Computational Statistics and Data Analysis 61 (2013) 174–186

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Automatic variable selection for longitudinal generalized linear models

Gaorong Li[a,*], Heng Lian[b], Sanying Feng[a], Lixing Zhu[c]

[a] College of Applied Sciences, Beijing University of Technology, Beijing 100124, PR China
[b] Division of Mathematical Sciences, SPMS, Nanyang Technological University, Singapore
[c] Department of Mathematics, Hong Kong Baptist University, Hong Kong, China

ARTICLE INFO

Article history:
Received 3 October 2011
Received in revised form 2 November 2012
Accepted 22 December 2012
Available online 29 December 2012

Keywords:
Generalized linear model
Longitudinal data
Automatic variable selection
Generalized estimating equations
Oracle property

ABSTRACT

We consider the problem of variable selection for the generalized linear models (GLMs) with longitudinal data. An automatic variable selection procedure is developed using smooth-threshold generalized estimating equations (SGEE). The proposed procedure automatically eliminates inactive predictors by setting the corresponding parameters to be zero, and simultaneously estimates the nonzero regression coefficients by solving the SGEE. The proposed method shares some of the desired features of existing variable selection methods: the resulting estimator enjoys the oracle property; the proposed procedure avoids the convex optimization problem and is flexible and easy to implement. Moreover, we propose a penalized weighted deviance criterion for a data-driven choice of the tuning parameters. Simulation studies are carried out to assess the performance of SGEE, and a real dataset is analyzed for further illustration.

1. Introduction

Generalized linear models (GLMs McCullagh and Nelder, 1989) extend the framework of linear models, by allowing for non-Gaussian data and nonlinear link functions. They have become a favored tool 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 generalized estimating equations (GEE) method was introduced in a seminal paper of Liang and Zeger (1986) as a useful extension of GLMs to correlated data, and has also become a very popular estimation method.

In the present paper, we consider the marginal longitudinal GLMs. Suppose that $\boldsymbol{Y}_i = (y_{i1}, \ldots, y_{im_i})^T$ is the multivariate response for the ith subject, and $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{im_i})^T$ is the $m_i \times p$ matrix of the covariates for the ith subject ($i = 1, \ldots, n$). Observations from different subjects are independent; but those from the same subjects are correlated. Assume that the mean of y_{it} is

$$E(y_{it}|\boldsymbol{x}_{it}) = g(\boldsymbol{\beta}^T\boldsymbol{x}_{it}), \quad i = 1, \ldots, n,\ t = 1, \ldots, m_i, \tag{1.1}$$

where $g(\cdot)$ is a known link function, $\beta = (\beta_1, \ldots, \beta_p)^T$ is the unknown parameter vector of interest, and $\boldsymbol{x}_{it} = (x_{it1}, \ldots, x_{itp})^T$ is the $p \times 1$ vector for $t = 1, \ldots, m_i$.

* Corresponding author.
E-mail addresses: ligaorong@gmail.com (G. Li), HengLian@ntu.edu.sg (H. Lian), fsy5801@sina.com (S. Feng), lzhu@hkbu.edu.hk (L. Zhu).

0167-9473/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2012.12.015

Note that the full likelihood for the model (1.1) is difficult to specify, particularly for correlated non-Gaussian data. Liang and Zeger (1986); also (Zeger and Liang, 1986) developed the GEE approach, a multivariate analogue of the quasi-likelihood, to estimate $\boldsymbol{\beta}$. A key advantage of the GEE approach is that it yields a consistent estimator even if the working correlation structure is misspecified. The GEE estimator is also asymptotically efficient if the correlation structure is indeed correctly specified. Qu et al. (2000) suggested the quadratic inference function to improve the efficiency of GEE, and Balan and Schiopu-Kratina (2005) also rigorously studied a closely related pseudo-likelihood framework for GEE and recommended a two-step estimation procedure. Wang (2011) developed an asymptotic theory for the GEE analysis of clustered binary data when the number of covariates grows to infinity with the number of clusters. Chiou and Müller (2005) proposed the estimated estimating equations (EEE) method based on semiparametric quasi-likelihood regression.

Recently, there has been considerable interest in investigating variable selection problems for GLMs. Variable selection is crucial in statistical modeling, but it is very difficult when an explicit likelihood function is unavailable. To do variable selection, Pan (2001) presented a modification to Akaike information criterion (AIC) called QIC that is obtained by using quasi-likelihood in lieu of likelihood under a strong assumption of working independence. Fu (2003) proposed a generalization of the bridge and Lasso penalties to GEE models. Cantoni et al. (2005) introduced a generalized version of Mallows' C_p to measure model adequacy in prediction. Wang and Qu (2009) developed a Bayesian information type of criterion that is based on the quadratic inference function which they called BIQIF. Xu et al. (in press) proposed a weighted least-squares (WLS) type function to study the longitudinal GLMs with a diverging number of parameters. Xu and Zhu (unpublished manuscript) extended the independence screening method to deal with the high dimensional longitudinal GLMs, and showed that the proposed method still had the so-called sure screening properties. Dziak (2006) generalized the Lasso and SCAD methods to the longitudinal GLMs and studied the $\sqrt{n}$ consistency, the asymptotic normality, and the oracle property of the penalized GEE estimator in Chapter 3 of his Ph.D Thesis. Wang et al. (2012) proposed the SCAD-penalized GEE for analyzing longitudinal data with high-dimensional covariates.

Various penalty functions have been used in the variable selection literature for linear regression models. Frank and Friedman (1993) considered the L_q penalty, which yields a "Bridge Regression". Tibshirani (1996) proposed the Lasso, which can be viewed as a solution to the penalized least squares with the L_1 penalty. Zou (2006) further developed the adaptive Lasso. Through combining both ridge (L_2) and lasso (L_1) penalty together, Zou and Hastie (2005) proposed the Elastic-Net, which also has the sparsity property, to solve the collinearity problems. Fan and Li (2001) proposed the SCAD penalty method and proved that the SCAD estimators enjoy the Oracle properties. All these variable selection procedures are based on penalized estimation using penalty functions, which have a singularity at zero. Consequently, these estimation procedures require convex optimization, which incurs a computational burden. To overcome this problem, Ueki (2009) developed a new variable selection procedure called the smooth-threshold estimating equations that can automatically eliminate irrelevant parameters by setting them as zero. In addition, the resulting estimator enjoys the oracle property in the sense that Fan and Li (2001) suggested.

In this paper we focus on marginal longitudinal generalized linear models and develop our variable selection technique for these models. Motivated by the idea of Ueki (2009), an automatic variable selection procedure is developed using smooth-threshold generalized estimating equations (SGEE). Even though the method is general enough, the details for longitudinal data setting still need to be worked out and the numerical performance examined in details, as we do here. First, one notable difficulty in our setting is that we have to treat the nuisance parameters ϕ and α involved in the working covariance matrix, which affect the final estimator of $\boldsymbol{\beta}$. Computationally, we need to update the values of these nuisance parameters together with the main parameter of interest. Theoretically, in the proof of our asymptotic results, we need to carefully take into account the fact that these nuisance parameters are estimated and explicitly consider their effect on the estimation of $\boldsymbol{\beta}$. Based on the method-of-moment estimators of the nuisance parameters, we propose an iterative algorithm to implement the procedures in Section 2 and obtain the efficient SGEE estimator of $\boldsymbol{\beta}$. Second, combining the GEE with the variable selection method in Ueki (2009), the proposed SGEE procedure not only inherits the advantages of GEE but also avoids the convex optimization problem, which exists in the penalized variable selection methods, such as Bridge regression (Frank and Friedman, 1993), Lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001), et al.. Further, the proposed procedure automatically eliminates the irrelevant parameters by setting them as zero, and simultaneously estimates the nonzero regression coefficients by solving the SGEE. Third, we propose a penalized weighted deviance criterion for the choice of the tuning parameters under the GEE models with longitudinal data. Therefore, the proposed method shares some of the desired features that existing variable selection methods enjoy: the resulting estimator enjoys the oracle property; the proposed procedure avoids the convex optimization problem; the proposed SGEE approach is flexible and easy to implement. Moreover, simulation studies are carried out to assess the performance of our method, and a real dataset is analyzed for further illustration.

The paper is organized as follows. In Section 2, we propose the smooth-threshold generalized estimating equations (SGEE) procedure to automatically eliminate the irrelevant parameters by setting them as zero, and simultaneously estimate the nonzero coefficients. In Section 3, the consistency and oracle property of the SGEE estimators are established. In Section 4, a data-driven penalized weighted deviance criterion is proposed to choose the tuning parameters, and a iterative algorithm is proposed to implement the procedures. In Section 5, some simulations are carried out to illustrate the efficacy of our method. A real data application is then presented to augment our theoretical results. Concluding remarks are presented in Section 6, and the technical details are presented in the Appendix.

2. Methodology

Throughout this paper, let $\boldsymbol{\beta}_0$ be the fixed true value of $\boldsymbol{\beta}$ and let $n \to \infty$ while the m_i are uniformly bounded. We partition $\boldsymbol{\beta}_0$ into active (nonzero) and inactive (zero) coefficients as follows: let $\mathcal{A}_0 = \{j : \beta_{0j} \neq 0\}$ and $\mathcal{A}_0^c = \{j : \beta_{0j} = 0\}$ be the complement of $\mathcal{A}$. Denote by $s = |\mathcal{A}_0|$ the number of true nonzero parameters.

Suppose that the population $(\boldsymbol{X}, \boldsymbol{Y})$ satisfies the marginal longitudinal generalized linear model (1.1). Then the mean of y_{it} is

$$\mu_{it} = E(y_{it}|\boldsymbol{x}_{it}) = g(\boldsymbol{\beta}^T\boldsymbol{x}_{it}), \tag{2.1}$$

and the variance of y_{it} is

$$\mathrm{Var}(y_{it}|\boldsymbol{x}_{it}) = \phi\nu(\mu_{it}), \tag{2.2}$$

where $\nu(\cdot)$ is a variance function, and ϕ is a scale parameter. We first introduce some notations. Let $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{im_i})^T$, $\boldsymbol{D}_i = -\partial\mu_i/\partial\beta$ be an $m_i \times p$ matrix, $\boldsymbol{A}_i$ be an $m_i \times m_i$ diagonal matrix with elements $\phi\nu(\mu_{it})$, and $\boldsymbol{R}_i(\alpha)$ be an $m_i \times m_i$ working correlation matrix, where α is a $q \times 1$ vector which fully characterizes $\boldsymbol{R}_i(\alpha)$. Define the following generalized estimating function

$$U(\boldsymbol{\beta}, \alpha) = \sum_{i=1}^{n} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \mu_i), \tag{2.3}$$

where $\boldsymbol{V}_i = \boldsymbol{A}_i^{1/2}\boldsymbol{R}_i(\alpha)\boldsymbol{A}_i^{1/2}$ is a working covariance matrix. Note that $\boldsymbol{V}_i$ will be equal to $\mathrm{Cov}(\boldsymbol{Y}_i)$ if $\boldsymbol{R}_i(\alpha)$ is indeed the true correlation matrix for $\boldsymbol{Y}_i$.

The main advantage of the GEE method is that it yields a consistent estimator even if the working correlation matrix is misspecified. For models using the canonical link (see McCullagh and Nelder, 1989), $\boldsymbol{D}_i = -\boldsymbol{A}_i\boldsymbol{X}_i$. For linear models, $\boldsymbol{A}_i = \boldsymbol{I}$. For instance, it is often convenient to use a working independence model where $\boldsymbol{R} = \boldsymbol{I}$. Some other popular choices include compound symmetry (CS) (i.e., exchangeable) with $\boldsymbol{R}_{ij} = \rho$ for any $i \neq j$ or first-order autoregressive (AR(1)) with $\boldsymbol{R}_{ij} = \rho^{|i-j|}$, where $\boldsymbol{R}_{ij}$ denotes the (i, j)th element of $\boldsymbol{R}$. Liang and Zeger (1986) suggested approximating $\boldsymbol{R}$ by a working correlation matrix $\hat{\boldsymbol{R}}$ involving only one or a few nuisance parameters α, using *ad hoc*, method-of-moments-like estimators for these α.

Motivated by the idea of Ueki (2009), we propose the following smooth-threshold generalized estimating equations (SGEE)

$$(\boldsymbol{I}_p - \boldsymbol{\Delta})U(\boldsymbol{\beta}, \alpha) + \boldsymbol{\Delta}\boldsymbol{\beta} = \boldsymbol{0}, \tag{2.4}$$

where $\boldsymbol{\Delta}$ is the diagonal matrix whose diagonal elements are $\boldsymbol{\delta} = (\delta_j)_{j=1,\ldots,p}$, and $\boldsymbol{I}_p$ is the p-dimensional identity matrix. Note that the jth SGEE with $\delta_j = 1$ reduces to $\beta_j = 0$. Therefore, SGEE (2.4) can yield a sparse solution. Unfortunately, we cannot directly obtain the estimator of $\boldsymbol{\beta}$ by solving (2.4). This is because the SGEE not only includes the unknown nuisance parameters α and ϕ, but also involves δ_j, which need be chosen using some data-driven criteria.

Since the $\boldsymbol{V}_i$'s are functions of both α and $\boldsymbol{\beta}$, they can be reexpressed as functions of $\boldsymbol{\beta}$ alone by first substituting a $\sqrt{n}$-consistent estimator, $\hat{\alpha}(\beta, \phi)$, in generalized estimating function $U(\boldsymbol{\beta}, \alpha)$ for α, and then replacing ϕ in $\hat{\alpha}$ by a $\sqrt{n}$-consistent estimators. For the choice of $\boldsymbol{\delta} = (\delta_j)_{j=1,\ldots,p}$, Ueki (2009) suggested that δ_j may be determined by the data, and can be chosen by $\hat{\delta}_j = \min(1, \lambda/|\hat{\beta}_j^{(0)}|^{1+\gamma})$ with an initial estimator $\hat{\beta}_j^{(0)}$. The initial estimator $\hat{\beta}_j^{(0)}$ can be obtained by solving the generalized estimating equations $U\{\boldsymbol{\beta}, \hat{\alpha}[\boldsymbol{\beta}, \hat{\phi}(\boldsymbol{\beta})]\} = \boldsymbol{0}$ for the full model. Note that this choice involves two tuning parameters (λ, γ). In Section 4, we will propose a penalized weighted deviance criterion to select the tuning parameters. Replacing Δ in (2.4) by $\widehat{\boldsymbol{\Delta}}$ with diagonal elements $\hat{\boldsymbol{\delta}} = (\hat{\delta}_j)_{j=1,\ldots,p}$, the SGEE becomes

$$(\boldsymbol{I}_p - \widehat{\boldsymbol{\Delta}})U\{\boldsymbol{\beta}, \hat{\alpha}[\boldsymbol{\beta}, \hat{\phi}(\boldsymbol{\beta})]\} + \widehat{\boldsymbol{\Delta}}\boldsymbol{\beta} = \boldsymbol{0}. \tag{2.5}$$

To solve the above SGEE for $\hat{\boldsymbol{\beta}}$, we need to iterate between a modified Fisher scoring for the regression coefficients and moment estimation of the correlation and scale parameters, α and ϕ (see (Liang and Zeger, 1986) for the given tuning parameters (λ, γ). Similarly, we can define the active set $\mathcal{A} = \{j : \hat{\delta}_j \neq 1\}$ which is the set of indices of nonzero parameters, where $\hat{\delta}_j = \min(1, \lambda/|\hat{\beta}_j^{(0)}|^{1+\gamma})$. The solution of (2.5) denoted by $\hat{\boldsymbol{\beta}}_{\lambda,\gamma}$ is called the *SGEE estimator*. Given current estimates $\hat{\alpha}$ and $\hat{\phi}$ of the nuisance parameters and the tuning parameters (λ, γ), we propose the following modified iterative procedure for $\boldsymbol{\beta}$:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\text{new}} = \hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\text{old}} &- \left\{\sum_{i=1}^{n} \boldsymbol{D}_{i,\mathcal{A}}^T(\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\text{old}})\widetilde{\boldsymbol{V}}_i^{-1}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\text{old}})\boldsymbol{D}_{i,\mathcal{A}}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\text{old}}) + \hat{\boldsymbol{G}}_{\mathcal{A}}\right\}^{-1} \\ &\times \left\{\sum_{i=1}^{n} \boldsymbol{D}_{i,\mathcal{A}}^T(\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\text{old}})\widetilde{\boldsymbol{V}}_i^{-1}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\text{old}})\Big(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\text{old}})\Big) + \hat{\boldsymbol{G}}_{\mathcal{A}}\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\text{old}}\right\}\end{aligned} \tag{2.6}$$

and

$$\hat{\boldsymbol{\beta}}_{\mathcal{A}^c,\widehat{\Delta}} = \mathbf{0}, \tag{2.7}$$

where $\widetilde{\boldsymbol{V}}_i(\boldsymbol{\beta}) = \boldsymbol{V}_i(\boldsymbol{\beta}, \hat{\alpha}\{\boldsymbol{\beta}, \hat{\phi}(\boldsymbol{\beta})\})$, $\hat{\boldsymbol{G}}_{\mathcal{A}} = (\boldsymbol{I}_{|\mathcal{A}|} - \widehat{\boldsymbol{\Delta}}_{\mathcal{A}})^{-1}\widehat{\boldsymbol{\Delta}}_{\mathcal{A}}$, $\boldsymbol{D}_{i,\mathcal{A}}(\boldsymbol{\beta}_{\mathcal{A}}) = -\partial\mu_i(\beta_{\mathcal{A}})/\partial\beta_{\mathcal{A}}$, and $\mu_i(\beta_{\mathcal{A}}) = g(\boldsymbol{X}_{i,\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}})$. In particular, for the longitudinal linear model, (2.6) and (2.7) can be reduced to

$$\hat{\boldsymbol{\beta}}_{\mathcal{A}} = \left\{\sum_{i=1}^{n} \boldsymbol{X}_{i,\mathcal{A}}^T \hat{\boldsymbol{V}}_i^{*-1}\boldsymbol{X}_{i,\mathcal{A}} + \hat{\boldsymbol{G}}_{\mathcal{A}}\right\}^{-1} \sum_{i=1}^{n} \boldsymbol{X}_{i,\mathcal{A}}^T \hat{\boldsymbol{V}}_i^{*-1}\boldsymbol{Y}_i, \quad \text{and} \quad \hat{\boldsymbol{\beta}}_{\mathcal{A}^c,\widehat{\Delta}} = \mathbf{0}, \tag{2.8}$$

where $\boldsymbol{V}_i^* = \mathrm{Cov}(\boldsymbol{Y}_i|\boldsymbol{X}_i)$, and $\hat{\boldsymbol{V}}_i^*$ is the estimator of $\boldsymbol{V}_i^*$ which can be estimated using the method of moments.

3. Asymptotic properties

In this section, we assume, under the regularity conditions, the initial estimator using the full model is consistent and asymptotically normally distributed by solving the GEE, that is $U(\boldsymbol{\beta}, \alpha) = 0$ (see Liang and Zeger, 1986). Following Fan and Li (2001), it is possible to prove the oracle properties for the SGEE estimators, including $\sqrt{n}$-consistency, variable selection consistency, and asymptotic normality.

Theorem 1 ($\sqrt{n}$-*consistency*)**.** *Under mild regularity conditions and given that:*
(1) $\hat{\alpha}$ *is* $\sqrt{n}$*-consistent given* $\boldsymbol{\beta}$ *and* ϕ*;*
(2) $\hat{\phi}$ *is* $\sqrt{n}$*-consistent given* $\boldsymbol{\beta}$*; and*
(3) $|\partial\hat{\alpha}(\boldsymbol{\beta}, \phi)/\partial\phi| \le H(\boldsymbol{Y}, \boldsymbol{\beta})$ *which is* $O_P(1)$*, where* $H(\cdot, \cdot)$ *is a function of the sample* $\boldsymbol{Y}$ *and* $\boldsymbol{\beta}$*.*

For any positive λ *and* γ *such that* $n^{1/2}\lambda \to 0$ *and* $n^{(1+\gamma)/2}\lambda \to \infty$ *as* $n \to \infty$*, there exists a sequence* $\hat{\boldsymbol{\beta}}_{\lambda,\gamma}$ *of the solutions of* (2.5) *such that* $\|\hat{\boldsymbol{\beta}}_{\lambda,\gamma} - \boldsymbol{\beta}_0\| = O_P(n^{1/2})$*.*

Theorem 2. *Suppose that the conditions of Theorem 1 hold, as* $n \to \infty$*, we have*
(i) *variable selection consistency, i.e.* $P(\mathcal{A} = \mathcal{A}_0) \to 1$*;*
(ii) *asymptotic normality, i.e.* $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\lambda,\gamma,\mathcal{A}_0} - \beta_{\mathcal{A}_0})$ *is asymptotically normally distributed with mean zero and covariance matrix* Φ *identical to that of the oracle estimator, where* Φ *is the limit in probability of, as* $n \to \infty$*,*

$$n\left\{\sum_{i=1}^{n} \boldsymbol{D}_{i,\mathcal{A}_0}^T \boldsymbol{V}_i^{-1}\boldsymbol{D}_{i,\mathcal{A}_0}\right\}^{-1} \left\{\sum_{i=1}^{n} \boldsymbol{D}_{i,\mathcal{A}_0}^T \boldsymbol{V}_i^{-1}\mathrm{Cov}(\boldsymbol{Y}_i)\boldsymbol{V}_i^{-1}\boldsymbol{D}_{i,\mathcal{A}_0}\right\} \left\{\sum_{i=1}^{n} \boldsymbol{D}_{i,\mathcal{A}_0}^T \boldsymbol{V}_i^{-1}\boldsymbol{D}_{i,\mathcal{A}_0}\right\}^{-1}.$$

Theorem 2 implies that the proposed automatic SGEE procedure is consistent in variable selection; it can identify the zero coefficients with probability tending to 1. By choosing appropriate tuning parameters, the SGEE estimators have the oracle property; that is, the asymptotic variance for the SGEE estimate is the same as what we would have if we knew in advance the correct submodel.

4. Issues in practical implementation

4.1. Tuning parameter selection

To implement the procedures described in Section 2, we need to choose the tuning parameters (λ, γ). One can select (λ, γ) by optimizing some data-driven criteria which balance goodness of fit and model complexity, such as the classical C_P, GCV or BIC. Fu (2003) considered how to choose the tuning parameters, and pointed out that it is difficult to extend these criteria to the GEE directly due to the lack of joint likelihood in the GEE models. Fu (2003) generalized the classical RSS to the following weighted deviance

$$\mathrm{WDev} = \sum_{i=1}^{n} \boldsymbol{r}_i^T \boldsymbol{R}_i^{-1}(\alpha)\boldsymbol{r}_i, \tag{4.1}$$

which takes into account correlations and allows non-Gaussian responses. Where $\boldsymbol{R}_i(\alpha)$ is the $m_i \times m_i$ working correlation matrix, and $\boldsymbol{r}_i$ are the deviance residuals (see McCullagh and Nelder, 1989; Agresti, 2002), although they could also be reasonably replaced by the Pearson residuals $\boldsymbol{A}^{-1/2}(\boldsymbol{Y} - \boldsymbol{\mu})$ for simplicity.

Based on the above discussion, we here propose the following penalized weighted deviance criterion:

$$\mathrm{PWD}_{(\lambda,\gamma)} = \mathrm{WDev} + \mathrm{DF}_{(\lambda,\gamma)} \log n, \tag{4.2}$$

where $\mathrm{DF}_{(\lambda,\gamma)} = \sum_{j=1}^{p} 1(\hat{\delta}_j \neq 1)$ denotes the number of nonzero parameters with $1(\cdot)$ the indicator function. We can choose (λ, γ) by minimizing the $\mathrm{PWD}_{(\lambda,\gamma)}$ (4.2). Advocating the penalized weighted deviance as selection criterion is based on our experience that it performs well in both simulations and real data examples (see Section 5).

4.2. Iterative algorithm

To obtain the *SGEE estimator* of $\boldsymbol{\beta}$ for solving (2.5) using the Newton–Raphson, Fisher scoring and iteratively reweighted least squares, we need obtain the $\sqrt{n}$-consistent estimators of the correlation parameters α and scale parameter ϕ. Therefore, we first discuss the estimation of the correlation parameters and scale parameter. The scale parameter can be estimated using Pearson's χ^2 method-of-moments (MOM) by

$$\hat{\phi} = \frac{1}{N-p}\sum_{i=1}^{n}\sum_{t=1}^{m_i}\frac{(y_{it}-g(\boldsymbol{x}_{it}^T\hat{\boldsymbol{\beta}}))^2}{\nu(g(\boldsymbol{x}_{it}^T\hat{\boldsymbol{\beta}}))}, \tag{4.3}$$

where $N = \sum_{i=1}^{n} m_i$ and $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$. Note that $\hat{\phi}\nu(g(\boldsymbol{x}_{it}^T\hat{\boldsymbol{\beta}}))$ is the tth diagonal element of $\hat{\boldsymbol{A}}_i$. It is easy to show that $\hat{\phi}$ is $\sqrt{n}$-consistent given that the fourth moments of the y_{it}'s are finite.

In practice, the correlation parameters are considered as nuisance parameters. To estimate α consistently, we use the corrected method-of-moment estimators of Liang and Zeger (1986). We define the (it)th Pearson residual as

$$\hat{r}_{it} = \frac{y_{it}-g(\boldsymbol{x}_{it}^T\hat{\boldsymbol{\beta}})}{\sqrt{\widehat{\mathrm{Var}}(y_{it})}}, \tag{4.4}$$

where $\widehat{\mathrm{Var}}(y_{it}) = \hat{\phi}\nu(g(\boldsymbol{x}_{it}^T\hat{\boldsymbol{\beta}}))$ is the estimated variance of y_{it}. Liang and Zeger (1986) suggested to estimate α by

$$\hat{\boldsymbol{R}}_{uv} = \frac{1}{df}\sum_{i=1}^{n}\hat{r}_{iu}\hat{r}_{iv}, \tag{4.5}$$

where $df = N-p$ is the degrees of freedom. The method-of-moments estimators of α for various correlation structures are given in Section 4 of Liang and Zeger (1986). For example, for the exchangeable working correlation (CS) structure, α can be estimated by

$$\hat{\alpha} = \frac{1}{\sum_{i=1}^{n}\frac{1}{2}m_i(m_i-1)-p}\sum_{i=1}^{n}\sum_{t>t'}\hat{r}_{it}\hat{r}_{it'}.$$

For the AR(1) working correlation structure, the $(t, t+1)$th element of $\boldsymbol{R}(\alpha)$ can be estimated by

$$\hat{\boldsymbol{R}}_{t,t+1} = \frac{1}{N-n-p}\sum_{i=1}^{n}\sum_{t=1}^{m_i-1}\hat{r}_{it}\hat{r}_{i(t+1)}.$$

Then, we propose the iterative algorithm to implement the procedures described in Section 2 as follows.

Step 1. Given an initial estimator $\hat{\boldsymbol{\beta}}^{(0)}$. Let $k = 0$.

Step 2. Based on (4.3) and (4.5), we estimate the correlation parameters α and scale parameter ϕ using the current estimate $\hat{\boldsymbol{\beta}}^{(k)}$, and compute the working covariance matrix $\boldsymbol{V}_i(\hat{\boldsymbol{\beta}}^{(k)}, \hat{\alpha}\{\hat{\boldsymbol{\beta}}^{(k)}, \hat{\phi}(\hat{\boldsymbol{\beta}}^{(k)})\}) = \hat{\boldsymbol{A}}_i^{1/2}\boldsymbol{R}_i(\hat{\alpha})\hat{\boldsymbol{A}}_i^{1/2}$. Meanwhile, we choose the tuning parameters (λ, γ) based on the penalized weighted deviance criterion (4.2).

Step 3. Update the estimator $\hat{\boldsymbol{\beta}}^{(k+1)}$ of $\boldsymbol{\beta}$ by solving the smooth-threshold generalized estimating Eq. (2.5).

Step 4. Iterate Step 2–Step 3 until convergence, and denote the final estimators of $\boldsymbol{\beta}$ as the *SGEE estimator*.

Note that the quantities $(\alpha, \phi, \lambda, \gamma)$ all change with iterations. However, we do not make this point explicit in our notations for simplicity. Instead, we implicitly regard $(\alpha, \phi, \lambda, \gamma)$ as functions of $\boldsymbol{\beta}$. Thus as the estimate of $\boldsymbol{\beta}$ changes in each iteration, these quantities also change.

In the initialization step, the initial estimator of $\boldsymbol{\beta}$ is in practice an important task. Theoretically, we require that the initial estimator is root-n consistent. The initial estimator not only affects the degree of sparsity of the solution and the accuracy of the final estimator, but also affects the speed of convergence of our iterative algorithm. In simulation studies, we use the GEE estimator as the initial estimate. The simulation results show that the proposed iterative algorithm is workable.

5. Numerical studies

5.1. Simulation studies

In this section, we report some simulation studies to illustrate the finite sample properties of the proposed SGEE procedure. Throughout the simulation studies, each dataset comprised $n = 100, 200$ and 400 subjects and $m_i \equiv 5$ observations per subject over time. For each case, we repeat the experiment M times and applied the penalized weighted deviance criterion (4.2) to select the tuning parameters. In the simulation studies, we measure the accuracy of estimation by the average mean square error (AMSE), which is $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|^2$ averaged over M simulated data sets. We consider the following three examples.

Table 1
Variable selections for linear model (5.1) using SGEE, when the correlation structure is correctly specified.

n	Method	α	CS			AR(1)		
			Correct	Incorrect	AMSE	Correct	Incorrect	AMSE
100	SGEE	0.3	5.817	0	0.004927	5.809	0	0.005017
		0.5	5.851	0	0.003790	5.839	0	0.003874
		0.7	5.874	0	0.002451	5.863	0	0.002709
	Oracle	0.3	6	0	0.003265	6	0	0.003421
		0.5	6	0	0.002389	6	0	0.002597
		0.7	6	0	0.001480	6	0	0.001603
200	SGEE	0.3	5.962	0	0.001892	5.876	0	0.001964
		0.5	5.986	0	0.001236	5.930	0	0.001387
		0.7	6	0	0.000933	5.998	0	0.000986
	Oracle	0.3	6	0	0.001561	6	0	0.001704
		0.5	6	0	0.001158	6	0	0.001262
		0.7	6	0	0.000754	6	0	0.000777
400	SGEE	0.3	6	0	0.000854	6	0	0.000876
		0.5	6	0	0.000649	6	0	0.000674
		0.7	6	0	0.000411	6	0	0.000426
	Oracle	0.3	6	0	0.000796	6	0	0.000851
		0.5	6	0	0.000564	6	0	0.000636
		0.7	6	0	0.000370	6	0	0.000386

Example 1 (*Continuous Responses*)**.** In this example, we first consider the linear model as a special case of GLM, and we specify $p = 10$ covariates with the true parameter $\boldsymbol{\beta}_0 = [1, 0.5, 0, 0, 0, -0.2, 0, 0, 0.4, 0]^T$, where four regression variables are significant, but the rest are not. The response variable is generated according to the model

$$y_{it} = \sum_{k=1}^{10} x_{it,k}\beta_{0k} + \epsilon_{it}, \quad i = 1, \dots, n, t = 1, \dots, 5, \tag{5.1}$$

where each covariate $\boldsymbol{x}_{i,k} = (x_{i1,k}, \dots, x_{i5,k})^T$ is independently generated from a multivariate normal distribution with mean $(0.1, 0.2, 0.3, 0.4, 0.5)^T$ and an identity covariance matrix, and the random error vectors $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \dots, \epsilon_{i5})^T$ are generated independently of the covariates from a five-dimensional normal distribution with mean 0, marginal variance 1 and a working correlation matrix $\boldsymbol{R}(\alpha)$. Consider two kinds of working correlation matrices: exchangeable working correlation (CS) with a correlation coefficient α and AR(1) working correlation structure with an auto-correlation coefficient α. For comparison, we take three different values of α: 0.3, 0.5 and 0.7. Based on the experiment time $M = 1000$, the simulation results are reported in Tables 1 and 2, for the correctly and incorrectly specified correlation structure respectively. In the tables, values in the column labeled "Correct" denote the average number of coefficients of the true zeros, correctly set to zero, and those in the column labeled "Incorrect" denote the average number of the true nonzeros incorrectly set to zero.

It is easy to see from Tables 1 and 2 that the proposed SGEE method is able to correctly identify the true submodel, and works remarkably well, even if the working correlation structure is misspecified. Not surprisingly using the correct correlation structure was better than using an incorrect correlation structure. We also note that the performance did not significantly depend on working covariance structure, despite the fairly strong ($\alpha = 0.7$) within-subject correlation parameter. The larger the sample size n, the better the proposed method performs.

Example 2 (*Discrete Responses*)**.** Consider the following logistic regression model. The response variable y_{it} is binary and its marginal expectation given $\boldsymbol{x}_{it}$ is

$$\mathrm{logit}(\mu_{it}) = \sum_{k=1}^{10} x_{it,k}\beta_{0k}, \quad i = 1, \dots, n, t = 1, \dots, 5, \tag{5.2}$$

where $\boldsymbol{\beta}_0 = [1, 0.5, 0, 0, 0, -0.2, 0, 0, 0.4, 0]^T$, and the covariate $\boldsymbol{x}_{it} = (x_{it,1}, \dots, x_{it,10})^T$ has a multivariate normal distribution with mean zero, marginal variance 0.2 and an AR(1) correlation matrix with autocorrelation coefficient 0.5. The binary response vector for each cluster has mean specified by (5.2) and an exchangeable correlation structure with correlation coefficient α or an AR(1) correlation structure with an auto-correlation coefficient α. Three values of α are considered: $\alpha = 0.3, 0.5$ and 0.7. Such correlated binary data are generated using R code with the correlated random binary data generator provided by Oman (2009). The experiment is repeated $M = 1000$ times, and the simulation results are reported in Tables 3 and 4.

As the results in Tables 3 and 4 are substantively similar to the previous example, they are not discussed further.

Example 3 (*High-dimensional Setup*)**.** In this example, we discuss how the proposed SGEE procedure can be applied to the "large n, diverging p" setup for longitudinal GLMs. In addition, we also compare the proposed method with the existing

Table 2
Variable selections for linear model (5.1) using SGEE, when the correlation structure is incorrectly specified. The term "CS.AR(1)" means estimation with the fitted misspecified AR(1) correlation structure, while "AR(1).CS" means estimation with the fitted misspecified CS correlation structure.

n	Method	α	CS.AR(1)			AR(1).CS		
			Correct	Incorrect	AMSE	Correct	Incorrect	AMSE
100	SGEE	0.3	5.775	0	0.005173	5.717	0	0.005458
		0.5	5.828	0	0.004650	5.787	0	0.004762
		0.7	5.870	0	0.002704	5.855	0	0.003173
	Oracle	0.3	6	0	0.003501	6	0	0.003648
		0.5	6	0	0.003092	6	0	0.003187
		0.7	6	0	0.001966	6	0	0.002262
200	SGEE	0.3	5.951	0	0.001917	5.928	0	0.002124
		0.5	5.973	0	0.001530	5.957	0	0.001769
		0.7	5.993	0	0.001083	5.982	0	0.001131
	Oracle	0.3	6	0	0.001761	6	0	0.001769
		0.5	6	0	0.001491	6	0	0.001564
		0.7	6	0	0.000963	6	0	0.001158
400	SGEE	0.3	6	0	0.000951	6	0	0.000968
		0.5	6	0	0.000769	6	0	0.000816
		0.7	6	0	0.000585	6	0	0.000594
	Oracle	0.3	6	0	0.000901	6	0	0.000915
		0.5	6	0	0.000736	6	0	0.000788
		0.7	6	0	0.000469	6	0	0.000569

Table 3
Variable selections for generalized linear model (5.2) using SGEE, when the correlation structure is correctly specified.

n	Method	α	CS			AR(1)		
			Correct	Incorrect	AMSE	Correct	Incorrect	AMSE
100	SGEE	0.3	5.555	0.071	0.041723	5.532	0.075	0.042717
		0.5	5.586	0.064	0.040882	5.563	0.069	0.042596
		0.7	5.641	0.054	0.040683	5.601	0.061	0.042058
	Oracle	0.3	6	0	0.038013	6	0	0.040818
		0.5	6	0	0.037901	6	0	0.039470
		0.7	6	0	0.037458	6	0	0.039092
200	SGEE	0.3	5.687	0.037	0.023053	5.670	0.038	0.024436
		0.5	5.824	0.029	0.022596	5.708	0.035	0.023726
		0.7	5.894	0.016	0.022173	5.879	0.019	0.023560
	Oracle	0.3	6	0	0.019206	6	0	0.021787
		0.5	6	0	0.019162	6	0	0.021596
		0.7	6	0	0.019101	6	0	0.021209
400	SGEE	0.3	5.964	0	0.011588	5.951	0	0.012047
		0.5	5.970	0	0.011492	5.968	0	0.011683
		0.7	6.000	0	0.011152	5.985	0	0.011359
	Oracle	0.3	6	0	0.009421	6	0	0.009593
		0.5	6	0	0.009245	6	0	0.009251
		0.7	6	0	0.009047	6	0	0.009222

methods, such as SCAD-based penalized GEE (SCAD-GEE) proposed in Wang et al. (2012) and Lasso-based penalized GEE (Lasso-GEE) proposed in Fu (2003) and Dziak (2006). We consider the following high-dimensional logistic model. The response variable y_{it} is binary and its marginal expectation given $\boldsymbol{x}_{it}$ is

$$\mathrm{logit}(\mu_{it}) = \sum_{k=1}^{p} x_{it,k}\beta_{0k}, \quad i = 1, \ldots, n, t = 1, \ldots, 5, \tag{5.3}$$

where $\boldsymbol{\beta}_0$ is a p-dimensional vector of parameters with $p = \lfloor 4n^{1/3} \rfloor - 5$ for $n = 100, 200$ and 400, and $\lfloor s \rfloor$ denotes the largest integer not greater than s. The covariate vectors $\boldsymbol{x}_{it}$ are i.i.d. from normal distribution $N_p(\mathbf{0}_p, \Sigma)$ with Σ whose (i, j)th element is equal to $0.5^{|i-j|}$. The true coefficient vector is $\boldsymbol{\beta}_0 = [0.6\boldsymbol{1}_d, \mathbf{0}_{p-d}]^T$, where $d = \lfloor p/5 \rfloor$ and $\boldsymbol{1}_m/\mathbf{0}_m$ denotes a m-vector of 1s/0s. The binary response vector for each cluster has mean specified by (5.3) and an AR(1) correlation structure with correlation coefficient α. Three values of α are considered: $\alpha = 0.3$, 0.5 and 0.7, and the experiment is repeated $M = 1000$ times. The summary of simulation results is reported in Table 5.

Several observations can be found from Table 5. First, all of the three variable selection procedures are able to correctly identify the true submodel even if the working correlation structure is misspecified. Second, the proposed SGEE method

Table 4
Variable selections for generalized linear model (5.2) using SGEE, when the correlation structure is incorrectly specified. The term "CS.AR(1)" means estimation with the fitted misspecified AR(1) correlation structure, while "AR(1).CS" means estimation with the fitted misspecified CS correlation structure.

n	Method	α	CS.AR(1)			AR(1).CS		
			Correct	Incorrect	AMSE	Correct	Incorrect	AMSE
100	SGEE	0.3	5.535	0.081	0.048246	5.503	0.085	0.049576
		0.5	5.582	0.065	0.047681	5.560	0.075	0.049036
		0.7	5.598	0.053	0.047132	5.585	0.063	0.048377
	Oracle	0.3	6	0	0.041513	6	0	0.043816
		0.5	6	0	0.041478	6	0	0.043384
		0.7	6	0	0.041109	6	0	0.043138
200	SGEE	0.3	5.676	0.041	0.027412	5.666	0.050	0.028073
		0.5	5.740	0.033	0.027259	5.704	0.037	0.027635
		0.7	5.871	0.023	0.026921	5.859	0.027	0.027147
	Oracle	0.3	6	0	0.024257	6	0	0.026242
		0.5	6	0	0.024143	6	0	0.025712
		0.7	6	0	0.023746	6	0	0.025152
400	SGEE	0.3	5.929	0	0.014547	5.904	0	0.014918
		0.5	5.946	0	0.014392	5.932	0	0.014631
		0.7	5.977	0	0.013738	5.971	0	0.014122
	Oracle	0.3	6	0	0.012578	6	0	0.012927
		0.5	6	0	0.012191	6	0	0.012914
		0.7	6	0	0.011615	6	0	0.012911

performs significantly better and has the smaller AMSE than the SCAD-GEE and Lasso-GEE methods. Third, it is worth mentioning that the SCAD-GEE method significantly reduces the AMSE and its results become comparable when the sample size increases. In addition, the SGEE and the SCAD-GEE procedures perform closely to the oracle GEE.

5.2. *Application to real data*

We now illustrate the proposed SGEE method through an application to a real dataset (Petkau et al., 2004; Petkau and White, 2003), and was previously analyzed in the book of Song (2007). The real data concerns a longitudinal clinical trial to assess the effects of neutralizing antibodies on interferon beta-1b (IFNB) in relapsing–remitting multiple sclerosis (MS), which is a disease that destroys the myelin sheath that surrounds the nerves. The data are from a Magnetic Resonance Imaging (MRI) sub-study of the Betaseron clinical trial conducted at the University of British Columbia in relapsing–remitting multiple sclerosis involving 50 patients, each of whom visits the university every six weeks. The patients were randomized into three treatment groups, with allocation of 17 patients being treated by placebo, 17 by low dose, and 16 by high dose. There exist the missing values in this dataset, we should analyze the unbalanced longitudinal data using the proposed method and compare it with the SCAD-GEE and the Lasso-GEE procedures.

For the analysis of this real dataset, the binary response variable is Exacerbation, which refers to whether an exacerbation appeared since the previous MRI scan, with 1 for yes and 0 for no. Seven explanatory variables are recorded: Treatment (Trt), Time (T) in weeks, Squared time (T^2), Age, Gender, Duration of disease (Dur) in years, and an additional baseline covariate initial EDSS (Expanded Disability Status Scale) scores. Similar to the analysis idea in Song (2007, Page 172), instead of treating the three dosage levels (placebo, low, and high dosage of the drug treatment) as one ordinal covariate, the placebo group should be treated as a comparison group and two dummy variables for the Trt are set as follows

$$\mathrm{Ltrt} = \begin{cases} 1, & \text{Low Dose} \\ 0, & \text{Otherwise,} \end{cases} \qquad \mathrm{Htrt} = \begin{cases} 1, & \text{High Dose} \\ 0, & \text{Otherwise.} \end{cases}$$

We consider the following marginal logistic model for this data:

$$\mathrm{logit}(\mu_{ij}) = \beta_0 + \beta_1 \mathrm{T}_j + \beta_2 \mathrm{T}_j^2 + \beta_3 \mathrm{Age}_i + \beta_4 \mathrm{Gender}_i + \beta_5 \mathrm{Dur}_i + \beta_6 \mathrm{EDSS}_i + \beta_7 \mathrm{Ltrt}_i + \beta_8 \mathrm{Htrt}_i, \tag{5.4}$$

where μ_{ij} is the probability of exacerbation at visit j for subject i. Two correlation structures (exchangeable (CS) and AR(1)) are considered in this analysis. Table 6 reports the estimated coefficients and the standard errors.

From Table 6, we can see that Duration of disease (Dur), EDSS and Htrt are statistically significant variables for all of the three variable selection procedures, and Gender has a positive impact on Exacerbation based on the proposed SGEE method. The effects of these variables T, T^2, Age and Ltrt are eliminated from the model as they are not significant in the analysis. Similar to the analysis of Song (2007), based on the results of the SGEE method for CS correlation structure, one unit increase

Table 5
Variable selections for high-dimensional logistic model (5.3) when the true correlation structure is exchangeable. The term "AR(1).AR(1)" means estimation with the fitted true specified AR(1) correlation structure, while "AR(1).CS" means estimation with the fitted misspecified CS correlation structure.

(n, p, d)	Method	α	AR(1).AR(1)			AR(1).CS		
			Correct	Incorrect	AMSE	Correct	Incorrect	AMSE
(100,13,2)		0.3	10.902	0.005	0.0424	10.898	0.005	0.0425
	SGEE	0.5	10.941	0.004	0.0368	10.940	0.004	0.0382
		0.7	10.970	0.002	0.0312	10.969	0.005	0.0330
		0.3	10.805	0.003	0.0741	10.801	0.003	0.0772
	SCAD-GEE	0.5	10.818	0.001	0.0649	10.817	0.003	0.0710
		0.7	10.889	0.000	0.0531	10.886	0.000	0.0610
		0.3	10.953	0.000	0.1298	10.949	0.000	0.1327
	Lasso-GEE	0.5	10.957	0.000	0.1216	10.957	0.001	0.1273
		0.7	10.946	0.000	0.1067	10.937	0.000	0.1130
		0.3	11	0	0.0292	11	0	0.0294
	Oracle	0.5	11	0	0.0292	11	0	0.0299
		0.7	11	0	0.0275	11	0	0.0273
(200,18,3)		0.3	14.939	0.001	0.0312	14.894	0.001	0.0342
	SGEE	0.5	14.961	0.001	0.0285	14.914	0.000	0.0320
		0.7	14.980	0.000	0.0254	14.938	0.000	0.0279
		0.3	14.918	0.001	0.0364	14.913	0.002	0.0369
	SCAD-GEE	0.5	14.936	0.001	0.0332	14.935	0.001	0.0354
		0.7	14.959	0.000	0.0297	14.956	0.000	0.0307
		0.3	14.913	0.000	0.1036	14.912	0.000	0.1053
	Lasso-GEE	0.5	14.935	0.000	0.0986	14.924	0.000	0.1022
		0.7	14.938	0.000	0.0880	14.925	0.000	0.0917
		0.3	15	0	0.0264	15	0	0.0268
	Oracle	0.5	15	0	0.0254	15	0	0.0261
		0.7	15	0	0.0239	15	0	0.0235
(400,24,4)		0.3	20.000	0.000	0.0204	20.000	0.000	0.0206
	SGEE	0.5	20.000	0.000	0.0197	20.000	0.000	0.0202
		0.7	20.000	0.000	0.0189	20.000	0.000	0.0195
		0.3	19.970	0.000	0.0208	19.969	0.000	0.0216
	SCAD-GEE	0.5	19.988	0.000	0.0215	19.975	0.000	0.0232
		0.7	19.991	0.000	0.0196	19.981	0.000	0.0202
		0.3	19.744	0.000	0.0709	19.676	0.000	0.0773
	Lasso-GEE	0.5	19.926	0.000	0.0770	19.914	0.000	0.0948
		0.7	19.974	0.000	0.0808	19.971	0.000	0.1124
		0.3	20	0	0.0195	20	0	0.0198
	Oracle	0.5	20	0	0.0189	20	0	0.0194
		0.7	20	0	0.0188	20	0	0.0190

in EDSS will result in an increase in the odds of exacerbation by $\exp(0.2714) = 1.3118$. In addition, we also can see that the odds ratio of exacerbation is $\exp(-0.0454) = 0.9556152$ between a patient who had a disease history of $T + 1$ years and a patient who had a disease history of T years. Therefore, these findings are close to the existing analysis in Song (2007).

6. Concluding remarks

The article develops a SGEE procedure for automatic variable selection in the marginal longitudinal generalized linear models that allows for non-Gaussian data and nonlinear link functions. This approach is flexible, conceptually simple and easy to implement. The estimation procedure can be implemented in an iterative algorithm that alternates between a modified Fisher scoring for the regression coefficients and moment estimation of the correlation and scale parameters, α and ϕ for given tuning parameters (λ, γ) by solving the smooth-threshold generalized estimating equations (SGEE). The proposed procedure automatically eliminates the irrelevant parameters by setting them as zero, and simultaneously estimates the nonzero regression coefficients. It is noteworthy that the proposed procedure avoids the convex optimization problem, and the resulting estimator enjoys the oracle property followed in Fan and Li (2001).

The methods described here will be easily extended to various statistical models based on estimating equations. These and other extensions are the subject of ongoing research.

For GLMs with longitudinal data with the number of variables p being larger than n, it is an interesting future research topic using the SGEE procedure. From the independence screening method (Xu and Zhu, unpublished manuscript), we may expect that the SGEE procedure will work well. The idea is to use first the independence screening to the model dimension down to a number smaller than to sample size in a probability approaching one, and then the SGEE procedure can efficiently be used for remaining variables.

Table 6
Estimates and standard errors for the real data.

Coefficients	GEE		SGEE	
	CS	AR(1)	CS	AR(1)
Intercept	−1.8131(1.5530)	−1.8692(0.2916)	−1.9872(0.4460)	−1.8426(0.4460)
T	−0.0313(0.0102)	−0.0184(0.0782)	0 (−)	0 (−)
T^2	0.0002(0.0001)	0.0001(0.2914)	0 (−)	0 (−)
Age	0.0138(0.0433)	−0.0019(0.2916)	0 (−)	0 (−)
Gender	0.1155(0.7778)	0.2729(0.0782)	0.1556 (0.3294)	0.1225(0.3294)
Dur	−0.0485(0.0549)	−0.0415(0.2914)	−0.0454(0.0240)	−0.0484(0.0240)
EDSS	0.4017(0.0828)	0.4613(0.2916)	0.2714 (0.0790)	0.2365(0.0790)
Ltrt	0.0200(0.7864)	−0.1402(0.0782)	0 (−)	0 (−)
Htrt	−0.8320(0.7994)	−0.5140(0.2914)	−0.3195(0.3036)	−0.3017(0.3036)
Coefficients	SCAD-GEE		Lasso-GEE	
	CS	AR(1)	CS	AR(1)
Intercept	−2.1259(0.2558)	−1.8422(0.2547)	−1.3334(0.4847)	−2.0415(0.5093)
T	−0.0017(0.0018)	−0.0014(0.0017)	−0.0291(0.0118)	−0.0081(0.0042)
T^2	0 (−)	0 (−)	0 (−)	0 (−)
Age	0 (−)	0 (−)	0 (−)	0 (−)
Gender	0 (−)	0 (−)	0 (−)	0.0621(0.1228)
Dur	−0.0104(0.0046)	−0.0268(0.0121)	−0.0320(0.0321)	−0.0398(0.0201)
EDSS	0.2567(0.0665)	0.2258(0.0695)	0.3789(0.1180)	0.4501(0.1315)
Ltrt	0 (−)	0 (−)	0 (−)	0 (−)
Htrt	−0.0342(0.0257)	−0.2519(0.2059)	−0.8575(0.4570)	−0.3583(0.2575)

Acknowledgments

Gaorong Li's research was supported by the National Nature Science Foundation of China (11101014), the Specialized Research Fund for the Doctoral Program of Higher Education of China (20101103120016), Funding Project for Academic Human Resources Development in Institutions of Higher Learning Under the Jurisdiction of Beijing Municipality (PHR20110822), Training Programme Foundation for the Beijing Municipal Excellent Talents (2010D005015000002), the Fundamental Research Foundation of Beijing University of Technology (X4006013201101) and Program for JingHua Talents in Beijing University of Technology. Heng Lian's research was supported by Singapore MOE Tier 1 RG 36/09. Lixing Zhu's research was supported by a grant from the Research Grants Council of Hong Kong, and a FRG grant from Hong Kong Baptist University, Hong Kong. The authors would like to thank the Editor, an associate editor, and the referees for their helpful comments that helped to improve an earlier version of this article. We are also grateful to Associate Professor Jianhui Zhou for providing the R code for the SCAD-GEE method.

Appendix. Proof of the theorems

In this Appendix, we will prove the main results stated in Section 3.

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

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

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

$$\begin{aligned} n^{-1/2}\boldsymbol{u}^T S_n\Big(\boldsymbol{\beta}_0 + n^{-1/2}\boldsymbol{u}\Big) &= n^{-1/2}\boldsymbol{u}^T S_n(\boldsymbol{\beta}_0) + n^{-1}\boldsymbol{u}^T \frac{\partial}{\partial \boldsymbol{\beta}} S_n(\widetilde{\boldsymbol{\beta}})\boldsymbol{u} \\ &=: I_{n1} + I_{n2}, \end{aligned}$$

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

$$\begin{aligned} I_{n1} &= n^{-1/2}\boldsymbol{u}^T(\boldsymbol{I}_p - \widehat{\boldsymbol{\Delta}})U(\boldsymbol{\beta}_0, \alpha) + n^{-1/2}\boldsymbol{u}^T(\boldsymbol{I}_p - \widehat{\boldsymbol{\Delta}})[U(\boldsymbol{\beta}_0, \alpha^*(\boldsymbol{\beta}_0)) - U(\boldsymbol{\beta}_0, \alpha)] + n^{-1/2}\boldsymbol{u}^T\widehat{\boldsymbol{\Delta}}\boldsymbol{\beta}_0 \\ &=: I_{n11} + I_{n12} + I_{n13}. \end{aligned}$$

By the Cauchy–Schwarz inequality, we can derive that

$$\begin{aligned}|I_{n11}| &\le n^{-1/2}\|\boldsymbol{u}^T(\boldsymbol{I}_p-\widehat{\boldsymbol{\Delta}})\|\|U(\boldsymbol{\beta}_0,\alpha)\| \\ &\le n^{-1/2}(1-\min_{j\in\mathcal{A}}\hat{\delta}_j(\lambda,\gamma))\|\boldsymbol{u}\|\|U(\boldsymbol{\beta}_0,\alpha)\|.\end{aligned} \tag{A.2}$$

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

$$\begin{aligned}P(\hat{\delta}_j(\lambda,\gamma)>n^{-1/2}\varepsilon) &= P\Big(\lambda/|\hat{\beta}_j^{(0)}|^{1+\gamma}>n^{-1/2}\varepsilon\Big)=P\Big((\lambda n^{1/2}/\varepsilon)^{1/(1+\gamma)}>|\hat{\beta}_j^{(0)}|\Big) \\ &\le P\Big((\lambda n^{1/2}/\varepsilon)^{1/(1+\gamma)}>\min_{j\in\mathcal{A}_0}|\beta_{0j}|-O_P(n^{-1/2})\Big)\to 0,\end{aligned} \tag{A.3}$$

which implies that $\hat{\delta}_j(\lambda,\gamma)=o_P(n^{-1/2})$ for each $j\in\mathcal{A}_0$. Therefore, we have that $\min_{j\in\mathcal{A}}\hat{\delta}_j(\lambda,\gamma)=o_P(n^{-1/2})$. By this, (A.2)–(A.3), and similar to the proof of Theorem 3.6 in Wang (2011), we can obtain that $|I_{n11}|=O_P(1)\|\boldsymbol{u}\|-o_P(n^{-1/2})\|\boldsymbol{u}\|$. For I_{n12}, using the conditions (1)–(3) and Taylor expansion for fixed $\boldsymbol{\beta}_0$, we have

$$\begin{aligned}U(\boldsymbol{\beta}_0,\alpha^*(\boldsymbol{\beta}_0))-U(\boldsymbol{\beta}_0,\alpha) &= \frac{\partial}{\partial\alpha}U(\boldsymbol{\beta}_0,\alpha)(\alpha^*-\alpha)+o_P(1) \\ &= \frac{\partial}{\partial\alpha}U(\boldsymbol{\beta}_0,\alpha)[\alpha(\boldsymbol{\beta}_0,\hat{\phi}(\boldsymbol{\beta}_0))-\hat{\alpha}(\boldsymbol{\beta}_0,\phi)+\hat{\alpha}(\boldsymbol{\beta}_{\mathcal{A}_0},\phi)-\alpha]+o_P(1) \\ &= \frac{\partial}{\partial\alpha}U(\boldsymbol{\beta}_0,\alpha)\Big[\frac{\partial\hat{\alpha}(\boldsymbol{\beta}_0,\phi^*)}{\partial\phi}(\hat{\phi}-\phi)+\hat{\alpha}(\boldsymbol{\beta}_{\mathcal{A}_0},\phi)-\alpha\Big]+o_P(1)=o_P(1),\end{aligned}$$

where ϕ^* lies between ϕ and $\hat{\phi}$. By the above result and using the similar argument of I_{n11}, we obtain that $|I_{n12}|=o_P(n^{-1/2})\|\boldsymbol{u}\|$. Since $\hat{\delta}_j=\min\{1,\lambda/|\hat{\beta}_j^0|^{1+\gamma}\}$, we have $|I_{n13}|\le n^{-1/2}\|\boldsymbol{u}\|\|\boldsymbol{\beta}_0\|=O_P(n^{-1/2})\|\boldsymbol{u}\|$. Hence $|I_{n1}|=O_P(1)\|\boldsymbol{u}\|$. Now consider I_{n2}, we can derive that

$$\begin{aligned}I_{n2} &= n^{-1}\boldsymbol{u}^T\frac{\partial}{\partial\boldsymbol{\beta}}S_n(\widetilde{\boldsymbol{\beta}})\boldsymbol{u} \\ &= \boldsymbol{u}^T(\boldsymbol{I}_p-\widehat{\boldsymbol{\Delta}})\Big[\frac{1}{n}\sum_{i=1}^n\boldsymbol{D}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{D}_i\Big]\boldsymbol{u}+n^{-1}\boldsymbol{u}^T\frac{\partial}{\partial\boldsymbol{\beta}}[U(\widetilde{\boldsymbol{\beta}},\alpha^*(\widetilde{\boldsymbol{\beta}}))-U(\widetilde{\boldsymbol{\beta}},\alpha)]\boldsymbol{u}+n^{-1}\boldsymbol{u}^T\widehat{\boldsymbol{\Delta}}\boldsymbol{u} \\ &=: I_{n21}+I_{n22}+I_{n23}.\end{aligned}$$

Using the above same argument, it is easy to show that $I_{n22}=o_P(1)\|\boldsymbol{u}\|^2$ and $|I_{n23}|=O_P(n^{-1})\|\boldsymbol{u}\|^2$. Thus, for sufficiently large n, $n^{-1/2}\boldsymbol{u}^TS_n(\boldsymbol{\beta}_0+n^{-1/2}\boldsymbol{u})$ on $\{\boldsymbol{\beta}_0+n^{-1/2}\boldsymbol{u}:\|\boldsymbol{u}\|=C\}$ is asymptotically dominated in probability by I_{n21}, which is positive for the sufficiently large C. □

Proof of Theorem 2. By Theorem 2 in Liang and Zeger (1986), it is known that the initial estimator $\hat{\boldsymbol{\beta}}^{(0)}$ obtained by solving the GEE $U\{\boldsymbol{\beta},\hat{\alpha}[\boldsymbol{\beta},\hat{\phi}(\boldsymbol{\beta})]\}=\boldsymbol{0}$ is $\sqrt{n}$-consistent. Note that $n^{(1+\gamma)/2}\lambda\to\infty$, we can derive that

$$\sum_{j\in\mathcal{A}_0^c}P\Big(\lambda/|\hat{\beta}_j^{(0)}|^{1+\gamma}<1\Big)\le\lambda^{-1}O(n^{-(1+\gamma)/2})\to 0, \tag{A.4}$$

which implies that

$$P(\hat{\delta}_j=1\quad\text{for all } j\in\mathcal{A}_0^c)\to 1. \tag{A.5}$$

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

$$\begin{aligned}P(\hat{\delta}_j>n^{-1/2}\varepsilon) &= P\Big(\lambda/|\hat{\beta}_j^{(0)}|^{1+\gamma}>n^{-1/2}\varepsilon\Big)=P\Big((\lambda n^{1/2}/\varepsilon)^{1/(1+\gamma)}>|\hat{\beta}_j^{(0)}|\Big) \\ &\le P\Big((\lambda n^{1/2}/\varepsilon)^{1/(1+\gamma)}>\min_{j\in\mathcal{A}_0}|\beta_{0j}|-O_P(n^{-1/2})\Big)\to 0,\end{aligned} \tag{A.6}$$

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

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

$$(\boldsymbol{I}_{|\mathcal{A}_0|}-\widehat{\boldsymbol{\Delta}}_{\mathcal{A}_0})U\{\hat{\boldsymbol{\beta}}_{\mathcal{A}_0},\hat{\alpha}[\hat{\boldsymbol{\beta}}_{\mathcal{A}_0},\hat{\phi}(\hat{\boldsymbol{\beta}}_{\mathcal{A}_0})]\}+\widehat{\boldsymbol{\Delta}}_{\mathcal{A}_0}\hat{\boldsymbol{\beta}}_{\mathcal{A}_0}=0. \tag{A.7}$$

Let $U(\boldsymbol{\beta},\alpha)=\sum_{i=1}^n U_i(\boldsymbol{\beta},\alpha)$ and $\alpha^*(\boldsymbol{\beta})=\hat{\alpha}\{\boldsymbol{\beta},\hat{\phi}(\boldsymbol{\beta})\}$, under some regularity conditions, and applying a Taylor expansion to (A.7), it is easy to show that $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{A}_0}-\boldsymbol{\beta}_{\mathcal{A}_0})$ can be approximated by

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

where

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol{\beta}_{\mathcal{A}_0}}\Big(U_i\{\boldsymbol{\beta}_{\mathcal{A}_0},\alpha^*(\boldsymbol{\beta}_{\mathcal{A}_0})\}\Big)&=\frac{\partial U_i\{\boldsymbol{\beta}_{\mathcal{A}_0},\alpha^*(\boldsymbol{\beta}_{\mathcal{A}_0})\}}{\partial\boldsymbol{\beta}_{\mathcal{A}_0}}+\frac{\partial U_i\{\boldsymbol{\beta}_{\mathcal{A}_0},\alpha^*(\boldsymbol{\beta}_{\mathcal{A}_0})\}}{\partial\alpha^*}\frac{\partial\alpha^*(\boldsymbol{\beta}_{\mathcal{A}_0})}{\partial\boldsymbol{\beta}_{\mathcal{A}_0}}\\&=:I_i+J_iK.\end{aligned}\tag{A.8}$$

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

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n U_i\{\boldsymbol{\beta}_{\mathcal{A}_0},\alpha^*(\boldsymbol{\beta}_{\mathcal{A}_0})\}=:I^*+J^*K^*+o_P(1),\tag{A.9}$$

where $I^*=\frac{1}{\sqrt{n}}\sum_{i=1}^n U_i\{\boldsymbol{\beta}_{\mathcal{A}_0},\alpha\}$, $J^*=\frac{1}{n}\sum_{i=1}^n\frac{\partial U_i(\boldsymbol{\beta}_{\mathcal{A}_0},\alpha)}{\partial\alpha}$ and $K^*=\sqrt{n}[\alpha^*(\boldsymbol{\beta}_{\mathcal{A}_0})-\alpha]$. Note that $\partial U_i(\boldsymbol{\beta}_{\mathcal{A}_0},\alpha)/\partial\alpha$ is a linear function of $\boldsymbol{Y}_i-\boldsymbol{\mu}_i(\boldsymbol{\beta}_{\mathcal{A}_0})$ and $E(\boldsymbol{Y}_i-\boldsymbol{\mu}_i(\boldsymbol{\beta}_{\mathcal{A}_0}))=0$, it is easy to prove that $J^*=o_P(1)$. By conditions (1)–(3), we have

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

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

$$\begin{aligned}\left\|\frac{1}{\sqrt{n}}\hat{G}_{\mathcal{A}_0}\boldsymbol{\beta}_{\mathcal{A}_0}\right\|^2&\le\frac{1}{n\Big\{1-\max\limits_{j\in\mathcal{A}_0}\hat{\delta}_j(\lambda,\gamma)\Big\}^2}\sum_{j\in\mathcal{A}_0}\frac{(\lambda\beta_j)^2}{\hat{\beta}_j^{(0)2(1+\gamma)}}\\&=\frac{\lambda^2}{n\Big\{1-\max\limits_{j\in\mathcal{A}_0}\hat{\delta}_j(\lambda,\gamma)\Big\}^2}\sum_{j\in\mathcal{A}_0}\left|\hat{\beta}_j^{(0)(-\gamma)}+(\beta_j-\hat{\beta}_j^{(0)})\hat{\beta}_j^{(0)(-\gamma-1)}\right|^2\\&=O_P(n^{-1}\lambda^2)\sum_{j\in\mathcal{A}_0}(2|\hat{\beta}_j^{(0)}|^{-2\gamma}+2|(\beta_j-\hat{\beta}_j^{(0)})\hat{\beta}_j^{(0)(-\gamma-1)}|^2)\\&\le O_P(n^{-1}\lambda^2)\Big(2s\min_{j\in\mathcal{A}_0}|\hat{\beta}_j^{(0)}|^{-2\gamma}+2\min_{j\in\mathcal{A}_0}|\hat{\beta}_j^{(0)}|^{-2\gamma-2}\|\beta_{\mathcal{A}_0}-\hat{\beta}_{\mathcal{A}_0}^{(0)}\|^2\Big)\\&=O_P\Big((\sqrt{n}\lambda)^2n^{-2}\tau^{-2\gamma}s\Big)(1+O_P(\tau^{-2}n^{-1}))=o_P(n^{-2}),\end{aligned}$$

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

$$\left\|\frac{1}{n}\hat{G}_{\mathcal{A}_0}\right\|^2=O_P\Big((\sqrt{n}\lambda)^2n^{-3}\tau^{-2\gamma-2}\Big)=o_P(n^{-3}).$$

Similarly, it is easy to show that $\sum_{i=1}^n J_i=o_P(n)$ and $K=O_P(1)$. Note that $\boldsymbol{D}_{i,\mathcal{A}_0}(\boldsymbol{\beta}_{\mathcal{A}_0})=-\partial\mu_i(\beta_{\mathcal{A}_0})/\partial\beta_{\mathcal{A}_0}$, thus we can prove that $-\frac{1}{\sqrt{n}}\sum_{i=1}^n U_i\{\boldsymbol{\beta}_{\mathcal{A}_0},\alpha^*(\boldsymbol{\beta}_{\mathcal{A}_0})\}$ is asymptotically equivalent to $-I^*$ whose asymptotic distribution is multivariate Gaussian with mean zero and covariance matrix

$$\lim_{n\to\infty}\left\{\frac{1}{n}\sum_{i=1}^n\boldsymbol{D}_{i,\mathcal{A}_0}^T\boldsymbol{V}_i^{-1}\mathrm{Cov}(\boldsymbol{Y}_i)\boldsymbol{V}_i^{-1}\boldsymbol{D}_{i,\mathcal{A}_0}\right\}.$$

Moreover, as $n\to\infty$, $\frac{1}{n}\sum_{i=1}^n I_i\to\frac{1}{n}\sum_{i=1}^n\boldsymbol{D}_{i,\mathcal{A}_0}^T\boldsymbol{V}_i^{-1}\boldsymbol{D}_{i,\mathcal{A}_0}$. We complete the proof of (ii). □

References

Agresti, A., 2002. Categorical Data Analysis, 2nd ed. Wiley Interscience, Hoboken, NJ.
Balan, R.M., Schiopu-Kratina, I., 2005. Asymptotic results with generalized estimating equations for longitudinal data. Ann. Statist. 32, 522–541.
Cantoni, E., Mills Flemming, J., Ronchetti, E., 2005. Variable selection for marginal longitudinal generalized linear models. Biometrics 61 (2), 507–514.
Chiou, J.M., Müller, H.G., 2005. Estimated estimating equations: semiparametric inference for clustered and longitudinal data. J. Roy. Statist. Soc. B 67 (4), 531–553.
Dziak, J.J., 2006. Penalized quadratic inference functions for variable selection in longitudinal research. Ph.D Thesis, the Pennsylvania State University (https://etda.libraries.psu.edu/paper/7084/).
Fan, J.Q., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96, 1348–1360.
Frank, I.E., Friedman, J.H., 1993. A statistical view of some chemometrics regression tools (with discussion). Technometrics 35, 109–148.
Fu, W.J., 2003. Penalized estimating equations. Biometrics 59, 126–132.
Liang, K.-Y., Zeger, S.L., 1986. Longitudinal data analysis using generalized linear models. Biometrika 73, 13–22.
McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, 2nd ed. Chapman & Hall, London.
Oman, S.D., 2009. Easily simulated multivariate binary distributions with given positive and negative correlations. Comput. Statis. Data Anal. 53, 999–1005.
Pan, W., 2001. Akaike's information criterion in generalized estimating equations. Biometrics 57, 120–125.
Petkau, A.J., White, R.A., 2003. Statistical approaches to assessing the effects of neutralizing antibodies: IFNbeta-1b in the pivotal trial of relapsing-remitting multiple sclerosis. Neurology 61, 35–37.
Petkau, A.J., White, R.A., Ebers, G.C., Reder, A.T., Sibley, W.A., Lublin, F.D., Paty, D.W., 2004. Longitudinal analyses of the effects of neutralizing antibodies on interferon beta-1b in relapsing-remitting multiple sclerosis. Mult. Sclerosis 10, 126–138.
Qu, A., Lindsay, B.G., Li, B., 2000. Improving generalised estimating equations using quadratic inference functions. Biometrika 87, 823–836.
Song, P.X.-K., 2007. Correlated Data Analysis: Modeling, Analytics and Applications. Springer, New York.
Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. Ser. B 58, 267–288.
Ueki, M., 2009. A note on automatic variable selection using smooth-threshold estimating equations. Biometrika 96 (4), 1005–1011.
Wang, L., 2011. GEE analysis of clustered binary data with diverging number of covariates. Ann. Statist. 39, 389–417.
Wang, L., Qu, A., 2009. Consistent model selection and data-driven smooth tests for longitudinal data in the estimating equations approach. J. R. Statist. Soc. B 71, 177–190.
Wang, L., Zhou, J., Qu, A., 2012. Penalized generalized estimating equations for high-dimensional longitudinal data analysis. Biometrics 68 (2), 353–360.
Xu, P.R., Fu, W., Zhu, L.X., 2012. Shrinkage estimation analysis of correlated binary data with a diverging number of parameters. Sci. China Ser. A: Mathematics (in press).
Xu, P.R., Zhu, L.X., 2010. Sure independence screening for marginal longitudinal generalized linear models (unpublished manuscript).
Zeger, S.L., Liang, K.-Y., 1986. Longitudinal data analysis for discrete and continuous outcomes. Biometrics 42, 121–130.
Zou, H., 2006. The adaptive Lasso and its oracle properties. J. Amer. Statist. Assoc. 101, 1418–1429.
Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. Roy. Statist. Soc. B 67, 301–320.