Statistica Sinica **23** (2013), 901-927
doi:http://dx.doi.org/10.5705/ss.2011.058

MODEL SELECTION FOR CORRELATED DATA WITH DIVERGING NUMBER OF PARAMETERS

Hyunkeun Cho and Annie Qu

University of Illinois at Urbana-Champaign

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 propose the penalized quadratic inference function to perform model selection and estimation simultaneously in the framework of a diverging number of regression parameters. The penalized quadratic inference function can easily take correlation information from clustered data into account, yet it does not require specifying the likelihood function. This is advantageous compared to existing model selection methods for discrete data with large cluster size. In addition, the proposed approach enjoys the oracle property; it is able to identify non-zero components consistently with probability tending to 1, and any finite linear combination of the estimated non-zero components has an asymptotic normal distribution. We propose an efficient algorithm by selecting an effective tuning parameter to solve the penalized quadratic inference function. Monte Carlo simulation studies have the proposed method selecting the correct model with a high frequency and estimating covariate effects accurately even when the dimension of parameters is high. We illustrate the proposed approach by analyzing periodontal disease data.

Key words and phrases: Diverging number of parameters, longitudinal data, model selection, oracle property, quadratic inference function, SCAD.

1. Introduction

Longitudinal data arise frequently in biomedical and health studies in which repeated measurements from the same subject are correlated. The correlated nature of longitudinal data makes it difficult to specify the full likelihood function when responses are non-normal. Liang and Zeger (1986) developed the generalized estimating equation (GEE) for correlated data; it only requires the first two moments and a working correlation matrix involving a small number of nuisance parameters. Although the GEE yields a consistent estimator even if the working correlation structure is misspecified, the estimator can be inefficient under the misspecified correlation structure. Qu, Lindsay, and Li (2000) proposed the quadratic inference function (QIF) to improve the efficiency of the GEE when the working correlation is misspecified, in addition to providing an inference function for model diagnostic tests and goodness-of-fit tests.

Variable selection is fundamental in extracting important predictors from large data sets when the covariates are high-dimensional, as inclusion of high-dimensional redundant variables can hinder efficient estimation and inference for the non-zero coefficients. In the longitudinal data framework, several variable selection methods for marginal models have been developed. Pan (2001) proposed an extension of the Akaike information criterion (Akaike (1973)) by applying the quasilikelihood to the GEE, assuming independent working correlation. Cantoni, Flemming, and Ronchetti (2005) proposed a generalized version of Mallows' C_p (Mallows (1973)) by minimizing the prediction error. However, the asymptotic properties of these model selection procedures have not been well studied. Wang and Qu (2009) developed a Bayesian information type of criterion (Schwarz (1978)) based on the quadratic inference function to incorporate correlation information. These approaches are the best sub-set selection approaches and have been shown to be consistent for model selection. However, the L_0-based penalty can be computationally intensive and unstable when the dimension of covariates is high. Fu (2003) applied the bridge penalty model to the GEE and Xu et al. (2010) introduced the adaptive LASSO (Zou (2006)) for the GEE setting. Dziak (2006) and Dziak, Li, and Qu (2009) discussed the SCAD penalty for GEE and QIF model selection for longitudinal data. These methods are able to perform model selection and parameter estimation simultaneously. However, most of the theory and implementation is restricted to a fixed dimension of parameters.

Despite the importance of model selection in high-dimensional settings (Fan and Li (2006); Fan and Lv (2010)), model selection for longitudinal discrete data is not well studied when the dimension of parameters diverges. This is probably due to the challenge of specifying the likelihood function for correlated discrete data. Wang, Zhou, and Qu (2012) developed the penalized generalized estimating equation (PGEE) for model selection when the number of parameters diverges, and this is based on the penalized estimating equation approach by Johnson, Lin, and Zeng (2008) in the framework of a diverging number of parameters by Wang (2011). However, in our simulation studies, we show that the penalized GEE tends to overfit the model regardless of whether the working correlation is correctly specified or not.

In this paper, we propose the penalized quadratic inference function (PQIF) approach for model selection in the longitudinal data setting. We show that, even when the number of parameters diverges as the sample size increases, the penalized QIF utilizing the smoothly clipped absolute deviation (SCAD) penalty function (Fan and Li (2001); Fan and Peng (2004)) possesses such desirable features of the SCAD as sparsity, unbiasedness, and continuity. The penalized QIF also enjoys the oracle property. That is, the proposed model selection is able to identify non-zero components correctly with probability tending to 1,

and any valid linear combination of the estimated non-zero components is the asymptotically normal.

One of the unique advantages of the penalized QIF approach for correlated data is that the correlation within subjects can be easily taken into account, as the working correlation can be approximated by a linear combination of known basis matrices. In addition, the nuisance parameters associated with the working correlation are not required to be estimated, as the minimization of the penalized QIF does not involve the nuisance parameters. This is especially advantageous when the dimension of estimated parameters is high, as reducing nuisance parameter estimation improves estimation efficiency and model selection performance significantly. Consequently, the penalized QIF outperforms the penalized GEE approach under any working correlation structure in our simulation studies. Furthermore, the penalized QIF only requires specifying the first two moments instead of the full likelihood function, and this is especially advantageous for discrete correlated data.

Another important advantage of our approach is in tuning parameter selection. The selection of the tuning parameter plays an important role in achieving desirable performance in model selection. We provide a more effective tuning parameter selection procedure based on the Bayesian information quadratic inference function criterion (BIQIF), and show that the proposed tuning parameter selector leads to consistent model selection and estimation for regression parameters. This is in contrast to the penalized GEE, which relies on cross-validation for tuning parameter selection. Our simulation studies for binary longitudinal data indicate that the penalized QIF is able to select the correct model with a higher frequency and provide a more efficient estimator, compared to the penalized GEE approach, when the dimensions of covariates and non-zero parameters increase as the sample size increases.

The paper is organized as follows. Section 2 briefly describes the quadratic inference function for longitudinal data. Section 3 introduces the penalized quadratic inference function and provides the asymptotic properties for variable selection when the number of parameters diverges. Section 4 presents two algorithms to implement the penalized QIF approach and a tuning parameter selector. Section 5 reports on simulation studies for binary responses and provides a data example from a periodontal disease study. The final section provides concluding remarks and discussion. All necessary lemmas and theoretical proofs are in the Appendix.

2. Quadratic Inference Function for Longitudinal Data

Suppose the response variable for the ith subject is measured m_i times, $y_i = (y_{i1}, \ldots, y_{im_i})^T$, where y_i's are independent identically distributed, $i = 1, \ldots, n$,

n is the sample size and m_i is the cluster size. The corresponding covariate $X_i = (X_{i1}, \ldots, X_{im_i})^T$ is $m_i \times p_n$-dimensional matrix for the ith subject. In the generalized linear model framework, the marginal mean of y_{ij} is specified as $\mu_{ij} = E(y_{ij}|X_{ij}) = \mu(X_{ij}^T \beta_n)$, where $\mu(\cdot)$ is the inverse link function and β_n is a p_n-dimensional parameter vector in the parameter space $\Omega_{p_n} \in \mathbf{R}^{p_n}$, p_n diverging as the sample size increases. Since the full likelihood function for correlated non-Gaussian data is rather difficult to specify when the cluster size is large, Liang and Zeger (1986) developed the generalized estimating equation (GEE) to obtain the β_n estimator by solving the equations

$$W_n(\beta_n) = \sum_{i=1}^{n} \dot{\mu}_i^T(\beta_n) V_i^{-1}(\beta_n)(y_i - \mu_i(\beta_n)) = 0, \tag{2.1}$$

where $\dot{\mu}_i = (\partial \mu_i / \partial \beta_n)$ is a $m_i \times p_n$ matrix, and $V_i = A_i^{1/2} R A_i^{1/2}$, with A_i the diagonal marginal variance matrix of y_i and R the working correlation matrix that involves a small number of correlation parameters. Although the GEE estimator is consistent and asymptotically normal even if the working correlation matrix is misspecified, the GEE estimator is not efficient under the misspecification of the working correlation.

To improve efficiency, Qu, Lindsay, and Li (2000) proposed the quadratic inference function for longitudinal data. They assume that the inverse of the working correlation can be approximated by a linear combination of several basis matrices, that is,

$$R^{-1} \approx \sum_{j=0}^{k} a_j M_j, \tag{2.2}$$

where M_0 is the identity matrix, $M_1, \ldots, M_k$ are basis matrices with 0 and 1 components and $a_0, \ldots, a_k$ are unknown coefficients. For example, if R corresponds to an exchangeable structure, then $R^{-1} = a_0 M_0 + a_1 M_1$, where a_0 and a_1 are constants associated with the exchangeable correlation parameter and the cluster size, and M_1 is a symmetric matrix with 0 on the diagonal and 1 elsewhere. If R has AR-1 structure, then $R^{-1} = a_0 M_0 + a_1 M_1 + a_2 M_2$, where a_0, a_1, and a_2 are constants associated with the AR-1 correlation parameter, M_1 is a symmetric matrix with 1 on the sub-diagonal entries and 0 elsewhere, and M_2 is a symmetric matrix with 1 on entries $(1,1)$ and (m_i, m_i). If there is no prior knowledge on the correlation structure, then a set of basis matrices containing 1 for (i,j) and (j,i) entries and 0 elsewhere can be used as a linear representation for R^{-1}.

Selecting the correct correlation matrix is fundamental to the QIF approach since it can improve the efficiency of the regression parameter estimators.

Zhou and Qu (2012) provide a model selection approach for selecting informative basis matrices that approximate the inverse of the true correlation structure. Their key idea is to approximate the empirical estimator of the correlation matrix by a linear combination of candidate basis matrices representing common correlation structures as well as mixtures of several correlation structures. They minimize the Euclidean distance between the estimating functions based on the empirical correlation matrix and candidate basis matrices, and penalize models involving too many matrices.

By replacing the inverse of the working correlation matrix with (2.2), the GEE in (2.1) can be approximated as a linear combination of the elements in the following extended score vector:

$$\bar{g}_n(\beta_n) = \frac{1}{n}\sum_{i=1}^{n} g_i(\beta_n) \approx \frac{1}{n}\begin{pmatrix} \sum_{i=1}^{n}(\dot{\mu}_i)^T A_i^{-1}(y_i-\mu_i) \\ \sum_{i=1}^{n}(\dot{\mu}_i)^T A_i^{-1/2} M_1 A_i^{-1/2}(y_i-\mu_i) \\ \vdots \\ \sum_{i=1}^{n}(\dot{\mu}_i)^T A_i^{-1/2} M_k A_i^{-1/2}(y_i-\mu_i) \end{pmatrix}. \tag{2.3}$$

Since it is impossible to set each equation in (2.3) to zero simultaneously in solving for β_n, as the dimension of the estimating equations exceeds the dimension of parameters, Qu, Lindsay, and Li (2000) applied the generalized method of moments (Hansen (1982)) to obtain an estimator of β_n by minimizing the quadratic inference function (QIF)

$$Q_n(\beta_n) = n\bar{g}_n(\beta_n)^T \bar{C}_n^{-1}(\beta_n)\bar{g}_n(\beta_n),$$

where $\bar{C}_n(\beta_n) = (1/n)\sum_{i=1}^{n} g_i(\beta_n)g_i^T(\beta_n)$ is the sample covariance matrix of g_i. Note that this minimization does not involve estimating the nuisance parameters $a_0, \ldots, a_k$ associated with the linear weights in (2.2). The quadratic inference function plays an inferential role similar to minus twice the log-likelihood function, and it possesses the same chi-squared asymptotic properties as in the likelihood ratio test. The QIF estimator is optimal in the sense that the asymptotic variance matrix of the estimator of β_n reaches the minimum among estimators solved by the same linear class of the estimating equations given in (2.3) (Qu, Lindsay, and Li (2000)).

3. A New Estimation Method and Theory

For correlated discrete data, existing approaches for model selection are rather limited due to the difficulty of specifying the full likelihood function. We propose a new variable selection approach based on the penalized quadratic inference function that can incorporate correlation information from clusters. The

proposed procedure can estimate parameters and select important variables simultaneously in the framework of a diverging number of covariates. Even when the dimension of parameters diverges as the sample size increases, the proposed model selection contains the sparsity property and shrinks the estimators of the non-signal components to zero. In addition, the non-zero components are selected correctly with probability tending to 1. We also show that the estimators of the non-zero components are consistent at the convergence rate of $\sqrt{n/p_n}$, and follow the normal distribution asymptotically.

Without loss of generality, the cluster sizes are taken to be equal, $m_i = m$, although the cluster size is unbalanced in our data example. Since the response variables are not necessarily continuous, we replace the typical least square function by the quadratic inference function since it is analogous to minus twice the log-likelihood. We define it as

$$S_n(\beta_n) = Q_n(\beta_n) + n\sum_{j=1}^{p_n} P_{\lambda_n}(|\beta_{nj}|). \tag{3.1}$$

Among several penalty functions $P_{\lambda_n}(\cdot)$, we choose the non-convex SCAD penalty function corresponding to

$$\begin{aligned} P_{\lambda_n}(|\beta_{nj}|) = {} & \lambda_n|\beta_{nj}|I(0 \le |\beta_{nj}| < \lambda_n) \\ & + \left\{ \frac{a\lambda_n(|\beta_{nj}| - \lambda_n) - (|\beta_{nj}|^2 - \lambda_n^2)/2}{(a-1)} + \lambda_n^2 \right\} I(\lambda_n \le |\beta_{nj}| < a\lambda_n) \\ & + \left\{ \frac{(a-1)\lambda_n^2}{2} + \lambda_n^2 \right\} I(|\beta_{nj}| \ge a\lambda_n), \end{aligned}$$

where $I(\cdot)$ is an indicator function, $\lambda_n > 0$ is a tuning parameter, and a constant a chosen to be 3.7 (Fan and Li (2001)). The SCAD penalty function has such desirable features as sparsity, unbiasedness, and continuity, while such penalty functions as bridge regression, LASSO, and hard thresholding fail to possess these three features simultaneously. For example, the bridge regression penalty (Frank and Friedman (1993)) does not satisfy the sparsity property, the LASSO penalty (Tibshirani (1996)) does not satisfy the unbiasedness property, and the hard thresholding penalty (Antoniadis (1997)) does not satisfy the continuity property. On the other hand, the adaptive LASSO (Zou (2006)); Zou and Zhang (2009)) does have all three features, and we apply the adaptive LASSO penalty for the proposed method in our simulation studies. The performance of the SCAD and the adaptive LASSO are quite comparable, as indicated in Section 5.1.

We obtain the estimator $\hat{\beta}_n$ by minimizing $S_n(\beta_n)$ in (3.1). Minimizing (3.1) ensures that the estimation and model selection procedures are efficient, since

correlations within the same cluster are taken into account for the first part of the objective function in (3.1). Model selection is more important, yet more challenging, when the dimension of the parameters increases as the sample size increases. Fan and Peng (2004) provide the asymptotic properties of model selection using the penalized likelihood function under the framework of a diverging number of parameters. We provide the asymptotic properties of model selection for longitudinal data without requiring the likelihood function when the number of parameters increases with the sample size.

We assume that there is a true model with the first q_n $(0 \le q_n \le p_n)$ predictors non-zero and the rest are zeros. The vector $\beta_n^* = (\beta_s^{*T}, \beta_{s^c}^{*T})^T$ is taken as the true parameter, where $\beta_s^* = (\beta_{n1}^*, \ldots, \beta_{nq_n}^*)^T$ is a non-zero coefficient vector and $\beta_{s^c}^* = (\beta_{n(q_n+1)}^*, \ldots, \beta_{np_n}^*)^T$ is a zero vector. Let $\hat{\beta}_n = (\hat{\beta}_s^T, \hat{\beta}_{s^c}^T)^T$ be an estimator of β_n that minimizes the penalized QIF in (3.1). Regularity conditions on the quadratic inference functions are imposed to establish the asymptotic properties of this estimator:

(A) The first derivative of the QIF satisfies

$$E\left\{\frac{\partial Q_n(\beta_n)}{\partial \beta_{nj}}\right\} = 0 \qquad \text{for } j = 1, \ldots, p_n,$$

and the second derivative of the QIF satisfies

$$E\left\{\frac{\partial^2 Q_n(\beta_n)}{\partial \beta_{nj}\partial \beta_{nk}}\right\}^2 < K_1 < \infty \qquad \text{for } j, k = 1, \ldots, p_n, \text{and a constant } K_1.$$

With $D_n(\beta_n) = E\{n^{-1}\nabla^2 Q_n(\beta_n)\}$, the eigenvalues of $D_n(\beta_n)$ are uniformly bounded by positive constants K_2 and K_3 for all n.

(B) The true parameter β_n is contained in a sufficiently large open subset ω_{p_n} of $\Omega_{p_n} \in \mathbf{R}^{p_n}$, and there exist constants M and K_4 such that

$$\left|\frac{\partial^3 Q_n(\beta_n)}{\partial \beta_{nj}\partial \beta_{nl}\partial \beta_{nk}}\right| \le M$$

for all β_n, and $E_{\beta_n}(M^2) < K_4 < \infty$ for all p_n and n.

(C) The parameter values $\beta_{n1}, \ldots, \beta_{nq_n}$ are such that $\min_{1 \le j \le q_n} |\beta_{nj}|/\lambda_n$ goes to ∞ as $n \to \infty$.

Conditions (A) and (B) require that the second and fourth moments of the quadratic inference function be bounded, and that the expectation of the second derivative of the QIF be positive definite with uniformly bounded eigenvalues; they are quite standard for estimating equation approaches, and can be verified

through the eigenvalues of the specified matrices. Condition (C) is easily satisfied as long as the tuning parameter is sufficiently small relative to non-zero coefficients. This type of assumption is standard in much of the model selection literature, e.g., Wang, Li, and Tsai (2007), Wang, Li, and Leng (2009), Zhang, Li, and Tsai (2010) and Gao et al. (2012). Fan and Peng (2004) also provided similar conditions for the penalized likelihood approach.

Further, condition (C) ensures that the penalized QIF possesses the oracle property, $\max\{P'_{\lambda_n}(|\beta_{nj}|) : \beta_{nj} \neq 0\} = 0$ and $\max\{P''_{\lambda_n}(|\beta_{nj}|) : \beta_{nj} \neq 0\} = 0$ when n is sufficiently large; consequently, the following regularity conditions for the SCAD penalty are satisfied

(D) $\liminf_{n\to\infty} \inf_{\theta\to 0^+} P'_{\lambda_n}(\theta)/\lambda_n > 0$;
(E) $\max\{P'_{\lambda_n}(|\beta_{nj}|) : \beta_{nj} \neq 0\} = o_p(1/\sqrt{np_n})$;
(F) $\max\{P''_{\lambda_n}(|\beta_{nj}|) : \beta_{nj} \neq 0\} = o_p(1/\sqrt{p_n})$.

These conditions ensure that the penalty functions possess desirable features such as sparsity, unbiasedness, and continuity for model selection. Specifically, (D) ensures that the penalized QIF estimator has the sparsity property since the penalty function is singular at the origin; (E) guarantees that the estimators for parameters with large magnitude are unbiased and retain asymptotic $\sqrt{n}$-consistency; (F) ensures that the first QIF term is dominant in the objective function (3.1).

Theorem 1. *If* (A)−(F) *hold and* $p_n = o(n^{1/4})$, *then there exists a local minimizer* $\hat{\beta}_n$ *of* $S(\beta_n)$ *such that* $\|\hat{\beta}_n - \beta_n^*\| = O_p\{\sqrt{p_n}(n^{-1/2} + a_n)\}$, *where* $a_n = max\{P'_{\lambda_n}(|\beta_{nj}|) : \beta_{nj} \neq 0\}$.

This result establishes a $\sqrt{n/p_n}$-consistency for the penalized quadratic inference function estimator; it holds as long as (C) is satisfied, since it ensures $a_n = 0$ when n is large.

In the following, we write

$$\mathbf{b}_n = \{P'_{\lambda_n}(|\beta_{n1}|)\text{sign}(\beta_{n1}), \ldots, P'_{\lambda_n}(|\beta_{np_n}|)\text{sign}(\beta_{nq_n})\}^T$$

and

$$\mathbf{\Sigma}_{\lambda_n} = diag\{P''_{\lambda_n}(\beta_{n1}), \ldots, P''_{\lambda_n}(\beta_{nq_n})\},$$

where $\text{sign}(\alpha) = I(\alpha > 0) - I(\alpha < 0)$.

Theorem 2. *Under* (A)−(F), *if* $p_n = o(n^{1/4})$, $\lambda_n \to 0$, *and* $\sqrt{n/p_n}\lambda_n \to \infty$ *as* $n \to \infty$, *then the estimator* $\hat{\beta}_n = (\hat{\beta}_s^T, \hat{\beta}_{s^c}^T)^T$ *satisfies the following, with probability tending to* 1.

(1) (*Sparsity*) $\hat{\beta}_{s^c} = 0$.

(2) (*Asymptotic normality*) *For any given* $d\times q_n$ *matrix* B_n *such that* $B_nB_n^T \to F$, *where* F *is a fixed dimensional constant matrix and* $D_n(\beta_s^*) = \mathrm{E}\{n^{-1}\nabla^2 Q_n(\beta_s^*)\}$,

$$\sqrt{n}B_nD_n^{-1/2}(\beta_s^*)\{D_n(\beta_s^*)+\mathbf{\Sigma}_{\lambda_n}\}\big[(\hat{\beta}_s-\beta_s^*)+\{D_n(\beta_s^*)+\mathbf{\Sigma}_{\lambda_n}\}^{-1}\boldsymbol{b}_n\big] \xrightarrow{\mathrm{d}} N(0,F).$$

In addition, if $\mathbf{\Sigma}_{\lambda_n} \to 0$ and $\mathbf{b}_n \to 0$ as $n \to \infty$, $\sqrt{n}B_nD_n^{1/2}(\beta_s^*)(\hat{\beta}_s - \beta_s^*) \xrightarrow{\mathrm{d}} N(0,F)$. Theorem 2 has the estimator of the penalized QIF as efficient as the oracle estimator that assumes the true model is known. The proofs of the two theorems and the necessary lemmas are in the Appendix.

4. Implementation

4.1. Local quadratic approximation

Since the SCAD penalty function is non-convex, we use the local quadratic approximation (Fan and Li (2001); Xue, Qu, and Zhou (2010)) to minimize the penalized quadratic inference function in (3.1) with the unpenalized QIF estimator as the initial value $\beta^{(0)}$. If $\beta^{(k)} = \big(\beta_1^{(k)},\ldots,\beta_{p_n}^{(k)}\big)^T$ is the estimator at the kth iteration and $\beta_j^{(k)}$ is close to 0, say $\big|\beta_j^{(k)}\big| < 10^{-4}$, then we set $\beta_j^{(k+1)}$ to 0. If $\beta_j^{(k+1)} \neq 0$ for $j = 1,\ldots,q_k$ and $\beta_j^{(k+1)} = 0$ for $j = q_{k+1},\ldots,p_n$, write $\beta^{(k+1)} = \Big(\big(\beta_s^{(k+1)}\big)^T, \big(\beta_{s^c}^{(k+1)}\big)^T\Big)^T$ where β_s^{k+1} is a vector containing the non-zero components and $\beta_{s^c}^{k+1}$ is a zero vector.

The local quadratic approximation is outlined as follows. For $\beta_j^{(k)} \neq 0$,

$$P_{\lambda_n}\big(|\beta_j|\big) \approx P_{\lambda_n}\big(|\beta_j^{(k)}|\big) + \frac{1}{2}\Big\{\frac{P'_{\lambda_n}\big(|\beta_j^{(k)}|\big)}{|\beta_j^{(k)}|}\Big\}\big(\beta_j^{(k)2} - \beta_j^2\big),$$

where $\beta_j \approx \beta_j^{(k)}$ and $P'_{\lambda_n}(|\beta_n|)$ is the first derivative of the SCAD penalty $P_{\lambda_n}(|\beta_n|)$,

$$P'_{\lambda_n}(|\beta_n|) = \lambda_n\Big\{I(|\beta_n| \le \lambda_n) + \frac{(a\lambda_n - |\beta_n|)_+}{(a-1)\lambda_n} I(|\beta_n| > \lambda_n)\Big\}.$$

Consequently, the penalized QIF in (3.1) can be approximated by

$$\begin{aligned} Q_n(\beta^{(k)}) &+ \nabla Q_n(\beta^{(k)})^T(\beta_s - \beta_s^{(k)}) + \frac{1}{2}(\beta_s - \beta_s^{(k)})^T\nabla^2 Q_n(\beta^{(k)})(\beta_s - \beta_s^{(k)}) \\ &+ \frac{1}{2}n\beta_s^T\Pi(\beta^{(k)})\beta_s, \end{aligned} \tag{4.1}$$

where β_s is a vector with non-zero components which has the same dimension of $\beta_s^{(k)}$, $\nabla Q_n(\beta^{(k)}) = \frac{\partial Q_n(\beta^{(k)})}{\partial \beta_s}$, $\nabla^2 Q_n(\beta^{(k)}) = \frac{\partial^2 Q_n(\beta^{(k)})}{\partial \beta_s \partial \beta_s^T}$, and $\Pi(\beta^{(k)}) =$

$diag\Big\{P'_{\lambda_n}(|\beta_1^{(k)}|)/|\beta_1^{(k)}|\,,\ldots,P'_{\lambda_n}(|\beta_{q_k}^{(k)}|)/|\beta_{q_k}^{(k)}|\Big\}$. The non-zero component $\beta_s^{(k+1)}$ at the $k+1$ step can be obtained by minimizing the quadratic function in (4.1) using the Newton-Raphson algorithm, which is equivalent to solving

$$\beta_s^{(k+1)} = \beta_s^{(k)} - \Big\{\nabla^2 Q_n(\beta^{(k)}) + n\Pi(\beta^{(k)})\Big\}^{-1}\Big\{\nabla Q_n(\beta^{(k)}) + n\Pi(\beta^{(k)})\beta^{(k)}\Big\}.$$

We iterate the above process to convergence, for example, when $\|\beta_s^{(k+1)} - \beta_s^{(k)}\| < 10^{-7}$.

4.2. Linear approximation method

We also consider an alternative algorithm based on the linear approximation for the first part of the PQIF in (3.1). This is analogous to Xu et al.'s (2010) linear approximation for the penalized GEE approach; however, their objective function and LASSO penalty function differ from ours. The key step here is to approximate the response $\mathbf{y}$ through linear approximation: $\mathbf{y} \approx \mu + \dot{\mu}(\hat{\beta}_Q)(\hat{\beta}_Q - \beta_n)$, where $\hat{\beta}_Q$ is the QIF estimator. One of the advantages of using the linear approximation approach is that the minimization of the penalized QIF can be solved using the *plus* package (Zhang (2007)) in R directly, since the first part of the objective function in (3.1) transforms to least squares.

For the extended score vector in (2.3), we replace $(\mathbf{y}_i - \mu_i)$ with $\dot{\mu}_i(\hat{\beta}_Q)(\hat{\beta}_Q - \beta_n)$, and therefore the extended score vector $g_i(\beta_n)$ can be expressed as

$$\begin{aligned} g_i(\beta_n) &\approx \begin{pmatrix} \dot{\mu}_i^T A_i^{-1}\dot{\mu}_i(\hat{\beta}_Q - \beta_n) \\ \dot{\mu}_i^T A_i^{-1/2} M_1 A_i^{-1/2}\dot{\mu}_i(\hat{\beta}_Q - \beta_n) \\ \vdots \\ \dot{\mu}_i^T A_i^{-1/2} M_k A_i^{-1/2}\dot{\mu}_i(\hat{\beta}_Q - \beta_n) \end{pmatrix} = \begin{pmatrix} \dot{\mu}_i^T A_i^{-1}\dot{\mu}_i \\ \dot{\mu}_i^T A_i^{-1/2} M_1 A_i^{-1/2}\dot{\mu}_i \\ \vdots \\ \dot{\mu}_i^T A_i^{-1/2} M_k A_i^{-1/2}\dot{\mu}_i \end{pmatrix} (\hat{\beta}_Q - \beta_n) \\ &= G_i(\hat{\beta}_Q - \beta_n). \end{aligned}$$

To simplify the notation, let $G = (G_1^T, G_2^T, \ldots, G_n^T)^T$ be a $(k+1)np \times p$ matrix and $\tilde{C}_n^{-1}$ be the $(k+1)np \times (k+1)np$ block diagonal matrix with each block matrix as $\bar{C}_n^{-1}$. The penalized QIF in (3.1) can be approximated by

$$\begin{aligned} S_n(\beta_n) &\approx \Big\{G(\hat{\beta}_Q)\hat{\beta}_Q - G(\hat{\beta}_Q)\beta_n\Big\}^T \tilde{C}_n^{-1}\Big\{G(\hat{\beta}_Q)\hat{\beta}_Q - G(\hat{\beta}_Q)\beta_n\Big\} + n\sum_{j=1}^{p_n} P_\lambda(|\beta_{nj}|) \\ &= \Big\{\tilde{C}_n^{-1/2}G(\hat{\beta}_Q)\hat{\beta}_Q - \tilde{C}_n^{-1/2}G(\hat{\beta}_Q)\beta_n\Big\}^T \Big\{\tilde{C}_n^{-1/2}G(\hat{\beta}_Q)\hat{\beta}_Q - \tilde{C}_n^{-1/2}G(\hat{\beta}_Q)\beta_n\Big\} \\ &\quad + n\sum_{j=1}^{p_n} P_\lambda(|\beta_{nj}|). \end{aligned}$$

Let $U = \tilde{C}_N^{-1/2} G(\hat{\beta}_Q)\hat{\beta}_Q$ and $T = \tilde{C}_N^{-1/2} G(\hat{\beta}_Q)$. Then the penalized QIF can be formulated as

$$S_n(\beta_n) \approx \big(U - T\beta_n\big)^T \big(U - T\beta_n\big) + n \sum_{j=1}^{p_n} P_\lambda(|\beta_{nj}|).$$

Here the *plus* package (Zhang (2007)) can be applied in R using the SCAD penalty.

In this way we approximate two parts of the objection function in (3.1). The local quadratic approximation method approximates the SCAD penalty function, while the linear approximation method approximates the first term of the QIF in (3.1). Based on our simulations, the local quadratic approximation approach performs better than the linear approximation method in terms of selecting the true model with a higher frequency, and with a smaller MSE for the estimators.

4.3. Tuning parameter selector

The performance of our method relies on the choice of a tuning parameter that is essential for model selection consistency and sparsity. Fan and Li (2001) proposed generalized cross-validation (GCV) to choose the regularization parameter. However, Wang, Li, and Tsai (2007) showed that the GCV approach sometimes tends to overfit the model and select null variables as non-zero components. In contrast, the Bayesian information criterion (BIC) is able to identify the true model consistently, and we adopt it based on the QIF as an objective function (BIQIF) (Wang and Qu (2009)). The BIQIF is defined as

$$\text{BIQIF}_{\lambda_n} = Q_n(\hat{\beta}_{\lambda_n}) + df_{\lambda_n} \log(n), \tag{4.2}$$

where $\hat{\beta}_{\lambda_n}$ is the marginal regression parameter estimated by minimizing the penalized QIF in (3.1) for a given λ_n, and df_{λ_n} is the number of non-zero coefficients in $\hat{\beta}_{\lambda_n}$. We choose the optimal tuning parameter λ_n by minimizing the BIQIF in (4.2).

To investigate consistency, let $\Upsilon = \{j_1, \ldots, j_q\}$ be an arbitrary candidate model that contains predictors $j_1, \ldots, j_q$ $(1 \le q \le p_n)$ and $\Upsilon_{\lambda_n} = \{j : \hat{\beta}_{nj} \neq 0\}$, where $\hat{\beta}_n$ is the estimator of the penalized QIF corresponding to the tuning parameter λ_n. Let $\Upsilon_F = \{1, \ldots, p_n\}$ and $\Upsilon_T = \{1, \ldots, q_n\}$ denote the full model and the true model respectively. An arbitrary candidate model Υ is overfitted if $\Upsilon \supset \Upsilon_T$ and $\Upsilon \neq \Upsilon_T$, underfitted if $\Upsilon \not\supseteq \Upsilon_T$. We take $\Lambda_- = \{\lambda_n \in \Lambda : \Upsilon \not\supseteq \Upsilon_T\}$, $\Lambda_0 = \{\lambda_n \in \Lambda : \Upsilon = \Upsilon_T\}$, and $\Lambda_+ = \{\lambda_n \in \Lambda : \Upsilon \supset \Upsilon_T \text{ and } \Upsilon \neq \Upsilon_T\}$ accordingly. We use similar arguments to those in Wang, Li, and Tsai (2007) to obtain the following.

Lemma 1. *If* (A)−(F) *hold,* $P(BIQIF_{\lambda_o} = BIQIF_{\Upsilon_T}) \longrightarrow 1$.

Lemma 2. *If* (A)−(F) *hold,* $P(\inf_{\lambda_n \in \Lambda_- \cup \Lambda_+} BIQIF_{\lambda_n} > BIQIF_{\lambda_o}) \longrightarrow 1$.

Lemmas 1 and 2 imply that, with probability tending to 1, the BIQIF procedure selects the tuning parameter λ_o that identifies the true model. Proofs are provided in the Appendix.

4.4. Unbalanced data implementation

In longitudinal studies, the data can be unbalanced as cluster size can vary for different subjects because of missing data. In the following, we provide a strategy to implement the proposed method for unbalanced data using a transformation matrix for each subject. Let H_i be a $m \times m_i$ transformation matrix of the ith subject, where m is the cluster size of the fully observed subject without missing data. The matrix H_i's are generated by deleting the columns of the $m \times m$ identity matrix corresponding to the missing measurements for the ith subject. Through the transformation, g_i in (2.3) is replaced by $g_i^* = \big\{(\dot{\mu}_i^*)^T (A_i^*)^{-1}(y_i^* - \mu_i^*), (\dot{\mu}_i^*)^T (A_i^*)^{-1/2} M_1 (A_i^*)^{-1/2}(y_i^* - \mu_i^*), \ldots, (\dot{\mu}_i^*)^T (A_i^*)^{-1/2} M_k (A_i^*)^{-1/2}(y_i^* - \mu_i^*)\big\}$, where $\dot{\mu}_i^* = H_i \dot{\mu}_i$, $\mu_i^* = H_i \mu_i$, $y_i^* = H_i y_i$, and $A_i^* = H_i A_i H_i^T$. The QIF estimator with unbalanced data is obtained based on the transformed extended score vector $\bar{g}_n^*(\beta_n) = (1/n)\sum_{i=1}^n g_i^*(\beta_n)$. Note that the values of $\dot{\mu}_i^*$ and $y_i^* - \mu_i^*$ are 0 corresponding to the missing observations, and thus the missing observations do not affect the estimation of β_n.

5. Numerical Studies

In this section, we examine the performance of the penalized QIF procedure with the three different penalty functions SCAD, LASSO, and Adaptive LASSO, and compare them with the penalized GEE with the SCAD penalty through simulation studies for correlated binary responses. We also compare these approaches using a data from a periodontal disease study.

5.1. Binary response

We generated the correlated binary response variable from a marginal logit model

$$\text{logit}(\mu_{ij}) = X_{ij}^T \beta, \qquad i = 1, \ldots, 400 \ \text{ and } j = 1, \ldots, 10,$$

where $X_{ij} = \big(x_{ij}^{(1)}, \ldots, x_{ij}^{(p_n)}\big)^T$ and $\beta = (\beta_1, \ldots, \beta_{p_n})^T$. Each covariate $x_{ij}^{(k)}$ was generated independently from a Uniform (0, 0.8) distribution for $k = 1, \ldots, q_n$ and a Uniform (0, 1) distribution for $k = q_n + 1, \ldots, p_n$. We chose the dimension of total covariates to be $p_n = 20$ and 50, the dimension of relevant covariates to be $q_n = 3$ and 6, and applied three types of working correlation

structure (independent, AR-1 and exchangeable) in the simulations. In the first simulation setting, the true $\beta = (0.8, -0.7, -0.6, 0, \ldots, 0)^T$ with $q_n = 3$. In the second, $\beta = (0.8, -0.8, 0.7, -0.7, 0.6, -0.6, 0, \ldots, 0)^T$ with $q_n = 6$. The R package *mvtBinaryEP* was applied to generate the correlated binary responses with an exchangeable correlation structure as the true structure, with correlation coefficients $\rho_1 = 0.4$ and $\rho_2 = 0.3$ for the first and second simulation settings, respectively.

To compare our approach to the penalized GEE approach, we first provide a brief description of the PGEE (Wang, Zhou, and Qu (2012)). It is defined as $F_n(\beta_n) = W_n(\beta_n) - n\mathbf{P}_{\lambda_\mathbf{n}}(|\beta_\mathbf{n}|)\text{sign}(\beta_\mathbf{n})$, where $W_n(\beta_n)$ is the GEE defined in (2.1), $\mathbf{P}_{\lambda_\mathbf{n}}(|\beta_\mathbf{n}|) = (\mathbf{P}_{\lambda_\mathbf{n}}(|\beta_{\mathbf{n1}}|), \ldots, \mathbf{P}_{\lambda_\mathbf{n}}(|\beta_{\mathbf{np_n}}|))^\mathbf{T}$ with $P_{\lambda_n}(\cdot)$ a SCAD penalty, and $\text{sign}(\beta_n) = (\text{sign}(\beta_{n1}), \ldots, \text{sign}(\beta_{np_n}))^T$; here we have employed the component-wise product of $\mathbf{P}_{\lambda_\mathbf{n}}(|\beta_\mathbf{n}|)$ and $\text{sign}(\beta_n)$. The penalized GEE estimator was obtained by solving the estimating equation $F_n(\beta_n) = 0$ through the combination of the minorization-maximization (MM) algorithm (Hunter and Li (2005)) and the Newton-Raphson algorithm. In addition, the estimator of the component β_k $(k = 1, \ldots, p_n)$ was set to zero if $|\hat{\beta}_k| < 10^{-3}$. To choose a proper tuning parameter λ_n, a 5-fold cross-validation method was implemented on the grid set $\{0.01, 0.02, \ldots, 0.10\}$.

The simulation results from the model selection and the mean square errors (MSE) of estimation are provided in Table 1. Table 1 illustrates the performance of the penalized QIF approach with the penalty functions of LASSO, adaptive LASSO (ALASSO), and SCAD. The SCAD penalty for the penalized QIF was carried out as SCAD^1 through a local quadratic approximation, and SCAD^2 through a linear approximation. We compare the penalized QIF to the penalized GEE using the SCAD penalty from 100 simulation runs. In addition, we also provide the standard QIF without penalization (QIF) and the QIF approach based on the oracle model (Oracle) that assumes the true model is known. Table 1 provides the proportions of times selecting only the relevant variables (EXACT), the relevant variables plus others (OVER), and only some relevant variables (UNDER). To illustrate estimation efficiency, we took $\text{MSE} = \sum_{i=1}^{100} \|\hat{\beta}^{(i)} - \beta\|^2/100q$, where $\hat{\beta}^{(i)}$ is the estimator from the ith simulation run, β is the true parameter, q is the dimension of β, and $\|\cdot\|$ denotes the Euclidean-norm.

Table 1 indicates that the penalized QIF methods based on SCAD^1, SCAD^2, and ALASSO select the correct model with higher frequencies and smaller MSEs under any working correlation structure. Specifically, SCAD^1 performs better than SCAD^2 in terms of EXACT and MSE under the true correlation structure, and SCAD^1 and SCAD^2 perform similarly under the misspecified correlation structures (except when $p_n = 50$ and $q_n = 3$). The performance of SCAD^1 and the adaptive LASSO are quite comparable under any working correlation

Table 1. Performance of penalized QIF with LASSO, adaptive LASSO (ALASSO), SCAD[1], SCAD[2], and penalized GEE (PGEE) using SCAD penalty, with three working correlation structures: IN (independent), AR (AR-1) and EX (exchangeable).

		$p_n = 20$				$p_n = 50$			
	Method	MSE	EXACT	OVER	UNDER	MSE	EXACT	OVER	UNDER
					$q_n = 3$				
	Oracle	0.0018	-	-	-	0.0008	-	-	-
	QIF	0.0130	0.00	1.00	0.00	0.0130	0.00	1.00	0.00
	SCAD[1]	0.0037	0.70	0.29	0.01	0.0018	0.61	0.39	0.00
IN	SCAD[2]	0.0035	**0.73**	0.26	0.01	0.0017	**0.71**	0.28	0.01
	ALASSO	0.0036	0.70	0.29	0.01	0.0017	0.62	0.37	0.01
	LASSO	0.0098	0.34	0.66	0.00	0.0056	0.29	0.71	0.00
	PGEE	0.0046	0.52	0.46	0.02	0.0018	0.57	0.41	0.02
	Oracle	0.0014	-	-	-	0.0006	-	-	-
	QIF	0.0108	0.00	1.00	0.00	0.0124	0.00	1.00	0.00
	SCAD[1]	0.0021	0.83	0.17	0.00	0.0010	0.77	0.23	0.00
AR	SCAD[2]	0.0021	**0.84**	0.16	0.00	0.0012	**0.85**	0.15	0.00
	ALASSO	0.0021	0.82	0.18	0.00	0.0010	0.76	0.24	0.00
	LASSO	0.0077	0.29	0.71	0.00	0.0047	0.39	0.60	0.01
	PGEE	0.0029	0.65	0.35	0.00	0.0011	0.62	0.38	0.00
	Oracle	0.0012	-	-	-	0.0006	-	-	-
	QIF	0.0091	0.00	1.00	0.00	0.0108	0.00	1.00	0.00
	SCAD[1]	0.0017	**0.85**	0.15	0.00	0.0008	**0.89**	0.11	0.00
EX	SCAD[2]	0.0021	0.79	0.21	0.00	0.0016	0.72	0.28	0.00
	ALASSO	0.0016	0.84	0.16	0.00	0.0009	0.76	0.24	0.00
	LASSO	0.0065	0.36	0.64	0.00	0.0032	0.37	0.63	0.00
	PGEE	0.0019	0.71	0.29	0.00	0.0010	0.67	0.33	0.00
					$q_n = 6$				
	Oracle	0.0060	-	-	-	0.0022	-	-	-
	QIF	0.0149	0.00	1.00	0.00	0.0138	0.00	1.00	0.00
	SCAD[1]	0.0086	**0.74**	0.17	0.09	0.0040	0.52	0.40	0.08
IN	SCAD[2]	0.0093	0.69	0.20	0.11	0.0047	**0.53**	0.33	0.14
	ALASSO	0.0090	**0.74**	0.18	0.08	0.0042	0.50	0.40	0.10
	LASSO	0.0202	0.14	0.83	0.03	0.0147	0.22	0.68	0.10
	PGEE	0.0117	0.19	0.72	0.09	0.0079	0.06	0.75	0.19
	Oracle	0.0058	-	-	-	0.0019	-	-	-
	QIF	0.0143	0.00	1.00	0.00	0.0142	0.00	1.00	0.00
	SCAD[1]	0.0075	0.75	0.20	0.05	0.0031	**0.69**	0.25	0.06
AR	SCAD[2]	0.0088	**0.76**	0.15	0.09	0.0041	0.62	0.28	0.10
	ALASSO	0.0077	0.74	0.22	0.04	0.0030	**0.69**	0.27	0.03
	LASSO	0.0179	0.21	0.75	0.04	0.0127	0.26	0.69	0.05
	PGEE	0.0101	0.32	0.60	0.08	0.0059	0.17	0.70	0.13
	Oracle	0.0045	-	-	-	0.0016	-	-	-
	QIF	0.0119	0.00	1.00	0.00	0.0131	0.00	1.00	0.00
	SCAD[1]	0.0055	**0.83**	0.14	0.03	0.0024	**0.72**	0.25	0.03
EX	SCAD[2]	0.0075	0.75	0.10	0.15	0.0044	0.64	0.26	0.10
	ALASSO	0.0056	**0.83**	0.16	0.01	0.0024	0.69	0.29	0.02
	LASSO	0.0144	0.25	0.75	0.00	0.0102	0.30	0.67	0.03
	PGEE	0.0070	0.50	0.45	0.05	0.0032	0.23	0.73	0.04

structure. In contrast, the PQIF using the LASSO penalty tends to overfit the model, and its MSEs are much larger compared to the others under any setting. In addition, the MSEs of the PGEE estimators are all greater than those of SCAD^1 and ALASSO, and the EXACT frequencies of selecting the true models using PGEE with the SCAD penalty are lower than those of the PQIF based on the SCAD and ALASSO penalties.

When the number of relevant variables doubles, the EXACT of the PQIF based on SCAD and ALASSO decreases about 18% in the worst case; however, the EXACT of the PGEE decreases much more significantly. In the worst case when $q_n = 6$ and $p_n = 50$, the PGEE selects the correct model less than 25% of the time under any working correlation structure. In addition, the proposed model selection performance is always better under the true correlation structure. For instance, the EXACT is around 70% under the true correlation structure, while it is around 50% under the independent structure when $q_n = 6$ and $p_n = 50$. This simulation also indicates that the proposed model selection method starts to break down when both q_n and p_n increase under misspecified correlation structures such as the independent structure.

In summary, our simulation results show that the penalized QIF approaches with the SCAD and ALASSO penalties outperform the penalized GEE with the SCAD under any given correlation structure for various dimension settings of parameters in general. The LASSO penalty is not competitive for model selection with diverging number of parameters. In general, SCAD^1 performs better than SCAD^2, because the linear approximation of SCAD^2 is for the first (dominant) term of the PQIF, while the quadratic approximation of SCAD^1 is for the second.

5.2. Periodontal disease data example

We illustrate the proposed penalized QIF method through performing model selection for an observational study of periodontal disease data (Stoner (2000)). The data contain patients with chronic periodontal disease who have participated in a dental insurance plan. Each patient had an initial periodontal exam between 1988 and 1992, and was followed annually for ten years. The data set consists of 791 patients with unequal cluster sizes varying from 1 to 10.

The binary response variable $y_{ij} = 1$ if the patient i at jth year has at least one surgical tooth extraction, and $y_{ij} = 0$ otherwise. There are 12 covariates of interest: patient gender (*gender*), patient age at time of initial exam (*age*), last date of enrollment in the insurance plan in fractional years since 1900 (*exit*), number of teeth present at time of initial exam (*teeth*), number of diseased sites (*sites*), mean pocket depth in diseased sites (*pddis*), mean pocket depth in all sites (*pdall*), year since initial exam (*year*), number of non-surgical periodontal procedures in a year (*nonsurg*), number of surgical periodontal procedures in

a year (*surg*), number of non-periodontal dental treatments in a year (*dent*), and number of non-periodontal dental preventive and diagnostic procedures in a year (*prev*). Although the variable *exit* is not related to the model selection, we included it as a null variable to examine whether it is selected by the proposed model selection procedures or not. The logit link function was imposed here for the binary responses.

We minimized the penalized QIF with the SCAD penalty applying the local quadratic approximation and the adaptive LASSO penalty to compare with the penalized GEE. Here the AR-1 working correlation structure was assumed for estimation and model selection; as each patient was followed up annually, the measurements are less likely to be correlated if they are further away in time. Although other types of working correlation structure can be applied to these data, the results are not reported here as the outcomes are quite similar. Based on the penalized QIF, we selected relevant covariates as *age*, *sites*, *pddis*, *pdall*, and *dent*. The rest of the covariates were not selected and *exit* was not selected, as expected.

We compare the penalized QIF with the penalized GEE approach (Wang, Zhou, and Qu (2012)) based on the AR-1 working correlation structure. The estimated coefficients of both methods are reported in Table 2 indicating that the coefficients of *age*, *pddis*, *pdall*, and *dent* are positive and the coefficient of the variable *sites* is negative. The penalized GEE selects the covariate *teeth*, while the penalized QIF does not. Overall, the results of the two methods for the periodontal disease data are quite comparable.

In order to evaluate the model selection performance when the dimension of covariates increases, we generated an additional 15 independent null variables from a Uniform (0,1) distribution. We applied the penalized QIF and the penalized GEE based on the AR-1 working correlation structure. Out of 100 runs, the penalized QIF selected at least one of fifteen null variables 11 times for the SCAD penalty and 13 times for the adaptive LASSO penalty, while the penalized GEE selected one of the null variables 36 times. Furthermore, the penalized QIF always selected the relevant covariates *age*, *sites*, *pddis*, *pdall*, and *dent*, while the penalized GEE selected three other covariates *year*, *nonsurg*, and *prev* twice, in addition to the 6 relevant variables, in 100 runs. In this example, the penalized GEE tended to overfit the model.

6. Discussion

In this paper, we propose a penalized quadratic inference function approach that enables one to perform model selection and parameter estimation simultaneously for correlated data in the framework of a diverging number of parameters. Our procedure is able to take into account correlation from clusters without

Table 2. For the periodontal disease study, the coefficients estimated by the unpenalized QIF (QIF), the penalized QIF with SCAD through a local quadratic approximation (SCAD), the adaptive LASSO (ALASSO), the unpenalized GEE (GEE), and the penalized GEE (PGEE).

	QIF	SCAD	ALASSO	GEE	PGEE
intercept	-8.284	-11.144	-10.824	-8.287	-9.125
gender	-0.002	0.000	0.000	0.034	0.000
age	0.016	0.009	0.006	0.012	0.009
exit	-0.032	0.000	0.000	-0.002	0.000
teeth	0.000	0.000	0.000	-0.027	-0.014
sites	-0.006	-0.006	-0.005	0.000	-0.003
pddis	0.704	0.715	0.605	0.567	0.545
pdall	0.833	0.871	0.826	0.551	0.668
year	0.018	0.000	0.000	-0.021	0.000
nonsurg	0.004	0.000	0.000	-0.039	0.000
surg	0.018	0.000	0.000	0.015	0.000
dent	0.124	0.115	0.128	0.110	0.106
prev	-0.152	0.000	0.000	-0.147	0.000

specifying the full likelihood function or estimating the correlation parameters. The method can easily be applied to correlated discrete responses as well as to continuous responses. Furthermore, our theoretical derivations indicate that the penalized QIF approach is consistent in model selection and possesses the oracle property. Our Monte Carlo simulation studies show that the penalized QIF outperforms the penalized GEE, selecting the true model more frequently.

It is important to point out that the first part of the objective function in the penalized GEE is the generalized estimating equation that is exactly 0 if there is no penalization. This imposes limited choices for selecting a tuning parameter as there is no likelihood function available. Consequently, the PGEE can only rely on the GCV as a tuning parameter selection criterion, which tends to overfit the model. By contrast, the first part of the PQIF is analog to minus twice the log-likelihood function, and therefore can be utilized for tuning parameter selection. We develop a BIC-type criterion for selecting a proper tuning parameter which leads to consistent model selection and estimation for regression parameters. It is also known that the BIC-type of criterion performs better than the GCV when the dimension of parameters is high (Wang, Li, and Leng (2009)). Therefore it is not surprising that the proposed model selection based on the BIC-type of criterion performs well in our numerical studies.

The proposed method is generally applicable for correlated data as long as the correlated measurements have the same correlation structure between clusters. This assumption is quite standard for marginal approaches, where the diagonal marginal variance matrix could be different for different clusters, but the

working correlation matrix is common for different clusters. When each subject is followed at irregular time points, we can apply semiparametric modeling and nonparametric functional data approaches, but this typically requires more data collection from each subject.

Recent work on handling irregularly observed longitudinal data includes Fan, Huang, and Li (2007) and Fan and Wu (2008) based on semiparametric modeling, and functional data such as James and Hastie (2001); James and Sugar (2003); Yao, Müller, and Wang (2005); Hall, Müller, and Wang (2006) and Jiang and Wang (2010). However, most of these are not suitable for discrete longitudinal responses. In addition, semiparametric modeling requires parametric modeling for the correlation function. A disadvantage of the parametric approach for the correlation function is that the estimation of the correlation might be nonexistent or inconsistent if the correlated structure is misspecified. To model the covariance function completely nonparametrically, Li (2011) develops the kernel covariance model in the framework of a generalized partially linear model and transforms the kernel covariance estimator into a positive semidefinite covariance estimator through spectral decomposition. Li's (2011) approach could be applicable for our method on dealing with irregularly observed longitudinal data, but further research on this topic is needed.

Acknowledgement

The authors are very grateful to the Co-Editor, two referees and an associate editor for their insightful comments and suggestions that have improved the manuscript significantly. Annie Qu's research was supported by a National Science Foundation Grant (DMS-0906660).

Appendix: Proofs of Theorems and Lemmas

Lemma 3. *If* (D) *holds,* $A_n(\beta_n) = \mathrm{E}\{n^{-1}\nabla Q_n(\beta_n)\} = 0$ *and*

$$\left\| \frac{1}{n} \nabla Q_n(\beta_n) \right\| = o_p(1).$$

Proof. By Chebyshev's inequality it follows that, for any ϵ,

$$P\left(\left\| \frac{1}{n} \nabla Q_n(\beta_n) - A_n(\beta_n) \right\| \geq \epsilon \right) \leq \frac{1}{n^2 \epsilon} E\left(\sum_{i=1}^{p_n} \left[\frac{\partial Q_n(\beta_n)}{\partial \beta_{ni}} - E\left\{ \frac{\partial Q_n(\beta_n)}{\partial \beta_{ni}} \right\} \right]^2 \right)$$
$$= \frac{p_n}{n} = o_p(1).$$

Lemma 4. *Under* (D)*, we have*

$$\left\| \frac{1}{n} \nabla^2 Q_n(\beta_n) - D_n(\beta_n) \right\| = o_p(p_n^{-1}).$$

Proof. By Chebyshev's inequality it follows that, for any ϵ,

$$\begin{aligned} &P\left(\left\|\frac{1}{n}\nabla^2 Q_n(\beta_n) - D_n(\beta_n)\right\| \geq \frac{\epsilon}{p_n}\right) \\ &\quad \leq \frac{p_n^2}{n^2\epsilon} E\left(\sum_{i,j=1}^{p_n}\left[\frac{\partial^2 Q_n(\beta_n)}{\partial\beta_{ni}\partial\beta_{nj}} - E\left\{\frac{\partial^2 Q_n(\beta_n)}{\partial\beta_{ni}\partial\beta_{nj}}\right\}\right]^2\right) \\ &\quad = \frac{p_n^4}{n} = o_p(1). \end{aligned}$$

Lemma 5. *Suppose the penalty function $P_{\lambda_n}(|\beta_n|)$ satisfies* (A)*, the QIF $Q_n(\beta_n)$ satisfies* (D)−(F)*, and there is an open subset ω_{q_n} of $\Omega_{q_n} \in \mathbf{R}^{q_n}$ that contains the true non-zero parameter point β_s^*. When $\lambda_n \to 0$, $\sqrt{n/p_n}\lambda_n \to \infty$ and $p_n^4/n \to 0$ as $n \to \infty$, for all the $\beta_s \in \omega_{q_n}$ that satisfy $\|\beta_s - \beta_s^*\| = O_p(\sqrt{p_n/n})$ and any constant K,*

$$S\{({\beta_s}^T, 0)^T\} = \min_{\|\beta_{s^c}\| \leq K(\sqrt{p_n/n})} S\{({\beta_s}^T, \beta_{s^c}^T)^T\}, \text{ with probability tending to } 1.$$

Proof. We take $\epsilon_n = K\sqrt{p_n/n}$. It is sufficient to prove that, with probability tending to 1 as $n \to \infty$, for all the β_s that satisfy $\beta_s - \beta_s^* = O_p(\sqrt{p_n/n})$, we have for $j = q_n + 1, \ldots, p_n$,

$$\begin{aligned} \frac{\partial S_n(\beta_n)}{\partial\beta_{nj}} &> 0 \qquad \text{for} \qquad 0 < \beta_{nj} < \epsilon_n, \\ \frac{\partial S_n(\beta_n)}{\partial\beta_{nj}} &< 0 \qquad \text{for} \qquad -\epsilon_n < \beta_{nj} < 0. \end{aligned}$$

By the Taylor expansion,

$$\begin{aligned} \frac{\partial S_n(\beta_n)}{\partial\beta_{nj}} &= \frac{\partial Q_n(\beta_n)}{\partial\beta_{nj}} + nP'_{\lambda_n}(|\beta_{nj}|)\text{sign}(\beta_{nj}) \\ &= \frac{\partial Q_n(\beta_n^*)}{\partial\beta_{nj}} + \sum_{l=1}^{p_n}\frac{\partial^2 Q_n(\beta_n^*)}{\partial\beta_{nj}\partial\beta_{nl}}(\beta_{nl} - \beta_{nl}^*) \\ &\quad + \sum_{l,k=1}^{p_n}\frac{\partial^3 Q_n(\dot{\beta}_n)}{\partial\beta_{nj}\partial\beta_{nl}\partial\beta_{nk}}(\beta_{nl} - \beta_{nl}^*)(\beta_{nk} - \beta_{nk}^*) + nP'_{\lambda_n}(|\beta_{nj}|)\text{sign}(\beta_{nj}) \\ &= I_1 + I_2 + I_3 + I_4, \end{aligned}$$

where $\dot{\beta}_n$ lies between β_n and β_n^*, and a standard argument gives

$$I_1 = O_p(\sqrt{n}) = O_p(\sqrt{np_n}). \tag{A.1}$$

The second term I_2 is

$$\begin{aligned} I_2 &= \sum_{l=1}^{p_n} \left[\frac{\partial^2 Q_n(\beta_n^*)}{\partial \beta_{nj} \partial \beta_{nl}} - E\left\{ \frac{\partial^2 Q_n(\beta_n^*)}{\partial \beta_{nj} \partial \beta_{nl}} \right\} \right] (\beta_{nl} - \beta_{nl}^*) \\ &\quad + \sum_{l=1}^{p_n} \frac{1}{n} E\left\{ \frac{\partial^2 Q_n(\beta_n^*)}{\partial \beta_{nj} \partial \beta_{nl}} \right\} n(\beta_{nl} - \beta_{nl}^*) \\ &= H_1 + H_2. \end{aligned}$$

Under (D), we obtain

$$\left(\sum_{l=1}^{p_n} \left[\frac{\partial^2 Q_n(\beta_n^*)}{\partial \beta_{nj} \partial \beta_{nl}} - E\left\{ \frac{\partial^2 Q_n(\beta_n^*)}{\partial \beta_{nj} \partial \beta_{nl}} \right\} \right]^2 \right)^{1/2} = O_p(\sqrt{np_n}),$$

and by $\|\beta_n - \beta_n^*\| = O_p(\sqrt{p_n/n})$, it follows that $H_1 = O_p(\sqrt{np_n})$. Moreover,

$$|H_2| = \left| \sum_{l=1}^{p_n} \frac{1}{n} E\left\{ \frac{\partial^2 Q_n(\beta_n^*)}{\partial \beta_{nj} \partial \beta_{nl}} \right\} n(\beta_{nl} - \beta_{nl}^*) \right| \leq nO_p(1)O_p(\sqrt{\frac{p_n}{n}}) = O_p(\sqrt{np_n}).$$

This yields

$$I_2 = O_p(\sqrt{np_n}). \tag{A.2}$$

We can write

$$\begin{aligned} I_3 &= \sum_{l,k=1}^{p_n} \left[\frac{\partial^3 Q_n(\dot{\beta}_n)}{\partial \beta_{nj} \partial \beta_{nl} \partial \beta_{nk}} - E\left\{ \frac{\partial^3 Q_n(\dot{\beta}_n)}{\partial \beta_{nj} \partial \beta_{nl} \partial \beta_{nk}} \right\} \right] (\beta_{nl} - \beta_{nl}^*)(\beta_{nk} - \beta_{nk}^*) \\ &\quad + \sum_{l,k=1}^{p_n} E\left\{ \frac{\partial^3 Q_n(\dot{\beta}_n)}{\partial \beta_{nj} \partial \beta_{nl} \partial \beta_{nk}} \right\} (\beta_{nl} - \beta_{nl}^*)(\beta_{nk} - \beta_{nk}^*) \\ &= H_3 + H_4. \end{aligned}$$

By the Cauchy-Schwarz inequality, we have

$$H_3^2 \leq \sum_{l,k=1}^{p_n} \left[\frac{\partial^3 Q_n(\dot{\beta}_n)}{\partial \beta_{nj} \partial \beta_{nl} \partial \beta_{nk}} - E\left\{ \frac{\partial^3 Q_n(\dot{\beta}_n)}{\partial \beta_{nj} \partial \beta_{nl} \partial \beta_{nk}} \right\} \right]^2 \|\beta_n - \beta_n^*\|^4.$$

Under (E) and (F),

$$H_3 = O_p\left\{ \left(np_n^2 \frac{p_n^2}{n^2} \right)^{1/2} \right\} = O_p\left\{ \left(\frac{p_n^4}{n} \right)^{1/2} \right\} = o_p(\sqrt{np_n}). \tag{A.3}$$

On the other hand, under (E),

$$|H_4| \leq K_1^{1/2} p_n^2 \|\beta_n - \beta_n^*\|^2 \leq K_1^{1/2} np_n \|\beta_n - \beta_n^*\|^2 = O_p(p_n^2) = o_p(\sqrt{np_n}). \tag{A.4}$$

From (A.1)-(A.4) we have

$$\begin{aligned}\frac{\partial S_n(\beta_n)}{\partial \beta_{nj}} =& O_p(\sqrt{np_n}) + O_p(\sqrt{np_n}) + o_p(\sqrt{np_n}) + nP'_{\lambda_n}(|\beta_{nj}|)\text{sign}(\beta_{nj}) \\ =& n\lambda_n\Big\{\frac{P'_{\lambda_n}(|\beta_{nj}|)}{\lambda_n}\text{sign}(\beta_{nj}) + O_p\Big(\frac{\sqrt{p_n}}{\sqrt{n}\lambda_n}\Big)\Big\}.\end{aligned}$$

By (A) and $\sqrt{p_n}/\sqrt{n}\lambda_n \to 0$, the sign of $\partial S_n(\beta_n)/\partial\beta_{nj}$ is entirely determined by the sign of β_{nj}.

Proof of Theorem 1. Suppose $\alpha_n = \sqrt{p_n}(n^{-1/2}+a_n)$. We want to show that for any given $\epsilon > 0$, there exists a constant K such that $P\big\{\inf_{\|\mathbf{u}\|=K} S_n(\beta_n^*+\alpha_n\mathbf{u}) > S_n(\beta_n^*)\big\} \geq 1-\epsilon$. This implies with probability at least $1-\epsilon$ that there exists a local minimum $\hat{\beta}_n$ in the ball $\{\beta_n^* + \alpha_n\mathbf{u} : \|\mathbf{u}\| \leq K\}$ such that $\|\hat{\beta}_n - \beta_n^*\| = O_p(\alpha_n)$. We write

$$\begin{aligned}G_n(\mathbf{u}) =& S_n(\beta_n^*) - S_n(\beta_n^* + \alpha_n\mathbf{u}) \\ =& Q_n(\beta_n^*) - Q_n(\beta_n^* + \alpha_n\mathbf{u}) + n\sum_{j=1}^{p_n}\{P_{\lambda_n}(|\beta_{nj}^*|) - P_{\lambda_n}(|\beta_{nj}^* + \alpha_n u_j|)\} \\ \leq& Q_n(\beta_n^*) - Q_n(\beta_n^* + \alpha_n\mathbf{u}) + n\sum_{j=1}^{q_n}\{P_{\lambda_n}(|\beta_{nj}^*|) - P_{\lambda_n}(|\beta_{nj}^* + \alpha_n u_j|)\} \\ =& (I) + (II).\end{aligned}$$

By the Taylor expansion,

$$\begin{aligned}(I) &= -\Big[\alpha_n\nabla^T Q_n(\beta_n^*)\mathbf{u} + \frac{1}{2}\mathbf{u}^T\nabla^2 Q_n(\beta_n^*)\mathbf{u}\alpha_n^2 + \frac{1}{6}\nabla^T\{\mathbf{u}^T\nabla^2 Q_n(\dot{\beta}_n)\mathbf{u}\}\mathbf{u}\alpha_n^3\Big] \\ &= -I_1 - I_2 - I_3,\end{aligned}$$

where the vector $\dot{\beta}_n$ lies between β_n^* and $\beta_n^* + \alpha_n\mathbf{u}$, and

$$\begin{aligned}(II) &= -\sum_{j=1}^{q_n}\big[n\alpha_n P'_{\lambda_n}(|\beta_{nj}^*|)\text{sign}(\beta_{nj}^*)u_j + n\alpha_n^2 P''_{\lambda_n}(|\beta_{nj}^*|)u_j^2\{1+o(1)\}\big] \\ &= -I_4 - I_5.\end{aligned}$$

By Lemma 1 and the Cauchy-Schwarz inequality, I_1 is bounded, as

$$\alpha_n\nabla^T Q_n(\beta_n^*)\mathbf{u} \leq \alpha_n\|\nabla^T Q_n(\beta_n^*)\|\|\mathbf{u}\| = O_p(\sqrt{np_n}\alpha_n)\|\mathbf{u}\| = O_p(n\alpha_n^2)\|\mathbf{u}\|.$$

Under (D) and by Lemma 2,

$$\begin{aligned}I_2 &= \frac{1}{2}\mathbf{u}^T\Big[\frac{1}{n}\nabla^2 Q_n(\beta_n^*) - \frac{1}{n}E\Big\{\nabla^2 Q_n(\beta_n^*)\Big\}\Big]\mathbf{u}n\alpha_n^2 + \frac{1}{2}\mathbf{u}^T E\Big\{\nabla^2 Q_n(\beta_n^*)\Big\}\mathbf{u}\alpha_n^2 \\ &= o_p(n\alpha_n^2)\|\mathbf{u}\|^2 + \frac{n\alpha_n^2}{2}\mathbf{u}^T D_n(\beta_n^*)\mathbf{u}.\end{aligned}$$

Under (C) and $p_n^2 a_n \to 0$ as $n \to \infty$, we have

$$|I_3| = \left|\frac{1}{6}\sum_{i,j,k=1}^{p_n} \frac{\partial Q_n(\dot{\beta}_n)}{\partial\beta_{ni}\partial\beta_{nj}\partial\beta_{nk}} u_i u_j u_k \alpha_n^3\right| \le \frac{1}{6} n \Big\{ \sum_{i,j,k=1}^{p_n} M^2 \Big\}^{1/2} \|\mathbf{u}\|^3 \alpha_n^3$$
$$= O_p(p_n^{3/2}\alpha_n) n\alpha_n^2 \|\mathbf{u}\|^3 = o_p(n\alpha_n^2)\|\mathbf{u}\|^3.$$

The terms I_4 and I_5 can be bounded as

$$|I_4| \le \sum_{j=1}^{q_n} |n\alpha_n P'_{\lambda_n}(|\beta_{nj}^*|)\mathrm{sign}(\beta_{nj}^*)u_j| \le n\alpha_n a_n \sum_{j=1}^{q_n} |u_j|$$
$$\le n\alpha_n a_n \sqrt{q_n}\|\mathbf{u}\| \le n\alpha_n^2 \|\mathbf{u}\|$$

and

$$I_5 = \sum_{j=1}^{q_n} n\alpha_n^2 P''_{\lambda_n}(\beta_{nj}^*) u_j^2 \{1 + o(1)\} \le 2\max_{1\le j\le q_n} P''_{\lambda_n}(|\beta_{nj}^*|) n\alpha_n^2 \|\mathbf{u}\|^2.$$

For a sufficiently large $\|\mathbf{u}\|$, all terms in (I) and (II) are dominated by I_2. Thus G_n is negative because $-I_2 < 0$.

Proof of Theorem 2. Theorem 1 shows that there is a local minimizer $\hat{\beta}_n$ of $S_n(\beta)$ and Lemma 3 proves the sparsity property. Next we prove the asymptotic normality. By the Taylor expansion on $\nabla S_n(\hat{\beta}_s)$ at point β_s^*, we have

$$\nabla S_n(\hat{\beta}_s) = \nabla Q_n(\beta_s^*) + \nabla^2 Q_n(\beta_s^*)(\hat{\beta}_s - \beta_s^*) + \frac{1}{2}(\hat{\beta}_s - \beta_s^*)^T \nabla^2 \big\{\nabla Q_n(\dot{\beta}_n)\big\}(\hat{\beta}_s - \beta_s^*)$$
$$+ \nabla P_{\lambda_n}(\beta_s^*) + \nabla^2 P_{\lambda_n}(\ddot{\beta}_n)(\hat{\beta}_s - \beta_s^*),$$

where $\dot{\beta}_n$ and $\ddot{\beta}_n$ lie between $\hat{\beta}_s$ and β_s^*. Because $\hat{\beta}_s$ is a local minimizer, $\nabla S_n(\hat{\beta}_s) = \mathbf{0}$, we obtain

$$\frac{1}{n}\Big[\nabla Q_n(\beta_s^*) + \frac{1}{2}(\hat{\beta}_s - \beta_s^*)^T \nabla^2 \big\{\nabla Q_n(\dot{\beta}_n)\big\}(\hat{\beta}_s - \beta_s^*)\Big]$$
$$= -\frac{1}{n}\Big[\{\nabla^2 Q_n(\beta_s^*) + \nabla^2 P_{\lambda_n}(\ddot{\beta}_n)\}(\hat{\beta}_s - \beta_s^*) + \nabla P_{\lambda_n}(\beta_s^*)\Big].$$

Let $\mathbf{Z} \cong (1/2)(\hat{\beta}_s - \beta_s^*)^T \nabla^2 \big\{\nabla Q_n(\dot{\beta}_n)\big\}(\hat{\beta}_s - \beta_s^*)$ and $\mathbf{W} \cong \nabla^2 Q_n(\beta_s^*) + \nabla^2 P_{\lambda_n}(\ddot{\beta}_n)$. By the Cauchy-Schwarz inequality and under (E) and (F), we have

$$\Big\|\frac{1}{n}\mathbf{Z}\Big\|^2 \le \frac{1}{n^2}\sum_{i=1}^{n} n\|\hat{\beta}_s - \beta_s^*\|^4 \sum_{j,l,k=1}^{q_n} M^2 = O_p\Big(\frac{p_n^2}{n^2}\Big) O_p(p_n^3) = o_p(n^{-1}). \quad \text{(A.5)}$$

By Lemma 2 and under (C) and (F), we obtain

$$\lambda_i\Big\{\frac{1}{n}\mathbf{W} - D_n(\beta_s^*) - \mathbf{\Sigma}_{\lambda_n}\Big\} = o_p(p_n^{-1/2}), \qquad \text{for } i = 1, \ldots, q_n,$$

where $\lambda_i(B)$ is the ith eigenvalue of a symmetric matrix B. If $\hat{\beta}_s - \beta_s^* = O_p(\sqrt{p_n/n})$, we have

$$\Big\{\frac{1}{n}\mathbf{W} - D_n(\beta_s^*) - \mathbf{\Sigma}_{\lambda_n}\Big\}(\hat{\beta}_s - \beta_s^*) = o_p(n^{-1/2}). \tag{A.6}$$

From (A.5) and (A.6) we obtain

$$\{D_n(\beta_s^*) + \mathbf{\Sigma}_{\lambda_n}\}(\hat{\beta}_s - \beta_s^*) + \mathbf{b}_n = -\frac{1}{n}\nabla Q_n(\beta_s^*) - o_p(n^{-1/2}), \tag{A.7}$$

and from (A.7) we have

$$\begin{aligned}
&\sqrt{n}B_n D_n^{-1/2}(\beta_s^*)\{D_n(\beta_s^*) + \mathbf{\Sigma}_{\lambda_n}\}\big[(\hat{\beta}_s - \beta_s^*) + \{D_n(\beta_s^*) + \mathbf{\Sigma}_{\lambda_n}\}^{-1}\mathbf{b}_n\big] \\
&= \sqrt{n}B_n D_n^{-1/2}(\beta_s^*)\big[\{D_n(\beta_s^*) + \mathbf{\Sigma}_{\lambda_n}\}(\hat{\beta}_s - \beta_s^*) + \mathbf{b}_n\big] \\
&= -\frac{1}{\sqrt{n}}B_n D_n^{-1/2}(\beta_s^*)\nabla Q_n(\beta_s^*) - o_p\{B_n D_n^{-1/2}(\beta_s^*)\}.
\end{aligned}$$

As the last term is $o_p(1)$, we only consider the first term denoted by

$$Y_{ni} = \frac{1}{\sqrt{n}}B_n D_n^{-1/2}(\beta_s^*)\nabla Q_{ni}(\beta_s^*), \qquad \text{for } i = 1, \ldots, n.$$

We show that Y_{ni} satisfies the conditions of the Lindeberg-Feller Central Limit Theorem. By Lemma 1, (D), and $B_n B_n^T \to F$, we have

$$\begin{aligned}
E\|Y_{n1}\|^4 &= \frac{1}{n^2}E\|B_n D_n^{-1/2}(\beta_s^*)\nabla Q_{ni}(\beta_s^*)\|^4 \\
&\le \frac{1}{n^2}\lambda_{max}(B_n B_n^T)\lambda_{max}\{D_n(\beta_s^*)\}E\|\nabla^T Q_n(\beta_s^*)\nabla Q_n(\beta_s^*)\|^2 \\
&= O(p_n^2 n^{-2}),
\end{aligned} \tag{A.8}$$

and by Chebyshev's inequality

$$P(\|Y_{n1}\| > \epsilon) \le \frac{E\|Y_{n1}\|^2}{\epsilon} \le \frac{E\|B_n D_n^{-1/2}(\beta_s^*)\nabla Q_{ni}(\beta_s^*)\|^2}{n\epsilon} = O(n^{-1}). \tag{A.9}$$

From (A.8) and (A.9) and $p_n^4/n \to 0$ as $n \to \infty$, we obtain

$$\begin{aligned}
\sum_{i=1}^n E\|Y_{ni}\|^2 \mathbf{1}\{\|Y_{ni}\| > \epsilon\} &\le n\{E\|Y_{n1}\|^4\}^{1/2}\{P(\|Y_{n1}\| > \epsilon)\}^{1/2} \\
&\le nO(p_n n^{-1})O(n^{-1/2}) = O(p_n n^{-1/2}) = o(1).
\end{aligned}$$

On the other hand, as $B_n B_n^T \to F$ we have

$$\sum_{i=1}^n cov(Y_{ni}) = n \cdot cov(Y_{n1}) = cov\{B_n D_n^{-1/2}(\beta_s^*)\nabla Q_n(\beta_s^*)\} \to F.$$

It follows that the Lindeberg condition is satisfied and then the Lindeberg-Feller central limit theorem gives

$$\sqrt{n}B_nD_n^{-1/2}(\beta_s^*)\{D_n(\beta_s^*)+\Sigma_{\lambda_n}\}\big[(\hat{\beta}_s-\beta_s^*)+\{D_n(\beta_s^*)+\mathbf{\Sigma}_{\lambda_n}\}^{-1}\mathbf{b}_n\big]\xrightarrow{\mathrm{d}}N(0,F).$$

Proof of Lemma 1. Let $\hat{\beta}_{n\lambda_o}=(\hat{\beta}_{s\lambda_o}^T,\hat{\beta}_{s^c\lambda_o}^T)^T$ be an estimator of $\beta_n=(\beta_s^T,\beta_{s^c}^T)^T$. The oracle property of the penalized QIF ensures that, with probability tending to 1, $\hat{\beta}_{s\lambda_o}$ satisfies

$$S_n'(\hat{\beta}_{s\lambda_o})=Q_n'(\hat{\beta}_{s\lambda_o})+\mathbf{b}_n(\hat{\beta}_{s\lambda_o})=0, \tag{A.10}$$

where $\mathbf{b}_n=\{P_{\lambda_n}'(|\beta_{n1}|)\mathrm{sign}(\beta_{n1}),\ldots,P_{\lambda_n}'(|\beta_{nq_n}|)\mathrm{sign}(\beta_{nq_n})\}^T$. By (F), $P(|\hat{\beta}_{s\lambda_o}|>a\lambda_o)\longrightarrow 1$, which implies that $P(\mathbf{b}_n(\hat{\beta}_{s\lambda_o})=0)\longrightarrow 1$. Therefore with probability tending to 1, (A.10) leads to $Q_n'(\hat{\beta}_{s\lambda_o})=0$. This implies that $\hat{\beta}_{s\lambda_o}$ is the same as $\hat{\beta}_s^*$, the oracle estimator for the non-zero coefficients. It immediately follows that, with probability tending to 1, $BIQIF_{\lambda_o}=Q_n'(\hat{\beta}_{s\lambda_o})+q_n\log(n)=Q_n'(\hat{\beta}_s^*)+q_n\log(n)=BIQIF_{\Upsilon_T}$.

Proof of Lemma 2. The proof of Lemma 2 consists of different cases for underfitted or overfitted models. We show that Lemma 2 holds for each case.

For underfitted models, it follows by Lemma 1 that

$$\begin{aligned}\frac{BIQIF_{\lambda_o}}{n} &= \bar{g}_n(\hat{\beta}_{\lambda_o})^T\bar{C}_n^{-1}(\hat{\beta}_{\lambda_o})\bar{g}_n(\hat{\beta}_{\lambda_o})+q_n\frac{\log(n)}{n}\\ &\xrightarrow{\mathrm{P}}\bar{g}_n(\beta_{\Upsilon_T})^T\bar{C}_n^{-1}(\beta_{\Upsilon_T})\bar{g}_n(\beta_{\Upsilon_T}).\end{aligned}$$

In addition, since $\Upsilon_\lambda\not\supseteq\Upsilon_T$, we have

$$\begin{aligned}\frac{BIQIF_{\lambda_n}}{n}&=\bar{g}_n(\hat{\beta}_{\lambda_n})^T\bar{C}_n^{-1}(\hat{\beta}_{\lambda_n})\bar{g}_n(\hat{\beta}_{\lambda_n})+df_{\lambda_n}\frac{\log(n)}{n}\geq\bar{g}_n(\hat{\beta}_{\lambda_n})^T\bar{C}_n^{-1}(\hat{\beta}_{\lambda_n})\bar{g}_n(\hat{\beta}_{\lambda_n})\\ &\geq\min_{\Upsilon:\Upsilon\not\supseteq\Upsilon_T}\bar{g}_n(\hat{\beta}_\Upsilon)^T\bar{C}_n^{-1}(\hat{\beta}_\Upsilon)\bar{g}_n(\hat{\beta}_\Upsilon)\\ &\xrightarrow{\mathrm{P}}\min_{\Upsilon:\Upsilon\not\supseteq\Upsilon_T}\bar{g}_n(\beta_\Upsilon)^T\bar{C}_n^{-1}(\beta_\Upsilon)\bar{g}_n(\beta_\Upsilon)>\bar{g}_n(\beta_{\Upsilon_T})^T\bar{C}_n^{-1}(\beta_{\Upsilon_T})\bar{g}_n(\beta_{\Upsilon_T}).\end{aligned}$$

Therefore,

$$P(\inf_{\lambda_n\in\Lambda_-}\frac{BIQIF_{\lambda_n}}{n}>\frac{BIQIF_{\lambda_o}}{n})=P(\inf_{\lambda_n\in\Lambda_-}BIQIF_{\lambda_n}>BIQIF_{\lambda_o})\longrightarrow 1.$$

For overfitted models, we have

$$\begin{aligned}\inf_{\lambda_n\in\Lambda_+}(BIQIF_{\lambda_n}-BIQIF_{\lambda_o})&=\inf_{\lambda_n\in\Lambda_+}(Q_n(\hat{\beta}_{\lambda_n})-Q_n(\hat{\beta}_{\lambda_o})+(df_{\lambda_n}-q_n))\log(n)\\ &\geq\inf_{\lambda_n\in\Lambda_+}(Q_n(\hat{\beta}_{\lambda_n})-Q_n(\hat{\beta}_{\lambda_o}))+\log(n)\\ &\geq\min_{\Upsilon:\Upsilon\supset\Upsilon_T}(Q_n(\hat{\beta}_\Upsilon)-Q_n(\hat{\beta}_{\Upsilon_T}))+\log(n).\end{aligned}$$

Since $Q_n(\hat{\beta}_\Upsilon) - Q_n(\hat{\beta}_{\Upsilon_T})$ has an asymptotic $\chi^2_{df_\Upsilon - q_n}$ distribution, $\min_{\Upsilon:\Upsilon\supset\Upsilon_T}$ $(Q_n(\hat{\beta}_\Upsilon) - Q_n(\hat{\beta}_{\Upsilon_T})) = O_p(1)$ and, with $\log(n)$ divergent, we have $P(\inf_{\lambda_n\in\Lambda_+}$ $BIQIF_{\lambda_n} > BIQIF_{\lambda_o}) \longrightarrow 1$.

Online Supplementary Materials

The R-coding for the binary simulation studies is given in the online supplemental material available at http://www.stat.sinica.edu/statistica. The R-coding for the model selection of the basis matrices for the correlation matrix, by Zhou and Qu (2012), is in the website, `https://publish.illinois.edu/anniequ/files/2013/01/Basis_matrices_selection.pdf`.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proc. Second International Symposium on Information Theory* (Edited by B. N. Petrov and F. Csaki), 267-281. Akademial Kiado, Budapest.

Antoniadis, A. (1997). Wavelets in statistics: A review (with discussion). *J. Italian Statist. Assoc.* **6**, 97-144.

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

Dziak, J. J. (2006). Penalized quadratic inference functions for variable selection in longitudinal research. Ph.D. dissertation, Pennsylvania State University, PA.

Dziak, J. J., Li, R. and Qu, A. (2009). An overview on quadratic inference function approaches for longitudinal data. *Frontiers of Statistics, Vol 1: New Developments in Biostatistics and Bioinformatics* (Edited by J. Fan, J. S. Liu and X. Lin), 49-72. World Scientific Publishing.

Fan, J., Huang, T. and Li, R. (2007). Analysis of longitudinal data with semiparametric estimation of covariance function. *J. Amer. Statist. Assoc.* **102**, 632-641.

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

Fan, J. and Li, R. (2006). Statistical challenges with high dimensionality: Feature selection in knowledge discovery. *Proceedings of the International Congress of Mathematicians* **3** (Edited by M. Sanz-Sole, J. Soria, J. L. Varona and J. Verdera), 595-622. European Mathematical Society.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high-dimensional feature space. *Statist. Sinica* **20**, 101-148.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32**, 928-961.

Fan, J. and Wu, Y. (2008). Semiparametric estimation of covariance matrices for longitudinal data. *J. Amer. Statist. Assoc.* **103**, 1520-1533.

Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35**, 109-148.

Fu, W. J. (2003). Penalized estimating equations. *Biometrics* **59**, 126-132.

Gao, X., Pu, Q., Wu, Y. and Xu, H. (2012). Tuning parameter selection for penalized likelihood estimation of gaussian graphical model. *Statist. Sinica* **22**, 1123-1146.

Hall, P., Müller, H. G. and Wang, J. L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Ann. Statist.* **34**, 1493-1517.

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

Hunter, D. R. and Li, R. (2005). Variable selection using MM algorithms. *Ann. Statist.* **33**, 1617-1642.

James, G. and Hastie, T. (2001). Functional linear discriminant analysis for irregularly sampled curves. *J. Roy. Statist. Soc. Ser. B* **63**, 533-550.

James, G. and Sugar, C. A. (2003). Clustering for sparsely sampled functional data. *J. Amer. Statist. Assoc.* **98**, 397-408.

Jiang, C. R. and Wang, J. L. (2010). Covariate adjusted functional principal components analysis for longitudinal data. *Ann. Statist.* **38**, 1194-1226.

Johnson, B., Lin, D. Y. and Zeng, D. (2008). Penalized estimating functions and variable selection in semiparametric regression models. *J. Amer. Statist. Assoc.* **103**, 672-680.

Li, Y. (2011). Efficient semiparametric regression for longitudinal data with nonparametric covariance estimation. *Biometrika* **98**, 355-370.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalised linear models. *Biometrika* **73**, 12-22.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics* **15**, 661-675.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics* **57**, 120-125.

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

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

Stoner, J. A. (2000). Analysis of clustered data: A combined estimating equations approach. Ph.D. dissertation, University of Washington, WA.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

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

Wang, H., Li, B. and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *J. Roy. Statist. Soc. Ser. B* **71**, 671-683.

Wang, H., Li, R. and Tsai, C. L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94**, 553-568.

Wang, L. and Qu, A. (2009). Consistent model selection and data-driven smooth tests for longitudinal data in the estimating equations approach. *J. Roy. Statist. Soc. Ser. B* **71**, 177-190.

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

Xu, P., Wu, P., Wang, T. and Zhu, L. (2010). A GEE based shrinkage estimation for the penalized linear model in longitudinal data analysis. Manuscript.

Xue, L., Qu, A. and Zhou, J. (2010). Consistent model selection for marginal generalized additive model for correlated data. *J. Amer. Statist. Assoc.* **105**, 1518-1530.

Yao, F., Müller, H. G. and Wang, J. L. (2005). Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.* **100**, 577-590.

Zhang, C. H. (2007). Penalized linear unbiased selection. Technical Report, No.2007-003, Department of Statistics, Rutgers University, NJ.

Zhang, Y., Li, R. and Tsai, C. L. (2010). Regularization parameter selections via generalized information criterion. *J. Amer. Statist. Assoc.* **105**, 312-323.

Zhou, J. and Qu, A. (2012). Informative estimation and selection of correlation structure for longitudinal data. *J. Amer. Statist. Assoc.* **107**, 701-710.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418-1429.

Zou, H. and Zhang, H. H. (2009). On the adaptive elastic net with a diverging number parameters. *Ann. Statist.* **37**, 1733-1751.

Department of Statistics, University of Illinois at Urbana-Champaign, 725 South Wright Street, Champaign, Illinois 61820 USA.

E-mail: cho75@illinois.edu

Department of Statistics, University of Illinois at Urbana-Champaign, 725 South Wright Street, Champaign, Illinois 61820 USA.

E-mail: anniequ@illinois.edu

(Received February 2011; accepted August 2012)