



RESEARCH ARTICLE

Robust approach for variable selection with high dimensional longitudinal data analysis

Liya Fu¹  | Jiaqi Li¹ | You-Gan Wang² 

¹School of Mathematics and Statistics,
Xi'an Jiaotong University, Xi'an, China

²School of Mathematical Sciences,
Queensland University of Technology,
Brisbane, Queensland, Australia

Correspondence

Liya Fu, School of Mathematics and
Statistics, Xi'an Jiaotong University, Xi'an,
Shaanxi, China.

Email: fuliya@mail.xjtu.edu.cn

Funding information

Natural Science Basic Research Plan in
Shaanxi Province of China, Grant/Award
Number: No.2018JQ1006; Australian
Research Council Discovery Project,
Grant/Award Number: DP160104292;
Natural Science Foundation of China,
Grant/Award Number: No.11871390

Abstract

This article proposes a new robust smooth-threshold estimating equation to select important variables and automatically estimate parameters for high dimensional longitudinal data. A novel working correlation matrix is proposed to capture correlations within the same subject. The proposed procedure works well when the number of covariates p_n increases as the number of subjects n increases. The proposed estimates are competitive with the estimates obtained with the true correlation structure, especially when the data are contaminated. Moreover, the proposed method is robust against outliers in the response variables and/or covariates. Furthermore, the oracle properties for robust smooth-threshold estimating equations under “large n , diverging p_n ” are established under some regularity conditions. Extensive simulation studies and a yeast cell cycle data are used to evaluate the performance of the proposed method, and results show that the proposed method is competitive with existing robust variable selection procedures.

KEYWORDS

automatic variable selection, high dimensional covariates, outliers, robustness, Tukey's biweight method, working correlation structure

1 | INTRODUCTION

Longitudinal data is usually collected by repeatedly observing the results for each subject at several points in time. It has been widely used in medical and economic research over the past decade. High-dimensional longitudinal data consisting of repeated measurements with a large number of covariates has become increasingly common in practical application. The number of covariates can be quite large, especially when the interactions of various factors are considered. Nevertheless, there is only a subset of covariates related to the response variables, and the redundant variables can affect the accuracy and efficiency of estimation. Therefore, it is important to develop a new methodology to select the important variables in high-dimensional longitudinal data.

To select the important variables in longitudinal data analysis, Pan¹ proposed a quasi-likelihood information criterion (QIC) based on an independence assumption, which can be used to select variables and working correlation matrices. Wang and Qu² combined the Bayesian information criterion with quadratic inference function, which does not require the full likelihood or quasi-likelihood. However, these two methods can be computationally intensive when the dimension of covariates is large. Tian et al³ extended the SCAD-penalized quadratic inference function to analyze semiparametric varying coefficient partially linear models, and simultaneously select significant variables in the parametric components and the nonparametric components. Li et al⁴ proposed an automatic variable selection procedure using

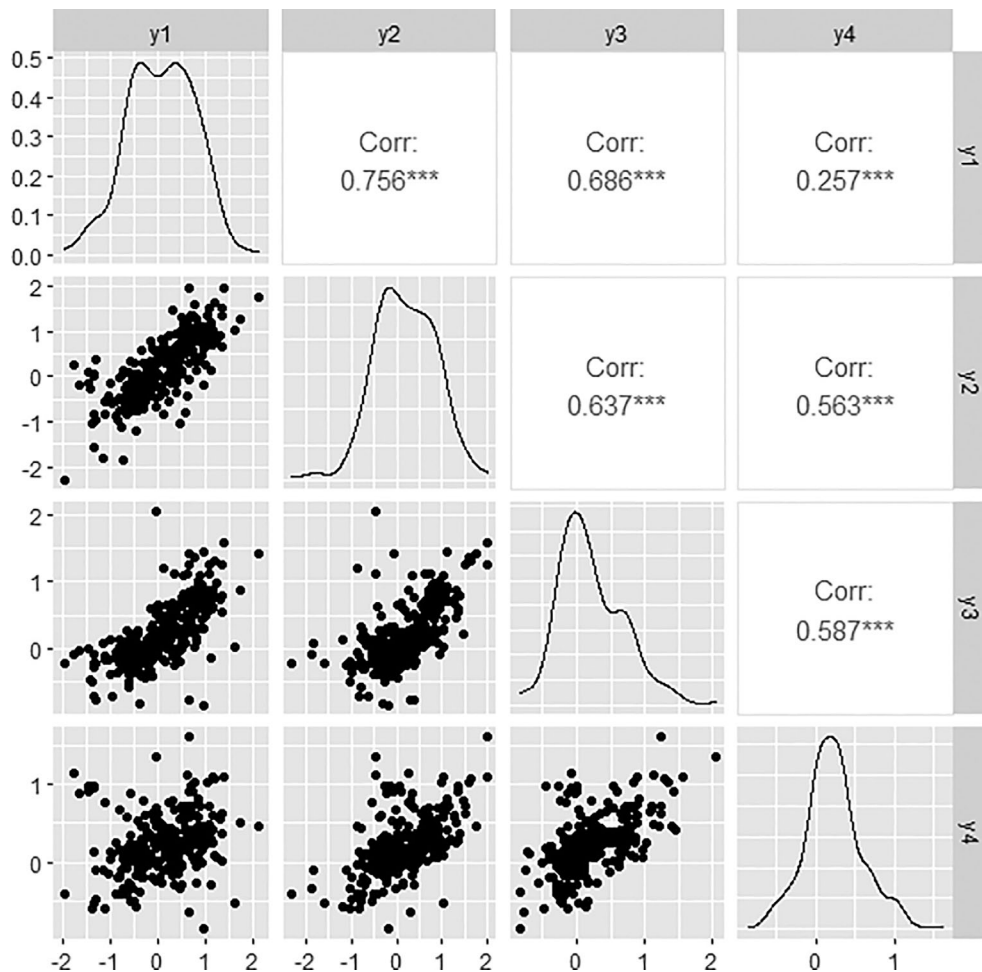


FIGURE 1 The correlation plots of the log-transformed gene expression level. Here “y1” represents the first observation, and so forth. The correlation coefficients among factors, the density maps of them, and the scatter plots of two factors lie on the upper right triangle, the diagonal, and the lower left triangle, respectively

smooth-threshold generalized estimating equations, which are based on the generalized estimating equations (GEE). Most of the above-mentioned methods only focused on the fixed dimension of covariates. Thus, Wang et al⁵ proposed a penalized GEE using a SCAD penalty and proved the asymptotic properties under the framework of large sample size n and diverging p_n , where p_n is the dimension of the covariates and is a function of the sample size n . The highlight of their procedure is that the consistency of model selection holds even when the working correlation structure is misspecified. However, the methods mentioned above are all based on the GEE. When the longitudinal data are contaminated or follow a heavy-tailed distribution, these methods are sensitive to response and/or covariates outliers. For example, in a large-scale yeast cell gene expression study reported by Spellman et al,⁶ genome-wide mRNA levels for 6178 yeast open reading frames that can determine which gene encode amino acids were recorded. The yeast cell cycle gene expression data cover approximately two cell-cycle periods and were collected at 7-minute intervals for 119 minutes, for a total of 18-time points measured at M/G1-G1-S-G2-M stages. Figure 1 reveals strong correlations, and the correlation matrix is neither a commonly used exchangeable nor autoregressive matrix. Furthermore, apart from the strong correlations, we find some outliers in the gene expression data (see Figure 2). Figure 3 indicates that abundant influence points occur in the observations of some important transcription factors, such as ASH1, MBP1, SWI4, and SWI6, which may lead to biased estimation and prediction. Some researchers have proposed various methods to identify important transcription factors (TFs) from a large set of transcription factors that are associated with gene expression levels and capture a complex relationship among those factors.^{5,7,8} Nevertheless, few researches focus on the robustness against outliers in observations, and most of them fail to capture the underlying correlation structure within gene expression level on multiple observations.

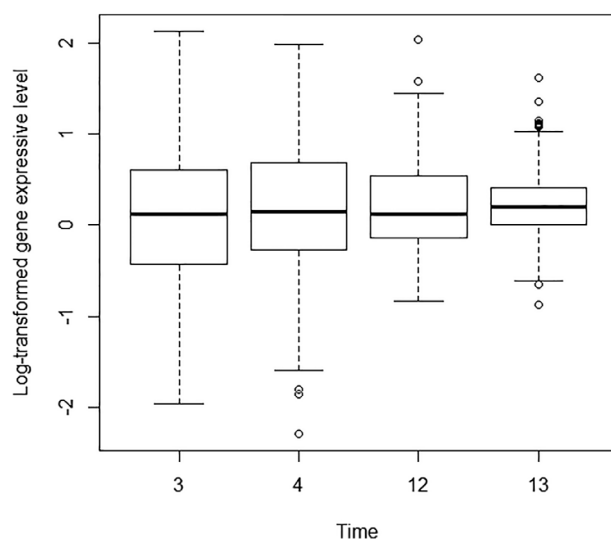


FIGURE 2 The boxplots of log-transformed gene expression level over four time points

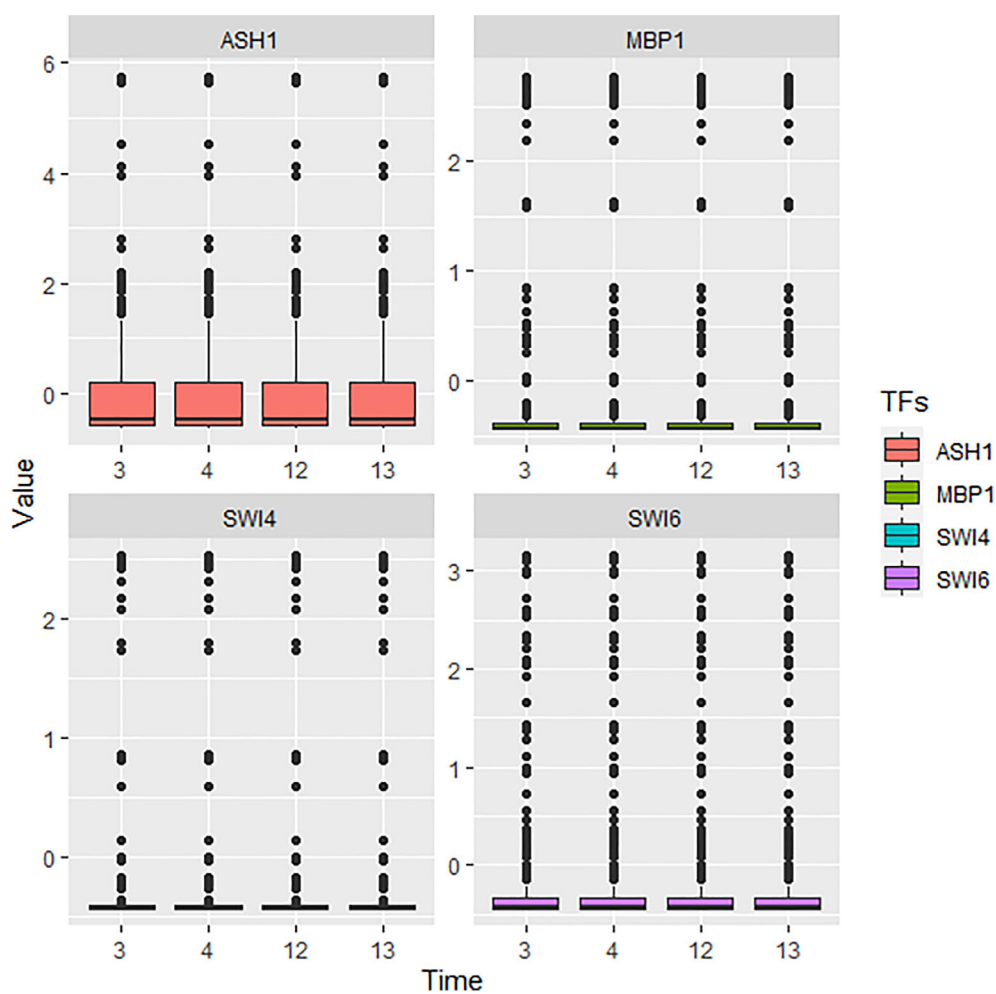


FIGURE 3 The boxplots of four important TFs: ASH1, MBP1, SWI4, and SWI6 over four time points [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com/doi/10.1002/sim.9213)]

Robust methods are desirable for contaminated data. Therefore, Fan et al⁹ proposed robust penalized estimating equations based on Huber's function for linear regression with longitudinal data, which is robust against outliers in response, but is sensitive to outliers in covariates. The regulated parameter in Huber's function is directly specified. Lv et al¹⁰ explored a weighted variable selection method based on an exponential squared loss¹¹ and a commonly used working correlation matrix for high dimensional longitudinal data, and they also provided a data-driven method to select tuning parameter in the exponential squared loss. These two methods are robust, but their work only looked at a specific case in which the variable dimension was no larger than the sample size, that is $p_n < n$.

In this article, we construct robust weighted estimating functions based on Tukey's biweight score equations, which are robust against outliers in response and/or covariates. Different from Li et al⁴ using robust residuals to estimate the correlation parameter, we propose a novel robust working correlation matrix to capture the correlations, which is more close to the true correlation matrix than the exchangeable and AR(1) correlation matrices, and performs competitively with the true correlation structure in variable selection. Following Li et al⁴ and Chang et al,¹² we establish robust smooth-threshold estimating equations for parameter estimation and variable selection. Furthermore, we prove the asymptotic properties of the proposed method under "large n and diverging p_n " setting. Robust estimating equations using bounded scores and leverage-based weights are robust against outliers and can reduce the bias when errors follow a heavy-tailed distribution. The proposed method can be applied to sparse marginal models under the large n small p_n , large n diverging p_n , and small n large p_n .

The rest of the article is organized as follows: In Section 2.1, we construct a robust estimating equation (RTGEE) for parameter estimation and variable selection. In Section 2.2, we apply an iterative algorithm to solve the smooth-threshold generalized estimating equations. In Section 2.3, we establish an effective criterion for tuning parameter selection. In Section 3, we establish the oracle properties of the proposed method. In Section 4, we carry out extensive simulation studies to evaluate the performance of the proposed method. In Section 5, we analyze a yeast cell cycle dataset to illustrate the proposed method. Finally, in Section 6, we draw some conclusions.

2 | ROBUST SMOOTH-THRESHOLD GEE

Suppose that $Y_i = (y_{i1}, \dots, y_{im_i})^T$ are measurements collected at times $(t_{i1}, \dots, t_{im_i})$ for the i th subject, where $i = 1, \dots, n$. Let $X_i = (x_{i1}, \dots, x_{im_i})$ be the corresponding covariate vector, in which $x_{ij} = (x_{ij1}, \dots, x_{ijp_n})^T$ is a $p_n \times 1$ vector. Assume that observations from the same subject are correlated, and observations from different subjects are independent. Denote the marginal mean of y_{ij} by $\mu_{ij} = E(y_{ij}|x_{ij}) = g(x_{ij}^T \beta)$, where $g(\cdot)$ is the inverse of the known link function, $\beta = (\beta_1, \dots, \beta_{p_n})^T$ is an unknown parameter vector, and variance of y_{ij} is $\text{Var}(y_{ij}|x_{ij}) = \phi v(\mu_{ij})$ with a variance function $v(\cdot)$ and a scale parameter ϕ . Let $\mu_i(\beta) = (\mu_{i1}, \dots, \mu_{im_i})^T$ and $A_i(\phi, \beta) = \phi \text{diag}(v(\mu_{i1}), \dots, v(\mu_{im_i}))$ be a diagonal matrix. For ease of notation, we use A_i for the rest of the article. The covariance matrix of Y_i is $\text{Cov}(Y_i) = A_i^{1/2} R_T A_i^{1/2}$, where R_T is the true correlation matrix of Y_i . Because R_T is unknown, Liang and Zeger¹³ proposed replacing R_T with a working correlation matrix $R(\alpha)$, where α is an unknown parameter vector and can be estimated using the residual-based moment method for a given working correlation matrix.¹³

2.1 | Methodology

We consider a new efficient and robust Tukey's biweight generalized estimating equation (RTGEE) for estimating the parameters in the marginal model with the longitudinal data,

$$U_n(\beta, \alpha) = \sum_{i=1}^n U_i(\beta) = \sum_{i=1}^n D_i^T V_i^{-1} h_i^b(\mu_i(\beta)) = 0, \quad (1)$$

where $D_i = \partial \mu_i(\beta) / \partial \beta$, and $V_i = R_i(\alpha) A_i^{\frac{1}{2}}$. The function $h_i^b(\mu_i(\beta)) = W_i [\tilde{\psi}_b(\mu_i(\beta)) - C_i(\mu_i(\beta))]$ with $C_i(\mu_i(\beta)) = E[\tilde{\psi}_b(\mu_i(\beta))]$, and W_i is a diagonal weight matrix used to downweight the effect of leverage points. One such leverage point, the j th element, is

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

where $r \geq 1$, b_0 is the 0.95 quantile of the χ^2 distribution with p_n degrees of freedom, and m_x and S_x are some robust estimators of the location and scale of x_{ij} . The robust function $\tilde{\psi}_b(\mu_i(\beta)) := \psi_b(A_i^{-1/2}(Y_i - \mu_i(\beta)))$ is given as follows:

$$\psi_b(u) = \begin{cases} u \left[1 - \left(\frac{u}{b} \right)^2 \right]^2 & \text{if } |u| \leq b, \\ 0 & \text{if } |u| > b \end{cases},$$

which is the derivative of Tukey's biweight loss function.

Guided by the idea of Ueki,¹⁴ we select important variables via an efficient and robust smooth-threshold GEE:

$$(I_{p_n} - \Delta) U_n(\beta, \alpha) + \Delta \beta = \mathbf{0}, \quad (2)$$

where I_{p_n} is the p_n -dimensional identity matrix, and $\Delta = \text{diag}\{\hat{\delta}_1, \hat{\delta}_2, \dots, \hat{\delta}_{p_n}\}$ is a diagonal matrix, in which $\hat{\delta}_j = \min \left\{ 1, \lambda / |\hat{\beta}_j^{(0)}|^{(1+\tau)} \right\}$ with a consistent estimator $\hat{\beta}_j^{(0)}$ of β_j . When $\hat{\delta}_j = 1$, we shrink $\hat{\beta}_j$ to zero and thus obtain a sparse estimator. The parameter τ can be selected among (0.5, 1, 2) according to a suggestion from numerical studies in Zou.¹⁵ In simulation studies, we found that $\tau = 1$ is highly effective for the numerical simulations.

To solve Equation (2), we need to specify the scale parameter ϕ and the working correlation matrix $R_i(\alpha)$. Let $\hat{\beta}$ be a consistent estimator of β . We use the robust median absolute deviation to estimate ϕ :¹⁶

$$\hat{\phi} = \left\{ 1.483 \text{ median } \left\{ \left| \hat{\eta}_{ij} - \text{median}(\hat{\eta}_{ij}) \right| \right\} \right\}^2, \quad (3)$$

where $\hat{\eta}_{ij} = (y_{ij} - \mu_{ij}(\hat{\beta})) / \sqrt{v(\mu_{ij})}$. Let $\hat{e}_i = (e_{i1}, \dots, e_{im_i})^T = A_i^{-1/2} (Y_i - \mu_i(\hat{\beta}))$ be the standardized Pearson residuals. For a chosen score function $\psi_b(\cdot)$, the corresponding robust residuals are denoted as $\psi_b(e_i) = \{\psi_b(e_{i1}), \dots, \psi_b(e_{im_i})\}^T$. Instead of estimating a constant correlation parameter for a specific correlation structure such as exchangeable and the first-order autoregressive correlation structures, here we use an unstructured correlation matrix $R(\alpha)$ as the working correlation matrix, in which $\alpha = (\alpha_{12}, \dots, \alpha_{1m_1}, \alpha_{23}, \dots, \alpha_{2m_2}, \dots, \alpha_{(m_i-1)m_i})^T$ is a $m_i(m_i - 1)/2$ -dimensional vector, and the element α_{kl} is the correlation coefficient of y_{ik} and y_{il} . We can utilize the moment method to estimate α_{kl} based on the robust residuals:

$$\hat{\alpha}_{kl} = \frac{1}{n} \sum_{i=1}^n \psi_b(\hat{e}_{ik}) \psi_b(\hat{e}_{il}).$$

Therefore, the working correlation matrix $R_i(\alpha)$ can be estimated by

$$\hat{R}_u = \frac{1}{n} \sum_{i=1}^n \psi_b(\hat{e}_i) \psi_b^T(\hat{e}_i). \quad (4)$$

To guarantee the diagonal elements of \hat{R}_u are equal to 1, and the off-diagonal elements of \hat{R}_u belong to $(-1, 1)$, we reconstruct the working correlation matrix estimate \hat{R}_u and propose the following matrix:

$$\hat{R}_{un} = \hat{B}_o^{-1/2} \hat{R}_u \hat{B}_o^{-1/2}, \quad (5)$$

where $\hat{B}_o = \text{diag}(\sum_{i=1}^n \psi_b^2(\hat{e}_{i1})/n, \sum_{i=1}^n \psi_b^2(\hat{e}_{i2})/n, \dots, \sum_{i=1}^n \psi_b^2(\hat{e}_{im_i})/n)$. Accordingly, we assign \hat{R}_{un} as an estimate of the working correlation matrix $R_i(\alpha)$. Hence, the diagonal elements of \hat{R}_{un} are equal to 1, and the off-diagonal elements of \hat{R}_{un} which are estimates of the vector α always lie in $(-1, 1)$ according to Cauchy-Schwarz inequality.

2.2 | Algorithm

To select the important variables and estimate the regression parameters in the marginal models, we follow a Fisher scoring iterative algorithm to implement the procedures as follows:

- Step 1. Give an initial estimator $\hat{\beta}^{(0)}$, for example, one can use the MM-estimate (see, Yohai¹⁷) as an initial value to ensure stability. Let $k = 0$.
- Step 2. Estimate the scale parameter $\hat{\phi}$ using (3) with the current estimator $\hat{\beta}^{(k)}$. Compute the working correlation matrix $R_i(\hat{\alpha})$ using (5), and we get

$$V_i \left(\mu_i \left(\hat{\beta}^{(k)} \right) \right) = R_i(\hat{\alpha}) \hat{A}_i^{1/2} \left(\hat{\phi}, \hat{\beta}^{(k)} \right).$$

- Step 3. For a given λ , we update the estimator of β via the following iterative formula:

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} - \left\{ \left(\sum_{i=1}^n D_i^T \Omega_i \left(\mu_i(\beta) \right) D_i + \hat{G} \right)^{-1} \left(U_n(\beta) + \hat{G}\beta \right) \right\}_{\beta=\hat{\beta}^{(k)}}, \quad (6)$$

where $\hat{G} = (I_{p_n} - \hat{\Delta})^{-1} \hat{\Delta}$, and $\Omega_i(\mu_i(\beta)) = V_i^{-1}(\mu_i(\beta)) \Gamma_i(\mu_i(\beta))$, in which

$$\Gamma_i(\mu_i(\beta)) = E \left[\dot{h}_i^b(\mu_i(\beta)) \right] = E \left[\partial h_i^b(\mu_i(\beta)) / \partial \mu_i \right] \Big|_{\mu_i=\mu_i(\beta)}.$$

- Step 4. Repeat Steps 2-3 until the algorithm converges. Here we set stop condition $\|\hat{\beta}^{(k+1)} - \hat{\beta}^{(k)}\|^2 < \epsilon$, where ϵ is a small number and takes a fixed value of $\epsilon = 10^{-8}$.

With the given λ and b , the corresponding estimator of β is denoted as $\hat{\beta}_\lambda^b$. According to the iterative algorithm mentioned above, we obtain a sandwich formula to estimate the asymptotic covariance matrix of $\hat{\beta}_\lambda^b$:

$$\text{Cov}(\hat{\beta}_\lambda^b) \approx \left[\hat{\Sigma}_n \left(\mu_i(\hat{\beta}_\lambda^b) \right) \right]^{-1} \hat{\mathbf{H}}_n \left(\mu_i(\hat{\beta}_\lambda^b) \right) \left[\hat{\Sigma}_n \left(\mu_i(\hat{\beta}_\lambda^b) \right) \right]^{-1}, \quad (7)$$

where

$$\hat{\mathbf{H}}_n \left(\mu_i(\hat{\beta}_\lambda^b) \right) = \sum_{i=1}^n D_i^T V_i^{-1} \left(\mu_i(\hat{\beta}_\lambda^b) \right) \left[h_i^b \left(\mu_i(\hat{\beta}_\lambda^b) \right) \left\{ h_i^b \left(\mu_i(\hat{\beta}_\lambda^b) \right) \right\}^T \right] V_i^{-1} \left(\mu_i(\hat{\beta}_\lambda^b) \right) D_i^T,$$

and

$$\hat{\Sigma}_n \left(\mu_i(\hat{\beta}_\lambda^b) \right) = \sum_{i=1}^n D_i^T V_i^{-1} \left(\mu_i(\hat{\beta}_\lambda^b) \right) \Gamma_i \left(\mu_i(\hat{\beta}_\lambda^b) \right) D_i.$$

2.3 | Selection of tuning parameters

To effectively select important variables using the proposed method, we need to choose proper tuning parameters b and λ as mentioned in Section 2.2, which determines the robustness of the estimator and consistency of variable selection respectively. For a given λ , we select the optimal parameter b in $\psi_b(u)$ from a series of candidates by minimizing the trace of the covariance matrix of $\hat{\beta}_\lambda^b$:

$$b_\lambda^{\text{opt}} = \min_b \text{trace}(\text{Cov}(\hat{\beta}_\lambda^b)) \quad (8)$$

The covariance matrix $\text{Cov}(\hat{\beta}_\lambda^b)$ can be obtained from (7).

For the regularization parameter λ selection, we adopt the PWD-type criterion proposed by Li et al⁴ to choose regularization parameter λ for (2):

$$\text{RPWD}_\lambda = \sum_{i=1}^n \{ h_i^{b_\lambda^{\text{opt}}}(\mu_i(\hat{\beta}_\lambda)) \}^T R_i^{-1}(\mu_i(\hat{\beta}_\lambda)) \{ h_i^{b_\lambda^{\text{opt}}}(\mu_i(\hat{\beta}_\lambda)) \} + df_\lambda \log(n), \quad (9)$$

where $\hat{\beta}_\lambda$ is the estimator of β for a given λ and the corresponding optimal b as in (8). Denote $df_\lambda = \sum_{j=1}^{p_n} 1(\hat{\delta}_j \neq 1)$ as the number of nonzero elements of the estimators. We choose λ , which corresponds to the minimizer of RPWD_λ , as an optimal value among a series of candidate values with a convergent solution $\hat{\beta}_\lambda^b$ under each λ and b values.

3 | ASYMPTOTIC PROPERTIES

In this section, we will establish large sample properties of the proposed estimator under a “large n , diverging p_n ” framework, which allows p_n to diverge to ∞ as n increases. The detailed proof of following Propositions are presented in the Appendix A in the Supplementary Information.

Let $\beta_0 = (\beta_{01}, \dots, \beta_{0p_n})^T$ be the true value of β , where $\beta \in \Theta$, $\Theta \subseteq \mathbb{R}^{p_n}$ is a bounded p_n -dimensional vector. Without loss of generality, we denote $\beta_0 = (\beta_{01}^T, \beta_{02}^T)^T$, where $\beta_{02} = \mathbf{0}$, and the elements of β_{01} are assumed to be nonzero in the dimension of s_n , which can also diverge with n . We partition β_0 into active (nonzero) coefficient sets $\mathcal{A}_0 = \{j : \beta_{0j} \neq 0\}$ with $|\mathcal{A}_0| = s_n$ and inactive (zero) coefficient sets $\mathcal{A}_0^c = \{j : \beta_{0j} = 0\}$. We define the active set $\mathcal{A} = \{j : \hat{\delta}_j \neq 1\}$ as the set of indices of nonzero estimated coefficients. Under the following conditions, we present the consistency of the proposed estimator.

C1. Assume x_{ij} for $1 \leq i \leq n$ and $1 \leq j \leq m_i$ satisfy $\sup_{i,j} \|x_{ij}\| = O(\sqrt{p_n})$.

C2. The unknown parameter β belongs to a compact subset $\Theta \subseteq \mathbb{R}^{p_n}$, and the true parameter value β_0 lies in the interior of Θ . Furthermore, we assume that the estimator of the correlation parameter vector $\hat{\alpha}$ is $\sqrt{p_n/n}$ -consistent given β and ϕ for some α , that is, $\|\hat{\alpha} - \alpha\| = O_p(\sqrt{p_n/n})$, and $|\partial \hat{\alpha}(\beta, \phi) / \partial \phi| \leq H(Y, \beta)$, where $H(\cdot, \cdot)$ is a bounded function for samples Y and β .

C3. Denote $\mathbf{X}_i^{h,b} = \mathbf{X}_i^T h_{0,i}^b(e_i)$. There exists finite positive constants $c_1 \leq c_2$ such that $\forall 1 \leq j \leq m_i$:

$$c_1 \leq \lambda_{\min} \left(n^{-1} \sum_{i=1}^n \mathbf{X}_i^{h,b} \left\{ \mathbf{X}_i^{h,b} \right\}^T \right) \leq \lambda_{\max} \left(n^{-1} \sum_{i=1}^n \mathbf{X}_i^{h,b} \left\{ \mathbf{X}_i^{h,b} \right\}^T \right) \leq c_2,$$

where $h_{0,i}^b(e_i) := h_i^b(e_i(\beta_0))$ with $e_i(\beta_0) = A_i^{-1/2}(Y_i - \mu_i(\beta_0))$ under true parameter β_0 . Similarly, the footnotes in $D_{0,i}$, $V_{0,i}$ and $\Gamma_{0,i}$ represent plugging in the true parameter β_0 .

C4. $\sup_{i \geq 1} E \|h_{0,i}^b(e_i)\|^{2+\delta} < \infty$ for some $\delta > 0$, and $0 < \sup_i \|E h_{0,i}^b(e_i) (h_{0,i}^b(e_i))^T\| < \infty$, where $h_i^b(e_i)$ centers by $\mu_i = \mu_i(\beta)$.

C5. There exists a positive constant c such that $0 < c \leq \inf_{i,j} v(\mu_{ij}) \leq \sup_{i,j} v(\mu_{ij}) < \infty$. The functions $C_{ij}(\mu_{ij}) = E \left[\psi_b \left((y_{ij} - \mu_{ij}) / \sqrt{v(\mu_{ij})} \right) \right]$, $v(\cdot)$ and $g(\cdot)$ have bounded second derivatives. The function $\psi_b(\cdot)$ is piecewise twice differentiable, and the second derivatives are bounded.

C6. Assume that $E \|U_n(\beta_0)\|^2 < \infty$ and there exists $\delta > 0$ such that

$$\lim_{n \rightarrow \infty} \frac{\sum_{i=1}^n E \|U_i(\beta_0)\|^{2+\delta}}{\left(E \|U_n(\beta_0)\|^2 \right)^{1+\delta/2}} = 0.$$

C7. Matrix

$$\Sigma = \lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n \left[D_{0,i}^T V_{0,i}^{-1} \Gamma_{0,i}(\mu_i(\beta_0)) D_{0,i} \right]$$

is positive definite. Matrix

$$B = \lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n D_{0,i}^T V_{0,i}^{-1} \text{cov}(h_i^b(\mu_i(\beta_0))) (V_{0,i}^{-1})^T D_{0,i}$$

is also positive definite.

C8. For any positive λ , τ , p_n , and s_n , $(p_n/n)^{(1+\tau)/2}\lambda^{-1} \rightarrow 0$, $(p_n/n)^{-1/2}\lambda \rightarrow 0$, $n^{-1/2}\lambda^2 = o(1)$, and $s_n n^{-1/2} = o(1)$, such as $s_n = O(n^{1/3})$.

Condition C1 is a common assumption in the M-estimator with diverging dimension,¹⁸ and it holds almost surely under some weak moment conditions for x_{ij} from spherically symmetric distributions. Condition C2 is established to ensure the $\sqrt{p_n/n}$ -consistency of $R_i(\hat{\alpha})$ in Section 2.1, which can be verified using similar analysis in He et al.¹⁹ Taking the spirit of lemma 3.7 in Wang,²⁰ we set condition C3, which is especially useful when establishing the asymptotic normality in Theorem 2. Similar to Lv et al.,¹⁰ conditions C4 and C5 can be easily checked under bounded score function $\psi_b(\cdot)$, and they are usually combined with conditions C6 and C7, which are also necessary for the central limit theory and hold in most cases. Condition C8 is established for exploring convergence rate and asymptotic properties, which controls the order of diverging number p_n and s_n precisely. Note that a preliminary $\sqrt{p_n/n}$ -consistent estimator β_0 is needed in both Theorems 1 and 2. When there are no outliers, it can be obtained by solving the generalized estimating equations under an independence working correlation structure as in Example 1 in Wang²⁰ when $p_n \rightarrow \infty$.

Theorem 1 (Consistency). *Suppose the regularity conditions C1-C8 hold, then we have*

$$\|\hat{\beta}_{\lambda,\tau} - \beta_0\| = O_p\left(\sqrt{p_n/n}\right).$$

Theorem 2 (Oracle properties). *Under conditions C1-C8, and if $n^{-1}p_n^3 = o(1)$, as $n \rightarrow \infty$, we have*

- (1) variable selection consistency, $P(\mathcal{A} = \mathcal{A}_0) \rightarrow 1$;
- (2) asymptotic normality: $\forall \alpha_n \in R^{s_n}$ such that $\|\alpha_n\| = 1$,

$$\sqrt{n}\alpha_n^T \mathbf{B}_{\mathcal{A}_0}^{-1/2} \Sigma_{\mathcal{A}_0} (\hat{\beta}_{\lambda,\tau,\mathcal{A}} - \beta_{\mathcal{A}_0}) \xrightarrow{d} N(0, 1),$$

where $\Sigma_{\mathcal{A}_0}$ and $\mathbf{B}_{\mathcal{A}_0}$ are the first $s_n \times s_n$ submatrices of Σ and \mathbf{B} .

Theorem 1 implies that our proposed estimator can achieve $\sqrt{p_n/n}$ -consistency. Theorem 2 shows that such consistent estimators possess the sparsity property and oracle property²¹ when we choose proper λ and τ . With probability approaching 1, our proposed method can correctly select the nonzero coefficients and estimate them as efficiently as if we know the correct submodel in advance.

4 | SIMULATION STUDIES

We conduct simulation studies to assess the performance of the proposed RTGEE method, the smooth-threshold generalized estimating equation (SGEE) proposed by Li et al.,⁴ the robust smooth-threshold generalized estimating equation (RSGEE) corresponding to Huber's score function, and the efficient and robust generalized estimating equation (ERSGEE) proposed by Lv et al.¹⁰ for continuous normal data and heavy-tailed data under setups $p_n < n$, large n and diverging p_n , and $p_n > n$. The default value of the tuning parameter in Huber's score function for RSGEE is 1.345, which can reach 95% efficiency when the data follow a normal distribution. We follow a similar tuning parameter selection criterion proposed in Lv et al.¹⁰ to choose an optimal parameter for ERSGEE, where the tuning parameter takes a range of candidates from set $\{2, 4, 6, 8, 10, 12, 14\}$, as suggested in Wang et al.¹¹ For the tuning parameter b in Tukey's biweight, a series of candidates have been listed in table 2 in Riani et al.,²² which are derived from a series of fixed asymptotic efficiency of the robust S-estimator of the regression coefficients in the Gaussian model, see, for example, Rousseeuw and Leroy.²³ We choose the b values that produce the corresponding asymptotic efficiency higher than 0.7 in our simulations, namely $b \in \{2.697, 2.897, 3.137, 3.444, 3.883, 4.03, 4.685, 5.9, 7.0414\}$. In our simulations, we search for the optimal regularization parameter λ via a grid of candidates, ranging from 0.008 to 0.08 with an interval of 0.008.

For each procedure, the true correlation structure of the response is exchangeable (EXC) with the correlation coefficient $\alpha = 0.7$. For each setup in the simulations, we generate 100 datasets and apply the iterative algorithm mentioned in Section 2.2 to estimate β and select important variables at the same time. Furthermore, we also consider the situation in which the true correlation structure is AR(1) with correlation parameter $\alpha = 0.7$ for the continuous data. The simulation

results show similar patterns and are presented in Tables 1-6 in the supplementary materials. Finally, we also consider the count data, and the results are listed in Tables 7-8 in the supplementary materials.

We compare these four methods under three working correlation matrices (EXC, AR(1), R_{un}) according to the following terms: the average number of correctly identified insignificant variables (C), the average number of incorrectly identified significant variables (IC), the correctly fitted odds (CF, the odd of identifying both significant variables and insignificant variables correctly over 100 simulations), the biases of estimators, the SD of estimators, the proportion of estimators fall into the 95% confidence interval (CI), the average of mean squared prediction error (AMSPE), the median of mean squared prediction error (MMSPE), where $MSPE = n^{-1} \sum_{i=1}^n (\hat{y}_i - y_i)^2$, and the average mean square error (AMSE), which is the average of $\|\hat{\beta} - \beta_0\|^2$ over 100 simulations. To demonstrate the efficiency of estimators, we compare the relative efficiency among three robust methods in Figures B1-B3 in the supplementary materials, which is defined as the ratio of the AMSE for SGEE to the AMSE for each robust method, from which a higher value represents higher efficiency. We also list the running time for each procedure in Tables 1-6 for further comparison, and we provide the mean values of the selected optimal parameters of b and λ in each scenario over 100 simulations in Tables 1-6 for reference. We present partial results in Tables 1-6 and more details can be found in Tables 9-14 in Appendix B in the supplementary information.

4.1 | Heavy-tailed continuous data

We generate the continuous data from the following model:

$$y_{ij} = x_{ij1}\beta_1 + x_{ij2}\beta_2 + \cdots + x_{ijp_n}\beta_{p_n} + \epsilon_{ij}, \quad i = 1, \dots, n, \quad j = 1, \dots, m_i. \quad (10)$$

Without loss of generality, we consider the balanced data with $m_i = 10$ for $i = 1, \dots, n$. Covariates $x_{ij} = (x_{ij1}, \dots, x_{ijp_n})^T$ follow a multivariate normal distribution with a mean of zero and the correlation between the k th and l th component of x_{ij} being $0.5^{|l-k|}$. The random error vectors $\epsilon_i = (\epsilon_{i1}, \dots, \epsilon_{i10})^T$ are generated from a multivariate Student's t -distribution with three degrees of freedom $T_3(0, R(\alpha))$. The true coefficients are assumed to be $\beta = (0.7, 0.7, -0.4, 0, \dots, 0)^T$ with nonzero coefficients $s_n = 3$ and $p_n - s_n$ coefficients being zero.

I. We first test performance when $p_n = 20$ and $n = 100$ under the following scenarios:

Case 1: There is no contamination on the dataset.

Case 2: We randomly add 20% y -outliers following $N(10, 1)$ on y_{ij} .

Case 3: We randomly add 10% x -outliers on x_{ij1} , following a Student's t -distribution with three degrees of freedom.

Meanwhile, we change response variables in the same way as Case 2.

II. Next we consider "large n and diverging p_n ".

The true coefficients are set as $\beta = (0.7, 0.7, -0.4, 0.7, 0.7, -0.4, \dots, \mathbf{0}_{p_n-s_n})$ with $p_n = [4n^{2/5}] - 5$ and the scale of nonzero coefficients $s_n = [p_n/5]$ for $n = 200$, where $[s]$ denotes the largest positive integer value not greater than s . In addition, the observation times m_i are randomly generated from 2 to 5. The other settings are the same as those in I.

III. We set $p_n = 300$ and $n = 100$, and the true coefficients vector $\beta = (0.7, 0.7, -0.4, 0, \dots, 0)^T$ is a p_n -dimensional vector with only three nonzero components. Other conditions are the same as I. To ensure the stability of simulations, we decrease the proportions of outliers as follows:

Case 2': We randomly add 10% y -outliers following $N(10, 1)$ on y_{ij} .

Case 3': We randomly add 10% x -outliers on x_{ij1} following a Student's t distribution with three degrees of freedom, and we randomly add 10% y -outliers, similar to Case 2'.

The simulation results for I, II, and III are presented in Tables 1-3, respectively. From Table 1, it is evident that non-robust SGEE has manifest shortcomings in variable selection compared with the other three robust methods according to the value of CF even in the no contamination case. When there are outliers in data sets, the defect of SGEE shows more clearly no matter in the variable selection or coefficient estimation. In contrast, the three robust methods (RSGEE, ERSAGEE, and RTGEE) perform well, even in a misspecified correlation structure. However, when adding outliers to the response variables, ERSAGEE and our proposed method, RTGEE, perform better in IC and CF than SGEE and RSGEE. RTGEE is more competitive with ERSAGEE in both coefficients estimation and variable selection when adding x -outliers and y -outliers simultaneously. RTGEE has a smaller estimation error (MMSPE) and higher CF than the other three methods. Furthermore, we find that, as a type of misspecified correlation structure, the results under R_{un} are superior over AR(1) and competitive with true correlation structure EXC, especially when there are outliers in the dataset. The R_{un}

TABLE 1 Correlated continuous data for $n > p_n$ ($p_n = 20$ and $n = 100$) with ϵ_{ij} following a $t(3)$ distribution: Comparison of SGEE, RSGEE, ERSAGE, and the proposed method RTGEE with three different working correlation matrices (exchangeable, AR(1) and unstructured)

Scenario	R	Method	β_1	β_2	β_3	MMSPE	No. of zeros			Time (m)	\bar{b}_{opt}	$\bar{\lambda}_{opt}$
			CI	CI	CI		C	IC	CF			
Case 1	EXC	SGEE	0.96	0.94	0.95	0.0029	16.65	0.05	0.74	0.97	5.233	0.060
		RSGEE	0.95	0.94	0.95	0.0016	16.91	0.00	0.92	6.10		
		ERSAGE	0.93	0.96	0.93	0.0020	16.96	0.00	0.96	38.45		
		RTGEE	0.92	0.95	0.97	0.0022	17.00	0.03	0.97	42.94		
	AR(1)	SGEE	0.95	0.94	0.96	0.0042	16.42	0.02	0.60	1.13	6.698	0.061
		RSGEE	0.96	0.95	0.98	0.0023	16.88	0.02	0.90	6.68		
		ERSAGE	0.95	0.94	0.91	0.0023	16.95	0.00	0.95	44.38		
		RTGEE	0.95	0.96	0.96	0.0028	17.00	0.04	0.96	60.18		
	R_{un}	SGEE	0.96	0.95	0.95	0.0025	16.55	0.05	0.70	1.54	6.655	0.053
		RSGEE	0.95	0.95	0.96	0.0017	16.92	0.00	0.93	8.52		
		ERSAGE	0.95	0.94	0.94	0.0017	16.96	0.00	0.96	57.60		
		RTGEE	0.97	0.96	0.96	0.0029	17.00	0.04	0.96	60.79		
Case 2	EXC	SGEE	0.95	0.93	0.92	0.0740	15.37	0.06	0.33	1.17	3.311	0.044
		RSGEE	0.97	0.95	0.97	0.0078	16.89	0.01	0.90	8.16		
		ERSAGE	0.93	0.95	0.95	0.0028	16.96	0.00	0.96	43.72		
		RTGEE	0.95	0.97	0.97	0.0021	16.95	0.02	0.95	43.89		
	AR(1)	SGEE	0.95	0.94	0.97	0.0807	15.06	0.00	0.27	1.15	3.563	0.046
		RSGEE	0.96	0.95	0.98	0.0095	16.90	0.01	0.91	9.06		
		ERSAGE	0.96	0.96	0.95	0.0036	16.96	0.00	0.96	48.32		
		RTGEE	0.98	0.93	0.95	0.0030	17.00	0.05	0.95	49.19		
	R_{un}	SGEE	0.92	0.91	0.93	0.0768	15.63	0.03	0.38	1.56	3.591	0.043
		RSGEE	0.96	0.95	0.98	0.0079	16.90	0.02	0.90	8.49		
		ERSAGE	0.94	0.96	0.96	0.0026	16.93	0.00	0.94	51.99		
		RTGEE	0.97	0.93	0.97	0.0023	16.99	0.03	0.96	66.02		
Case 3	EXC	SGEE	0.95	0.95	0.92	8.3204	15.94	0.05	0.42	0.87	3.319	0.065
		RSGEE	0.90	0.96	0.96	0.0084	16.90	0.04	0.88	7.96		
		ERSAGE	0.94	0.95	0.96	0.0048	16.90	0.00	0.90	36.75		
		RTGEE	0.96	0.93	0.94	0.0027	16.97	0.02	0.95	39.45		
	AR(1)	SGEE	0.94	0.92	0.93	0.0728	15.86	0.03	0.45	1.12	3.543	0.065
		RSGEE	0.93	0.96	0.95	0.0089	16.90	0.05	0.88	9.16		
		ERSAGE	0.94	0.95	0.97	0.0067	16.90	0.00	0.90	39.69		
		RTGEE	0.93	0.93	0.94	0.0030	16.97	0.04	0.94	46.07		
	R_{un}	SGEE	0.96	0.92	0.94	0.0770	15.87	0.04	0.42	1.60	3.573	0.063
		RSGEE	0.93	0.95	0.96	0.0088	16.94	0.04	0.92	9.23		
		ERSAGE	0.97	0.94	0.94	0.0046	16.91	0.00	0.91	51.98		
		RTGEE	0.96	0.91	0.93	0.0032	16.96	0.05	0.92	61.64		

Note: \bar{b}_{opt} and $\bar{\lambda}_{opt}$ are the mean values of the selected optimal b_{opt} and λ_{opt} of the proposed method based on 100 simulations.

TABLE 2 Correlated continuous data for large n and diverging p_n ($n = 200$ and $p_n = \lceil 4n^{2/5} \rceil - 5$) with ϵ_{ij} following a $t(3)$ distribution: Comparison of SGEE, RSGEE, ERSAGEE, and the proposed method RTGEE with three different working correlation matrices (exchangeable, AR(1) and unstructured)

Scenario	R	Method	β_1	β_2	β_3	MMSPE	No. of zeros			Time (m)	\bar{b}_{opt}	$\bar{\lambda}_{\text{opt}}$
			CI	CI	CI		C	IC	CF			
Case 1	EXC	SGEE	0.93	0.95	0.93	0.0301	21.58	0.06	0.46	13.71		
		RSGEE	0.96	0.95	0.94	0.0143	22.75	0.05	0.84	29.02		
		ERSAGEE	0.94	0.93	0.91	0.0158	22.81	0.07	0.88	170.10		
		RTGEE	0.95	0.93	0.95	0.0117	22.91	0.04	0.92	65.62	5.328	0.057
	AR(1)	SGEE	0.93	0.93	0.95	0.0274	21.48	0.05	0.45	14.89		
		RSGEE	0.93	0.95	0.94	0.0143	22.82	0.06	0.85	29.71		
		ERSAGEE	0.96	0.95	0.95	0.0164	22.51	0.00	0.83	176.20		
		RTGEE	0.96	0.95	0.91	0.0129	22.97	0.08	0.90	69.96	5.151	0.057
	R_{un}	SGEE	0.94	0.92	1.00	0.0214	22.67	0.39	0.48	62.08		
		RSGEE	0.95	0.95	0.94	0.0069	22.98	0.06	0.92	100.90		
		ERSAGEE	0.93	0.92	0.94	0.0118	22.93	0.06	0.89	548.00		
		RTGEE	0.94	0.96	0.94	0.0071	23.00	0.06	0.94	248.38	6.404	0.070
Case 2	EXC	SGEE	0.97	0.99	0.98	0.3731	19.05	0.01	0.15	10.19		
		RSGEE	0.96	0.97	0.98	0.0501	22.35	0.06	0.72	29.59		
		ERSAGEE	0.95	0.94	0.87	0.0218	22.94	0.12	0.84	163.70		
		RTGEE	0.93	0.95	0.92	0.0193	22.93	0.08	0.87	63.11	3.869	0.070
	AR(1)	SGEE	0.97	0.93	0.95	0.3781	18.84	0.01	0.10	15.13		
		RSGEE	0.92	0.96	0.94	0.0479	22.28	0.05	0.72	40.14		
		ERSAGEE	0.95	0.94	0.87	0.0213	22.91	0.12	0.84	209.24		
		RTGEE	0.96	0.95	0.93	0.0201	22.77	0.05	0.84	69.76	3.707	0.073
	R_{un}	SGEE	0.97	0.94	0.98	0.3429	18.82	0.03	0.13	51.78		
		RSGEE	0.96	0.95	0.87	0.0352	22.93	0.12	0.81	89.94		
		ERSAGEE	0.96	0.92	0.88	0.0174	22.97	0.12	0.85	404.80		
		RTGEE	0.97	0.95	0.92	0.0146	22.93	0.07	0.86	249.82	3.715	0.080
Case 3	EXC	SGEE	0.99	0.99	0.93	0.3062	18.86	0.02	0.10	11.47		
		RSGEE	0.96	0.95	0.98	0.0341	22.69	0.08	0.79	37.54		
		ERSAGEE	0.97	0.95	0.85	0.0213	23.00	0.14	0.86	202.10		
		RTGEE	0.97	0.97	0.92	0.0192	22.91	0.10	0.87	56.03	3.895	0.075
	AR(1)	SGEE	0.94	0.94	0.92	0.2950	18.82	0.01	0.08	12.77		
		RSGEE	0.96	0.95	0.91	0.0337	22.72	0.08	0.81	37.82		
		ERSAGEE	0.98	0.92	0.86	0.0183	22.93	0.14	0.83	206.20		
		RTGEE	0.97	0.96	0.94	0.0168	22.77	0.05	0.84	62.46	3.672	0.073
	R_{un}	SGEE	0.94	0.96	0.93	0.3040	18.73	0.01	0.13	60.61		
		RSGEE	0.95	0.94	0.87	0.0247	22.98	0.13	0.85	110.10		
		ERSAGEE	0.94	0.93	0.85	0.0141	23.00	0.14	0.86	476.10		
		RTGEE	0.95	0.93	0.92	0.0114	22.97	0.08	0.89	252.17	3.699	0.072

Note: \bar{b}_{opt} and $\bar{\lambda}_{\text{opt}}$ are the mean values of the selected optimal b_{opt} and λ_{opt} of the proposed method based on 100 simulations.

TABLE 3 Correlated continuous data for $p_n > n$ ($n = 100$ and $p_n = 300$) with ϵ_{ij} following a $t(3)$ distribution: Comparison of SGEE, RSGEE, ERSAGE, and the proposed method RTGEE with three different working correlation matrices (exchangeable, AR(1) and unstructured)

Scenario	R	Method	β_1	β_2	β_3	MMSPE	No. of zeros			Time (m)	\bar{b}_{opt}	$\bar{\lambda}_{opt}$
			CI	CI	CI		C	IC	CF			
Case 1	EXC	SGEE	0.93	0.95	0.96	0.0057	293.16	0.04	0.39	9.55		
		RSGEE	0.94	0.96	0.97	0.0015	296.83	0.03	0.82	23.36		
		ERSAGE	0.95	0.94	0.95	0.0021	297.00	0.05	0.95	194.60		
		RTGEE	0.95	0.94	0.96	0.0020	297.00	0.02	0.98	452.29	3.297	0.049
	AR(1)	SGEE	0.96	0.83	1.00	0.1159	295.53	0.54	0.30	9.94		
		RSGEE	0.97	0.97	0.97	0.0023	296.80	0.03	0.83	23.79		
		ERSAGE	0.95	0.94	0.97	0.0029	296.98	0.03	0.96	198.60		
		RTGEE	0.95	0.95	0.96	0.0025	297.00	0.04	0.96	464.31	2.697	0.078
	R_{un}	SGEE	0.94	0.96	0.95	0.0045	294.26	0.05	0.43	13.53		
		RSGEE	0.94	0.97	0.97	0.0015	296.82	0.03	0.84	23.68		
		ERSAGE	0.96	0.95	0.96	0.0026	296.98	0.04	0.95	267.80		
		RTGEE	0.95	0.96	0.96	0.0023	297.00	0.04	0.96	545.51	2.697	0.075
Case 2'	EXC	SGEE	0.95	0.96	0.89	0.3216	271.30	0.09	0.00	10.25		
		RSGEE	0.96	0.95	0.97	0.0034	296.67	0.03	0.75	25.90		
		ERSAGE	0.95	0.96	0.96	0.0027	296.90	0.04	0.88	199.60		
		RTGEE	0.94	0.94	0.94	0.0027	297.00	0.06	0.94	7.29	3.444	0.050
	AR(1)	SGEE	0.96	0.92	1.00	0.1445	289.53	0.49	0.02	10.00		
		RSGEE	0.97	0.93	0.90	0.0050	296.85	0.10	0.78	23.54		
		ERSAGE	0.97	0.90	0.87	0.0035	296.98	0.13	0.86	176.50		
		RTGEE	0.95	0.94	0.94	0.0034	296.92	0.05	0.90	975.76	2.968	0.057
	R_{un}	SGEE	0.97	0.95	0.90	0.3642	271.55	0.07	0.01	11.87		
		RSGEE	0.96	0.95	0.96	0.0039	296.66	0.04	0.73	24.90		
		ERSAGE	0.97	0.95	0.95	0.0031	296.90	0.05	0.87	231.90		
		RTGEE	0.93	0.94	0.94	0.0028	296.96	0.06	0.90	1053.98	3.030	0.061
Case 3'	EXC	SGEE	0.96	0.96	0.94	0.6168	242.35	0.00	0.01	8.61		
		RSGEE	0.94	0.95	0.94	0.0042	296.75	0.06	0.79	20.08		
		ERSAGE	0.93	0.92	0.93	0.0026	296.87	0.07	0.86	178.10		
		RTGEE	0.94	0.93	0.95	0.0026	296.99	0.03	0.96	708.04	3.292	0.054
	AR(1)	SGEE	0.94	0.95	0.93	0.7165	242.15	0.00	0.00	9.55		
		RSGEE	0.94	0.95	0.93	0.0052	296.78	0.07	0.79	19.46		
		ERSAGE	0.95	0.93	0.93	0.0028	296.87	0.07	0.86	194.30		
		RTGEE	0.96	0.93	0.93	0.0030	296.99	0.13	0.86	887.92	3.060	0.053
	R_{un}	SGEE	0.93	0.95	0.96	0.6998	242.27	0.00	0.00	11.52		
		RSGEE	0.94	0.94	0.93	0.0039	296.79	0.06	0.80	21.51		
		ERSAGE	0.94	0.93	0.93	0.0029	296.88	0.07	0.87	222.20		
		RTGEE	0.94	0.94	0.95	0.0027	297.00	0.12	0.88	1529.86	3.049	0.056

Note: \bar{b}_{opt} and $\bar{\lambda}_{opt}$ are the mean values of the selected optimal b_{opt} and λ_{opt} of the proposed method based on 100 simulations.

boosts the performance of non-robust SGEE in variable selection and significantly decreases prediction and estimation error compared to AR(1). Figure B1 depicts the relative efficiency for RSGEE, ERSAGE, and RTAGE under settings I. The left plot A shows that our proposed estimator has higher relative efficiency than the other two robust methods for contaminated heavy-tailed data, and it is more obvious when there are both x -outliers and y -outliers.

When p_n is diverging, the results in Table 2 indicate that the proposed method is comparable with RSGEE and ERSAGE in Cases 1 and 2 and performs better than RSGEE and ERSAGE when the covariates have outliers. In Case 3, our proposed method performs superiorly over the other methods regardless of the estimation or variable selection, which confirms that our proposed method has superiority under diverging p_n . We notice that the CI of our proposed estimator always flies floats around 95%, even with contamination, which implies the asymptotic normality of our proposed estimator and gives a numerical validation of Theorem 2 established in Section 3. It is appealing that almost all the methods perform better under our proposed working correlation structure R_{un} , even better than the estimated true correlation structure regardless of whether there are outliers. The proposed method outperforms when the data are contaminated (see Figure B2 in the supplementary materials).

From Table 3 and plot E in Figure B3, we can see that ERSAGE and RTAGE show superiority in variable selection compared to SGEE and RSGEE when there are no outliers. When outliers are added, our proposed method is superior to ERSAGE with a lower MMSPE and higher relatively efficiency, implying RTAGE can keep robustness against outliers even under $p_n > n$ and misspecified working correlation structure.

4.2 | Continuous normal data

We generate response variable according to model (10), and the covariates are generated in the same way as in Section 4.1. The random error vectors $\epsilon_i = (\epsilon_{i1}, \dots, \epsilon_{i10})$ are generated from a $N_{10}(0, R(\alpha))$ with the correlation coefficient $\alpha = 0.7$. Other settings are the same as those in Section 4.1, except that when testing the performance of the foregoing methods in a sparse model when $p_n > n$, we decrease the proportions of outliers as follows:

Case 2'': We randomly convert 10% of y_{ij} into $y_{ij} + 5$.

Case 3'': We artificially add 5% x -outliers on x_{ij1} following a $t(3)$ distribution as well as y -outliers which are the same with Case 2''.

The corresponding results are listed in Tables 4-6. From Table 4, when there are no outliers added in response and/or covariates, SGEE can perform well as expected, whereas, it is inferior to robust methods no matter in parameter estimation or variable selection in contaminated datasets. It is noticeable that our proposed unstructured working correlation matrix R_{un} performs competitively with the true correlation matrix, and it is superior to wrongly assigned working correlation matrix (eg, AR(1)), especially when there are outliers in response and/or covariates.

In Table 5, when p_n is diverging, SGEE can not perform as well as robust methods even if there are no outliers added. Among robust methods, ERSAGE and RTAGE perform similarly well and they are superior to RSGEE when there are outliers in datasets. Furthermore, RTAGE has significant advantages with higher relative efficiency than ERSAGE, especially under the wrongly assigned working correlation structure, according to plot D in Figure B2 in the supplementary materials. The reasonable performance of our proposed estimators in CI verifies the oracle properties again. It is noticeable that, although we allow a wider range of candidates for tuning parameter selection for RTAGE, the running time listed in Tables 2 and 5 shows the proposed method is superior to ERSAGE in the “diverging p_n ” setting, indicating a faster convergence rate.

When $p_n > n$, the results in Table 6 show that SGEE is affected by outliers more obviously in variable selection compared with robust methods. Although RTAGE and ERSAGE perform relatively similar well, where they are preferred methods when outliers are added in covariates, RTAGE enjoys superior prediction ability than ERSAGE with lower MMSPE.

5 | REAL DATA ANALYSIS

The cell cycle is one of the most important processes for cell growth, DNA replication, chromosome segregation, and daughter cells' division. Investigating the functions of gene expression during the cell cycle process can give an insight into how the cell cycle affects biological processes and cell cycle regulation. Transcription factors (TFs) are critical in the cell cycle process, where they have been shown to influence gene expression by regulating the flow of genetic information from

TABLE 4 Correlated continuous data with ϵ_{ij} following a normal distribution for $n > p_n$ ($p_n = 20$ and $n = 100$): Comparison of SGEE, RSGEE, ERSAGE, and the proposed method RTGEE with three different working correlation matrices (exchangeable, AR(1) and unstructured)

Scenario	R	Method	β_1	β_2	β_3	MMSPE	No. of Zeros			Time (m)	\bar{b}_{opt}	$\bar{\lambda}_{opt}$
			CI	CI	CI		C	IC	CF			
Case 1	EXC	SGEE	0.94	0.98	0.98	0.0008	16.99	0.02	0.97	0.93		
		RSGEE	0.96	0.98	0.98	0.0008	16.97	0.02	0.95	5.01		
		ERSAGE	0.95	0.97	0.96	0.0008	16.99	0.00	0.99	40.14		
		RTGEE	0.96	0.97	0.99	0.0009	17.00	0.01	0.99	52.72	7.041	0.052
	AR(1)	SGEE	0.96	0.97	0.99	0.0009	16.90	0.01	0.91	1.01		
		RSGEE	0.95	0.96	0.96	0.0011	16.97	0.04	0.93	6.94		
		ERSAGE	0.95	0.93	0.95	0.0013	17.00	0.00	1.00	43.59		
		RTGEE	0.96	0.99	0.99	0.0011	17.00	0.01	0.99	55.83	7.041	0.051
	R_{un}	SGEE	0.93	0.99	0.99	0.0009	16.97	0.01	0.96	1.59		
		RSGEE	0.96	0.98	0.98	0.0010	16.96	0.02	0.94	8.59		
		ERSAGE	0.95	0.98	0.95	0.0010	17.00	0.00	1.00	56.27		
		RTGEE	0.95	0.99	0.99	0.0009	17.00	0.01	0.99	61.11	7.041	0.057
Case 2	EXC	SGEE	0.95	0.94	0.91	0.0528	16.75	0.05	0.75	0.93		
		RSGEE	0.96	0.94	0.98	0.0046	16.94	0.01	0.93	6.39		
		ERSAGE	0.96	0.94	0.96	0.0017	16.99	0.00	0.99	36.17		
		RTGEE	0.96	0.94	0.96	0.0014	17.00	0.00	1.00	48.82	4.958	0.041
	AR(1)	SGEE	0.97	0.94	1.00	0.0779	16.94	0.27	0.67	1.11		
		RSGEE	0.96	0.95	0.97	0.0057	16.95	0.02	0.93	6.60		
		ERSAGE	0.92	0.96	0.93	0.0022	16.99	0.00	0.99	42.69		
		RTGEE	0.95	0.95	0.93	0.0021	16.98	0.00	0.99	52.40	4.841	0.042
	R_{un}	SGEE	0.96	0.96	0.92	0.0633	16.81	0.04	0.79	1.50		
		RSGEE	0.97	0.97	0.96	0.0051	16.96	0.01	0.95	7.18		
		ERSAGE	0.93	0.95	0.96	0.0018	16.99	0.00	0.99	50.09		
		RTGEE	0.95	0.95	0.95	0.0016	17.00	0.00	1.00	63.08	4.860	0.045
Case 3	EXC	SGEE	0.95	0.97	0.94	0.0537	16.66	0.04	0.75	1.03		
		RSGEE	0.95	0.97	0.96	0.0047	16.93	0.00	0.94	12.49		
		ERSAGE	0.94	0.97	0.95	0.0024	16.91	0.00	0.93	39.12		
		RTGEE	0.96	0.95	0.95	0.0017	17.00	0.04	0.96	47.20	4.668	0.067
	AR(1)	SGEE	0.92	0.98	0.94	0.0583	16.59	0.04	0.69	1.10		
		RSGEE	0.95	0.95	0.94	0.0048	16.86	0.00	0.88	12.24		
		ERSAGE	0.94	0.93	0.95	0.0024	16.90	0.00	0.90	42.72		
		RTGEE	0.94	0.95	0.95	0.0022	16.99	0.05	0.94	48.81	4.650	0.061
	R_{un}	SGEE	0.95	0.96	0.93	0.0577	16.56	0.04	0.68	1.58		
		RSGEE	0.96	0.95	0.96	0.0055	16.88	0.00	0.89	12.57		
		ERSAGE	0.95	0.94	0.95	0.0026	16.90	0.00	0.90	49.72		
		RTGEE	0.95	0.96	0.96	0.0022	17.00	0.04	0.96	62.89	4.655	0.070

Note: \bar{b}_{opt} and $\bar{\lambda}_{opt}$ are the mean values of the selected optimal b_{opt} and λ_{opt} of the proposed method based on 100 simulations.

TABLE 5 Correlated continuous data for large n and diverging p_n ($n = 200$ and $p_n = \lceil 4n^{2/5} \rceil - 5$) with ϵ_{ij} following a normal distribution: Comparison of SGEE, RSGEE, ERSAGEE, and the proposed method RTGEE with three different working correlation matrices (exchangeable, AR(1) and unstructured)

Scenario	R	Method	β_1	β_2	β_3	MMSPE	No. of Zeros			Time (m)	\bar{b}_{opt}	$\bar{\lambda}_{\text{opt}}$
			CI	CI	CI		C	IC	CF			
Case 1	EXC	SGEE	0.98	0.99	0.95	0.0068	22.48	0.03	0.69	14.34		
		RSGEE	0.93	0.95	0.99	0.0060	22.83	0.01	0.90	32.18		
		ERSAGEE	0.96	0.96	0.95	0.0111	22.93	0.02	0.95	187.00		
		RTGEE	0.95	0.98	0.97	0.0050	22.96	0.02	0.96	67.63	6.678	0.061
	AR(1)	SGEE	0.93	0.92	0.98	0.0069	22.47	0.02	0.69	19.35		
		RSGEE	0.92	0.93	0.98	0.0065	22.83	0.02	0.89	38.17		
		ERSAGEE	0.95	0.96	0.93	0.0117	22.94	0.01	0.94	232.90		
		RTGEE	0.93	0.91	0.99	0.0059	22.94	0.01	0.95	75.98	6.721	0.062
	R_{un}	SGEE	0.93	0.96	0.99	0.0039	22.64	0.01	0.76	11.25		
		RSGEE	0.94	0.94	0.99	0.0039	22.98	0.01	0.97	15.20		
		ERSAGEE	0.96	0.96	0.95	0.0079	22.91	0.05	0.94	568.50		
		RTGEE	0.93	0.96	0.99	0.0038	23.00	0.01	0.99	237.56	7.030	0.053
Case 2	EXC	SGEE	0.95	0.99	0.99	0.2582	21.37	0.04	0.36	12.35		
		RSGEE	0.95	0.97	0.97	0.0232	22.69	0.03	0.81	29.82		
		ERGEE	0.93	0.99	0.94	0.0101	22.92	0.05	0.91	182.70		
		RTGEE	0.99	0.99	0.99	0.0078	22.97	0.03	0.94	66.71	5.223	0.062
	AR(1)	SGEE	0.96	0.94	0.95	0.2511	21.17	0.01	0.35	14.25		
		RSGEE	0.93	0.96	0.96	0.0248	22.62	0.04	0.77	36.16		
		ERSAGEE	0.95	0.95	0.95	0.0096	22.95	0.05	0.91	200.60		
		RTGEE	0.94	0.92	0.97	0.0122	22.94	0.03	0.92	76.96	5.145	0.059
	R_{un}	SGEE	0.95	0.93	0.97	0.1980	21.32	0.02	0.41	54.59		
		RSGEE	0.94	0.95	0.96	0.0154	22.99	0.04	0.95	98.30		
		ERSAGEE	0.98	0.95	0.94	0.0088	23.00	0.06	0.94	431.80		
		RTGEE	0.95	0.93	0.95	0.0061	23.00	0.05	0.95	241.43	5.084	0.065
Case 3	EXC	SGEE	0.99	0.99	0.99	0.2307	21.09	0.03	0.23	10.74		
		RSGEE	0.96	0.99	0.93	0.0205	22.63	0.03	0.84	35.54		
		ERSAGEE	0.97	0.94	0.92	0.0094	23.00	0.07	0.93	186.40		
		RTGEE	0.99	0.99	0.96	0.0084	22.99	0.06	0.93	70.07	4.986	0.064
	AR(1)	SGEE	0.95	0.96	0.94	0.2386	21.09	0.04	0.25	16.57		
		RSGEE	0.94	0.95	0.93	0.0200	22.71	0.03	0.82	47.42		
		ERSAGEE	0.94	0.92	0.94	0.0087	23.00	0.06	0.94	254.90		
		RTGEE	0.95	0.92	0.94	0.0083	23.00	0.06	0.94	77.84	4.864	0.065
	R_{un}	SGEE	0.95	0.96	0.95	0.2146	21.07	0.04	0.25	59.36		
		RSGEE	0.94	0.97	0.95	0.0146	22.97	0.03	0.94	107.90		
		ERSAGEE	0.95	0.93	0.93	0.0067	23.00	0.07	0.93	446.10		
		RTGEE	0.95	0.96	0.96	0.0061	23.00	0.04	0.96	245.33	4.754	0.067

Note: \bar{b}_{opt} and $\bar{\lambda}_{\text{opt}}$ are the mean values of the selected optimal b_{opt} and λ_{opt} of the proposed method based on 100 simulations.

TABLE 6 Correlated continuous data for $p_n > n$ ($n = 100$ and $p_n = 300$) with ϵ_{ij} following a normal distribution: Comparison of SGEE, RSGEE, ERSAGE, and the proposed method RTGEE with three different working correlation matrices (exchangeable, AR(1) and unstructured)

Scenario	R	Method	β_1	β_2	β_3	MMSPE	No. of Zeros			Time (m)	\bar{b}_{opt}	$\bar{\lambda}_{opt}$
			CI	CI	CI		C	IC	CF			
Case 1	EXC	SGEE	0.96	0.95	0.95	0.0012	296.51	0.05	0.75	9.64		
		RSGEE	0.94	0.97	0.97	0.0012	296.90	0.03	0.89	18.93		
		ERSAGE	0.96	0.98	0.98	0.0023	297.00	0.02	0.98	184.10		
		RTGEE	0.96	0.95	0.98	0.0011	297.00	0.01	0.99	318.76	3.727	0.036
	AR(1)	SGEE	0.96	0.98	0.98	0.0017	296.36	0.02	0.72	11.79		
		RSGEE	0.97	0.95	0.95	0.0017	296.86	0.05	0.87	23.40		
		ERSAGE	0.93	0.97	0.98	0.0032	297.00	0.02	0.98	209.70		
		RTGEE	0.94	0.97	0.99	0.0016	296.99	0.01	0.98	695.23	3.298	0.032
	R_{un}	SGEE	0.96	0.97	0.97	0.0011	296.55	0.03	0.76	16.16		
		RSGEE	0.95	0.96	0.96	0.0011	296.89	0.04	0.87	25.17		
		ERSAGE	0.96	0.93	0.93	0.0023	297.00	0.07	0.93	282.20		
		RTGEE	0.94	0.96	0.99	0.0015	296.94	0.01	0.95	800.16	3.478	0.044
Case 2''	EXC	SGEE	0.95	0.94	0.98	0.0566	283.38	0.02	0.26	9.72		
		RSGEE	0.94	0.96	0.99	0.0020	296.81	0.01	0.87	21.78		
		ERSAGE	0.94	0.96	0.97	0.0018	296.99	0.03	0.96	167.80		
		RTGEE	0.97	0.95	0.97	0.0014	297.00	0.03	0.97	585.79	3.301	0.108
	AR(1)	SGEE	0.96	0.95	0.98	0.0374	277.70	0.02	0.25	10.11		
		RSGEE	0.96	0.95	0.97	0.0023	296.77	0.03	0.82	23.50		
		ERSAGE	0.96	0.95	0.97	0.0024	296.99	0.03	0.96	194.20		
		RTGEE	0.97	0.94	0.94	0.0019	296.97	0.00	0.98	915.43	2.974	0.041
	R_{un}	SGEE	0.94	0.95	0.97	0.0546	276.36	0.03	0.24	13.99		
		RSGEE	0.93	0.96	0.98	0.0019	296.84	0.02	0.87	25.81		
		ERSAGE	0.94	0.96	0.97	0.0022	296.99	0.03	0.96	184.60		
		RTGEE	0.98	0.95	0.97	0.0018	297.00	0.03	0.97	764.03	2.988	0.043
Case 3''	EXC	SGEE	0.96	0.92	0.91	0.0549	282.37	0.09	0.18	9.23		
		RSGEE	0.97	0.94	0.96	0.0024	296.75	0.04	0.78	20.29		
		ERSAGE	0.93	0.95	0.96	0.0020	296.97	0.04	0.94	171.03		
		RTGEE	0.96	0.96	0.96	0.0016	297.00	0.02	0.98	644.78	3.433	0.047
	AR(1)	SGEE	0.95	0.91	0.90	0.0407	286.46	0.09	0.23	9.66		
		RSGEE	0.96	0.93	0.96	0.0033	296.82	0.04	0.84	20.71		
		ERSAGE	0.97	0.95	0.96	0.0027	296.97	0.04	0.94	178.20		
		RTGEE	0.97	0.95	0.96	0.0022	297.00	0.02	0.98	811.67	3.039	0.045
	R_{un}	SGEE	0.95	0.93	0.90	0.0653	281.18	0.10	0.19	11.97		
		RSGEE	0.95	0.94	0.96	0.0028	296.86	0.04	0.84	21.68		
		ERSAGE	0.93	0.95	0.96	0.0025	296.97	0.04	0.94	195.90		
		RTGEE	0.94	0.96	0.96	0.0020	297.00	0.03	0.97	1114.45	3.047	0.052

Note: \bar{b}_{opt} and $\bar{\lambda}_{opt}$ are the mean values of the selected optimal b_{opt} and λ_{opt} of the proposed method based on 100 simulations.

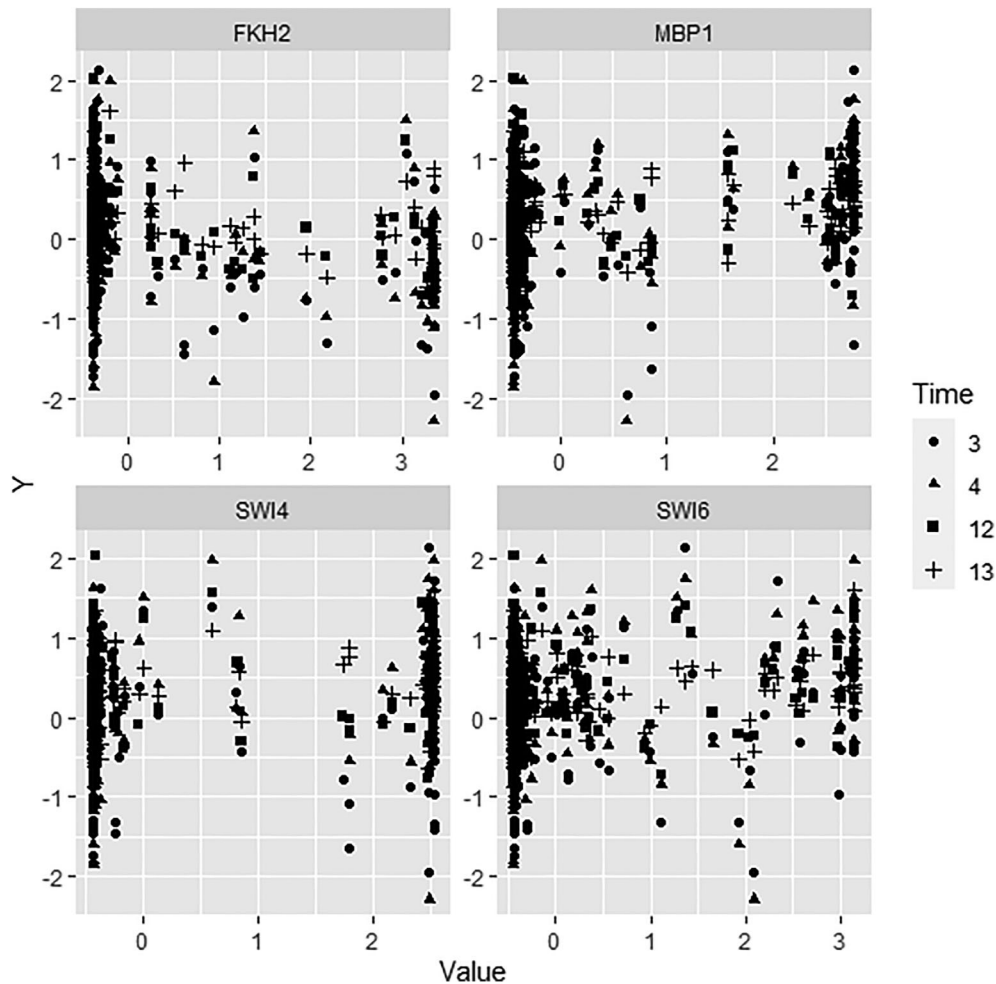


FIGURE 4 The scatter plots of gene expression level vs four important TFs: FKH2, MBP1, SWI4, and SWI6. Here “Y” represents the log-transformed gene expression level

DNA to mRNA during the cell cycle process. We are interested in selecting important TFs from a large set of candidates that are associated with yeast gene expression levels.

We apply the proposed RTGEE method to analyze the yeast cell cycle gene expression dataset, which was mentioned in Section 1. Our investigation indicates that log-transformed gene expression levels and observations of TFs contain many underlying outliers, thus it is worthwhile to reanalyze the yeast cell cycle via robust procedures. In this section, we apply SGEE, RSGEE, ERSAGEE, and RTGEE to the dataset of the G1 stage in a yeast cell cycle with 1132 observations (283 cell-cycled-regularized genes observed over 4-time points). The dataset is available in R package PGEE.

The scatter plot in Figure 4 depicts the complicated functional relationship among gene expression level and TFs, which is highly dependent on varying time, hence we consider following model, which is the same as Wang et al.,⁵

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \sum_{k=1}^{96} \beta_k x_{ik} + \epsilon_{ij}, \quad i = 1, \dots, 283, \quad j = 1, \dots, 4,$$

where the response variable y_{ij} is the log-transformed gene expression level of gene i measured at time point j , the covariates x_{ik} are the matching score of the binding probability of the k th transcription factor on the promoter region of the i th gene for $k = 1, \dots, 96$, and t_{ij} represents the time points. We consider three correlation structures: EXC, AR(1), and R_{un} for ϵ_{ij} . Table 7 summarizes the selected numbers of TFs and the mean squared error for cross validation procedures (MSE_{CV}) to assess the goodness of fit:

$$MSE_{CV} = \frac{1}{n} \sum_{i=1}^n \left\| Y_i - X_i \hat{\beta}_{(-i)} \right\|^2,$$

TABLE 7 The parameter estimates of selected TFs, the mean squared error for cross validation procedures under three correlation structures, and the running time (s means seconds) for four procedures in the yeast cell-cycle process

Covariates	SGEE			RSGEE			ERSGEE			RTGEE		
	EXC	AR(1)	R_{un}	EXC	AR(1)	R_{un}	EXC	AR(1)	R_{un}	EXC	AR(1)	R_{un}
intercept	0.098	0.105	0.068	0.113	0.121	0.098	0.119	0.134	0.099	0.126	0.137	0.087
time	0.010	0.008	0.010	0.007	0.006	0.008	0.006	0.004	0.008	0.006	0.004	0.009
ABF1	-0.048	-0.047	-0.045	0	0	0	0	0	0	0	0	0
ACE2	0.041	0.041	0.045	0	0	0	0	0	0	0	0	0
ASH1	-0.104	-0.094	-0.073	-0.113	-0.107	-0.101	-0.124	-0.123	-0.099	-0.125	-0.117	-0.092
CIN5	0.044	0.048	0.059	0	0	0	0	0	0	0	0	0
CUP9	-0.061	-0.050	-0.028	-0.058	-0.047	-0.042	-0.064	-0.055	-0.045	-0.057	-0.048	-0.035
FKH2	-0.111	-0.106	-0.097	-0.110	-0.102	-0.094	-0.117	-0.108	-0.092	-0.119	-0.106	-0.092
GAL4	-0.035	-0.020	-0.009	0	0	0	0	0	0	0	0	0
GAT3	0.493	0.459	0.436	0.434	0.422	0.385	0.441	0.411	0.389	0.443	0.427	0.399
GCR1	-0.071	-0.068	-0.066	-0.056	-0.056	-0.054	-0.051	-0.053	-0.055	-0.053	-0.056	-0.056
GCR2	-0.098	-0.086	-0.071	0.001	0.011	0.026	0.019	0.029	0.021	0.002	0.009	0.011
GLN3	0.033	0.040	0.049	-0.008	-0.002	0.005	-0.005	-0.000	0.004	-0.011	-0.004	0
GRF10.Pho2	-0.035	-0.035	-0.037	-0.018	-0.016	-0.008	-0.007	-0.001	-0.009	-0.009	-0.009	-0.017
HAP2	-0.179	-0.166	-0.079	-0.525	-0.466	-0.395	-0.454	-0.404	-0.295	-0.641	-0.548	-0.478
HAP3	-0.084	-0.082	-0.080	-0.032	-0.031	-0.029	-0.031	-0.030	-0.029	-0.031	-0.030	-0.029
IME4	0.142	0.134	0.057	0.498	0.444	0.375	0.422	0.377	0.273	0.614	0.527	0.462
IXR1	-0.059	-0.059	-0.060	0	0	0	0	0	0	0	0	0
MAC1	-0.022	-0.018	-0.013	-0.012	-0.010	-0.009	-0.016	-0.015	-0.011	-0.011	-0.009	-0.006
MBP1	0.106	0.099	0.083	0.133	0.125	0.117	0.134	0.124	0.119	0.134	0.126	0.118
MET31	-0.081	-0.080	-0.072	-0.072	-0.069	-0.067	-0.078	-0.079	-0.063	-0.080	-0.076	-0.064
MET4	-0.058	-0.058	-0.062	-0.049	-0.045	-0.030	-0.032	-0.020	-0.026	-0.030	-0.019	-0.043
MTH1	-0.024	-0.023	-0.024	0	0	0	0	0	0	0	0	0
NDD1	-0.100	-0.101	-0.107	-0.089	-0.095	-0.104	-0.073	-0.084	-0.107	-0.072	-0.090	-0.110
NRG1	0.073	0.063	0.040	0.060	0.047	0.041	0.069	0.056	0.044	0.059	0.045	0.032
PDR1	0.150	0.119	0.091	0.109	0.101	0.091	0.124	0.123	0.092	0.114	0.105	0.085
ROX1	0.094	0.089	0.079	0.077	0.072	0.067	0.075	0.072	0.070	0.076	0.075	0.070
RTG3	0.064	0.065	0.067	0	0	0	0	0	0	0	0	0
SRD1	0.056	0.047	0.039	-0.043	-0.050	-0.063	-0.060	-0.067	-0.059	-0.043	-0.047	-0.047
STB1	0.103	0.103	0.102	0.095	0.096	0.100	0.090	0.095	0.096	0.090	0.091	0.096
STP1	0.080	0.076	0.069	0.100	0.099	0.093	0.100	0.098	0.091	0.100	0.096	0.094
SWI4	0.051	0.049	0.045	0.066	0.069	0.073	0.069	0.077	0.070	0.070	0.074	0.068
SWI6	0.064	0.062	0.057	0.040	0.035	0.030	0.039	0.033	0.029	0.039	0.031	0.029
YAP5	-0.477	-0.439	-0.416	-0.406	-0.394	-0.358	-0.416	-0.388	-0.364	-0.415	-0.401	-0.374
YAP6	-0.053	-0.055	-0.061	0	0	0	0	0	0	0	0	0
ZAP1	-0.039	0	0	0	0	0	0	0	0	0	0	0
MSE _{CV}	1.780	1.802	1.847	2.232	2.202	2.186	2.070	2.042	1.969	1.862	1.837	1.858
Time (s)	2.417	2.270	3.604	11.011	11.434	20.251	48.713	49.774	85.775	40.735	40.044	73.160

where $\hat{\beta}_{(-i)}$ is the estimator obtained based on the data excluding the i th subject. The parameter estimates of the selected TFs are also given in Table 7.

The results indicate that the robust methods (RSGEE, ERSGEE, and RTGEE) select 25 TFs, while non-robust method (SGEE) select more TFs, from which, we find that significant TFs such as MBP1, SWI4, and SWI6 are selected by both robust and non-robust methods, which have been reported to function during the G1 stage in Simon et al.²⁴ In addition, TFs such as FKH2, GAT3, GCR2, NDD1, SRD1, STB1 are commonly selected by all of the methods, and they are also confirmed in Wang et al.⁵ ABF1 is selected by SGEE in Wang et al.,⁵ but not by the robust methods. Nevertheless, the transcription factor YAP5 noted as an important factor in Banerjee and Zhang²⁵ is consistently selected by all of the methods considered in our research, but not selected by Wang et al.⁵ Similarly, TFs such as MET31 and GCR1 selected by our robust methods have been verified in Tsai et al.²⁶ and Song et al.²⁷ respectively, while not been found in Wang et al.⁵

In particular, Table 7 presents the mean squared error (MSE_{CV}) and also gives the running time of procedures. For this dataset analysis, our proposed method, RTGEE, performs better than other robust methods with the lower mean squared error under EXC and AR(1), and ERSGEE performs better than RSGEE. All of the methods using R_{un} can be more competitive than other working correlation structures, though it can be more time-consuming. The longer running time of ERSGEE and our procedure RTGEE than other methods attributes to a wide range of tuning parameters used in Section 2.3. To the best of our experience over abundant simulations, we recommend $b = 7.0414$ for RTGEE in this yeast data analysis, which can also lead to a sufficient variable selection and help save a lot of time.

6 | CONCLUSIONS

This article develops a robust automatic variable selection procedure, RTGEE, in the longitudinal marginal models by utilizing the robustness of Tukey's Biweight criterion. A new robust working correlation structure is proposed to take into account the correlations, and is competitive with other misspecified working correlation structures. According to our simulation results, achieving high effectiveness and consistency in robust variable selection, the proposed working correlation can be a substitute for the true correlation structure. The correlation parameters in the proposed correlation matrix depend on the number of the repeated measurements m_i . When m_i is large compared to the sample size, the accuracy of the correlation matrix estimation will decrease, and the computation time will increase. Hence, the number of the repeated measurements cannot be too large. If $m = \max_i \{m_i\}$ is diverging or larger than the sample size, the computation and the theorems need to be restructured, which is left for the future work. We apply smooth-threshold estimating equations to select the important variables. This approach is conceptually simple, easy to implement, and does not need penalty functions. Furthermore, this approach eliminates the irrelevant parameters by shrinking them to zero and simultaneously estimates the nonzero coefficients. Previous researchers have proposed similar robust smooth-threshold estimating equations.^{9,10} Nevertheless, the robustness of our method is still competitive regardless of whether there are more severe or moderate setting conditions. Our numerical studies conclude that our proposed method is robust for both response variables and covariates in longitudinal marginal models. It is incredibly competitive under a heavy-tailed distribution, and it has broad prospects when the dimension of covariates is larger than the sample size.

Robust variable selection for ultrahigh-dimensional data is gaining more traction in the biomedical area, and in future research, our proposed method can be extended to cases where the dimension of covariates is in the exponential order of the sample size. However, some guiding theoretical research for parameter selection criterion needs to be conducted when applying our procedure to ultrahigh-dimensional data.

ACKNOWLEDGEMENTS

The authors thank the Associate Editor and referees for their constructive comments. This research was supported by the Natural Science Foundation of China (No. 11871390), and the Natural Science Basic Research Plan in Shaanxi Province of China (No. 2018JQ1006), the Australian Research Council Discovery Project (DP160104292).

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The authors confirm that the data supporting the findings of this study are available within the article and/or its supplementary materials.

ORCID

Liya Fu  <https://orcid.org/0000-0002-6789-442X>

You-Gan Wang  <https://orcid.org/0000-0003-0901-4671>

REFERENCES

1. Pan W. Akaike's information criterion in generalized estimating equations. *Biometrics*. 2001;57:120-125.
2. Wang L, Qu A. Consistent model selection and data-driven smooth tests for longitudinal data in the estimating equations approach. *JR Stat Soc Ser B*. 2009;71:177-190.
3. Tian R, Xue L, Liu C. Penalized quadratic inference functions for semiparametric varying coefficient partially linear models with longitudinal data. *J Multivar Anal*. 2014;132:94-110.
4. Li GR, Lian H, Feng SY, Zhu LX. Automatic variable selection for longitudinal generalized linear models. *Comput Stat Data Anal*. 2013;61:174-186.
5. Wang L, Zhou JH, Qu A. Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics*. 2012;68:353-360.
6. Spellman PT, Sherlock G, Zhang MQ, et al. Comprehensive identification of cell cycle-regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell*. 1998;9:3273-3297.
7. Luan YH, Li HZ. Clustering of time-course gene expression data using a mixed-effects model with B-splines. *Bioinformatics*. 2003;17:474-482.
8. Wang L, Chen G, Li H. Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics*. 2007;23:1486-1494.
9. Fan YL, Qin GY, Zhu ZY. Variable selection in robust regression models for longitudinal data. *J Multivar Anal*. 2012;109:156-167.
10. Lv J, Yang H, Guo CH. An efficient and robust variable selection method for longitudinal generalized linear models. *Comput Stat Data Anal*. 2015;82:74-88.
11. Wang XQ, Jiang YL, Huang M, Zhang HP. Robust variable selection with exponential squared loss. *J Am Stat Assoc*. 2013;108:632-643.
12. Chang L, Roberts S, Welsh A. Robust lasso regression using Tukey's biweight criterion. *Technometrics*. 2018;60:36-47.
13. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986;73:13-22.
14. Ueki M. A note on automatic variable selection using smooth-threshold estimating equations. *Biometrics*. 2009;96:1005-1011.
15. Zou H. The adaptive lasso and its oracle properties. *J Am Stat Assoc*. 2006;101:1418-1429.
16. Wang YG, Lin X, Zhu M. Robust estimation functions and bias correction for longitudinal data analysis. *Biometrics*. 2005;61:684-691.
17. Yohai VJ. High breakdown-point and high efficiency robust estimates for regression. *Ann Stat*. 1987;15:642-656.
18. Portnoy S. Asymptotic behavior of M estimators of p regression parameters when p^2/n is large. II. normal approximation. *Ann Stat*. 1985;13:1403-1417.
19. He XM, Fung WK, Zhu ZY. Robust estimation in generalized partial linear models for clustered data. *J Amer Stat Assoc*. 2005;100(472):1176-1184.
20. Wang L. GEE analysis of clustered binary data with diverging number of covariates. *Ann Stat*. 2011;39:389-417.
21. Fan JQ, Li RZ. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc*. 2001;96:1348-1360.
22. Riani M, Cerioli A, Torti F. On consistency factors and efficiency of robust S-estimators. *TEST*. 2014;23:356-387.
23. Rousseeuw PJ, Leroy AM. *Robust Regression and Outlier Detection*. New York, NY: Wiley; 1987.
24. Simon I, Barnett J, Hannett N, et al. Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell*. 2001;106:697-708.
25. Banerjee N, Zhang MQ. Identifying cooperativity among transcription factors controlling the cell cycle in yeast. *Nucleic Acids Res*. 2003;31:7024-7031.
26. Tsai HK, Lu HHS, Li WH. Statistical methods for identifying yeast cell cycle transcription factors. *Proc Natl Acad Sci*. 2005;102:13532-13537.
27. Song R, Yi F, Zou H. On varying-coefficient independence screening for high-dimensional varying-coefficient models. *Stat Sin*. 2014;24:1735-1752.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Fu L, Li J, Wang Y-G. Robust approach for variable selection with high dimensional longitudinal data analysis. *Statistics in Medicine*. 2021;40(30):6835-6854. doi: 10.1002/sim.9213