Computational Statistics and Data Analysis 82 (2015) 74–88

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

An efficient and robust variable selection method for longitudinal generalized linear models

Jing Lv [*], Hu Yang, Chaohui Guo

College of Mathematics and Statistics, Chongqing University, Chongqing, 401331, China

HIGHLIGHTS

- We develop a new efficient and robust variable selection approach for generalized linear models with longitudinal data.
- The root n-consistency and asymptotic normality of the proposed estimators are established.
- An efficient algorithm is proposed to implement the procedures.
- Simulation studies and a real data example have shown that our proposed estimators are superior to some recently developed variable selection methods.

ARTICLE INFO

Article history:
Received 24 April 2014
Received in revised form 17 July 2014
Accepted 12 August 2014
Available online 22 August 2014

Keywords:
Efficiency
Generalized estimating equations
Generalized linear model
Longitudinal data
Oracle property
Robust estimation

ABSTRACT

This paper presents a new efficient and robust smooth-threshold generalized estimating equations for generalized linear models (GLMs) with longitudinal data. The proposed method is based on a bounded exponential score function and leverage-based weights to achieve robustness against outliers both in the response and the covariate domain. Our motivation for the new variable selection procedure is that it enables us to achieve better robustness and efficiency by introducing an additional tuning parameter γ which can be automatically selected using the observed data. Moreover, its performance is near optimal and superior to some recently developed variable selection methods. Under some regularity conditions, the resulting estimator possesses the consistency in variable selection and the oracle property in estimation. Finally, simulation studies and a detailed real data analysis are carried out to assess and illustrate the finite sample performance, which show that the proposed method works better than other existing methods, in particular, when many outliers are included.

1. Introduction

Longitudinal data sets arise frequently from many subject-matter studies, such as medical and public health studies. Generalized linear models (McCullagh and Nelder, 1989) are powerful and popular technique for modeling clustered and longitudinal data, in particular, for repeated or correlated non-Gaussian data, such as Binomial or Poisson type response that is commonly encountered in longitudinal studies. The challenge in analyzing longitudinal data is that it is difficult to specify the full likelihood function when responses are non-normal. This motivated Liang and Zeger (1986) to develop an approach of generalized estimating equations (GEE) which is a milestone in the development of methodology for longitudinal data analysis. Moreover, the well-known GEE approach only requires specification of marginal mean and covariance function.

[*] Corresponding author. Tel.: +86 13594603914.
E-mail address: cqulv2015@126.com (J. Lv).

http://dx.doi.org/10.1016/j.csda.2014.08.006

Recent research works on the GEE method include Xie and Yang (2003), Wang et al. (2005a), Balan and Schiopu-Kratina (2005) and Wang (2011) and so on. However, the GEE method is in principle very similar to the weighted least squares method, which does not possess robust property. In the longitudinal data set, one outlier in the subject level may generate a set of outliers in the sample due to repeated measurements. Hence, robustness against outliers is a very important issue in longitudinal studies. Recently, the traditional robust M-estimations (e.g. Huber's estimation) for longitudinal data have attracted much attention and have been discussed in many literatures. An incomplete list of recent works on the robust GEE methods include Fan et al. (2012), He et al. (2005), Qin and Zhu (2007), Qin et al. (2009), Wang et al. (2005b) and Zheng et al. (2013). These above papers all use the Mallows-type weights to downweight the effect of leverage points and adopt the Huber's score function on the Pearson residuals to dampen the effect of outliers in the response.

Although the Huber's score function is a robust modeling tool, it has limitation in terms of efficiency of estimation, which stimulates a lot of work about finding other bounded score functions to achieve better robustness and efficiency. Here we only list a few. Yao et al. (2012) investigated a new estimation method for the classical nonparametric model based on a local modal regression (LMR). Liu et al. (2013) and Zhang et al. (2013) extended the LMR method to single index models and semiparametric partially linear varying coefficient models respectively. The outstanding merit of the procedure is that it can achieve both robustness and efficiency by introducing an additional tuning parameter. Wang et al. (2013) adopted a similar view and proposed a class of robust regression estimators based on exponential squared loss. To be more specific, for the linear regression model $y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \varepsilon_i$, they estimated the regression parameter $\boldsymbol{\beta}$ by minimizing

$$Q_\gamma(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left(1 - \varphi_\gamma(t_i)\right), \tag{1}$$

where $\varphi_\gamma(t_i) = \exp\left(-t_i^2/\gamma\right)$ with $t_i = y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}$, $\gamma > 0$ determines the degree of robustness of the estimation. If γ is large, we have $1 - \exp\left(-t^2/\gamma\right) \approx t^2/\gamma$. Thereby, the new estimators are similar to the least squares estimators. For observations with large absolute values of $t_i = y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}$, we can use a smaller γ to downweight the influence of an outlier on the estimators. Obviously, minimizing the objective function (1) is equivalent to solve the following estimating equations

$$\sum_{i=1}^{n}\boldsymbol{x}_i\psi_\gamma(t_i) = \boldsymbol{0}, \tag{2}$$

where $\psi_\gamma(t) = \dot{\varphi}_\gamma(t) = -\frac{2t}{\gamma}\exp\left(-t^2/\gamma\right)$, $\psi_\gamma(\cdot)$ is the first derivative of $\varphi_\gamma(\cdot)$. Note that $\psi_\gamma(t)$ is also a bounded score function since $\psi_\gamma(t)$ will go to zero when t approaches infinity. Wang et al. (2013) pointed out that their method loses less efficiency of estimation compared with other existing robust methods, e.g. Huber's estimate (Huber, 1973), quantile regression estimate (Koenker, 2005), composite quantile regression estimate (Zou and Yuan, 2008), etc. However, as discussed above literatures, the new approach was only considered for independent data. In this paper, we first extend the bounded exponential score function $\psi_\gamma(t)$ to longitudinal data analysis for achieving a more robust and effective estimate.

Variable selection is a technique of selecting a subset of relevant covariates for constructing reliable statistical models. Various penalty functions (such as Lasso, adaptive Lasso, SCAD) have been used to select significant variables among all the candidate variables for independent data. But variable selection is also a fundamentally important issue for the analysis of longitudinal data, which could greatly enhance the prediction performance of the fitted model and select significant variables. For example, Fan et al. (2012) proposed penalized robust estimating equations with the SCAD penalty for longitudinal linear models. Wang et al. (2012) considered the SCAD-penalized GEE for analyzing longitudinal data with high-dimensional covariates. These variable selection methods mentioned above are based on estimating equations and the SCAD penalty function which is singular at zero. Thereby, these estimation procedures require convex optimization, which incurs a computational burden. Ueki (2009) developed smooth-threshold estimating equations that can automatically eliminate irrelevant parameters by setting them as zero. As far as we know, Ueki's method is in principle applicable to the procedures based on estimating equations. Moreover, this approach possesses the oracle property in the sense that Fan and Li (2001) suggested. Motivated by the idea of Ueki (2009) and Li et al. (2013) developed smooth-threshold generalized estimating equations (SGEE) for longitudinal generalized linear models. In this paper, we focus on marginal longitudinal generalized linear models and develop an efficient and robust variable selection method based on the bounded exponential score function and smooth-threshold estimating equations. Our contributions are the following: (i) we establish root n-consistency and asymptotic normality of estimators. (ii) We use the Mallows-type weights to downweight the effect of leverage points and adopt the bounded exponential score function $\psi_\gamma(t)$ on the Pearson residuals to dampen the effect of outliers in the response. (iii) The proposed method can automatically eliminate inactive predictors by setting the corresponding parameters to be zero and estimate nonzero coefficients simultaneously.

The rest of the paper is organized as follows. The main results are described in Section 2, including the efficient and robust smooth-threshold estimating equations (ERSGEE) and their asymptotic properties and influence function. In Section 3, an efficient algorithm is proposed to implement the procedures. Moreover, we discuss how to select the tuning parameter γ and other regularization parameters so that the corresponding ERSGEE estimators are robust and sparse. In Section 4, we apply a number of simulations to assess the finite sample performance of our method by comparing with other variable selection procedures. A real data analysis is also presented in this section to augment our theoretical results. Some concluding remarks are given in Section 5. The proofs of the main results are relegated to the Appendix.

2. Methodology

2.1. New efficient and robust smooth-threshold estimating equations

Suppose the response variable for the ith subject is measured m_i times, $\boldsymbol{Y}_i = \left(y_{i1}, \ldots, y_{im_i}\right)^T$, $i = 1, \ldots, n$, n is the sample size and m_i is the cluster size for a total of $N = \sum_{i=1}^n m_i$ observations. The corresponding covariate $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{im_i})^T$ is an $m_i \times p$ matrix for the ith subject. Let $\boldsymbol{\varepsilon}_i = \left(\varepsilon_{i1}, \ldots, \varepsilon_{im_i}\right)^T$ be a random error vector and $\boldsymbol{\beta}$ be a $p \times 1$ regression coefficient. For a marginal longitudinal linear regression model $\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon_i}$, Fan et al. (2012) chose a bounded Huber's score function and proposed the following robust estimating equations

$$U_n^{\mathrm{REE}}(\boldsymbol{\beta}, \alpha) = \sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{V}_i^{-1} \boldsymbol{h}_i(\mathbf{X}_i\boldsymbol{\beta}) = \mathbf{0}, \tag{3}$$

where $\mathbf{V}_i = \mathbf{R}_i(\alpha)\mathbf{A}_i^{1/2}$, $\mathbf{A}_i = \sigma^2 I_{m_i}$ (σ^2 is variance of ε_{ij}) and $\mathbf{R}_i(\alpha)$ is the working correlation matrix that involves unknown parameter vector α. The core of the estimating equations are $\mathbf{h}_i(\mathbf{X}_i\boldsymbol{\beta}) = \mathbf{W}_i\left(\kappa\left(\boldsymbol{A}_i^{-1/2}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})\right) - C_i\right)$, where $\kappa(r) = \min(c, \max(-c, r))$ and $C_i = E\left\{\kappa\left(\boldsymbol{A}_i^{-1/2}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})\right)\right\}$ is used to ensure Fisher consistency of the estimator. The Huber's score function $\kappa(\cdot)$ is chosen to downweight the influence of outliers in the response variable. The weight matrix $\mathbf{W}_i = \mathrm{diag}(w_{i1}, \ldots, w_{im_i})$ is used to bound the effect of leverage points in the covariate space. Many different types of weights could be designed. A common choice is Mallows-type weight function which is a function of the Mahalanobis distance, that is,

$$w_{ij} = w\left(\boldsymbol{x}_{ij}\right) = \min\left\{1, \left(\frac{b_0}{\left(\boldsymbol{x}_{ij} - \boldsymbol{m}_x\right)^T \mathbf{S}_x^{-1}\left(\boldsymbol{x}_{ij} - \boldsymbol{m}_x\right)}\right)^{\rho/2}\right\} \tag{4}$$

with $\rho \geq 1$, $\boldsymbol{m}_x$ and $\boldsymbol{S}_x$ are some robust estimates of the location and scale of $\boldsymbol{x}_{ij}$. One can use high breakdown point location and scatter estimators such as Minimum Covariance Determinant (MCD) and Minimum Volume Ellipsoid (MVE). b_0 is the 0.95 quantile of the chi-square distribution with p degrees of freedom.

Although the bounded Huber's score function $\kappa(\cdot)$ may dampen the effect of outliers in the response, and result in a robust estimator by solving the estimation equations (3), it will lose some efficiency in some cases. That is, the robust Huber's estimator is slightly less efficient than non-robust estimation when there are no outliers, e.g. Fan et al. (2012), He et al. (2005), Qin and Zhu (2007) and Qin et al. (2009), etc. Moreover, estimation equations (3) are only suitable for marginal longitudinal linear regression models. Hence, it is necessary to look for a new efficient and robust method for marginal longitudinal GLMs. In this paper, we focus on the marginal longitudinal GLMs. Denote the first two marginal moments of y_{ij} by $\mu_{ij} = E\left(y_{ij}|\boldsymbol{x}_{ij}\right) = g\left(\boldsymbol{\beta}^T\boldsymbol{x}_{ij}\right)$, $\mathrm{Var}\left(y_{ij}|\boldsymbol{x}_{ij}\right) = \phi v\left(\mu_{ij}\right)$, $i = 1, \ldots, n, j = 1, \ldots, m_i$. Here, $v(\cdot)$ is a variance function, and ϕ is a scale parameter. Motivated by (2) and (3), we propose the following efficient and robust generalized estimating equations (ERGEE)

$$U_n^{\mathrm{ERGEE}}(\boldsymbol{\beta}, \alpha) = \sum_{i=1}^{n} \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{h}_i^{\gamma}\left(\boldsymbol{\mu}_i(\boldsymbol{\beta})\right), \tag{5}$$

where $\mathbf{V}_i = \mathbf{R}_i(\alpha)\mathbf{A}_i^{1/2}$, $\boldsymbol{\mu}_i = \left(\mu_{i1}, \ldots, \mu_{im_i}\right)^T$, $\mathbf{A}_i = \phi\mathrm{diag}\left(v\left(\mu_{i1}\right), \ldots, v\left(\mu_{im_i}\right)\right)$, $\mathbf{D}_i = \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}$ is an $m_i \times p$ matrix, $\boldsymbol{h}_i^{\gamma}\left(\boldsymbol{\mu}_i\right) = \boldsymbol{W}_i\left[\psi_\gamma\left(\boldsymbol{\mu}_i(\boldsymbol{\beta})\right) - C_i\left(\boldsymbol{\mu}_i(\boldsymbol{\beta})\right)\right]$ with $\psi_\gamma\left(\boldsymbol{\mu}_i\right) = \psi_\gamma\left(\mathbf{A}_i^{-1/2}\left(\mathbf{Y}_i - \boldsymbol{\mu}_i\right)\right)$ and $C_i\left(\boldsymbol{\mu}_i\right) = E\left[\psi_\gamma\left(\boldsymbol{\mu}_i\right)\right]$ which is developed to ensure the Fisher consistency of the estimate. Other definitions are the same as those in the estimation equations (2) and (3). For models using the canonical link (see McCullagh and Nelder, 1989), we have $\boldsymbol{D}_i = \boldsymbol{A}_i\boldsymbol{X}_i$.

We assume that some components of $\boldsymbol{\beta}$ are zero in the true model. Our main goal is to identify the zero coefficients consistently and estimate the nonzero coefficients efficiently and robustly. Motivated by the idea of Ueki (2009) and Li et al. (2013), we consider the following efficient and robust smooth-threshold generalized estimating equations (ERSGEE)

$$\left(\boldsymbol{I}_p - \boldsymbol{\Delta}\right)U_n^{\mathrm{ERGEE}}(\boldsymbol{\beta}, \alpha) + \boldsymbol{\Delta}\boldsymbol{\beta} = \mathbf{0}, \tag{6}$$

where $\boldsymbol{\Delta}$ is the diagonal matrix with diagonal elements being $\delta = \left(\delta_j\right)_{j=1,\ldots,p}$ and $\boldsymbol{I}_p$ is the p-dimensional identity matrix. When $\delta_j = 1$ in the ERSGEE, it reduces to $\beta_j = 0$, then we can obtain a sparse solution. But we cannot directly obtain the estimator of $\boldsymbol{\beta}$ by solving (6). The main reason is that the ERSGEE not only includes the unknown nuisance parameters α and ϕ, but also involves δ_j and tuning parameter γ, which need to be determined by the observed data. Detailed discussions of selecting these parameters are listed in Section 3.

Since the $\mathbf{V}_i$'s are functions of α, ϕ and $\boldsymbol{\beta}$, they can be reexpressed as functions of $\boldsymbol{\beta}$ alone by first substituting a $\sqrt{n}$-consistent estimator, $\hat{\alpha}(\boldsymbol{\beta}, \phi)$ in the efficient and robust generalized estimating function $U_n^{\mathrm{ERGEE}}(\boldsymbol{\beta}, \alpha)$ for α, and then replacing ϕ in $\hat{\alpha}$ by a $\sqrt{n}$-consistent estimator. Follow Li et al. (2013)'s suggestion, we choose $\hat{\delta}_j = \min\left\{1, \lambda/\left|\hat{\beta}_j^{(0)}\right|^{(1+\tau)}\right\}$

with an initial estimator $\hat{\beta}_j^{(0)}$ which can be obtained by solving $U_n^{\text{ERGEE}}\left(\boldsymbol{\beta}, \hat{\alpha}[\boldsymbol{\beta}, \hat{\phi}(\boldsymbol{\beta})]\right) = \mathbf{0}$ for the full model. Replacing $\boldsymbol{\Delta}$ in (6) by $\hat{\boldsymbol{\Delta}}$ with diagonal elements $\hat{\delta} = (\hat{\delta}_j)_{j=1,\ldots,p}$, Eqs. (6) become

$$\left(\boldsymbol{I}_p - \hat{\boldsymbol{\Delta}}\right) U_n^{\text{ERGEE}}\left(\boldsymbol{\beta}, \hat{\alpha}(\boldsymbol{\beta}, \hat{\phi}(\boldsymbol{\beta}))\right) + \hat{\boldsymbol{\Delta}}\boldsymbol{\beta} = \mathbf{0}. \tag{7}$$

The solution of (7) denoted by $\hat{\boldsymbol{\beta}}_{\lambda,\tau}$ is called the ERSGEE estimator.

2.2. Asymptotic properties of the ERSGEE estimator

In order to study the large sample properties of our proposed estimator, we need to make some assumptions. Firstly, we assume that $\boldsymbol{\beta} \in \boldsymbol{\Theta}$, $\boldsymbol{\Theta} \subseteq \mathbb{R}^p$, where $\boldsymbol{\Theta}$ is bounded subset in $\mathbb{R}^p$. Let $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0p})^T$ be the true value of $\boldsymbol{\beta}$. Without loss of generality, we partition $\boldsymbol{\beta}_0$ into active (nonzero) and inactive (zero) coefficients as follows: let $\mathscr{A}_0 = \left\{j : \beta_{0j} \neq 0\right\}$ and $\mathscr{A}_0^c = \left\{j : \beta_{0j} = 0\right\}$ be the complement of $\mathscr{A}_0$. Denote by $s = |\mathscr{A}_0|$ the number of true nonzero parameters. Throughout the paper, we use $\|\cdot\|$ to denote the Euclidean norm. We define the active set $\mathscr{A} = \left\{j : \hat{\delta}_j \neq 1\right\}$ as the set of indices of nonzero parameters, where $\hat{\delta}_j = \min\left\{1, \lambda/\left|\hat{\beta}_j^{(0)}\right|^{(1+\tau)}\right\}$. Theorem 1 below presents the consistency of the ERSGEE estimators.

Theorem 1. *Suppose that the regularity conditions* (C1)–(C8) *in the Appendix hold, we have*

$$\|\hat{\boldsymbol{\beta}}_{\lambda,\tau} - \boldsymbol{\beta}_0\| = O_p(n^{-1/2}).$$

Furthermore, Theorem 2 below shows that such consistent estimators presented in Theorem 1 possess the sparsity property and the estimators of nonzero coefficients have the same asymptotic distribution as that based on the correct submodel.

Theorem 2. *Suppose that the conditions of Theorem 1 hold, as $n \to \infty$, we have*

(a) *variable selection consistency, i.e.* $P(\mathscr{A} = \mathscr{A}_0) \to 1$,
(b) *asymptotic normality, i.e.*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\lambda,\tau,\mathscr{A}_0} - \boldsymbol{\beta}_{\mathscr{A}_0}) \xrightarrow{d} N(0, \boldsymbol{\Sigma}_{\mathscr{A}_0}^{-1}\mathbf{B}_{\mathscr{A}_0}\boldsymbol{\Sigma}_{\mathscr{A}_0}^{-1}),$$

where $\boldsymbol{\Sigma}_{\mathscr{A}_0}$ and $\mathbf{B}_{\mathscr{A}_0}$ are the first $s \times s$ submatrices of $\boldsymbol{\Sigma}$ and $\mathbf{B}$. The definitions of $\boldsymbol{\Sigma}$ and $\mathbf{B}$ are given in the conditions (C6) *and* (C7) *respectively in the Appendix.*

Remark 1. Theorem 1 implies that our proposed approach can achieve the $\sqrt{n}$ consistency of the regression coefficient estimation. Theorem 2 suggests that the ERSGEE possesses oracle property (Fan and Li, 2001) when we choose proper λ and τ. That is, with probability approaching 1, the ERSGEE can correctly select the nonzero coefficients, and estimate them as efficiently as the ERGEE does as if we knew in advance the correct submodel.

2.3. Influence function

Now, we study the robustness property of the ERSGEE. As we know, for robust estimation, the influence function is a convenient and essential tool to measure the local robustness (Hampel et al., 1986), which characters the impact of an infinitesimal proportion of contamination. The influence function of a statistical function T at the model distribution F_0 is defined by

$$IF(\boldsymbol{z}; T, F_0) = \lim_{\varepsilon \to 0} \frac{T((1-\varepsilon)F_0 + \varepsilon\Delta_{(\mathbf{x}^*, \boldsymbol{y}^*)}) - T(F_0)}{\varepsilon}, \tag{8}$$

where F_0 is the joint cumulative distribution function of the underlying distribution without contamination, $\Delta_{(\mathbf{x}^*, \boldsymbol{y}^*)}$ represents the point mass probability distribution at a contaminated point $\boldsymbol{z} = (\mathbf{x}^*, \boldsymbol{y}^*)$, and ε is a constant and $\varepsilon \in (0, 1)$.

If an estimator is robust, $IF(\boldsymbol{z}; T, F_0)$ should not be arbitrarily large for any value of $\boldsymbol{z}$. That is, $IF(\boldsymbol{z}; T, F_0)$ should be bounded for all values of $\boldsymbol{z}$ if the estimator is robust (Hampel et al., 1986). Note that $\hat{\boldsymbol{\beta}}_{\lambda,\tau}$ is a solution of the score function

$$\varsigma(\boldsymbol{X}, \boldsymbol{Y}; \boldsymbol{\beta}) = U_n^{\text{ERGEE}}(\boldsymbol{\beta}, \alpha(\boldsymbol{\beta})) + (\boldsymbol{I}_p - \hat{\boldsymbol{\Delta}})^{-1}\hat{\boldsymbol{\Delta}}\boldsymbol{\beta}.$$

We first deduce the influence function of the statistical function corresponding to the estimation of the parameter $\boldsymbol{\beta}$. Thereby, the statistical function relative to the ERSGEE estimator is given by $\boldsymbol{\beta}_{\text{ERSGEE}}(F)$, which is a solution of $E_F[\varsigma(\boldsymbol{X}, \boldsymbol{Y}; \boldsymbol{\beta})] = \mathbf{0}$ for any distribution F of the variables $(\boldsymbol{X}, \boldsymbol{Y})$. At the model distribution F_0, it follows that $\boldsymbol{\beta}_{\text{ERSGEE}}(F_0) = \boldsymbol{\beta}_0$, with $\boldsymbol{\beta}_0$ the true parameter vector. From the literature of Hampel et al. (1986), Croux et al. (2013) and references in, we have

$$IF(\boldsymbol{z}; \boldsymbol{\beta}_{\text{ERSGEE}}, F_0) = -E_{F_0}\left[\left.\frac{\partial\varsigma(\boldsymbol{X}, \boldsymbol{Y}; \boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right]^{-1} \varsigma(\boldsymbol{z}; \boldsymbol{\beta}_0). \tag{9}$$

Note that the first factor in (9) is a constant square matrix, which is independent of $\boldsymbol{z}$. Thus the shape of the influence function is mainly determined by the second factor $\varsigma(\boldsymbol{z};\boldsymbol{\beta}_0)=\sum_{i=1}^n \boldsymbol{D}_{0i}^{*T}\boldsymbol{V}_{0i}^{*-1}\boldsymbol{W}_i^*\left[\psi_\gamma\left(\boldsymbol{\mu}_i^*\left(\boldsymbol{\beta}_0\right)\right)-C_i\left(\boldsymbol{\mu}_i^*\left(\boldsymbol{\beta}_0\right)\right)\right]+\hat{\boldsymbol{G}}\boldsymbol{\beta}_0$, where $\hat{\mathbf{G}}=\left(\boldsymbol{I}_p-\hat{\boldsymbol{\Delta}}\right)^{-1}\hat{\boldsymbol{\Delta}}$, $\mathbf{D}_{0,i}^*$, $\mathbf{V}_{0,i}^*$ and $\boldsymbol{W}_i^*$ are evaluated at $\boldsymbol{Y}=\boldsymbol{y}^*$, $\boldsymbol{X}=\mathbf{x}^*$ and $\boldsymbol{\mu}_i(\boldsymbol{\beta})=\boldsymbol{\mu}_i^*(\boldsymbol{\beta}_0)$, with jth component of $\boldsymbol{\mu}_i^*(\boldsymbol{\beta}_0)$ being $g(\boldsymbol{\beta}_0^T\mathbf{x}_{ij}^*)$.

By the proof of Theorem 2, we know that $\left\|\frac{1}{\sqrt{n}}\hat{\boldsymbol{G}}\boldsymbol{\beta}_0\right\|^2=\left\|\frac{1}{\sqrt{n}}\hat{\boldsymbol{G}}_{\mathscr{A}_0}\boldsymbol{\beta}_{\mathscr{A}_0}\right\|^2=o_p(n^{-2})$, so $\left\|\hat{\boldsymbol{G}}\boldsymbol{\beta}_0\right\|=o_p(n^{-1/2})$. Meanwhile, $\psi_\gamma(\cdot)$ is a bounded score function and V_i is bounded based on condition (C4) in the Appendix. Therefore, it is easy to show that the influence function of the ERSGEE estimator is bounded in $\boldsymbol{z}$ since $\mathbf{x}_i^{*T}\mathbf{W}_i^*$ is bounded by the definition of the Mallows-type weight function (4) in Section 2.1. This indicates our proposed method is a robust estimate. By the proof of Theorem 2, we also know that $\hat{\delta}_j=o_p(n^{-1/2})$ for $j\in\mathscr{A}_0$ and $P\left\{\hat{\delta}_j=1 \text{ for all } j\in\mathscr{A}_0^c\right\}\to 1$. Furthermore, we have $E_{F_0}\left[\left.\frac{\partial\varsigma(\boldsymbol{X},\boldsymbol{Y};\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right]^{-1}=[\boldsymbol{\Sigma}+\hat{\boldsymbol{G}}]^{-1}$ with $\hat{G}_j\to\infty$ for $j\in\mathscr{A}_0^c$. Therefore, the corresponding influence function of the zero coefficients are zero. Moreover, the influence function of the nonzero coefficients has the following form: $IF(\boldsymbol{z};\boldsymbol{\beta}_{\text{ERSGEE}},F_0)=-\boldsymbol{\Sigma}_{\mathscr{A}_0}^{-1}\sum_{i=1}^n\boldsymbol{D}_{0i,\mathscr{A}_0}^{*T}\boldsymbol{V}_{0i,\mathscr{A}_0}^{*-1}\boldsymbol{h}_{0i,\mathscr{A}_0}^{*\gamma}$, where $\boldsymbol{\Sigma}_{\mathscr{A}_0}$ are the first $s\times s$ submatrices of $\boldsymbol{\Sigma}$ which is given in the condition (C6) in the Appendix. $\boldsymbol{D}_{0i,\mathscr{A}_0}^*$ and $\boldsymbol{V}_{0i,\mathscr{A}_0}^*$ are the first $s\times s$ submatrices of $\boldsymbol{D}_{0i}^*$ and $\boldsymbol{V}_{0i}^*$, and $\boldsymbol{h}_{0i,\mathscr{A}_0}^*$ is the first s components of $\boldsymbol{W}_i^*\left[\psi_\gamma\left(\boldsymbol{\mu}_i^*\left(\boldsymbol{\beta}_0\right)\right)-C_i\left(\boldsymbol{\mu}_i^*\left(\boldsymbol{\beta}_0\right)\right)\right]$.

3. Issues in practical implementation

3.1. Nuisance parameters

To obtain the ERSGEE estimator $\hat{\boldsymbol{\beta}}$ for solving (7) using the Fisher scoring method, we need to obtain the $\sqrt{n}$-consistent estimators of the correlation parameter α and scale parameter ϕ. Therefore, we first discuss the estimations of the correlation parameter and scale parameter. To proceed, let $\boldsymbol{e}_i=\left(e_{i1},\ldots,e_{im_i}\right)^T=(\phi\boldsymbol{A}_i)^{-1/2}\left(\boldsymbol{Y}_i-\boldsymbol{\mu}_i\right)$ be the standardized residuals. For a chosen function ψ_γ, we denote $\psi_\gamma\left(\boldsymbol{e}_i\right)=\{\psi_\gamma\left(e_{i1}\right),\ldots,\psi_\gamma\left(e_{i,m_i}\right)\}$ as robust residuals. Similar to Wang et al. (2005b), we can obtain a robust estimate of ϕ through the median absolute deviation

$$\hat{\phi}=\left\{1.483\text{ median}\left\{\left|\hat{\xi}_{ij}-\text{median}\left(\hat{\xi}_{ij}\right)\right|\right\}\right\}^2, \tag{10}$$

where $\hat{\xi}_{ij}=A_{ij}^{-1/2}\left(Y_{ij}-\mu_{ij}(\hat{\boldsymbol{\beta}})\right)$ is the Pearson residual and $\hat{\boldsymbol{\beta}}$ is the current estimate of $\boldsymbol{\beta}$. To obtain a robust estimate of α, we use the robust moment estimator motivated by Wang et al. (2005b). For example, for the exchangeable working correlation structure, α can be estimated by

$$\hat{\alpha}=\frac{1}{nH^2}\sum_{i=1}^n\frac{1}{m_i\left(m_i-1\right)}\sum_{j\neq k}\psi_\gamma\left(e_{ij}\right)\psi_\gamma\left(e_{ik}\right). \tag{11}$$

For the first order autoregressive working correlation structure, α can be estimated by

$$\hat{\alpha}=\frac{1}{nH^2}\sum_{i=1}^n\frac{1}{m_i-1}\sum_{t\leq m_i-1}\psi_\gamma\left(e_{it}\right)\psi_\gamma\left(e_{i,t+1}\right), \tag{12}$$

where H^2 is the mean of $\psi_\gamma^2\left(e_{ij}\right)$, $i=1,\ldots,n$ and $t=1,\ldots,m_i$.

3.2. Algorithm

In this subsection, we first apply a Fisher scoring procedure to achieve numerical solutions of estimating equations $U_n^{\text{ERGEE}}\left(\boldsymbol{\beta},\hat{\alpha}[\boldsymbol{\beta},\hat{\phi}(\boldsymbol{\beta})]\right)=\mathbf{0}$ for null model and obtain the optimal tuning parameter γ_{opt}. More specifically, we propose the iterative algorithm to implement the procedures as follows.

Step 1. Given an initial estimator $\tilde{\boldsymbol{\beta}}^{(0)}$, where the choice of $\tilde{\boldsymbol{\beta}}^{(0)}$ will be defined later. Let $k=0$.

Step 2. We estimate the correlation parameter α and scale parameter ϕ using the current estimate $\tilde{\boldsymbol{\beta}}^{(k)}$ and compute the working covariance matrix $\mathbf{V}_i\{\boldsymbol{\mu}_i(\tilde{\boldsymbol{\beta}}^{(k)}),\hat{\alpha}[\tilde{\boldsymbol{\beta}}^{(k)},\hat{\phi}(\tilde{\boldsymbol{\beta}}^{(k)})]\}=\mathbf{R}_i(\hat{\alpha})\hat{\boldsymbol{A}}_i^{1/2}$. Meanwhile, we find the tuning parameter γ controls the degree of robustness and efficiency of the proposed estimator. Similar to Wang et al. (2013), we propose a data-driven procedure to get $\gamma_{\text{opt}}^{(k)}$ by minimizing $\det(\text{Cov}(\tilde{\boldsymbol{\beta}}^{(k)}))$, where $\det(\cdot)$ denotes the determinant operator and $\text{Cov}(\tilde{\boldsymbol{\beta}})$ is defined as follows:

$$\text{Cov}(\tilde{\boldsymbol{\beta}})=\left[\hat{\boldsymbol{\Sigma}}_n\left(\boldsymbol{\mu}_i(\tilde{\boldsymbol{\beta}})\right)\right]^{-1}\hat{\boldsymbol{B}}_n\left(\boldsymbol{\mu}_i(\tilde{\boldsymbol{\beta}})\right)\left[\hat{\boldsymbol{\Sigma}}_n\left(\boldsymbol{\mu}_i(\tilde{\boldsymbol{\beta}})\right)\right]^{-1}, \tag{13}$$

where $\hat{\boldsymbol{\Sigma}}_n\left(\boldsymbol{\mu}_i(\tilde{\boldsymbol{\beta}})\right)$ and $\hat{\boldsymbol{B}}_n\left(\boldsymbol{\mu}_i(\tilde{\boldsymbol{\beta}})\right)$ are defined by

$$\hat{\boldsymbol{\Sigma}}_n\left(\boldsymbol{\mu}_i(\tilde{\boldsymbol{\beta}})\right)=\sum_{i=1}^{n}\mathbf{D}_i^T\left(\boldsymbol{\mu}_i(\tilde{\boldsymbol{\beta}})\right)\mathbf{V}_i^{-1}\left(\boldsymbol{\mu}_i(\tilde{\boldsymbol{\beta}})\right)\boldsymbol{\Gamma}_i\left(\boldsymbol{\mu}_i(\tilde{\boldsymbol{\beta}})\right)\mathbf{D}_i\left(\boldsymbol{\mu}_i(\tilde{\boldsymbol{\beta}})\right),$$

with $\boldsymbol{\Gamma}_i(\boldsymbol{\mu}_i(\boldsymbol{\beta}))=E\dot{\mathbf{h}}_i^{\gamma}(\boldsymbol{\mu}_i(\boldsymbol{\beta}))=E\partial\mathbf{h}_i^{\gamma}(\boldsymbol{\mu}_i(\boldsymbol{\beta}))/\partial(\boldsymbol{\mu}_i)|_{\boldsymbol{\mu}_i=\boldsymbol{\mu}_i(\boldsymbol{\beta})}$. And

$$\hat{\boldsymbol{B}}_n\left(\boldsymbol{\mu}_i(\tilde{\boldsymbol{\beta}})\right)=\sum_{i=1}^{n}\boldsymbol{D}_i^T\left(\boldsymbol{\mu}_i(\tilde{\boldsymbol{\beta}})\right)\mathbf{V}_i^{-1}\left(\boldsymbol{\mu}_i(\tilde{\boldsymbol{\beta}})\right)\left[\mathbf{h}_i^{\gamma}\left(\boldsymbol{\mu}_i(\tilde{\boldsymbol{\beta}})\right)\left\{\mathbf{h}_i^{\gamma}\left(\boldsymbol{\mu}_i(\tilde{\boldsymbol{\beta}})\right)\right\}^T\right]\mathbf{V}_i^{-1}\left(\boldsymbol{\mu}_i(\tilde{\boldsymbol{\beta}})\right)\boldsymbol{D}_i^T\left(\boldsymbol{\mu}_i(\tilde{\boldsymbol{\beta}})\right).$$

Step 3. Update the estimator $\tilde{\boldsymbol{\beta}}^{(k+1)}$ of $\boldsymbol{\beta}$ by using the following iterative procedure

$$\tilde{\boldsymbol{\beta}}^{(k+1)}=\tilde{\boldsymbol{\beta}}^{(k)}-\left.\left\{\left(\sum_{i=1}^{n}\boldsymbol{X}_i^T\boldsymbol{\Omega}_i(\boldsymbol{\mu}_i(\boldsymbol{\beta}))\boldsymbol{X}_i\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{X}_i^T\mathbf{A}_i^T(\boldsymbol{\mu}_i(\boldsymbol{\beta}))\boldsymbol{V}_i^{-1}(\boldsymbol{\mu}_i(\boldsymbol{\beta}))\boldsymbol{h}_i^{\gamma}(\boldsymbol{\mu}_i(\boldsymbol{\beta}))\boldsymbol{\beta}\right)\right\}\right|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}^{(k)}}, \tag{14}$$

where $\boldsymbol{\Omega}(\boldsymbol{\mu}_i(\boldsymbol{\beta}))=\mathbf{A}_i^T(\boldsymbol{\mu}_i(\boldsymbol{\beta}))\mathbf{V}_i^{-1}(\boldsymbol{\mu}_i(\boldsymbol{\beta}))\boldsymbol{\Gamma}_i(\boldsymbol{\mu}_i(\boldsymbol{\beta}))\mathbf{A}_i(\boldsymbol{\mu}_i(\boldsymbol{\beta}))$.

Step 4. Set $k\leftarrow k+1$ and iterate *Step* 2–*Step* 3 until convergence, and we define the final estimators of $\boldsymbol{\beta}$ and γ as $\tilde{\boldsymbol{\beta}}$ and $\tilde{\gamma}_{\text{opt}}$.

Remark 2. Note that the quantities (α,ϕ,γ) all change with iterations. For simplicity, we implicitly regard (α,ϕ,γ) as functions of $\boldsymbol{\beta}$. Thus as the estimate of $\boldsymbol{\beta}$ changes in each iteration, these quantities also change. The tuning parameter γ controls the degree of robustness and efficiency of the proposed regression estimator. To select γ, we propose a data-driven procedure that yields both high robustness and high efficiency simultaneously. Hence, we choose the tuning parameter γ by minimizing the determinant of asymptotic covariance matrix in Step 2 for attaining high efficiency. Since the calculation of $\det(\text{Cov}(\tilde{\boldsymbol{\beta}}^{(k)}))$ depends on estimate $\tilde{\boldsymbol{\beta}}^{(k)}$, we update $\tilde{\boldsymbol{\beta}}^{(k)}$ in Step 3 and repeat the algorithm until convergence.

Remark 3. In the initialization step, we need a initial estimator $\tilde{\boldsymbol{\beta}}^{(0)}$ which is an important task in practice. In simulation studies, we use RGEE estimator as the initial value for solving the above iterative algorithm. The RGEE estimator is based on the Huber's score function and its definition in detail is given in Example 1 in Section 4.1. Another initial value for the RGEE estimator is obtained by the GEE estimator with the independent correlation structure, which is $\sqrt{n}$ consistent. In addition, if $\left\|\tilde{\boldsymbol{\beta}}^{(k+1)}-\tilde{\boldsymbol{\beta}}^{(k)}\right\|^2$ is smaller than a cutoff value $\epsilon>0$ (such as 10^{-6}), then we stop the iteration. Our numerical experiences in Section 4 indicate that the convergence criterion is met within 30 iterations.

To decrease the computational burden, we use the $\tilde{\gamma}_{\text{opt}}$ obtained above as the optimal tuning parameter in achieving numerical solutions of estimating equations (7). Compared to the solutions of $U_n^{\text{ERGEE}}\left(\boldsymbol{\beta},\hat{\alpha}[\boldsymbol{\beta},\hat{\phi}(\boldsymbol{\beta})]\right)=\mathbf{0}$, the only difference of obtaining solutions of (7) is to replace (14) by the following formula

$$\hat{\boldsymbol{\beta}}^{(k+1)}=\hat{\boldsymbol{\beta}}^{(k)}-\left.\left\{\left(\sum_{i=1}^{n}\boldsymbol{X}_i^T\boldsymbol{\Omega}_i(\boldsymbol{\mu}_i(\boldsymbol{\beta}))\boldsymbol{X}_i+\hat{\mathbf{G}}\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{X}_i^T\mathbf{A}_i^T(\boldsymbol{\mu}_i(\boldsymbol{\beta}))\boldsymbol{V}_i^{-1}(\boldsymbol{\mu}_i(\boldsymbol{\beta}))\boldsymbol{h}_i^{\gamma}(\boldsymbol{\mu}_i(\boldsymbol{\beta}))+\hat{\mathbf{G}}\boldsymbol{\beta}\right)\right\}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^{(k)}}, \tag{15}$$

where $\hat{\mathbf{G}}=\left(\boldsymbol{I}_p-\hat{\boldsymbol{\Delta}}\right)^{-1}\hat{\boldsymbol{\Delta}}$. Then we denote the final estimators of $\boldsymbol{\beta}$ by solving the estimation equation (7) as $\hat{\boldsymbol{\beta}}$.

Remark 4. In simulation studies, to obtain the ERSGEE estimator $\hat{\boldsymbol{\beta}}$ by solving (7), we also need a initial estimator $\hat{\boldsymbol{\beta}}^{(0)}$ which can be obtained by using the ERGEE estimator $\tilde{\boldsymbol{\beta}}$. From our simulation experiences, we can see that this robust method provides an useful and stable initial estimator $\hat{\boldsymbol{\beta}}^{(0)}$ which is not sensitive to outliers.

3.3. Regularization parameters selection

To implement the procedures described in Section 3.2, we need to choose the regularization parameters (λ,τ). One can select (λ,τ) by optimizing some data-driven criteria which can balance goodness of fit and model complexity. In order to simplify the calculation, we adopt $\tau=1$ in our simulation studies. Following Li et al. (2013), we use PWD-type criterion to choose these parameters. That is, we choose λ as the minimizer of

$$PWD_{\lambda}=W\text{Dev}+df_{\lambda}\log(n), \tag{16}$$

where $W\text{Dev}=\sum_{i=1}^{n}\left\{\mathbf{h}_i^{\gamma}\left(\boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}}_{\lambda})\right)\right\}^T\mathbf{R}_i^{-1}\left(\boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}}_{\lambda})\right)\mathbf{h}_i^{\gamma}\left(\boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}}_{\lambda})\right)$, $\hat{\boldsymbol{\beta}}_{\lambda}$ is the ERSGEE estimator for given λ and $df_{\lambda}=\sum_{j=1}^{p}1(\hat{\delta}_j\neq 1)$ denotes the number of nonzero parameters with $1(\cdot)$ denoting the indicator function. The selected λ minimizes the PWD_{λ}.

4. Numerical studies

In this section, we consider simulation studies to assess the performance of the proposed estimators from three aspects: (1) the effect of using the true working correlation matrix as compared with misspecified working correlation structure, (2) robustness against outliers, and (3) efficiency comparison with some recently developed methods. We also apply our proposed method to a real-world longitudinal data set.

4.1. Simulation studies

Example 1. In this example, we consider the marginal longitudinal linear regression model as well as the marginal longitudinal Poisson regression model. The main goal is to compare the proposed ERGEE by solving Eqs. (5) with the conventional GEE (Li et al., 2013) and RGEE (Fan et al., 2012). RGEE is defined through the same Eqs. (5) except that the bounded exponential score function $\psi_\gamma(t)$ is replaced by Huber's score function $\kappa(\cdot)$. Firstly, the continuous response variable is generated according to the model

$$y_{ij} = x_{ij}^{(1)}\beta_{01} + x_{ij}^{(2)}\beta_{02} + x_{ij}^{(3)}\beta_{03} + \varepsilon_{ij}, \quad i = 1, \ldots, n, j = 1, \ldots, 5, \tag{17}$$

where $\beta_{01} = 0.7, \beta_{02} = 0.7$ and $\beta_{03} = -0.4$, x_{ij} are drawn from a standard normal distribution with the correlation between the kth and lth component of x_{ij} being $0.5^{|l-k|}$ and the random error vectors $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{i5})^T$ are generated independently of the covariates from a five-dimensional standard normal distribution with exchangeable true correlation structure, correlation coefficient $\alpha = 0.7$.

Secondly, we generate data from the correlated multiple Poisson distribution. Suppose that the true marginal mean of count response variable y_{ij} is defined by

$$\log\left(\mu_{ij}\right) = x_{ij}^{(1)}\beta_{01} + x_{ij}^{(2)}\beta_{02} + x_{ij}^{(3)}\beta_{03}, \quad i = 1, \ldots, n,\ j = 1, \ldots, 5. \tag{18}$$

For the covariates, we generate $x_{ij}^{(k)}$ independently such that $x_{ij}^{(k)} \sim$ Uniform $(1, 2), k = 1, 2, 3$. Similar to the model (17), we assume that the true correlation structure within the ith subject is exchangeable correlation structure with correlation coefficient 0.7. The regression coefficients of model (18) are the same as those of the model (17). The response variable is conducted using multivariate Poisson data generator proposed by Yahav and Shmueli (2012). The replication time is 200 for each experiment.

To illustrate the effect incorporating within-cluster correlation on estimation efficiency, two kinds of working correlation matrices are considered: exchangeable working correlation (exch) and first order autoregressive working correlation structure (ar1). In our simulations, the ρ in the weight function w_{ij} is chosen to be $\rho = 1$ and the constant c in Huber's function is chosen to be 2. The simulation including contaminated cases are conducted to assess the performance of the proposed robust method. To be specific, we denote *Case*0 as no contamination situation and consider the following two types of contaminations

*Case*1: randomly choose 10% of y_{ij} to be $y_{ij} + 10$;
*Case*2: randomly choose 2% of $x_{ij}^{(1)}$ to be $x_{ij}^{(1)} + 3$ and 10% of y_{ij} to be $y_{ij} + 10$.

In addition, we note that model (18) is more challenging than model (17) since the Poisson response contains much less information than the continuous response. Therefore, we consider less outliers for model (18) with the following two types of contaminations

*Case*1′: randomly choose 5% of y_{ij} to be $y_{ij} + 10$;
*Case*2′: randomly choose 1% of $x_{ij}^{(1)}$ to be $x_{ij}^{(1)} + 1$ and 5% of y_{ij} to be $y_{ij} + 10$.

We use the average mean square error (AMSE) to measure the accuracy of estimation, which is $\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right\|^2$ averaged over 200 simulated data sets.

The AMSE and the bias and the corresponding sample standard deviations (SD) of estimates with $n = 100$ are displayed in Tables 1 and 2. Several observations can be found from the above two tables. Firstly, under *Case*0, our proposed ERGEE performs equally well as the GEE estimate since their bias, SD and AMSE have little difference, and the ERGEE estimate generally achieves smaller SD than that of the RGEE estimate. These results indicate that our proposed method seems to perform no worse than GEE estimate by introducing an additional tuning parameter γ which controls the degree of robustness and efficiency for the estimators. That is, to some extent, the ERGEE estimator will lose less efficiency when the data contains no outliers compared with other robust estimators. Secondly, the GEE method can impact greatly in terms of SD and AMSE when data contain outliers which implies that it is not a robust approach. Furthermore, the ERGEE estimate performs obviously better than the other two methods for the contaminated data. Thus, we can conclude that the proposed estimation procedure can achieve better robustness and efficiency and outperform both GEE and RGEE, in particular, when many outliers are included.

Example 2. In this example, we investigate the performance of the proposed ERSGEE estimate and compare it with existing variable selection methods, such as SGEE proposed in Li et al. (2013) and RSGEE which is similar to that of Fan et al. (2012).

Table 1
Simulation results for the marginal longitudinal linear regression model in Example 1 with $n = 100$.

			β_1		β_2		β_3		AMSE
			Bias	SD	Bias	SD	Bias	SD	
*Case*0	ar1	GEE	−0.0040	0.0429	−0.0014	0.0487	−0.0027	0.0438	0.0061
		RGEE	−0.0037	0.0415	−0.0020	0.0478	−0.0012	0.0459	0.0061
		ERGEE	−0.0041	0.0414	−0.0013	0.0471	−0.0014	0.0449	0.0059
	exch	GEE	−0.0045	0.0390	−0.0011	0.0449	−0.0030	0.0412	0.0052
		RGEE	−0.0043	0.0388	−0.0017	0.0461	−0.0017	0.0436	0.0055
		ERGEE	−0.0045	0.0388	−0.0012	0.0457	−0.0020	0.0429	0.0054
*Case*1	ar1	GEE	0.0001	0.1737	0.0031	0.1750	0.0097	0.1739	0.0906
		RGEE	−0.0021	0.0626	0.0013	0.0669	0.0009	0.0627	0.0126
		ERGEE	−0.0031	0.0464	0.0065	0.0505	−0.0028	0.0456	0.0068
	exch	GEE	0.0020	0.1700	0.0034	0.1721	0.0045	0.1721	0.0877
		RGEE	−0.0011	0.0608	0.0012	0.0648	0.0006	0.0607	0.0115
		ERGEE	−0.0023	0.0457	0.0002	0.0477	−0.0020	0.0425	0.0062
*Case*2	ar1	GEE	−0.0885	0.1464	0.0472	0.1763	−0.0200	0.1690	0.0912
		RGEE	−0.0872	0.0631	0.0452	0.0690	−0.0095	0.0704	0.0234
		ERGEE	−0.0882	0.0530	0.0441	0.0561	−0.0052	0.0519	0.0184
	exch	GEE	−0.0937	0.1443	0.0486	0.1743	−0.0179	0.1677	0.0904
		RGEE	−0.0901	0.0623	0.0457	0.0670	−0.0082	0.0667	0.0230
		ERGEE	−0.0880	0.0511	0.0435	0.0538	−0.0040	0.0504	0.0176

Table 2
Simulation results for the marginal longitudinal Poisson regression model in Example 1 with $n = 100$.

			β_1		β_2		β_3		AMSE
			Bias	SD	Bias	SD	Bias	SD	
*Case*0	ar1	GEE	0.0065	0.0409	−0.0067	0.0385	−0.0014	0.0528	0.0060
		RGEE	0.0068	0.0411	−0.0069	0.0391	−0.0002	0.0527	0.0061
		ERGEE	0.0063	0.0411	−0.0063	0.0385	−0.0019	0.0521	0.0060
	exch	GEE	0.0032	0.0365	−0.0054	0.0338	−0.0002	0.0460	0.0046
		RGEE	0.0029	0.0365	−0.0049	0.0347	−0.0002	0.0449	0.0046
		ERGEE	0.0026	0.0359	−0.0045	0.0344	−0.0005	0.0449	0.0046
*Case*1′	ar1	GEE	−0.0012	0.0714	−0.0266	0.0799	0.0784	0.0858	0.0255
		RGEE	0.0099	0.0534	−0.0057	0.0542	0.0252	0.0628	0.0104
		ERGEE	0.0133	0.0517	0.0046	0.0487	−0.0080	0.0552	0.0083
	exch	GEE	−0.0076	0.0681	−0.0250	0.0748	0.0743	0.0823	0.0231
		RGEE	0.0085	0.0513	−0.0061	0.0496	0.0261	0.0583	0.0092
		ERGEE	0.0139	0.0500	0.0038	0.0436	0.0261	0.0505	0.0071
*Case*2′	ar1	GEE	−0.0866	0.0742	0.0156	0.0731	0.1044	0.0844	0.0364
		RGEE	−0.0506	0.0557	0.0230	0.0574	0.0480	0.0619	0.0155
		ERGEE	−0.0380	0.0519	0.0310	0.0570	0.0092	0.0587	0.0119
	exch	GEE	−0.0858	0.0653	0.0119	0.0691	0.1046	0.0761	0.0331
		RGEE	−0.0495	0.0488	0.0200	0.0533	0.0464	0.0571	0.0134
		ERGEE	−0.0382	0.0447	0.0288	0.0531	0.0088	0.0546	0.0101

Note that SGEE is obtained by solving the estimating equations (2.4) of Li et al. (2013) and the criterion (4.1) in their article is used to select regularization parameter λ. RSGEE is defined through the same Eqs. (6) except that the bounded exponential score function $\psi_\gamma(t)$ is replaced by Huber's score function $\kappa(\cdot)$, and we replace $\psi_\gamma(t)$ in the model selection criterion (16) by $\kappa(\cdot)$ to select the optimal regularization parameter λ for RSGEE. We also use the two above models (17) and (18) except that the true regression coefficients are replaced by $\boldsymbol{\beta}_0 = (0.7, 0.7, -0.4, 0, 0, 0, 0, 0, 0, 0)^T$, and other settings are the same. For sake of comparison, we adopt the notation (C, IC, CF) to identify the variable selection results. The C means the average number of zero regression coefficients that are correctly estimated as zero, IC presents the average number of non-zero regression coefficients incorrectly set to zero, and correct fit (CF) depicts the proportion of times that the correct model is selected.

Tables 3 and 4 summarize model selection properties of SGEE, RSGEE, ERSGEE and their corresponding oracle's methods for two different working correlation matrices and three different cases. Several observations can be seen from the two tables. Firstly, under *Case*0, AMSE of all methods decrease and their CF approach to 1 and the values in the column labeled C become more and more closer to the true number of zero regression coefficients in the models as the number of subjects n increases. These are consistent with the oracle property. Moreover, ERSGEE seems to perform no worse than SGEE in terms of CF, although acceptable loss of efficiency in our method can be detected from slightly larger AMSE (total) in Table 3. Secondly, for the contaminated data (*Case*1 and *Case*2), our method apparently outperforms the other two methods in terms of estimation efficiency and variable selection. For example, in Table 3, CF of ERSGEE is about 90% with the moderate sample

Table 3
Variable selection results for the marginal longitudinal linear regression model in Example 2.

	n	Method	ar1					exch				
			C	IC	CF	oracle	total	C	IC	CF	oracle	total
*Case*0	50	SGEE	6.95	0.03	0.91	0.0116	0.0219	6.99	0.04	0.94	0.0097	0.0207
		RSGEE	6.71	0.16	0.63	0.0115	0.0473	6.76	0.15	0.65	0.0100	0.0445
		ERSGEE	6.90	0.03	0.91	0.0108	0.0411	6.92	0.05	0.91	0.0099	0.0398
	100	SGEE	7	0.04	0.96	0.0066	0.0159	7	0.04	0.96	0.0054	0.0145
		RSGEE	6.84	0.13	0.74	0.0067	0.0322	6.85	0.13	0.74	0.0058	0.0317
		ERSGEE	6.94	0.02	0.94	0.0066	0.0264	6.96	0.03	0.94	0.0058	0.0244
	200	SGEE	7	0.02	0.98	0.0035	0.0085	7	0.02	0.98	0.0028	0.0081
		RSGEE	6.96	0.07	0.90	0.0034	0.0189	6.98	0.08	0.91	0.0030	0.0182
		ERSGEE	7	0.02	0.98	0.0032	0.0136	7	0.01	0.99	0.0028	0.0102
*Case*1	50	SGEE	6.26	1.18	0.11	0.1864	0.5949	6.47	1.18	0.13	0.1812	0.5610
		RSGEE	6.40	0.08	0.49	0.0258	0.0524	6.50	0.08	0.57	0.0237	0.0490
		ERSGEE	6.85	0.09	0.80	0.0155	0.0498	6.87	0.09	0.80	0.0136	0.0419
	100	SGEE	6.69	1.13	0.13	0.1118	0.4529	6.68	1.09	0.15	0.1063	0.4371
		RSGEE	6.43	0.05	0.55	0.0176	0.0333	6.45	0.03	0.61	0.0156	0.0281
		ERSGEE	6.86	0.06	0.82	0.0091	0.0270	6.92	0.03	0.90	0.0078	0.0171
	200	SGEE	6.93	1.12	0.11	0.0479	0.3435	6.86	1.05	0.13	0.0469	0.3400
		RSGEE	6.68	0.02	0.73	0.0074	0.0135	6.72	0.02	0.76	0.0066	0.0130
		ERSGEE	6.95	0.02	0.94	0.0043	0.0103	6.95	0.02	0.94	0.0036	0.0101
*Case*2	50	SGEE	6.06	1.03	0.11	0.1787	0.5455	6.28	1.05	0.12	0.1696	0.5424
		RSGEE	6.36	0.13	0.46	0.0385	0.0802	6.41	0.13	0.47	0.0356	0.0751
		ERSGEE	6.81	0.12	0.75	0.0263	0.0640	6.88	0.11	0.79	0.0246	0.0622
	100	SGEE	6.76	0.92	0.19	0.0979	0.3549	6.70	0.85	0.20	0.0962	0.3522
		RSGEE	6.45	0.05	0.56	0.0223	0.0446	6.50	0.07	0.58	0.0222	0.0441
		ERSGEE	6.94	0.06	0.87	0.0168	0.0387	6.95	0.03	0.88	0.0163	0.0381
	200	SGEE	6.80	0.87	0.18	0.0524	0.2741	6.67	0.83	0.19	0.0513	0.2594
		RSGEE	6.62	0.02	0.67	0.0163	0.0251	6.62	0.01	0.72	0.0160	0.0240
		ERSGEE	6.95	0.03	0.92	0.0134	0.0242	6.95	0.02	0.92	0.0127	0.0239

Notation: oracle $= \sum_{k=1}^{s}$ AMSE$(\hat{\beta}_k)$; total $= \sum_{k=1}^{p}$ AMSE$(\hat{\beta}_k)$.

size $n = 200$, which indicates ERSGEE can correctly identify the true model under contaminated data, but CF of the other two methods are no more than 80%. Hence, ERSGEE estimate has great gain of efficiency over RSGEE method. Thirdly, it is not surprised that the performances of variable selection procedures and estimation efficiency based on the correct correlation structure (exch) work better than those based on the incorrect correlation structure (ar1). However, we also note that the performance does not significantly depend on working covariance structure. Finally, for a given correlation structure and contamination situation, the larger the sample size n, the better all methods perform.

Example 3. In this example, we discuss how the proposed ERSGEE procedure can be applied to the "large n, diverging p" setup for contaminated longitudinal data. We use the same model as in (17) except that the regression coefficients are reset as $\boldsymbol{\beta}_0 = \left(0.3 \times \mathbf{1}_{s_n}, \mathbf{0}_{p_n - s_n}\right)^T$ with $p_n = [4n^{2/5}] - 5$ and $s_n = [p_n/5]$ for $n = 100, 200$ and 400, where $[s]$ denotes the largest integer not greater than s and $\mathbf{1}_m/\mathbf{0}_m$ defines a m-vector of 1s/0s. In addition, we assume each subject is supposed to be measured at scheduled time points $\{1, 2, \ldots, 5\}$. In this example, we handle unbalanced data, where the number of repeated measurements m_i is randomly chosen such that each time point has a 10% probability of being skipped, resulting in different m_i for each subject. The detailed simulation results are listed in Table 5. From Table 5, we can see that the proposed method is still able to correctly identify the true submodel and works remarkably well even for the high dimension case with outliers. This shows that our method may be extended to the scenario with increased number of variables, which is a quite challenging issue, and the study of our method with the diverging variables in theory needs further investigation in the near future.

In summary, the simulation studies demonstrate that the proposed method is not only able to accommodate the effect of outliers and improve the efficiency of parameter estimation but also can perform robust variable selection under contamination.

4.2. Real data analysis

We now apply the proposed method to analyze a real data set from an epileptic seizure study. Details of the study design can be found in Wang et al. (2005b) and Xu and Zhu (2012). In this study, the response variable is the number of seizures in a two-week period and our concern here is whether the drug helps to reduce the rate of epileptic seizures or not. Following Wang et al. (2005b), we consider the same covariates including treatment (0 for placebo, 1 for drug), logarithm of age (log age), baseline seizure count (which is divided by 4 and then log-transformed), and the interaction between treatment and baseline seizures. As in Wang et al. (2005b), we analyze this data set with ar1 working correlation assumption under

Table 4
Variable selection results for the marginal longitudinal Poisson regression model in Example 2.

	n	Method	ar1					exch				
			C	IC	CF	oracle	total	C	IC	CF	oracle	total
*Case*0	50	SGEE	6.67	0.02	0.75	0.0096	0.0240	6.80	0.04	0.81	0.0079	0.0215
		RSGEE	6.72	0.03	0.75	0.0096	0.0249	6.82	0	0.82	0.0079	0.0222
		ERSGEE	6.74	0	0.78	0.0096	0.0248	6.83	0	0.86	0.0078	0.0212
	100	SGEE	6.85	0.02	0.84	0.0056	0.0130	6.94	0.03	0.90	0.0047	0.0105
		RSGEE	6.84	0.02	0.84	0.0056	0.0128	6.96	0.01	0.91	0.0049	0.0109
		ERSGEE	6.87	0	0.90	0.0055	0.0126	6.91	0	0.90	0.0048	0.0090
	200	SGEE	6.97	0	0.97	0.0029	0.0036	7	0	1	0.0020	0.0033
		RSGEE	6.96	0	0.96	0.0029	0.0038	6.99	0	0.99	0.0022	0.0035
		ERSGEE	6.96	0	0.96	0.0028	0.0038	6.99	0	0.99	0.0021	0.0034
*Case*1′	50	SGEE	5.87	0.41	0.08	0.0324	0.1701	6.28	0.58	0.09	0.0295	0.1699
		RSGEE	6.55	0.05	0.56	0.0158	0.0406	6.61	0.03	0.60	0.0140	0.0280
		ERSGEE	6.54	0.03	0.59	0.0149	0.0306	6.54	0	0.61	0.0128	0.0224
	100	SGEE	6.63	0.68	0.08	0.0237	0.1652	6.76	0.73	0.09	0.0235	0.1625
		RSGEE	6.76	0.01	0.77	0.0103	0.0175	6.82	0.05	0.79	0.0093	0.0171
		ERSGEE	6.78	0	0.82	0.0088	0.0123	6.79	0	0.83	0.0074	0.0098
	200	SGEE	6.75	0.75	0.07	0.0170	0.1656	6.89	0.85	0.08	0.0166	0.1610
		RSGEE	6.87	0	0.88	0.0061	0.0075	6.96	0	0.90	0.0055	0.0073
		ERSGEE	6.92	0	0.92	0.0049	0.0057	6.92	0	0.92	0.0042	0.0052
*Case*2′	50	SGEE	5.45	0.31	0.07	0.0460	0.1748	5.42	0.38	0.03	0.0463	0.1743
		RSGEE	6.65	0.08	0.63	0.0219	0.0537	6.72	0.09	0.65	0.0197	0.0472
		ERSGEE	6.63	0.06	0.63	0.0187	0.0428	6.65	0.03	0.65	0.0152	0.0305
	100	SGEE	5.68	0.30	0.08	0.0324	0.1307	5.92	0.35	0.11	0.0323	0.1283
		RSGEE	6.65	0.01	0.69	0.0146	0.0264	6.81	0.04	0.79	0.0134	0.0261
		ERSGEE	6.69	0.01	0.73	0.0120	0.0203	6.79	0	0.81	0.0103	0.0152
	200	SGEE	6.65	0.07	0.06	0.0264	0.1647	6.86	0.83	0.08	0.0263	0.1848
		RSGEE	6.81	0	0.83	0.0091	0.0140	6.93	0.01	0.92	0.0084	0.0132
		ERSGEE	6.88	0	0.89	0.0065	0.0089	6.96	0	0.96	0.0057	0.0073

Notation: oracle $= \sum_{k=1}^{s} \text{AMSE}(\hat{\beta}_k)$; total $= \sum_{k=1}^{p} \text{AMSE}(\hat{\beta}_k)$.

Table 5
Variable selection results for the high-dimensional marginal longitudinal linear regression model in Example 3.

	(n, p_n, s_n)	ar1					exch				
		C	IC	CF	oracle	total	C	IC	CF	oracle	total
*Case*0	(100, 20, 4)	15.70	0.08	0.75	0.0095	0.0297	15.68	0.06	0.77	0.0087	0.0279
	(200, 28, 5)	22.64	0.02	0.77	0.0062	0.0130	22.73	0.02	0.79	0.0054	0.0107
	(400, 38, 7)	30.74	0	0.82	0.0042	0.0058	30.82	0	0.86	0.0037	0.0048
*Case*1	(100, 20, 4)	15.57	0.12	0.61	0.0117	0.0310	15.56	0.07	0.65	0.0102	0.0282
	(200, 28, 5)	22.61	0.03	0.69	0.0073	0.0137	22.65	0.03	0.72	0.0066	0.0122
	(400, 38, 7)	30.71	0.01	0.76	0.0059	0.0085	30.65	0	0.78	0.0050	0.0060
*Case*2	(100, 20, 4)	15.72	0.26	0.56	0.0136	0.0481	15.70	0.25	0.59	0.0125	0.0464
	(200, 28, 5)	22.65	0.13	0.64	0.0111	0.0307	22.69	0.12	0.69	0.0100	0.0280
	(400, 38, 7)	30.66	0.03	0.72	0.0078	0.0138	30.72	0.03	0.73	0.0074	0.0133

Notation: oracle $= \sum_{k=1}^{s_n} \text{AMSE}(\hat{\beta}_k)$; total $= \sum_{k=1}^{p_n} \text{AMSE}(\hat{\beta}_k)$.

the framework of a longitudinal Poisson regression model. But the number of covariates is only 5 (including intercept), it cannot be used to carry out variable selection. To further demonstrate the effectiveness of the proposed ERSGEE method, we adopt a similar strategy of Guo et al. (2013) by adding 10 irrelevant variables which are randomly drawn from Uniform(0, 1) to the longitudinal Poisson regression model. Accordingly, there are 15 covariates, among which the last 10 covariates (denoted as $r_1, \ldots, r_{10}$) are known to be irrelevant variables, and the first 5 variables might be relevant to the response. We fit the 15-dimensional sparse longitudinal Poisson regression model using the three methods which include the SGEE, RSGEE and ERSGEE. To verify our efficient and robust variable selection procedure, we re-analyzed the data set by including some outliers in the response variable. Two cases were considered in the analysis, a single outlier and 2% contaminated observations. For each case, the outliers were randomly generated by increasing the value y_{ij} to $y_{ij} + 100$. The similar strategy was adopted by Yao and Wang (2013). The mean of regression coefficient estimates and sample standard deviations (in parentheses) for the first 5 variables are presented in Table 6. The results of the appearance frequency of the 15 variables, avg.Size (averaged number of selected variables over the 200 replications) and the correlation between each estimated direction and the direction of SGEE without outliers (denoted by $\text{corr}(\hat{\beta}, \hat{\beta}^{\text{sgee0}})$) are reported in Table 7.

Although we cannot know the true parameter, we can compare the change of the estimated regression coefficients for different cases, for example, no outliers, a single outlier and 2% outliers. If the change is very small, we have reasons to

Table 6
The mean of regression coefficient estimates and sample standard deviations (in parentheses) for epileptic data using the ar1 working correlation over the 200 replications.

Outliers	Method	Intercept	Treatment	Log(age)	Log(baseline)	Interaction
No outlier	SGEE	−2.380(0.260)	−1.283(0.142)	0.960(0.031)	0.765(0.053)	0.522(0.062)
	RSGEE	−2.536(0.191)	−1.261(0.213)	0.992(0.024)	0.780(0.054)	0.464(0.101)
	ERSGEE	−1.723(0.173)	−0.898(0.185)	0.988(0.012)	0.541(0.050)	0.312(0.103)
Single outlier	SGEE	−1.594(2.430)	−1.099(0.112)	0.846(0.398)	0.620(0.664)	0.446(0.448)
	RSGEE	−2.516(0.211)	−1.251(0.228)	0.992(0.026)	0.776(0.062)	0.460(0.108)
	ERSGEE	−1.747(0.174)	−0.891(0.212)	0.988(0.014)	0.548(0.049)	0.305(0.116)
2% outliers	SGEE	−0.499(2.259)	−0.818(1.069)	0.702(0.392)	0.448(0.587)	0.334(0.428)
	RSGEE	−2.473(0.329)	−1.193(0.321)	0.995(0.041)	0.764(0.107)	0.434(0.152)
	ERSGEE	−1.757(0.227)	−0.900(0.205)	0.990(0.015)	0.549(0.065)	0.310(0.115)

Table 7
Variable selection results for epileptic data using the ar1 working correlation with 200 replications.

Variables	No outlier			Single outlier			2% outliers		
	SGEE	RSGEE	ERSGEE	SGEE	RSGEE	ERSGEE	SGEE	RSGEE	ERSGEE
Intercept	1	1	1	0.980	1	1	0.880	1	1
Treatment	1	1	1	0.920	1	1	0.765	1	1
Log(age)	1	1	1	0.885	1	1	0.840	1	1
Log(baseline)	1	1	1	0.820	1	1	0.630	1	1
Interaction	1	0.970	0.980	0.655	0.960	0.955	0.490	0.915	0.970
r_1	0.205	0.045	0.025	0.225	0.050	0.040	0.325	0.055	0.035
r_2	0.180	0.050	0.015	0.240	0.060	0.025	0.305	0.045	0.030
r_3	0.175	0.070	0.045	0.215	0.095	0.050	0.305	0.090	0.035
r_4	0.200	0.085	0.015	0.220	0.050	0.025	0.335	0.065	0.030
r_5	0.185	0.040	0.010	0.225	0.050	0.010	0.380	0.060	0.010
r_6	0.210	0.045	0.040	0.200	0.055	0.030	0.345	0.065	0.040
r_7	0.170	0.055	0.030	0.170	0.080	0.040	0.325	0.080	0.050
r_8	0.180	0.070	0.050	0.190	0.075	0.060	0.320	0.065	0.060
r_9	0.190	0.060	0.010	0.195	0.040	0	0.325	0.055	0.015
r_{10}	0.195	0.055	0.025	0.240	0.060	0.010	0.340	0.070	0.030
avg.size	6.89	5.55	5.25	6.38	5.58	5.25	6.91	5.57	5.31
corr($\hat{\beta}$, $\hat{\beta}^{\text{sgee0}}$)	1	0.985	0.976	0.688	0.985	0.976	0.370	0.981	0.975

The upper panel presents the selected frequency for the 15 variables which consist of the real 5 variables and other 10 irrelevant variables, and the lower panel presents the averaged number of selected variables from the 15 variables and the correlation between each estimated direction and the direction of SGEE estimate without outliers.

believe that the corresponding method is robust. As we can see from Table 6, there is little difference for estimated regression coefficients of RSGEE and ERSGEE under different types of contaminations. But the regression coefficients of SGEE vary hugely from no outliers to 2% outliers. These results indicate RSGEE and ERSGEE are robust estimators but SGEE is not. Meanwhile, we also find the sample standard errors for the estimated coefficients of our method are almost uniformly smaller than those of the RSGEE estimator, which clearly show that our method is more efficient.

From Table 7, we can see that the first five variables have very high selected frequencies for all methods, which indicates that these variables may be very significant in longitudinal Poisson regression model. In addition, SGEE is clearly affected in direction estimation by adding outliers since its correlation has changed significantly but the other two approaches are not. More importantly, our proposed method results in a final model with smaller size.

From the real data analysis, again, we come to a conclusion that our proposed method is not only an efficient variable selection method but also has good robustness for longitudinal generalized linear models even if many outliers are included.

5. Concluding remarks

In this paper, we develop an efficient and robust smooth-threshold estimating equations for automatic variable selection in the marginal longitudinal generalized linear models that allow for non-Gaussian data and nonlinear link functions. Our proposed estimate and variable selection method are able to control the influence of outliers which have been checked by simulation studies and a real data analysis. The distinguishing characteristic of the proposed method is that it introduces additional tuning parameter γ which can be automatically selected using the observed data in order to achieve both robustness and efficiency. Moreover, we have proved that our estimate is consistent in variable selection and enjoys the oracle property followed in Fan and Li (2001). In addition, our estimation procedure can be implemented using an iterative algorithm that alternates between a modified Fisher scoring for the regression coefficients and robust moment estimation of the correlation and scale parameters α and ϕ. Specifically, based on the discussion in Section 2.3, we know that the influence function is bounded with respect to outliers in either the response or the covariate domain.

There are some directions to further extend this work. Firstly, when the number of covariates p_n increases as the number of clusters n increases (such as, Wang et al., 2012 and Cho and Qu, 2013), it is an interesting future research topic for us to use the ERSGEE procedure. Secondly, we can apply our method to other marginal longitudinal semiparametric models, such as single-index models, partially linear single-index models and varying coefficient models, to obtain a robust and efficient estimation.

Acknowledgments

The authors are very grateful to the editor, associate editor, and three anonymous referees for their detailed comments on the earlier version of the manuscript, which led to a much improved paper. This work is supported by the National Natural Science Foundation of China (Grant No. 11171361) and Ph.D. Programs Foundation of Ministry of Education of China (Grant No. 20110191110033).

Appendix

The following conditions are needed to establish the asymptotic properties.

(C1) The covariate vectors are fixed and the first four moments of y_{ij} exist. Also, for each i, $\{m_i\}$ is a bounded sequence of positive integers.

(C2) We assume that the estimated correlation parameter vector $\hat{\alpha}$ is $n^{1/2}$-consistent given $\boldsymbol{\beta}$ and ϕ for some α; that is, $n^{1/2}(\hat{\alpha}-\alpha)=O_p(1)$ for some α, $\hat{\phi}$ is $n^{1/2}$-consistent given $\boldsymbol{\beta}$, and $\left|\partial\hat{\alpha}(\boldsymbol{\beta},\phi)/\partial\phi\right| \le H(\mathbf{Y},\boldsymbol{\beta})$ which is $O_p(1)$, where $H(\cdot,\cdot)$ is a function of the sample $\mathbf{Y}$ and $\boldsymbol{\beta}$.

(C3) $\sup_{i\ge 1} E\left\|\mathbf{h}^{\gamma}_{0,i}(\mathbf{e}_i)\right\|^{2+\delta} < \infty$ for some $\delta > 0$, and $E\mathbf{h}^{\gamma}_{0,i}(\mathbf{e}_i)(\mathbf{h}^{\gamma}_{0,i}(\mathbf{e}_i))^T = \mathbf{F}_i > 0$ with $\sup_i \|\mathbf{F}_i\| < \infty$, where $\mathbf{h}^{\gamma}_{0,i}(\mathbf{e}_i)$ is similar to $\mathbf{h}^{\gamma}_i(\mathbf{e}_i)$, but the former centers $\mathbf{Y}_i$ by it true mean $\boldsymbol{\mu}_{0,i}$, whereas the latter involves centering by $\boldsymbol{\mu}_i$.

(C4) There exists a positive constant c such that $0 < c \le \inf_{i,j} v(\mu_{ij}) \le \sup_{i,j} v(\mu_{ij}) < \infty$. The functions $C_{ij}(\mu_{ij}) = E\psi_{\gamma}\left(A_{ij}^{-1/2}(y_{ij}-\mu_{ij})\right)$, $v(\cdot)$ and $g(\cdot)$ have bounded second derivatives. The function $\psi_{\gamma}(\cdot)$ is piecewise twice differentiable, and the second derivatives are bounded.

(C5) Let $U_n^{\mathrm{ERGEE}}(\boldsymbol{\beta},\alpha) = \sum_{i=1}^n U_i(\boldsymbol{\beta},\alpha)$. $E\left\|U_n^{\mathrm{ERGEE}}(\boldsymbol{\beta}_0,\alpha)\right\|^2 < \infty$, and exist $\delta > 0$, such that,

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} E\left\|U_i(\boldsymbol{\beta}_0,\alpha)\right\|^{2+\delta}}{\left(\left(E\left\|U_n^{\mathrm{ERGEE}}(\boldsymbol{\beta}_0,\alpha)\right\|^2\right)^{1/2}\right)^{2+\delta}} = 0.$$

(C6) Matrix $\boldsymbol{\Sigma} = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n[\mathbf{D}_{0,i}^T\mathbf{V}_{0,i}^{-1}\boldsymbol{\Gamma}_{0,i}(\boldsymbol{\mu}_i(\boldsymbol{\beta}_0))\mathbf{D}_{0,i}^T]$ is positive definite, where $\boldsymbol{\Gamma}_i(\boldsymbol{\mu}_i(\boldsymbol{\beta})) = E\dot{\mathbf{h}}_i^{\gamma}(\boldsymbol{\mu}_i(\boldsymbol{\beta})) = E\partial\mathbf{h}_i^{\gamma}(\boldsymbol{\mu}_i(\boldsymbol{\beta}))/\partial(\boldsymbol{\mu}_i)|_{\boldsymbol{\mu}_i=\boldsymbol{\mu}_i(\boldsymbol{\beta})}$.

(C7) Matrix $\mathbf{B} = \lim_{n\to\infty} n^{-1}\sum_{i=1}^n \mathbf{D}_{0,i}^T\mathbf{V}_{0,i}^{-1}\mathrm{cov}(\mathbf{h}_i^{\gamma}(\boldsymbol{\mu}_i(\boldsymbol{\beta}_0)))(\mathbf{V}_{0,i}^{-1})^T\mathbf{D}_{0,i}$ is positive definite.

(C8) For any positive λ and τ such that $n^{1/2}\lambda \to 0$ and $n^{(1+\tau)/2}\lambda \to \infty$.

Remark 5. Under (C1), the total sample size N is of the same order as the number of subjects n. The estimators $\hat{\alpha}$ and $\hat{\phi}$ given in Section 3.1 satisfy (C2). The conditions (C3)–(C4) imposed on the score function ψ_{γ} are usually easy to check. When ψ_{γ} is bounded, as in Section 2.1, condition (C3) holds automatically. The assumptions (C5)–(C7) are usual conditions for central limit theory and they are expected to hold under general design, where $\mathbf{D}_{0,i}$, $\mathbf{V}_{0,i}$ and $\boldsymbol{\Gamma}_{0,i}$ are evaluated at $\boldsymbol{\mu}_i(\boldsymbol{\beta}) = \boldsymbol{\mu}_i(\boldsymbol{\beta}_0)$.

Proof of Theorem 1. Let $S_n(\boldsymbol{\beta}) = (\boldsymbol{I}_p - \hat{\boldsymbol{\Delta}})U_n^{\mathrm{ERGEE}}(\boldsymbol{\beta},\alpha^*(\boldsymbol{\beta})) + \hat{\boldsymbol{\Delta}}\boldsymbol{\beta}$, where $\alpha^*(\boldsymbol{\beta}) = \hat{\alpha}\left\{\boldsymbol{\beta},\hat{\phi}(\boldsymbol{\beta})\right\}$. It suffices to prove that $\forall\varepsilon > 0$, there exists a constant $C > 0$, such that

$$P\left(\sup_{\|\mathbf{u}\|=C} n^{-1/2}\mathbf{u}^T S_n(\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u}) > 0\right) \ge 1-\varepsilon \tag{A.1}$$

for n large enough. This will imply that there exists a local solution to the equation $S_n(\boldsymbol{\beta}) = 0$ such that $\|\hat{\boldsymbol{\beta}}_{\lambda,\tau} - \boldsymbol{\beta}_0\| = O_p(n^{-1/2})$ with probability at least $1-\varepsilon$. The proof follows that of Theorem 3.6 in Wang (2011), we will evaluate the sign of $n^{-1/2}\mathbf{u}^T S_n(\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u})$ in the ball $\left\{\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u} : \|\mathbf{u}\| = C\right\}$. Note that

$$\begin{aligned} n^{-1/2}\mathbf{u}^T S_n(\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u}) &= n^{-1/2}\mathbf{u}^T S_n(\boldsymbol{\beta}_0) + \frac{1}{n}\mathbf{u}^T\frac{\partial}{\partial\boldsymbol{\beta}}S_n(\tilde{\boldsymbol{\beta}})\mathbf{u} \\ &\triangleq I_{n1} + I_{n2}, \end{aligned} \tag{A.2}$$

where $\tilde{\boldsymbol{\beta}}$ lies between $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u}$. Next we will consider I_{n1} and I_{n2} respectively. For I_{n1}, by some elementary calculations, we have

$$\begin{aligned} I_{n1} &= n^{-1/2}\mathbf{u}^T(\boldsymbol{I}_p - \hat{\boldsymbol{\Delta}})U_n^{\text{ERGEE}}(\boldsymbol{\beta}_0, \alpha) + n^{-1/2}\mathbf{u}^T(\boldsymbol{I}_p - \hat{\boldsymbol{\Delta}})[U_n^{\text{ERGEE}}(\boldsymbol{\beta}_0, \alpha^*(\boldsymbol{\beta}_0)) - U_n^{\text{ERGEE}}(\boldsymbol{\beta}_0, \alpha)] + n^{-1/2}\mathbf{u}^T\hat{\boldsymbol{\Delta}}\boldsymbol{\beta}_0 \\ &\triangleq I_{n11} + I_{n12} + I_{n13}. \end{aligned} \tag{A.3}$$

Using the Cauchy–Schwarz inequality, we have

$$\begin{aligned} |I_{n11}| &\le n^{-1/2}\left\|\mathbf{u}^T(\boldsymbol{I}_p - \hat{\boldsymbol{\Delta}})\right\| \, \left\|U_n^{\text{ERGEE}}(\boldsymbol{\beta}_0, \alpha)\right\| \\ &\le n^{-1/2}(1 - \min_{j\in\mathscr{A}} \hat{\delta}_j(\lambda, \tau))\, \|\mathbf{u}\| \, \left\|U_n^{\text{ERGEE}}(\boldsymbol{\beta}_0, \alpha)\right\|. \end{aligned} \tag{A.4}$$

Since $\min\limits_{j\in\mathscr{A}} \hat{\delta}_j(\lambda, \tau) \le \min\limits_{j\in\mathscr{A}_0} \hat{\delta}_j(\lambda, \tau)$, we only need to obtain the convergence rate of $\min\limits_{j\in\mathscr{A}_0} \hat{\delta}_j(\lambda, \tau)$. Assume that $\hat{\boldsymbol{\beta}}^{(0)}$ is the initial estimator, and is $\sqrt{n}$-consistent. By using the condition $n^{1/2}\lambda \to 0$, for $\varepsilon > 0$ and $j \in \mathscr{A}_0$, we have

$$\begin{aligned} P\left(\hat{\delta}_j(\lambda, \tau) > n^{-1/2}\varepsilon\right) &= P\left(\lambda / \left|\hat{\beta}_j^{(0)}\right|^{1+\tau} > n^{-1/2}\varepsilon\right) \\ &= P\left((n^{1/2}\lambda/\varepsilon)^{1/(1+\tau)} > \left|\hat{\beta}_j^{(0)}\right|\right) \\ &\le P\left((n^{1/2}\lambda/\varepsilon)^{1/(1+\tau)} > \min_{j\in\mathscr{A}_0}\left|\beta_{0j}\right| - O_p(n^{-1/2})\right) \to 0, \end{aligned} \tag{A.5}$$

which implies that $\hat{\delta}_j(\lambda, \tau) = o_p(n^{-1/2})$ for each $j \in \mathscr{A}_0$. Therefore, we have that $\min_{j\in\mathscr{A}} \hat{\delta}_j(\lambda, \tau) = o_p(n^{-1/2})$, together with (A.4) and (A.5), and similar to the proof of Theorem 3.6 in Wang (2011), we can obtain that $|I_{n11}| = O_p(1)\,\|\mathbf{u}\| - o_p(n^{-1/2})\,\|\mathbf{u}\|$. For I_{n12}, using the condition (C2) and Taylor expansion for fixed $\boldsymbol{\beta}_0$, we have

$$\begin{aligned} U_n^{\text{ERGEE}}(\boldsymbol{\beta}_0, \alpha^*(\boldsymbol{\beta}_0)) - U_n^{\text{ERGEE}}(\boldsymbol{\beta}_0, \alpha) &= \frac{\partial}{\partial\alpha}U_n^{\text{ERGEE}}(\boldsymbol{\beta}_0, \alpha)[\alpha^* - \alpha] + o_p(1) \\ &= \frac{\partial}{\partial\alpha}U_n^{\text{ERGEE}}(\boldsymbol{\beta}_0, \alpha)[\hat{\alpha}(\boldsymbol{\beta}_0, \hat{\phi}(\boldsymbol{\beta}_0)) - \hat{\alpha}(\boldsymbol{\beta}_0, \phi) + \hat{\alpha}(\boldsymbol{\beta}_0, \phi) - \alpha] + o_p(1) \\ &= \frac{\partial}{\partial\alpha}U_n^{\text{ERGEE}}(\boldsymbol{\beta}_0, \alpha)\left[\frac{\partial\hat{\alpha}(\boldsymbol{\beta}_0, \phi^*)}{\partial\phi}(\hat{\phi} - \phi) + \hat{\alpha}(\boldsymbol{\beta}_0, \phi) - \alpha\right] + o_p(1) \\ &= o_p(1), \end{aligned} \tag{A.6}$$

where ϕ^* lies between ϕ and $\hat{\phi}$. With the same argument as I_{n11}, together with the above result, we can derive that $|I_{n12}| = o_p(n^{-1/2})\,\|\mathbf{u}\|$. Since $\hat{\delta}_j = \min\left\{1, \lambda/\left|\hat{\beta}_j^{(0)}\right|^{1+\tau}\right\}$, we have $|I_{n13}| \le n^{-1/2}\,\|\mathbf{u}\|\,\left\|\boldsymbol{\beta}_0\right\| = O_p(n^{-1/2}\,\|\mathbf{u}\|)$. Hence $|I_{n1}| = O_p(1)$ $\|\mathbf{u}\|$. For I_{n2}, we can derive that

$$\begin{aligned} I_{n2} &= n^{-1}\mathbf{u}^T\frac{\partial}{\partial\beta}S_n(\tilde{\boldsymbol{\beta}})u \\ &= \mathbf{u}^T(\boldsymbol{I}_p - \hat{\boldsymbol{\Delta}})\left[\frac{\sum\limits_{i=1}^n \mathbf{D}_{0,i}^T\mathbf{V}_{0,i}^{-1}\boldsymbol{\Gamma}_{0,i}(\boldsymbol{\mu}_i(\boldsymbol{\beta}_0))\mathbf{D}_{0,i}}{n}\right]\mathbf{u} + n^{-1}\mathbf{u}^T(\boldsymbol{I}_p - \hat{\boldsymbol{\Delta}})\frac{\partial}{\partial\boldsymbol{\beta}}\left[U_n^{\text{ERGEE}}(\tilde{\boldsymbol{\beta}}, \alpha^*(\tilde{\boldsymbol{\beta}})) - U_n^{\text{ERGEE}}(\tilde{\boldsymbol{\beta}}, \alpha)\right]\mathbf{u} \\ &\quad + n^{-1}\mathbf{u}^T\hat{\boldsymbol{\Delta}}\mathbf{u} + o_p(1)\|\mathbf{u}\|^2 \\ &\triangleq I_{n21} + I_{n22} + I_{n23} + o_p(1)\|\mathbf{u}\|^2. \end{aligned} \tag{A.7}$$

Using the above same argument, it is easy to show that $I_{n22} = o_p(1)\,\|\mathbf{u}\|^2$ and $I_{n23} = O_p(n^{-1})\,\|\mathbf{u}\|^2$. By condition (C6), thus, for sufficiently large n, $n^{-1/2}\mathbf{u}^T S_n(\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u})$ is asymptotically dominated in probability by I_{n21} on $\left\{\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u} : \|\mathbf{u}\| = C\right\}$, which is positive for the sufficiently large C. □

Proof of Theorem 2. With the similar argument as the proof of Theorem 1, we can show that the initial estimator $\hat{\boldsymbol{\beta}}^{(0)}$ obtained by solving the ERGEE $U_n^{\text{ERGEE}}\left(\boldsymbol{\beta}, \hat{\alpha}(\boldsymbol{\beta}, \hat{\phi}(\boldsymbol{\beta}))\right) = \mathbf{0}$ is $\sqrt{n}$-consistent. By the definition of $\mathscr{A}_0^c$ in Section 2.2, we have $\beta_{0j} = 0$ for $j \in \mathscr{A}_0^c$. For any given $j \in \mathscr{A}_0^c$, we have that the initial estimator $\hat{\beta}_j^{(0)}$ satisfies $\left|\hat{\beta}_j^{(0)} - \beta_{0j}\right| = \left|\hat{\beta}_j^{(0)}\right| = O_p(n^{-1/2})$, together with $n^{(1+\tau)/2}\lambda \to \infty$, we can derive that

$$\begin{aligned} P\left(\lambda / \left|\hat{\beta}_j^{(0)}\right|^{1+\tau} < 1\right) &= P\left(\left|\hat{\beta}_j^{(0)}\right|^{1+\tau} > \lambda\right) \\ &\le \lambda^{-1} E\left(\left|\hat{\beta}_j^{(0)}\right|^{1+\tau}\right) \\ &= \lambda^{-1} O(n^{-(1+\tau)/2}) \to 0, \end{aligned} \tag{A.8}$$

where the first inequality applies Markov's inequality and the second equality follows from the fact $\left|\hat{\beta}_j^{(0)}\right| = O_p(n^{-1/2})$ for $j \in \mathscr{A}_0^c$. This implies that

$$P\left\{\hat{\delta}_j = 1 \text{ for all } j \in \mathscr{A}_0^c\right\} \to 1. \tag{A.9}$$

On the other hand, by the condition $n^{1/2}\lambda \to 0$, for $\varepsilon > 0$ and $j \in \mathscr{A}_0$, we have

$$\begin{aligned} P\left(\hat{\delta}_j > n^{-1/2}\varepsilon\right) &= P\left(\lambda / \left|\hat{\beta}_j^{(0)}\right|^{1+\tau} > n^{-1/2}\varepsilon\right) \\ &= P\left((n^{1/2}\lambda/\varepsilon)^{1/(1+\tau)} > \left|\hat{\beta}_j^{(0)}\right|\right) \\ &\le P\left((n^{1/2}\lambda/\varepsilon)^{1/(1+\tau)} > \min_{j\in\mathscr{A}_0} \left|\beta_{0j}\right| - O_p(n^{-1/2})\right) \to 0, \end{aligned} \tag{A.10}$$

which implies that $\hat{\delta}_j = o_p(n^{-1/2})$ for each $j \in \mathscr{A}_0$. Therefore, we prove that $P\left\{\hat{\delta}_j < 1 \text{ for all } j \in \mathscr{A}_0\right\} \to 1$. Thus, we complete the proof of (a).

Next we will prove (b). As shown in (a), $\hat{\beta}_j = 0$ for $j \in \mathscr{A}_0^c$ with probability tending to 1. At the same time, with probability tending to 1, $\hat{\boldsymbol{\beta}}_{\mathscr{A}_0}$ satisfies the efficient and robust smooth-threshold generalized estimating equations (ERSGEE)

$$(\boldsymbol{I}_{|\mathscr{A}_0|} - \hat{\boldsymbol{\Delta}}_{\mathscr{A}_0}) U_n^{\mathrm{ERGEE}}\left(\hat{\boldsymbol{\beta}}_{\mathscr{A}_0}, \hat{\alpha}[\hat{\boldsymbol{\beta}}_{\mathscr{A}_0}, \hat{\phi}(\boldsymbol{\beta}_{\mathscr{A}_0})]\right) + \hat{\boldsymbol{\Delta}}_{\mathscr{A}_0}\hat{\boldsymbol{\beta}}_{\mathscr{A}_0} = \boldsymbol{0}. \tag{A.11}$$

Under some regularity conditions, and applying the Taylor expansion to (A.11) at $\boldsymbol{\beta}_{\mathscr{A}_0}$, it is easy to show that $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathscr{A}_0} - \boldsymbol{\beta}_{\mathscr{A}_0})$ can be approximated by

$$\left[\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\beta}_{\mathscr{A}_0}} \left(U_i\{\boldsymbol{\beta}_{\mathscr{A}_0}, \alpha^*(\boldsymbol{\beta}_{\mathscr{A}_0})\}\right) + \frac{1}{n}\hat{G}_{\mathscr{A}_0}\right]^{-1} \left[-\frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_i\{\boldsymbol{\beta}_{\mathscr{A}_0}, \alpha^*(\boldsymbol{\beta}_{\mathscr{A}_0})\} - \frac{1}{\sqrt{n}}\hat{G}_{\mathscr{A}_0}\boldsymbol{\beta}_{\mathscr{A}_0}\right], \tag{A.12}$$

where

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\beta}_{\mathscr{A}_0}} \left(U_i\{\boldsymbol{\beta}_{\mathscr{A}_0}, \alpha^*(\boldsymbol{\beta}_{\mathscr{A}_0})\}\right) &= \frac{\partial U_i\{\boldsymbol{\beta}_{\mathscr{A}_0}, \alpha^*(\boldsymbol{\beta}_{\mathscr{A}_0})\}}{\partial \boldsymbol{\beta}_{\mathscr{A}_0}} + \frac{\partial U_i\{\boldsymbol{\beta}_{\mathscr{A}_0}, \alpha^*(\boldsymbol{\beta}_{\mathscr{A}_0})\}}{\partial \alpha^*} \frac{\partial \alpha^*(\boldsymbol{\beta}_{\mathscr{A}_0})}{\partial \boldsymbol{\beta}_{\mathscr{A}_0}} \\ &\triangleq I_i + J_i K. \end{aligned}$$

For fixed $\boldsymbol{\beta}_{\mathscr{A}_0}$ and again applying the Taylor expansion, we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_i\{\boldsymbol{\beta}_{\mathscr{A}_0}, \alpha^*(\boldsymbol{\beta}_{\mathscr{A}_0})\} \triangleq I^* + J^* K^* + o_p(1), \tag{A.13}$$

where $I^* = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_i\{\boldsymbol{\beta}_{\mathscr{A}_0}, \alpha\}$, $J^* = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial U_i\{\boldsymbol{\beta}_{\mathscr{A}_0}, \alpha\}}{\partial \alpha}$ and $K^* = \sqrt{n}(\alpha^*(\boldsymbol{\beta}_{\mathscr{A}_0}) - \alpha)$. Note that $\partial U_i\{\boldsymbol{\beta}_{\mathscr{A}_0}, \alpha\}/\partial\alpha$ is a linear function of $\mathbf{h}_i^{\gamma}(\boldsymbol{\mu}_i(\boldsymbol{\beta}_{\mathscr{A}_0}))$ and $E\mathbf{h}_i^{\gamma}(\boldsymbol{\mu}_i(\boldsymbol{\beta}_{\mathscr{A}_0})) = 0$, it is easy to prove that $J^* = o_p(1)$. By conditions (C2), we have

$$\begin{aligned} K^* &= \sqrt{n}[\alpha^*(\boldsymbol{\beta}_{\mathscr{A}_0}) - \alpha] \\ &= \sqrt{n}[\hat{\alpha}(\boldsymbol{\beta}_{\mathscr{A}_0}, \hat{\phi}(\boldsymbol{\beta}_{\mathscr{A}_0})) - \alpha] \\ &= \sqrt{n}[\hat{\alpha}(\boldsymbol{\beta}_{\mathscr{A}_0}, \hat{\phi}(\boldsymbol{\beta}_{\mathscr{A}_0})) - \hat{\alpha}(\boldsymbol{\beta}_{\mathscr{A}_0}, \phi) + \hat{\alpha}(\boldsymbol{\beta}_{\mathscr{A}_0}, \phi) - \alpha] \\ &= \sqrt{n}\left[\frac{\partial \hat{\alpha}(\boldsymbol{\beta}_{\mathscr{A}_0}, \phi^*)}{\partial \phi}(\hat{\phi} - \phi) + \hat{\alpha}(\boldsymbol{\beta}_{\mathscr{A}_0}, \phi) - \alpha\right] \\ &= O_p(1), \end{aligned} \tag{A.14}$$

where ϕ^* is between ϕ and $\hat{\phi}$. On the other hand,

$$\begin{aligned}
\left\| \frac{1}{\sqrt{n}} \hat{G}_{\mathscr{A}_0} \boldsymbol{\beta}_{\mathscr{A}_0} \right\|^2 &\leq \frac{1}{n\left\{1 - \max_{j\in\mathscr{A}_0} \hat{\delta}_j(\lambda,\tau)\right\}^2} \sum_{j\in\mathscr{A}_0} \frac{(\lambda\beta_j)^2}{\hat{\beta}_j^{(0)2(1+\tau)}} \\
&= \frac{\lambda^2}{n\left\{1 - \max_{j\in\mathscr{A}_0} \hat{\delta}_j(\lambda,\tau)\right\}^2} \sum_{j\in\mathscr{A}_0} \left| \hat{\beta}_j^{(0)(-\tau)} + (\beta_j - \hat{\beta}_j^{(0)}) \hat{\beta}_j^{(0)(-1-\tau)} \right|^2 \\
&\leq O_p(n^{-1}\lambda^2) \sum_{j\in\mathscr{A}_0} \left(2\left|\hat{\beta}_j^{(0)}\right|^{-2\tau} + 2\left| (\beta_j - \hat{\beta}_j^{(0)}) \hat{\beta}_j^{(0)(-1-\tau)} \right|^2 \right) \\
&\leq O_p(n^{-1}\lambda^2) \left(2s \min_{j\in\mathscr{A}_0} \left|\hat{\beta}_j^{(0)}\right|^{-2\tau} + 2 \min_{j\in\mathscr{A}_0} \left|\hat{\beta}_j^{(0)}\right|^{2(-1-\tau)} \left\| \boldsymbol{\beta}_{\mathscr{A}_0} - \hat{\boldsymbol{\beta}}_{\mathscr{A}_0}^{(0)} \right\|^2 \right) \\
&= O_p\left\{ ((\sqrt{n}\lambda)^2 n^{-2} \kappa^{-2\tau} s)(1 + O_p(\kappa^{-2} n^{-1})) \right\} = o_p(n^{-2}),
\end{aligned}$$

where $\kappa = \min_{j\in\mathscr{A}_0} \left|\hat{\beta}_j^{(0)}\right|$. Using the same argument, we obtain that

$$\left\| \frac{1}{n} \hat{G}_{\mathscr{A}_0} \right\|^2 = O_p\left\{ ((\sqrt{n}\lambda)^2 n^{-3} \kappa^{-2\tau-2}) \right\} = o_p(n^{-3}).$$

Similarly, it is easy to show that $\sum_{i=1}^n J_i = o_p(n)$ and $K = O_p(1)$. Note that $\mathbf{D}_{i,\mathscr{A}_0} = \partial \boldsymbol{\mu}_i(\boldsymbol{\beta}_{\mathscr{A}_0}) / \partial \boldsymbol{\beta}_{\mathscr{A}_0}$ and Conditions (C5)–(C7), we can prove that $-\frac{1}{\sqrt{n}} \sum_{i=1}^n U_i\{\boldsymbol{\beta}_{\mathscr{A}_0}, \hat{\alpha}(\boldsymbol{\beta}_{\mathscr{A}_0})\}$ is asymptotically equivalent to $-I^*$ whose asymptotic distribution is multivariate Gaussian with mean zero and covariance matrix

$$\lim_{n\to\infty} n^{-1} \sum_{i=1}^n \mathbf{D}_{i,\mathscr{A}_0}^T \mathbf{V}_{i,\mathscr{A}_0}^{-1} \mathrm{cov}(\mathbf{h}_i^\gamma(\boldsymbol{\mu}_i(\boldsymbol{\beta}_{\mathscr{A}_0}))) (\mathbf{V}_{i,\mathscr{A}_0}^{-1})^T \mathbf{D}_{i,\mathscr{A}_0} = \mathbf{B}_{\mathscr{A}_0}.$$

Moreover, as $n \to \infty$, $n^{-1} \sum_{i=1}^n I_i \to \boldsymbol{\Sigma}_{\mathscr{A}_0}$. We complete the proof (b). □

References

Balan, R.M., Schiopu-Kratina, I., 2005. Asymptotic results with generalized estimating equations for longitudinal data. Ann. Statist. 32, 522–541.

Cho, H., Qu, A., 2013. Model selection for correlated data with a diverging number of regression parameters. Statist. Sinica 23, 901–927.

Croux, C., Haesbroeck, G., Ruwet, C., 2013. Robust estimation for ordinal regression. J. Statist. Plann. Inference 143, 1486–1499.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96, 1348–1360.

Fan, Y., Qin, G., Zhu, Z., 2012. Variable selection in robust regression models for longitudinal data. J. Multivariate Anal. 109, 156–167.

Guo, J., Tang, M., Tian, M., Zhu, K., 2013. Variable selection in high-dimensional partially linear additive models for composite quantile regression. Comput. Statist. Data Anal. 65, 56–67.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A., 1986. Robust Statistics: The Approach Based on Influence Functions. John Wiley, New York.

He, X., Fung, W.K., Zhu, Z., 2005. Robust estimation in generalized partial linear models for clustered data. J. Amer. Statist. Assoc. 472, 1176–1184.

Huber, P.J., 1973. Robust regression: asymptotics, conjectures and Monte Carlo. Ann. Statist. 1, 799–821.

Koenker, R., 2005. Quantile Regression. Cambridge University Press.

Li, G., Lian, H., Feng, S., Zhu, L., 2013. Automatic variable selection for longitudinal generalized linear models. Comput. Statist. Data Anal. 61, 174–186.

Liang, K.Y., Zeger, S.L., 1986. Longitudinal data analysis using generalized linear models. Biometrika 73, 13–22.

Liu, J., Zhang, R., Zhao, W., Lv, Y., 2013. A robust and efficient estimation method for single index models. J. Multivariate Anal. 122, 226–238.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, second ed. Chapman & Hall, London.

Qin, G., Zhu, Z., 2007. Robust estimation in generalized semiparametric mixed models for longitudinal data. J. Multivariate Anal. 98, 1658–1683.

Qin, G., Zhu, Z., Fung, W.K., 2009. Robust estimation of covariance parameters in partial linear model for longitudinal data. J. Statist. Plann. Inference 139, 558–570.

Ueki, M., 2009. A note on automatic variable selection using smooth-threshold estimating equations. Biometrika 96, 1005–1011.

Wang, L., 2011. GEE analysis of clustered binary data with diverging number of covariates. Ann. Statist. 39, 389–417.

Wang, N., Carroll, R.J., Lin, X., 2005a. Efficient semiparametric marginal estimation for longitudinal/ clustered data. J. Amer. Statist. Assoc. 100, 147–157.

Wang, X., Jiang, Y., Huang, M., Zhang, H., 2013. Robust variable selection with exponential squared loss. J. Amer. Statist. Assoc. 108, 632–643.

Wang, Y.G., Lin, X., Zhu, M., 2005b. Robust estimation functions and bias correction for longitudinal data analysis. Biometrics 61, 684–691.

Wang, L., Zhou, J., Qu, A., 2012. Penalized generalized estimating equations for high-dimensional longitudinal data analysis. Biometrics 68, 353–360.

Xie, M., Yang, Y., 2003. Asymptotics for generalized estimating equations with large cluster sizes. Ann. Statist. 31, 310–347.

Xu, P., Zhu, L., 2012. Estimation for a marginal generalized single-index longitudinal model. J. Multivariate Anal. 105, 285–299.

Yahav, I., Shmueli, G., 2012. On generating multivariate Poisson data in management science applications. Appl. Stoch. Models Bus. Ind. 28, 91–102.

Yao, W., Lindsay, B., Li, R., 2012. Local modal regression. J. Nonparametr. Stat. 24, 647–663.

Yao, W., Wang, Q., 2013. Robust variable selection through MAVE. Comput. Statist. Data Anal. 63, 42–49.

Zhang, R., Zhao, W., Liu, J., 2013. Robust estimation and variable selection for semiparametric partially linear varying coefficient model based on modal regression. J. Nonparametr. Stat. 25, 523–544.

Zheng, X., Fung, W.K., Zhu, Z., 2013. Robust estimation in joint mean-covariance regression model for longitudinal data. Ann. Inst. Stat. Math. 65, 617–638.

Zou, H., Yuan, M., 2008. Composite quantile regression and the oracle model selection theory. Ann. Statist. 36, 1108–1126.