Open Journal of Statistics, **2014, 4, 38-48**
Published Online February 2014 (http://www.scirp.org/journal/ojs)
http://dx.doi.org/10.4236/ojs.2014.41005

Automatic Variable Selection for High-Dimensional Linear Models with Longitudinal Data

Ruiqin Tian, Liugen Xue
College of Applied Sciences, Beijing University of Technology, Beijing, China
Email: tianruiqin@emails.bjut.edu.cn

Received November 6, 2013; revised December 6, 2013; accepted December 13, 2013

ABSTRACT

High-dimensional longitudinal data arise frequently in biomedical and genomic research. It is important to select relevant covariates when the dimension of the parameters diverges as the sample size increases. We consider the problem of variable selection in high-dimensional linear models with longitudinal data. A new variable selection procedure is proposed using the smooth-threshold generalized estimating equation and quadratic inference functions (SGEE-QIF) to incorporate correlation information. 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-QIF. The proposed procedure avoids the convex optimization problem and is flexible and easy to implement. We establish the asymptotic properties in a high-dimensional framework where the number of covariates p_n increases as the number of cluster n increases. Extensive Monte Carlo simulation studies are conducted to examine the finite sample performance of the proposed variable selection procedure.

KEYWORDS

Variable Selection; Diverging Number of Parameters; Longitudinal Data; Quadratic Inference Functions; Generalized Estimating Equation

1. Introduction

Longitudinal data arise frequently in biomedical and health studies in which repeated measurements form the same subject are correlated. A major aspect of longitudinal data is the within subject correlation among the repeated measurements. Ignoring this within subject correlation causes a loss of efficiency in general problems. One of the commonly used regression methods for analyzing longitudinal data is generalized estimating equations (Liang and Zeger, [1]). Generalized estimating equations (GEE) using a working correlation matrix with nuisance parameters estimate regression parameters consistently even when the correlation structure is misspecified. However, under such misspecification, the estimator can be inefficient. For this reason, Qu *et al.* [2] proposed a method of quadratic inference functions (QIF). It avoids estimating the nuisance correlation structure parameters by assuming that the inverse of working correlation matrix can be approximated by a linear combination of several known basis matrices. The QIF can efficiently take the within subject correlation into account and is more efficient than the GEE approach when the working correlation is misspecified. The QIF estimator is also more robust against contamination when there are outlying observations (Qu and Song, [3]).

High-dimensional longitudinal data, which consist of repeated measurements on a large number of covariates, arise frequently from health and medical studies. Thus, it is important for statisticians to develop a new statistical methodology and theory of variable selection and estimation for high-dimensional longitudinal data, which can reduce the modeling bias. Generally speaking, most of the variable selection procedures are based on pena-

lized estimation using penalty functions. Such as, L_q penalty (Frank and Friedman, [4]), LASSO penalty (Tibshirani, [5]), SCAD penalty (Fan and Li, [6]), and so on. In the longitudinal data framework, Pan [7] proposed an extension of the Akaike information criterion (Akaike, [8]) by applying the quasi-likelihood to the GEE, assuming independent working correlation. Wang and Qu [9] developed a Bayesian information type of criterion (Schwarz, [10]) based on the quadratic inference functions. Fu [11] applied the bridge penalty model to the GEE and Xu *et al.* [12] introduced the adaptive lasso for the GEE setting, and references therein. These methods are able to perform variable selection and parameter estimation simultaneously. However, most of the theory and implementation is restricted to a fixed dimension of parameters.

Despite the importance of variable selection in high-dimensional settings (Fan and Li, [13]; Fan and Lv, [14]), variable selection for longitudinal data that take into consideration the correlation information is not well studied when the dimension of parameters diverges. In this paper we use the smooth-threshold generalized estimating equation based on quadratic inference functions (SGEE-QIF) to the high-dimensional longitudinal data. The proposed procedure automatically eliminates the irrelevant parameters by setting them as zero, and simultaneously estimates the nonzero regression coefficients by solving the SGEE-QIF. Compared to the shrinkage methods and the existing research findings reviewed above, our method offers the following improvements: 1) the proposed procedure avoids the convex optimization problem; 2) the proposed SGEE-QIF approach is flexible and easy to implement; 3) the proposed method is easy to deal with the longitudinal correlation structure and extend the estimating equations approach to high-dimensional longitudinal data.

The rest of this paper is organized as follows. In Section 2 we first propose variable selection procedures for high-dimensional linear models with longitudinal data, and asymptotic properties of the resulting estimators. In Section 3 we give the computation of the estimators as well as the choice of the tuning parameters. In Section 4 we carry out simulation studies to assess the finite sample performance of the method. Some assumptions and the technical proofs of all asymptotic results are provided in the **Appendix**.

2. Automatic Variable Selection Procedure

2.1. Model and Notation

We consider a longitudinal study with n subjects and m_i observations over time for the i th subject $(i=1,\cdots,n)$ for a total of $N=\sum_{i=1}^{n} m_i$ observation. Each observation consists of a response variable Y_{ij} and a covariate vector $X_{ij}\in R^{p_n}$ taken from the i th subject at time t_{ij}. We assume that the full data set $\left\{\left(X_{ij},Y_{ij}\right), i=1,\cdots,n,\ j=1,\cdots,m_i\right\}$ is observed and can be modelled as

$$Y_{ij}=X_{ij}^{\mathrm{T}}\beta+\varepsilon_{ij}, i=1,\cdots,n,\ j=1,\cdots,m_i, \tag{2.1}$$

where β is a $p_n\times 1$ vector of unknown regression coefficients, and p_n diverging as the sample size increases. ε_{ij} is random error with $E\left(\varepsilon_{ij}\middle|X_{ij}\right)=0$. In addition, we give assumptions on the first two moments of the observations $\left\{Y_{ij}\right\}$, that is, $E\left(Y_{ij}\right)=X_{ij}^{\mathrm{T}}\beta$ and $\mathrm{Var}\left(Y_{ij}\right)=v\left(X_{ij}^{\mathrm{T}}\beta\right)$, where $v(\cdot)$ is a known variance function.

2.2. Quadratic Inference Functions

Denote $Y_i=\left(Y_{i1},\cdots,Y_{im_i}\right)^{\mathrm{T}}$, and write X_i in a similar fashion. Following Liang and Zeger [1], a GEE can be used to estimate the regression parameters, β,

$$\sum_{i=1}^{n} X_i^{\mathrm{T}}V_i^{-1}\left(Y_i-X_i\beta\right)=0, \tag{2.2}$$

where V_i is the covariance matrix of Y_i. The matrix V_i is often modelled as $A_i^{1/2}R(\alpha)A_i^{1/2}$, where A_i is a diagonal matrix representing the variance of Y_{ij}, that is, $A_i=\mathrm{diag}\left(\mathrm{Var}\left(Y_{i1}\right),\cdots,\mathrm{Var}\left(Y_{im_i}\right)\right)$, $R(\alpha)$ is a common working correlation depending on a set of unknown nuisance parameters α. Based on the estimation theory associated with the working correlation structure, the GEE estimator of the regression coefficient proposed by Liang and Zeger [1] is consistent if consistent estimators of the nuisance parameters α can be obtained. For suggested methods for estimating α, see Liang and Zeger [1]. However, even in some simple cases,

consistent estimators of α do not always exist (Crowder [15]). To avoid this drawback, Qu *et al.* [2] suggested that the inverse of the working correlation matrix, $R^{-1}(\alpha)$ is approximated by a linear combination of basis matrices, $M_i, i=1,\cdots,s$, such as

$$R^{-1}(\alpha) \approx \sum_{k=1}^{s} a_k M_k, \tag{2.3}$$

where $M_1,\cdots,M_s$ are known matrices, and $a_1,\cdots,a_s$ are unknown constants. This is a sufficiently rich class that accommodates, or at least approximates, the correlation structures most commonly used. Details of utilizing a linear combination of some basic matrices to model the inverse of working correlation can be found in Qu *et al.* [2].

Substituting (2.3) to (2.2), we get the following class of estimating functions:

$$\sum_{i=1}^{n} X_i^{\mathrm{T}} A_i^{-1/2} \left(a_1 M_1 + \cdots + a_s M_s\right) A_i^{-1/2} \left(Y_i - X_i\beta\right) = 0. \tag{2.4}$$

Instead of estimating parameters $a=(a_1,\cdots,a_s)^{\mathrm{T}}$ directly, they recognized that a GEE (2.4) is equivalent to solving the linear combination of a vector of estimating equations:

$$g_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} g_i(\beta), \tag{2.5}$$

where

$$g_i(\beta) = \begin{pmatrix} X_i^{\mathrm{T}} A_i^{-1/2} M_1 A_i^{-1/2} \left(Y_i - X_i\beta\right) \\ \vdots \\ X_i^{\mathrm{T}} A_i^{-1/2} M_s A_i^{-1/2} \left(Y_i - X_i\beta\right) \end{pmatrix}$$

However, (2.5) does not work because the dimension of $g_n(\beta)$ is obviously greater than the number of unknown parameters. Using the idea of generalized method of moments (Hansen, [16]), Qu *et al.* [2] defined the quadratic inference functions (QIF),

$$Q_n(\beta) = g_n^{\mathrm{T}}(\beta)\Omega_n^{-1}(\beta) g_n(\beta) \tag{2.6}$$

where

$$\Omega_n(\beta) = \frac{1}{n^2}\sum_{i=1}^{n} g_i(\beta) g_i^{\mathrm{T}}(\beta)$$

Note that Ω_n depends on β. The QIF estimate $\hat{\beta}_n$ is then given by

$$\hat{\beta}_n = \arg\min_{\beta} Q_n(\beta).$$

Then, based on (2.6), according to Qu *et al.* [2], the corresponding estimating equation for β is

$$U_n(\beta) = n\dot{g}_n^{\mathrm{T}}\Omega_n^{-1} g_n, \tag{2.7}$$

where $\dot{g}_n$ is the $sp_n \times p_n$ matrix $\left\{\partial g_n / \partial \beta^{\mathrm{T}}\right\}$.

2.3. Smooth-Threshold Generalized Estimating Equations Based on QIF

Variable selection is an important topic in high dimensional regression analysis and most of the variable selection procedures are based on penalized estimation using penalty functions. Because of these variable selection procedures using penalty function have a singularity at zero. So, these procedure require convex optimization, which incurs a computational burden. To overcome this problem, Ueki [17] developed a automatic variable selection procedure that can automatically eliminate irrelevant parameters by setting them as zero. The method is easily implemented without solving any convex optimization problems. Motivated by this idea we propose the following smooth-threshold generalized estimating equations based on quadratic inference functions (SGEE-QIF)

$$\left(I_{p_n} - \Delta\right) U_n(\beta) + \Delta\beta = 0, \tag{2.8}$$

where Δ is the diagonal matrix whose diagonal elements are $\delta=\left(\delta_j\right)_{j=1,\cdots,p_n}$, and I_{p_n} is the p_n dimensional identity matrix. Note that the j th SGEE-QIF with $\delta_j=1$ reduces to $\beta_j=0$. Therefore, SGEE-QIF (2.8) can yield a sparse solution. Unfortunately, we cannot directly obtain the estimator of β by solving (2.8). This is because the SGEE-QIF involves δ_j, which need to be chosen using some data-driven criteria. For the choice of $\delta=\left(\delta_j\right)_{j=1,\cdots,p_n}$, Ueki [17] suggested that δ_j may be determined by the data, and can be chosen by $\hat{\delta}_j=\min\left(1,\lambda\Big/\left|\hat{\beta}_j^{(0)}\right|^{1+\gamma}\right)$ with an initial estimator $\hat{\beta}_j^{(0)}$. The initial estimator $\hat{\beta}_j^{(0)}$ can be obtained by solving the QIF (2.6) for the full model. Note that this choice involves two tuning parameters (λ,γ). In Section 4, following the idea of Ueki [17], we use the BIC-type criterion to select the tuning parameters.

Replacing Δ in (2.8) by $\hat{\Delta}$ with diagonal elements $\hat{\delta}=\left(\hat{\delta}_j\right)_{j=1,\cdots,p_n}$. The SGEE-QIF becomes

$$\left(I_{p_n}-\hat{\Delta}\right)U_n(\beta)+\hat{\Delta}\beta=0. \tag{2.9}$$

The solution of (2.9) denoted by $\hat{\beta}_{\lambda,\gamma}$ is called the SGEE-QIF estimator.

2.4. Asymptotic Properties

We next study the asymptotic properties of the smooth-threshold estimator. Let β_0 be the fixed true value of β. Denote $\mathrm{A}_0=\left\{j:\beta_{0j}\neq 0\right\}$ and $\mathrm{A}_0^c=\left\{j:\beta_{0j}=0\right\}$. Denote by $s=\left|\mathrm{A}_0\right|$ the number of true nonzero parameters. s may be fixed or grow with n. We assume, under the regularity conditions, the initial QIF estimator $\hat{\beta}_0$ obtained by solving the QIF (2.6) satisfies $\left\|\hat{\beta}_0-\beta_0\right\|=O_p\left(\sqrt{\frac{p_n}{n}}\right)$ when $p_n\to\infty$ and $p_n^2n^{-1}=o(1)$, where β_0 is the true value of β. Following Fan and Peng [18] and Wang [19], it is possible to prove the oracle properties for the SGEE-QIF estimators, including $\sqrt{n/p_n}$ consistency, variable selection consistency and asymptotic normality.

To obtain the asymptotic properties in the paper, we require the following regularity conditions:

(C1). The parameter space S is compact, and β_0 is an interior point of S.

(C2). $E\varepsilon_i^4<\infty$, $EX_{ij(r)}<\infty$, $i=1,\cdots,n$, $j=1,\cdots,m_i$, $r=1,\cdots,p_n$, where $EX_{ij(r)}$ is the r th component of X_{ij}.

(C3). $\lim\limits_{n\to\infty}\frac{1}{n}\sum\limits_{i=1}^{n}E\begin{pmatrix}X_{i,\mathrm{A}_0}^{\mathrm{T}}A_i^{-1/2}M_1A_i^{-1/2}X_{i,\mathrm{A}_0}\\ \vdots\\ X_{i,\mathrm{A}_0}^{\mathrm{T}}A_i^{-1/2}M_sA_i^{-1/2}X_{i,\mathrm{A}_0}\end{pmatrix}\equiv J_0$.

(C4). The weighting matrix $\Omega_n(\beta)$ converges almost surely to a constant matrix Ω_0, where Ω_0 is invertible. Furthermore, the first and second partial derivatives of Ω_n in β are all $O_p(1)$.

(C5). $Q_n(\beta)$ is a twice differentiable function of β. Furthermore the third derivatives of $Q_n(\beta)$ are $O_p(n)$.

(C6). All the variance matrixes $A_i\geq 0$, and $\sup_i\left\|A_i\right\|<\infty$.

(C7). $\sup_{i,j}\left\|X_{ij}\right\|=O\left(\sqrt{p_n}\right)$.

(C8). $E\left\|M_n\left(\beta_{\mathrm{A}_0}\right)\dot{g}_n^{\mathrm{T}}\left(\beta_{\mathrm{A}_0}\right)\Omega_0^{-1}\left(\beta_{\mathrm{A}_0}\right)\right\|^2=O_p\left(p_n^2\right)$.

We define the active set $\mathrm{A}=\left\{j:\hat{\delta}_j\neq 1\right\}$ which is the set of indices of nonzero parameters, where $\hat{\delta}_j=\min\left(1,\lambda\Big/\left|\hat{\beta}_j^{(0)}\right|^{1+\gamma}\right),\ j=1,\cdots,p_n$. The following theorem gives the consistency of the SGEE-QIF estimators.

Theorem 1. Under conditions C1-C8, for any positive λ and γ, such that $p_n^2n^{-1}=o(1)$, $n^{1/2}\lambda\to 0$ and $\left(n/p_n\right)^{(1+\gamma)/2}\lambda\to\infty$, as $n\to\infty$. There exists a sequence $\hat{\beta}$ of the solutions of (2.9) such that

$$\left\|\hat{\beta}-\beta_0\right\| = O_p\left(\sqrt{\frac{p_n}{n}}\right).$$

Furthermore, we show that such consistent estimators must possess the sparsity property and the estimators for nonzero coefficients have the same asymptotic distribution as that based on the correct submodel.

Theorem 2. Suppose that the conditions of Theorem 1 hold, if $n^{-1}p_n^3 = o(1)$, as $n \to \infty$, we have

1) Variable selection consistency, *i.e.*

$$P\left(A = A_0\right) \to 1$$

2) Asymptotic normality, *i.e.*

$$\sqrt{n}A_n M_n^{-1/2}\left(\beta_{A_0}\right) J_0\left(\hat{\beta}_{A_0} - \beta_{A_0}\right) \xrightarrow{L} N\left(0, G\right)$$

where A_n is a $q \times s$ matrix such that $A_n A_n^{\mathrm{T}} \to G$, G is a $q \times q$ non-negative symmetric matrix, $M_n\left(\beta_{A_0}\right) = J_0^{\mathrm{T}}\Omega_0^{-1}\left(\beta_{A_0}\right)J_0$, and "$\xrightarrow{L}$" represents the convergence in distribution.

Remark: Theorems 1 and 2 imply that the proposed SGEE-QIF procedure is consistent in variable selection, it can identify the zero coefficients with probability tending to 1. By choosing appropriate tuning parameters, the SGEE-QIF estimators have the oracle property; that is, the asymptotic variances for the SGEE-QIF estimators are the same as what we would have if we knew in advance the correct submodel.

3. Computation

3.1. Algorithm

Next, we propose the iterative algorithm to implement the procedures as follows:

Step 1, Calculate the initial estimates $\hat{\beta}^{(0)}$ of β by solving the initial QIF (2.5) estimator. Let $k = 0$.

Step 2, By using the current estimate $\hat{\beta}^{(k)}$, we choose the tuning parameters (λ, γ) by the BIC criterion.

Step 3, Update the estimator of β as follows:

$$\hat{\beta}_A^{(k+1)} = \hat{\beta}_A^{(k)} - \left\{n\dot{g}_{n,A}^{\mathrm{T}}\left(\hat{\beta}_A^{(k)}\right)\Omega_n^{-1}\left(\hat{\beta}_A^{(k)}\right)\dot{g}_{n,A}\left(\hat{\beta}_A^{(k)}\right) + \hat{G}_A\right\}^{-1} \times \left\{\dot{g}_{n,A}^{\mathrm{T}}\left(\hat{\beta}_A^{(k)}\right)\Omega_n^{-1}\left(\hat{\beta}_A^{(k)}\right)g_{n,A}\left(\hat{\beta}_A^{(k)}\right) + \hat{G}_A\hat{\beta}_A^{(k)}\right\},$$

$$\hat{\beta}_{A^c,\hat{\Delta}} = 0$$

where $g_{n,A}\left(\beta_A\right) = \dfrac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} X_{i,A}^{\mathrm{T}}A_i^{-1/2}M_1A_i^{-1/2}\left(Y_i - X_{i,A}\beta_A\right) \\ \vdots \\ X_{i,A}^{\mathrm{T}}A_i^{-1/2}M_sA_i^{-1/2}\left(Y_i - X_{i,A}\beta_A\right)\end{pmatrix}$

and $\hat{G}_A = \left(I_{|A|} - \hat{\Delta}_A\right)^{-1}\hat{\Delta}_A$.

Step 4, Iterate Steps 2-3 until convergence, and denote the final estimators of β as the SGEE-QIF estimator.

3.2. Choosing the Tuning Parameters

To implement the procedures described above, we need to choose the tuning parameters (λ, γ). Following Ueki [17], we use BIC-type criterion to choose these two parameters. That is, we choose (λ, γ) as the minimizer of

$$BIC_{\lambda,\gamma} = Q_n\left(\hat{\beta}_{\lambda,\gamma}\right) + df_{\lambda,\gamma}\log(n),$$

where $\hat{\beta}_{\lambda,\gamma}$ is the SGEE-QIF estimator for given (λ, γ), $df_{\lambda,\gamma}$ is simply the number of nonzero coefficient $\hat{\beta}$. The selected (λ, γ) minimizes the $BIC_{\lambda,\gamma}$.

3.3. Choosing the Basis Matrices

The choice of basis matrices M_k in (2.2) is not difficult, especially for those special correlation structures which are frequently used. If we assume $R(\alpha)$ is the first-order autoregressive correlation matrix. The exact inversion $R^{-1}(\alpha)$ can be written as a linear combination of three basis matrices, they are M_1, M_2 and M_3,

where M_1 is the identity matrix, M_2 has 1 on two main off-diagonals and 0 elsewhere, and M_3 has 1 on the corners (1, 1) and (m, m), and 0 elsewhere. Suppose $R(\alpha)$ is an exchangeable working correlation matrix, it has 1 on the diagonal, and α everywhere off the diagonal. Then $R^{-1}(\alpha)$ can be written as a linear combination of two basis matrices, M_1 is the identity matrix and M_2 is a matrix with 0 on the diagonal and 1 off the diagonal. More details about choosing the basis matrices can be seen in Zhou and Qu [20].

4. Simulation Studies

In this section we conduct a simulation study to assess the finite sample performance of the proposed procedures. In the simulation study, the performance of estimator $\hat{\beta}$ will be assessed by using the average the mean square error (AMSE), defined as $\left\|\hat{\beta}-\beta_0\right\|^2$ averaged over 500 times simulated data sets.

We simulate data from the model (1.1), where $\beta_0=\left(\beta_1,\cdots,\beta_{p_n}\right)^{\mathrm{T}}$ with $\beta_1=2.8$, $\beta_2=-1.8$ and $\beta_3=3.8$. While the remaining coefficients, corresponding to the irrelevant variables, are given by zeros. In addition, let $p_n=\lfloor 4n^{1/3}\rfloor$, where $\lfloor u\rfloor$ denotes the largest integer not greater than u. To perform this simulation, we take the covariates $X_{ij}\ (j=1,\cdots,5)$ from a multivariate normal distribution with mean zero, marginal variance 1 and correlation 0.5. The response variable Y_{ij} is generated according to the model. And error vector $\varepsilon_i=\left(\varepsilon_{i1},\cdots,\varepsilon_{i5}\right)^{\mathrm{T}}\sim N\left(0,\sigma^2\mathrm{Corr}\left(\varepsilon_i,\alpha\right)\right)$, where $\sigma^2=1$ and $\mathrm{Corr}\left(\varepsilon_i,\alpha\right)$ is a known correlation matrix with parameter α used to determine the strength of with-subject dependence. Here we consider ε_{ij} has the first-order autoregressive (AR-1) or compound symmetry (CS) correlation (*i.e.* exchangeable correlation) structure with $\alpha=0.7$. In the following simulations, we make 500 simulation runs and take $n=60$ and 120.

In the simulation study, for each simulated data set, we compare the estimation accuracy and model selection properties of the SGEE-QIF method, the SCAD-penalized QIF and the Lasso-penalized QIF for two different working correlations. For each of these methods, the average of zero coefficients over the 500 simulated data sets is reported in **Tables 1** and **2**. Note that "Correct" in tables means the average number of zero regression

Table 1. Variable selections for the parametric components under different methods when the correlation structure is correctly specified.

n	p_n	Method	CS			AR(1)		
			AMSE	Correct	Incorrect	AMSE	Correct	Incorrect
60	15	SGEE-QIF	0.0062	12.0000	0	0.0066	12.0000	0
		SCAD	0.0064	11.9020	0	0.0070	11.9280	0
		Lasso	0.0067	11.8520	0	0.0074	11.9080	0
120	19	SGEE-QIF	0.0028	16.0000	0	0.0029	16.0000	0
		SCAD	0.0028	16.0000	0	0.0030	16.0000	0
		Lasso	0.0031	16.0000	0	0.0033	15.9980	0

Table 2. Variable selections for the parametric components under different methods 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	p_n	Method	CS.AR(1)			AR(1).CS		
			AMSE	Correct	Incorrect	AMSE	Correct	Incorrect
60	15	SGEE-QIF	0.0083	12.0000	0	0.0094	12.0000	0
		SCAD	0.0087	11.8040	0	0.0094	11.9180	0
		Lasso	0.0092	11.7140	0	0.0099	11.8560	0
120	19	SGEE-QIF	0.0035	16.0000	0	0.0043	16.0000	0
		SCAD	0.0035	15.9960	0	0.0043	15.9980	0
		Lasso	0.0038	15.9880	0	0.0046	15.9960	0

coefficients that are correctly estimated as zero, and "Incorrect" depicts the average number of non-zero regression coefficients that are erroneously set to zero. At the same time, we also examine the effect of using a misspecified correlation structure in the model, which are also reported in **Tables 1** and **2**. From **Tables 1** and **2**, we can make the following observations.

1) For the parametric component, the performances of variable selection procedures become better and better as n increases. For example, the values in the column labeled "Correct" become more and more closer to the true number of zero regression coefficients in the models.

2) Compared with the penalized QIF based on Lasso and SCAD, SGEE-QIF performs satisfactory in terms of variable selection.

3) It is not surprised that the performances of variable selection procedures based on the correct correlation structure work better than based on the incorrect correlation structure. However, we also note that the performance does not significantly depend on working covariance structure.

5. Discussion

In this paper, we have proposed the smooth-threshold generalized estimating equation based on quadratic inference function (SGEE-QIF) to the high-dimensional longitudinal data. Our method incorporates the within subject correlation structure of the longitudinal to automatically eliminate the irrelevant parameters by setting them as zero, and simultaneously estimates the nonzero regression coefficients. As a future research topic, it is interesting to consider the variable selection for the high/ultra-high dimension varying coefficient models with longitudinal data.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (11171012, 11331011), the Science and technology project of the faculty adviser of excellent PHD degree thesis of Beijing (20111000503), the Beijing Municipal Education Commission Foundation (KM201110005029), and the Beijing municipal key disciplines (No.006000541213009).

REFERENCES

[1] K. L. Liang and S. L. Zeger, "Longitudinal Data Analysis Using Generalised Estimating Equations," *Biometrika*, Vol. 73, No. 1, 1986, pp. 13-22. http://dx.doi.org/10.1093/biomet/73.1.13

[2] A. Qu, B. G. Lindsay and B. Li, "Improving Generalized Estimating Equations Using Quadratic Inference Functions," *Biometrika*, Vol. 87, No. 4, 2000, pp. 823-836. http://dx.doi.org/10.1093/biomet/87.4.823

[3] A. Qu and P. X. K. Song, "Assessing Robustness of Generalized Estimating Equations and Quadratic Inference Functions," *Biometrika*, Vol. 91, No. 2, 2004, pp. 447-459. http://dx.doi.org/10.1093/biomet/91.2.447

[4] I. E. Frank and J. H. Friedman, "A Statistical View of Some Chemometrics Regression Tools (with Discussion)," *Technometrics*, Vol. 35, No. 2, 1993, pp. 109-148. http://dx.doi.org/10.1080/00401706.1993.10485033

[5] R. Tibshirani, "Regression Shrinkage and Selection via the LASSO," *Journal of Royal Statistical Society, Series B*, Vol. 58, No. 1, 1996, pp. 267-288.

[6] J. Q. Fan and R. Li, "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of American Statistical Association*, Vol. 96, No. 456, 2001, pp. 1348-1360. http://dx.doi.org/10.1198/016214501753382273

[7] W. Pan, "Akaike's Information Criterion in Generalized Estimating Equations," *Biometrics*, Vol. 57, No. 1, 2001, pp. 120-125. http://dx.doi.org/10.1111/j.0006-341X.2001.00120.x

[8] H. Akaike, "Information Theory and an Extension of the Maximum Likelihood Principle," *Proceedings of the* 2*nd International Symposium on Information Theory*, Budapest, 1973, pp. 267-281.

[9] L. Wang and A. Qu, "Consistent Model Selection and Data-Driven Smooth Tests for Longitudinal Data in the Estimating Equations Approach," *Journal of the Royal Statistical Society*: *Series B*, Vol. 71, No. 1, 2009, pp. 177-190. http://dx.doi.org/10.1111/j.1467-9868.2008.00679.x

[10] G. Schwarz, "Estimating the Dimension of a Model," *The Annals of Statistics*, Vol. 6, No. 2, 1978, pp. 461-464. http://dx.doi.org/10.1214/aos/1176344136

[11] W. J. Fu, "Penalized Estimating Equation," *Biometrics*, Vol. 59, No. 1, 2003, pp. 126-132. http://dx.doi.org/10.1111/1541-0420.00015

[12] P. R. Xu, W. Fu and L. X. Zhu, "Shrinkage Estimation Analysis of Correlated Binary Data with a Diverging Number of Para-

meters," *Science China Mathematics*, Vol. 56, No. 2, 2013, pp. 359-377. http://dx.doi.org/10.1007/s11425-012-4564-y

[13] J. Fan and R. Li, "Statistical Challenges with High Dimensionality: Feature Selection in Knowledge Discovery," *Proceedings of the International Congress of Mathematicians*, Vol. 3, European Mathematical Society, 2006, pp. 595-622.

[14] J. Fan and J. Lv, "A Selective Overview of Variable Selection in High-Dimensional Feature Space," *Statistica Sinica*, Vol. 20, No. 1, 2009, pp. 101-148.

[15] M. Crowder, "On the Use of a Working Correlation Matrix in Using Generalised Linear Models for Repeated Measures," *Biometrika*, Vol. 82, No. 2, 1995, pp. 407-410. http://dx.doi.org/10.1093/biomet/82.2.407

[16] L. Hansen, "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, Vol. 50, No. 4, 1982, pp. 1029-1054. http://dx.doi.org/10.2307/1912775

[17] M. Ueki, "A Note on Automatic Variable Selection Using Smooth-Threshold Estimating Equations," *Biometrika*, Vol. 96, No. 4, 2009, pp. 1005-1011. http://dx.doi.org/10.1093/biomet/asp060

[18] J. Fan and H. Peng, "Nonconcave Penalized Likelihood with a Diverging Number of Parameters," *The Annals of Statistics*, Vol. 32, No. 3, 2004, pp. 928-961. http://dx.doi.org/10.1214/009053604000000256

[19] L. Wang, "GEE Analysis of Clustered Binary Data with Diverging Number of Covariates," *The Annals of Statistics*, Vol. 39, No. 1, 2011, pp. 389-417. http://dx.doi.org/10.1214/10-AOS846

[20] J. Zhou and A. Qu, "Informative Estimation and Selection of Correlation Structure for Longitudinal Data," *Journal of the American Statistical Association*, Vol. 107, No. 498, 2012, pp. 701-710. http://dx.doi.org/10.1080/01621459.2012.682534

[21] J. J. Dziak, "Penalized Quadratic Inference Functions for Variable Selection in Longitudinal Research," Ph.D Thesis, The Pennsylvania State University, 2006. http://sites.stat.psu.edu/~jdziak/DziakDissert.pdf

Appendix. Proof of Theorems

Proof of Theorem 1

Let $S_n(\beta)=\left(I_{p_n}-\hat{\Delta}\right)U_n(\beta)+\hat{\Delta}\beta$. It suffices to prove for any $\varepsilon>0$, there is a constant $c>0$, such that

$$P\left(\sup_{\|u\|=c}\sqrt{p_n/n}u^{\mathrm{T}}S_n\left(\beta_0+\sqrt{p_n/n}u\right)>0\right)\geq 1-\varepsilon \tag{A.1}$$

for n enough. This will imply that there exists a local solution to the equation $S_n(\beta)=0$. Such that $\left\|\hat{\beta}-\beta_0\right\|=O_p\left(\sqrt{p_n/n}\right)$ with probability at least $1-\varepsilon$. The proof follows that of Theorem 3.6 in Wang [19], we will evaluate the sign of $\sqrt{p_n/n}u^{\mathrm{T}}S_n\left(\beta_0+\sqrt{p_n/n}u\right)$ in the ball of $\left\{\beta_0+\sqrt{p_n/n}u:\|u\|=c\right\}$. Note that

$$\sqrt{p_n/n}u^{\mathrm{T}}S_n\left(\beta_0+\sqrt{p_n/n}u\right)=\sqrt{p_n/n}u^{\mathrm{T}}S_n(\beta_0)+p_n/nu^{T}\frac{\partial}{\partial\beta}S_n\left(\tilde{\beta}\right)u\equiv I_{n1}+I_{n2}, \tag{A.2}$$

where $\tilde{\beta}$ lies between β_0 and $\beta_0+\sqrt{p_n/n}u$. Next we will consider I_{n1} and I_{n2} respectively. For I_{n1}, by Cauchy-Schwarz inequality, we can derive that

$$|I_{n1}|\leq\sqrt{p_n/n}\left\|u^{\mathrm{T}}\left(I_{p_n}-\hat{\Delta}\right)\right\|\left\|U_n(\beta_0)\right\|\leq\sqrt{p_n/n}\left(1-\min_{j\in A}\hat{\delta}_j(\lambda,\gamma)\right)\|u\|\left\|U_n(\beta_0)\right\|.$$

According to condition (C7), consider the k th $(k=1,\cdots,m)$ block of $g_n(\beta)$, $B_k=\frac{1}{n}\sum_{i=1}^{n}X_i^{\mathrm{T}}A_i^{-1/2}M_kA_i^{-1/2}\left(Y_i-X_i\beta\right)$, let $B_{k,s}$ denote the s th components B_k, by some elementary calculation, we obtain

$$E\left(B_{k,s}^2\right)=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{v=1}^{m_i}X_{ij,s}\sigma_i^{jv}\right)^2E\varepsilon_{ij}^2=O\left(p_n/n\right),$$

where σ_i^{jv} is the (j,v) th element of $A_i^{-1/2}M_kA_i^{-1/2}$, hence, we have $g_n(\beta)=O_p\left(\sqrt{p_n/n}\right)$. By C3 and C4, we obtain,

$$\left\|U_n(\beta_0)\right\|=\left\|\dot{g}_n(\beta)\Omega_n^{-1}(\beta)g_n(\beta)\right\|=O_p\left(\sqrt{p_n/n}\right). \tag{A.3}$$

Since $\min_{j\in A}\hat{\delta}_j(\lambda,\gamma)\leq\min_{j\in A_0}\hat{\delta}_j(\lambda,\gamma)$, we only need to obtain the convergence rate of $\min_{j\in A_0}\hat{\delta}_j(\lambda,\gamma)$. Assume that $\hat{\beta}_0$ is the initial estimator, and is $\sqrt{n/p_n}$-consistent. By using the condition $n^{1/2}\lambda\to 0$ for any $\varepsilon>0$ and $j\in A_0$, we have

$$P\left(\hat{\delta}_j(\lambda,\gamma)>n^{-1/2}\varepsilon\right)=P\left(\lambda\Big/\left|\hat{\beta}_j^{(0)}\right|^{1+\gamma}>n^{-1/2}\varepsilon\right)=P\left(\left(\lambda n^{1/2}/\varepsilon\right)^{1/(1+\gamma)}>\left|\hat{\beta}_j^{(0)}\right|\right)$$

$$\leq P\left(\left(\lambda n^{1/2}/\varepsilon\right)^{1/(1+\gamma)}>\min_{j\in A_0}\left|\beta_{oj}\right|-O_p\left(\sqrt{p_n/n}\right)\right)\to 0,$$

which implies that $\hat{\delta}_j(\lambda,\gamma)=o_p\left(n^{-1/2}\right)$ for each $j\in A_0$. Therefore, we have that $\min_{j\in A}\hat{\delta}_j(\lambda,\gamma)=o_p\left(n^{-1/2}\right)$. Hence, by (A.3), we obtain that

$$|I_{n1}|=O_p\left(p_n/n\right)\|u\|-o_p\left(p_nn^{-3/2}\right)\|u\|. \tag{A.4}$$

Now, we consider I_{n2}, we can derive that

$$I_{n2}=\frac{p_n}{n}u^{\mathrm{T}}\frac{\partial}{\partial\beta}S_n\left(\tilde{\beta}\right)u=\frac{p_n}{n}u^{\mathrm{T}}\left(I_{p_n}-\hat{\Delta}\right)\dot{g}_n\Omega_n^{-1}\dot{g}_nu+\frac{p_n}{n}u^{\mathrm{T}}\hat{\Delta}u=I_{n21}+I_{n22}.$$

By above result, we get $|I_{n22}|=O_p\left(p_n/n\right)\|u\|^2$. Thus, it is easy to show that for sufficiently large n, $\sqrt{p_n/n}u^{\mathrm{T}}S_n\left(\beta_0+\sqrt{p_n/n}u\right)$ on the ball of $\left\{\beta_0+\sqrt{p_n/n}u:\|u\|=c\right\}$ is asymptotically dominated in probability

by I_{n21}, which is positive for the sufficiently large c. The proof of Theorem 1 is completed.

Proof of Theorem 2

According to Dziak [21], it is known that the initial estimator $\hat{\beta}^{(0)}$ obtained by solving the QIF is $\sqrt{n/p_n}$-consistent. For any given $j \in A_0^c$, by $(n/p_n)^{(1+\gamma)/2}\lambda \to \infty$, then,

$$\begin{aligned} &P\left(\lambda \Big/ \left|\hat{\beta}_j^{(0)}\right|^{1+\gamma} < 1\right) = p\left(\lambda^{2/(1+\gamma)} < \left|\hat{\beta}_j^{(0)}\right|^2\right) = P\left(\lambda^{2/(1+\gamma)} < \max_j \left|\hat{\beta}_j^{(0)}\right|^2\right) \le P\left(\lambda^{2/(1+\gamma)} < \left\|\hat{\beta}^{(0)}\right\|^2\right) \\ &= P\left(\lambda^{2/(1+\gamma)} n/p_n < O(1)\right) \to 0, \end{aligned} \tag{A.5}$$

which implies that

$$P\left(\hat{\delta}_j = 1 \text{ for all } j \in A_0^c\right) \to 1 \tag{A.6}$$

On the other hand, by the condition $n^{1/2}\lambda \to 0$, for any $\varepsilon > 0$, and $j \in A_0$, we have

$$\begin{aligned} &P\left(\hat{\delta}_j > n^{-1/2}\varepsilon\right) = P\left(\lambda \Big/ \left|\hat{\beta}_j^{(0)}\right|^{1+\gamma} > n^{-1/2}\varepsilon\right) = P\left(\left(\lambda n^{1/2}/\varepsilon\right)^{1/(1+\gamma)} > \left|\hat{\beta}_j^{(0)}\right|\right) \\ &\le P\left(\left(\lambda n^{1/2}/\varepsilon\right)^{1/(1+\gamma)} > \min_{j\in A_0}\left|\beta_{0j}\right| - O_p\left(\sqrt{p_n/n}\right)\right) \to 0, \end{aligned}$$

which implies that $\hat{\delta}_j = o_p\left(n^{-1/2}\right)$ for each $j \in A_0$. Therefore, we prove that $P\left(\hat{\delta}_j < 1 \text{ for all } j \in A_0\right) \to 1$.

Thus, we complete the proof of 1).

Next, we will prove 2). As shown in 1), $\hat{\beta}_j = 0$ for $j \in A_0$ with probability tending to 1. At the same time, with probability tending to 1, $\hat{\beta}_{A_0}$ satisfies the smooth threshold generalized estimating equations based on quadratic inference functions (SGEE-QIF)

$$\left(I_{|A_0|} - \hat{\Delta}_{A_0}\right)U_n\left(\hat{\beta}_{A_0}\right) + \hat{\Delta}_{A_0}\hat{\beta}_{A_0} = 0. \tag{A.7}$$

Applying a Taylor expansion to (A.7) at β_{A_0}, it easy to show that

$$o_p(1) = \frac{1}{\sqrt{n}}U_n\left(\beta_{A_0}\right) + \frac{1}{\sqrt{n}}\frac{\partial}{\partial\beta_{A_0}}U_n\left(\beta_{A_0}\right)\left(\hat{\beta}_{A_0} - \beta_{A_0}\right) + \frac{1}{\sqrt{n}}\hat{G}_{A_0}\hat{\beta}_{A_0}$$

where $\hat{G}_{A_0} = \left(I_{|A_0|} - \hat{\Delta}_{A_0}\right)^{-1}\hat{\Delta}_{A_0}$. Since, we have

$$\begin{aligned} &\left\|\frac{1}{\sqrt{n}}\hat{G}_{A_0}\hat{\beta}_{A_0}\right\| \le \frac{1}{n\left\{1 - \max_{j\in A_0}\hat{\delta}_j(\lambda,\gamma)\right\}^2}\sum_{j\in A_0}\frac{\left(\lambda\beta_j\right)^2}{\hat{\beta}_j^{(0)2(1+\gamma)}} = \frac{\lambda^2}{n\left\{1 - \max_{j\in A_0}\hat{\delta}_j(\lambda,\gamma)\right\}^2}\sum_{j\in A_0}\left|\hat{\beta}_j^{(0)(-\gamma)} + \left(\beta_j - \hat{\beta}_j^{(0)}\right)\hat{\beta}_j^{(0)(-\gamma-1)}\right|^2 \\ &= O_p\left(n^{-1}\lambda^2\right)\sum_{j\in A_0}\left(2\left|\hat{\beta}_j^{(0)}\right|^{-2\gamma} + 2\left|\left(\beta_j - \hat{\beta}_j^{(0)}\right)\hat{\beta}_j^{(0)(-\gamma-1)}\right|^2\right) \le O_p\left(n^{-1}\lambda^2\right)\left(2s\min_{j\in A_0}\left|\hat{\beta}_j^{(0)}\right|^{-2\gamma}\right. \\ &\left. + 2\min_{j\in A_0}\left|\hat{\beta}_j^{(0)}\right|^{-2\gamma-2}\left\|\beta_{A_0} - \hat{\beta}_{A_0}\right\|^2\right) = O_p\left(\left(\sqrt{n}\lambda\right)^2 n^{-2}\tau^{-2\gamma}s\right)\left(1 + O_p\left(\tau^{-2}p_n n^{-1}\right)\right) = o_p\left(n^{-2}\right), \end{aligned}$$

where $\tau = \min_{j\in A_0}\left|\hat{\beta}_j^{(0)}\right|$. Then, if $n^{-1/2}A_n M_n^{-1/2}\left(\beta_{A_0}\right)U_n\left(\beta_{A_0}\right) \xrightarrow{L} N(0,G)$ holds, by the Slutsky theorem, we can prove Theorem 2 (2). We write $n^{-1/2}A_n M_n^{-1/2}\left(\beta_{A_0}\right)U_n\left(\beta_{A_0}\right) = \sum_{i=1}^{n} Z_{ni}$,

where

$$Z_{ni} = \frac{1}{\sqrt{n}}A_n M_n^{-1/2}\left(\beta_{A_0}\right)\dot{g}_n^{\mathrm{T}}\left(\beta_{A_0}\right)\Omega_0^{-1}\left(\beta_{A_0}\right)g_n\left(\beta_{A_0}\right).$$

Since $M_n\left(\beta_{A_0}\right)=\mathrm{Cov}\left(\dot{g}_n^{\mathrm{T}}\left(\beta_{A_0}\right)\Omega_0^{-1}\left(\beta_{A_0}\right)g_n\left(\beta_{A_0}\right)\right)$, we have

$$\mathrm{Cov}\left(\sum_{i=1}^{n}Z_{ni}\right)=C\,\mathrm{ov}\left(\frac{1}{\sqrt{n}}A_nM_n^{-1/2}\left(\beta_{A_0}\right)g_n^{\mathrm{T}}\left(\beta_{A_0}\right)\Omega_0^{-1}\left(\beta_{A_0}\right)\dot{g}_n\left(\beta_{A_0}\right)\right)\to G$$

To establish the asymptotic normality, it suffices to check the Lindeberg condition, *i.e.*, for any ε,

$$\sum_{i=1}^{n}E\left\|Z_{ni}\right\|^2I\left\{\left\|Z_{ni}\right\|>\varepsilon\right\}\to 0. \tag{A.8}$$

Using the Cauchy-Schwarz inequality, we have

$$\sum_{i=1}^{n}E\left\|Z_{ni}\right\|^2I\left\{\left\|Z_{ni}\right\|>\varepsilon\right\}=nE\left\|Z_{ni}\right\|^2\,I\left\{\left\|Z_{ni}\right\|>\varepsilon\right\}\le n\left\{E\left\|Z_{ni}\right\|^4\right\}^{1/2}\left\{P\left(\left\|Z_{ni}\right\|>\varepsilon\right)\right\}^{1/2}.$$

By Bhebyshev's inequality,

$$P\left(\left\|Z_{n1}\right\|>\varepsilon\right)\le\varepsilon^{-2}E\left\|Z_{n1}\right\|^2=\varepsilon^{-2}E\left\|\frac{1}{n}A_nM_n\left(\beta_{A_0}\right)\dot{g}_n^{\mathrm{T}}\left(\beta_{A_0}\right)\Omega_0^{-1}\left(\beta_{A_0}\right)g_n\left(\beta_{A_0}\right)\right\|^2=O\left(\frac{p_n^2}{n^2}\right),$$

and

$$\begin{aligned}E\left\|Z_{n1}\right\|^4&=\frac{1}{n^2}E\left\|A_nM_n\left(\beta_{A_0}\right)\dot{g}_n^{\mathrm{T}}\left(\beta_{A_0}\right)\Omega_0^{-1}\left(\beta_{A_0}\right)g_n\left(\beta_{A_0}\right)\right\|^4\\&\le\frac{1}{n^2}\lambda_{\max}\left(A_nA_n^{\mathrm{T}}\right)\lambda_{\max}\left(M_n\left(\beta_{A_0}\right)\dot{g}_n^{\mathrm{T}}\left(\beta_{A_0}\right)\Omega_0^{-1}\left(\beta_{A_0}\right)\right)E\left(g_n\left(\beta_{A_0}\right)g_n^{\mathrm{T}}\left(\beta_{A_0}\right)\right)^2=O\left(\frac{p_n^2}{n^2}\right)\end{aligned}$$

Thus, we have

$$\sum_{i=1}^{n}E\left\|Z_{ni}\right\|^2\,I\left\{\left\|Z_{ni}\right\|>\varepsilon\right\}=O\left(n\frac{p_n^2}{n^2}\frac{p_n}{n}\right)=o(1).$$

Therefore, Z_{ni} satisfies the conditions of the Lindeberg-Feller central limit theorem. Hence, the proof of Theorem 2 is completed.