Robust statistical inference for longitudinal data with nonignorable dropouts

Yujing Shao, Wei Ma and Lei Wang

School of Statistics and Data Science & LPMC, Nankai University, Tianjin, People’s Republic of China

ABSTRACT
In this paper, we propose robust statistical inference and variable selection method for generalized linear models that accommodate the outliers, nonignorable dropouts and within-subject correlations. The purpose of our study is threefold. First, we construct the robust and bias-corrected generalized estimating equations (GEEs) by combining the Mallows-type weights, Huber’s score function and inverse probability weighting approaches to against the influence of outliers and account for nonignorable dropouts. Subsequently, the generalized method of moments is utilized to estimate the parameters in the nonignorable dropout propensity based on sufficient instrumental estimating equations. Second, in order to incorporate the within-subject correlations under an informative working correlation structure, we borrow the idea of quadratic inference function and hybrid-GEE to obtain the improved empirical likelihood procedures. The asymptotic properties of the proposed estimators and their confidence regions are derived. Third, the robust variable selection and algorithm are investigated. We evaluate the performance of proposed estimators through simulation and illustrate our method in an application to HIV-CD4 data.

1. Introduction
Longitudinal data arises frequently in medicine, population health, biological researches, economics, social sciences and so on. The observations are often collected repeatedly from every sampled subject at many time points, and thus intrinsically correlated. Our study is motivated by the following longitudinal data from the AIDS Clinical Trial Group 193A (ACTG 193A), which was a study of HIV-AIDS patients with advanced immune suppression. The data set can be accessed at http://www.hsph.harvard.edu/fitzmaur/ala/cd4.txt. For this HIV clinical trial, the CD4 cell count is of prime interest which decreases as HIV progresses. After the treatments were applied, the CD4 cell count was scheduled to be collected from each patient in every 8 weeks. However, because of adverse events, low-grade toxic reactions, the desire to seek other therapies, death and some other reasons, the
dropout rates of the first four follow-up times are 31.5%, 42.4%, 55.4% and 65.3%, respectively. Previous experiences from doctors indicate that a steep decline in the CD4 cell count indicates the disease progression, and patients with low CD4 cell counts are more likely to drop out from the scheduled study visits as compared to patients with normal CD4. Therefore, nonresponse of the CD4 cell count is likely related to itself and is nonignorable [1]. Our purpose is to examine whether the CD4 counts of young patients are more likely to decrease.

For the longitudinal data, a major aspect is how to take into account the within-subject correlation structure to improve estimation efficiency. A naive and simple way is to use a working independent model [2,3], but it ignores the correlation structure and may lose efficiency when strong correlations exist. In order to incorporate the within-subject correlation, Liang and Zeger [4] proposed generalized estimating equations (GEEs) through a working correlation matrix. However, two challenges related to GEEs have not been solved yet. First, it is difficult to describe and specify the underlying within-subject covariance matrix, which is unknown in practice and suffers from the positive-definiteness constraint. Second, the GEEs method is closely related to the weighted least squares method, which is not robust when the observed data contains outliers, such that the resulting estimators may be biased and their confidence regions may be greatly lengthened in the direction of the outliers [5]. To solve the first issue, Huang et al. [6], Bai et al. [7], Fu and Wang [8] approximated the covariance matrices with basis functions or possible dependence structures. Leung et al. [9] proposed a hybrid-GEE method that combines multiple GEEs based on different working correlation matrices. Leng et al. [10], Zhang and Leng [11], Zhang et al. [12] and Lv et al. [13] applied the Cholesky decomposition to obtain the within-subject covariance matrix. Li and Pan [14] and Leng and Zhang [15] constructed estimating functions by the quadratic inference function (QIF). Xu et al. [16] proposed a combined multiple likelihood estimating procedure based on three dynamic covariance models. Among these methods, the QIF and hybrid-GEE approaches have recently received considerable attention due to their flexibility and simplicity [17,18]. For the second challenge, the traditional robust M-estimation for longitudinal data have been discussed, by using the Mallows-type weights to downweight the effect of leverage points and adopting the Huber’s score function on the Pearson residuals to dampen the effect of outliers; see He et al. [19], Wang et al. [20], Qin and Zhu [21], Qin et al. [22], Fan et al. [23], Zheng et al. [24] and so on.

Moreover, the CD4 cell count drops out of the study prior to the end of the study. It is well known that under nonignorable missing responses, the complete case analysis can not be trusted. The majority of existing methods only naturally accommodate ignorable dropout or missing at random (MAR; [25]) assumption. For the ACTG 193A data, however, the dropout is nonignorable or missing not at random (MNAR) such that developing valid methodologies for statistical analysis with nonignorable dropout is challenging [26]. In this case, the population parameters are, in general, not identifiable [27] if there is no assumption imposed, and estimators based on the assumption of ignorable dropouts may have large biases. Thus, methods which are very different from those for an ignorable propensity have to be applied; for example, see Molenberghs and Kenward [28], Kim and Yu [29], Wang et al. [30], Shao and Wang [31], Bindele and Zhao [32] and so on. Furthermore, when the dimension of covariates is high, the generalized linear models (GLMs) often include
many irrelevant covariates. In practice, it is useful to find which covariates are relevant for prediction, both for better interpretation of the model and for better efficiency of the estimators [33].

The existence of nonignorable dropouts and outliers and their impacts on the inference motivate us to find a robust and efficient estimation approach. To the best of our knowledge, this problem has not previously been investigated. Our contributions of this paper are in three aspects.

1) We impose a parametric model on the dropout propensity and then use a nonresponse instrument, which is a useful covariate vector that can be excluded from the propensity given the response and other covariates, to deal with the identifiability issue [26,30]. The main technique is to create sufficient estimating equations to estimate the parametric propensity based on a given instrument. In specific, we apply the generalized method of moments (GMM; [34]) to estimate the propensity.

2) Once the propensity is estimated, we propose the improved robust and bias-corrected GEEs in the presence of outliers and nonignorable dropouts through the following three steps. First, in order to against the influence of outliers, the Huber’s score function and Mallows-type weights are applied to build the robust GEEs. Second, the bias-corrected GEEs are constructed by the inverse propensity weighting (IPW; [35]) to account for nonignorable dropouts. Third, in conjunction with the quadratic inference function (QIF; [36]) and hybrid-GEE [9] methods, we construct two classes of improved estimators that can incorporate the within-subject correlations under an informative working correlation structure. In specific, the proposed QIF procedure is based on the matrix expansion idea, which neither assumes the exact knowledge of the within-subject covariance matrix nor estimates the parameters of the within-subject covariance matrix. Alternatively, the hybrid-GEE method combines multiple GEEs based on different working correlation models to improve the estimation efficiency. The resulting two robust EL ratios are two different weighted sum of chi-square random variables asymptotically, which can be used to construct two corresponding confidence regions, respectively.

3) For the robust variable selection, we propose the penalized robust EL approach by combining the profile robust EL and the smoothly clipped absolute deviation (SCAD; [37]) methods together in Section 3. We show that the proposed variable selection method can efficiently select significant variables and estimate parameters simultaneously. Furthermore, the resulting estimators based on the QIF and hybrid-GEE methods are consistent and have the oracle property. In addition, an algorithm based on the local quadratic approximation (LQA) is proposed.

The rest of this paper is organized as follows. We propose the improved robust estimators and variable selection approaches based on the QIF and hybrid-GEE methods in Section 2. The asymptotic theories are investigated in Section 3. Simulation studies are given in Sections 4 and 5 analyzes the ACTG 193A data for illustration. Some discussions can be found in Section 6. All technical details and additional simulation results are provided in the Supplementary Material.
2. Methodology

2.1. Robust GEEs

To illustrate our proposed methods, we first review the robust GEEs without missing data. Let \( y_i = (y_{i1}, y_{i2}, \ldots, y_{im_i})^T \) be a \( m_i \)-dimensional vector of the \( i \)th subject’s response which may drop out, \( x_i = (x_{i1}, \ldots, x_{im_i})^T \) be an always observed \((m_i \times p)\)-dimensional matrix of covariates associated with \( y_i, i = 1, \ldots, n \) and \( j = 1, \ldots, m_i \). Here, \( m_i \) is called as the size of the \( i \)th subject. We consider that \( y_{ij} \) are modelled by

\[
g(\mu_{ij}) = x_i^T \beta, \quad \text{Var}(y_{ij}) = \phi v(\mu_{ij}),
\]

where \( \beta \) is a \( p \)-dimensional parameter vector, \( g(\cdot) \) is a known link function, \( \mu_{ij} = E(y_{ij}) \), \( \phi \) is a dispersion parameter, \( v(\cdot) \) is a known variance function and \( a^T \) is the transpose of \( a \).

Without loss of generality, we consider the balanced data with the same subject size \( m_i = m \) and the unbalanced situation will be discussed in Section 6 later. Motivated by Lv et al. [38] and Qin et al. [39], denote \( \mu_i = (\mu_{i1}, \ldots, \mu_{im_i})^T \) and define the core robust estimating equations

\[
T_i(\mu_i(\beta), \phi) = W_i \{ \psi(A_i^{-1/2}(\phi)(y_i - \mu_i)) \} - C_i,
\]

where \( W_i = \text{diag}(w_{ij}, j = 1, 2, \ldots, m) \) is a weight matrix to bound the effect of leverage points in the covariate space, \( A_i(\phi) = \text{diag}(\text{Var}(y_{i1}), \ldots, \text{Var}(y_{im_i})) = \phi \text{diag}(v(\mu_{i1}), \ldots, v(\mu_{im_i})) \), \( \psi(\cdot) \) is a bounded function to downweight the influence of outliers, and the correction term \( C_i = E[\psi(A_i^{-1/2}(\phi)(y_i - \mu_i))] \) is used to ensure Fisher consistency of the estimator. In specific, \( \psi(\cdot) \) is the Huber’s score function \( \psi(x) = \min\{c, \max\{-c, x\}\} \), where the tuning constant \( c \) is used to balance the efficiency and robustness of the estimator. While, the weight \( w_{ij} \) is considered based on the Mahalanobis distance [40] as follows:

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

where \( b_0 \) is the 0.95 quantile of the chi-square distribution with the degree of freedom equal to the dimension of \( x_{ij}, m_x \) and \( S_x \) are robust estimates of location and scale of \( x_{ij} \), such as minimum volume ellipsoid (MVE) estimates of Rousseeuw and van Zomeren [41], Lv et al. [38] and Qin et al. [39]. By using the Huber’s score function and weight matrix, the influence of the outliers will be reduced.

2.2. Robust and bias-corrected GEEs

Motivated by the ACTG 193A data, let \( r_i = (r_{i1}, r_{i2}, \ldots, r_{im_i})^T \) be the vector of response indicators, where \( r_{ij} = 1 \) if \( y_{ij} \) is observed and \( r_{ij} = 0 \) if \( y_{ij}, \ldots, y_{im_i} \) are not observed, and our interest is to estimate the unknown \( \beta \) based on \( (y_i, x_i, r_i) \) for \( i = 1, \ldots, n \). To take into account the nonignorable dropouts, define \( \pi_{ij} = \text{Pr}(r_{ij} = 1 | x_i, y_i) \). Subsequently, we
propose the robust and bias-corrected GEEs as follows:

\[
\sum_{i=1}^{n} D_i^T V_i^{-1} S_i T_i(\mu_i(\beta), \phi) = 0,
\]

where \( D_i = \partial \mu_i / \partial \beta \), \( V_i \) is the covariance matrix of \((y_i - \mu_i)\) and \( S_i = \text{diag}(r_{ij} / \pi_{ij}, j = 1, 2, \ldots, m) \). Moreover, it can be verified that the inverse of covariance matrix \( V_i^{-1} \) can be decomposed as \( A_i^{-1/2}(\phi) \Phi^{-1} A_i^{-1/2}(\phi) \) with \( \Phi \) being an \((m \times m)\)-dimensional true correlation matrix. Unfortunately, \( \Phi \) is unknown in practice and one has to use a working correlation matrix \( R \). Some common working correlation structures include independent structure, compound symmetry (CS) and first-order autoregressive (AR(1)). Then, the above equations can be written as

\[
\sum_{i=1}^{n} D_i^T A_i^{-1/2}(\phi) R^{-1} A_i^{-1/2}(\phi) S_i T_i(\mu_i(\beta), \phi) = 0.
\]

To obtain the estimator of \( \beta \), we first need to get the consistent estimators of \( \pi_{ij} \) and \( \phi \), denoted as \( \hat{\pi}_{ij} \) and \( \hat{\phi} \), respectively. Using the identifiability and estimation approaches proposed in Wang et al. [26], we assume that \( x_{ij} \) can be decomposed as two parts \( u_{ij} \) and \( z_{ij} \), where \( u_{ij} \) is continuously distributed and \( z_{ij} \) can be continuous, discrete, or mixed. Denote \( \overrightarrow{y}_{ij} = (y_{i1}, \ldots, y_{ij})^T \), \( \overrightarrow{u}_{ij} = (u_{i1}^T, \ldots, u_{ij}^T)^T \) and \( \overrightarrow{z}_{ij} = (z_{i1}^T, \ldots, z_{ij}^T)^T \) as the histories of \( y_{ij} \), \( u_{ij} \) and \( z_{ij} \) up to and including cycle \( j \), respectively. According to Diggle and Kenward [42], we impose a parametric model on the propensity as follows:

\[
\text{Pr}(r_{ij} = 1 | r_{i(j-1)} = 1, x_{ij}, y_{ij}) = \text{Pr}(r_{ij} = 1 | r_{i(j-1)} = 1, \overrightarrow{u}_{ij}, \overrightarrow{y}_{ij}) = \Psi(\alpha_j + y_j^T \mathcal{O}_{ij}), \quad j = 1, \ldots, m,
\]

where \( \mathcal{O}_{ij} = (\overrightarrow{u}_{ij}^T, \overrightarrow{y}_{ij})^T \), \( \xi_j = (\alpha_j, y_j^T) \) is a vector of unknown parameters, \( \Psi \) is a known monotone function and \( r_{i0} \) is always defined to be 1. Given \( \overrightarrow{y}_{ij} \) and \( \overrightarrow{u}_{ij} \), the instruments \( \overrightarrow{z}_{ij} \) can be excluded from the nonresponse propensity, which are used to create sufficient estimation equations for estimating the propensity [26,30].

Under the model (2), \( \pi_{ij} = \Pr(r_{ij} = 1 | \overrightarrow{u}_{ij}, \overrightarrow{y}_{ij}) = \prod_{t=1}^{j} \Pr(r_{it} = 1 | r_{i(t-1)} = 1, \overrightarrow{u}_{it}, \overrightarrow{y}_{it}) = \prod_{t=1}^{j} \Psi(\gamma_{i1} + y_{i2}^T \mathcal{O}_{it}) \equiv \pi_{ij}(\theta_j), \) where \( \theta_j = (\xi_j^T, \ldots, \xi_m^T)^T \) for \( j = 1, \ldots, m \) and \( \theta = (\xi_1^T, \ldots, \xi_m^T)^T \). Given a consistent estimator \( \hat{\theta}_j \) of \( \theta_j \), we have \( \hat{\pi}_{ij} = \pi_{ij}(\hat{\theta}_j) = \prod_{t=1}^{j} \Psi(\gamma_{i1} + y_{i2}^T \mathcal{O}_{it}) \). Similar to He et al. [19] and Wang et al. [20], a consistent estimator of \( \phi \) can be obtained as

\[
\hat{\phi} = \left\{ 1.483 \text{median}\left[ |r_{ij}\hat{\pi}_{ij}^{-1}(y_{ij} - \mu_{ij}(\beta^*)) - \text{median}\{r_{ij}\hat{\pi}_{ij}^{-1}(y_{ij} - \mu_{ij}(\beta^*))\}| \right] \right\}^2,
\]

where \( \beta^* \) is an initial estimate of \( \beta \) by using the MM method, i.e., a robust and efficient estimator in the presence of outliers [43,44].

Once we have \( \hat{\pi}_{ij} \) and \( \hat{\phi} \), the unknown \( A_i^{-1/2}(\phi) \), \( S_i(\theta) \) and \( T_i(\mu_i(\beta), \phi) \) can be estimated by \( \hat{A}_i^{-1/2} = A_i^{-1/2}(\hat{\phi}) \), \( \hat{S}_i = \text{diag}(r_{ij} / \pi_{ij}(\hat{\theta}_j), j = 1, 2, \ldots, m) \) and \( \hat{T}_i(\mu_i(\beta)) = \).
Let \( T_i(\mu_i(\beta), \hat{\phi}) \), respectively. Then, we define the robust and unbiased estimating equations for \( \beta \) as follows:

\[
\hat{U}_n(\beta) = \hat{U}_n(\beta, \hat{\theta}, \hat{\phi}) = \sum_{i=1}^{n} D_i^T \hat{\Lambda}_i^{-1/2} R^{-1} \hat{\Lambda}_i^{-1/2} \hat{S}_i \hat{T}_i(\mu_i(\beta)) = 0.
\]

### 2.3. Robust EL inference based on QIF and hybrid-GEE

In order to improve the efficiency, we borrow the matrix expansion idea of Qu et al. [36] and propose the quadratic inference function (QIF) by assuming that the inverse of the working correlation matrix \( R^{-1} \) can be approximated by a linear combination of several basis matrices, that is

\[
R^{-1} = \sum_{k=1}^{l} b_k B_k,
\]

where \( B_k \in \mathbb{R}^{m \times m}, \ k = 1, 2, \ldots, l, \) are symmetric basic matrices depending on the particular choice of \( R^{-1} \) and \( b_1, \ldots, b_l \) are unknown coefficients. For example, if a working correlation structure is CS, then \( R^{-1} = b_1 B_1 + b_2 B_2 \) with an identity matrix \( B_1 \) and a symmetric matrix \( B_2 \) having 0 on the diagonal and 1 elsewhere. If \( R \) corresponds to AR(1), \( R^{-1} = b_1 B_1 + b_2 B_2 + b_3 B_3 \) with an identity matrix \( B_1 \), a symmetric matrix \( B_2 \) having 1 on the sub-diagonal entries and 0 elsewhere, and a symmetric matrix \( B_3 \) having 1 in elements \((1, 1)\) and \((m, m)\), and 0 elsewhere. More details can be found in Qu et al. [36] and Cho and Qu [17]. Therefore, by substituting (5) into (4), it leads to

\[
\sum_{i=1}^{n} D_i^T \hat{\Lambda}_i^{-1/2} (b_1 B_1 + \cdots + b_l B_l) \hat{\Lambda}_i^{-1/2} \hat{S}_i \hat{T}_i(\mu_i(\beta)) = 0. \tag{6}
\]

Consequently, the equations (6) can be approximated as a linear combination of elements \( \hat{g}_i(\beta) \) for \( i = 1, \ldots, n \), where

\[
\hat{g}_i(\beta) = \begin{pmatrix} D_i^T \hat{\Lambda}_i^{-1/2} B_1 \hat{\Lambda}_i^{-1/2} \hat{S}_i \hat{T}_i(\mu_i(\beta)) \\ D_i^T \hat{\Lambda}_i^{-1/2} B_2 \hat{\Lambda}_i^{-1/2} \hat{S}_i \hat{T}_i(\mu_i(\beta)) \\ \vdots \\ D_i^T \hat{\Lambda}_i^{-1/2} B_l \hat{\Lambda}_i^{-1/2} \hat{S}_i \hat{T}_i(\mu_i(\beta)) \end{pmatrix}. \tag{7}
\]

Notice that the functions \( \hat{g}_i(\beta) \) do not involve the parameters \( b_1, \ldots, b_l \), such that the estimation of these parameters is not required and \( \hat{g}_i(\beta) \) are the overdetermined estimating equations with \( pl \) functions. Thus, the EL method is proposed for the inference of \( \beta \) under some regular conditions. Let \( p_i \) represent the probability weight allocated to \( \hat{g}_i(\beta) \) and the empirical log-likelihood ratio function is defined as follows:

\[
\hat{R}_Q(\beta) = -2 \sup \left\{ \sum_{i=1}^{n} \log (np_i) : p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i \hat{g}_i(\beta) = 0 \right\}.
\]

The maximum EL estimator based on \( \hat{g}_i(\beta) \), denoted as \( \hat{\beta}_Q \), can be obtained by

\[
\hat{\beta}_Q = \arg \min_{\beta} \hat{R}_Q(\beta).
\]
Alternatively, Liang and Zeger [4] considered $R = R(\alpha)$ with unknown nuisance parameter $\alpha$ and Leung et al. [9] proposed a hybrid method that combines multiple GEEs based on different and linearly independent choices of $R(\alpha)$, say $R_k(\alpha)$, $k = 1, 2, \ldots, q$, to improve the efficiency as follows:

$$
\hat{h}_i(\beta) = \left( D_i^T \hat{A}_i^{-1/2} R_1^{-1}(\alpha) \hat{A}_i^{-1/2} S_i \hat{T}_i(\mu_i(\beta)) \right). 
$$

(8)

In practice, a few popular choices of $R(\alpha)$ may be used, i.e., CS and AR(1), and different values of $\alpha$ are specified by Leung et al. [9] to model the degree of intra-subject correlation. Similarly, the corresponding empirical log-likelihood ratio function is defined as follows:

$$
\hat{R}_H(\beta) = -2 \sup \left\{ \sum_{i=1}^n \log \left( np_i \right) : p_i \geq 0, \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i \hat{h}_i(\beta) = 0 \right\},
$$

and the maximum EL estimator based on $\hat{h}_i(\beta)$, denoted as $\hat{\beta}_H$, is obtained by

$$
\hat{\beta}_H = \arg \min_{\beta} \hat{R}_H(\beta).
$$

**Algorithm 1** Robust EL estimation

1: Obtain the consistent estimate of $\hat{\vartheta}_j$ by (11) for $j = 1, \ldots, m$.

2: Calculate the estimated $\hat{\pi}_ij(\hat{\vartheta}_j)$ and $\hat{\phi}$ by (2) and (3). Obtain $\hat{A}_i^{-1/2} = A_i^{-1/2}(\hat{\phi})$, $\hat{S}_i = \text{diag}\{r_{ij}/\pi_{ij}(\hat{\vartheta}_j), j = 1, 2, \ldots, m\}$ and $\hat{T}_i(\mu_i(\beta)) = T_i(\mu_i(\beta), \hat{\phi})$.

3: Construct functions $\hat{g}_i(\beta)$ and $\hat{h}_i(\beta)$ in (7) and (8), respectively, based on the QIF and hybrid-GEE.

4: Obtain the maximum EL estimators $\hat{\beta}_Q$ and $\hat{\beta}_H$ from $\hat{R}_Q(\beta)$ and $\hat{R}_H(\beta)$, respectively.

**2.4. Variable selection**

When the dimension of covariates is high, we attend to identify the zero coefficients consistently and estimate the nonzero coefficients efficiently and robustly. Therefore, we use penalized empirical likelihood (PEL) to do variable selection as follows:

$$
\hat{R}_p(\beta) = 2 \sum_{i=1}^n \log \left\{ 1 + \lambda^T \hat{\eta}_i(\beta) \right\} + n \sum_{j=1}^p p_\nu(\| \beta_j \|),
$$

(9)

where $\hat{\eta}$ can be $\hat{g}$ and $\hat{h}$, $p_\nu(\cdot)$ is a penalty function with tuning parameter $\nu$ which can shrink small components of coefficients to zero. In this paper, we use the SCAD penalty,
while other choices of penalty could possibly also be entertained. The first derivative of the SCAD is

\[ p'_v(t) = v \left\{ I(t \leq v) + \frac{(av - t)_+}{(a - 1)v} I(t > v) \right\}, \quad t > 0, \]

where \( a > 2 \) and \( v > 0 \) are tuning parameters. As suggested by Fan and Li [37], \( a = 3.7 \) is recommended. To choose the optimal tuning parameter \( v \), three information criteria: BIC of Schwarz [45], BICC of Wang et al. [46] and EBIC of Chen and Chen [47], are considered as follows:

\[
\begin{align*}
BIC(v) &= -2\hat{R}_p(\beta_v) + \log(n)df_v, \\
BICC(v) &= -2\hat{R}_p(\beta_v) + \max\{1, \log \log(p)\} \log(n)df_v, \\
EBIC(v) &= -2\hat{R}_p(\beta_v) + \log(n) + 2 \log(p)df_v,
\end{align*}
\]

where \( \beta_v \) is the estimate of \( \beta \) based on the QIF or hybrid-GEE methods with \( v \) being the tuning parameter, and \( df_v \) is the number of nonzero coefficients in \( \beta_v \).

### 3. Asymptotic theories

Denote \( \hat{g}_i(\beta) = g_i(\beta, \hat{\theta}, \hat{\phi}) \), \( \hat{h}_i(\beta) = h_i(\beta, \hat{\theta}, \hat{\phi}) \), \( g_i(\beta, \theta_0) = g_i(\beta, \theta_0, \phi_0) \) and \( h_i(\beta, \theta_0) = h_i(\beta, \theta_0, \phi_0) \), where \( \beta_0 \) and \( \theta_0 \) are the true values of \( \beta \) and \( \theta \) respectively. Define \( \Delta_g = E[\partial g_i(\beta_0, \theta_0)/\partial \beta] \), \( \Lambda_g = E[g_i(\beta_0, \theta_0)g_i^T(\beta_0, \theta_0)] \) and \( \Gamma_g = \{\Delta_g^T \Lambda_g^{-1} \Delta_g\}^{-1} \Delta_g^T \Lambda_g^{-1}; \Delta_h = E[\partial h_i(\beta_0, \theta_0)/\partial \beta] \), \( \Lambda_h = E[h_i(\beta_0, \theta_0)h_i^T(\beta_0, \theta_0)] \) and \( \Gamma_h = \{\Delta_h^T \Lambda_h^{-1} \Delta_h\}^{-1} \Delta_h^T \Lambda_h^{-1}. \)

Denote \( H_g = E[\partial g_i(\beta_0, \theta_0)/\partial \theta] \) and \( H_h = E[\partial h_i(\beta_0, \theta_0)/\partial \theta] \).

**Theorem 3.1:** Assume that the propensity model (2) and the conditions (C1)–(C7) in the Supplementary Material hold, as \( n \to \infty \), we have

\[
\sqrt{n}(\hat{\beta}_Q - \beta_0) \to N(0, \Gamma_g \Sigma_g \Gamma_g^T), \quad \sqrt{n}(\hat{\beta}_h - \beta_0) \to N(0, \Gamma_h \Sigma_h \Gamma_h^T),
\]

in distribution, where \( \Sigma_g = \Lambda_g + H_g \Sigma H_g^T \) and \( \Sigma_h = \Lambda_h + H_h \Sigma H_h^T \).

**Remark 3.1:** If \( \pi_{ij} \) is known, we have \( H_g = H_h = 0 \), then the asymptotic covariance matrices of \( \hat{\beta}_Q \) and \( \hat{\beta}_h \) can be simplified as \( \{\Delta_g^T \Lambda_g^{-1} \Delta_g\}^{-1} \) and \( \{\Delta_h^T \Lambda_h^{-1} \Delta_h\}^{-1} \), respectively.

**Remark 3.2:** By the plug-in method [48], we can obtain

\[
\begin{align*}
\hat{\Sigma}_g &= \hat{\Lambda}_g + \hat{H}_g \hat{\Sigma} \hat{H}_g^T, \quad \hat{\Lambda}_g = \frac{1}{n} \sum_{i=1}^n \hat{g}_i(\hat{\beta}_Q, \hat{\theta}) \hat{g}_i^T(\hat{\beta}_Q, \hat{\theta}), \\
\hat{\Lambda}_g &= \frac{1}{n} \sum_{i=1}^n \partial \hat{g}_i(\hat{\beta}_Q, \hat{\theta})/\partial \beta, \\
\hat{\Gamma}_g &= \{\Delta_g^T \Lambda_g^{-1} \Delta_g\}^{-1} \Delta_g^T \Lambda_g^{-1}, \quad \hat{H}_g = \frac{1}{n} \sum_{i=1}^n \partial \hat{g}_i(\hat{\beta}_Q, \hat{\theta})/\partial \theta,
\end{align*}
\]
\[
\hat{Y}_j = \frac{1}{n} \sum_{i=1}^{n} \partial s_j(y_i, x_i, r_i, \hat{\xi}_j) / \partial \xi, \quad \hat{\Omega}_j = \frac{1}{n} \sum_{i=1}^{n} s_j(y_i, x_i, r_i, \hat{\xi}_j) s_j(y_i, x_i, r_i, \hat{\xi}_j)^T, \\
\hat{\Sigma} = \text{diag}((\mathbf{Y}_j^T \mathbf{\Omega}_j \mathbf{Y}_j)^{-1}, j = 1, \ldots, m).
\]

It can be proved that \( \hat{\Sigma}_g \) and \( \hat{\Gamma}_g \) are consistent estimators of \( \Sigma_g \) and \( \Gamma_g \), respectively. Once we have \( \hat{\Sigma}_g \) and \( \hat{\Gamma}_g \), the variance matrix \( \Gamma_g \Sigma_g \Gamma^T_g \) can be estimated by \( \hat{\Gamma}_g \hat{\Sigma}_g \hat{\Gamma}_g^T \). Similarly, we can obtain the variance matrix estimator for \( \hat{\beta}_H \). Hence, Theorem 3.1 can be used to construct a normal-approximation-based confidence region.

As in Theorems 2 and 4 of Tang et al. [48] and many others, for simplicity we only need to study the asymptotic properties of \( \hat{R}_Q(\beta_0) \) and \( \hat{R}_H(\beta_0) \). Compared to the standard empirical log-likelihood ratio without missing data, the main difference is that both \( \hat{g}_i(\beta) \) and \( \hat{h}_i(\beta) \), \( i = 1, \ldots, n \), are not independent and identically distributed. Hence, the asymptotic distributions of \( \hat{R}_Q(\beta_0) \) and \( \hat{R}_H(\beta_0) \) may not be standard chi-squares. Actually, we will show that \( \hat{R}_Q(\beta_0) \) and \( \hat{R}_H(\beta_0) \) are two different weighted sum of chi-square random variables asymptotically.

**Theorem 3.2:** Under the conditions in Theorem 3.1, as \( n \to \infty \), we have

\[
\hat{R}_Q(\beta_0) \to \rho_1 \chi^2_1 + \rho_2 \chi^2_1 + \cdots + \rho_p \chi^2_{1}, \\
\hat{R}_H(\beta_0) \to \varrho_1 \chi^2_1 + \varrho_2 \chi^2_1 + \cdots + \varrho_{pq} \chi^2_{1},
\]

where \( \chi^2_1 \) is the standard \( \chi^2 \) distribution with one degree, the weights \( \rho_i \) and \( \varrho_j \) are eigenvalues of \( \Lambda_g^{-1} \Sigma_g \) and \( \Lambda_h^{-1} \Sigma_h \), respectively, \( i = 1, \ldots, pl \) and \( j = 1, \ldots, pq \).

**Remark 3.3:** In general, when the parameters are over-identified, one can define the empirical log-likelihood ratio functions \( W_Q(\beta_0) = \hat{R}_Q(\beta_0) - \hat{R}_Q(\hat{\beta}_Q) \) and \( W_H(\beta_0) = \hat{R}_H(\beta_0) - \hat{R}_H(\hat{\beta}_H) \) [49]. For nonignorable missing data, however, the Wilks’s theorem based on the EL does not hold any more [48]. Lemmas 1 and 2 in the Supplementary Material reveal the reasons why the limiting distributions of \( \hat{R}_Q(\beta_0) \) and \( \hat{R}_H(\beta_0) \) do not follow the standard chi-square. Thus, it also can be shown that \( W_Q(\beta_0) \) and \( W_H(\beta_0) \) do not follow Wilks’s phenomenon.

**Remark 3.4:** When there are no missing data, we have \( \Lambda_g = \Sigma_g \) and \( \Lambda_h = \Sigma_h \) such that both \( \Lambda_g^{-1} \Sigma_g \) and \( \Lambda_h^{-1} \Sigma_h \) equal to the identity matrix, which makes the Wilks’s theorem hold. This is the same as the result of Li and Pan [14]. Moreover, Theorem 3.2 can be used to test the hypothesis \( H_0: \beta = \beta_0 \) and construct the confidence region for \( \beta_0 \).

Let \( r_Q(\beta_0) = pl/\text{tr}(\Lambda_g^{-1} \Sigma_g) \) and \( r_H(\beta_0) = pq/\text{tr}(\Lambda_h^{-1} \Sigma_h) \) be the adjustment factors.

Along the lines of Rao and Scott [50], we have the following corollary.

**Corollary 3.3:** Under the conditions of Theorem 2.1, as \( n \to \infty \), we obtain

\[
\hat{R}_Q(\beta_0) r_Q(\beta_0) \to \chi^2_{pl}, \quad \hat{R}_H(\beta_0) r_H(\beta_0) \to \chi^2_{pq}.
\]
Remark 3.5: To construct the confidence regions of $\hat{\beta}_Q$ according to Corollary 3.3, 100(1 - $\alpha$)% confidence region for $\hat{\beta}_Q$ is give by

$$CI_Q(\alpha) = \{ \beta, r(\hat{\beta})\hat{R}_Q(\beta) < \chi^2_{p,1-\alpha} \},$$

where $r_Q(\hat{\beta}) = pl/tr(\hat{\Lambda}_g^{-1}\hat{\Sigma}_g)$. Similarly, we can obtain the confidence regions of $\hat{\beta}_H$.

Let $\mathcal{A} = \{ j : \beta_j \neq 0 \}$ be the set of nonzero components of true parameter vector $\beta_0$ with its cardinality as $d = |\mathcal{A}|$. Without loss of generality, the parameter vector can be partitioned as two parts $\beta = (\beta_1^T, \beta_2^T)^T$, where $\beta_1 \in R^d$ and $\beta_2 \in R^{p-d}$. Through the Equation (9), we can perform variable selection and also produce robust estimators of the nonzero components. Denote $\hat{\beta}_Q = (\hat{\beta}_Q^1, \hat{\beta}_Q^2)^T$ and $\hat{\beta}_H = (\hat{\beta}_H^1, \hat{\beta}_H^2)^T$ as the resulting penalized robust estimators based on the QIF and hybrid-GEE methods in (9), respectively. Here, the true parameter $\beta_0 = (\beta_1^T, 0^T)^T$.

Theorem 3.4: Under the regularity conditions in Theorem 3.1, we further assume conditions (C8)–(C9) hold. As $n \to \infty$, the estimators $\hat{\beta}_Q$ and $\hat{\beta}_H$ satisfies

(i) (Selection consistency): With probability tending to 1, $\hat{\beta}_Q^2 = 0$ and $\hat{\beta}_H^2 = 0$;

(ii) (Asymptotic normality):

$$\sqrt{n}(\hat{\beta}_Q - \beta_1) \to N(0, \Gamma_g^{(11)}\Sigma_g^{(11)}[\Gamma_g^{(11)}]^T),$$

$$\sqrt{n}(\hat{\beta}_H - \beta_1) \to N(0, \Gamma_h^{(11)}\Sigma_h^{(11)}[\Gamma_h^{(11)}]^T),$$

where $\Lambda_g^{(11)}$ and $\Lambda_h^{(11)}$ are $dl \times dl$ and $dq \times dq$ sub-matrices of $\Lambda_g$ and $\Lambda_h$, $\Delta_g^{(12)}$ and $\Delta_h^{(12)}$ are $dl \times d$ and $dq \times d$ sub-matrices of $\Gamma_g$ and $\Gamma_h$, $\Sigma_g^{(11)}$ and $\Sigma_h^{(11)}$ are $dl \times dl$ and $dq \times dq$ sub-matrices of $\Sigma_g$ and $\Sigma_h$.

$$\Gamma_g^{(11)} = \{[\Lambda_g^{(12)}]^T[\Lambda_g^{(11)}]^{-1}[\Lambda_g^{(12)}]^{-1}\}^{-1}[\Lambda_g^{(12)}]^T[\Lambda_g^{(11)}]^{-1},$$

$$\Gamma_h^{(11)} = \{[\Delta_h^{(12)}]^T[\Lambda_h^{(11)}]^{-1}\}^{-1}[\Delta_h^{(12)}]^T[\Lambda_h^{(11)}]^{-1}.$$

4. Simulation studies

We conduct simulation studies to examine the finite-sample performance of the following estimators of $\beta$:

(a) the proposed robust QIF estimators based on $\hat{g}_i(\beta)$ in (7) and the robust hybrid-GEE estimators based on $\hat{h}_i(\beta)$ in (8) with nonignorable dropout propensity $\pi_{ij}(\hat{\theta}_j)$ in $\hat{S}_i$, and the GMM estimator $\hat{\theta}_j$ obtained by (11). Computation details can be seen in section 4.1. Here, $R^{-1}$ in the QIF estimators are based on two common working correlation choices: AR(1) and CS, which are denoted as R-QIF_{AR(1)} and R-QIF_{CS}; two different choices of $\{R_1(\alpha), R_2(\alpha)\} = \{\text{AR(1)}, \text{CS}\}$ with $\alpha = 0.4$ and $0.7$ are used in the hybrid-GEE method, which are denoted as R-Hybrid_{0.4} and R-Hybrid_{0.7}. In addition, the non-robust QIF and hybrid-GEE estimators based on $\hat{g}_i(\beta)$ and $\hat{h}_i(\beta)$ in (7)
and (8) with $T_i(\mu_i(\beta)) = y_i - \mu_i$ are also obtained, which are denoted as QIFAR(1), QIFCS, Hybrid0.4 and Hybrid0.7, respectively.

(b) the robust MAR estimator (denoted as R-MAR) based on (5) with the ignorable nominal level 1 squared error (RMSE) based on the R-FULL method, coverage probability (CP) at the matric efficiency, we compute the simulated bias, standard deviation (SD), relative mean squared error (RMSE) based on the R-FULL method, coverage probability (CP) at the nominal level 1 - $\alpha = 0.99$ by Corollary 2.3. All results are based on 1000 simulation replications with the sample sizes $n = 200$ and 500.

4.1. Correctly specified propensity

In the first simulation, the indicators for the missingness $r_i = (r_{i1}, r_{i2}, r_{i3}, r_{i4})^T$ are generated from the following nonignorable dropout choice:

$$(M1) : \Pr(r_{ij} = 1 \mid r_{i(j-1)} = 1, \overrightarrow{u}_{ij}, \overrightarrow{y}_{ij}) = 1 / \{1 + \exp(\alpha_j + \gamma_{j1}x_{ij1} + \gamma_{j2}y_{ij})\},$$

with $\alpha_j = -1.6 + 0.2(j - 1)$, $\gamma_{j1} = 0.2 + 0.2(j - 1)$ and $\gamma_{j2} = -0.4$. For $j = 1, \ldots, 4$, the unconditional dropout percentages for four time points under different scenarios are about 19%, 37%, 56% and 72%. It can be shown that $u_{ij} = x_{ij1}$, the instrument variable $z_{ij} = x_{ij2}$ and
the working propensity model is correct. To estimate $\xi_j$, we define the following estimating equations:

$$s_j(y_i, x_i, r_i, \xi_j) = r_{i(j-1)} \left\{ \frac{r_{ij}}{\Psi(\alpha_j + \gamma_1 x_{ij1} + \gamma_2 y_{ij})} - 1 \right\} \begin{pmatrix} 1, \mathbf{u}_i^T, \mathbf{z}_i^T, \mathbf{y}_i^T_{i(j-1)} \end{pmatrix}^T.$$

If $\xi_j^0$ is the true value of $\xi_j$, it can be verified that $E\{s_j(y_i, x_i, r_i, \xi_j^0)\} = 0$. The efficient two-step GMM [34] estimator of $\xi_j$ is

$$\hat{\xi}_j = \arg\min_{\xi_j} \bar{s}_j(\xi_j)^T \hat{\Omega}_j^{-1} \bar{s}_j(\xi_j), \quad (11)$$

where $\bar{s}_j(\xi_j) = n^{-1} \sum_{i=1}^n s_j(y_i, x_i, r_i, \xi_j)$, $\hat{\Omega}_j^{-1}$ is the inverse of the matrix $n^{-1} \sum_{i=1}^n s_j(y_i, x_i, r_i, \hat{\xi}_j^{(1)} s_j(y_i, x_i, r_i, \hat{\xi}_j^{(1)})^T$, and $\hat{\xi}_j^{(1)} = \arg\min_{\xi_j} \bar{s}_j(\xi_j)^T \bar{s}_j(\xi_j)$. To see the robustness of the proposed estimators, we consider the following three cases:

(Case0) no contamination based on observed $y_{ij}$ and $x_{ij}$;
(Case1) randomly choose 10% of observed $y_{ij}$ to be $y_{ij} + 10$;
(Case2) randomly choose 2% of observed $x_{ij}$ to be $x_{ij} + 2$ and 10% of observed $y_{ij}$ to be $y_{ij} + 10$.

To save space, Tables 1–2 only report the simulation results under multivariate $t$ errors with the AR(1) and CS correlation structures and the rest of the results are provided in Tables S1–S4 of the Supplementary Material. A few conclusions can be drawn from the results:

1. When there are no outliers (Case0), the proposed robust QIF and hybrid-GEE estimates are almost the same as those non-robust QIF and hybrid-GEE estimates in terms of the biases, SDs, RMSEs and CPs. These results indicate that our proposed robust estimators seem to perform no worse than non-robust estimators by introducing the weight matrix and the Huber’s score function, which control the degree of robustness and efficiency. While, the MAR estimates based on ignorable dropout have large biases and RMSEs under nonignorable missing models, which makes their CPs much lower.

2. In the presence of outliers (Case1 and Case2), the non-robust estimates can impact greatly in terms of the biases, SDs, RMSEs and CPs, while the proposed estimates based on the robust QIF and hybrid-GEE methods still perform well, which indicates that the influence of outliers in responses and/or covariates can be reduced. The biases and RMSEs of the MAR estimates become more larger due to nonignorable missing models and outliers. Among the two contaminated cases, the SDs and RMSEs of all estimates increase as the contamination level increases, but the proposed robust QIF and hybrid-GEE estimates have much smaller magnitudes of increase.

3. Among the two robust QIF estimators, it can be seen that the estimates R-QIF$\text{AR}(1)$ have smaller or comparable RMSEs than these based on the estimates R-QIF$\text{CS}$, which indicates that the AR(1) structure best approximates the true correlation structure. Among the two robust hybrid-GEE estimators, the estimates R-Hybrid$0.4$ have smaller
In these second simulation, we investigate the performance of the proposed estimators when errors with the AR(1) correlation matrix do not affect the consistence, but will affect the efficiency.

(4) When $n = 500$, the SDs and RMSEs become smaller and the proposed four estimators have similar performance.

### 4.2. Misspecified propensity

In the second simulation, we investigate the performance of the proposed estimators when the propensity is misspecified. In specific, we consider the same settings as in the first simulation with normal and multivariate $t$ errors, expect that

\[ \text{(M2): } \Pr(r_{ij} = 1 | r_{i(j-1)}) = 1/\{1 + \exp(\alpha_2 + \gamma_1 \sin(x_{ij1}) + \gamma_2 y_{ij})\}. \]
In the third simulation, we assess the finite sample performance of variable selection based on the proposed estimators with SCAD penalty in terms of model complexity (sparsity), and the similar results as in the first simulation. The proposed estimators have robust and efficient results, even the working dropout propensity model is wrong.

### 4.3. Variable selection

In the third simulation, we assess the finite sample performance of variable selection based on the proposed estimators with SCAD penalty in terms of model complexity (sparsity), and the similar results as in the first simulation. The proposed estimators have robust and efficient results, even the working dropout propensity model is wrong.

---

**Table 2.** Biases, standard deviations (in parentheses), relative mean squared errors (RMSEs) and coverage probabilities (CPs) in the first simulation under multivariate $t$ errors with the CS correlation structure.

| Methods | $\beta_1$ | $\beta_2$ | RMSE | CP   | $\beta_1$ | $\beta_2$ | RMSE | CP   |
|---------|----------|----------|------|------|----------|----------|------|------|
| Case0   | FULL     | -0.002(0.045) | 0.002(0.054) | 1.013 | 0.991 | 0.000(0.030) | 0.000(0.032) | 1.017 | 0.993 |
| MAR     | 0.124(0.074) | -0.155(0.097) | 11.25 | 0.931 | 0.132(0.045) | -0.164(0.063) | 26.99 | 0.560 |
| QIF $\theta_{AR}(1)$ | 0.031(0.077) | -0.035(0.096) | 3.597 | 0.975 | 0.018(0.052) | -0.025(0.065) | 4.211 | 0.973 |
| QIF $\theta_{CS}$ | 0.019(0.084) | -0.021(0.103) | 3.813 | 0.979 | 0.012(0.054) | -0.017(0.068) | 4.294 | 0.976 |
| Hybrid $\theta_{4}$ | 0.018(0.083) | -0.022(0.101) | 3.708 | 0.989 | 0.012(0.055) | -0.018(0.067) | 4.310 | 0.976 |
| Hybrid $\theta_{7}$ | 0.013(0.095) | -0.016(0.105) | 4.231 | 0.983 | 0.011(0.060) | -0.016(0.069) | 4.670 | 0.979 |
| R-FULL  | -0.002(0.045) | 0.001(0.053) | 1.00 | 0.992 | -0.001(0.029) | 0.000(0.032) | 1.000 | 0.995 |
| R-MAR   | 0.118(0.073) | -0.150(0.099) | 10.64 | 0.960 | 0.125(0.045) | -0.154(0.062) | 24.35 | 0.666 |
| R-QIF $\theta_{AR}(1)$ | 0.022(0.080) | -0.028(0.102) | 3.749 | 0.985 | 0.011(0.051) | -0.016(0.064) | 3.758 | 0.987 |
| R-QIF $\theta_{CS}$ | 0.015(0.084) | -0.016(0.108) | 3.939 | 0.987 | 0.007(0.052) | -0.011(0.064) | 3.732 | 0.985 |
| R-Hybrid $\theta_{4}$ | 0.015(0.084) | -0.017(0.105) | 3.830 | 0.989 | 0.007(0.052) | -0.012(0.063) | 3.672 | 0.984 |
| R-Hybrid $\theta_{7}$ | 0.012(0.097) | -0.014(0.111) | 4.536 | 0.987 | 0.007(0.059) | -0.011(0.066) | 4.311 | 0.987 |

Here, the unconditional dropout percentages for four time points are about 18%, 35%, 52% and 66% for $j = 1, 2, 3, 4$. Use the same method to estimate $\xi_j$. In this case, the working propensity model is misspecified so that we can see the robustness of the proposed estimators. Simulation results under multivariate $t$ errors with the AR(1) and CS correlation structure are presented in Tables 3–4. The results under normal errors with the AR(1) and CS correlation structure are given in Tables S5–S6 of the Supplementary Material. We have the similar results as in the first simulation. The proposed estimators have robust and efficient results, even the working dropout propensity model is wrong.
model error and model selection accuracy. We consider
\[ y_i = x_i \beta + \epsilon_i, \quad i = 1, 2, \ldots, n, \]

where \( x_i = (x_{i1}, \ldots, x_{id})^T \) is a \((4 \times p)\)-dimensional matrix of covariates, \( x_{i1}, \ldots, x_{id} \) are from a \( p \)-dimensional normal distribution having mean \((1, \ldots, 1)^T\) and covariance matrix \( \Gamma \) with \( \Gamma_{jj} = 1 \) and \( \Gamma_{jj'} = 0.4 \) for \( 1 \leq j < j' \leq p \), \( \epsilon_i \) are considered to be AR(1) and CS errors, respectively. The true value of \( \beta = (3, 1.5, 2, 0, \ldots, 0) \). The indicators for the missingness \( r_i = (r_{i1}, r_{i2}, r_{i3}, r_{i4})^T \) are generated from
\[
\Pr(r_{ij} = 1 \mid r_{i(j-1)} = 1, \vec{u}_{ij}, \vec{y}_{ij}) = 1/(1 + \exp(\alpha_j + y_{ij}^T \beta_j + y_{j2}y_{ij})),
\]
Table 4. Biases, standard deviations (in parentheses), relative mean squared errors (RMSEs) and coverage probabilities (CPs) in the second simulation under multivariate t errors with the CS correlation structure.

| Methods  | $\beta_1$ | $\beta_2$ | RMSE | CP      | $\beta_1$ | $\beta_2$ | RMSE | CP      |
|----------|-----------|-----------|------|---------|-----------|-----------|------|---------|
| Case0    | 0.002(0.046) | -0.003(0.054) | 0.964 | 0.989 | 0.000(0.028) | -0.001(0.034) | 0.952 | 0.988 |
| FULL     | 0.084(0.071) | -0.150(0.092) | 8.095 | 0.927 | 0.084(0.044) | -0.147(0.057) | 16.58 | 0.754 |
| MAR      | 0.017(0.076) | -0.040(0.098) | 3.245 | 0.968 | 0.008(0.050) | -0.022(0.064) | 3.491 | 0.971 |
| QIFAR(1) | 0.010(0.078) | -0.029(0.102) | 3.284 | 0.972 | 0.003(0.052) | -0.016(0.065) | 3.493 | 0.974 |
| QIFCS    | 0.006(0.078) | -0.029(0.102) | 3.304 | 0.974 | 0.003(0.052) | -0.016(0.066) | 3.601 | 0.976 |
| Hybrid0.4| -0.006(0.087) | -0.022(0.106) | 3.648 | 0.983 | -0.003(0.058) | -0.012(0.065) | 3.788 | 0.984 |
| Hybrid0.7| -0.006(0.087) | -0.022(0.106) | 3.648 | 0.983 | -0.003(0.058) | -0.012(0.065) | 3.788 | 0.984 |
| R-FULL   | 0.002(0.047) | -0.003(0.055) | 1.000 | 0.989 | 0.000(0.029) | -0.001(0.035) | 1.000 | 0.988 |
| R-MAR    | 0.081(0.072) | -0.145(0.093) | 7.802 | 0.950 | 0.080(0.044) | -0.140(0.057) | 15.35 | 0.801 |
| R-QIFAR1 | 0.012(0.078) | -0.033(0.102) | 3.293 | 0.983 | 0.005(0.050) | -0.016(0.064) | 3.395 | 0.987 |
| R-QIFCS  | 0.007(0.078) | -0.023(0.104) | 3.288 | 0.982 | 0.002(0.049) | -0.012(0.063) | 3.226 | 0.988 |
| R-Hybrid0.4| 0.002(0.082) | -0.023(0.105) | 3.444 | 0.984 | 0.002(0.050) | -0.011(0.063) | 3.239 | 0.991 |
| R-Hybrid0.7| -0.008(0.091) | -0.018(0.109) | 3.862 | 0.990 | -0.002(0.056) | -0.008(0.064) | 3.568 | 0.988 |

where $u_{ij} = (x_{ij1}, \ldots, x_{ijp})^T$, $\alpha_j = -0.8 + 0.2(j - 1)$, $y_{j1}^T = (-0.1, 0.1, \ldots, -0.1, 0.1, \ldots)$ and $\gamma_{j2} = -0.4 + 0.1(j - 1)$ for $j = 1, 2, 3, 4$. In order to investigate the performance of the proposed estimators, the contamination situation Case1 is considered.

For the proposed method with SCAD penalty, we apply the local quadratic approximation (LQA) to the penalty function as discussed in Fan and Li [37]. That is,

$$p_v(|\beta|) \approx p_v(|\beta_0|) + p_v^\prime(|\beta_0|)/|\beta_0|(|\beta|_2^2 - |\beta_0|_2^2)$$

for $|\beta|_2 \approx |\beta_0|_2$.

To iterate on $\beta$ directly, we adopt an approximate algorithm based on the full model expression. Given an initial $\beta^{(0)}$, the optimization can be carried out by a modified Newton-Raphson algorithm. For $k = 1, 2, \ldots$, an iterative sequence

$$\beta^{(k+1)} = \beta^{(k)} + [Z_1(\beta^{(k)}) + \sum_{v} \beta^{(k)}]^{-1}Z_2(\beta^{(k)} - U_v(\beta^{(k)})),$$
are generated, where \( \Sigma_{v, \beta} = \text{diag}[p'_v(|\beta_1|)/|\beta_1|, \ldots, p'_v(|\beta_p|)/|\beta_p|] \), \( U_v(\beta) = \Sigma_{v, \beta} \beta \) and

\[
Z_1(\beta) = \frac{\partial T_{2n}(\beta, 0)}{\partial \lambda} \left\{ \frac{\partial T_{1n}(\beta, 0)}{\partial \lambda} \right\}^{-1} \frac{\partial T_{1n}(\beta, 0)}{\partial \beta}, \\
Z_2(\beta) = \frac{\partial T_{2n}(\beta, 0)}{\partial \lambda} \left\{ \frac{\partial T_{1n}(\beta, 0)}{\partial \lambda} \right\}^{-1} T_{1n}(\beta, 0),
\]

with

\[
\frac{\partial T_{1n}(\beta, 0)}{\partial \beta} = \frac{\partial T_{2n}(\beta, 0)}{\partial \lambda} = \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{\eta}_i(\beta)}{\hat{\beta}_i}, \\
\frac{\partial T_{1n}(\beta, 0)}{\partial \lambda} = - \frac{1}{n} \sum_{i=1}^{n} \hat{\eta}_i(\beta) \hat{\lambda}(\beta)^T, \\
T_{1n}(\beta, \lambda) = \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{\eta}_i(\beta)}{1 + \lambda T \hat{\eta}_i(\beta)}, \\
T_{2n}(\beta, \lambda) = \frac{1}{n} \sum_{i=1}^{n} \left[ \frac{\hat{\eta}_i(\beta) / \hat{\beta}_i^T \hat{\lambda}}{1 + \lambda T \hat{\eta}_i(\beta)} \right].
\]

The iteration stops when solutions converge to a satisfying precision. For a prespecified value \( \zeta \), set \( \hat{\beta}_j = 0 \) if \( |\hat{\beta}_j| < \zeta \) and we apply the algorithm in Owen [5] to compute \( \hat{\lambda} \).

Table 5 only reports the results for \( n = 200, 500 \) and \( p = 10, 20 \) with the AR(1) error based on three information criteria: BIC, BICC and EBIC, for selecting the tuning parameter \( \nu \). The rest of the results under the CS error are presented in Table S7 of the Supplementary Material. We obtain the mean square errors (MSE) defined by MSE(\( \hat{\beta} \)) = (\( \hat{\beta} - \beta \)) \( ^T \) (\( \hat{\beta} - \beta \)). Columns ‘C’ and ‘IC’ are measures of model complexity, with ‘C’ representing the average number of nonzero coefficients correctly estimated to be nonzero, and ‘IC’ representing the average number of zero coefficients incorrectly estimated to be nonzero. The simulated results of the oracle model (i.e., the model using the true predictors) are also reported.

From Tables 5 and S7, it can be seen that: (1) the proposed variable selection methods can select all three true predictors and the average numbers of zero coefficients incorrectly estimated to be nonzero are close to zero in most of cases; (2) the simulated MSEs of the proposed methods based on the BIC, BICC and EBIC are close to that of the oracle EL, especially for larger sample sizes; (3) in terms of MSEs and ICs, it is interesting to note that the BIC and BICC have similar performance and the EBIC has the best performance in most of cases. These findings imply that the model selection results based on the proposed approaches are satisfactory and the selected models are very close to the true model; (4) based on these results, in practice we recommend to use the information criteria EBIC for selecting \( \nu \).

5. Application to HIV-CD4 data

For illustration, we apply the proposed methods to the AIDS Clinical Trial Group 193A described in Section 1. In this study, the CD4 counts were collected from 316 patients who took the daily regimen containing 600 mg of zidovudine plus 2.25 mg of zalcitabine. We consider the first four follow-up times, 8, 16, 24, 32, as four time points \( j = 1, 2, 3, 4 \), and use the CD4 counts in four time intervals, (4, 12], (12,20], (20, 28], (28, 36], as the study variable \( y_{ij} \) for \( j = 1, 2, 3, 4 \), because the realised follow-up time points might be a little different from the scheduled time points. A few patients had more than one measurement.
Table 5. Mean square errors (MSE) and variable selection results under AR(1) errors.

| p   | Methods      | Criteria | MSE  | C   | IC  | MSE  | C   | IC  |
|-----|--------------|----------|------|-----|-----|------|-----|-----|
| 10  | R-QIF_{AR(1)} BIC | 0.024    | 3    | 0.669 | 0.007 | 3 | 0 |
|     |              | BICC     | 0.024 | 3 | 0.667 | 0.007 | 3 | 0 |
|     |              | EBIC     | 0.022 | 3 | 0.033 | 0.007 | 3 | 0 |
|     | R-QIF_{CS}   BIC | 0.024    | 3    | 0.060 | 0.007 | 3 | 0.001 |
|     |              | BICC     | 0.023 | 3 | 0.052 | 0.007 | 3 | 0.001 |
|     |              | EBIC     | 0.021 | 3 | 0.017 | 0.007 | 3 | 0.001 |
|     | R-Hybrid_{0.4} BIC | 0.022    | 3    | 0.067 | 0.007 | 3 | 0 |
|     |              | BICC     | 0.022 | 3 | 0.061 | 0.007 | 3 | 0 |
|     |              | EBIC     | 0.019 | 3 | 0.015 | 0.007 | 3 | 0 |
|     | R-Hybrid_{0.7} BIC | 0.025    | 3    | 0.084 | 0.007 | 3 | 0 |
|     |              | BICC     | 0.024 | 3 | 0.078 | 0.007 | 3 | 0 |
|     |              | EBIC     | 0.021 | 3 | 0.026 | 0.007 | 3 | 0 |
|     | Oracle       | 0.009    | 3    | 0    | 0.004 | 3 | 0 |
| 20  | R-QIF_{AR(1)} BIC | 0.139    | 3    | 0.774 | 0.048 | 3 | 0.143 |
|     |              | BICC     | 0.137 | 3 | 0.749 | 0.047 | 3 | 0.121 |
|     |              | EBIC     | 0.122 | 3 | 0.598 | 0.047 | 3 | 0.106 |
|     | R-QIF_{CS}   BIC | 0.162    | 3    | 0.792 | 0.055 | 3 | 0.146 |
|     |              | BICC     | 0.144 | 3 | 0.593 | 0.054 | 3 | 0.134 |
|     |              | EBIC     | 0.130 | 3 | 0.299 | 0.052 | 3 | 0.075 |
|     | R-Hybrid_{0.4} BIC | 0.102    | 3    | 0.714 | 0.020 | 3 | 0.028 |
|     |              | BICC     | 0.095 | 3 | 0.615 | 0.020 | 3 | 0.025 |
|     |              | EBIC     | 0.073 | 3 | 0.281 | 0.019 | 3 | 0.012 |
|     | R-Hybrid_{0.7} BIC | 0.135    | 3    | 0.849 | 0.016 | 3 | 0.028 |
|     |              | BICC     | 0.115 | 3 | 0.596 | 0.016 | 3 | 0.025 |
|     |              | EBIC     | 0.085 | 3 | 0.251 | 0.015 | 3 | 0.012 |
|     | Oracle       | 0.217    | 3    | 0    | 0.174 | 3 | 0 |

in one time interval, in which case we use the last record in that interval as \( y_{ij} \) at time point \( j \). Some patients returned to the study after they dropped out of the study. For simplicity, the measurements after they dropped out of the study are not used in the analysis. There are two continuous covariates: age \((x_{ij1})\) and follow-up time \((x_{ij2})\). We use the working propensity model (10) and \( \Psi(\cdot) = [1 + \exp(\cdot)]^{-1} \).

Because of adverse events, low-grade toxic reactions, declining CD4 cell counts before the study end, the desire to seek other therapies, death and some other reasons, patients are willing to dropout. Hence, the follow-up times are treated as covariates \( u_{ij} \) may affect the dropout. The ages are always observed and thus are used as instruments \( z_{ij} \).

The point estimates and their normal-approximation-based confidence intervals by bootstrap with replication size 300 are reported in Table 6. It can be seen that: (1) all proposed estimates of \( \beta_1 \) are statistically significant negative, which is reasonable since we have known that the number of CD4 counts of these patients keeps decreasing as time goes on and the trends become worse for those with lower CD4 counts. However, all the confidence intervals of the non-robust estimates include 0, which indicates the time effect on the number of CD4 counts is not significant; (2) the proposed estimates of \( \beta_2 \) are statistically significant positive, which indicates that patients with earlier ages infected by the HIV are more likely to have lower CD4 counts. On the contrary, all the confidence intervals of the non-robust estimates include 0 and show the infected ages effect on the number of CD4 counts is not significant; (3) the estimates based on the MNAR and complete-case (CC) assumptions are different in most of cases. Therefore, the ignorable dropout assumption...
Table 6. Estimates, standard errors (SE) and confidence intervals (CI) for the HIV-CD4 data based on the proposed methods when the instrument variable is age.

| Method         | MNAR | CC       |
|----------------|------|----------|
|                | β₁   | β₂   | β₀   | β₁   | β₂   | β₀   |
| QIFCS          | -0.238 | 0.715 | 0.148 | -0.147 | 0.660 | 0.165 |
| SE             | 0.216 | 0.543 | 1.239 | 0.131 | 0.457 | 0.821 |
| CI             | (-0.662, 0.186) | (-0.349, 1.780) | (-2.280, 2.576) | (-0.404, 0.110) | (-0.236, 1.557) | (-1.444, 1.774) |
| QIFAR(1)       | -0.225 | 0.661 | 0.311 | -0.150 | 0.646 | 0.228 |
| SE             | 0.186 | 0.502 | 1.167 | 0.113 | 0.454 | 0.888 |
| CI             | (-0.589, 0.140) | (-0.323, 1.644) | (-1.977, 2.598) | (-0.370, 0.071) | (-0.244, 1.536) | (-1.513, 1.968) |
| Hybrid0.7      | -0.253 | 0.748 | 0.074 | -0.174 | 0.689 | 0.138 |
| SE             | 0.172 | 0.556 | 1.274 | 0.114 | 0.479 | 0.909 |
| CI             | (-0.589, 0.084) | (-0.343, 1.838) | (-2.424, 2.572) | (-0.397, 0.048) | (-0.249, 1.627) | (-1.644, 1.920) |
| Hybrid0.4      | -0.256 | 0.769 | 0.004 | -0.173 | 0.682 | 0.157 |
| SE             | 0.182 | 0.572 | 1.288 | 0.113 | 0.469 | 0.870 |
| CI             | (-0.612, 0.100) | (-0.353, 1.890) | (-2.528, 2.520) | (-0.395, 0.049) | (-0.238, 1.602) | (-1.548, 1.863) |
| R-QIFCS        | -0.377 | 0.746 | 1.236 | -0.223 | 0.785 | 0.696 |
| SE             | 0.150 | 0.267 | 0.821 | 0.075 | 0.222 | 0.778 |
| CI             | (-0.671, -0.084) | (0.223, 1.269) | (-0.373, 2.845) | (-0.371, -0.076) | (0.350, 1.219) | (-0.830, 2.221) |
| R-QIFAR(1)     | -0.374 | 0.724 | 1.038 | -0.204 | 0.735 | 0.827 |
| SE             | 0.162 | 0.303 | 0.913 | 0.086 | 0.234 | 0.861 |
| CI             | (-0.692, -0.056) | (0.130, 1.317) | (-0.480, 3.097) | (-0.372, -0.037) | (0.276, 1.194) | (-0.862, 2.515) |
| R-Hybrid0.7    | -0.376 | 0.727 | 1.309 | -0.240 | 0.789 | 0.725 |
| SE             | 0.097 | 0.249 | 0.851 | 0.061 | 0.239 | 0.868 |
| CI             | (-0.567, -0.185) | (0.238, 1.215) | (-0.358, 2.976) | (-0.359, -0.121) | (0.322, 1.257) | (-0.976, 2.425) |
| R-Hybrid0.4    | -0.367 | 0.756 | 1.166 | -0.236 | 0.795 | 0.691 |
| SE             | 0.087 | 0.245 | 0.869 | 0.062 | 0.241 | 0.883 |
| CI             | (-0.537, -0.197) | (0.276, 1.236) | (-0.536, 2.869) | (-0.357, -0.115) | (0.322, 1.268) | (-1.039, 2.421) |

is questionable; (4) compared with the non-robust estimates, our proposed estimates have smaller standard errors and shorter confidence intervals.

A sensitivity analysis is implemented by choosing the follow-up times as the instrument \( z_{ij} \) and the ages as the covariates \( u_{ij} \), respectively. The estimates and standard errors are reported in Table S8 in the Supplementary Material. It can be seen that our proposed estimates using \( z_{ij} = \text{age} \) have much smaller standard errors, which also indicates that we should use the age as the instrument variable.

6. Discussion

In this paper, we first make use of the IPW method and an instrument variable to deal with the nonignorable dropout. To achieve the robustness against outliers, the Huber’s score function and Mallows-type weights are then applied to build the robust and bias-corrected GEEs. Based on the QIF and hybrid-GEE, we construct two classes of improved estimators that can incorporate the within-subject correlations under an informative working correlation structure. The existing two-step GMM and EL approaches are used to obtain the proposed estimators. In addition, we propose the robust variable selection through the penalized robust EL by combining the profile robust EL and the SCAD methods together. The simulation results show that our proposed estimators perform well even when the working dropout propensity model is wrong; the proposed variable selection methods can efficiently select significant variables and estimate parameters simultaneously.

The above methods are presented with balanced data, that is, \( m_i = m \). In practice, longitudinal data may not be measured with the same subject size, and could be unbalanced due
to experimental constraints. To configure the proposed methods for unbalanced data, we apply the transformation matrix to each subject. Similar in Zhou and Qu [51], we create the largest subject with a size \( m \), which contains time points for all possible measurements, and assume that fully observed clusters contain \( m \) observations. We define the \( m \times m_i \) transformation matrix \( M_i \) for the \( i \)th subject by removing the columns of the \( m \times m_i \) identity matrix, where the removed columns correspond to unmeasured data/time points for the \( i \)th subject. Through the transformation, \( \hat{g}_i(\beta) \) is replaced by

\[
\hat{g}_i(\beta) = \begin{pmatrix}
(D_i^*)^T(\hat{A}_i^*)^{-1/2}B_1(\hat{A}_i^*)^{-1/2}S_i^T \hat{T}_i(\mu_i^*(\beta)) \\
\vdots \\
(D_i^*)^T(\hat{A}_i^*)^{-1/2}B_i(\hat{A}_i^*)^{-1/2}S_i^T \hat{T}_i(\mu_i^*(\beta))
\end{pmatrix},
\]

where \( y_i^* = M_i y_i \), \( \mu_i^* = M_i \mu_i \), \( D_i^* = M_i D_i \), \( \hat{A}_i^* = M_i \hat{A}_i M_i^T \), \( \hat{S}_i^* = M_i \hat{S}_i M_i^T \), \( W_i^* = M_i W_i \), \( M_i^T \), \( C_i^* = M_i C_i \). It can be seen that the components in \( y_i^* \) are the same as in \( y_i \) for responses for \( j \leq m_i \) but are 0 for \( j > m_i \), and similarly for \( \mu_i^* \), \( D_i^* \) and \( C_i^* \) which do not affect the estimation of \( \beta \). It can be also seen that the main diagonal elements in the matrix \( \hat{A}_i^* \) are the same as in \( \hat{A}_i \) when \( j \leq m_i \) but are 0 when \( j > m_i \) and similarly for \( \hat{S}_i^* \) and \( W_i^* \), which do not affect the estimation of \( \beta \) either. Correspondingly, \( \hat{h}_i(\beta) \) is replaced by

\[
\hat{h}_i^*(\beta) = \begin{pmatrix}
(D_i^*)^T(\hat{A}_i^*)^{-1/2}R_1^{-1}(\alpha)(\hat{A}_i^*)^{-1/2}S_i^T \hat{T}_i(\mu_i^*(\beta)) \\
\vdots \\
(D_i^*)^T(\hat{A}_i^*)^{-1/2}R_q^{-1}(\alpha)(\hat{A}_i^*)^{-1/2}S_i^T \hat{T}_i(\mu_i^*(\beta))
\end{pmatrix}.
\]

Therefore, for unbalanced data with dropout, parameter estimation and variable selection also can be implemented using our proposed methods in Sections 2. We can show that the asymptotic results of Theorems 3.1, 3.2 and 3.4 still hold for unequal subject sizes using the similar way in the proofs of the theorems. When \( y_{ij} \) is obtained under some treatments, baseline covariates measured prior to treatments, such as age group, gender, race, and education level, are related to the study variable but are unrelated with the propensity, when the study variable and other covariates are conditioned, and thus can be considered as instruments. A fully observed proxy or a mismeasured version of the missing response may be treated as an instrument. A sensitivity analysis of studying some or all different decompositions of \( x_{ij} = (u_{ij}, z_{ij}) \).

Several further problems need to be investigated. First, the dimension of the covariates in regression models is assumed to be fixed, thus it is interesting to extend our approach with diverging numbers of parameters [52]. Moreover, the efficiency can further be improved by incorporating population level information based on empirical likelihood approach. Second, in our simulations the constant \( c \) is fixed. Actually, if the errors are normally distributed and there is no contamination, the best choice of \( c \) is \( \infty \); if the errors follow a heavy-tailed distribution, \( c \) should be chosen to be a small positive value. Thus, the choice of tuning constant \( c \) in Huber’s score function may impact the estimation efficiency such that we may propose a Huber’s score function with a data-dependent tuning constant as in Wang et al. [53] as follows. Once \( \hat{\phi} \) is obtained, the data-dependent value \( c \)
can be estimated by
\[
\hat{\tau}(c) = \frac{\left( \sum_{i=1}^{n} \sum_{j=1}^{m} I(|\hat{e}_{ij}| \leq c) \right)^2}{nm \sum_{i=1}^{n} \sum_{j=1}^{m} \left[ I(|\hat{e}_{ij}| \leq c) \psi^2(\hat{e}_{ij}) + c^2 I(|\hat{e}_{ij}| > c) \right]}
\]
where \(\hat{e}_{ij} = \hat{A}_{ij}^{-1/2}(y_{ij} - x_{ij}^T \beta^*)\) with an initial estimate \(\beta^*\). Note that \(\hat{\tau}(c)\) is not a continuous function of \(c\), in practice, the optimal \(c\) can be determined by \(\hat{\tau}(c)\) in the range of \(c \in [0, 3]\) and we obtain the \(c\) value that makes \(\hat{\tau}(c)\) the largest. Third, we may consider the model with heterogeneous effects and apply concave penalty functions to subgroup analysis [54]. Fourth, the efficiency of proposed GMM estimator \(\hat{\xi}_j\) depends on the choice of \(s_j(y_i, x_i, r_i, \xi_j)\), which may not be optimal since we only use the first order moments of data, i.e., \((1, \overrightarrow{u}_{ij}, \overrightarrow{z}_{ij}, \overrightarrow{y}_{i(j-1)})^T\). Other moments or characteristics of may provide more information and, hence, result in more efficient GMM estimators.

**Acknowledgments**

We are grateful to the Editor, an associate editor and two anonymous referees for their insightful comments and suggestions, which have led to significant improvements.

**Disclosure statement**

No potential conflict of interest was reported by the author(s).

**Funding**

This work was supported by the National Natural Science Foundation of China [Grant Nos. 11871287, 11771144, 11801359], Fundamental Research Funds for the Central Universities [Grant No. ZB22000102] and the Key Laboratory for Medical Data Analysis and Statistical Research of Tianjin.

**References**

[1] Yuan Y, Yin G. Bayesian quantile regression for longitudinal studies with nonignorable missing data. Biometrics. 2010;66:105–114.

[2] Xue L, Zhu L. Empirical likelihood semiparametric regression analysis for longitudinal data. Biometrika. 2007;94:921–937.

[3] You J, Chen G, Zhou Y. Block empirical likelihood for longitudinal partially linear regression models. Canad J Statist. 2006;34:79–96.

[4] Liang K, Zeger S. Longitudinal data analysis using generalized linear models. Biometrika. 1986;73:13–22.

[5] Owen A. Empirical likelihood. Boca Raton (FL): Chapman and Hall/CRC Press; 2001.

[6] Huang J, Liu L, Liu N. Estimation of large covariance matrices of longitudinal data with basis function approximations. J Comput Graph Stat. 2007;16:189–209.

[7] Bai Y, Fung WK, Zhu Z. Weighted empirical likelihood for generalized linear models with longitudinal data. J Stat Plan Inference. 2010;140:3446–3456.

[8] Fu L, Wang Y. Quantile regression for longitudinal data with a working correlation model. Comput Statist Data Anal. 2012;56:2526–2538.

[9] Leung D, Wang Y, Zhu M. Efficient parameter estimation in longitudinal data analysis using a hybrid-GEE method. Biostatistics. 2009;10:436–445.
[10] Leng C, Zhang W, Pan J. Semiparametric mean-covariance regression analysis for longitudinal data. J Am Stat Assoc. 2010;105:181–193.

[11] Zhang W, Leng C. A moving average Cholesky factor model in covariance modelling for longitudinal data. Biometrika. 2011;99:141–150.

[12] Zhang W, Leng C, Tang CY. A joint modelling approach for longitudinal studies. J R Stat Soc Ser B (Stat Methodol). 2015;77:219–238.

[13] Ly J, Guo C, Yang H, et al. A moving average Cholesky factor model in covariance modeling for composite quantile regression with longitudinal data. Comput Statist Data Anal. 2017;112:129–144.

[14] Li D, Pan J. Empirical likelihood for generalized linear models with longitudinal data. J Multivar Anal. 2013;114:63–73.

[15] Leng C, Zhang W. Smoothing combined estimating equations in quantile regression for longitudinal data. Stat Comput. 2014;24:123–136.

[16] Xu L, Tang ML, Chen Z. Analysis of longitudinal data by combining multiple dynamic covariance models. Stat Interface. 2019;12:479–487.

[17] Cho H, Qu A. Efficient estimation for longitudinal data by combining large-dimensional moment conditions. Electron J Stat. 2015;9:1315–1334.

[18] Zhao P, Tang N, Jiang D. Efficient inverse probability weighting method for quantile regression with nonignorable missing data. Statistics. 2017;51:363–386.

[19] He X, Fung W, Zhu Z. Robust estimation in generalized partial linear models for clustered data. J Am Stat Assoc. 2005;100(472):1176–1184.

[20] Wang Y, Lin X, Zhu M. Robust estimation functions and bias correction for longitudinal data analysis. Biometrics. 2005;61:684–691.

[21] Qin G, Zhu Z. Robust estimation in generalized semiparametric mixed models for longitudinal data. J Multivar Anal. 2007;98:1658–1683.

[22] Qin G, Zhu Z, Fung W. Robust estimation of covariance parameters in partial liner model for longitudinal data. J Stat Plan Inference. 2009;139:558–570.

[23] Fan Y, Qin G, Zhu Z. Variable selection in robust regression models for longitudinal data. J Multivar Anal. 2012;109:156–167.

[24] Zheng R, Fung W, Zhu Z. Robust estimation in joint mean-covariance regression model for longitudinal data. Ann Inst Stat Math. 2013;65:617–638.

[25] Little RJA, Rubin DB. Statistical analysis with missing data. 2nd ed. New York: Wiley; 2002.

[26] Wang L, Qi C, Shao J. Model-assisted regression estimators for longitudinal data with nonignorable dropout. Int Stat Rev. 2019;87:S121–S138.

[27] Fitzmaurice GM, Molenberghs G, Lipsitz SR. Regression models for longitudinal binary responses with informative drop-outs. J R Stat Soc. 1995;57:691–704.

[28] Molenberghs G, Kenward M. Missing data in clinical studies. Chichester, West Sussex: John Wiley and Sons; 2007.

[29] Kim JK, Yu CL. A semiparametric estimation of mean functionals with nonignorable missing data. J Am Stat Assoc. 2011;106:157–165.

[30] Wang S, Shao J, Kim JK. An instrumental variable approach for identification and estimation with nonignorable nonresponse. Stat Sin. 2014;24:1097–1116.

[31] Shao J, Wang L. Semiparametric inverse propensity weighting for nonignorable missing data. Biometrika. 2016;103:175–187.

[32] Bindele F, Zhao Y. Rank-based estimating equation with non-ignorable missing responses via empirical likelihood. Stat Sin. 2018;28:1787–1820.

[33] Cantoni E, Flemming JM, Ronchetti E. Variable selection for marginal longitudinal generalized linear models. Biometrics. 2005;61:507–514.

[34] Hansen LP. Large sample properties of generalized method of moments estimators. Econometrica. 1982;50:1029–1054.

[35] Robins JM, Rotnitzky A, Zhao L. Estimation of regression coefficients when some regressors are not always observed. J Am Stat Assoc. 1994;89:846–866.

[36] Qu A, Lindsay BG, Li B. Improving generalised estimating equations using quadratic inference functions. Biometrika. 2000;87:823–836.
[37] Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc. 2001;96:1348–1360.
[38] Lv J, Yang H, Guo C. An efficient and robust variable selection method for longitudinal generalized linear models. Comput Statist Data Anal. 2014;82:74–88.
[39] Qin G, Zhang J, Zhu Z, et al. Robust estimation of partially linear models for longitudinal data with dropouts and measurement error. Stat Med. 2016;35:5401–5416.
[40] Sinha SK. Robust inference in generalized linear models for longitudinal data. J Statist. 2006;34:261–278.
[41] Rousseeuw PJ, van Zomeren BC. Unmasking multivariate outliers and leverage points. J Am Stat Assoc. 1990;85:633–639.
[42] Diggle P, Kenward MG. Informative drop-out in longitudinal data analysis (with discussion). J R Stat Soc Ser C (Appl Statist). 1994;43:49–93.
[43] Jiang Y, Wang Y, Fu L, et al. Robust estimation using modified Huber’s functions with new tails. Technometrics. 2019;61:111–122.
[44] Yohai VJ. High breakdown-point and high efficiency robust estimates for regression. Ann Statist. 1987;15(2):642–656.
[45] Schwarz G. Estimating the dimension of a model. Ann Statist. 1978;6:461–464.
[46] Wang H, Li B, Leng C. Shrinkage tuning parameter selection with a diverging number of parameters. J R Stat Soc Ser B (Stat Methodol). 2009;71:671–683.
[47] Chen J, Chen Z. Extended Bayesian information criterion for model selection with large sample space. Biometrika. 2008;95:759–771.
[48] Tang N, Zhao P, Zhu H. Empirical likelihood for estimating equations with nonignorably missing data. Stat Sin. 2014;24:723–747.
[49] Qin J, Lawless J. Empirical likelihood and general estimating equations. Ann Statist. 1994;22:300–325.
[50] Rao JNK, Scott AJ. The analysis of categorical data from complex sample surveys: chi-squared tests for goodness of fit and independence in two-way tables. J Am Stat Assoc. 1981;76:221–230.
[51] Zhou J, Qu A. Informative estimation and selection of correlation structure for longitudinal data. J Am Stat Assoc. 2012;107:701–710.
[52] Tang CY, Leng C. Penalized high dimensional empirical likelihood. Biometrika. 2010;97:905–920.
[53] Wang Y, Lin X, Bai Z. Robust estimation using the huber function with a data-dependent tuning constant. J Comput Graph Stat. 2007;16:468–481.
[54] Huang J, Ma S. A concave pairwise fusion approach to subgroup analysis. J Am Stat Assoc. 2017;112:420–423.