for the within-subject covariate. The GEE estimates are the most efficient. When \( \sigma^2 = 9 \), CPs and CNs of IND are close to those of WIL, GR, and HBR, but ICs and MME of IND are larger than those of WIL, GR, and HBR.

\( \odot \) Springer
Table 1 Bias, the relative efficiencies (Eff) of WIL, GR, HBR and gee.orcal to IND for the first three parameters, mean of relative model errors (MME), the average number of three nonzero coefficients improperly estimated to be zero (IC), the average number of the true zero coefficients (CN), and the percentile of identifying the true model (CP) are presented for Case 1 when $\beta_{10} = (3.0, 1.5, 2.0)$

| Case 1: $\epsilon_i \sim N(0, \sigma^2 R(\alpha))$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | Eff | $\beta_1$ | $\beta_2$ | $\beta_3$ | MME | IC | CN | CP |
|---|---|---|---|---|---|---|---|---|---|---|---|
| IND $\sigma^2 = 1 \alpha = 0.5$ | 0.003 | 0.003 | 0.003 | 1.000 | 1.000 | 1.000 | 0.022 | 0 | 4.366 | 61 |
| WIL | 0.002 | 0.002 | 0.002 | 1.121 | 1.067 | 0.939 | 0.018 | 0 | 4.764 | 85 |
| GR | 0.003 | 0.002 | 0.002 | 1.113 | 1.055 | 0.928 | 0.019 | 0 | 4.792 | 86 |
| HBR | 0.002 | 0.002 | 0.002 | 1.120 | 1.067 | 0.939 | 0.018 | 0 | 4.766 | 85 |
| gee.orcal | 0.001 | 0.001 | 0.001 | 1.157 | 1.083 | 1.864 | 0.017 | 0 | 100 |
| IND $\sigma^2 = 1 \alpha = 0.8$ | 0.003 | 0.003 | 0.003 | 1.000 | 1.000 | 1.000 | 0.034 | 0 | 4.486 | 65 |
| WIL | 0.001 | 0.002 | 0.002 | 1.086 | 1.110 | 0.910 | 0.032 | 0 | 4.916 | 93 |
| GR | 0.001 | 0.002 | 0.002 | 1.085 | 1.097 | 0.904 | 0.032 | 0 | 4.912 | 92 |
| HBR | 0.001 | 0.002 | 0.002 | 1.086 | 1.109 | 0.909 | 0.032 | 0 | 4.914 | 93 |
| gee.orcal | 0.003 | 0.003 | 0.003 | 1.099 | 1.121 | 4.158 | 0.026 | 0 | 100 |
| IND $\sigma^2 = 9 \alpha = 0.5$ | 0.003 | 0.003 | 0.003 | 1.000 | 1.000 | 1.000 | 0.185 | 0.008 | 4.702 | 84 |
| WIL | 0.009 | 0.013 | 0.007 | 1.090 | 1.094 | 0.904 | 0.171 | 0.006 | 4.864 | 88 |
| GR | 0.008 | 0.011 | 0.003 | 1.084 | 1.089 | 0.896 | 0.172 | 0.006 | 4.854 | 88 |
| HBR | 0.006 | 0.013 | 0.007 | 1.090 | 1.094 | 0.904 | 0.171 | 0.006 | 4.862 | 88 |
| gee.orcal | 0.001 | 0.011 | 0.008 | 1.125 | 1.151 | 1.815 | 0.156 | 0 | 100 |
Table 1 continued

| Case 1: $\epsilon_i \sim N(0, \sigma^2 R(\alpha))$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | MME | IC | CN | CP |
|------------------------------------------------|----------|----------|----------|----------|----------|----------|-----|----|----|----|
| Bias | Eff | MME | IC | CN | CP |
| $\sigma^2 = 9 \alpha = 0.8$ | | | | | | | | | | | |
| IND | 0.002 | −0.009 | −0.020 | 1.000 | 1.000 | 1.000 | 0.235 | 0.024 | 4.730 | 84 |
| WIL | −0.005 | 0.001 | −0.011 | 1.133 | 1.284 | 0.935 | 0.198 | 0.010 | 4.864 | 88 |
| GR | −0.006 | 0.000 | −0.012 | 1.103 | 1.265 | 0.928 | 0.194 | 0.010 | 4.894 | 90 |
| HBR | −0.005 | 0.001 | −0.011 | 1.133 | 1.283 | 0.934 | 0.198 | 0.010 | 4.864 | 88 |
| gee.orcal | −0.004 | 0.005 | −0.007 | 1.195 | 1.443 | 3.912 | 0.188 | 0 | 5 | 100 |
Table 2  Bias, the relative efficiencies (Eff) of WIL, GR, HBR and gee.orcal to IND for the first three parameters, mean of relative model errors (MME), the average number of three nonzero coefficients improperly estimated to be zero (IC), the average number of the true zero coefficients (CN), and the percentile of identifying the true model (CP) are presented for Case 2 when $\beta_{10} = (3.0, 1.5, 2.0)$

| $\sigma^2 = 1$ $\alpha = 0.5$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | MME | IC | CN | CP |
|-------------------------------|---------|---------|---------|---------|---------|---------|------|----|----|----|
| IND                           | -0.000 | 0.007   | 0.007   | 1.000   | 1.000   | 1.000   | 0.044 | 0  | 1.000 | 4.530 | 72  |
| WIL                           | -0.000 | 0.007   | 0.006   | 1.082   | 1.020   | 0.931   | 0.038 | 0  | 0.926 | 4.830 | 88  |
| GR                            | -0.001 | 0.007   | 0.008   | 1.072   | 1.011   | 0.926   | 0.038 | 0  | 0.926 | 4.842 | 88  |
| HBR                           | -0.000 | 0.008   | 0.007   | 1.103   | 1.043   | 0.946   | 0.039 | 0  | 0.946 | 4.838 | 88  |
| gee.orcal                     | 0.000  | 0.000   | 0.015   | 0.197   | 0.178   | 0.410   | 0.077 | 0  | 0.410 | 5    | 100 |
| $\sigma^2 = 1$ $\alpha = 0.8$| $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | MME | IC | CN | CP |
| IND                           | -0.010 | -0.018  | 0.015   | 1.000   | 1.000   | 1.000   | 0.057 | 0  | 1.000 | 4.602 | 75  |
| WIL                           | -0.012 | -0.013  | 0.016   | 1.087   | 1.190   | 0.887   | 0.050 | 0  | 0.887 | 4.874 | 90  |
| GR                            | -0.012 | -0.013  | 0.016   | 1.076   | 1.196   | 0.896   | 0.048 | 0  | 0.896 | 4.876 | 90  |
| HBR                           | -0.012 | -0.013  | 0.016   | 1.109   | 1.222   | 0.906   | 0.049 | 0  | 0.906 | 4.866 | 89  |
| gee.orcal                     | -0.018 | -0.011  | 0.012   | 0.233   | 0.314   | 0.868   | 0.136 | 0  | 0.868 | 5    | 100 |
| $\sigma^2 = 9$ $\alpha = 0.5$| $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | MME | IC | CN | CP |
| IND                           | -0.008 | -0.151  | -0.233  | 1.000   | 1.000   | 1.000   | 0.544 | 0.300 | 0.300 | 4.820 | 67  |
| WIL                           | -0.014 | -0.152  | -0.236  | 1.020   | 1.028   | 0.964   | 0.510 | 0.306 | 0.306 | 4.896 | 68  |
| GR                            | -0.018 | -0.162  | -0.254  | 1.020   | 0.991   | 0.903   | 0.623 | 0.326 | 0.326 | 4.892 | 67  |
| HBR                           | -0.014 | -0.121  | -0.172  | 1.140   | 1.166   | 1.199   | 0.480 | 0.244 | 0.244 | 4.894 | 72  |
| gee.orcal                     | 0.045  | -0.063  | -0.016  | 0.237   | 0.303   | 1.075   | 0.993 | 0  | 0.993 | 5    | 100 |
Table 2 continued

Case 2: $\epsilon_i \sim T_2(0, \sigma^2 R(\alpha))$

| Case          | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | MME | IC | CN | CP |
|---------------|------------|------------|------------|------------|------------|------------|-----|----|----|----|
| \(\sigma^2 = 9\) \(\alpha = 0.8\) |            |            |            |            |            |            |     |    |    |    |
| IND           | 0.007      | -0.174     | -0.200     | 1.000      | 1.000      | 1.000      | 0.742 | 0.360 | 4.802 | 63 |
| WIL           | 0.010      | -0.194     | -0.214     | 1.092      | 1.009      | 0.940      | 0.679 | 0.380 | 4.902 | 62 |
| GR            | 0.010      | -0.193     | -0.205     | 1.126      | 0.999      | 0.957      | 0.647 | 0.376 | 4.894 | 61 |
| HBR           | 0.008      | -0.173     | -0.171     | 1.111      | 1.088      | 1.080      | 0.660 | 0.334 | 4.904 | 65 |
| gee.orcal     | 0.055      | 0.015      | 0.023      | 0.278      | 0.477      | 3.818      | 0.783 | 0   | 5   | 100 |
Table 3  Bias, the relative efficiencies (Eff) of WIL, GR, HBR and gee.orcal to IND for the first three parameters, mean of relative model errors (MME), the average number of three nonzero coefficients improperly estimated to be zero (IC), the average number of the true zero coefficients (CN), and the percentile of identifying the true model (CP) are presented for Case 3 when $\beta_{10} = (3.0, 1.5, 2.0)$.

| Case 3: $\epsilon_i \sim N(0, \sigma^2 R(\alpha))$ $X_{ik}$ are contaminated by outliers | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\bar{\beta}_1$ | $\bar{\beta}_2$ | $\bar{\beta}_3$ | MME | IC | CN | CP |
|---|---|---|---|---|---|---|---|---|---|---|
| IND | $\sigma^2 = 1$ $\alpha = 0.5$ | -0.445 | -0.417 | -1.032 | 1.000 | 1.000 | 1.000 | 2.071 | 0.010 | 4.464 | 65 |
| WIL | -0.444 | -0.416 | -1.032 | 1.005 | 1.002 | 0.992 | 1.925 | 0.014 | 4.942 | 94 |
| GR | -0.150 | -0.150 | -0.172 | 7.074 | 5.345 | 25.455 | 0.138 | 0.000 | 4.890 | 93 |
| HBR | -0.190 | -0.262 | -0.434 | 4.610 | 2.241 | 4.919 | 0.458 | 0.000 | 4.914 | 95 |
| gee.orcal | -2.884 | -1.449 | -1.345 | 0.032 | 0.103 | 0.644 | 20.396 | 0 | 5 | 100 |
| IND | $\sigma^2 = 1$ $\alpha = 0.8$ | -0.441 | -0.386 | -0.992 | 1.000 | 1.000 | 1.000 | 1.787 | 0.002 | 4.418 | 65 |
| WIL | -0.441 | -0.383 | -0.992 | 1.005 | 1.017 | 0.993 | 1.829 | 0.004 | 4.946 | 96 |
| GR | -0.157 | -0.138 | -0.172 | 5.529 | 4.957 | 22.965 | 0.145 | 0.000 | 4.916 | 94 |
| HBR | -0.194 | -0.236 | -0.416 | 3.942 | 2.292 | 4.884 | 0.407 | 0.000 | 4.930 | 95 |
| gee.orcal | -2.910 | -1.463 | -1.337 | 0.031 | 0.092 | 0.603 | 19.627 | 0 | 5 | 100 |
| IND | $\sigma^2 = 9$ $\alpha = 0.5$ | -0.921 | -0.684 | -1.577 | 1.000 | 1.000 | 1.000 | 5.785 | 0.632 | 4.812 | 37 |
| WIL | -0.923 | -0.667 | -1.586 | 0.998 | 1.054 | 0.985 | 5.946 | 0.660 | 4.982 | 41 |
| GR | -0.429 | -0.386 | -0.459 | 3.120 | 2.249 | 7.643 | 1.230 | 0.056 | 4.906 | 86 |
| HBR | -0.767 | -0.634 | -1.419 | 1.319 | 1.145 | 1.176 | 4.819 | 0.514 | 4.974 | 53 |
| gee.orcal | -2.722 | -1.358 | -1.354 | 0.144 | 0.337 | 1.438 | 18.078 | 0 | 5 | 100 |
Table 3 continued

Case 3: $\epsilon_i \sim N(0, \sigma^2 R(\alpha))$, $X_{ik}$ are contaminated by outliers

|       | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_1$ | $\beta_2$ | $\beta_3$ |
|-------|-----------|-----------|-----------|-----------|-----------|-----------|
|       | Bias      | Eff       | Eff       | MME       | IC        | CN        | CP        |
|       |           |           |           |           |           |           |           |
| $\sigma^2 = 9$, $\alpha = 0.8$ |
| IND   | -0.939    | -0.672    | -1.543    | 1.000     | 1.000     | 1.000     | 6.167     | 0.610     | 4.824     | 41        |
| WIL   | -0.929    | -0.646    | -1.565    | 1.032     | 1.079     | 0.970     | 5.865     | 0.636     | 4.980     | 43        |
| GR    | -0.437    | -0.374    | -0.488    | 3.007     | 2.027     | 6.582     | 1.386     | 0.074     | 4.930     | 87        |
| HBR   | -0.772    | -0.615    | -1.402    | 1.340     | 1.154     | 1.157     | 5.033     | 0.498     | 4.968     | 53        |
| gee.orcal | -2.856   | -1.423    | -1.352    | 0.140     | 0.319     | 1.397     | 19.048    | 0         | 5         | 100       |
Table 4  Bias, the relative efficiencies (Eff) of WIL, GR, HBR and gee.orcal to IND for the first three parameters, mean of relative model errors (MME), the average number of three nonzero coefficients improperly estimated to be zero (IC), the average number of the true zero coefficients (CN), and the percentile of identifying the true model (CP) are presented for Case 4 when $\beta_{10} = (3.0, 1.5, 2.0)$

| Case 4: $\epsilon_i \sim N(0, \sigma^2 R(\alpha))$ | $Y_{ik}$ and $X_{ik}$ are contaminated by outliers |
|---------------------------------|---------------------------------|---------------------------------|---------------------------------|
| $\sigma^2 = 1 \alpha = 0.5$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | Eff | MME | IC | CN | CP |
| IND | -0.399 | -0.341 | -0.971 | 1.000 | 1.000 | 1.000 | 1.663 | 0.008 | 4.436 | 66 |
| WIL | -0.400 | -0.335 | -0.966 | 0.991 | 1.026 | 0.995 | 1.618 | 0.016 | 4.914 | 93 |
| GR | -0.100 | -0.108 | -0.138 | 7.723 | 5.189 | 26.782 | 0.107 | 0.000 | 4.886 | 93 |
| HBR | -0.152 | -0.221 | -0.436 | 4.612 | 2.026 | 4.225 | 0.400 | 0.000 | 4.922 | 95 |
| gee.orcal | -2.761 | -1.372 | -1.297 | 0.032 | 0.086 | 0.616 | 16.802 | 0 | 5 | 100 |

| $\sigma^2 = 1 \alpha = 0.8$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | Eff | MME | IC | CN | CP |
| IND | -0.381 | -0.366 | -0.967 | 1.000 | 1.000 | 1.000 | 1.442 | 0.020 | 4.460 | 62 |
| WIL | -0.381 | -0.367 | -0.967 | 0.996 | 0.992 | 0.998 | 1.572 | 0.024 | 4.952 | 94 |
| GR | -0.086 | -0.124 | -0.135 | 6.573 | 4.815 | 29.625 | 0.106 | 0.000 | 4.906 | 92 |
| HBR | -0.138 | -0.240 | -0.428 | 4.338 | 1.989 | 4.458 | 0.339 | 0.000 | 4.942 | 95 |
| gee.orcal | -2.795 | -1.403 | -1.297 | 0.029 | 0.098 | 0.610 | 17.761 | 0 | 5 | 100 |

| $\sigma^2 = 9 \alpha = 0.5$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | Eff | MME | IC | CN | CP |
| IND | -0.850 | -0.633 | -1.538 | 1.000 | 1.000 | 1.000 | 5.315 | 0.670 | 4.884 | 37 |
| WIL | -0.847 | -0.620 | -1.553 | 1.021 | 1.041 | 0.977 | 5.548 | 0.700 | 4.994 | 41 |
| GR | -0.325 | -0.289 | -0.417 | 3.613 | 2.564 | 6.332 | 0.984 | 0.082 | 4.914 | 85 |
| HBR | -0.689 | -0.579 | -1.406 | 1.359 | 1.162 | 1.128 | 4.703 | 0.574 | 4.992 | 50 |
| gee.orcal | -2.633 | -1.306 | -1.297 | 0.137 | 0.348 | 1.509 | 17.055 | 0 | 5 | 100 |
**Table 4 continued**

Case 4: $\epsilon_i \sim N(0, \sigma^2 R(\alpha))$, $Y_{ik}$ and $X_{ik}$ are contaminated by outliers

|                     | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | MME | IC  | CN  | CP  |
|---------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----|-----|-----|-----|
| Bias                | Eff       | Eff       | Eff       | Eff       | Eff       | Eff       |     |     |     |     |
| $\sigma^2 = 9, \alpha = 0.8$ |           |           |           |           |           |           |     |     |     |     |
| IND                 | -0.819    | -0.666    | -1.546    | 1.000     | 1.000     | 1.000     | 5.275| 0.708| 4.812| 32  |
| WIL                 | -0.820    | -0.656    | -1.552    | 1.011     | 1.038     | 0.986     | 5.312| 0.730| 4.974| 36  |
| GR                  | -0.298    | -0.328    | -0.409    | 3.113     | 2.939     | 7.103     | 0.872| 0.106| 4.946| 85  |
| HBR                 | -0.652    | -0.615    | -1.396    | 1.331     | 1.125     | 1.156     | 4.663| 0.596| 4.972| 47  |
| gee.orcal           | -2.800    | -1.409    | -1.315    | 0.119     | 0.340     | 1.492     | 18.766| 0   | 5   | 100 |
Table 5  The upper panel is penalized robust estimates with four different weights: IND, WIL, GR, and HBR. The lower panel is the selected frequencies for the 31 variables which consist of the real 11 variables and other 20 irrelevant variables based on 500 replications.

|               | IND  | WIL  | GR   | HBR  |
|---------------|------|------|------|------|
| Intercept     | 0.9894 | 1.0119 | 1.0395 | 1.0408 |
| Age           | 0.1026 | 0.0000 | 0.0000 | 0.0000 |
| BMI           | -0.0824 | 0.0000 | 0.0000 | 0.0000 |
| Time          | -0.9097 | -0.8169 | 0.5676 | -0.9258 |
| Age*BMI       | -0.0182 | 0.0000 | 0.0000 | 0.0000 |
| Age*Time      | -0.0512 | 0.0000 | 0.0000 | 0.0000 |
| BMI*Time      | 0.0095 | -0.2094 | 0.0000 | 0.0000 |
| Time²         | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| Time³         | 4.3983 | 4.6429 | 2.8507 | 4.5700 |
| Time⁴         | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| Time⁵         | -3.0093 | -3.2146 | -2.5363 | -3.1894 |

Containing 20 irrelevant covariates

|               | 1     | 1     | 1     | 1     |
|---------------|-------|-------|-------|-------|
| Intercept     | 1.000 | 1.000 | 0.836 | 0.962 |
| Age           | 0.466 | 0.000 | 0.000 | 0.006 |
| BMI           | 0.232 | 0.000 | 0.002 | 0.000 |
| Time          | 0.334 | 0.942 | 0.732 | 0.752 |
| Age*BMI       | 0.042 | 0.254 | 0.598 | 0.240 |
| Age*Time      | 0.650 | 0.130 | 0.532 | 0.284 |
| BMI*Time      | 0.886 | 0.478 | 0.224 | 0.334 |
| Time²         | 0.004 | 0.000 | 0.132 | 0.092 |
| Time³         | 1.000 | 1.000 | 0.836 | 0.962 |
| Time⁴         | 0.008 | 0.000 | 0.228 | 0.092 |
| Time⁵         | 0.976 | 1.000 | 0.690 | 0.784 |
| Irrelevant x₁ | 0.190 | 0.000 | 0.002 | 0.002 |
| Irrelevant x₂ | 0.178 | 0.000 | 0.000 | 0.000 |
| Irrelevant x₃ | 0.188 | 0.000 | 0.002 | 0.000 |
| Irrelevant x₄ | 0.190 | 0.002 | 0.002 | 0.004 |
| Irrelevant x₅ | 0.168 | 0.000 | 0.000 | 0.000 |
| Irrelevant x₆ | 0.164 | 0.000 | 0.000 | 0.004 |
| Irrelevant x₇ | 0.196 | 0.000 | 0.000 | 0.002 |
| Irrelevant x₈ | 0.188 | 0.000 | 0.000 | 0.002 |
| Irrelevant x₉ | 0.156 | 0.000 | 0.000 | 0.000 |
| Irrelevant x₁₀| 0.194 | 0.002 | 0.000 | 0.002 |
| Irrelevant x₁₁| 0.188 | 0.000 | 0.000 | 0.000 |
| Irrelevant x₁₂| 0.160 | 0.002 | 0.002 | 0.002 |
| Irrelevant x₁₃| 0.170 | 0.000 | 0.006 | 0.006 |
| Irrelevant x₁₄| 0.184 | 0.000 | 0.004 | 0.004 |
| Irrelevant x₁₅| 0.182 | 0.002 | 0.004 | 0.006 |
| Irrelevant x₁₆| 0.174 | 0.000 | 0.000 | 0.002 |
Table 5  continued

|       | IND | WIL | GR  | HBR |
|-------|-----|-----|-----|-----|
| Irrelevant $x_{17}$ | 0.192 | 0.002 | 0.002 | 0.000 |
| Irrelevant $x_{18}$ | 0.188 | 0.000 | 0.004 | 0.000 |
| Irrelevant $x_{19}$ | 0.184 | 0.000 | 0.002 | 0.004 |
| Irrelevant $x_{20}$ | 0.184 | 0.002 | 0.006 | 0.002 |

Fig. 1  Boxplots of the log-transformed progesterone for 34 women

Fig. 2  The left panel is predictions for log progesterone based on four different weights for 10 covariates. The right panel is predictions for log progesterone after adding outliers to age.
When the joint distribution of error terms are heavy-tails (Table 2), the rank-based methods IND, WIL, GR, HBR have the same pattern as those in Case 1 (Table 1) for $\sigma^2 = 1$, but they are much better than the GEE method in terms of efficiency and model error. When $\sigma^2 = 9$, HBR outperforms in terms of CN, CP, IC, and efficiency. When covariates are contaminated by outliers (Table 3), the estimates obtained from all the methods are biased, and the GEE method has much larger model errors. When $\sigma^2 = 1$, GR and HBR have much higher efficiency and much smaller model errors than IND, WIL, and gee.orcal. When $\sigma^2 = 9$, GR has significantly higher CP and efficiency, and has much smaller IC and model errors than all of others.

### 4 Hormone study

In this section, we illustrate the proposed methods by analyzing the longitudinal progesterone data (Sowers et al. 1998), which has been analyzed in some literature (Fan et al. 2012; Fung et al. 2002; Zhang et al. 1998). In this study, a total of 492 urine samples were collected from 34 women (with menstrual cycles) aged between 27 and 45 years, and urinary progesterone was assayed on alternate days. Each woman contributed between 11 and 28 observations over a period of time; hence the data is unbalanced. One purpose of the study was to test the effects of age and body mass index (BMI) on women’s progesterone levels after an appropriate adjustment of their menstrual cycles.

Let $Y$ be the log-transformed progesterone level. Figure 1 indicates that some outliers exist, which coincides with the findings of Fung et al. (2002). We check the data and find that one woman’s BMI (her ID number is 21,208, and has 20 observations) exceeds 38 and 11.8% of women had BMI over 30. In our model, covariates include age, BMI, and time effects. We also considered their interaction effects. The model is given as follows:

$$Y_{ik} = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{BMI}_i + \beta_3 \text{Time}_{ik} + \beta_4 \text{Age}_i \times \text{BMI}_i + \beta_5 \text{Age}_i \times \text{Time}_{ik} + \beta_6 \text{BMI}_i \times \text{Time}_{ik} + \beta_7 \text{Time}^2_{ik} + \beta_8 \text{Time}^3_{ik} + \beta_9 \text{Time}^4_{ik} + \beta_{10} \text{Time}^5_{ik} + \epsilon_{ik}.$$  

In the model, $\text{Age}_i$, $\text{BMI}_i$, and $\text{Age}_i \times \text{BMI}_i$ are subject-level covariates. The terms, including time effect, are within-cluster covariates. All of the covariates are standardized in the computation. The results presented on the upper panel in Table 5 indicate that WIL, GR, and HBR select intercepts, Time, $\text{Time}^3$, and $\text{Time}^5$, which is consistent with the findings by Zhang et al. (1998) and Fung et al. (2002), where age and BMI were found to have no significant effect on progesterone level. IND selected eight other covariates, including age and BMI, except $\text{Time}^2$ and $\text{Time}^4$. The left panel in Fig. 2 indicates that the predictions of IND, WIL, and HBR is better than that of GR. The right panel in Fig. 2 indicates that the predictions of WIL and HBR are good, and IND and GR performed poorly after adding some outliers to age.

To further demonstrate the performance of the methods based on different weights, we also consider the regression that contains 20 additional irrelevant variables randomly generated from the standard normal distribution. The lower panel in Table 5 presents the selected frequencies for the 31 variables which consist of the real 11
variables and other 20 irrelevant variables based on 500 replications. As we can see, intercept, Time, Time^3 and Time^5 have very high selected frequencies for all the methods after adding the 20 irrelevant variables, which indicates that these variables may be very significant in this regression model. In addition, IND has higher frequencies of choosing the 20 irrelevant variables than GR, WIL and HBR methods, which indicates that IND is clearly affected by the adding irrelevant variables and tends to overfitting. According to the analysis made above, we prefer WIL and HBR methods to determine this hormone data. The dataset for this hormone study is available from the corresponding author on reasonable request.

5 Conclusion

In this paper, we have provided efficient and doubly robust methods for variable selection and parameter estimation in longitudinal data analysis. The objective functions are based on the ranks of the pairwise residuals, and weight \( w_{ij} \) captures the effects of the within-subject correlations and varying cluster sizes; hence, the proposed methods are efficient and robust when responses deviate from the Gaussian distribution or contain underlying outliers, or a strong within-correlation exists. Moreover, the GR and HBR weights automatically downweight the outliers existing in the covariates. Therefore, the proposed methods are doubly robust. Furthermore, the calculation of the proposed methods can be easily implemented in the statistical software R. According to the simulation results and the real data analysis for weight \( b_{ikj} \), we propose using the HBR weight when response distribution is heavy-tails or contaminated by outliers and using the GR weight when covariates contain outliers. If there is no evidence that outliers exist in response or covariates, the the Wilcoxon weight is preferable. The GR weight may depend on the selection of the tuning parameters \( c \) and \( \kappa \), and the cross-validation method can be utilized to choose them, but this remains unexplored in our paper. It is worth noting that the performance of the proposed methods depends on the covariate type. The weight \( w_{ij} \) only improves the efficiency of parameter estimates in cluster-level covariates. We will seek a more efficient method for within-cluster covariates in future work.

Acknowledgements The authors thank the Associate Editor and referees for their constructive comments. We would like to express gratitude to Professor Xihong Lin for providing the hormone data. This research was supported by the National Science Foundation of China (No. 11871390), the Fundamental Research Funds for the Central Universities (No. xjj2017180), and the Natural Science Foundation of Shaanxi Province (No. 2018JQ1006), Zhejiang Science Grant (Project KZS1905002) and the Australian Research Council Discovery Project (DP160104292).

Compliance with ethical standard

Conflict of interest The authors declare that they have no conflict of interest.

Appendix

C1 Cluster size \( n_i \) is bound.
C2 $\epsilon_{ik}$ is continuous, and the median of any pairwise difference $\epsilon_{ik} - \epsilon_{kl}$ is zero.
C3 Density function $f_{ik}$ is absolutely continuous.
C4 Matrix $D$ is a positive definite matrix.
C5 $\liminf_{N \to +\infty} \liminf_{\beta \to 0^+} P'_\lambda(\beta)/\lambda > 0$, and $\max_{1 \leq s \leq p} \{ P''_{\lambda}(|\beta_s|) : \beta_s \neq 0 \} \to 0$.

Let $f_{ik}$ and $F_{ik}$ be the density and cumulative distribution functions of $\epsilon_{ik}$, respectively. Assume that $b_{ij}$ and $w_{ij}$ are given. Define

$$D = M^{-2} \sum_{i=1}^{N} \sum_{j<i}^{N} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} b_{ikjl} w_{ij}(X_{ik} - X_{jl})(X_{ik} - X_{jl})^T \int f_{ik} dF_{ij},$$

and

$$U(\beta) = M^{-2} \sum_{i=1}^{N} \sum_{j<i}^{N} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} b_{ikjl} w_{ij}(X_{ik} - X_{jl}) \text{sign}(\epsilon_{ik} - \epsilon_{jl}).$$

Before proving the theorem, we give the following lemmas.

**Lemma 1** Under conditions C1–C3, $\sqrt{M} U_1(\beta_T)$ converges in distribution to $N(0, V)$, where $V = \lim_{N \to +\infty} M^{-3} \sum_{i=1}^{N} \zeta_i^T \zeta_i$, and $\zeta_i = \sum_{j<i} \sum_{k,l} b_{ikjl} w_{ij}(X_{ik} - X_{jl})(F_{jl}(\epsilon_{ik}) - 1/2)$.

**Lemma 2** Under conditions C1–C4, $U(\beta)$ is asymptotic linearity, that is,

$$\sup_{||b - \beta|| \leq M^{-1/2} \eta} \left\| \sqrt{M} \{ U(b) - U(\beta) \} - M^{1/2} D(b - \beta) \right\| = o_p(1 + \sqrt{M} ||b - \beta||).$$

**Proof of Lemmas 1 and 2** can refer to Jung and Ying (2003) and Wang and Zhao (2008).

**Lemma 3** Under conditions C1–C4, if $\lambda \to 0$, then the estimator $\hat{\beta}$ obtained from $Q_W(\beta)$ satisfies $||\hat{\beta} - \beta_T|| = O_p(M^{-1/2})$, where $\beta_T$ is the true value of $\beta$.

**Proof** We will prove that, for $\forall \epsilon > 0$, there exists a large constant $C$ that satisfies

$$P \left( \inf_{||u|| = C} Q_W(\beta_T + M^{-1/2} u) > Q_W(\beta_T) \right) \geq 1 - \epsilon, \quad (4)$$

where $u = (u_1, \ldots, u_p)^T$. Because $Q_W(\beta)$ is convex in $\beta$, the estimator $\hat{\beta}$ lies in the ball $\{ \beta_T + N^{-1/2} u : ||u|| \leq C \}$. According to Sievers (1983) and Lemma 2,

$$L_W(\beta_T + M^{-1/2} u) - L_W(\beta_T) = -u^T M^{-1/2} U(\beta_T) + \frac{1}{2} M^{-1} u^T D(\beta_T) u + o_p(1).$$

Springer
Therefore,

\[
Q_W (\beta_T + M^{-1/2} u) - Q_W (\beta_T) = L_W (\beta_T + N^{-1/2} u) - L_W (\beta_T) \\
+ \sum_{s=1}^{p} P'_x (|\beta^0_s|) \{|\beta_s + N^{-1/2} u_s | - |\beta_s|\} \\
\geq -u^T M^{-1/2} U (\beta_T) + \frac{1}{2} M^{-1/2} u^T D (\beta_T) u \\
-M^{-1/2} \sum_{s=1}^{p} P'_x (|\beta^0_s|) |u_s| + o_p (1).
\]

According to Lemma 1, \( M^{-1/2} U (\beta_T) = O_p (M^{-1/2}) \). Because \( D (\beta_T) \) is a positive definite matrix (C4), \( u^T D (\beta_T) u > 0 \). Furthermore, \( P'_x (|\beta^0_s|) \rightarrow P'_x (|\beta_s|) \) in probability, and \( P'_x (|\beta_s|) = P'_x (|\beta_s|) \mathbb{I} (|\beta_s| \leq a \lambda) \), hence for \( \varepsilon > 0 \), \( P (P'_x (|\beta_s|) > \varepsilon) \leq P (|\beta_s| \leq a \lambda) \rightarrow 0 \) as \( \lambda \rightarrow 0 \). Therefore, by choosing a sufficiently \( C \), the sign of \( Q_W (\beta_T + M^{-1/2} u) - Q_W (\beta_T) \) is dominated by the second term on the right-hand side, and (4) holds.

**Lemma 4** If \( \lambda \rightarrow 0 \), and \( \sqrt{N \lambda} \rightarrow +\infty \) as \( N \rightarrow +\infty \), for any \( \beta_d \) satisfying \( ||\beta_{10} - \beta_{10}^*|| = o_p (M^{-1/2}) \) and any constant \( C \),

\[
Q_W \left( \begin{array}{c} \beta_{10} \\ 0 \end{array} \right) = \min_{||\beta_{20}|| \leq CM^{-1/2}} Q_W \left( \begin{array}{c} \beta_{10} \\ \beta_{20} \end{array} \right).
\]

**Proof of Lemma 4** Because \( Q_W (\beta) \) is a convex, piecewise linear of \( \beta \), it is sufficient to show that with probability tending to 1 as \( N \rightarrow \infty \), for any \( \beta_d \) satisfying \( ||\beta_{10} - \beta_{10}^*|| = O_p (M^{-1/2}) \) and for any small \( \epsilon_M = CM^{-1/2} \), and \( s = d + 1, \ldots, p \),

\[
\frac{\partial Q (\beta)}{\partial \beta_s} > 0 \quad \text{for} \quad 0 < \beta_s < \epsilon_M \\
\frac{\partial Q (\beta)}{\partial \beta_s} < 0 \quad \text{for} \quad -\epsilon_M < \beta_s < 0.
\]

Note that

\[
\frac{\partial Q_W (\beta)}{\partial \beta_s} = \frac{\partial L_W (\beta)}{\partial \beta_s} + P'_x (|\beta^0_s|) \text{sign} (\beta_s) \\
= -M^{-2} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} b_{ikjl} w_{ij} (X_{iks} - X_{jls}) \text{sign} (\epsilon_{ikjl}) \\
+ P'_x (|\beta^0_s|) \text{sign} (\beta_s) \\
= -U_s (\beta) + P'_x (|\beta^0_s|) \text{sign} (\beta_s)
\]

where \( U_s (\beta) \) is the \( s \)th element of \( U (\beta) \), and \( \epsilon_{ikjl} = \epsilon_{ik} - \epsilon_{jl} \). According to Lemam2, we have \( \sqrt{M} (U_s (\beta) - U_s (\beta_T)) = \sqrt{M} \sum_{l=1}^{d} D_{sl} (\beta_l - \beta^*_l) + o_p (1 + M^{1/2} |\beta_s - \beta^*_s|) \), where \( D_{sl} \) is the \((s, l)\)th element of \( D \), and \( \beta^*_s \) is the \(s\)th element of the true value.
of $\beta$. According to Lemma 1, $U_s(\beta_T) = O_p(M^{-1/2})$. Thus, for any $\beta_l$ satisfying $\|\beta_l - \beta^*_l\| = O_p(M^{-1/2})$, we have

$$\frac{\partial Q(\beta)}{\partial \beta_s} = P'_\beta(|\hat{\beta}_s^0|)\text{sign}(\beta_s) + O_p((d + 1)M^{-1/2}/\lambda) + o_p(M^{-1/2}/\lambda)$$

Because $\hat{\beta}_s^0$ is a consistent estimate of $\beta_s$, $P'_\beta(|\hat{\beta}_s^0|) = \lambda P'_\beta(|\beta_s|)/\lambda + o_p(1)$. Under condition C5, $\lim \inf_{N \to +\infty} \lim \inf_{\beta \to 0+} P'_\beta(\beta_s)/\lambda > 0$. Therefore, as $\lambda \to 0$ and $\sqrt{M} \lambda \to +\infty$, the sign of the derivative is completely determined by that of $\beta_j$. This completes the Proof of Lemma 4.

**Proof of Theorem** It follows by Lemma 4 that part (1) of the theorem holds. Now, we prove part (2) of the theorem. Similarly as in Lemmas 2 and 3, it can be shown that there exists a $\hat{\beta}_{10}$, a $\sqrt{M}$-consistent local minimizer of $Q(\hat{\beta}_{10})$, which satisfies the equations

$$\frac{\partial Q(\beta)}{\partial \beta_s} \bigg|_{\hat{\beta}_{10}} = 0 \quad \text{for} \ s = 1, \ldots, d.$$

Similar to the Proof of Lemma 4,

$$\frac{\partial Q(\beta)}{\partial \beta_s} = -U_s(\beta) + P'_\beta(|\hat{\beta}_s^0|)\text{sign}(\beta_s), \quad \text{for} \ s = 1, \ldots, d.$$

Hence, $\sqrt{M}U_s(\hat{\beta}) = \sqrt{M}P'_\beta(|\hat{\beta}_s^0|)\text{sign}(\hat{\beta}_s) = 0$, for $s = 1, \ldots, d$. According to Lemma 2, we have

$$\sqrt{M}U_s(\hat{\beta}) = -\sqrt{M}U_s(\beta) + \sqrt{M} \sum_{l=1}^d D_{sl}(\hat{\beta}_l - \beta_l^*) + o_p(1 + M^{1/2}|\hat{\beta}_s - \beta_s^*|)$$

$$= \sqrt{M}P'_\beta(|\beta_s|)\text{sign}(\beta_s) + \sqrt{M}P''_\beta(|\beta_s|)\text{sign}(\beta_s)(\hat{\beta}_s - \beta_s^*).$$

Therefore,

$$\sqrt{M}(\hat{\beta}_{10} - \beta_{10} - (D_{11} + \Sigma_{11})^{-1}P'_\beta(|\beta_{10}|)\text{sign}(\beta_{10})$$

$$= -(D_{11} + \Sigma_{11})^{-1}\sqrt{M}U_R(\beta) + o_p(1),$$

where $D_{11}$ and $U_R(\beta)$ correspond to the first $d \times d$ submatrix of $D$ and the first $d$ elements of $U(\beta)$ with $\beta^T = (\beta_{10}, 0_{p-d})$, and $P'_\beta(|\beta_{10}|)\text{sign}(\beta_{10}) = (P'_\beta(|\beta_1|)\text{sign}(\beta_1), \ldots, P'_\beta(|\beta_d|)\text{sign}(\beta_d))^T$. According to conditions C4 and Lemma 1,

$$\sqrt{M}(D_{11} + \Sigma_{11})(\hat{\beta}_{10} - \beta_{10} - (D_{11} + \Sigma_{11})^{-1}P'_\beta(|\beta_{10}|)\text{sign}(\beta_{10}) \to N(0, V_{11}),$$

where $V_{11}$ is the first $d \times d$ submatrix of $V$. \hfill \Box
References

Chang WH, McKean JW, Naranjo JD, Sheather SJ (1999) High-breakdown rank regression. J Am Stat Assoc 94:205–219
Cho H-K, Qu A (2013) Model selection for correlated data with diverging number of parameters. Stat Sinica 23:901–927
Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc 96:1348–1360
Fan J, Li R (2004) New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. J Am Stat Assoc 99:710–723
Fan Y, Qin G, Zhu Z (2012) Variable selection in robust regression models for longitudinal data. J Multivar Anal 109:156–167
Fu LY, Wang Y-G, Bai Z (2010) Rank regression for analysis of clustered data: A natural induced smoothing approach. Comput Stat Data Anal 54:1036–1050
Fu LY, Wang Y-G (2012) Efficient estimation for rank-based regression with clustered data. Biometrics 68:1074–1082
Fu LY, Wang Y-G (2018) Variable selection in rank regression for analyzing longitudinal data. Stat Methods Med Res 27(8):2447–2458
Fung K-W, Zhu ZY, Wei BC, He XM (2002) Inference diagnostics and outlier tests for semiparametric mixed models. J Royal Stat Soc Ser B 64:565–579
Guo CH, Yang H, Lv J (2014) Robust variable selection in semiparametric mean-covariance regression for longitudinal data analysis. Appl Math Comput 245:343–356
Jaeckel LA (1972) Estimating regression coefficients by minimizing the dispersion of the residuals. Ann Math Stat 43:1449–1458
Jung SH, Ying Z (2003) Rank-based regression with repeated measurement data. Biometrika 90:732–740
Koenker R (2005) Quantile Regression. Cambridge University Press, Cambridge
Liang KY, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrika 73:13–22
Lv J, Yang H, Guo CH (2015) An efficient and robust variable selection method for longitudinal generalized linear models. Comput Stat Data Anal 82:74–88
Naranjo J, McKean JW, Sheather SJ, Hettmansperger TP (1994) The use and interpretation of rank-based residuals. Nonparametr Stat 3:323–341
Ni X, Zhang D, Zhang HH (2010) Variable selection for semiparametric mixed models in longitudinal studies. Biometrics 66:79–88
Rousseeuw PJ, Zomeren BCV (1990) Unmasking multivariate outliers and leverage points. J Am Stat Assoc 85:633–639
Sievers GL (1983) A weighted dispersion function for estimation in linear models. Commun Stat Theory Methods 12:1161–1179
Sowers MF, Crutchfield M, Randolph JF, Shapiro B, Zhang B, Pietra ML, Schork MA (1998) Urinary ovarian and gonadotrophin hormone levels in premenopausal women with low bone mass. J Bone Mining Res 13:1191–1202
Terpstra JT, McKean JW (2005) Rank-based reanalysis of linear models using R. J Stat Softw 14:1–26
Wang L, Li R (2009) Weighted Wilcoxon-type smoothly clipped absolute deviation method. Biometrics 65:564–571
Wang L, Zhou J, Qu A (2012) Penalized generalized estimating equations for high-dimensional longitudinal data analysis. Biometrics 68:353–360
Wang XQ, Jiang YL, Huang M, Zhang HP (2013) Robust variable selection with exponential squared loss. J Am Stat Assoc 108:632–643
Wang Y-G, Carey V (2003) Working correlation structure misspecification, estimation and covariate design: Implications for generalised estimating equations performance. Biometrika 90:29–41
Wang Y-G, Zhao YD (2008) Weighted rank regression for clustered data analysis. Biometrics 64:39–45
Xu JF, Leng CL, Ying Z (2010) Rank-based variable selection with censored data. Stat Comput 20:165–176
Yang H, Guo CH, Lv J (2015) SCAD penalized rank regression with a diverging number of parameters. J Multivar Anal 133:321–333
Zhang D, Lin XH, Raz J, Sowers MF (1998) Semiparametric stochastic mixed models for longitudinal data. J Am Stat Assoc 93:710–719
Zou H, Li R (2008) One-step sparse estimates in noncave penalized likelihood models. Ann Stat 36:1509–1566

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Liya Fu¹ · Zhuoran Yang¹ · Fengjing Cai² · You-Gan Wang³

Fengjing Cai
cyclie@163.com

You-Gan Wang
you-gan.wang@qut.edu.au

Liya Fu
fuliya@mail.xjtu.edu.cn

Zhuoran Yang
yzr741@stu.xjtu.edu.cn

¹ School of Mathematics and Statistics, Xi’an Jiaotong University, Xi’an, People’s Republic of China
² College of Mathematics, Wenzhou University, Wenzhou, People’s Republic of China
³ School of Mathematical Science, Queensland University of Technology, Brisbane, Australia