Mathematics

http://www.aimspress.com/journal/Math

AIMS Mathematics, 7(4): 7199–7211.
DOI:10.3934/math.2022402
Received: 01 November 2021
Revised: 18 January 2022
Accepted: 26 January 2022
Published: 09 February 2022

Research article

Simultaneous variable selection and estimation for longitudinal ordinal data with a diverging number of covariates

Xianbin Chen and Juliang Yin*

School of Economics and Statistics, Guangzhou University, Guangzhou, 510006, China

* **Correspondence:** Email: yin_juliang@hotmail.com.

Abstract: In this paper, we study the problem of simultaneous variable selection and estimation for longitudinal ordinal data with high-dimensional covariates. Using the penalized generalized estimation equation (GEE) method, we obtain some asymptotic properties for these types of data in the case that the dimension of the covariates p_n tends to infinity as the number of cluster n approaches to infinity. More precisely, under appropriate regular conditions, all the covariates with zero coefficients can be examined simultaneously with probability tending to 1, and the estimator of the non-zero coefficients exhibits the asymptotic Oracle properties. Finally, we also perform some Monte Carlo studies to illustrate the theoretical analysis. The main result in this paper extends the elegant work of Wang et al. [1] to the multinomial response variable case.

Keywords: longitudinal ordinal data; penalized generalized estimating equations; high-dimensional covariates; consistency; asymptotic normality
Mathematics Subject Classification: 62E20, 62F12, 62J12

1. Introduction

High-dimensional longitudinal data, comprising repeated observations with a diverging number of parameters, are widely used in bioscience and public health studies. A representative example is the gene expression experiment, in which the data set includes a large number of covariates (see, [1]). In fact, the number of collected variables may not be large, but when various interactive effects are considered, the actual number of predictors in the statistical model will be larger and should be fitted using high-dimensional covariates (see, for example, [2]). This may lead to a more complicated model with many insignificant variables, resulting in the model having less predictive power and being difficult to interpret. Therefore, variable selection plays an important role in high-dimensional statistical modeling.

There has been considerable studies on variable selection for longitudinal data. The challenge of analyzing longitudinal data is that, it is difficult to specify full likelihood functions, especially for correlated non-Gaussian data. If the distributions are not available, traditional likelihood-based model selection criteria, such as Akaike's Information Criterion [3] and Bayes's Information Criterion [4], cannot be employed directly for model selection. To solve this problem, Pan [5] proposed a modification of the Akaike Information criterion (AIC), named QIC, which was obtained by replacing likelihood with quasi-likelihood under the strong assumption of work independence. Nevertheless, ignoring the intrinsic correlation may lead to inefficient estimation and poor prediction capability. Fu [6] investigated the bridge penalty model for the estimating equations in general. Cantoni et al. [7] introduced a generalized version of Mallows' C_P (GC_p), which can be applied to the parametric and nonparametric model. Wang and Qu [8] discussed a BIC-type model selection criterion with the help of quadratic inference function, and their procedure could select the most simplest correct model with probability approaching 1. Yang et al. [9] considered a frequentist model average estimator for longitudinal data based on generalized estimating equations. Chen et al. [10] combined adaptive sampling with sequential method for modeling correlated response data. Most work in the literature, however, focus on the case where the number of covariates is fixed. Therefore, it is necessary to develop new statistical methods and variable selection theories for high-dimensional longitudinal data.

The aim of this paper is to study the problem of variable selection for longitudinal ordinal data with a diverging number of covariates. It should be noted that longitudinal ordinal data have been studied well by many authors. We here only mention Williamson et al. [11], Lipsitz et al. [12], Lin and Chen [13], who all investigated the asymptotic theory by means of GEE method. The GEE approach has an advantage of producing a consistent estimator even if the working correlation structure is incorrectly specified (see, for example, [14]). However, to the best of my knowledge, there is no existing literature on the variable selection for longitudinal ordinal data with high-dimensional covariates. Motivated by the idea of Wang et al. [1], we use the penalized GEE with a nonconvex penalty function to do selection for such data. Similar to GEE, the penalized GEE procedure only requires the first two marginal moments and a working correlation matrix to be specified. It does not require the full joint likelihood for high-dimensional correlated data; this is particularly advantageous for modeling related discrete response data.

Under the assumption of a sparse marginal model, we show that the penalized GEE technique can correctly select the zero coefficients with probability converging to 1 and the estimator of the non-zero coefficients can perform as well as if the true model is known in advance. That is, the resulting estimator enjoys the Oracle properties proposed by Fan and Li [15].

The remainder of this paper is organized as follows. In Section 2, we give a brief overview of the GEE approach and introduce penalized GEE for longitudinal ordinal data. The asymptotic properties of high-dimensional penalized GEE are established in Section 3. Detailed proofs of the main results are provided in Section 4. To verify the main results, the Monte Carlo simulations are conducted in Section 5.

Notation: Throughout this paper, superscript "T" always denotes the transpose of a vector or a matrix. $\mathrm{Tr}(\boldsymbol{A})$ is the trace of a matrix $\boldsymbol{A}$. Moreover, we use $||\cdot||$ to denote the Frobenius norm. For a matrix $\boldsymbol{D}$, $||\boldsymbol{D}|| = ||\boldsymbol{D}||_F = [\mathrm{Tr}(\boldsymbol{D}\boldsymbol{D}^T)]^{1/2}$. Particularly, for a vector $\boldsymbol{x}$, $||\boldsymbol{x}|| = ||\boldsymbol{x}||_F = ||\boldsymbol{x}||_2 = (x_1^2 + \cdots + x_n^2)^{1/2}$. For any vector $\boldsymbol{v}$, $\mathrm{diag}(\boldsymbol{v})$ represents a diagonal matrix whose diagonal elements are the elements of $\boldsymbol{v}$.

2. Preliminaries

2.1. Generalized estimating equations for longitudinal ordinal data

Let Y_{ij} be longitudinal ordinal responses with $q+1$ categories, and $\boldsymbol{x}_{ij}$ denote $\bar{p}_n$-dimensional covariate vectors for subject i at occasion j for $i = 1, ..., n$ and $j = 1, ..., m_i$. We assume that the observations on different subjects are independent and the observations on the same subject are correlated. For simplicity, we assume equal occasions, $m_i = m$. Define $\boldsymbol{y}_{ij}$ as a vector of q-dimensional variables, where $\boldsymbol{y}_{ij} = (y_{ij1}, \cdots, y_{ijq})^T$ with $y_{ijr} = 1$ if response $Y_{ij} = r$ and 0 else. Moreover, let $\boldsymbol{\pi}_{ij}$ and η_{ijr} be the vector of marginal probabilities and marginal cumulative probabilities, respectively, where $\boldsymbol{\pi}_{ij} = (\pi_{ij1}, \cdots, \pi_{ijq})^T$ with $\pi_{ijr} = P(Y_{ij} = r|\boldsymbol{x}_{ij})$, and $\eta_{ijr} = P(Y_{ij} \le r|\boldsymbol{x}_{ij}) = \sum_{l=1}^{r} \pi_{ijl}$.

We consider the following cumulative logit model (see, for example, [13]):

$$\operatorname{logit}(\eta_{ijr}) = \log \frac{\eta_{ijr}}{1-\eta_{ijr}} = \phi_r + \boldsymbol{x}_{ij}^T \boldsymbol{\delta} = \gamma_{ijr}, \quad r = 1, ..., q, \tag{2.1}$$

where the intercepts $\phi_1, \cdots, \phi_q$ satisfy $\phi_1 \le \cdots \le \phi_q$, $\boldsymbol{\delta}$ is the vector of regression coefficients, and γ_{ijr} is the rth element of a q-dimensional linear predictor $\boldsymbol{\gamma}_{ij} = (\gamma_{ij1}, \cdots, \gamma_{ijq})^T$. Clearly, $\pi_{ij1} = \eta_{ij1}$ and

$$\pi_{ijr} = \frac{\exp(\gamma_{ijr})}{1+\exp(\gamma_{ijr})} - \frac{\exp(\gamma_{ij,r-1})}{1+\exp(\gamma_{ij,r-1})}, \quad r = 2, \cdots, q. \tag{2.2}$$

Then, the linear predictor $\boldsymbol{\gamma}_{ij}$ can be rewritten as $\boldsymbol{\gamma}_{ij} = \boldsymbol{X}_{ij}^T \boldsymbol{\beta}_n$, where $\boldsymbol{\beta}_n = (\phi_1, \cdots, \phi_q, \boldsymbol{\delta}^T)^T$ is the $p_n \times 1$ vector of parameters, and

$$\boldsymbol{X}_{ij}^T = \begin{bmatrix} 1 & & & \boldsymbol{x}_{ij}^T \\ & \ddots & & \vdots \\ & & 1 & \boldsymbol{x}_{ij}^T \end{bmatrix}_{q\times(q+\bar{p}_n)}.$$

is the $q \times p_n$ design matrix. It is clear that there exists a q-dimensional link function $\boldsymbol{g}$ such that $\boldsymbol{g}(\boldsymbol{\pi}_{ij}) = \boldsymbol{X}_{ij}^T \boldsymbol{\beta}_n$ (see, for example, [16], p.73–84).

Let the responses, the marginal probabilities and the design matrix for cluster i be denoted by $\boldsymbol{Y}_i = (\boldsymbol{y}_{i1}^T, \cdots, \boldsymbol{y}_{im}^T)_{qm\times1}^T$, $\boldsymbol{\pi}_i = (\boldsymbol{\pi}_{i1}^T, \cdots, \boldsymbol{\pi}_{im}^T)_{qm\times1}^T$ and $\boldsymbol{X}_i = (\boldsymbol{X}_{i1}, \cdots, \boldsymbol{X}_{im})_{qm\times p_n}^T$, respectively. The generalized estimating equations (see, [14]) is defined as follows:

$$\sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\pi}_i^T}{\partial \boldsymbol{\beta}_n}\right) \boldsymbol{V}_i^{-1}(\boldsymbol{\beta}_n, \tau)(\boldsymbol{Y}_i - \boldsymbol{\pi}_i(\boldsymbol{\beta}_n)) = 0, \tag{2.3}$$

where $\boldsymbol{V}_i(\boldsymbol{\beta}_n, \tau) \approx \operatorname{Cov}(Y_i)$ is a working covariance matrix of Y_i.

However, since the true $\boldsymbol{V}_i(\boldsymbol{\beta}_n, \tau)$ is generally difficult to obtain in practice, the working covariance matrix is usually specified by virtue of a working correlation matrix $\boldsymbol{R}(\tau)$: $\boldsymbol{V}_i(\boldsymbol{\beta}_n, \tau) = \boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n)\boldsymbol{R}(\tau)\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_n)$, where

$$\boldsymbol{A}_i = \operatorname{diag}[\{\pi_{i11}(1-\pi_{i11})\}, \cdots, \{\pi_{imq}(1-\pi_{imq})\}],$$

also

$$\boldsymbol{A}_i^{1/2} = \operatorname{diag}[\{\pi_{i11}(1-\pi_{i11})\}^{1/2}, \cdots, \{\pi_{imq}(1-\pi_{imq})\}^{1/2}],$$

and $\boldsymbol{R}(\tau)$ is an optional working correlation matrix, which may include a nuisance parameter (or parameter vector) τ. Indeed, the estimator of $\boldsymbol{R}(\tau)$ could be obtained using a local odds ratios GEE method [17]. It should be mentioned that, if $\boldsymbol{R}(\tau)$ is equal to the true correlation matrix $\boldsymbol{R}_0$, then $\boldsymbol{V}_i(\boldsymbol{\beta}_n, \tau) = \mathrm{Cov}(\boldsymbol{Y}_i)$ at the true parameter $\boldsymbol{\beta}_{n0}$.

Let $\hat{\boldsymbol{R}}$ represent the estimated working correlation matrix. It follows from (2.3) that the generalized estimating equations of longitudinal ordinal data have the following form:

$$\boldsymbol{S}_n(\boldsymbol{\beta}_n) = \sum_{i=1}^{n} \boldsymbol{X}_i^T \boldsymbol{H}_i(\boldsymbol{\beta}_n) \boldsymbol{A}_i^{-1/2}(\boldsymbol{\beta}_n) \hat{\boldsymbol{R}}^{-1} \boldsymbol{A}_i^{-1/2}(\boldsymbol{\beta}_n)(\boldsymbol{Y}_i - \boldsymbol{\pi}_i(\boldsymbol{\beta}_n)) = \boldsymbol{0}, \tag{2.4}$$

where $\boldsymbol{H}_i(\boldsymbol{\beta}_n) = \partial \boldsymbol{\pi}_i^T / \partial \boldsymbol{\gamma}_i = \mathrm{diag}[\boldsymbol{H}_{i1}(\boldsymbol{\beta}_n), \cdots, \boldsymbol{H}_{im}(\boldsymbol{\beta}_n)]$ with $\boldsymbol{\gamma}_i = (\boldsymbol{\gamma}_{i1}^T, \cdots, \boldsymbol{\gamma}_{im}^T)^T$ and the diagonal block matrix $\boldsymbol{H}_{ij}(\boldsymbol{\beta}_n) = \partial \boldsymbol{\pi}_{ij}^T / \partial \boldsymbol{\gamma}_{ij}$. For more details, the reader could refers to Williamson et al. [11].

Remark 1. In this paper, we suppose q and m are fixed, but the dimension p_n of the covariates is infinity. It is easy to see that $\bar{p}_n$ is also infinity from the definition of $\boldsymbol{\beta}_n$.

Remark 2. Note that, the cumulative logit link function is adopted for analyzing longitudinal ordinal data. In fact, the adjacent-categories logit link function also can be employed to conduct corresponding analysis (see, [17]).

Remark 3. When we study the asymptotic properties of longitudinal ordinal data, as described in Section 2.1, we need to adopt multidimensional dummy variables to represent the categories of response variables $\boldsymbol{y}_{ij}$. Since the response variables $\boldsymbol{y}_{ij}$ are multidimensional, it is more difficult to prove the asymptotic properties than Wang et al. (2012).

2.2. *Penalized generalized estimating equations*

In this subsection we turn our attention to the penalized GEE for simultaneous estimation and variable selection. Consider the following penalized GEE model:

$$\boldsymbol{U}_n(\boldsymbol{\beta}_n) = \boldsymbol{S}_n(\boldsymbol{\beta}_n) - \boldsymbol{q}_{\zeta_n}(|\boldsymbol{\beta}_n|)\mathrm{sign}(\boldsymbol{\beta}_n), \tag{2.5}$$

where

$$\boldsymbol{S}_n(\boldsymbol{\beta}_n) = n^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i^T \boldsymbol{H}_i(\boldsymbol{\beta}_n) \boldsymbol{A}_i^{-1/2}(\boldsymbol{\beta}_n) \hat{\boldsymbol{R}}^{-1} \boldsymbol{A}_i^{-1/2}(\boldsymbol{\beta}_n)(\boldsymbol{Y}_i - \boldsymbol{\pi}_i(\boldsymbol{\beta}_n)) = \boldsymbol{0} \tag{2.6}$$

are the GEE, $\boldsymbol{q}_{\zeta_n}(|\boldsymbol{\beta}_n|) = (q_{\zeta_n}(|\beta_{n1}|), \cdots, q_{\zeta_n}(|\beta_{np_n}|))^T$ is a p_n-dimensional vector of penalty functions, and $\mathrm{sign}(\boldsymbol{\beta}_n) = (\mathrm{sign}(\beta_{n1}), \cdots, \mathrm{sign}(\beta_{np_n}))^T$ with $\mathrm{sign}(t) = I(t > 0) - I(t < 0)$. Moreover, $\boldsymbol{q}_{\zeta_n}(|\boldsymbol{\beta}_n|)\mathrm{sign}(\boldsymbol{\beta}_n)$ denotes the component-wise product. The thresholding parameter ζ_n controls the size of shrinkage.

Because $\boldsymbol{U}_n(\boldsymbol{\beta}_n)$ contains discontinuous points, the exact solution of $\boldsymbol{U}_n(\boldsymbol{\beta}_n) = \boldsymbol{0}$ may not exist. Similar to the ideas proposed by Wang et al. [1], we define the penalized GEE estimator $\hat{\boldsymbol{\beta}}_n$ as an approximate solution: $\boldsymbol{U}_n(\hat{\boldsymbol{\beta}}_n) = o(a_n)$ for a sequence $a_n \to 0$. The rate of a_n will be made clarified in Theorem 1 below.

Different penalty functions can be chose in nature; however, in this article, we consider the nonconvex smoothly clipped absolute deviation (SCAD) penalty, which is given by

$$q_{\zeta_n}(\theta) = \zeta_n \{ I(\theta \le \zeta_n) + \frac{(a\zeta_n - \theta)_+}{(a-1)\zeta_n} I(\theta > \zeta_n) \}$$

for $\theta \geq 0$ and some $a > 2$. According to Fan and Li [15], the SCAD penalty simultaneously possesses three attractive properties of variable selection: unbiasedness, sparsity, and continuity. Compared to other penalty functions, the LASSO penalty (L_1 penalty) does not satisfy the unbiasedness condition, the L_q penalty with $q > 1$ does not satisfy the sparsity condition, and the L_q penalty with $0 \leq q < 1$ does not satisfy the continuity condition. What's more, following the recommendation of [15], we choose $a = 3.7$.

It is not hard to see that the penalty function $q_{\zeta_n}(|\beta_{nj}|)$ is zero for a large value of $|\beta_{nj}|$ and is comparatively large for a small value of $|\beta_{nj}|$. Consequently, the generalized estimating function $S_{nj}(\boldsymbol{\beta}_n)$, the jth component of $\boldsymbol{S}_n(\boldsymbol{\beta}_n)$, is not penalized if β_{nj} is large in magnitude; however, if β_{nj} is close to 0, the penalty of $q_{\zeta_n}(|\beta_{nj}|)$ increases and forces its estimate to decrease to zero. Once the estimated coefficient is reduced to zero, it is eliminated from the final selected model; thus, this method significantly reduces the computational burden.

3. Asymptotic theory for high-dimensional penalized GEE

In this section, we establish the asymptotic theory of the penalized GEE estimator with a diverging number of parameters. We denote the true value $\boldsymbol{\beta}_{n0}$ by $\boldsymbol{\beta}_{n0} = (\boldsymbol{\beta}_{n10}^T, \boldsymbol{\beta}_{n20}^T)^T$. Without loss of generality, it is assumed that $\boldsymbol{\beta}_{n20} = \boldsymbol{0}$ and that the elements of $\boldsymbol{\beta}_{n10}$ are all nonzero. Besides, the covariates matrix is divided into $\boldsymbol{X}_i = (\boldsymbol{X}_{i1}, \boldsymbol{X}_{i2})$ accordingly. We also denote the dimension of $\boldsymbol{\beta}_{n10}$ by s_n, where s_n may be fixed or grow with n.

Before we present the main result of the theorem, we first state some regularity conditions as follows:

(C1) $\|\boldsymbol{X}_{ij}\|$, $1 \leq i \leq n$, $1 \leq j \leq m$, are uniformly bounded;

(C2) The unknown parameter $\boldsymbol{\beta}_n$ belongs to a compact subset $\mathbb{B} \subseteq R^{p_n}$, the true parameter value $\boldsymbol{\beta}_{n0}$ lies in the interior of $\mathbb{B}$; further, recall that, $\boldsymbol{\pi}_{ij}$ be the vector of marginal probabilities, where $\boldsymbol{\pi}_{ij} = (\pi_{ij1}, \cdots, \pi_{ijq})^T$ with $\pi_{ijr} = P(Y_{ij} = r|\boldsymbol{x}_{ij})$, and there exist two positive constants b_1, b_2, such that $0 < b_1 \leq \pi_{ijr} \leq b_2 < 1$, for $i = 1, \cdots, n$, $j = 1, \cdots, m$, $r = 1, \cdots, q$;

(C3) Let $B_n = \{\boldsymbol{\beta}_n : \|\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}\| \leq \Delta\sqrt{p_n/n}\}$, then $\|\boldsymbol{\pi}_{ij}^{[1]}(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_n)\|$ is uniformly bounded away from zero and $+\infty$ on B_n; $\|\pi_{ij}^{[2]}(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_n)\|$ and $\|\boldsymbol{\pi}_{ij}^{[3]}(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_n)\|$ are both uniformly bounded by a finite positive constant M_1 on B_n, where $\boldsymbol{\pi}_{ij}^{[h]}(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_n)$ is the h-order partial derivative of $\boldsymbol{\pi}_{ij}$ with respect to $\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_n$, for $i = 1, \cdots, n$, $j = 1, \cdots, m$, $h = 1, 2, 3$;

(C4) The true correlation matrix $\boldsymbol{R}_0$ has eigenvalues bounded away from zero and $+\infty$; $\hat{\boldsymbol{R}}$ satisfies $\|\hat{\boldsymbol{R}}^{-1} - \bar{\boldsymbol{R}}^{-1}\| = O_p(\sqrt{p_n/n})$, where $\bar{\boldsymbol{R}}$ is a constant positive definite matrix with eigenvalues bounded away from zero and $+\infty$; $\bar{\boldsymbol{R}}$ is not required to be the true correlation matrix $\boldsymbol{R}_0$;

(C5) There exist two positive constants, c_1 and c_2, such that

$$c_1 \leq \lambda_{min}(n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_i^T\boldsymbol{X}_i) \leq \lambda_{max}(n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_i^T\boldsymbol{X}_i) \leq c_2;$$

where λ_{min}(resp. λ_{max}) denotes the minimum (resp. maximum) eigenvalue of the matrix;

(C6) Assuming $\min_{1\leq j\leq s_n}|\beta_{n0j}|/\zeta_n \to 0$ as $n \to \infty$ and $s_n^3 n^{-1} = o(1)$, $\zeta_n \to 0$, $s_n(\log n)^2 = o(n\zeta_n^2)$, $\log p_n(\log n)^2 = o(n\zeta_n^2)$, $s_n^2(\log n)^4 = o(n\zeta_n^2)$ and $p_n s_n^4(\log n)^6 = o(n^2\zeta_n^2)$.

Remark 4. Conditions (C2) and (C4)–(C5) are the same as the assumptions of Wang [2, p.394–395]. In addition, conditions (C1) and (C3) are also the same as the hypothesis of Wang et al. [1], however, conditions (C6) is similar to that of (A7) in Wang et al. [1] with a slightly difference.

Remark 5. [18, p.50] If we suppose $\boldsymbol{S} = (s_{ij})_{m\times n}$ and $\boldsymbol{T} = (t_{kl})_{p\times q}$ are both matrices, then

$$\frac{\partial \boldsymbol{S}}{\partial \boldsymbol{T}} = \begin{pmatrix} \frac{\partial s_{11}}{\partial t_{11}} & \frac{\partial s_{11}}{\partial t_{12}} & \cdots & \frac{\partial s_{11}}{\partial t_{pq}} \\ \frac{\partial s_{12}}{\partial t_{11}} & \frac{\partial s_{12}}{\partial t_{12}} & \cdots & \frac{\partial s_{12}}{\partial t_{pq}} \\ \vdots & \vdots & & \vdots \\ \frac{\partial s_{mn}}{\partial t_{11}} & \frac{\partial s_{mn}}{\partial t_{12}} & \cdots & \frac{\partial s_{mn}}{\partial t_{pq}} \end{pmatrix}_{mn\times pq}.$$

Note that, $\boldsymbol{\pi}_{ij}^{[h]}(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_n)$ of condition (C3) are some matrices with different dimensions for $h = 1, 2, 3$.

Now, let us present the main result in this study:

Theorem 1. If assumptions (C1)–(C6) hold, then there exists an approximate penalized GEE solution $\hat{\boldsymbol{\beta}}_n = (\hat{\boldsymbol{\beta}}_{n1}^T, \hat{\boldsymbol{\beta}}_{n2}^T)^T$ such that

(1)

$$P(|U_{nj}(\hat{\boldsymbol{\beta}}_n)| = 0, \quad j = 1\cdots, s_n) \to 1 \tag{3.1}$$

$$P(|U_{nj}(\hat{\boldsymbol{\beta}}_n)| \leq \frac{\zeta_n}{\log n}, \quad j = s_n + 1\cdots, p_n) \to 1 \tag{3.2}$$

(2) $P(\hat{\boldsymbol{\beta}}_{n2} = \boldsymbol{0}) \to 1$,
(3) $\forall \boldsymbol{\alpha}_n \in R^{s_n}$ such that $||\boldsymbol{\alpha}_n|| = 1$, we have

$$\boldsymbol{\alpha}_n^T \bar{\boldsymbol{M}}_{n1}^{-1/2}(\boldsymbol{\beta}_{n0})\bar{\boldsymbol{H}}_{n1}(\boldsymbol{\beta}_{n0})(\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}) \xrightarrow{d} N(0, 1),$$

where

$$\begin{aligned} \bar{\boldsymbol{M}}_{n1}(\boldsymbol{\beta}_{n0}) &= \sum_{i=1}^n \boldsymbol{X}_{i1}^T\boldsymbol{H}_i(\boldsymbol{\beta}_{n0})\boldsymbol{A}_i^{-1/2}(\boldsymbol{\beta}_{n0})\bar{\boldsymbol{R}}^{-1}\boldsymbol{R}_0\bar{\boldsymbol{R}}^{-1}\boldsymbol{A}_i^{-1/2}(\boldsymbol{\beta}_{n0})\boldsymbol{H}_i(\boldsymbol{\beta}_{n0})\boldsymbol{X}_{i1}. \\ \bar{\boldsymbol{H}}_{n1}(\boldsymbol{\beta}_{n0}) &= \sum_{i=1}^n \boldsymbol{X}_{i1}^T\boldsymbol{H}_i(\boldsymbol{\beta}_{n0})\boldsymbol{A}_i^{-1/2}(\boldsymbol{\beta}_{n0})\bar{\boldsymbol{R}}^{-1}\boldsymbol{A}_i^{-1/2}(\boldsymbol{\beta}_{n0})\boldsymbol{H}_i^T(\boldsymbol{\beta}_{n0})\boldsymbol{X}_{i1}. \end{aligned}$$

Remark 6. Properties (2) and (3) in Theorem 1 are usually called the oracle property of variable selection. Namely, when the true parameters have some zero coefficients, they are estimated as 0 with probability approaching to one, and the nonzero coefficients are estimated efficiently as if the correct submodel is known. Further, property (1) shows more clear and accurate description for the approximate solution of the penalized GEE. In (3.2), we let $a_n = \frac{\zeta_n}{\log n}$, in nature, another sequence $a_n \to 0$ could be considered.

4. Proof of main results

Throughout the proof, let C be any positive constant independent of n whose value may change from one expression to another. In addition, $\boldsymbol{e}_k$ and $\boldsymbol{o}_k$ denote unit vector of length p_n and qm whose kth entry is 1 and all other entries are 0, respectively.

In order to prove the asymptotic properties of penalized GEE estimator, the essential idea is to approximate $\boldsymbol{S}_n(\boldsymbol{\beta}_n)$ by $\bar{\boldsymbol{S}}_n(\boldsymbol{\beta}_n)$, whose moments are easier to evaluate, where

$$\bar{\boldsymbol{S}}_n(\boldsymbol{\beta}_n) = n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_i^T\boldsymbol{H}_i(\boldsymbol{\beta}_n)\boldsymbol{A}_i^{-1/2}(\boldsymbol{\beta}_n)\bar{\boldsymbol{R}}^{-1}\boldsymbol{A}_i^{-1/2}(\boldsymbol{\beta}_n)(\boldsymbol{Y}_i - \boldsymbol{\pi}_i(\boldsymbol{\beta}_n)).$$

We write $\bar{\boldsymbol{S}}_n(\boldsymbol{\beta}_n) = (\bar{\boldsymbol{S}}_{n1}(\boldsymbol{\beta}_n), \cdots, \bar{\boldsymbol{S}}_{np_n}(\boldsymbol{\beta}_n))^T$, where $\bar{\boldsymbol{S}}_{nk}(\boldsymbol{\beta}_n) = \boldsymbol{e}_k^T\bar{\boldsymbol{S}}_n(\boldsymbol{\beta}_n)$.

The approach we adopt here is based on some ideas in Wang et al. [1]. As a preparation, we first present the following lemmas.

Lemma 1.

$$\frac{\partial\bar{\boldsymbol{S}}_{nk}(\boldsymbol{\beta}_n)}{\partial\boldsymbol{\beta}_n^T} = \bar{\boldsymbol{G}}_{nk}(\boldsymbol{\beta}_n) + \bar{\boldsymbol{B}}_{nk}(\boldsymbol{\beta}_n) + \bar{\boldsymbol{L}}_{nk}(\boldsymbol{\beta}_n) + \bar{\boldsymbol{T}}_{nk}(\boldsymbol{\beta}_n), \tag{4.1}$$

where

$$\begin{aligned}
\bar{\boldsymbol{G}}_{nk}(\boldsymbol{\beta}_n) &= -n^{-1}\sum_{i=1}^{n}\boldsymbol{e}_k^T\boldsymbol{X}_i^T\boldsymbol{H}_i(\boldsymbol{\beta}_n)\boldsymbol{A}_i^{-1/2}(\boldsymbol{\beta}_n)\bar{\boldsymbol{R}}^{-1}\boldsymbol{A}_i^{-1/2}(\boldsymbol{\beta}_n)\boldsymbol{H}_i^T(\boldsymbol{\beta}_n)\boldsymbol{X}_i,\\
\bar{\boldsymbol{B}}_{nk}(\boldsymbol{\beta}_n) &= n^{-1}\sum_{i=1}^{n}\sum_{j_1=1}^{qm}\sum_{j_2=1}^{qm}\boldsymbol{e}_k^T\boldsymbol{X}_i^T\boldsymbol{H}_i(\boldsymbol{\beta}_n)\boldsymbol{A}_i^{-1/2}(\boldsymbol{\beta}_n)\bar{\boldsymbol{R}}^{-1}\boldsymbol{o}_{j_1}\boldsymbol{o}_{j_2}^T(\boldsymbol{Y}_i - \boldsymbol{\pi}_i(\boldsymbol{\beta}_n))\boldsymbol{\kappa}_{i,j_1j_2}^T(\boldsymbol{\beta}_n)\boldsymbol{X}_i,\\
\bar{\boldsymbol{L}}_{nk}(\boldsymbol{\beta}_n) &= n^{-1}\sum_{i=1}^{n}\sum_{j_1=1}^{qm}\sum_{j_2=1}^{qm}\boldsymbol{e}_k^T\boldsymbol{X}_i^T\boldsymbol{H}_i(\boldsymbol{\beta}_n)\boldsymbol{o}_{j_1}\boldsymbol{o}_{j_2}^T\bar{\boldsymbol{R}}^{-1}\boldsymbol{A}_i^{-1/2}(\boldsymbol{\beta}_n)(\boldsymbol{Y}_i - \boldsymbol{\pi}_i(\boldsymbol{\beta}_n))\boldsymbol{\kappa}_{i,j_1j_2}^T(\boldsymbol{\beta}_n)\boldsymbol{X}_i,\\
\bar{\boldsymbol{T}}_{nk}(\boldsymbol{\beta}_n) &= n^{-1}\sum_{i=1}^{n}\sum_{j_1=1}^{qm}\sum_{j_2=1}^{qm}\boldsymbol{e}_k^T\boldsymbol{X}_i^T\boldsymbol{o}_{j_1}\boldsymbol{o}_{j_2}^T\boldsymbol{A}_i^{-1/2}(\boldsymbol{\beta}_n)\bar{\boldsymbol{R}}^{-1}\boldsymbol{A}_i^{-1/2}(\boldsymbol{\beta}_n)(\boldsymbol{Y}_i - \boldsymbol{\pi}_i(\boldsymbol{\beta}_n))\boldsymbol{\nu}_{i,j_1j_2}^T(\boldsymbol{\beta}_n)\boldsymbol{X}_i,
\end{aligned}$$

with

$$\boldsymbol{\kappa}_{i,j_1j_2}(\boldsymbol{\beta}_n) = \frac{\partial[\boldsymbol{o}_{j_1}^T\boldsymbol{A}_i^{-1/2}(\boldsymbol{\beta}_n)\boldsymbol{o}_{j_2}]}{\partial\boldsymbol{X}_i\boldsymbol{\beta}_n}, \quad \boldsymbol{\nu}_{i,j_1j_2}(\boldsymbol{\beta}_n) = \frac{\partial[\boldsymbol{o}_{j_1}^T\boldsymbol{H}_i(\boldsymbol{\beta}_n)\boldsymbol{o}_{j_2}]}{\partial\boldsymbol{X}_i\boldsymbol{\beta}_n}.$$

Proof. The decomposition is analogous to that of Lemma 4.2 in Chen and Yin [20] with some modifications, and is thus omitted.

Remark 7. It is not hard to see that $\boldsymbol{\kappa}_{i,j_1j_2}(\boldsymbol{\beta}_n)$ and $\boldsymbol{\nu}_{i,j_1j_2}(\boldsymbol{\beta}_n)$ are the row vectors, respectively. Moreover, it follows from condition (C3) that $\|\boldsymbol{\kappa}_{i,j_1j_2}(\boldsymbol{\beta}_n)\| = O(1)$ and $\|\boldsymbol{\nu}_{i,j_1j_2}(\boldsymbol{\beta}_n)\| = O(1)$.

Lemma 2. (Bernstein's inequality) Let $Y_1, \cdots, Y_n$ be independent random variables with mean zero such that

$$\mathrm{E}|Y_i|^l \le l!M^{l-2}B_i/2,$$

for every $l \ge 2$, all i, and some positive constants M and B_i. Then

$$\mathrm{P}\left(|Y_1 + \cdots + Y_n| > v\right) \le 2\exp\left(-\frac{1}{2}\frac{v^2}{B + Mv}\right),$$

for $B \geq B_1 + \cdots + B_n$.

Proof. For a detailed proof, the reader could refer to the Lemma 2.2.11 of [21].

Proof of Theorem 1. Let $\hat{\boldsymbol{\beta}}_n = (\hat{\boldsymbol{\beta}}_{n1}^T, \mathbf{0}^T)^T$ be the oracle estimator. We shall prove that $\hat{\boldsymbol{\beta}}_n$ satisfies properties (1)–(3) of Theorem 1. It is easy to see that properties (2) and (3) follow by virtue of the definition of $\hat{\boldsymbol{\beta}}_n$ and the results in [20]. Next, we show that $\hat{\boldsymbol{\beta}}_n$ satisfies (3.1) and (3.2).

Proof of (3.1). From the definition of $\hat{\boldsymbol{\beta}}_n$, we can obtain that $S_{nj}(\hat{\boldsymbol{\beta}}_n) = 0,\ j = 1, \cdots, s_n$. It is sufficient to prove that $\mathrm{P}(|\beta_{nj}| \geq a\zeta_n,\ j = 1, \cdots, s_n) \to 1$, as this implies the penalty function to be zero with probability approaching one. It is clear that

$$\min_{1\leq j\leq s_n} |\hat{\beta}_{nj}| \geq \min_{1\leq j\leq s_n} |\beta_{n0j}| - \max_{1\leq j\leq s_n} |\beta_{n0j} - \hat{\beta}_{nj}| \geq \min_{1\leq j\leq s_n} |\beta_{n0j}| - \|\boldsymbol{\beta}_{n10} - \hat{\boldsymbol{\beta}}_{n10}\|.$$

Owing to [20], we have that

$$\|\boldsymbol{\beta}_{n10} - \hat{\boldsymbol{\beta}}_{n10}\| = O_p(\sqrt{s_n/n}). \tag{4.2}$$

Combining $\min_{1\leq j\leq s_n} |\beta_{n0j}|/\zeta_n \to \infty$ with $\|\boldsymbol{\beta}_{n10} - \hat{\boldsymbol{\beta}}_{n10}\| = o(\zeta_n)$, we obtain

$$\mathrm{P}\left(\min_{1\leq j\leq s_n} |\beta_{n0j}| - \|\boldsymbol{\beta}_{n10} - \hat{\boldsymbol{\beta}}_{n10}\| \geq a\zeta_n\right) = \mathrm{P}\left(\|\boldsymbol{\beta}_{n10} - \hat{\boldsymbol{\beta}}_{n10}\| \leq \min_{1\leq j\leq s_n} |\beta_{n0j}| - a\zeta_n\right) \to 1.$$

The proof of (3.1) is therefore complete.

Proof of (3.2). Using the definition of $\hat{\boldsymbol{\beta}}_n$, we can derive that $q_{\zeta_n}(\hat{\beta}_{nk}) \cdot \mathrm{sign}(\hat{\beta}_{nk}) = 0$, for $k = s_n + 1, \cdots, p_n$. To prove (3.2), it suffices to show that

$$\mathrm{P}\left(\max_{s_n+1\leq k\leq p_n} |S_{nk}(\hat{\boldsymbol{\beta}}_n)| \leq \frac{\zeta_n}{\log n}\right) \to 1, \tag{4.3}$$

which is implied by

$$\mathrm{P}\left(\max_{s_n+1\leq k\leq p_n} |S_{nk}(\hat{\boldsymbol{\beta}}_n) - \bar{S}_{nk}(\hat{\boldsymbol{\beta}}_n)| \geq \frac{\zeta_n}{2\log n}\right) \to 0, \tag{4.4}$$

$$\mathrm{P}\left(\max_{s_n+1\leq k\leq p_n} |\bar{S}_{nk}(\hat{\boldsymbol{\beta}}_n)| \geq \frac{\zeta_n}{2\log n}\right) \to 0. \tag{4.5}$$

For (4.4), it follows from conditions (C1)–(C4) and (C6) that

$$\begin{aligned}
&\mathrm{P}\left(\max_{s_n+1\leq k\leq p_n} n^{-1} \sum_{i=1}^{n} \boldsymbol{e}_k^T \boldsymbol{X}_i^T \boldsymbol{H}_i(\hat{\boldsymbol{\beta}}_n) \boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_n)[\boldsymbol{R}^{-1} - \bar{\boldsymbol{R}}^{-1}] \cdot \boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_n)(\boldsymbol{Y}_i - \boldsymbol{\pi}_i(\hat{\boldsymbol{\beta}}_n)) > \tfrac{\zeta_n}{2\log n}\right)\\
&\leq \mathrm{P}\left(\max_{s_n+1\leq k\leq p_n} n^{-1} \sum_{i=1}^{n} \|\boldsymbol{e}_k^T \boldsymbol{X}_i^T \boldsymbol{H}_i(\hat{\boldsymbol{\beta}}_n) \boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_n)\| \cdot \|\boldsymbol{R}^{-1} - \bar{\boldsymbol{R}}^{-1}\| \cdot \|\boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_n)(\boldsymbol{Y}_i - \boldsymbol{\pi}_i(\hat{\boldsymbol{\beta}}_n))\| > \tfrac{\zeta_n}{2\log n}\right)\\
&\leq \mathrm{P}\left(\|\boldsymbol{R}^{-1} - \bar{\boldsymbol{R}}^{-1}\| \cdot n^{-1} \sum_{i=1}^{n} \left(\max_{s_n+1\leq k\leq p_n} \|\boldsymbol{e}_k^T \boldsymbol{X}_i^T \boldsymbol{H}_i(\hat{\boldsymbol{\beta}}_n) \boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_n)\|\right) \cdot \|\boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_n)(\boldsymbol{Y}_i - \boldsymbol{\pi}_i(\hat{\boldsymbol{\beta}}_n))\| > \tfrac{\zeta_n}{2\log n}\right)\\
&\leq \mathrm{P}\left(n^{-1} \sum_{i=1}^{n} \|\boldsymbol{\epsilon}_i(\hat{\boldsymbol{\beta}}_n)\| > \tfrac{C\zeta_n\sqrt{n}}{2\sqrt{s_n}\log n}\right)\\
&\leq C\frac{n^{-1}\sum_{i=1}^{n} \mathrm{E}\big(\|\boldsymbol{\epsilon}_i(\hat{\boldsymbol{\beta}}_n)\|\big)\sqrt{s_n}\log n}{\zeta_n\sqrt{n}}\\
&= O\big(\tfrac{\sqrt{s_n}\log n}{\zeta_n\sqrt{n}}\big) = o(1),
\end{aligned}$$

where $\boldsymbol{\epsilon}_i(\hat{\boldsymbol{\beta}}_n) = \boldsymbol{Y}_i - \boldsymbol{\pi}_i(\hat{\boldsymbol{\beta}}_n)$.

In what follows, in order to prove (4.5), we utilize the Taylor expansion as follows:

$$\bar{S}_{nk}(\hat{\boldsymbol{\beta}}_n) = \bar{S}_{nk}(\boldsymbol{\beta}_{n0}) + \frac{\partial \bar{S}_{nk}(\boldsymbol{\beta}_{n0})}{\partial \boldsymbol{\beta}_n^T}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}) + (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0})^T \frac{\partial^2 \bar{S}_{nk}(\boldsymbol{\beta}_n^*)}{\partial \boldsymbol{\beta}_n \partial \boldsymbol{\beta}_n^T}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}), \tag{4.6}$$

where $\boldsymbol{\beta}_n^*$ is between $\boldsymbol{\beta}_{n0}$ and $\hat{\boldsymbol{\beta}}_n$. We denote $\frac{\partial \bar{S}_{nk}(\boldsymbol{\beta}_{n0})}{\partial \boldsymbol{\beta}_n^T}$ and $\frac{\partial^2 \bar{S}_{nk}(\boldsymbol{\beta}_n^*)}{\partial \boldsymbol{\beta}_n \partial \boldsymbol{\beta}_n^T}$ by $\bar{\boldsymbol{\Gamma}}_k(\boldsymbol{\beta}_n)$ and $\bar{\boldsymbol{\Lambda}}_k(\boldsymbol{\beta}_n)$, respectively. Furthermore, let $\bar{\boldsymbol{\Gamma}}_{k1}(\boldsymbol{\beta}_n)$ represent the subvector, which is consisted by the first s_n elements of $\bar{\boldsymbol{\Gamma}}_k(\boldsymbol{\beta}_n)$, and let $\bar{\boldsymbol{\Lambda}}_{k1}(\boldsymbol{\beta}_n)$ be the $s_n \times s_n$ submatrix in the upper-left corner of $\bar{\boldsymbol{\Lambda}}_k(\boldsymbol{\beta}_n)$.

Because $\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0} = ((\hat{\boldsymbol{\beta}}_{n1} - \boldsymbol{\beta}_{n10})^T, \boldsymbol{0}^T)^T$, (4.6) can be written as

$$\bar{S}_{nk}(\hat{\boldsymbol{\beta}}_n) = \bar{S}_{nk}(\hat{\boldsymbol{\beta}}_{n0}) + \bar{\boldsymbol{\Gamma}}_{k1}(\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_{n1} - \boldsymbol{\beta}_{n10}) + (\hat{\boldsymbol{\beta}}_{n1} - \boldsymbol{\beta}_{n10})^T \bar{\boldsymbol{\Lambda}}_{k1}(\boldsymbol{\beta}_n^*)(\hat{\boldsymbol{\beta}}_{n1} - \boldsymbol{\beta}_{n10}). \tag{4.7}$$

It is obvious that

$$\begin{aligned} \mathrm{P}\left(\max_{s_n+1\le k\le p_n} |\bar{S}_{nk}(\hat{\boldsymbol{\beta}}_{n0})| > \tfrac{\zeta_n}{2\log n}\right) &\le \mathrm{P}\left(\max_{s_n+1\le k\le p_n} |\bar{S}_{nk}(\hat{\boldsymbol{\beta}}_{n0})| > \tfrac{\zeta_n}{6\log n}\right) + \mathrm{P}\left(\max_{s_n+1\le k\le p_n} |\bar{\boldsymbol{\Gamma}}_{k1}(\boldsymbol{\beta}_{n0})(\hat{\boldsymbol{\beta}}_{n1} - \boldsymbol{\beta}_{n10})| > \tfrac{\zeta_n}{6\log n}\right) \\ &\quad + \mathrm{P}\left(\max_{s_n+1\le k\le p_n} |(\hat{\boldsymbol{\beta}}_{n1} - \boldsymbol{\beta}_{n10})^T \bar{\boldsymbol{\Lambda}}_{k1}(\boldsymbol{\beta}_n^*)(\hat{\boldsymbol{\beta}}_{n1} - \boldsymbol{\beta}_{n10})| > \tfrac{\zeta_n}{6\log n}\right) = K_{n1} + K_{n2} + K_{n3}. \end{aligned}$$

Thus (4.5) is implied by $K_{ni} = o(1)$, $i = 1, 2, 3$.

We first consider K_{n1}. Note that

$$I_{n1} \le \sum_{k=s_n+1}^{p_n} \mathrm{P}\left(|\bar{S}_{nk}(\hat{\boldsymbol{\beta}}_{n0})| > \frac{\zeta_n}{6\log n}\right).$$

We write $\bar{S}_{nk}(\boldsymbol{\beta}_{n0}) = n^{-1} \sum_{i=1}^{n} Z_i$, where $Z_i = \boldsymbol{e}_k^T \boldsymbol{X}_i' \boldsymbol{H}_i(\boldsymbol{\beta}_{n0}) \boldsymbol{A}_i^{-1/2}(\boldsymbol{\beta}_{n0}) \bar{\boldsymbol{R}}^{-1} \boldsymbol{\epsilon}_i(\boldsymbol{\beta}_{n0})$ are independent mean zero random variables. On one hand, by conditions (C1)–(C4), we have

$$|Z_i(\boldsymbol{\beta}_{n0})| < C.$$

On the other hand, $\forall\, l \ge 2$, we get

$$\mathrm{E}|Z_i|^l \le l! M^{l-2} \delta/2,$$

for some constants $M > 0$ and $\delta > 0$. Therefore, the Z_i satisfy the conditions of Bernstein's inequality. Applying Lemma 2, we immediately obtain

$$\begin{aligned} \mathrm{P}\left(|\bar{S}_{nk}(\hat{\boldsymbol{\beta}}_{n0})| > \tfrac{\zeta_n}{6\log n}\right) &\le 2\exp\left[-\tfrac{1}{2}\tfrac{n^2\zeta_n^2/(36(\log n)^2)}{n\delta + M{*}n\zeta_n/(6\log n)}\right] \\ &\le 2\exp\left[-C\tfrac{n\zeta_n^2}{(\log n)^2}\right]. \end{aligned}$$

It is obvious that

$$I_{n1} \le 2\exp\left[\log p_n - C\frac{n\zeta_n^2}{(\log n)^2}\right] = o(1),$$

with the help of $\log p_n = o\left(n\zeta_n^2/(\log n)^2\right)$ by condition (C6). This implies that $K_{n1} = o(1)$.

Next we'll prove that $K_{n2} = o(1)$. It follows from (4.2) and Lemma 1 that

$$\begin{aligned}
K_{n2} &= \mathrm{P}\left(\max_{s_n+1\le k\le p_n} |\bar{\boldsymbol{\Gamma}}_{k1}(\boldsymbol{\beta}_{n0})(\hat{\boldsymbol{\beta}}_{n1}-\boldsymbol{\beta}_{n10})| > \tfrac{\zeta_n}{6\log n}\right)\\
&\le \mathrm{P}\left(\max_{s_n+1\le k\le p_n} |\bar{\boldsymbol{\Gamma}}_{k1}(\boldsymbol{\beta}_{n0})(\hat{\boldsymbol{\beta}}_{n1}-\boldsymbol{\beta}_{n10})| > \tfrac{\zeta_n}{6\log n}, \|\hat{\boldsymbol{\beta}}_{n1}-\boldsymbol{\beta}_{n10}\| \le \sqrt{s_n/n}\log n\right) + \mathrm{P}\left(\|\hat{\boldsymbol{\beta}}_{n1}-\boldsymbol{\beta}_{n10}\| > \sqrt{s_n/n}\log n\right)\\
&\le \mathrm{P}\left(\max_{s_n+1\le k\le p_n} \|\bar{\boldsymbol{\Gamma}}_{k1}(\boldsymbol{\beta}_{n0})\| > \tfrac{\zeta_n\sqrt{n}}{6\sqrt{s_n}(\log n)^2}\right) + o(1)\\
&\le \mathrm{P}\left(\max_{s_n+1\le k\le p_n} \|\bar{\boldsymbol{G}}_{nk1}(\boldsymbol{\beta}_{n0})\| > \tfrac{\zeta_n\sqrt{n}}{24\sqrt{s_n}(\log n)^2}\right) + \mathrm{P}\left(\max_{s_n+1\le k\le p_n} \|\bar{\boldsymbol{B}}_{nk1}(\boldsymbol{\beta}_{n0})\| > \tfrac{\zeta_n\sqrt{n}}{24\sqrt{s_n}(\log n)^2}\right)\\
&\quad +\mathrm{P}\left(\max_{s_n+1\le k\le p_n} \|\bar{\boldsymbol{L}}_{nk1}(\boldsymbol{\beta}_{n0})\| > \tfrac{\zeta_n\sqrt{n}}{24\sqrt{s_n}(\log n)^2}\right) + \mathrm{P}\left(\max_{s_n+1\le k\le p_n} \|\bar{\boldsymbol{T}}_{nk1}(\boldsymbol{\beta}_{n0})\| > \tfrac{\zeta_n\sqrt{n}}{24\sqrt{s_n}(\log n)^2}\right) + o(1)\\
&= K_{n21} + K_{n22} + K_{n23} + +K_{n24} + o(1),
\end{aligned}$$

where $\bar{\boldsymbol{G}}_{nk1} = (\bar{G}_{nk1}, \cdots, \bar{G}_{nks_n})^T$ denotes the subvector of $\bar{\boldsymbol{G}}_{nk}$ which consists its first s_n elements, $\bar{\boldsymbol{B}}_{nk1}$, $\bar{\boldsymbol{L}}_{nk1}$ and $\bar{\boldsymbol{T}}_{nk1}$ are defined similarly. It is clear that $|\bar{G}_{nkj}(\beta_{n0})|$ is uniformly bounded, thus $\max_{s_n+1\le k\le p_n} \mathrm{E}\|\bar{\boldsymbol{G}}_{nk1}(\boldsymbol{\beta}_{n0})\|^2 = \max_{s_n+1\le k\le p_n} \mathrm{E}\left(\sum_{j=1}^{s_n} \bar{G}_{nkj}(\boldsymbol{\beta}_{n0})\right) \le Cs_n$. To evaluate K_{n21}, a combination of conditions (C1)–(C4), $s_n^2(\log n)^4 = o(n\zeta_n^2)$ and Markov's inequality yields that

$$\begin{aligned}
K_{n21} &= \mathrm{P}\left(\max_{s_n+1\le k\le p_n} \|\bar{\boldsymbol{G}}_{nk1}(\boldsymbol{\beta}_{n0})\| > \tfrac{\zeta_n\sqrt{n}}{24\sqrt{s_n}(\log n)^2}\right)\\
&\le \sum_{k=s_n+1}^{p_n} \mathrm{P}\left(\|\bar{\boldsymbol{G}}_{nk1}(\boldsymbol{\beta}_{n0})\| > \tfrac{\zeta_n\sqrt{n}}{24\sqrt{s_n}(\log n)^2}\right)\\
&\le \frac{\mathrm{E}\|\bar{\boldsymbol{G}}_{nk1}(\boldsymbol{\beta}_{n0})\|^2 \cdot 576 s_n(\log n)^4}{n\zeta_n^2}\\
&\le C\frac{s_n^2(\log n)^4}{n\zeta_n^2} = o(1).
\end{aligned}$$

By the same arguments, we can prove that $K_{n22} = o(1)$, $K_{n23} = o(1)$ and $K_{n24} = o(1)$, respectively. We thus have $K_{n2} = o(1)$.

Finally, we verify that $K_{n3} = o(1)$. Applying Markov's inequality, we then get

$$\begin{aligned}
K_{n3} &= \mathrm{P}\left(\max_{s_n+1\le k\le p_n} |(\hat{\boldsymbol{\beta}}_{n1}-\boldsymbol{\beta}_{n10})^T\bar{\boldsymbol{\Lambda}}_{k1}(\boldsymbol{\beta}_n^*)(\hat{\boldsymbol{\beta}}_{n1}-\boldsymbol{\beta}_{n10})| > \tfrac{\zeta_n}{6\log n}\right)\\
&\le \mathrm{P}\left(\max_{s_n+1\le k\le p_n} |(\hat{\boldsymbol{\beta}}_{n1}-\boldsymbol{\beta}_{n10})^T\bar{\boldsymbol{\Lambda}}_{k1}(\boldsymbol{\beta}_n^*)(\hat{\boldsymbol{\beta}}_{n1}-\boldsymbol{\beta}_{n10})| > \tfrac{\zeta_n}{6\log n}, \|\hat{\boldsymbol{\beta}}_{n1}-\boldsymbol{\beta}_{n10}\| \le \sqrt{s_n/n}\log n\right)\\
&\quad +\mathrm{P}\left(\|\hat{\boldsymbol{\beta}}_{n1}-\boldsymbol{\beta}_{n10}\| > \sqrt{s_n/n}\log n\right)\\
&\le \sum_{k=s_n+1}^{p_n} \mathrm{P}\left(\|\bar{\boldsymbol{\Lambda}}_{k1}(\boldsymbol{\beta}_n^*)\| > \tfrac{n\zeta_n}{6s_n(\log n)^3}\right) + o(1)\\
&\le \sum_{k=s_n+1}^{p_n} \frac{36\ \mathrm{E}\left[\|\bar{\boldsymbol{\Lambda}}_{k1}(\boldsymbol{\beta}_n^*)\|^2\right]\cdot s_n^2(\log n)^6}{n^2\zeta_n^2} + o(1).
\end{aligned}$$

By means of conditions (C1)–(C4), we obtain

$$\mathrm{E}\left[\|\boldsymbol{\Lambda}_{k1}(\boldsymbol{\beta}_n^*)\|^2\right] = \mathrm{E}\left[\mathrm{Tr}\left(\boldsymbol{\Lambda}_{k1}(\boldsymbol{\beta}_n^*)\boldsymbol{\Lambda}_{k1}(\boldsymbol{\beta}_n^*)^T\right)\right] = \mathrm{E}\left[\sum_{t=1}^{s_n}\sum_{j=1}^{s_n}\left(\frac{\partial^2\bar{S}_{nk}(\boldsymbol{\beta}_n^*)}{\partial\beta_{nj}\partial\beta_{nt}}\right)^2\right] \le Cs_n^2,$$

which, together with $p_n s_n^4(\log n)^6/(n^2\zeta_n^2) = o(1)$, gives the required result.

Summarizing the above, this proof is completed. □

5. Monte Carlo simulations

Numerical studies are conducted in this section to demonstrate the main results. For the sake of simplicity, we consider the following cumulative logit model for longitudinal ordinal responses with two categories:

$$\log \frac{\pi_{ij}}{1-\pi_{ij}} = \phi_1 + \boldsymbol{x}_{ij}^T\boldsymbol{\delta} = \boldsymbol{X}_{ij}^T\boldsymbol{\beta}_{n0},\ i = 1, \ldots, 400;\ j = 1, ..., 4;$$

where the true parameters $\boldsymbol{\beta}_{n0} = (\phi_1, \boldsymbol{\delta}^T)^T = (0, 0.5, 0.5, 0, 0, ..., 0)^T$ is a 50-dimensional vector of parameters, $\boldsymbol{X}_{ij}^T = (1, \boldsymbol{x}_{ij}^T) = (1, x_{ij,1}, ..., x_{ij,49})$ is the 1×50 design vector. Further, $\boldsymbol{x}_{ij}^T = (x_{ij,1}, ..., x_{ij,49})$ has a multivariate normal distribution with mean zero, marginal variance 1 and an AR-1 correlation matrix with autocorrelation coefficient 0.3. Moreover, given on $\boldsymbol{x}_{ij}$ and Y_{ij} are determined by

$$Y_{ij} = r \quad \Leftrightarrow \quad \phi_{r-1} < \epsilon_{ij} - \boldsymbol{x}_{ij}^T\boldsymbol{\delta} \leq \phi_r,$$

for $r = 1, 2$, $-\infty = \phi_0 \leq \phi_1 \leq \phi_2 = \infty$ and *i.i.d.* latent vector $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \epsilon_{i2}, \epsilon_{i3}, \epsilon_{i4})^T$ from a tetra-variate normal distribution with mean vector $\mathbf{0}$, the marginal variance matrix $\boldsymbol{I}$ and an exchangeable correlation structure with correlation coefficient $\rho = 0.5$. Such correlated ordinal data can be generated from Touloumis [19].

We compare penalized GEE approach with the unpenalized GEE and the oracle GEE (i.e., the GEE with the true marginal regression model is known). To illustrate the influence of intra-cluster correlation on estimation efficiency, we consider three different working correlation structures: independence, exchangeable and first-order autoregressive (AR-1). The modified Newton-Raphson algorithm in page 355 of Wang et al. [1] was adopted to estimate $\boldsymbol{\beta}_n$. It is worth pointing out that, a fourfold cross-validation was used to estimate the tuning parameter ζ_n in the SCAD penalty function. At the end of the iteration, if an estimated coefficient has magnitude below the cut-off value 10^{-3}, it was considered as zero.

Additionally, we used the estimated mean squared error (MSE) to evaluate the estimation accuracy, which is defined by $||\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}||^2$. Moreover, to evaluate model selection performance, we adopt the notation (C, IC) to show variable selection results. C is the average number of non-zero coefficients correctly estimated to be non-zero, and IC denotes the average number of zero coefficients incorrectly estimated to be non-zero. All results were based on 100 replicate simulations.

Table 1 summarize the estimation accuracy and model selection properties of the penalized GEE, the unpenalized GEE and the oracle GEE for three different working correlation matrices. We observed that the performance of the penalized GEE procedure was comparable to that of the oracle GEE, and significantly reduced the MSE of the unpenalized GEE estimator. Using the true correlation structure (exchangeable) in penalized GEE gives the smallest MSE. Furthermore, we observed that the unpenalized GEE generally did not lead to a sparse model. The penalized GEE successfully selects all covariates with non-zero coefficients and contains a fairly small number of IC. Similar results were also observed with exchangeable correlation coefficient $\rho = 0.3$ and $\rho = 0.8$, respectively, which are not reported here.

Table 1. Simulation results ($n = 400$, $p_n = 50$) for GEE, oracle GEE and penalized GEE with three different working correlation matrices (independence, exchangeable, and AR-1).

	MSE	C	IC
GEE.independence	0.2026	2.00	47.36
GEE.exchangeable	0.1653	2.00	47.26
GEE.AR-1	0.1788	2.00	47.25
oracle.independence	0.0166	2.00	0.00
oracle.exchangeable	0.0129	2.00	0.00
oracle.AR-1	0.0137	2.00	0.00
PGEE.independence	0.0718	2.00	0.94
PGEE.exchangeable	0.0441	2.00	1.50
PGEE.AR-1	0.0484	2.00	1.51

Acknowledgments

This work is partly supported by the Natural Science Foundation of China under Grant 61973096 and by GDUPS [Guangdong Province Universities and Colleges Pearl River Scholar Funded Scheme] (2019).

Conflict of interest

The authors declare no conflict of interest.

References

1. L. Wang, J. H. Zhou, A. N. Qu, Penalized generalized estimating equations for high-dimensional longitudinal data analysis, *Biometrics,* **68** (2012), 353–360. http://dx.doi.org/10.1111/j.1541-0420.2011.01678.x
2. L. Wang, GEE analysis of clustered binary data with diverging number of covariates, *Ann. Stat.,* **39** (2011), 389–417. https://doi.org/10.1214/10-AOS846
3. H. Akaike, A new look at the statistical model identification, *IEEE. T. Automat. Contr.* **19** (1974), 716–723. http://dx.doi.org/10.1109/tac.1974.1100705
4. G. Schwarz, Estimating the dimension of a model, *Ann. Stat.,* **6** (1978), 461–464. http://dx.doi.org/10.1214/aos/1176344136
5. W. Pan, Akaike's information criterion in generalized estimating equations, *Biometrics,* **57** (2001), 120–125. https://doi.org/10.1111/j.0006-341X.2001.00120.x
6. W. J. Fu, Penalized estimating equations, *Biometrics,* **59** (2003), 126–132. http://dx.doi.org/10.1111/1541-0420.00015
7. E. Cantoni, J. M. Flemming, E. Ronchetti, Variable selection for marginal longitudinal generalized linear models, *Biometrics,* **61** (2005), 507–514. http://dx.doi.org/10.1111/j.1541-0420.2005.00331.x

8. L. Wang, A. N. Qu, Consistent model selection and data-driven smooth tests for longitudinal data in the estimating equations approach, *J. Roy. Statist. Soc.,* **71** (2009), 177–190. https://doi.org/10.1111/j.1467-9868.2008.00679.x

9. H. Yang, P. Lin, G. H. Zou, H. Liang, Variable selection and model averaging for longitudinal data incorporating GEE approach, *Stat. Sinica*, **27** (2017), 389–413. http://dx.doi.org/10.5705/ss.2013.277

10. Z. M. Chen, Z. F. Wang, Y. Ivan Chang, Sequential adaptive variables and subject selection for GEE methods, *Biometrics,* **76** (2020), 496–507. http://dx.doi.org/10.1111/biom.13160

11. J. M. Williamson, H. M. Lin, H. X. Barnhart, A classification statistic for GEE categorical response models, *Journal of Data Science*, **1** (2003), 149–165. http://dx.doi.org/10.6339/JDS.2003.01(2).106

12. S. R. Lipsitz, K. Kim, L. P. Zhao, Analysis of repeated categorical data using generalized estimating equations, *Stat. Med.,* **13** (1994), 1149–1163. https://doi.org/10.1002/sim.4780131106

13. K. C. Lin, Y. J. Chen, Assessing GEE models with longitudinal ordinal data by global odds ratio, *Int. Statistical Inst.: Proc. 58th World Statistical Congress,* (2011), 5763–5768.

14. K. Y. Liang, S. L. Zeger, Longitudinal data analysis using generalized linear models, *Biometrika,* **73** (1986), 13–22. https://doi.org/10.1093/biomet/73.1.13

15. J. Q. Fan, R. Z. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, *J. Am. Stat. Assoc.,* **96** (2001), 1348–1360. https://doi.org/10.1198/016214501753382273

16. L. Fahrmeir, G. Tutz, *Multivariate statistcal modelling based on generalized linear models*, New York: Springer, 1994. https://doi.org/10.1007/978-1-4899-0010-4

17. A. Touloumis, A. Agresti, M. Kateri, GEE for multinomial responses using a local odds ratios parameterization, *Biometrics,* **69** (2013), 633–640. http://dx.doi.org/10.1111/biom.12054

18. S. G. Wang, J. H. Shi, S. J. Yin, M. X. Wu, *Introduction to linear models. 3rd ed*, Beijing: Science Press, 2004.

19. A. Touloumis, Simulating correlated binary and multinomial responses under marginal model specification: the SimCorMultRes package, *The R Journal,* **8** (2016), 79–91. http://dx.doi.org/10.32614/RJ-2016-034

20. X. B. Chen, J. L. Yin, Asymptotic properties of GEE estimator for clustered ordinal data with high-dimensional covariates, *Commun. Stat.-Theor. M.,* (2021). http://dx.doi.org/10.1080/03610926.2021.1934029

21. V. D. Vaart, J. Wellner, *Weak convergence and empirical processes: with applications to statistics*, New York: Springer, 1996.