Annals of the Institute of Statistical Mathematics (2021) 73:623–647
https://doi.org/10.1007/s10463-020-00761-4

Improved empirical likelihood inference and variable selection for generalized linear models with longitudinal nonignorable dropouts

Lei Wang[1] · Wei Ma[1]

Received: 28 May 2019 / Revised: 16 April 2020 / Published online: 27 August 2020

Abstract

In this paper, we propose improved statistical inference and variable selection methods for generalized linear models based on empirical likelihood approach that accommodates both the within-subject correlations and nonignorable dropouts. We first apply the generalized method of moments to estimate the parameters in the nonignorable dropout propensity based on an instrument. The inverse probability weighting is applied to obtain the bias-corrected generalized estimating equations (GEEs), and then we borrow the idea of quadratic inference function and hybrid GEE to construct the empirical likelihood procedures for longitudinal data with nonignorable dropouts, respectively. Two different classes of estimators and their confidence regions are derived. Further, the penalized EL method and algorithm for variable selection are investigated. The finite-sample performance of the proposed estimators is studied through simulation, and an application to HIV-CD4 data set is also presented.

Keywords Inverse probability weighting · Missing not at random · Nonresponse instrument · Quadratic inference function · Variable selection

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10463-020-00761-4) contains supplementary material, which is available to authorized users.

✉ Lei Wang
lwangstat@nankai.edu.cn

Wei Ma
maweiha@gmail.com

[1] School of Statistics and Data Science, LPMC & KLMDASR, Nankai University, Tianjin 300071, China

1 Introduction

In research areas such as medicine, population health, economics, social sciences and sample surveys, data are often collected from every sampled subject at many time points, which are referred to as longitudinal data. Let $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \dots, y_{im_i})^T$ be a m_i dimensional vector of the ith subject's response and $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}, \dots, \boldsymbol{x}_{im_i})^T$ be a $(m_i \times p)$-dimensional matrix of covariates associated with $\boldsymbol{y}_i$, $i = 1, \dots, n$, where m_i is also called as the cluster size for the ith cluster. Assume that the first and second moments of y_{ij} are modeled by

$$g(\mu_{ij}) = \boldsymbol{x}_{ij}^T \boldsymbol{\beta}, \quad Var(y_{ij}) = \phi v(\mu_{ij}), \tag{1}$$

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

For longitudinal data, it has been recognized that the within-cluster correlation structure plays an important role and a major aspect is how to take into account the correlation structure to improve estimation efficiency. However, since the underlying correlation structure is difficult to describe and specify, a naive and simple way is to use a working model, see You et al. (2006) and Xue and Zhu (2007) and references therein, which may lose some efficiency when strong correlations exist. To overcome this issue, generalized estimating equations (GEEs) proposed by Liang and Zeger (1986) is a popular approach through a working correlation matrix to incorporate the correlation. Recently, Huang et al. (2007) approximated the covariance matrices with basis functions. Bai et al. (2010) proposed the weighted empirical likelihood (EL) to incorporate the possible dependence. Fu and Wang (2012) introduced a combination of between- and within-subject estimating functions based on an exchangeable correlation structure assumption. Li and Pan (2013) and Leng and Zhang (2014) constructed estimating functions by the quadratic inference function (QIF). Alternatively, Leng et al. (2010), Zhang and Leng (2011), Zhang et al. (2015) and Lv et al. (2017) applied the Cholesky decomposition to obtain the within-subject covariance matrix. To get more efficient estimators, Xu et al. (2019) proposed a combined multiple likelihood estimating procedure based on three well-known dynamic covariance models, while Leung et al. (2009) considered a hybrid method that combines multiple GEEs based on different working correlation matrices. Moreover, the GLM may include many irrelevant covariates, especially when the dimension of covariates is not low. In this case, it is important to find which covariates are relevant for prediction, both for better interpretation of the model and for better efficiency of the estimator (Cantoni et al. 2005).

In this paper, we consider the situation where $\boldsymbol{x}_i$ is always observed, but subjects $\boldsymbol{y}_i$ may drop out prior to the end of the study. Let $\boldsymbol{r}_i = (r_{i1}, r_{i2}, \dots, r_{im_i})^T$ be the vector of response indicators, where $r_{ij} = 1$ if y_{ij} is observed and $r_{ij} = 0$ if $y_{ij}, \dots, y_{im_i}$ are not observed. Dropout is ignorable if the dropout propensity $p(\boldsymbol{r}_i | \boldsymbol{x}_i, \boldsymbol{y}_i)$ is a function of the observed values (Little and Rubin 2002), where $p(\cdot|\cdot)$ is a generic notation for conditional distribution or density. Otherwise, dropout is nonignorable or missing not at random (MNAR). The majority of existing methods take the framework of the GEE only naturally accommodates missing at random (MAR) or ignorable dropout. However, in practice, the dropout is often nonignorable (Wang et al. 2019), and developing valid methodologies for statistical analysis with nonignorable dropout is always challenging, since some parameters are not identifiable if there is no assumption imposed, see Molenberghs and Kenward (2007), Kim and Yu (2011), Wang et al. (2014) and Shao and Wang (2016). One of the two key assumptions for identifiability (Wang et al. 2014) is that $\boldsymbol{x}_i$ can be decomposed as two parts $\boldsymbol{x}_i = (\boldsymbol{u}_i, z_i)$, and z_i is unrelated to dropout propensity conditioned on $(\boldsymbol{u}_i, \boldsymbol{y}_i)$, that is, $p(\boldsymbol{r}_i|\boldsymbol{x}_i, \boldsymbol{y}_i) = p(\boldsymbol{r}_i|\boldsymbol{u}_i, \boldsymbol{y}_i)$. Such a covariate z_i is used to create more estimation equations for estimating the propensity and ensures that the propensity is identifiable, and is referred to as a dropout instrument (Wang et al. 2019). For example, in a study of mental health of children in Connecticut (Zahner et al. 1992), researchers were interested in evaluating the prevalence of students with abnormal psychopathological status based on their teachers assessment, which was subject to missingness. As indicated by Ibrahim et al. (2001), the teachers response rate may be related to her assessment of the student but is unlikely to be related to a separate parent report after conditioning on the teachers assessment and fully observed covariates; moreover, the parent report is likely highly correlated with that of the teacher. In this case, the parental assessment constitutes an instrument variable (Miao and Tchetgen Tchetgen 2016). The second key assumption on identifiability is that $p(\boldsymbol{r}_i|\boldsymbol{u}_i, \boldsymbol{y}_i)$ has a parametric form. Details are given in Sect. 2, where we apply the generalized method of moments (GMM; Hansen 1982) to estimate the propensity.

Our contributions of this paper are in three aspects. First, we use a covariate not involved in the propensity to deal with the identifiability issue and such a covariate is called nonresponse instrument (Wang et al. 2014; Shao and Wang 2016; Wang et al. 2019). Secondly, by constructing the bias-corrected GEEs based on the inverse propensity weighting (IPW; Robins et al. 1994) in conjunction with quadratic inference function (QIF; Qu et al. 2000) and hybrid GEE (Leung et al. 2009) methods, we propose two classes of estimators which can incorporate the within-subject correlations under an informative working correlation structure and account for nonignorable dropouts. Finally, for variable selection, we propose the penalized EL approach by combining the profile EL and the smoothly clipped absolute deviation (SCAD; Fan and Li 2001) method together in Sect. 4.

In specific, the proposed QIF procedure is based on the matrix expansion idea, which neither assumes the exact knowledge of the true correlation structure nor estimates the parameters of the correlation structure. Alternatively, the hybrid GEE method combines multiple GEEs based on different working correlation models to improve the estimation efficiency of the GEE method in Liang and Zeger (1986). The resulting EL ratios are shown to have different asymptotically weighted sum Chi-squares, which can be used to construct the corresponding confidence regions. Furthermore, it can be seen that penalized EL efficiently selects significant variables and estimates parameters simultaneously. With a proper choice of the tuning parameters, the penalized estimators based on the QIF and hybrid GEE methods are consistent and have the oracle property. The penalized EL method can make inference for the parameters in the selected model without estimating their estimators' covariance. In addition, we propose an algorithm for computing the penalized EL estimators by the local quadratic approximation. The proposed EL inference procedure is readily implemented by existing R packages.

The rest of this paper is organized as follows. After presenting the parametric dropout propensity and instrument approach, we construct the proposed estimators based on the QIF and hybrid GEE methods in Section 2 and investigate the statistical properties in Section 3. In Section 4, we introduce the penalized EL estimators and the algorithm for variable selection. We discuss the unbalanced data case in Sect. 5. Simulation studies are given in Section 6. Section 7 analyzes the AIDS Clinical Trial Group 193A data for illustration. Some discussions can be found in Sect. 8. All technical details are provided in the Supplementary Material.

2 Methodology

2.1 Nonignorable dropout and bias-corrected GEE

We first consider the longitudinal data are balanced with the same cluster size, i.e., $m_i = m$, while the unbalanced longitudinal data will be investigated in Section 5 later. As we discussed in Section 1, to address the identifiability problem, $\boldsymbol{x}_i$ can be decomposed as two parts, i.e., $\boldsymbol{x}_i = (\boldsymbol{u}_i, z_i)$. Furthermore, for longitudinal $\boldsymbol{y}_i$, it is reasonable to assume that the dropout at time point j is unrelated to the future values $y_{i(j+1)}, \dots, y_{im}$ (Diggle and Kenward 1994). Thus, we have

$$\begin{aligned} &\Pr(r_{ij} = 1 | r_{i(j-1)} = 1, \boldsymbol{x}_i, \boldsymbol{y}_i) = \Pr(r_{ij} = 1 | r_{i(j-1)} = 1, \vec{\boldsymbol{u}}_{ij}, \vec{\boldsymbol{y}}_{ij}), \\ &\Pr(r_{ij} = 1 | r_{i(j-1)} = 0, \boldsymbol{x}_i, \boldsymbol{y}_i) = 0, \text{ for } j = 1, \dots, m, \end{aligned} \tag{2}$$

where $\vec{\boldsymbol{u}}_{ij} = (\boldsymbol{u}_{i1}^T, \dots, \boldsymbol{u}_{ij}^T)^T$, $\vec{\boldsymbol{z}}_{ij} = (z_{i1}^T, \dots, z_{ij}^T)^T$ and $\vec{\boldsymbol{y}}_{ij} = (y_{i1}, \dots, y_{ij})^T$ are denoted as the histories $\boldsymbol{u}_{ij}$, z_{ij} and y_{ij} up to and including cycle j, respectively. The first line in (2) indicates that dropout is nonignorable, i.e., the probability of observing y_{ij} at time j depends on y_{ij} regardless of whether y_{ij} is observed or not; the second line reflects the dropout or monotone missing data pattern. Further, we assume that the dropout propensity in (2) has a parametric form,

$$\Pr(r_{ij} = 1 | r_{i(j-1)} = 1, \vec{\boldsymbol{u}}_{ij}, \vec{\boldsymbol{y}}_{ij}) = \Psi(\alpha_j + \boldsymbol{\gamma}_j^T \mathcal{O}_{ij}), \ j = 1, \dots, m, \tag{3}$$

where $\mathcal{O}_{ij} = (\vec{\boldsymbol{u}}_{ij}^T, \vec{\boldsymbol{y}}_{ij}^T)^T$, α_j is unknown parameter, $\boldsymbol{\gamma}_j$ is a column vector of unknown parameters, Ψ is a known monotone function defined on [0, 1] and r_{i0} is always defined to be 1. Popular choices of Ψ are the logistic function with $\Psi(t) = \{1 + \exp(t)\}^{-1}$ and the probit function with Ψ being the standard normal distribution function. In applications, we may consider some special cases of (3). For example, Tang et al. (2003) considered that

$$\Pr(r_{ij} = 1 | r_{i(j-1)} = 1, \vec{\boldsymbol{u}}_{ij}, \vec{\boldsymbol{y}}_{ij}) = \Psi(\alpha_j + \gamma_j y_{ij}), \ j = 1, \dots, m. \tag{4}$$

The following assumption between (3) and (4) can also be considered,

$$\Pr(r_{ij}=1|r_{i(j-1)}=1,\vec{\boldsymbol{u}}_{ij},\vec{\boldsymbol{y}}_{ij})=\Psi(\alpha_j+\boldsymbol{\gamma}_{j1}^T\boldsymbol{u}_{ij}+\gamma_{j2}y_{ij}),\ j=1,\dots,m. \tag{5}$$

Model (5) is used in our simulation studies.

For $j=1,\dots,m$, write $\boldsymbol{\theta}_j=(\alpha_j,\boldsymbol{\gamma}_j^T)^T$ and define the following estimating equations

$$\boldsymbol{s}_j(\boldsymbol{y}_i,\boldsymbol{x}_i,\boldsymbol{r}_i,\boldsymbol{\theta}_j)=r_{i(j-1)}\left\{\frac{r_{ij}}{\Psi(\alpha_j+\boldsymbol{\gamma}_j^T\mathcal{O}_{ij})}-1\right\}\left(1,\vec{\boldsymbol{u}}_{ij}^T,\vec{\boldsymbol{z}}_{ij}^T,\vec{\boldsymbol{y}}_{i(j-1)}^T\right)^T. \tag{6}$$

If $\boldsymbol{\theta}_j^0$ is the true value of $\boldsymbol{\theta}_j$, it can be verified that $E\{\boldsymbol{s}_j(\boldsymbol{y}_i,\boldsymbol{x}_i,\boldsymbol{r}_i,\boldsymbol{\theta}_j^0)\}=0$. The efficient two-step GMM (Hansen 1982) estimator of $\boldsymbol{\theta}_j$ is

$$\hat{\boldsymbol{\theta}}_j=\text{argmin}_{\boldsymbol{\theta}_j}\bar{\boldsymbol{s}}_j(\boldsymbol{\theta}_j)^T\hat{\boldsymbol{\Omega}}_j^{-1}\bar{\boldsymbol{s}}_j(\boldsymbol{\theta}_j), \tag{7}$$

where $\hat{\boldsymbol{\Omega}}_j^{-1}$ is the inverse of the matrix $n^{-1}\sum_{i=1}^n\boldsymbol{s}_j(\boldsymbol{y}_i,\boldsymbol{x}_i,\boldsymbol{r}_i,\hat{\boldsymbol{\theta}}_j^{(1)})\boldsymbol{s}_j(\boldsymbol{y}_i,\boldsymbol{x}_i,\boldsymbol{r}_i,\hat{\boldsymbol{\theta}}_j^{(1)})^T$, $\hat{\boldsymbol{\theta}}_j^{(1)}=\text{argmin}_{\boldsymbol{\theta}_j}\bar{\boldsymbol{s}}_j(\boldsymbol{\theta}_j)^T\bar{\boldsymbol{s}}_j(\boldsymbol{\theta}_j)$ and $\bar{\boldsymbol{s}}_j(\boldsymbol{\theta}_j)=n^{-1}\sum_{i=1}^n\boldsymbol{s}_j(\boldsymbol{y}_i,\boldsymbol{x}_i,\boldsymbol{r}_i,\boldsymbol{\theta}_j)$. For any $j=1,\dots,m$, let $\boldsymbol{\Theta}_j=(\boldsymbol{\theta}_1^T,\dots,\boldsymbol{\theta}_j^T)^T$ be the joint parameters vector up to and including cycle j. Define $\pi_{ij}=\Pr(r_{ij}=1|\boldsymbol{x}_i,\boldsymbol{y}_i)=\Pr(r_{ij}=1|\vec{\boldsymbol{u}}_{ij},\vec{\boldsymbol{y}}_{ij})$. Then, under the model (3), $\pi_{ij}=\prod_{t=1}^j\Pr\{r_{it}=1|r_{i(t-1)}=1,\vec{\boldsymbol{u}}_{it},\vec{\boldsymbol{y}}_{it}\}=\prod_{t=1}^j\Psi(\alpha_t+\boldsymbol{\gamma}_t^T\mathcal{O}_{it})\triangleq\pi_{ij}(\boldsymbol{\Theta}_j)$, which can be estimated by

$$\pi_{ij}(\hat{\boldsymbol{\Theta}}_j)=\prod_{t=1}^j\Psi(\hat{\alpha}_t+\hat{\boldsymbol{\gamma}}_t^T\mathcal{O}_{it}),$$

where $\hat{\boldsymbol{\Theta}}_j=(\hat{\boldsymbol{\theta}}_1^T,\dots,\hat{\boldsymbol{\theta}}_j^T)^T$ are the GMM estimators under the dropout propensity model (3). Motivated by Liang and Zeger (1986), the bias-corrected GEE can be written as

$$\sum_{i=1}^n\dot{\boldsymbol{\mu}}_i^T\boldsymbol{V}_i^{-1}\hat{\boldsymbol{W}}_i(\boldsymbol{y}_i-\boldsymbol{\mu}_i)=0, \tag{8}$$

where $\hat{\boldsymbol{W}}_i=diag(r_{i1}/\hat{\pi}_{i1},\dots,r_{im}/\hat{\pi}_{im})$, $\boldsymbol{V}_i$ is the covariance matrix of $(\boldsymbol{y}_i-\boldsymbol{\mu}_i)$, $\dot{\boldsymbol{\mu}}_i=\partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}$, $\boldsymbol{\mu}_i=(\mu_{i1},\cdots,\mu_{im})^T$. The inverse of covariance matrix $\boldsymbol{V}_i^{-1}$ can be decomposed as $\boldsymbol{A}_i^{-1/2}\boldsymbol{\Phi}_i^{-1}\boldsymbol{A}_i^{-1/2}$, with $\boldsymbol{A}_i=diag\{Var(y_{i1}),\cdots,Var(y_{im})\}$ being a $(m\times m)$-dimensional diagonal marginal variance matrix of $(\boldsymbol{y}_i-\boldsymbol{\mu}_i)$ and $\boldsymbol{\Phi}_i$ being an $(m\times m)$-dimensional true correlation matrix. In practice, $\boldsymbol{\Phi}_i$ is unknown and a working correlation structure, donated by $\boldsymbol{R}_i$, is utilized. Some common working correlation structures include independent structure, compound symmetry (CS) and first-order autoregressive (AR(1)). If the working covariance matrix $\boldsymbol{R}_i=\boldsymbol{I}_m$, the $m\times m$ identity matrix, it assumes working independence structure; when $\boldsymbol{R}_i=\boldsymbol{\Phi}_i$, it assumes the true within-subject correlation structure for longitudinal data.

2.2 EL inference based on QIF and hybrid GEE

Since the working covariance matrix $\boldsymbol{R}_i^{-1}$ is unknown in practice, misspecification of the working covariance matrix $\boldsymbol{R}_i^{-1}$ will lead to less efficient GLM estimators. To improve the efficiency of estimation, we borrow the matrix expansion idea of Qu et al. (2000) and propose the quadratic inference function (QIF) by assuming that the inverse of the working correlation $\boldsymbol{R}_i^{-1}$ can be approximated by a linear combination of several basis matrices, that is,

$$\boldsymbol{R}_i^{-1} = \sum_{j=1}^{q} b_j \boldsymbol{B}_j, \tag{9}$$

where $\boldsymbol{B}_1, \dots, \boldsymbol{B}_q$ are $(m \times m)$-dimensional symmetric basic matrices depending on the particular choice of $\boldsymbol{R}_i^{-1}$ and $b_1, \dots, b_q$ are unknown coefficients. For example, if a working correlation structure is CS, then $\boldsymbol{R}_i^{-1} = b_1\boldsymbol{B}_1 + b_2\boldsymbol{B}_2$, where $\boldsymbol{B}_1$ is an identity matrix and $\boldsymbol{B}_2$ is a symmetric matrix with 0 on the diagonal and 1 elsewhere. The coefficients b_0 and b_1 are parameters associated with the CS correlation. If $\boldsymbol{R}_i^{-1}$ corresponds to AR(1), $\boldsymbol{R}_i^{-1} = b_1\boldsymbol{B}_1 + b_2\boldsymbol{B}_2 + b_3\boldsymbol{B}_3$, where $\boldsymbol{B}_1$ is an identity matrix, $\boldsymbol{B}_2$ is a symmetric matrix with 1 on the sub-diagonal entries and 0 elsewhere, and $\boldsymbol{B}_3$ is a symmetric matrix with 1 in elements (1, 1) and (m, m), and 0 elsewhere. More details can be found in Qu et al. (2000) and Cho and Qu (2015).

Substituting (9) into (8) leads to

$$\sum_{i=1}^{n} \dot{\boldsymbol{\mu}}_i^T \boldsymbol{A}_i^{-1/2}(b_1\boldsymbol{B}_1 + \cdots + b_q\boldsymbol{B}_q)\boldsymbol{A}_i^{-1/2}\hat{\mathbf{W}}_i(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = 0. \tag{10}$$

Consequently, Eq. (10) can be approximated as a linear combination of elements, $\hat{\boldsymbol{g}}_i(\boldsymbol{\beta})$, for $i = 1, \dots, n$, where

$$\hat{\boldsymbol{g}}_i(\boldsymbol{\beta}) = \begin{pmatrix} \dot{\boldsymbol{\mu}}_i^T \boldsymbol{A}_i^{-1/2}\boldsymbol{B}_1\boldsymbol{A}_i^{-1/2}\hat{\mathbf{W}}_i(\boldsymbol{y}_i - \boldsymbol{\mu}_i) \\ \vdots \\ \dot{\boldsymbol{\mu}}_i^T \boldsymbol{A}_i^{-1/2}\boldsymbol{B}_q\boldsymbol{A}_i^{-1/2}\hat{\mathbf{W}}_i(\boldsymbol{y}_i - \boldsymbol{\mu}_i) \end{pmatrix}. \tag{11}$$

Note that estimation of the parameters $b_1, \dots, b_q$ is not required, since the function $\hat{\boldsymbol{g}}_i(\boldsymbol{\beta})$ does not involve the parameters, and $\hat{\boldsymbol{g}}_i(\boldsymbol{\beta})$ is an overdetermined equations with pq variate function. Thus, we propose to apply the following EL for the inference of $\boldsymbol{\beta}$ under some regular conditions. Let p_i represent the probability weight allocated to $\hat{\boldsymbol{g}}_i(\boldsymbol{\beta})$, $i = 1, \dots, n$. The empirical log-likelihood ratio function for $\boldsymbol{\beta}$ based on the QIF approach is defined as

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

By using the Lagrange multiplier method, $\hat{R}_Q(\boldsymbol{\beta})$ can be represented as

$$\hat{R}_Q(\boldsymbol{\beta}) = 2\sum_{i=1}^{n} \log\{1 + \boldsymbol{\lambda}^T \hat{\boldsymbol{g}}_i(\boldsymbol{\beta})\},$$

where $\boldsymbol{\lambda}^T$ is the root of the following equation:

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\hat{\boldsymbol{g}}_i(\boldsymbol{\beta})}{1 + \lambda^T \hat{\boldsymbol{g}}_i(\boldsymbol{\beta})} = 0.$$

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

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

Alternatively, Liang and Zeger (1986) assumed that the matrix $\boldsymbol{V}_i$ can be expressed in terms of a working correlation matrix $\boldsymbol{R}(\alpha)$ as $\boldsymbol{V}_i = \boldsymbol{A}_i^{1/2}\boldsymbol{R}(\alpha)\boldsymbol{A}_i^{1/2}$, where α is some unknown nuisance parameter. Thus, one can obtain the following GEE,

$$\sum_{i=1}^{n} \dot{\boldsymbol{\mu}}_i^T \boldsymbol{A}_i^{-1/2} \boldsymbol{R}^{-1}(\alpha) \boldsymbol{A}_i^{-1/2} \hat{\boldsymbol{W}}_i(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = 0. \tag{12}$$

Note that, if the working correlation $\boldsymbol{R}(\alpha)$ is misspecified, the resulting estimator of the parameters $\boldsymbol{\beta}$ based on (12) is still consistent, but it may not be efficient. In order to improve the efficiency, motivated by Leung et al. (2009), we propose a hybrid method that combines multiple GEEs based on different and linearly independent choices of $\boldsymbol{R}(\alpha)$, say $\boldsymbol{R}^l(\alpha)$, $l = 1, \cdots, L$. Let

$$\hat{\boldsymbol{h}}_i(\boldsymbol{\beta}) = \begin{pmatrix} \dot{\boldsymbol{\mu}}_i^T \boldsymbol{A}_i^{-1/2} \{\boldsymbol{R}^1(\alpha)\}^{-1} \boldsymbol{A}_i^{-1/2} \hat{\boldsymbol{W}}_i(\boldsymbol{y}_i - \boldsymbol{\mu}_i) \\ \vdots \\ \dot{\boldsymbol{\mu}}_i^T \boldsymbol{A}_i^{-1/2} \{\boldsymbol{R}^L(\alpha)\}^{-1} \boldsymbol{A}_i^{-1/2} \hat{\boldsymbol{W}}_i(\boldsymbol{y}_i - \boldsymbol{\mu}_i) \end{pmatrix}, \tag{13}$$

and p_i represent the probability weight allocated to $\hat{\boldsymbol{h}}_i(\boldsymbol{\beta})$, $i = 1, \ldots, n$. The empirical log-likelihood ratio function for $\boldsymbol{\beta}$ based on the hybrid GEE approach is defined as

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

By using the Lagrange multiplier method, $\hat{R}_H(\boldsymbol{\beta})$ can be represented as

$$\hat{R}_H(\boldsymbol{\beta}) = 2\sum_{i=1}^{n} \log\{1 + \boldsymbol{\lambda}^T \hat{\boldsymbol{h}}_i(\boldsymbol{\beta})\},$$

where $\boldsymbol{\lambda}^T$ is the root of the following equation:

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\hat{\boldsymbol{h}}_i(\boldsymbol{\beta})}{1 + \lambda^T \hat{\boldsymbol{h}}_i(\boldsymbol{\beta})} = 0.$$

The maximum EL estimator based on $\hat{\boldsymbol{h}}_i(\boldsymbol{\beta})$, denoted as $\hat{\boldsymbol{\beta}}_H$, can be obtained as below:

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

In practice, a few popular choices of $\boldsymbol{R}^{-1}(\alpha)$ can be applied, and we use the CS andAR(1) in the simulations.

3 Asymptotic theories

Assume $\boldsymbol{\beta}^0$ and $\boldsymbol{\Theta}_m^0$ are the true values of $\boldsymbol{\beta}$ and $\boldsymbol{\Theta}_m$, respectively. Note that, $\hat{\boldsymbol{g}}_i(\boldsymbol{\beta}) = \boldsymbol{g}_i(\hat{\boldsymbol{\Theta}}_m, \boldsymbol{\beta})$ and $\hat{\boldsymbol{h}}_i(\boldsymbol{\beta}) = \boldsymbol{h}_i(\hat{\boldsymbol{\Theta}}_m, \boldsymbol{\beta})$. Subsequently, define $\boldsymbol{\Delta}_{\boldsymbol{g}} = E[\partial \boldsymbol{g}_i(\boldsymbol{\Theta}_m^0, \boldsymbol{\beta}^0)/\partial \boldsymbol{\beta}]$, $\boldsymbol{\Lambda}_{\boldsymbol{g}} = E\{\boldsymbol{g}_i(\boldsymbol{\Theta}_m^0, \boldsymbol{\beta}^0)[\boldsymbol{g}_i(\boldsymbol{\Theta}_m^0, \boldsymbol{\beta}^0)]^T\}$, $\boldsymbol{\Gamma}_{\boldsymbol{g}} = \{\boldsymbol{\Delta}_{\boldsymbol{g}}^T \boldsymbol{\Lambda}_{\boldsymbol{g}}^{-1} \boldsymbol{\Delta}_{\boldsymbol{g}}\}^{-1} \boldsymbol{\Delta}_{\boldsymbol{g}}^T \boldsymbol{\Lambda}_{\boldsymbol{g}}^{-1}$, $\boldsymbol{\Delta}_{\boldsymbol{h}} = E[\partial \boldsymbol{h}_i(\boldsymbol{\Theta}_m^0, \boldsymbol{\beta}^0)/\partial \boldsymbol{\beta}]$, $\boldsymbol{\Lambda}_{\boldsymbol{h}} = E\{\boldsymbol{h}_i(\boldsymbol{\Theta}_m^0, \boldsymbol{\beta}^0)[\boldsymbol{h}_i(\boldsymbol{\Theta}_m^0, \boldsymbol{\beta}^0)]^T\}$ and $\boldsymbol{\Gamma}_{\boldsymbol{h}} = \{\boldsymbol{\Delta}_{\boldsymbol{h}}^T \boldsymbol{\Lambda}_{\boldsymbol{h}}^{-1} \boldsymbol{\Delta}_{\boldsymbol{h}}\}^{-1} \boldsymbol{\Delta}_{\boldsymbol{h}}^T \boldsymbol{\Lambda}_{\boldsymbol{h}}^{-1}$.

Theorem 1 *Suppose that $\boldsymbol{\theta}_j^0$ is the unique solution to $E\{\boldsymbol{s}_j(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{r}_i, \boldsymbol{\theta}_j)\} = 0$ and models* (1–2) *hold, $\boldsymbol{\Omega}_j = E\{\boldsymbol{s}_j(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{r}_i, \boldsymbol{\theta}_j^0)\boldsymbol{s}_j(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{r}_i, \boldsymbol{\theta}_j^0)^T\}$ is positive definite and the matrix $\boldsymbol{\Upsilon}_j = E[\partial \boldsymbol{s}_j(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{r}_i, \boldsymbol{\theta}_j^0)/\partial \boldsymbol{\theta}_j]$ is of full rank. As $n \to \infty$, $\sqrt{n}(\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j^0) \to N(0, (\boldsymbol{\Upsilon}_j^T \boldsymbol{\Omega}_j \boldsymbol{\Upsilon}_j)^{-1})$ and $\sqrt{n}(\hat{\boldsymbol{\Theta}}_m - \boldsymbol{\Theta}_m) \to N(0, \boldsymbol{\Sigma})$ in distributions. Under the conditions* (C1–C4) *in the Supplementary Material, as $n \to \infty$, we have*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_Q - \boldsymbol{\beta}^0) \longrightarrow N(0, \boldsymbol{\Gamma}_{\boldsymbol{g}} \boldsymbol{\Sigma}_{\boldsymbol{g}} \boldsymbol{\Gamma}_{\boldsymbol{g}}^T), \quad \sqrt{n}(\hat{\boldsymbol{\beta}}_H - \boldsymbol{\beta}^0) \longrightarrow N(0, \boldsymbol{\Gamma}_{\boldsymbol{h}} \boldsymbol{\Sigma}_{\boldsymbol{h}} \boldsymbol{\Gamma}_{\boldsymbol{h}}^T),$$

where $\boldsymbol{\Sigma}_{\boldsymbol{g}} = \boldsymbol{\Lambda}_{\boldsymbol{g}} + E[\partial \boldsymbol{g}_i(\boldsymbol{\Theta}_m^0, \boldsymbol{\beta}^0)/\partial \boldsymbol{\Theta}_m]\boldsymbol{\Sigma} E^T[\partial \boldsymbol{g}_i(\boldsymbol{\Theta}_m^0, \boldsymbol{\beta}^0)/\partial \boldsymbol{\Theta}_m]$ and $\boldsymbol{\Sigma}_{\boldsymbol{h}} = \boldsymbol{\Lambda}_{\boldsymbol{h}} + E[\partial \boldsymbol{h}_i(\boldsymbol{\Theta}_m^0, \boldsymbol{\beta}^0)/\partial \boldsymbol{\Theta}_m]\boldsymbol{\Sigma} E^T[\partial \boldsymbol{h}_i(\boldsymbol{\Theta}_m^0, \boldsymbol{\beta}^0)/\partial \boldsymbol{\Theta}_m]$.

Remark 1 If π_{ij} is known, it can be verified that $E[\partial \boldsymbol{g}_i(\boldsymbol{\Theta}_m^0, \boldsymbol{\beta}^0)/\partial \boldsymbol{\Theta}_m] = 0$ and $E[\partial \boldsymbol{h}_i(\boldsymbol{\Theta}_m^0, \boldsymbol{\beta}^0)/\partial \boldsymbol{\Theta}_m] = 0$, and the asymptotic covariance matrices of $\hat{\boldsymbol{\beta}}_Q$ and $\hat{\boldsymbol{\beta}}_H$ can be simplified as $\{\boldsymbol{\Delta}_{\boldsymbol{g}}^T \boldsymbol{\Lambda}_{\boldsymbol{g}}^{-1} \boldsymbol{\Delta}_{\boldsymbol{g}}\}^{-1}$ and $\{\boldsymbol{\Delta}_{\boldsymbol{h}}^T \boldsymbol{\Lambda}_{\boldsymbol{h}}^{-1} \boldsymbol{\Delta}_{\boldsymbol{h}}\}^{-1}$, respectively. When there is no missing data, it means $\pi_{ij} = 1$ and the estimating equations are the same as the equations in Li and Pan (2013) and Leung et al. (2009), respectively. In addition, Theorem 1 can be used to construct normal-approximation-based confidence regions.

Next, we will study the asymptotic properties of $\hat{R}_Q(\boldsymbol{\beta}^0)$ and $\hat{R}_H(\boldsymbol{\beta}^0)$. Compared to the standard empirical log-likelihood ratio without missing data, the main difference is that the $\hat{\boldsymbol{g}}_i(\boldsymbol{\beta}^0)$ and $\hat{\boldsymbol{h}}_i(\boldsymbol{\beta}^0)$, $i = 1, \ldots, n$, are not independent and identically distributed. Hence, the asymptotic distributions of $\hat{R}_Q(\boldsymbol{\beta}^0)$ and $\hat{R}_H(\boldsymbol{\beta}^0)$ may not be standard Chi-squares. Actually, we will show that $\hat{R}_Q(\boldsymbol{\beta}^0)$ and $\hat{R}_H(\boldsymbol{\beta}^0)$ are asymptotically two different weighted sum Chi-squares.

Theorem 2 *Under the regularity conditions in Theorem 1, as $n \to \infty$, we have*

$$\hat{R}_Q(\boldsymbol{\beta}^0) \longrightarrow \rho_1 w_1 + \rho_2 w_2 + \ldots + \rho_{psq} w_{pq},$$
$$\hat{R}_H(\boldsymbol{\beta}^0) \longrightarrow \varrho_1 \varpi_1 + \varrho_2 \varpi_2 + \ldots + \varrho_{pL} \varpi_{pL},$$

where w_l and ϖ_s are independent and follow the standard χ^2 distribution with one degree, the weights ρ_l and ϱ_s are eigenvalues of $\boldsymbol{\Lambda}_{\boldsymbol{g}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{g}}$ and $\boldsymbol{\Lambda}_{\boldsymbol{h}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{h}}$, respectively, $l = 1, \ldots, pq$ and $s = 1, \ldots, pL$.

Remark 2 When there is no missing data, according to Li and Pan (2013), it can be shown that the Wilks theorem holds. However, compared to the standard empirical log-likelihood ratio without missing data, the main difference is that the proposed $\hat{\boldsymbol{g}}_i(\boldsymbol{\beta}^0)$ and $\hat{\boldsymbol{h}}_i(\boldsymbol{\beta}^0)$ are not independent and identically distributed. As a result, the asymptotic distributions of $\hat{R}_Q(\boldsymbol{\beta}^0)$ and $\hat{R}_H(\boldsymbol{\beta}^0)$ may not be the standard Chi-square and the Wilks's theorem breaks down. To be specific, Lemmas 1 and 2 in the Supplementary Material reveal the reasons why Wilks's theorem does not hold. On the other hand, when there is no missing data, we have $\boldsymbol{\Lambda}_{\boldsymbol{g}} = \boldsymbol{\Sigma}_{\boldsymbol{g}}$ and $\boldsymbol{\Lambda}_{\boldsymbol{h}} = \boldsymbol{\Sigma}_{\boldsymbol{h}}$ due to the fact that $E[\partial \boldsymbol{g}_i(\boldsymbol{\Theta}_m^0, \boldsymbol{\beta}^0)/\partial \boldsymbol{\Theta}_m] = 0$ and $E[\partial \boldsymbol{h}_i(\boldsymbol{\Theta}_m^0, \boldsymbol{\beta}^0)/\partial \boldsymbol{\Theta}_m] = 0$, such that both $\boldsymbol{\Lambda}_{\boldsymbol{g}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{g}}$ and $\boldsymbol{\Lambda}_{\boldsymbol{h}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{h}}$ equal to the identity matrix, which makes the Wilks's theorem hold. This is the same as the result of Li and Pan (2013). Moreover, Theorem 2 can be used to test the hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}^0$ and construct the confidence region for $\boldsymbol{\beta}^0$.

Let $r_Q(\boldsymbol{\beta}^0) = (pq)/\mathrm{tr}\{\boldsymbol{\Lambda}_{\boldsymbol{g}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{g}}\}$ and $r_H(\boldsymbol{\beta}^0) = (pL)/\mathrm{tr}\{\boldsymbol{\Lambda}_{\boldsymbol{h}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{h}}\}$ be the adjustment factors. Along the lines of Rao and Scott (1981), we have the following corollary.

Corollary 1 *Under the conditions of Theorem 1, as $n \to \infty$, we obtain*

$$\hat{R}_Q(\boldsymbol{\beta}^0) r_Q(\boldsymbol{\beta}^0) \longrightarrow \chi^2_{pq}, \ \hat{R}_H(\boldsymbol{\beta}^0) r_H(\boldsymbol{\beta}^0) \longrightarrow \chi^2_{pL}.$$

To construct the confidence regions of $\boldsymbol{\beta}$, we propose to obtain the estimators $\hat{\boldsymbol{\Lambda}}_{\boldsymbol{g}}^{-1}$, $\hat{\boldsymbol{\Lambda}}_{\boldsymbol{h}}^{-1}$, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{g}}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{h}}$ of $\boldsymbol{\Lambda}_{\boldsymbol{g}}^{-1}$, $\boldsymbol{\Lambda}_{\boldsymbol{h}}^{-1}$, $\boldsymbol{\Sigma}_{\boldsymbol{g}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{h}}$ by the plug-in method, and then obtain the consistent estimators $\hat{\rho}_1, \ldots, \hat{\rho}_{pq}$ and $\hat{\varrho}_1, \ldots, \hat{\varrho}_{pL}$ of $\rho_1, \ldots, \rho_{pq}$ and $\varrho_1, \ldots, \varrho_{pL}$, respectively. Let c_α^Q and c_α^H be the $1-\alpha$ quantiles of $\hat{\rho}_1 w_1 + \ldots + \hat{\rho}_{pq} w_{pq}$ and $\hat{\varrho}_1 w_1 + \ldots + \hat{\varrho}_{pL} w_{pL}$ for $0 < \alpha < 1$, respectively. According to Theorem 2, the approximate $100(1-\alpha)\%$ confidence regions for $\boldsymbol{\beta}$ based on the QIF and hybrid GEE methods are given by

$$\mathrm{CI}_1^Q(\alpha) = \{\boldsymbol{\beta} : \hat{R}_Q(\boldsymbol{\beta}) < c_\alpha^Q\}, \ \mathrm{CI}_1^H(\alpha) = \{\boldsymbol{\beta} : \hat{R}_H(\boldsymbol{\beta}) < c_\alpha^H\}.$$

Alternatively, based on Corollary 1, the $100(1-\alpha)\%$ confidence regions can also be obtained by

$$\mathrm{CI}_2^Q(\alpha) = \{\boldsymbol{\beta} : r_Q(\hat{\boldsymbol{\beta}})\hat{R}_Q(\boldsymbol{\beta}) < \chi^2_{pq,1-\alpha}\}, \ \mathrm{CI}_2^H(\alpha) = \{\boldsymbol{\beta} : r_H(\hat{\boldsymbol{\beta}})\hat{R}_H(\boldsymbol{\beta}) < \chi^2_{pL,1-\alpha}\}$$

where $r_Q(\hat{\boldsymbol{\beta}}) = (pq)/\mathrm{tr}(\hat{\boldsymbol{\Lambda}}_{\boldsymbol{g}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{g}})$ and $r_H(\hat{\boldsymbol{\beta}}) = (pL)/\mathrm{tr}(\hat{\boldsymbol{\Lambda}}_{\boldsymbol{h}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{h}})$.

4 Variable selection

When the dimension of covariate $\boldsymbol{x}_{ij}$ is high, in order to build robust models and identify relevant predictors to the response variable, variable selection in the GLM should be considered. For this purpose, we propose the penalized empirical likelihood (PEL) by combining the profile EL method and the smoothly clipped absolute deviation (SCAD) together. The PEL estimator is defined to be the minimizer of the following objective function, which is still denoted as $\hat{\boldsymbol{\beta}}$ for simplicity.

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

where $\hat{\boldsymbol{\eta}}_i(\boldsymbol{\beta}) = \hat{\boldsymbol{g}}_i(\boldsymbol{\beta})$ or $\hat{\boldsymbol{h}}_i(\boldsymbol{\beta})$, $p_\nu(t)$ is a penalty function with tuning parameter ν. We use the SCAD penalty, which is defined in terms of its first derivative and is symmetric around the origin. For $t > 0$, its first derivative is

$$p'_\nu(t) = \nu\{I(t \le \nu) + \frac{(a\nu - t)_+}{(a-1)\nu} I(t > \nu)\},$$

where $a > 2$ and $\nu > 0$ are tuning parameters. We choose $a = 3.7$ suggested by Fan and Li (2001).

Let $\mathcal{A}$ be the set of nonzero components of true parameter vector $\boldsymbol{\beta}^0$ and its cardinality as $d = |\mathcal{A}|$. Without loss of generality, one can partition the parameter vector as $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$, where $\boldsymbol{\beta}_1 \in R^d$ and $\boldsymbol{\beta}_2 \in R^{p-d}$. Hence, the true parameter $\boldsymbol{\beta}^0 = (\boldsymbol{\beta}_1^{0T}, 0^T)^T$, and we write $\hat{\boldsymbol{\beta}}_Q = (\hat{\boldsymbol{\beta}}_{Q_1}^T, \hat{\boldsymbol{\beta}}_{Q_2}^T)^T$ and $\hat{\boldsymbol{\beta}}_H = (\hat{\boldsymbol{\beta}}_{H_1}^T, \hat{\boldsymbol{\beta}}_{H_2}^T)^T$ as the resulting penalized estimators based on the QIF and hybrid GEE methods, respectively. The following theorem shows the selection consistency and asymptotic efficiency of the proposed PEL estimators $\hat{\boldsymbol{\beta}}_Q$ and $\hat{\boldsymbol{\beta}}_H$.

Theorem 3 *Under the regularity conditions in Theorem 1, we further assume conditions* (C5)-(C6) *hold. As $n \to \infty$, the estimators $\hat{\boldsymbol{\beta}}_Q$ and $\hat{\boldsymbol{\beta}}_H$ satisfies*

(i) *(Selection consistency): With probability tending to 1, $\hat{\boldsymbol{\beta}}_{Q_2} = 0$ and $\hat{\boldsymbol{\beta}}_{H_2} = 0$;*
(ii) *(Asymptotic efficiency):*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{Q_1} - \boldsymbol{\beta}_1^0) \longrightarrow N(0, \boldsymbol{\Gamma}_{\boldsymbol{g}}^{(11)} \boldsymbol{\Sigma}_{\boldsymbol{g}}^{(11)} \{\boldsymbol{\Gamma}_{\boldsymbol{g}}^{(11)}\}^T),$$
$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{H_1} - \boldsymbol{\beta}_1^0) \longrightarrow N(0, \boldsymbol{\Gamma}_{\boldsymbol{h}}^{(11)} \boldsymbol{\Sigma}_{\boldsymbol{h}}^{(11)} \{\boldsymbol{\Gamma}_{\boldsymbol{h}}^{(11)}\}^T),$$

where $\boldsymbol{\Lambda}_{\boldsymbol{g}}^{(11)}$ and $\boldsymbol{\Lambda}_{\boldsymbol{h}}^{(11)}$ are $dq \times dq$ and $dL \times dL$ submatrices of $\boldsymbol{\Lambda}_{\boldsymbol{g}}$ and $\boldsymbol{\Lambda}_{\boldsymbol{h}}$, $\boldsymbol{\Delta}_{\boldsymbol{g}}^{(12)}$ and $\boldsymbol{\Delta}_{\boldsymbol{h}}^{(12)}$ are $dq \times d$ and $dL \times d$ submatrices of $\boldsymbol{\Gamma}_{\boldsymbol{g}}$ and $\boldsymbol{\Gamma}_{\boldsymbol{h}}$, $\boldsymbol{\Sigma}_{\boldsymbol{g}}^{(11)}$ and $\boldsymbol{\Sigma}_{\boldsymbol{h}}^{(11)}$ are $dq \times dq$ and $dL \times dL$ submatrices of $\boldsymbol{\Sigma}_{\boldsymbol{g}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{h}}$, $\boldsymbol{\Gamma}_{\boldsymbol{g}}^{(11)} = [\{\boldsymbol{\Delta}_{\boldsymbol{g}}^{(12)}\}^T\{\boldsymbol{\Lambda}_{\boldsymbol{g}}^{(11)}\}^{-1}\{\boldsymbol{\Delta}_{\boldsymbol{g}}^{(12)}\}]^{-1}\{\boldsymbol{\Delta}_{\boldsymbol{g}}^{(12)}\}^T\{\boldsymbol{\Lambda}_{\boldsymbol{g}}^{(11)}\}^{-1}$, $\boldsymbol{\Gamma}_{\boldsymbol{h}} = [\{\boldsymbol{\Delta}_{\boldsymbol{h}}^{(12)}\}^T\{\boldsymbol{\Lambda}_{\boldsymbol{h}}^{(11)}\}^{-1}\{\boldsymbol{\Delta}_{\boldsymbol{h}}^{(12)}\}]^{-1}\{\boldsymbol{\Delta}_{\boldsymbol{h}}^{(12)}\}^T\{\boldsymbol{\Lambda}_{\boldsymbol{h}}^{(11)}\}^{-1}$. More details can be seen in the Supplementary Material.

For the proposed method with SCAD penalty, we apply the local quadratic approximation (LQA) to the penalty function as discussed in Fan and Li (2001). That is,

$$p_\nu(|\beta_j|) \approx p_\nu(|\beta_{j0}|) + p'_\nu(|\beta_{j0}|)/|\beta_{j0}|(|\beta_j|^2 - |\beta_{j0}|^2), \quad \text{for } |\beta_j| \approx |\beta_{j0}|.$$

To iterate on $\boldsymbol{\beta}$ directly, we adopt an approximate algorithm by using the full model expression. Suppose that we are given an initial $\boldsymbol{\beta}^{(0)}$, the solution of the estimating equations $\sum_{i=1}^n \hat{\boldsymbol{\eta}}_i(\boldsymbol{\beta}) = 0$. Then, the optimization of the PEL function can be carried out using a modified Newton-Raphson algorithm. That is to say, for $k = 0, 1, 2, \ldots,$ we generate an iterative sequence as

$$\begin{aligned}
\boldsymbol{\beta}^{(k+1)} &= \boldsymbol{\beta}^{(k)} + \{\boldsymbol{Z}_1(\boldsymbol{\beta}^{(k)}) + \boldsymbol{\Sigma}_{\nu,\boldsymbol{\beta}^{(k)}}\}^{-1}\{\boldsymbol{Z}_2(\boldsymbol{\beta}^{(k)}) - \boldsymbol{U}_\nu(\boldsymbol{\beta}^{(k)})\},\\
\boldsymbol{\Sigma}_{\nu,\boldsymbol{\beta}} &= \operatorname{diag}\{p'_\nu(|\beta_1|)/|\beta_1|, \ldots, p'_\nu(|\beta_p|)/|\beta_p|\},
\end{aligned}$$

where $\boldsymbol{U}_\nu(\boldsymbol{\beta}) = \boldsymbol{\Sigma}_{\nu,\boldsymbol{\beta}}\boldsymbol{\beta}$ and

$$\begin{aligned}
\boldsymbol{Z}_1(\boldsymbol{\beta}) &= \frac{\partial \boldsymbol{T}_{2n}(\boldsymbol{\beta}, 0)}{\partial \boldsymbol{\lambda}}\left\{\frac{\partial \boldsymbol{T}_{1n}(\boldsymbol{\beta}, 0)}{\partial \boldsymbol{\lambda}}\right\}^{-1}\frac{\partial \boldsymbol{T}_{1n}(\boldsymbol{\beta}, 0)}{\partial \boldsymbol{\beta}},\\
\boldsymbol{Z}_2(\boldsymbol{\beta}) &= \frac{\partial \boldsymbol{T}_{2n}(\boldsymbol{\beta}, 0)}{\partial \boldsymbol{\lambda}}\left\{\frac{\partial \boldsymbol{T}_{1n}(\boldsymbol{\beta}, 0)}{\partial \boldsymbol{\lambda}}\right\}^{-1}\boldsymbol{T}_{1n}(\boldsymbol{\beta}, 0),
\end{aligned}$$

with

$$\begin{aligned}
\frac{\partial \boldsymbol{T}_{1n}(\boldsymbol{\beta}, 0)}{\partial \boldsymbol{\beta}} &= \frac{\partial \boldsymbol{T}_{2n}(\boldsymbol{\beta}, 0)}{\partial \boldsymbol{\lambda}} = \frac{1}{n}\sum_{i=1}^n \frac{\partial \hat{\boldsymbol{\eta}}_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}, \quad \frac{\partial \boldsymbol{T}_{1n}(\boldsymbol{\beta}, 0)}{\partial \boldsymbol{\lambda}^T} = -\frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{\eta}}_i(\boldsymbol{\beta})\hat{\boldsymbol{\eta}}_i(\boldsymbol{\beta})^T,\\
\boldsymbol{T}_{1n}(\boldsymbol{\beta}, \boldsymbol{\lambda}) &= \frac{1}{n}\sum_{i=1}^n \frac{\hat{\boldsymbol{\eta}}_i(\boldsymbol{\beta})}{1 + \boldsymbol{\lambda}^T\hat{\boldsymbol{\eta}}_i(\boldsymbol{\beta})}, \quad \boldsymbol{T}_{2n}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \frac{1}{n}\sum_{i=1}^n \frac{\{\partial \hat{\boldsymbol{\eta}}_i(\boldsymbol{\beta})/\partial \boldsymbol{\beta}\}^T\boldsymbol{\lambda}}{1 + \boldsymbol{\lambda}^T\hat{\boldsymbol{\eta}}_i(\boldsymbol{\beta})}.
\end{aligned}$$

We can stop the iteration when solutions converge to a satisfying precision. If $\hat{\beta}_j$ is very close to zero, say $|\hat{\beta}_j| < \zeta$ (a prespecified value), we set $\hat{\beta}_j = 0$ and apply the algorithm in Owen (2001) to compute $\hat{\boldsymbol{\lambda}}$.

To choose the optimal value for the tuning parameter, we combine our variable selection method with three information criteria: BIC of Schwarz (1978), BICC of Wang et al. (2009) and EBIC of Chen and Chen (2008). Three BIC-type criteria are defined as follows:

$$\begin{aligned}
BIC(\nu) &= -2\hat{R}_p(\boldsymbol{\beta}_\nu) + \log(n)df_\nu,\\
BICC(\nu) &= -2\hat{R}_p(\boldsymbol{\beta}_\nu) + \max\{1, \log\log(p)\}\log(n)df_\nu,\\
EBIC(\nu) &= -2\hat{R}_p(\boldsymbol{\beta}_\nu) + [\log(n) + 2\log(p)]df_\nu,
\end{aligned}$$

where $\boldsymbol{\beta}_\nu = \boldsymbol{\beta}_Q$ or $\boldsymbol{\beta}_H$ is the estimate of $\boldsymbol{\beta}$ based on the QIF or hybrid GEE methods with ν being the tuning parameter, and df_ν is the number of nonzero coefficients in $\boldsymbol{\beta}_\nu$.

5 Implementation with unbalanced data

The above methods are presented with balanced data, that is, $m_i = m$. In practice, longitudinal data may not be measured with the same cluster size, and could be unbalanced due to experimental constraints. To configure the proposed methods for unbalanced data, we apply the transformation matrix to each cluster. As in Zhou and Qu (2012), we create the largest cluster 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 $\boldsymbol{T}_i$ for the ith cluster by removing the columns of the $m \times m$ identity matrix, where the removed columns correspond to unmeasured data/time points for the ith subject. Through the transformation, $\hat{\boldsymbol{g}}_i(\boldsymbol{\beta})$ is replaced by

$$\hat{\boldsymbol{g}}_i^*(\boldsymbol{\beta}) = \begin{pmatrix} (\dot{\boldsymbol{\mu}}_i^*)^T (\boldsymbol{A}_i^*)^{-1/2} \boldsymbol{B}_1 (\boldsymbol{A}_i^*)^{-1/2} \hat{\boldsymbol{W}}_i (\boldsymbol{y}_i^* - \boldsymbol{\mu}_i^*) \\ \vdots \\ (\dot{\boldsymbol{\mu}}_i^*)^T (\boldsymbol{A}_i^*)^{-1/2} \boldsymbol{B}_q (\boldsymbol{A}_i^*)^{-1/2} \hat{\boldsymbol{W}}_i (\boldsymbol{y}_i^* - \boldsymbol{\mu}_i^*) \end{pmatrix},$$

where $\boldsymbol{y}_i^* = \boldsymbol{T}_i \boldsymbol{y}_i$, $\boldsymbol{\mu}_i^* = \boldsymbol{T}_i \boldsymbol{\mu}_i$, $\dot{\boldsymbol{\mu}}_i^* = \boldsymbol{T}_i \dot{\boldsymbol{\mu}}_i$, $\boldsymbol{A}_i^* = \boldsymbol{T}_i \boldsymbol{A}_i \boldsymbol{T}_i^T$. It can be seen that the components in $\boldsymbol{y}_i^*$ are the same as in $\boldsymbol{y}_i$ for responses for $j \le m_i$ but are 0 for $j > m_i$, and similarly for $\boldsymbol{\mu}_i^*$ and $\dot{\boldsymbol{\mu}}_i^*$, which do not affect the estimation of $\boldsymbol{\beta}$. Correspondingly, $\hat{\boldsymbol{h}}_i(\boldsymbol{\beta})$ is replaced by

$$\hat{\boldsymbol{h}}_i^*(\boldsymbol{\beta}) = \begin{pmatrix} (\dot{\boldsymbol{\mu}}_i^*)^T (\boldsymbol{A}_i^*)^{-1/2} \{\boldsymbol{R}^1(\alpha)\}^{-1} (\boldsymbol{A}_i^*)^{-1/2} \hat{\boldsymbol{W}}_i (\boldsymbol{y}_i^* - \boldsymbol{\mu}_i^*) \\ \vdots \\ (\dot{\boldsymbol{\mu}}_i^*)^T (\boldsymbol{A}_i^*)^{-1/2} \{\boldsymbol{R}^L(\alpha)\}^{-1} (\boldsymbol{A}_i^*)^{-1/2} \hat{\boldsymbol{W}}_i (\boldsymbol{y}_i^* - \boldsymbol{\mu}_i^*) \end{pmatrix}.$$

Therefore, for unbalanced data with dropout, parameter estimation and variable selection also can be implemented using our proposed methods in Sections 3–4. We can show that the asymptotic results of Theorems 1–3 still hold for unequal cluster sizes using the similar way in the proofs of the theorems.

6 Simulation studies

6.1 The QIF and hybrid GEE-based estimators

In the first simulation, we consider

$$y_{ij} = \beta_1 x_{ij1} + \beta_2 x_{ij2} + \varepsilon_{ij}, \tag{14}$$

where $x_{ij1} \sim N(1,1)$, $x_{ij2} \sim N(0,1)$ and $\mathrm{Cov}(x_{ij1}, x_{ij2}) = \sigma$, the random errors $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3}, \varepsilon_{i4})^T$ are generated from normal distributions $N(0, \Sigma)$. Here, we consider the AR(1) errors with $\Sigma_{jj'} = 4\rho^{|j-j'|}$ and CS errors with $\Sigma_{jj'} = 4\rho$ for $j \ne j$ and $\Sigma_{jj} = 4$ for $j, j' = 1, \ldots, 4$. In addition, two correlation structures are considered: (i) $\boldsymbol{\varepsilon}_i$ are strongly correlated, i.e., $\rho = 0.7$; (ii) $\boldsymbol{\varepsilon}_i$ are moderately correlated, i.e., $\rho = 0.4$. Set the true value $(\beta_1, \beta_2) = (1, 2)$. The missing indicators $\boldsymbol{r}_i = (r_{i1}, r_{i2}, r_{i3}, r_{i4})^T$ are generated from the following nonignorable dropout choice:

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

with $\alpha_j = -1.2$, $\gamma_{j1} = 0.2j$ and $\gamma_{j2} = -0.4 + 0.1(j-1)$. For $j = 1, \ldots, 4$, the coefficients were chosen so that the unconditional dropout percentages for four time points under different scenarios are about 25%, 44%, 61% and 75%. In addition, we consider two skewed distributions for the errors $\boldsymbol{\varepsilon}_i$, i.e., $\varepsilon_{ij} = \text{Exp}(1) - 1$ and $\boldsymbol{\Gamma}$ (1,1)-1 with AR(1) covariance matrix $\Sigma_{jj'} = 4\rho^{|j-j'|}$, by using R packages *simstudy* and *copula*.

To evaluate the estimation efficiency of the proposed approach, we compute the simulated relative bias and variance of the estimators based on the following six GLM estimators of $\boldsymbol{\beta}$.

(a) the proposed QIF estimator based on $\hat{\boldsymbol{g}}_i(\boldsymbol{\beta})$ in (11) and the hybrid GEE estimator based on $\hat{\boldsymbol{h}}_i(\boldsymbol{\beta})$ in (13) with nonignorable dropout propensity $\pi_{ij}(\hat{\boldsymbol{\Theta}}_j)$ in $\hat{\boldsymbol{W}}_i$ and the GMM estimator $\hat{\boldsymbol{\Theta}}_j$ obtained by (7). Here, $\boldsymbol{R}_i^{-1}$ in the QIF estimator are based on two common working correlation choices: AR(1) and CS, which are denoted as $\text{QIF}_{\text{AR(1)}}$ and QIF_{CS}; two different choices of $\{\boldsymbol{R}^1(\alpha), \boldsymbol{R}^2(\alpha)\}$ =(AR(1), CS) with $\alpha = 0.4$ and 0.7 are used in the hybrid GEE method, which are denoted as $\text{Hybrid}_{0.4}$ and $\text{Hybrid}_{0.7}$.
(b) the naive MNAR estimator based on (8) with an independent working correlation structure, i.e., $\boldsymbol{V}_i = \boldsymbol{I}_m$, which is denoted as MNAR_{IND}.
(c) the MAR estimator based on (8) with ignorable dropout $\pi_{ij}(\hat{\Upsilon}_j) = \pi_{ij}(\vec{\boldsymbol{x}}_{ij}, \hat{\Upsilon}_j)$ in $\hat{\boldsymbol{W}}_i$. Here, the ignorable dropout propensity $\Pr(r_{ij} = 1 | r_{i(j-1)} = 1, \boldsymbol{x}_{ij})$ is imposed by a parametric linear logistic regression and the GMM estimator $\hat{\Upsilon}_j$ is obtained similarly by (7);
(d) the complete case (CC) estimator based on (8) with $\hat{\boldsymbol{W}}_i = diag\{r_{i1}, \ldots, r_{im}\}$;
(e) the full sample (FULL) estimator based on (8) with $\hat{\boldsymbol{W}}_i = \boldsymbol{I}_m$ when there is no missing data, which is used as a gold standard.

In the estimators (c–e), the true values of $\boldsymbol{V}_i$ are used to obtain the best results. To apply the propose method, we use the working propensity model (5) and $\boldsymbol{\Psi}(\cdot) = [1 + \exp(\cdot)]^{-1}$. It can be seen that $u_{ij} = x_{ij1}$ and the instrument variable $z_{ij} = x_{ij2}$. We further examine the confidence regions of two dimensional $\boldsymbol{\beta}$ in terms of the coverage probability (CP). In particular, the EL confidence regions based on the proposed methods are obtained by $\text{CI}_2^Q(\alpha)$ and $\text{CI}_2^H(\alpha)$ in Section 3, the EL confidence region based on the estimator (c) is obtained similarly by $\text{CI}_2^Q(\alpha)$ with the ignorable dropout propensity, and the EL confidence regions based on the estimators (d–e) are obtained by $\text{CI}(\alpha) = \{\boldsymbol{\beta} : \hat{R}(\boldsymbol{\beta}) < \chi^2_{1-\alpha}(p)\}$ with $\hat{\boldsymbol{g}}_i(\boldsymbol{\beta})$ in (11) replaced by the corresponding estimating equations under the CC and FULL methods, respectively. According to Qin and Lawless (1994), only the full sample method can produce correct EL confidence regions. Simulation results are presented in Tables 1, 2, 3 and 4, and a few conclusions can be drawn from the simulation results.

(1) The naive MNAR estimator, the proposed estimators based on QIF and hybrid GEE methods are unbiased. On the other hand, the CC estimators are biased

Table 1 Relative biases, standard deviations (in parentheses) and coverage probabilities in the first simulation under normal errors with AR(1) structure

(σ, ρ)	Methods	$n = 200$			$n = 500$		
		β_1	β_2	CP	β_1	β_2	CP
(0.9, 0.4)	CC	0.257(0.103)	−0.155(0.136)	0.416	0.255(0.070)	−0.153(0.089)	0.048
	MAR	0.268(0.128)	−0.151(0.155)	0.520	0.269(0.086)	−0.150(0.101)	0.120
	FULL	0.001(0.075)	−0.002(0.093)	0.955	0.001(0.051)	0.001(0.059)	0.948
	MNAR_{IND}	0.053(0.228)	−0.047(0.270)	0.935	0.023(0.151)	−0.008(0.173)	0.947
	$\text{QIF}_{\text{AR(1)}}$	0.057(0.167)	−0.027(0.188)	0.931	0.025(0.117)	−0.010(0.131)	0.942
	QIF_{CS}	0.042(0.181)	−0.020(0.202)	0.948	0.016(0.123)	−0.005(0.135)	0.950
	$\text{Hybrid}_{0.4}$	0.037(0.172)	−0.015(0.200)	0.951	0.016(0.118)	−0.005(0.131)	0.957
	$\text{Hybrid}_{0.7}$	0.021(0.202)	−0.008(0.226)	0.952	0.008(0.133)	0.001(0.141)	0.958
(0.9, 0.7)	CC	0.185(0.121)	−0.111(0.134)	0.663	0.184(0.071)	−0.111(0.085)	0.326
	MAR	0.226(0.144)	−0.120(0.151)	0.682	0.231(0.088)	−0.121(0.097)	0.312
	FULL	0.001(0.077)	−0.002(0.083)	0.953	0.001(0.048)	0.001(0.053)	0.954
	MNAR_{IND}	0.047(0.246)	−0.012(0.259)	0.942	−0.003(0.140)	0.006(0.167)	0.951
	$\text{QIF}_{\text{AR(1)}}$	0.054(0.164)	−0.021(0.183)	0.945	0.014(0.105)	−0.005(0.117)	0.960
	QIF_{CS}	0.040(0.183)	−0.017(0.203)	0.948	0.001(0.122)	0.002(0.135)	0.952
	$\text{Hybrid}_{0.4}$	0.043(0.180)	−0.016(0.196)	0.949	0.006(0.112)	0.001(0.123)	0.958
	$\text{Hybrid}_{0.7}$	0.028(0.175)	−0.007(0.188)	0.947	0.003(0.115)	0.002(0.126)	0.952
(0.6, 0.4)	CC	0.144(0.088)	−0.078(0.115)	0.636	0.151(0.055)	−0.083(0.072)	0.252
	MAR	0.163(0.111)	−0.078(0.130)	0.699	0.170(0.072)	−0.085(0.081)	0.340
	FULL	0.001(0.059)	0.001(0.069)	0.957	0.001(0.038)	0.001(0.045)	0.960
	MNAR_{IND}	0.012(0.142)	0.002(0.174)	0.934	0.010(0.094)	−0.011(0.108)	0.951
	$\text{QIF}_{\text{AR(1)}}$	0.026(0.124)	−0.006(0.145)	0.920	0.016(0.079)	−0.009(0.090)	0.948
	QIF_{CS}	0.017(0.129)	−0.001(0.156)	0.947	0.010(0.084)	−0.006(0.096)	0.930
	$\text{Hybrid}_{0.4}$	0.013(0.123)	0.001(0.145)	0.947	0.010(0.081)	−0.006(0.094)	0.946
	$\text{Hybrid}_{0.7}$	0.005(0.144)	0.004(0.160)	0.940	0.010(0.094)	−0.006(0.100)	0.958
(0.6, 0.7)	CC	0.076(0.083)	−0.052(0.096)	0.855	0.077(0.052)	−0.048(0.059)	0.702
	MAR	0.113(0.107)	−0.060(0.113)	0.836	0.117(0.064)	−0.057(0.070)	0.592
	FULL	0.001(0.049)	−0.002(0.054)	0.950	0.001(0.032)	0.001(0.034)	0.961
	MNAR_{IND}	0.018(0.138)	0.013(0.159)	0.937	0.009(0.091)	−0.009(0.106)	0.951
	$\text{QIF}_{\text{AR(1)}}$	0.030(0.118)	−0.013(0.130)	0.913	0.015(0.076)	−0.005(0.090)	0.944
	QIF_{CS}	0.016(0.122)	−0.006(0.143)	0.925	0.013(0.080)	−0.003(0.091)	0.944
	$\text{Hybrid}_{0.4}$	0.021(0.117)	−0.009(0.131)	0.922	0.012(0.078)	−0.003(0.086)	0.946
	$\text{Hybrid}_{0.7}$	0.014(0.122)	−0.006(0.131)	0.939	0.009(0.077)	−0.002(0.080)	0.944

due to the fact that missing is not completely at random; the estimators based on ignorable dropout also have large biases. When the covariates are strongly correlated ($\sigma = 0.9$), the biases of the CC and MAR estimates become larger. Moreover, it also shows robustness of the proposed estimators which are less sensitivity to the error distributions $\boldsymbol{\varepsilon}_i$ and correlation structures σ.

Table 2 Relative biases, standard deviations (in parentheses) and coverage probabilities in the first simulation under normal errors with CS structure

(σ, ρ)	Methods	$n = 200$			$n = 500$		
		β_1	β_2	CP	β_1	β_2	CP
(0.9, 0.4)	CC	0.266(0.115)	−0.107(0.149)	0.453	0.257(0.073)	−0.155(0.093)	0.056
	MAR	0.300(0.139)	−0.162(0.167)	0.489	0.294(0.086)	−0.157(0.101)	0.118
	FULL	0.005(0.082)	−0.003(0.098)	0.953	0.001(0.052)	0.001(0.059)	0.960
	MNAR_{IND}	0.038(0.219)	−0.017(0.260)	0.933	−0.013(0.156)	0.013(0.179)	0.947
	$\text{QIF}_{\text{AR(1)}}$	0.058(0.169)	−0.028(0.201)	0.939	0.025(0.113)	−0.011(0.134)	0.952
	QIF_{CS}	0.044(0.175)	−0.019(0.204)	0.944	0.015(0.117)	−0.005(0.134)	0.952
	$\text{Hybrid}_{0.4}$	0.034(0.181)	−0.012(0.210)	0.950	0.015(0.112)	−0.003(0.132)	0.958
	$\text{Hybrid}_{0.7}$	0.019(0.202)	−0.005(0.229)	0.958	0.007(0.129)	0.003(0.148)	0.957
(0.9, 0.7)	CC	0.188(0.134)	−0.123(0.149)	0.686	0.185(0.080)	−0.119(0.090)	0.368
	MAR	0.246(0.158)	−0.127(0.165)	0.695	0.248(0.095)	−0.124(0.102)	0.382
	FULL	0.001(0.081)	0.001(0.087)	0.946	0.001(0.050)	0.001(0.053)	0.956
	MNAR_{IND}	0.058(0.280)	−0.032(0.321)	0.943	0.019(0.157)	−0.017(0.193)	0.946
	$\text{QIF}_{\text{AR(1)}}$	0.065(0.166)	−0.031(0.189)	0.941	0.033(0.106)	−0.010(0.118)	0.955
	QIF_{CS}	0.041(0.175)	−0.018(0.199)	0.937	0.021(0.113)	−0.005(0.129)	0.950
	$\text{Hybrid}_{0.4}$	0.047(0.171)	−0.021(0.198)	0.945	0.025(0.113)	−0.009(0.124)	0.949
	$\text{Hybrid}_{0.7}$	0.033(0.184)	−0.012(0.207)	0.951	0.017(0.113)	−0.004(0.125)	0.953
(0.6, 0.4)	CC	0.132(0.091)	−0.075(0.115)	0.710	0.128(0.057)	−0.075(0.070)	0.404
	MAR	0.166(0.112)	−0.079(0.132)	0.684	0.163(0.070)	−0.078(0.082)	0.376
	FULL	−0.001(0.061)	0.001(0.071)	0.945	−0.005(0.038)	0.001(0.044)	0.960
	MNAR_{IND}	0.013(0.155)	0.006(0.177)	0.924	0.006(0.094)	−0.013(0.113)	0.945
	$\text{QIF}_{\text{AR(1)}}$	0.025(0.128)	−0.008(0.146)	0.913	0.011(0.077)	−0.005(0.090)	0.946
	QIF_{CS}	0.013(0.128)	−0.002(0.145)	0.935	0.001(0.082)	0.001(0.094)	0.942
	$\text{Hybrid}_{0.4}$	0.011(0.127)	−0.001(0.146)	0.944	0.001(0.082)	0.001(0.095)	0.946
	$\text{Hybrid}_{0.7}$	0.002(0.146)	0.004(0.154)	0.949	−0.001(0.088)	0.001(0.098)	0.952
(0.6, 0.7)	CC	0.062(0.078)	−0.048(0.089)	0.842	0.065(0.058)	−0.052(0.064)	0.760
	MAR	0.110(0.099)	−0.050(0.107)	0.839	0.117(0.069)	−0.054(0.074)	0.734
	FULL	−0.002(0.043)	0.001(0.049)	0.951	0.003(0.033)	−0.001(0.035)	0.950
	MNAR_{IND}	0.007(0.178)	−0.002(0.188)	0.913	0.015(0.102)	−0.008(0.119)	0.928
	$\text{QIF}_{\text{AR(1)}}$	0.033(0.102)	−0.010(0.116)	0.898	0.025(0.076)	−0.010(0.083)	0.932
	QIF_{CS}	0.016(0.111)	−0.002(0.125)	0.917	0.016(0.082)	−0.004(0.091)	0.914
	$\text{Hybrid}_{0.4}$	0.024(0.105)	−0.006(0.120)	0.936	0.019(0.082)	−0.006(0.087)	0.918
	$\text{Hybrid}_{0.7}$	0.020(0.109)	−0.002(0.119)	0.950	0.014(0.081)	−0.004(0.086)	0.928

(2) Compared with the naive MNAR estimator, the proposed four estimators have smaller variances. Among the two QIF estimators, it can be seen that the estimates $\text{QIF}_{\text{AR(1)}}$ have smaller or comparable variances than these based on QIF_{CS}; Among the two hybrid GEE estimators, the estimates based on $\text{Hybrid}_{0.4}$ have smaller variances when ρ is small and $\text{Hybrid}_{0.7}$ performs better when ρ is large. The within-subject correlations involved with the quantile regression are sign

Table 3 Relative biases, standard deviations (in parentheses) and coverage probabilities in the first simulation under errors ε_{ij} = Exp(1) − 1 with AR(1) structure

(σ, ρ)	Methods	$n = 200$			$n = 500$		
		β_1	β_2	CP	β_1	β_2	CP
(0.9, 0.4)	CC	0.211(0.123)	−0.130(0.158)	0.602	0.207(0.079)	−0.129(0.107)	0.258
	MAR	0.235(0.158)	−0.130(0.183)	0.676	0.232(0.094)	−0.130(0.116)	0.296
	FULL	0.000(0.080)	0.001(0.095)	0.950	0.000(0.048)	0.000(0.060)	0.952
	MNAR_{IND}	0.081(0.217)	−0.035(0.247)	0.950	0.027(0.145)	−0.010(0.146)	0.952
	$\text{QIF}_{\text{AR(1)}}$	0.039(0.157)	−0.015(0.182)	0.960	0.011(0.101)	−0.004(0.113)	0.958
	QIF_{CS}	0.066(0.183)	−0.029(0.216)	0.948	0.026(0.131)	−0.010(0.137)	0.932
	$\text{Hybrid}_{0.4}$	0.052(0.178)	−0.021(0.211)	0.962	0.024(0.133)	−0.007(0.123)	0.948
	$\text{Hybrid}_{0.7}$	0.022(0.192)	−0.008(0.210)	0.956	0.014(0.127)	−0.005(0.134)	0.960
(0.9, 0.7)	CC	0.159(0.137)	−0.100(0.158)	0.776	0.158(0.083)	−0.102(0.102)	0.532
	MAR	0.215(0.179)	−0.113(0.196)	0.758	0.213(0.103)	−0.114(0.116)	0.518
	FULL	0.002(0.079)	−0.001(0.083)	0.948	−0.004(0.048)	0.001(0.053)	0.944
	MNAR_{IND}	0.039(0.194)	−0.017(0.248)	0.952	0.008(0.113)	−0.003(0.124)	0.951
	$\text{QIF}_{\text{AR(1)}}$	0.018(0.151)	−0.006(0.163)	0.962	0.000(0.086)	−0.002(0.097)	0.958
	QIF_{CS}	0.034(0.170)	−0.011(0.183)	0.954	0.004(0.093)	−0.003(0.105)	0.958
	$\text{Hybrid}_{0.4}$	0.029(0.165)	−0.011(0.181)	0.957	0.005(0.094)	−0.003(0.104)	0.952
	$\text{Hybrid}_{0.7}$	0.012(0.163)	−0.004(0.179)	0.958	0.000(0.094)	−0.001(0.100)	0.954
(0.6, 0.4)	CC	0.131(0.101)	−0.076(0.132)	0.742	0.123(0.064)	−0.070(0.082)	0.497
	MAR	0.156(0.131)	−0.077(0.159)	0.774	0.149(0.084)	−0.073(0.099)	0.526
	FULL	0.003(0.058)	0.000(0.072)	0.960	−0.001(0.038)	0.000(0.047)	0.950
	MNAR_{IND}	0.024(0.136)	0.004(0.143)	0.954	0.007(0.093)	0.001(0.106)	0.954
	$\text{QIF}_{\text{AR(1)}}$	0.014(0.108)	−0.004(0.128)	0.936	0.004(0.069)	0.003(0.080)	0.950
	QIF_{CS}	0.025(0.115)	−0.004(0.133)	0.908	0.009(0.078)	0.001(0.088)	0.952
	$\text{Hybrid}_{0.4}$	0.021(0.118)	−0.001(0.134)	0.952	0.006(0.075)	0.002(0.084)	0.931
	$\text{Hybrid}_{0.7}$	0.015(0.127)	0.000(0.142)	0.940	0.003(0.081)	0.003(0.085)	0.929
(0.6, 0.7)	CC	0.064(0.093)	−0.048(0.116)	0.854	0.068(0.062)	−0.047(0.073)	0.778
	MAR	0.106(0.126)	−0.057(0.145)	0.838	0.111(0.081)	−0.054(0.090)	0.712
	FULL	−0.001(0.051)	0.000(0.057)	0.954	0.001(0.034)	−0.001(0.035)	0.938
	MNAR_{IND}	0.007(0.124)	−0.001(0.132)	0.952	0.002(0.074)	0.002(0.084)	0.958
	$\text{QIF}_{\text{AR(1)}}$	0.003(0.090)	0.001(0.105)	0.958	0.005(0.062)	−0.001(0.067)	0.958
	QIF_{CS}	0.007(0.101)	−0.003(0.120)	0.936	0.005(0.068)	0.001(0.073)	0.962
	$\text{Hybrid}_{0.4}$	0.007(0.098)	−0.002(0.111)	0.948	0.006(0.066)	0.001(0.070)	0.946
	$\text{Hybrid}_{0.7}$	0.002(0.096)	0.001(0.111)	0.944	0.004(0.065)	0.001(0.068)	0.946

correlations such that the true correlation structure is a toeplitz with $(m - 1)$ number of parameters. Among the two common working correlation choices, the AR(1) structure best approximates the true correlation structure. These findings are consistent with our theoretical result that the choice of correlation matrix does not affect the consistence, but will affect the efficiency. Moreover, the vari-

Table 4 Relative biases, standard deviations (in parentheses) and coverage probabilities in the first simulation under errors $\varepsilon_{ij} = \Gamma(1,1) - 1$ with AR(1) structure

(σ, ρ)	Methods	$n = 200$			$n = 500$		
		β_1	β_2	CP	β_1	β_2	CP
(0.9, 0.4)	CC	0.226(0.130)	−0.139(0.173)	0.598	0.235(0.086)	−0.143(0.115)	0.186
	MAR	0.273(0.169)	−0.150(0.200)	0.641	0.280(0.103)	−0.153(0.128)	0.216
	FULL	−0.004(0.084)	0.002(0.099)	0.938	0.001(0.055)	0.001(0.069)	0.927
	MNAR_{IND}	0.065(0.190)	−0.025(0.224)	0.948	0.021(0.120)	−0.007(0.133)	0.957
	$\text{QIF}_{\text{AR(1)}}$	0.038(0.171)	−0.017(0.202)	0.928	0.020(0.127)	−0.009(0.148)	0.962
	QIF_{CS}	0.051(0.190)	−0.023(0.213)	0.950	0.022(0.120)	−0.008(0.139)	0.967
	$\text{Hybrid}_{0.4}$	0.044(0.184)	−0.019(0.213)	0.935	0.022(0.117)	−0.008(0.132)	0.962
	$\text{Hybrid}_{0.7}$	0.027(0.205)	−0.011(0.239)	0.946	0.016(0.131)	−0.005(0.151)	0.950
(0.9, 0.7)	CC	0.176(0.145)	−0.107(0.178)	0.755	0.172(0.097)	−0.107(0.112)	0.511
	MAR	0.254(0.193)	−0.127(0.216)	0.745	0.253(0.125)	−0.127(0.136)	0.457
	FULL	0.001(0.088)	0.001(0.097)	0.946	0.001(0.058)	−0.001(0.061)	0.932
	MNAR_{IND}	0.035(0.201)	−0.015(0.246)	0.947	0.006(0.112)	−0.002(0.129)	0.952
	$\text{QIF}_{\text{AR(1)}}$	0.016(0.156)	−0.005(0.176)	0.936	0.007(0.094)	−0.002(0.105)	0.948
	QIF_{CS}	0.024(0.165)	−0.007(0.181)	0.946	0.009(0.098)	−0.001(0.105)	0.953
	$\text{Hybrid}_{0.4}$	0.029(0.163)	−0.009(0.179)	0.950	0.008(0.098)	−0.002(0.106)	0.950
	$\text{Hybrid}_{0.7}$	0.011(0.153)	−0.001(0.172)	0.946	0.001(0.094)	0.002(0.102)	0.948
(0.6, 0.4)	CC	0.147(0.105)	−0.081(0.129)	0.701	0.136(0.063)	−0.077(0.085)	0.471
	MAR	0.187(0.144)	−0.089(0.158)	0.701	0.180(0.080)	−0.084(0.098)	0.416
	FULL	0.005(0.064)	−0.002(0.071)	0.946	0.001(0.040)	0.000(0.050)	0.938
	MNAR_{IND}	0.023(0.133)	−0.003(0.147)	0.947	0.003(0.071)	−0.001(0.081)	0.953
	$\text{QIF}_{\text{AR(1)}}$	0.017(0.112)	−0.003(0.123)	0.921	0.006(0.063)	−0.001(0.077)	0.949
	QIF_{CS}	0.022(0.117)	−0.003(0.121)	0.939	0.007(0.066)	−0.001(0.078)	0.951
	$\text{Hybrid}_{0.4}$	0.021(0.127)	−0.003(0.125)	0.945	0.010(0.082)	−0.001(0.082)	0.951
	$\text{Hybrid}_{0.7}$	0.012(0.137)	−0.002(0.134)	0.941	0.007(0.071)	−0.001(0.082)	0.964
(0.6, 0.7)	CC	0.074(0.111)	−0.048(0.128)	0.842	0.074(0.065)	−0.050(0.072)	0.758
	MAR	0.126(0.145)	−0.056(0.158)	0.843	0.130(0.087)	−0.059(0.092)	0.678
	FULL	0.001(0.062)	0.000(0.067)	0.933	0.000(0.036)	−0.001(0.040)	0.944
	MNAR_{IND}	0.011(0.119)	−0.002(0.137)	0.942	−0.002(0.076)	0.000(0.084)	0.951
	$\text{QIF}_{\text{AR(1)}}$	0.006(0.099)	0.001(0.110)	0.933	0.001(0.061)	0.001(0.067)	0.940
	QIF_{CS}	0.006(0.101)	0.001(0.114)	0.935	0.001(0.063)	0.002(0.068)	0.940
	$\text{Hybrid}_{0.4}$	0.009(0.101)	0.001(0.111)	0.947	0.001(0.063)	0.002(0.067)	0.957
	$\text{Hybrid}_{0.7}$	0.004(0.104)	0.003(0.111)	0.928	0.001(0.061)	0.002(0.066)	0.959

ances become smaller when the mean response rate or the sample size is larger, and become larger when σ increases.

(3) The coverage probabilities based on the proposed estimators are close to the nominal level, and are quite comparable to the FULL method assuming no missing data. It can be seen that the coverage probabilities based on the MAR and

CC methods have undercoverage. The poor performance is due to the large bias and the fact that the corresponding asymptotic distribution of $\hat{R}(\boldsymbol{\beta}_0)$ is not χ^2_2.
(4) When $n = 500$, the proposed four estimators have similar performance.

In the second simulation, we consider the similar settings as in the first simulation, but investigate the performance of the proposed estimators when the propensity is misspecified. In specific, we consider

$$\Pr(r_{ij} = 1 | r_{i(j-1)} = 1, \vec{\boldsymbol{u}}_{ij}, \vec{\boldsymbol{y}}_{ij}) = 1/\{1 + \exp(\alpha_j + \gamma_{j1}\sin(x_{ij1}) + \gamma_{j2}y_{ij})\},$$

with the same α_j, γ_{j1} and γ_{j2} as in the first simulation. In this case, however, the working model was misspecified so that we can see the robustness of the proposed estimators. For $j = 1, ..., 4$, the coefficients were chosen so that the unconditional dropout percentages for four time points are about 32%, 54%, 69% and 80%.

Tables 5-6 report the simulation results, and we have the similar results as in the first simulation. The proposed estimators have negligible biases, even the working dropout propensity model is wrong. In conclusion, the above two simulations suggest that the proposed estimators not only have good point estimates, but also are robust against propensity model specifications and common error distributions.

6.2 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), model error and model selection accuracy. We consider

$$\boldsymbol{y}_i = \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i, \quad i = 1, 2, \ldots, n,$$

where $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{i4})^T$ is a $(4 \times p)$-dimensional matrix of covariates, $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{i4}$ are from a p-dimensional normal distribution having mean $(1, \ldots, 1)^T$ and covariance matrix $\boldsymbol{\Gamma}$ with $\boldsymbol{\Gamma}_{jj} = 1$ and $\boldsymbol{\Gamma}_{jj'} = 0.6$ for $1 \le j < j' \le p$. The true value of $\boldsymbol{\beta} = (3, 1.5, 2, 0, \ldots, 0)$ and $\boldsymbol{\epsilon}_i$ are from the same normal errors as in Section 5.1 with $\rho = 0.7$. The missing indicators $\boldsymbol{r}_i = (r_{i1}, r_{i2}, r_{i3}, r_{i4})^T$ are generated from

$$\Pr(r_{ij} = 1 | r_{i(j-1)} = 1, \vec{\boldsymbol{u}}_{ij}, \vec{\boldsymbol{y}}_{ij}) = 1/\{1 + \exp(\alpha_j + \boldsymbol{\gamma}_{j1}^T\boldsymbol{u}_{ij} + \gamma_{j2}y_{ij})\},$$

where $\boldsymbol{u}_{ij} = (x_{ij3}, \ldots, x_{ijp})^T$, $\alpha_j = -0.8 + 0.2(j-1)$, $\boldsymbol{\gamma}_{j1}^T = (-0.1, 0.1, \ldots, -0.1, 0.1, \ldots)$ and $\gamma_{j2} = -0.4 + 0.1(j-1)$ for $j = 1, 2, 3, 4$.

Our penalized EL method is combined with three information criteria: BIC, BICC and EBIC, for selecting the tuning parameter ν. Table 7 reports the results for $n = 200$, 500 and $p = 10$, 20, 50. We obtain the mean square errors (MSE) defined by $\mathrm{MSE}(\hat{\boldsymbol{\beta}}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$. 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

Table 5 Relative biases, standard deviations (in parentheses) and coverage probabilities in the second simulation under normal errors with AR(1) structure

(σ, ρ)	Methods	$n = 200$			$n = 500$		
		β_1	β_2	CP	β_1	β_2	CP
(0.9, 0.4)	CC	0.225(0.105)	−0.147(0.131)	0.477	0.223(0.067)	−0.144(0.091)	0.078
	MAR	0.234(0.121)	−0.147(0.150)	0.546	0.230(0.078)	−0.144(0.102)	0.150
	FULL	0.001(0.079)	0.001(0.093)	0.941	−0.002(0.051)	0.001(0.062)	0.954
	MNAR_{IND}	0.019(0.183)	0.021(0.230)	0.943	0.009(0.121)	−0.002(0.150)	0.951
	$\text{QIF}_{\text{AR(1)}}$	0.035(0.147)	−0.022(0.183)	0.956	0.009(0.101)	−0.005(0.125)	0.948
	QIF_{CS}	0.022(0.159)	−0.014(0.194)	0.942	0.001(0.107)	0.001(0.130)	0.954
	$\text{Hybrid}_{0.4}$	0.019(0.151)	−0.012(0.189)	0.953	0.002(0.102)	0.001(0.125)	0.956
	$\text{Hybrid}_{0.7}$	0.001(0.176)	−0.003(0.202)	0.957	−0.008(0.116)	0.005(0.135)	0.958
(0.9, 0.7)	CC	0.170(0.108)	−0.109(0.128)	0.699	0.166(0.073)	−0.105(0.087)	0.342
	MAR	0.199(0.125)	−0.119(0.141)	0.684	0.193(0.084)	−0.115(0.095)	0.318
	FULL	0.002(0.076)	−0.002(0.083)	0.950	−0.002(0.051)	0.001(0.053)	0.948
	MNAR_{IND}	0.043(0.172)	−0.037(0.210)	0.945	−0.005(0.125)	0.004(0.145)	0.953
	$\text{QIF}_{\text{AR(1)}}$	0.036(0.144)	−0.021(0.168)	0.944	0.004(0.102)	−0.002(0.119)	0.956
	QIF_{CS}	0.019(0.157)	−0.011(0.184)	0.947	−0.002(0.105)	0.002(0.123)	0.954
	$\text{Hybrid}_{0.4}$	0.022(0.155)	−0.015(0.181)	0.941	−0.001(0.105)	0.001(0.121)	0.952
	$\text{Hybrid}_{0.7}$	0.011(0.154)	−0.008(0.178)	0.955	−0.006(0.103)	0.003(0.120)	0.950
(0.6, 0.4)	CC	0.119(0.082)	−0.072(0.109)	0.711	0.120(0.050)	−0.073(0.068)	0.372
	MAR	0.126(0.095)	−0.079(0.123)	0.735	0.126(0.058)	−0.079(0.075)	0.422
	FULL	−0.003(0.059)	0.002(0.073)	0.958	0.001(0.039)	0.001(0.047)	0.952
	MNAR_{IND}	−0.005(0.144)	0.008(0.154)	0.932	−0.007(0.080)	−0.001(0.097)	0.941
	$\text{QIF}_{\text{AR(1)}}$	0.011(0.106)	−0.007(0.136)	0.927	0.002(0.069)	−0.002(0.085)	0.938
	QIF_{CS}	0.001(0.110)	0.001(0.144)	0.931	−0.002(0.071)	−0.001(0.087)	0.936
	$\text{Hybrid}_{0.4}$	−0.003(0.108)	0.001(0.138)	0.943	−0.004(0.069)	0.002(0.086)	0.944
	$\text{Hybrid}_{0.7}$	−0.009(0.122)	0.003(0.147)	0.940	−0.008(0.079)	0.003(0.092)	0.942
(0.6, 0.7)	CC	0.063(0.076)	−0.043(0.089)	0.843	0.063(0.048)	−0.046(0.055)	0.696
	MAR	0.078(0.087)	−0.050(0.098)	0.830	0.077(0.053)	−0.054(0.061)	0.668
	FULL	0.001(0.050)	0.001(0.054)	0.943	0.001(0.031)	−0.002(0.034)	0.932
	MNAR_{IND}	0.007(0.123)	−0.002(0.151)	0.934	0.009(0.094)	−0.004(0.104)	0.945
	$\text{QIF}_{\text{AR(1)}}$	0.016(0.099)	−0.007(0.117)	0.908	−0.001(0.064)	−0.003(0.072)	0.940
	QIF_{CS}	0.005(0.114)	−0.002(0.131)	0.927	−0.009(0.076)	0.001(0.087)	0.954
	$\text{Hybrid}_{0.4}$	0.010(0.108)	−0.004(0.122)	0.934	−0.005(0.070)	−0.003(0.076)	0.942
	$\text{Hybrid}_{0.7}$	0.001(0.106)	0.001(0.121)	0.945	– 0.007(0.067)	– 0.001(0.074)	0.948

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 Table 7, 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)

Table 6 Relative biases, standard deviations (in parentheses) and coverage probabilities in the second simulation under normal errors with CS structure

(σ, ρ)	Methods	$n = 200$			$n = 500$		
		β_1	β_2	CP	β_1	β_2	CP
(0.9, 0.4)	CC	0.227(0.107)	− 0.144(0.140)	0.529	0.225(0.068)	− 0.144(0.088)	0.110
	MAR	0.231(0.125)	− 0.143(0.157)	0.556	0.233(0.077)	− 0.144(0.097)	0.124
	FULL	− 0.001(0.081)	0.002(0.094)	0.948	0.001(0.051)	0.001(0.062)	0.944
	$\mathrm{MNAR}_{\mathrm{IND}}$	0.004(0.210)	− 0.008(0.271)	0.943	0.010(0.144)	− 0.002(0.160)	0.955
	$\mathrm{QIF}_{\mathrm{AR(1)}}$	0.023(0.155)	− 0.014(0.188)	0.939	0.009(0.098)	− 0.005(0.122)	0.952
	$\mathrm{QIF}_{\mathrm{CS}}$	0.012(0.163)	− 0.006(0.201)	0.945	0.001(0.104)	0.001(0.127)	0.950
	$\mathrm{Hybrid}_{0.4}$	0.007(0.160)	− 0.003(0.195)	0.957	0.002(0.099)	0.001(0.120)	0.958
	$\mathrm{Hybrid}_{0.7}$	− 0.013(0.185)	0.006(0.210)	0.957	− 0.006(0.113)	0.004(0.128)	0.960
(0.9, 0.7)	CC	0.176(0.106)	− 0.109(0.131)	0.717	0.171(0.069)	− 0.107(0.080)	0.332
	MAR	0.206(0.124)	− 0.120(0.148)	0.696	0.198(0.079)	− 0.117(0.089)	0.356
	FULL	0.003(0.072)	0.001(0.080)	0.957	0.003(0.045)	− 0.001(0.051)	0.958
	$\mathrm{MNAR}_{\mathrm{IND}}$	0.018(0.248)	− 0.015(0.276)	0.953	0.019(0.186)	− 0.010(0.201)	0.955
	$\mathrm{QIF}_{\mathrm{AR(1)}}$	0.037(0.139)	− 0.021(0.168)	0.937	0.013(0.098)	− 0.009(0.114)	0.958
	$\mathrm{QIF}_{\mathrm{CS}}$	0.021(0.158)	− 0.012(0.188)	0.949	0.009(0.100)	− 0.007(0.118)	0.952
	$\mathrm{Hybrid}_{0.4}$	0.026(0.155)	− 0.014(0.188)	0.937	0.012(0.094)	− 0.008(0.113)	0.954
	$\mathrm{Hybrid}_{0.7}$	0.021(0.157)	− 0.011(0.185)	0.949	0.003(0.096)	− 0.002(0.111)	0.956
(0.6, 0.4)	CC	0.112(0.083)	− 0.073(0.114)	0.733	0.114(0.052)	− 0.070(0.072)	0.474
	MAR	0.131(0.096)	− 0.080(0.125)	0.719	0.132(0.059)	− 0.077(0.078)	0.454
	FULL	0.001(0.058)	− 0.002(0.073)	0.952	0.002(0.037)	0.001(0.045)	0.944
	$\mathrm{MNAR}_{\mathrm{IND}}$	0.008(0.134)	− 0.010(0.148)	0.939	0.005(0.093)	− 0.004(0.115)	0.951
	$\mathrm{QIF}_{\mathrm{AR(1)}}$	0.014(0.111)	− 0.010(0.138)	0.916	0.014(0.065)	− 0.006(0.085)	0.946
	$\mathrm{QIF}_{\mathrm{CS}}$	0.005(0.109)	− 0.003(0.140)	0.928	0.006(0.068)	− 0.002(0.089)	0.940
	$\mathrm{Hybrid}_{0.4}$	0.006(0.108)	− 0.003(0.140)	0.944	0.003(0.071)	0.001(0.090)	0.954
	$\mathrm{Hybrid}_{0.7}$	0.001(0.126)	0.001(0.146)	0.951	0.001(0.081)	0.001(0.093)	0.946
(0.6, 0.7)	CC	0.057(0.081)	− 0.045(0.093)	0.864	0.059(0.049)	− 0.047(0.058)	0.720
	MAR	0.079(0.095)	− 0.049(0.104)	0.869	0.082(0.058)	− 0.052(0.063)	0.688
	FULL	− 0.002(0.050)	0.001(0.054)	0.939	− 0.002(0.034)	0.001(0.035)	0.950
	$\mathrm{MNAR}_{\mathrm{IND}}$	0.010(0.142)	0.001(0.162)	0.943	− 0.002(0.100)	− 0.008(0.110)	0.954
	$\mathrm{QIF}_{\mathrm{AR(1)}}$	0.020(0.105)	− 0.008(0.125)	0.911	0.004(0.066)	− 0.006(0.074)	0.946
	$\mathrm{QIF}_{\mathrm{CS}}$	0.009(0.112)	0.001(0.130)	0.930	− 0.002(0.069)	− 0.001(0.078)	0.952
	$\mathrm{Hybrid}_{0.4}$	0.013(0.107)	− 0.004(0.125)	0.933	− 0.001(0.069)	− 0.004(0.076)	0.954
	$\mathrm{Hybrid}_{0.7}$	0.006(0.107)	− 0.001(0.121)	0.940	− 0.001(0.071)	− 0.003(0.077)	0.932

The simulated MSEs of the proposed methods based on BIC, BICC and EBIC are close to that of 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

Table 7 Mean square errors (MSE) and variable selection results

p	Methods	Criteria	$n = 200$			$n = 500$		
			MSE	C	IC	MSE	C	IC
10	$\text{QIF}_{\text{AR(1)}}$	BIC	0.086	3	0.229	0.042	3	0.065
		BICC	0.082	3	0.178	0.041	3	0.053
		EBIC	0.072	3	0.072	0.042	3	0.016
	QIF_{CS}	BIC	0.084	3	0.269	0.030	3	0.081
		BICC	0.081	3	0.231	0.027	3	0.059
		EBIC	0.075	3	0.145	0.025	3	0.037
	$\text{Hybrid}_{0.4}$	BIC	0.072	3	0.387	0.021	3	0.052
		BICC	0.067	3	0.306	0.019	3	0.040
		EBIC	0.047	3	0.080	0.018	3	0.021
	$\text{Hybrid}_{0.7}$	BIC	0.070	3	0.380	0.019	3	0.035
		BICC	0.065	3	0.307	0.018	3	0.030
		EBIC	0.046	3	0.086	0.017	3	0.010
20	$\text{QIF}_{\text{AR(1)}}$	BIC	0.145	3	0.447	0.127	3	0.153
		BICC	0.141	3	0.402	0.126	3	0.118
		EBIC	0.131	3	0.259	0.129	3	0.082
	QIF_{CS}	BIC	0.187	3	0.934	0.080	3	0.225
		BICC	0.181	3	0.892	0.078	3	0.210
		EBIC	0.175	3	0.824	0.075	3	0.148
	$\text{Hybrid}_{0.4}$	BIC	0.116	3	0.785	0.031	3	0.128
		BICC	0.109	3	0.700	0.030	3	0.118
		EBIC	0.084	3	0.379	0.027	3	0.051
	$\text{Hybrid}_{0.7}$	BIC	0.120	3	0.862	0.024	3	0.097
		BICC	0.109	3	0.713	0.023	3	0.077
		EBIC	0.082	3	0.308	0.021	3	0.041
50	$\text{QIF}_{\text{AR(1)}}$	BIC	0.385	3	1.213	0.186	3	0.463
		BICC	0.327	3	1.144	0.185	3	0.391
		EBIC	0.313	3	1.002	0.185	3	0.247
	QIF_{CS}	BIC	0.422	3	1.735	0.169	3	0.549
		BICC	0.419	3	1.528	0.167	3	0.445
		EBIC	0.418	3	1.431	0.163	3	0.402
	$\text{Hybrid}_{0.4}$	BIC	0.274	3	1.398	0.077	3	0.369
		BICC	0.275	3	1.315	0.077	3	0.352
		EBIC	0.267	3	1.240	0.078	3	0.327
	$\text{Hybrid}_{0.7}$	BIC	0.347	3	1.234	0.071	3	0.284
		BICC	0.340	3	1.128	0.067	3	0.223
		EBIC	0.323	3	1.066	0.065	3	0.241
	Oracle		0.012	3	0	0.005	3	0

these results, in practice, we recommend to use the information criteria EBIC for selecting ν.

7 Application to HIV-CD4 data

For illustration, we apply the proposed estimators to a longitudinal data from the AIDS Clinical Trial Group 193A, which was a study of HIV-AIDS patients with advanced immune suppression. In this study, the patients were taken the daily regimen containing 600 mg of zidovudine plus 2.25 mg of zalcitabine. The data set can be accessed at http://www.hsph.harvard.edu/fitzmaur/ala/cd4.txt.

For the HIV clinical trial, the CD4 cell count is of prime interest which decreases as HIV progresses. In this study, the CD4 counts were collected from 316 patients before the treatments were applied (baseline measurements), and we use their records in the analysis. After the treatments were applied, the CD4 count was scheduled to be collected from each patient in every 8 weeks. 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 realized follow-up time points might be a little different from the scheduled time points. A few patients had more than one measurement 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}) and the dropout rates are 31.5%, 42.4%, 55.4% and 65.3%, respectively.

Previous experiences from doctors indicate that, at time point j, the HIV infected patients with low CD4 counts nearby time point j are more likely to drop out. That is, dropout at time point j is related with y_{ij} and and it can be nonignorable. Thus, we use the working propensity model (5) and $\Psi(\cdot) = [1 + \exp(\cdot)]^{-1}$. Also, the follow-up times are treated as covariates u_{ij} may affect the dropout. The ages are always observed and thus can be used as instruments z_{ij}. The purpose of this study is to examine whether the CD4 counts of young patients are more likely to decrease.

The estimates are reported in Table 8. It can be seen that: (1) All estimates of β_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. (2) The estimates of β_2

Table 8 Estimates for the HIV-CD4 data based on QIF and hybrid GEE methods

R_i^{-1}	MNAR			CC		
	β_0	β_1	β_2	β_0	β_1	β_2
$\text{QIF}_{\text{AR(1)}}$	0.310	− 0.280	0.912	0.498	− 0.222	0.838
QIF_{CS}	0.209	− 0.269	0.934	0.442	− 0.220	0.852
$\text{Hybrid}_{0.4}$	0.278	− 0.279	0.922	0.420	− 0.228	0.865
$\text{Hybrid}_{0.7}$	0.212	− 0.274	0.939	0.355	− 0.216	0.878

are statistically significant positive, which indicates that patients with earlier ages infected by the HIV are more likely to have lower CD4 counts. (3) The proposed four estimates based on MNAR and CC assumptions are different in most of cases. Therefore, the ignorable dropout assumption is questionable.

8 Summary

Handling longitudinal data with nonignorable dropout is a challenging problem, mainly due to the issue of identifiability of the nonresponse propensity and how to incorporate the within-subject correlations. We use a parametric propensity model and the GMM approach making use of a nonresponse instrument to identify unknown parameters in the propensity. The inverse probability weighting is applied to construct the unbiased GEE, and then the matrix expansion idea of QIF and hybrid GEE methods are used to approximate the working correlation. Two classes of improved estimators and confidence regions for GLM are derived based on EL method. Further, the penalized EL method and algorithm for variable selection are investigated.

Some interesting issues still merit further research. Firstly, the proposed method relies on the assumption that the dropout propensity models is correct and an instrument exists. However, it is hard to check this assumption in the presence of nonignorable missing data. Hence, propensity model selection or model averaging methods should be considered. In addition, note that in the real data set, we use age as instrument, because when the CD4 counts are included in the dropout propensity, it is reasonable to believe that age does not add more information to the missing mechanism. In some other applications, some baseline measurements prior to the treatments and categorical covariates such as age group, gender, race and education level are related to the study variable $\boldsymbol{y}_i$, but one or several of them may be unrelated with the propensity when $\boldsymbol{y}_i$ and other covariates are conditioned, which may be considered as an instrument. To make sure an instrument exists, these baseline measurements and categorical covariates should be included in $\boldsymbol{x}_i$. Secondly, the efficiency of proposed GMM estimators $\hat{\boldsymbol{\theta}}_j$ may be improved. When $\boldsymbol{y}_i$ is univariate, several different approaches of determining the optimal estimating equations were proposed by Ai et al. (2018) to achieve the semiparametric efficiency bound. Zhao et al. (2017) also proposed the maximum likelihood estimation, semiparametric likelihood estimation and EL-based IPW approaches to estimate the unknown parameters in the propensity. Thirdly, we focus on parametric propensity models (2), while an extension of our approach to semiparametric dropout propensity models described in Kim and Yu (2011) and Shao and Wang (2016). The nonparametric component in the propensity can be profiled using a kernel-type estimator. It is also of interest to investigate the composite quantile regression (Zou and Yuan 2008) procedure to achieve robustness and estimation consistency. Some further research will be conducted.

Acknowledgements We are grateful to the Editor, the associate editor and two anonymous referees for their insightful comments and suggestions, which have led to significant improvements. Our research was supported by the National Natural Science Foundation of China (11871287, 11501208, 11771144, 11801359), the Natural Science Foundation of Tianjin (18JCYBJC41100), the Fundamental Research Funds for the Central Universities, the Key Laboratory for Medical Data Analysis and Statistical Research of Tianjin. The two authors contributed equally to this work.

References

Ai, C., Linton, O., Zhang, Z. (2018). A simple and efficient estimation method for models with nonignorable missing data. *Statistica Sinica*, to appear.

Bai, Y., Fung, W. K., Zhu, Z. (2010). Weighted empirical likelihood for generalized linear models with longitudinal data. *Journal of Statistical Planning and Inference*, *140*, 3446–3456.

Cantoni, E., Flemming, J. M., Ronchetti, E. (2005). Variable selection for marginal longitudinal generalized linear models. *Biometrics*, *61*, 507–514.

Chen, J., Chen, Z. (2008). Extended Bayesian information criterion for model selection with large sample space. *Biometrika*, *95*, 759–771.

Cho, H., Qu, A. (2015). Efficient estimation for longitudinal data by combining large-dimensional moment conditions. *Electronic Journal of Statistics*, *9*, 1315–1334.

Diggle, P., Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis (with discussion). *Journal of the Royal Statistical Society Series C (Applied Statistics)*, *43*, 49–93.

Fan, J., Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, *96*, 1348–1360.

Fu, L., Wang, Y. (2012). Quantile regression for longitudinal data with a working correlation model. *Computational Statistics and Data Analysis*, *56*, 2526–2538.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, *50*, 1029–1054.

Huang, J., Liu, L., Liu, N. (2007). Estimation of large covariance matrices of longitudinal data with basis function approximations. *Journal of Computational and Graphical Statistics*, *16*, 189–209.

Ibrahim, J. G., Lipsitz, S. R., Horton, N. (2001). Using auxiliary data for parameter estimation with nonignorably missing outcomes. *Journal of the Royal Statistical Society*: *Series C (Applied Statistics)*, *50*, 361–373.

Kim, J. K., Yu, C. L. (2011). A semiparametric estimation of mean functionals with nonignorable missing data. *Journal of the American Statistical Association*, *106*, 157–165.

Leng, C., Zhang, W. (2014). Smoothing combined estimating equations in quantile regression for longitudinal data. *Statistics and Computing*, *24*, 123–136.

Leng, C., Zhang, W., Pan, J. (2010). Semiparametric mean-covariance regression analysis for longitudinal data. *Journal of the American Statistical Association*, *105*, 181–193.

Leung, D., Wang, Y., Zhu, M. (2009). Efficient parameter estimation in longitudinal data analysis using a hybrid GEE method. *Biostatistics*, *10*, 436–445.

Li, D., Pan, J. (2013). Empirical likelihood for generalized linear models with longitudinal data. *Journal of Multivariate Analysis*, *114*, 63–73.

Liang, K., Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*, 13–22.

Little, R. J. A., Rubin, D. B. (2002). *Statistical Analysis with Missing Data*2nd ed. New York: Wiley.

Lv, J., Guo, C., Yang, H., Li, Y. (2017). A moving average Cholesky factor model in covariance modeling for composite quantile regression with longitudinal data. *Computational Statistics and Data Analysis*, *112*, 129–144.

Miao, W., Tchetgen Tchetgen, E. J. (2016). On varieties of doubly robust estimators under missingness not at random with a shadow variable. *Biometrika*, *103*, 475–482.

Molenberghs, G., Kenward, M. (2007). *Missing Data in Clinical Studies*. West Sussex: John Wiley and Sons.

Owen, A. (2001). *Empirical Likelihood*. Boca Raton, FL: Chapman and Hall/CRC Press.

Qin, J., Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, *22*, 300–325.

Qu, A., Lindsay, B. G., Li, B. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika*, *87*, 823–836.

Rao, J. N. K., Scott, A. J. (1981). The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American statistical Association*, *76*, 221–230.
Robins, J. M., Rotnitzky, A., Zhao, L. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, *89*, 846–866.
Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of statistics*, *6*, 461–464.
Shao, J., Wang, L. (2016). Semiparametric inverse propensity weighting for nonignorable missing data. *Biometrika*, *103*, 175–187.
Tang, G., Little, R. J. A., Raghunathan, T. E. (2003). Analysis of multivariate missing data with non-ignorable nonresponse. *Biometrika*, *90*, 747–764.
Wang, H., Li, B., Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society*: *Series B (Statistical Methodology)*, *71*, 671–683.
Wang, L., Qi, C., Shao, J. (2019). Model-assisted regression estimators for longitudinal data with non-ignorable dropout. *International Statistical Review*, *87*, S121–S138.
Wang, S., Shao, J., Kim, J. K. (2014). An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica*, *24*, 1097–1116.
Xu, L., Tang, M. L., Chen, Z. (2019). Analysis of longitudinal data by combining multiple dynamic covariance models. *Statistics and Its Interface*, *12*, 479–487.
Xue, L., Zhu, L. (2007). Empirical likelihood semiparametric regression analysis for longitudinal data. *Biometrika*, *94*, 921–937.
You, J., Chen, G., Zhou, Y. (2006). Block empirical likelihood for longitudinal partially linear regression models. *Canadian Journal of Statistics*, *34*, 79–96.
Zahner, G. E., Pawelkiewicz, W., DeFrancesco, J. J., Adnopoz, J. (1992). Children's mental health service needs and utilization patterns in an urban community: An epidemiological assessment. *Journal of the American Academy of Child and Adolescent Psychiatry*, *31*, 951–960.
Zhang, W., Leng, C. (2011). A moving average Cholesky factor model in covariance modelling for longitudinal data. *Biometrika*, *99*, 141–150.
Zhang, W., Leng, C., Tang, C. Y. (2015). A joint modelling approach for longitudinal studies. *Journal of the Royal Statistical Society*: *Series B (Statistical Methodology)*, *77*, 219–238.
Zhao, P., Tang, N., Jiang, D. (2017). Efficient inverse probability weighting method for quantile regression with nonignorable missing data. *Statistics*, *51*, 363–386.
Zhou, J., Qu, A. (2012). Informative estimation and selection of correlation structure for longitudinal data. *Journal of the American statistical Association*, *107*, 701–710.
Zou, H., Yuan, M. (2008). Composite quantile regression and the oracle model selection theory. *The Annals of Statistics*, *36*, 1108–1126.