MODEL SELECTION FOR CORRELATED DATA AND MOMENT SELECTION FROM HIGH-DIMENSIONAL MOMENT CONDITIONS

BY

HYUN KEUN CHO

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Statistics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2013

Urbana, Illinois

Doctoral Committee:

Professor Annie Qu, Chair
Professor Jeffrey Douglas
Professor Xiaofeng Shao
Professor Douglas Simpson

Abstract

High-dimensional correlated data arise frequently in many studies. My primary research interests lie broadly in statistical methodology for correlated data such as longitudinal data and panel data. In this thesis, we address two important but challenging issues: model selection for correlated data with diverging number of parameters and consistent moment selection from high-dimensional moment conditions.

Longitudinal data arise frequently in biomedical and genomic research where repeated measurements within subjects are correlated. 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.

The generalized method of moments (GMM) approach combines moment conditions optimally to obtain efficient estimation without specifying the full likelihood function. However, the GMM estimator could be infeasible when the number of moment conditions exceeds the sample size. This

research intends to address issues arising from the motivating problem where the dimension of estimating equations or moment conditions far exceeds the sample size, such as in selecting informative correlation structure or modeling for dynamic panel data. We propose a Bayesian information type of criterion to select the optimal number of linear combinations of moment conditions. In theory, we show that the proposed criterion leads to consistent selection of the number of principal components for the weighting matrix in the GMM. Monte Carlo studies indicate that the proposed method outperforms existing methods in the sense of reducing bias and improving the efficiency of estimation. We also illustrate a real data example for moment selection using dynamic panel data models.

I dedicate this dissertation to my parents.

You have always been, and will always be, my inspiration.

Acknowledgments

Foremost, I would like to express my sincere gratitude to my advisor Dr. Annie Qu for her continuous support of my Ph.D. study and research, for her patience, motivation, enthusiasm, and immense knowledge. Her guidance helped me throughout my period of research and thesis writing. I could not have imagined having a better advisor and mentor for my Ph.D. study.

A special thanks to my thesis committee: Dr. Jeffrey Douglas, Dr. Xiaofeng Shao, and Dr. Douglas Simpson for providing valuable feedback and counsel during particularly difficult personal times that threatened to halt my entire research. Their encouraging words helped me work through all the distractions and noise. In addition to my advisor and committee, I would like to acknowledge my collaborators, Dr. Peng Wang and Dr. Jianhui Zhou for useful discussions about estimation procedures for longitudinal data, permission to use simulation code in my thesis as well as valuable and pleasant collaborative experiences.

Many thanks are given to my fellow students at the Department of Statistics. Particularly, I would like to thank Yeonjoo Park, Peibei Shi, Xinxin Shu, Samuel Ye, and Xianyang Zhang for their support, encouragement and friendship. And to the wonderful staff, Melissa Banks, you have always gone far beyond your job description to help us as graduate students. You are always available to take care of my myriad of any issues. Your efforts and time have been greatly appreciated.

Finally and most importantly, I want to thank my family for their constant support and prayers throughout my life but particulary during my years of study and research. Their unconditional love, encouragement, and dedication carried me on through difficult times.

Table of Contents

Chapter 1

Introduction

Correlated data arise frequently in many studies where repeated measurements are taken from the same subject over time. Variable selection is fundamental to extracting important relevant predictors from large data sets, as inclusion of high-dimensional redundant variables can hinder efficient estimation and inference for the non-zero coefficients. In the longitudinal data framework, however, the research on variable selections is still limited when the dimension of parameters diverges.

In Chapter 2, we propose a penalized quadratic inference function (QIF) 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 possesses desirable features of the SCAD such 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 follows the asymptotic normal distribution.

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.

Another important advantage of our approach is in tuning parameter selection. The selection of the tuning parameter plays an important role in order to achieve optimal performance in model

selection. We provide an effective tuning parameter selection procedure based on the Bayesian information quadratic inference function criterion, and show that the proposed tuning parameter selector leads to consistent model selection and estimation for regression parameters.

In Chapter 3, we are motivated by the problem where the dimension of estimating equations or moment conditions far exceeds the sample size. For example, for correlated data, the dimension of moment conditions depends on the number of basis matrices associated with the inverse of the correlation matrix, and can be larger than the sample size. For dynamic panel data example, a large dimension of valid moment conditions can be generated based on the first-order moments.

The generalized method of moments (GMM, Hansen, 1982) is widely applicable when the likelihood function is difficult to specify, while moment conditions are easy to formulate. The GMM is powerful as it optimally combines valid moment conditions, and so is able to achieve estimation efficiency. However, the GMM could perform poorly if there are too many moment conditions relative to the sample size, due to limitation in finite samples.

To solve this problem, we examine the key element of the GMM: the optimal weighting matrix which is the inverse of the sample covariance matrix of moment conditions. In fact, the sample covariance matrix could be problematic when the dimension of the matrix is large due to the following two reasons: i) the sample covariance matrix is not of full rank if the dimension of moment conditions exceeds the sample size; ii) even if the sample covariance matrix is invertible, the estimation of its inverse could be biased with high variation when the number of moment conditions is close to the sample size. The singularity problem of the sample covariance matrix makes the GMM estimator infeasible or unstable.

We propose a new objective function based on a Bayesian information type of criterion which selects an optimal number of linear combinations of the moment conditions. This allows one to reduce the dimensionality of available moment conditions while retaining most of the important information from data. In theory, we show that the proposed criterion can select an optimal number of principal components consistently without loss of efficiency, when both the number of moment conditions and the sample size go to infinity. The proposed criterion can be applied to estimate the inverse of the covariance matrix in high-dimensional data settings, in addition to solving moment selection problems arising from dynamic panel data.

Chapter 2

Model Selection for Correlated Data with Diverging Number of Parameters

2.1 Introduction

Longitudinal data arise frequently in biomedical and health studies where repeated measurements are taken from the same subject. 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, which 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 when the covariates are high-dimensional. Including high-dimensional redundant variables can hinder efficient estimation and distort inference for the non-zero coefficients for high-dimensional data. 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.

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 rest of Chapter 2 is organized as follows. Section 2.2 briefly describes the quadratic inference function for longitudinal data. Section 2.3 introduces the penalized quadratic inference function and provides the asymptotic properties for variable selection when the number of parameters diverges. Section 2.4 presents two algorithms to implement the penalized QIF approach and a tuning parameter selector. Section 2.5 reports on simulation studies for binary responses and provides a data example from a periodontal disease study. Section 2.6 provides concluding remarks and discussion. All necessary lemmas and theoretical proofs are in Section 2.7.

2.2 Estimation Procedures 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 im-

prove 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 $\bar{C}_n(\beta_n) = \frac{1}{n}\sum_{i=1}^{n} g_i(\beta_n) g_i^T(\beta_n)$ 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).

2.3 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

$$S_n(\beta_n) = Q_n(\beta_n) + n\sum_{j=1}^{p_n} P_{\lambda_n}(|\beta_{nj}|). \tag{2.4}$$

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

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

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 2.5.1.

We obtain the estimator $\hat{\beta}_n$ by minimizing $S_n(\beta_n)$ in (2.4). Minimizing (2.4) 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 (2.4). 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 $(0 \leq q_n \leq p_n)$ 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 (2.4). 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

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

and the second derivative of the QIF satisfies

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

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

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

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 $\min_{1 \leq j \leq q_n} |\beta_{nj}|/\lambda_n$ 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 (2.4).

Theorem 2.1. *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 $\mathbf{b}_n = \{P'_{\lambda_n}(|\beta_{n1}|)\text{sign}(\beta_{n1}), ..., P'_{\lambda_n}(|\beta_{np_n}|)\text{sign}(\beta_{nq_n})\}^T$ and $\mathbf{\Sigma}_{\lambda_n} = diag\{P''_{\lambda_n}(\beta_{n1}), ..., P''_{\lambda_n}(\beta_{nq_n})\}$, where $\text{sign}(\alpha) = I(\alpha > 0) - I(\alpha < 0)$.

Theorem 2.2. *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 $B_n B_n^T \to F$, where F is a fixed dimensional constant matrix and $D_n(\beta_s^*) = \text{E}\{n^{-1}\nabla^2 Q_n(\beta_s^*)\}$,

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

In addition, if $\mathbf{\Sigma}_{\lambda_n} \to 0$ and $\mathbf{b}_n \to 0$ as $n \to \infty$, $\sqrt{n}B_n D_n^{1/2}(\beta_s^*)(\hat{\beta}_s - \beta_s^*) \xrightarrow{\text{d}} N(0, F)$. Theorem 2.2 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 Section 2.7.

2.4 Implementation

2.4.1 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 (2.4) with the unpenalized QIF estimator as the initial value $\beta^{(0)}$. If $\beta^{(k)} = \big(\beta_1^{(k)}, ..., \beta_{p_n}^{(k)}\big)^T$ 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 $j = 1, ..., q_k$ and $\beta_j^{(k+1)} = 0$ for $j = q_{k+1}, ..., p_n$, 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 $\beta_j^{(k)} \neq 0$,

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

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

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

Consequently, the penalized QIF in (2.4) can be approximated by

$$Q_n(\beta^{(k)}) + \nabla Q_n(\beta^{(k)})^T(\beta_s - \beta_s^{(k)}) + \frac{1}{2}(\beta_s - \beta_s^{(k)})^T \nabla^2 Q_n(\beta^{(k)})(\beta_s - \beta_s^{(k)}) + \frac{1}{2} n \beta_s^T \Pi(\beta^{(k)}) \beta_s,$$

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 $\Pi(\beta^{(k)}) = diag\Big\{ P'_{\lambda_n}(|\beta_1^{(k)}|)/|\beta_1^{(k)}| \ , ..., P'_{\lambda_n}(|\beta_{q_k}^{(k)}|)/|\beta_{q_k}^{(k)}| \Big\}$. The non-zero component $\beta_s^{(k+1)}$ at the $k+1$ step can be obtained by minimizing the quadratic function in (2.4.1) 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 $\|\beta_s^{(k+1)} - \beta_s^{(k)}\| < 10^{-7}$.

2.4.2 Linear Approximation Method

We also consider an alternative algorithm based on the linear approximation for the first part of the PQIF in (2.4). 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 (2.4) 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 (2.4) can be approximated by

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

Let $U = \tilde{C}_N^{-\frac{1}{2}} G(\hat{\beta}_Q)\hat{\beta}_Q$ and $T = \tilde{C}_N^{-\frac{1}{2}} G(\hat{\beta}_Q)$. 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 can be applied in R using the SCAD penalty.

In this way we approximate two parts of the objection function in (2.4). The local quadratic approximation method approximates the SCAD penalty function, while the linear approximation method approximates the first term of the QIF in (2.4). 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.

2.4.3 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

$$\text{BIQIF}_{\lambda_n} = Q_n(\hat{\beta}_{\lambda_n}) + df_{\lambda_n} \log(n), \tag{2.5}$$

where $\hat{\beta}_{\lambda_n}$ is the marginal regression parameter estimated by minimizing the penalized QIF in (2.4) 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 (2.5).

To investigate consistency, let $\Upsilon = \{j_1, ..., j_q\}$ be an arbitrary candidate model that contains predictors $j_1, \ldots, j_q$ $(1 \leq q \leq p_n)$ 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 $\Upsilon_F = \{1, ..., p_n\}$ and $\Upsilon_T = \{1, ..., q_n\}$ 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 2.1. *If (A)-(F) hold,* $P(BIQIF_{\lambda_o} = BIQIF_{\Upsilon_T}) \longrightarrow 1$.

Lemma 2.2. *If (A)-(F) hold,* $P(inf_{\lambda_n \in \Lambda_- \cup \Lambda_+} BIQIF_{\lambda_n} > BIQIF_{\lambda_o}) \longrightarrow 1$.

Lemmas 2.1 and 2.2 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 Section 2.7.

2.4.4 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 transfor-

mation, 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 $\dot{\mu}_i^* = H_i\dot{\mu}_i$, $\mu_i^* = H_i\mu_i$, $y_i^* = H_iy_i$, and $A_i^* = H_iA_iH_i^T$. The QIF estimator with unbalanced data is obtained based on the transformed extended score vector $\bar{g}_n^*(\beta_n) = \frac{1}{n}\sum_{i=1}^n g_i^*(\beta_n)$. 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.

2.5 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.

2.5.1 Correlated Binary Response

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

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

where $X_{ij} = \big(x_{ij}^{(1)}, ..., x_{ij}^{(p_n)}\big)^T$ and $\beta = (\beta_1, ..., \beta_{p_n})^T$. 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), $\mathbf{P}_{\lambda_\mathbf{n}}(|\beta_\mathbf{n}|) = (\mathbf{P}_{\lambda_\mathbf{n}}(|\beta_{\mathbf{n1}}|), ..., \mathbf{P}_{\lambda_\mathbf{n}}(|\beta_{\mathbf{np_n}}|))^\mathbf{T}$ with $P_{\lambda_n}(\cdot)$ a SCAD penalty, and $\text{sign}(\beta_n) = (\text{sign}(\beta_{n1}) \ , ..., \text{sign}(\beta_{np_n}))^T$; 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 $(k = 1, ..., p_n)$ 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 $\{0.01, 0.02, ..., 0.10\}$.

The simulation results from the model selection and the mean square errors (MSE) of estimation are provided in Table 2.1. Table 2.1 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 2.1 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 2.1 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 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.

2.5.2 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 2.2 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.

2.6 Discussion

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 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.

2.7 Proofs of Theorems and Lemmas

Lemma 2.3 *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,

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

Lemma 2.4 *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,

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

Lemma 2.5 *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,*

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

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 $j = q_n + 1, ..., p_n$,

$$\frac{\partial S_n(\beta_n)}{\partial \beta_{nj}} > 0 \qquad \text{for} \qquad 0 < \beta_{nj} < \epsilon_n,$$

$$\frac{\partial S_n(\beta_n)}{\partial \beta_{nj}} < 0 \qquad \text{for} \qquad -\epsilon_n < \beta_{nj} < 0.$$

By the Taylor expansion,

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

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

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

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,

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

This yields

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

We can write

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

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),

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

On the other hand, under (E),

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

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

Proof of Theorem 2.1

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

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

By Lemma 2.1 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 2.2,

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

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

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

The terms I_4 and I_5 can be bounded as

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

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

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 2.2

Theorem 2.1 shows that there is a local minimizer $\hat{\beta}_n$ of $S_n(\beta)$ and Lemma 2.3 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

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

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

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

Let $\mathbf{Z} \cong \frac{1}{2}(\hat{\beta}_s - \beta_s^*)^T \nabla^2 \big\{\nabla Q_n(\dot{\beta}_n)\big\}(\hat{\beta}_s - \beta_s^*)$ 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

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

By Lemma 2.2 and under (C) and (F), we obtain

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

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

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

We show that Y_{ni} satisfies the conditions of the Lindeberg-Feller Central Limit Theorem. By Lemma 2.1, (D), and $B_nB_n^T\to F$, we have

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

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

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

Proof of Lemma 2.1

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 $\mathbf{b}_n = \{P_{\lambda_n}'(|\beta_{n1}|)\mathrm{sign}(\beta_{n1}), ..., P_{\lambda_n}'(|\beta_{nq_n}|)\mathrm{sign}(\beta_{nq_n})\}^T$. 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 2.2

The proof of Lemma 2.2 consists of different cases for underfitted or overfitted models. We show that Lemma 2.2 holds for each case.

For underfitted models, it follows by Lemma 2.1 that

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

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

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

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 simulation studies for binary responses is given in the online supplemental material available at http://www.stat.sinica.edu/statistica.

Table 2.1: 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

Table 2.2: 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

Chapter 3

Consistent Moment Selection from High-Dimensional Moment Conditions

3.1 Introduction

The generalized method of moments (GMM, Hansen, 1982) is widely applicable when the likelihood function is difficult to specify, while moment conditions are easy to formulate. The GMM is powerful as it optimally combines valid moment conditions and is able to achieve estimation efficiency. However, the GMM could perform poorly if there are too many moment conditions relative to the sample size, due to limitation in finite samples (Newey and Smith, 2004). We are motivated by the problem where the dimension of estimating equations or moment conditions far exceeds the sample size. For example, in modeling dynamic panel data, a large dimension of valid moment conditions can be generated based on the first-order moments (Anderson and Hsiao, 1981; Han, Orea, and Schmidt, 2005; Han and Phillips, 2006). For longitudinal data, the dimension of moment conditions depends on the number of basis matrices to approximate an inverse of the correlation matrix (Qu, Lindsay, and Li, 2000), which can be larger than the sample size.

The key component of the GMM is the optimal weighting matrix, which is the inverse of the sample covariance matrix of moment conditions. However, the sample covariance matrix could be problematic when the dimension is large due to the following two reasons: i) the sample covariance matrix is not of full rank if the dimension of moment conditions exceeds the sample size; ii) even if the sample covariance matrix is invertible, the estimation of its inverse could be biased with high variation when the number of moment conditions is close to the sample size. Donald and Newey (2001) and Donald, Imbens, and Newey (2009) proposed selecting moment conditions based on the criterion of minimizing the mean square error of the estimator. However, their criterion involves inverting the sample covariance matrix, which could be infeasible if the covariance matrix

is ill-conditioned, as indicated in the above two cases.

In recent years, estimating the covariance matrix Σ and its inverse has drawn a lot of attention for the high-dimensional data setting. For example, Bickel and Levina (2008), Rothman, Levina, and Zhu (2009) and Cai and Liu (2011) proposed element-wise shrinkage and thresholding procedures to estimate Σ^{-1}. Meinshausen and Bühlmann (2006) introduced neighborhood selection for high-dimensional graphs via the lasso penalty. In addition, Friedman, Hastie, and Tibshirani (2007), Peng et al. (2009), Witten, Friedman, and Simon (2011) and Danaher, Wang, and Witten (2012) solved the graphical lasso problem through estimating the precision matrix Σ^{-1}. Most of these methods utilize sparsity structure assuming that the majority of off-diagonal elements are zero; however, they do not provide strategies for solving the matrix singularity problem.

To estimate the large dimensional covariance matrix under a more general framework without sparsity assumptions, various dimension reduction strategies through matrix decomposition have been proposed. For example, Wu and Pourahmadi (2003), Huang et al. (2006) and Pourahmadi (2007) employed regularized regression based on a modified Cholesky decomposition; Magdon-Ismail and Purnell (2011) applied a low-rank perturbation of a diagonal matrix to estimate Σ^{-1} for a Gaussian mixture model; and Fan, Fan, and Lv (2008) and Fan, Liao, and Mincheva (2011) developed a factor model to estimate the invertible covariance matrix. In addition, Luo (2011) proposed a general framework for low-rank approximation and sparse covariance structures simultaneously. However, these methods do not directly address how to extract important information from a large-dimensional matrix which is either singular or close to singular.

The singularity problem of the sample covariance makes the GMM estimator infeasible or unstable. When there are many valid moment conditions available, subset moment selection methods have been developed. Gallant and Tauchen (1996), Andrews and Lu (2001), Donald and Newey (2001), Donald, Imbens, and Newey (2009) and Okui (2009) proposed to eliminate the least useful moment conditions to reduce the overall number of moment conditions. However, selecting a subset of moment conditions requires prior information of the moment conditions. In addition, information from the unselected moment conditions is lost for parameter estimation.

To circumvent this problem, Doran and Schmidt (2006) proposed to combine all available moment conditions using principle components analysis. They apply spectral decomposition of

the covariance matrix for the moment conditions, and select the first several largest eigenvalues which contribute a fixed percentage of the total sum of eigenvalues. However, the fixed percentage criterion is arbitrary, and does not guarantee desirable asymptotic properties such as estimation consistency and efficiency. This is also confirmed by our simulations studies, in that their selection criterion procedure produces less accurate and efficient estimators compared to the oracle estimator when the true information is known. This is probably because the fixed selection criterion is not adaptive to different variations of data, and therefore their selected weighting matrix estimation could be far off from the optimal weighting matrix.

We propose a new objective function based on a Bayesian information type of criterion (Schwarz, 1978) which selects an optimal number of linear combinations of the moment conditions. In theory, we show that the proposed criterion can select the number of principal components consistently without loss of efficiency, when both the number of moment conditions and the sample size go to infinity. In addition to solving moment selection problems such as in the GMM, the proposed criterion can also be applied to estimate the inverse of the covariance matrix arising from high-dimensional data applications. The proposed method performs well in the sense of reducing bias and improving efficiency of the GMM estimation, and is especially effective when the dimension of moment conditions is high. Furthermore, it is capable of incorporating a set of preselected important moment conditions in addition to selecting the optimal linear combinations of the remaining moment conditions. Including the preselected moment conditions has the advantage of preventing any information loss from these important moment conditions.

The rest of Chapter 3 is organized as follows. Section 3.2 describes dynamic panel data models and the quadratic inference function for correlated data. Section 3.3 introduces a new moment selection method which provides an objective moment selection criterion and its asymptotic properties. Section 3.4 illustrates various simulation studies and compares different methods using the dynamic panel model for Fortune 500 data. Section 3.5 provides concluding remarks and discussion. All proofs of the lemma and theory are provided in Section 3.6.

3.2 Dynamic Panel Data Models and Quadratic Inference Function

In this section, we illustrate two motivating examples where the dimension of moment conditions exceeds the sample size. The first one is motivated by generating valid moment conditions in dynamic panel data models, and the second one is motivated by correlation structure selection using the quadratic inference function.

3.2.1 Simple Dynamic Panel Models

Dynamic panel data models are widely used in economics applications (e.g., Arellano and Bond, 1991; Blundell and Bond, 1998; Bond, 2002). The important feature of dynamic panel models is that the lagged values observed from the previous responses can also be included as part of the explanatory variables. In addition, errors in the dynamic panel models are assumed to contain time-invariant subject-specific effects in addition to the random errors. We will provide the background and construction of many moment conditions arising from the standard assumptions of dynamic panel modeling.

Suppose the dependent variable y_{ij} for the ith subject ($i = 1, ..., n$) is repeated measured at time points $j = 1, ..., m$, where observations from different subjects are independent. Without loss of generality, we provide a simple dynamic panel model whose only explanatory variable is the previous dependent variable y_{ij-1},

$$y_{ij} = \rho y_{ij-1} + u_{ij},$$

where $u_{ij} = \eta_i + \varepsilon_{ij}$ and $|\rho| < 1$.

Even for the above simple dynamic panel model, we can generate many moment conditions based on the following assumptions commonly adopted in the dynamic panel data literature:

$(\boldsymbol{A1})$ $E(\eta_i) = 0$ and $E(\varepsilon_{ij}) = 0$,

$(\boldsymbol{A2})$ ε_{ij}'s are mutually uncorrelated,

$(\boldsymbol{A3})$ ε_{ij}'s are uncorrelated with y_{i0} and η_i.

Assumptions $(\boldsymbol{A1})$-$(\boldsymbol{A3})$ generate $m(m-1)/2$ orthogonal moment conditions which are linear

functions of parameter ρ, that is,

$$E(y_{ih}\Delta u_{ij}) = 0, \quad \text{for } j = 2, ..., m \text{ and } h = 0, ..., j-2, \tag{3.1}$$

where $\Delta u_{ij} = u_{ij} - u_{ij-1}$. There are additional $(m-1)$ non-linear moment conditions associated with the orthogonality condition:

$$E(u_{im}\Delta u_{ij}) = 0, \quad \text{for } j = 2, ..., m. \tag{3.2}$$

In addition to (***A1***)-(***A3***), we can include the following model assumptions to improve the efficiency of the parameter estimators.

(*A4*) The ε_{ij}'s are homoskedastic, that is, var(ε_{ij}) is the same for all i and j.

This implies non-linear moment conditions:

$$E(u_{ij}^2) \text{ is the same for } j = 1, ..., m. \tag{3.3}$$

Based on assumptions (***A1***)-(***A4***), Ahn and Schmidt (1997) and Blundell and Bond (1998) further generate more moment conditions:

$$E(y_{ij}\Delta u_{ij-1} - y_{ij+1}\Delta u_{ij+2}) = 0, \quad \text{for } j = 1, ..., m-2, \tag{3.4}$$

$$E(\Delta y_{ij-1} u_{ij}) = 0, \quad \text{for } j = 2, ..., m, \tag{3.5}$$

$$\text{and } E(\bar{u}_i \Delta u_{ij+1}) = 0, \quad \text{for } j = 1, ..., m-1, \tag{3.6}$$

where $\bar{u}_i = \frac{1}{m}\sum_{j=1}^{m} u_{ij}$.

Furthermore, the first observed response variable is assumed to be mean zero, that is:

(*A5*) For all i, $E(y_{i0}) = 0$.

Therefore, the first-order m moment conditions satisfy

$$E(u_{ij}) = 0, \quad \text{for } j = 1, ..., m. \tag{3.7}$$

In summary, the total number of moment conditions from (3.1)-(3.7) is $m(m-1)/2 + 3(m-1) + 2m + (m-2)$ or $(m^2 + 11m - 10)/2$.

3.2.2 Dynamic Panel Data Models with Exogenous Variables

We now consider dynamic panel data models with exogenous variables X_{ij},

$$y_{ij} = \rho y_{ij-1} + X_{ij}^T \beta + u_{ij}, \quad \text{for } i = 1, ..., n \text{ and } j = 1, ..., m, \tag{3.8}$$

where $u_{ij} = \eta_i + \varepsilon_{ij}$ and X_{ij} is a p-dimensional vector of explanatory variables. For simplicity, we express (3.8) in matrix form,

$$Y = \rho Y_{-1} + X\beta + u, \tag{3.9}$$

where $u = \eta + \varepsilon$ and Y_{-1} is an n-dimensional vector of the previous observed responses. To distinguish different sources of associations between exogenous variables and error terms, we partition exogenous variables X into X_1 and X_2, where the dimensions of X_1 and X_2 are p_1 and p_2, respectively. The assumptions for exogenous variables are

(*A6*) For all i, j and s, X_{ij} is uncorrelated with ε_{is},

(*A7*) For all i and j, X_{1ij} is uncorrelated with η_i,

(*A8*) For all i and j, $E(X_{2ij}\eta_i)$ is the same.

Assumption (***A6***) implies that X is strongly exogenous with respect to ε, and assumptions (***A7***) and (***A8***) also imply that X_1 is uncorrelated with η, and X_2 and η have the same magnitude of association at different times. These assumptions generate the following moment conditions:

$$E(X_{1ij}u_{ih}) = 0, \quad \text{for } j = 1, ..., m \text{ and } h = 1, ..., m, \tag{3.10}$$

$$E(X_{2ij}u_{ih}) \text{ is the same}, \quad \text{for } j = 1, ..., m \text{ and } h = 1, ..., m. \tag{3.11}$$

Note that there are $p_1 m^2$ and $p_2 m(m-1)$ valid moment conditions generated by (3.10) and (3.11), respectively. Therefore, for the dynamic panel data model in (3.9) with exogenous variables, (3.10) and (3.11) provide additional moment conditions in conjunction with (3.1)-(3.7) under the simple dynamic panel model assumptions.

The existing literature on dynamic panel data models (Arellano and Bond, 1991; Arellano and Bover, 1995; Ahn, Lee, and Schmidt, 2001; Lai, Small, and Liu, 2008) mainly focuses on the cases with large sample size when the cluster size is small. Under this ideal setting, Blundell and Bond (1998) showed that the GMM estimator utilizing all valid moment conditions is asymptotically more efficient than the GMM estimator based on a subset of valid moment conditions. However, this is under the assumption that the sample size is much larger than the number of moment conditions. When the cluster size is large compared to the sample size, the dimension of moment conditions could increase exponentially and the GMM estimator could be unstable. We will provide another motivating problem on correlation structure modeling to illustrate that the dimension of moment conditions could exceed the sample size easily.

3.2.3 Quadratic Inference Function

For correlated data, utilizing an accurate correlation structure for correlated observations is essential for improving the efficiency of regression parameter estimators. Liang and Zeger (1986) proposed the generalized estimating equations approach, which requires only a few nuisance parameters to specify a common working correlation structure. However, a common working correlation structure does not represent the true correlation structure sufficiently well, especially when the cluster size is large. It is well-known that when the correlation structure is misspecified, the GEE estimator can be inefficient. Qu, Lindsay, and Li (2000) proposed the quadratic inference function to improve the efficiency of parameter estimation when the working correlation is misspecified.

Let the response variable for the ith subject $y_i = (y_{i1}, \ldots, y_{im})^T$ which is repeatedly measured for m times, where the y_i's are independent identically distributed, $i = 1, \ldots, n$, n is the sample size and m is the cluster size. The corresponding covariate for the ith subject is $X_i = (X_{i1}, \ldots, X_{im})^T$, which is $m \times p$-dimensional. For the generalized linear model, the marginal mean of y_{ij} can be specified as $\mu_{ij} = E(y_{ij}|X_{ij}) = \mu(X_{ij}^T\beta)$, where $\mu(\cdot)$ is an inverse link function and β is a p-dimensional parameter vector. The GEE is a marginal approach for estimating β by solving the equations

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

where $\mu_i = (\mu_{i1}, ..., \mu_{im})^T$, $\dot{\mu}_i = (\partial\mu_i/\partial\beta)$, $V_i = A_i^{1/2} R A_i^{1/2}$, A_i is the diagonal marginal variance

matrix of y_i and R is the common working correlation matrix for all subjects.

To improve the efficiency of GEE estimation, Qu, Lindsay, and Li (2000) approximated the inverse of the working correlation by a linear combination of basis matrices,

$$R^{-1} \approx \sum_{j=0}^{q} a_j M_j, \tag{3.12}$$

where M_0 is the identity matrix, M_j's are basis matrices with 0 and 1 components and a_j's are unknown coefficients. Therefore the GEE can be approximated as a linear combination of the elements in the following extended score vector

$$G_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} g_i(\beta), \tag{3.13}$$

where

$$g_i = \begin{pmatrix} \dot{\mu}_i^T A_i^{-1}(y_i - \mu_i) \\ \dot{\mu}_i^T A_i^{-1/2} M_1 A_i^{-1/2}(y_i - \mu_i) \\ \vdots \\ \dot{\mu}_i^T A_i^{-1/2} M_q A_i^{-1/2}(y_i - \mu_i) \end{pmatrix}.$$

However, it is impossible to set each equation in (3.13) to zero simultaneously in solving β, as the dimension of the estimating equations exceeds the dimension of parameters. Qu, Lindsay, and Li (2000) proposed the quadratic inference function (QIF) to obtain an estimator of β by minimizing the generalized method of moments (Hansen, 1982),

$$Q_n(\beta) = n G_n(\beta)^T W^{-1}(\beta) G_n(\beta), \tag{3.14}$$

where $W^{-1}(\beta) = \left[E\{g_i(\beta) g_i(\beta)^T\}\right]^{-1}$ is a weighting matrix and $W(\beta)$ can be estimated consistently by the sample covariance matrix $C(\beta) = \frac{1}{n} \sum_{i=1}^{n} g_i(\beta) g_i^T(\beta)$.

The pre-specified basis matrices are useful to approximate the working correlation matrix R if the inverse of the correlation structure has a linear representation in (3.12). However, this requires prior information on the basis matrices. Suppose the prior information for correlation

structure is unknown; we can use a linear representation of a complete set of basis matrices which contains 1 for the (i, j) and (j, i) entries and 0 elsewhere, which can handle any form of the correlation matrix. Alternatively, we can also represent R^{-1} through spectral decomposition of $C(\beta) = \sum_{j=1}^{m} \lambda_j e_j e_j^T$, where the candidate basis matrices contain $e_j e_j^T$, and e_j is the jth eigenvector of $C(\beta)$ for $j = 1, ..., m$. However, when the cluster size m is large, the number of moment conditions mp increases as m increases. This leads to the over-identified problem. Consequently, the weighting matrix $C(\beta)$ could be singular or close to singular and the QIF estimator could be extremely unstable. In the following section, we propose a new moment selection approach which incorporates all valid moment conditions when the number of moment conditions is greater than the sample size.

3.3 A New Moment Selection Method and Theory

Hansen (1982) showed that the generalized method of moments (GMM) is effective on optimally combining k-dimensional moment conditions through minimizing the objective function, which has similar form as (3.14). The GMM estimator is efficient in the sense that the asymptotic covariance matrix of $\hat{\beta}$ reaches the minimum among estimators solved by the same linear class of the estimating equations $G_n(\beta)$ given in (3.13).

When the number of moment conditions is much larger than the number of parameters ($k \gg p$), this could lead to the overly identified problem, and the GMM estimator could perform poorly if the sample size is small. In particular, when there are more moment conditions than the sample size, the GMM estimator will be unstable due to the singularity problem of the sample covariance matrix $C(\beta)$. Selecting a subset of valid moment conditions has been well-studied, while selection among many valid moment conditions is less developed. We propose a new objective function which selects an optimal number of linear combinations among all valid moment conditions.

Let $g_i = \left(g_{1i}^T, g_{2i}^T\right)^T$ be k-dimensional moment conditions for the ith subject, where $\dim(g_{1i}) = s$, $\dim(g_{2i}) = w = k - s$, and the moment conditions g_{1i} contain important information from the data based on some prior information. Here s is smaller than the sample size n. The key element in our development is to keep the first set of important moment conditions g_{1i} and extract most of the available information from the remaining moment conditions g_{2i} for parameter estimation.

Even if the prior information for important moment condition is not available, our method is still applicable for extracting information from the entire moment conditions with $s = 0$ and $w = k$.

To simplify our notations, let $G_n = (G_1^T, G_2^T)^T$, which contains two sets of moment conditions, and G_n is defined in (3.13). Let C_{11}, C_{12} and C_{22} be the block matrices of the covariance matrix of G_n corresponding to the variances and covariance of these two sets of moment conditions. We first orthogonalize G_2 against the main moment conditions G_1; this allows us to separate the contribution of two sets of moment conditions more clearly for estimation. The orthogonalized moment conditions corresponding to G_2 against G_1 is calculated by $G_2^* = G_2 - C_{21}C_{11}^{-1}G_1$. The moment conditions after orthogonalization can be represented as

$$\begin{pmatrix} G_1 \\ G_2^* \end{pmatrix} = \begin{pmatrix} I_1 & \mathbf{0} \\ -C_{21}C_{11}^{-1} & I_2 \end{pmatrix} \begin{pmatrix} G_1 \\ G_2 \end{pmatrix},$$

where I_1 and I_2 are identity matrices with $s \times s$ and $w \times w$ dimensions.

In order to extract important information from the orthogonalized moment conditions G_2^*, we calculate the covariance matrix of G_2^* by $C_2^* = C_{22} - C_{21}C_{11}^{-1}C_{12}$. Then we decompose the sample covariance matrix C_2^* through spectral decomposition based on $C_2^* = \sum_{j=1}^{w} \lambda_j e_j e_j^T$, where e_j is the jth eigenvector of C_2^* corresponding to the jth largest eigenvalue λ_j. If the first t principal components are selected, then we are able to reduce w moment conditions to t orthogonal linear combinations of moment conditions G_2^*. The final selected moment conditions $G^*(\beta)$ incorporating t principal components of C_2^* are:

$$G^* = \begin{pmatrix} I_1 & \mathbf{0} \\ \mathbf{0} & U \end{pmatrix} \begin{pmatrix} G_1 \\ G_2^* \end{pmatrix},$$

where U is the matrix containing t eigenvectors $(e_1, \ldots, e_t)^T$. Consequently, the GMM estimator of the proposed approach can be obtained via minimizing (3.14) based on the selected moment conditions G^*, where the dimension of G^* is reduced to $s + t$.

The crucial step here is to select t so as to ensure that most of the information from the remaining moment conditions can be captured. We propose a Bayesian information type of criterion to select

the number of moment conditions t among G_2^* through minimizing the objective function

$$\mathrm{J}(t) = \frac{\mathrm{tr}\{C_2^* - \tilde{C}(t)\}}{\mathrm{tr}(C_2^*)} + t\frac{log(nw)}{nw}, \tag{3.15}$$

where $\tilde{C}(t) = \sum_{j=1}^{t} \lambda_j e_j e_j^T$. Note that the first term of (3.15) measures the difference between the empirical covariance matrix calculated from the moment conditions G_2^* and the covariance matrix based on the selected combination of moment conditions. The second term of (3.15) is a penalty function of both n and w, which ensures an appropriate convergence rate for consistent moment selection. This is in contrast to the standard BIC, which is either a function of n or a function of w.

The advantage of the proposed procedure is that it does not require inversion of the sample covariance matrix C_2^*. This is quite useful when the dimension of moment conditions is high and the inversion of the high-dimensional covariance matrix is infeasible. Note that our approach is very different from Donald and Newey (2001) and Donald, Imbens, and Newey's (2009) minimizing the mean square error criterion, which requires the inverse of the sample covariance matrix.

We investigate the asymptotic properties for selection of t when the number of moment conditions and the sample size both increase; here we do not need to impose a restriction between n and w. In particular, we investigate whether the criterion (3.15) leads to a consistent selection of the optimal number principal components t_0. The following lemma provides the asymptotic rate of convergence for the estimated covariance matrix using t_0 selected principal components.

Lemma 3.1. *There exists t_0 such that $\|C_2^* - \tilde{C}(t_0)\| = O_p\{1/max(n,w)\}$, where $\|X\|$ is defined as $\sqrt{tr(X^TX)/IJ}$ and $I \times J$ is the dimension of matrix X.*

Lemma 3.1 indicates that the discrepancy (in matrix norms) between the estimated covariance matrix $\tilde{C}(t_0)$ and the covariance matrix C_2^* for G_2^* converges to 0 as $n, w \to \infty$. This implies that there is no efficiency loss if the optimal number of principle components t_0 is selected as the estimated covariance matrix $\left\{\dot{G}_2^* \tilde{C}(t_0)^{-1} \dot{G}_2^*\right\}^{-1}$ converges to the asymptotic covariance matrix $\left\{\dot{G}_2^* C_2^{*-1} \dot{G}_2^*\right\}^{-1}$, where $\dot{G}_2^* = (\partial G_2^*/\partial\beta)$. The following theorem shows that the optimal number of principal components t_0 can be consistently selected based on criterion (3.15) when both the number of moment conditions and the sample size go to infinity.

Theorem 3.1. *There exists a minimizer* $\hat{t}$ *of* $J(t)$ *in* (3.15) *such that*

$$\lim_{n,w\to\infty} Prob\big[\hat{t} = t_0\big] = 1.$$

Note that the choice of a penalty function plays an important role in identifying the number of moment conditions consistently. Here the penalty term in (3.15) vanishes at an appropriate rate such that the number of linear combinations of moment conditions are consistently selected with probability tending to 1. The above asymptotic property also ensures that the the covariance matrix C_2^* can be consistently estimated. Consequently, the new weighting matrix in GMM enables one to combine all valid moment conditions optimally without loss of efficiency. The proofs of Lemma 3.1 and Theorem 3.1 are provided in Section 3.6.

3.4 Numerical Studies

In this section, we examine the performance of the proposed method through simulation studies, and compare it with Doran and Schmidt's (2006) approach and the GEE estimators under three working correlation structures: independent structure (denoted by IND), exchangeable correlation structure (denoted by EX), and AR-1 correlation structure. In addition, we apply the proposed method using a dynamic panel model for Fortune 500 data.

3.4.1 Correlated Continuous Response

We generate the correlated continuous response variable from a marginal model

$$y_{ij} = X_{ij}^T\beta + \varepsilon_{ij}, \quad \text{for } i = 1, ..., n \text{ and } j = 1, ..., m,$$

where $X_{ij} = \big(x_{ij}^{(1)}, x_{ij}^{(2)}\big)^T$, $x_{ij}^{(1)} = \frac{j}{m} + N\Big(0, \frac{1}{m}\Big)$, $x_{ij}^{(2)} = \Big(\frac{m-j}{m}\Big)^2 + N\Big(0, \frac{1}{m}\Big)$, $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{im})^T \sim N(0, R)$, $\beta = (\beta_1, \beta_2)^T = (1, 1)^T$, n is the sample size and m is the cluster size. We generate repeated responses with varying cluster size of $m = 25$, 50 and 100; and the sample size ranges from $n = 50$, 100 and 500.

We design two simulation settings based on different correlation structures. The first setting has a three-block diagonal correlation matrix with the exchangeable correlation structure. The

dimensions of three block correlation matrices are $\frac{m}{5} \times \frac{m}{5}$, $\frac{3m}{5} \times \frac{3m}{5}$ and $\frac{m}{5} \times \frac{m}{5}$ respectively, and the correlation coefficient is 0.6. This setting allows one to compare the proposed estimators to the GEE estimators using the true correlation structure, which is denoted by "Oracle." The other setting has a slightly more complicated correlation structure, where the first block has a $\frac{3m}{5} \times \frac{3m}{5}$ exchangeable structure with correlation parameter 0.7, the second block has an $\frac{m}{5} \times \frac{m}{5}$ AR-1 structure with correlation 0.6, and the third block has an $\frac{m}{5} \times \frac{m}{5}$ exchangeable structure with correlation 0.8. In the second setting, the "Oracle" estimator is not provided, since the GEE estimator under a more complicated correlation structure is not available.

The basis matrices are obtained via an eigenvector decomposition, $R^{-1} \approx a_0 I + \sum_{j=1}^{m} a_j M_j$, where $M_j = e_j e_j^T$ and e_j is the eigenvector corresponding to the jth largest eigenvalue of the sample correlation matrix of y_i. There are a total of $m+1$ basis matrices. When $n = 50$, the number of moment conditions $k = (m+1)p$ exceeds the sample size for any given cluster size of 25, 50 and 100. That is, the QIF estimator constructed from moment conditions using all eigenvector bases is infeasible since the sample covariance matrix of the moment conditions is singular.

We compare the performance of the proposed method to Doran and Schmidt's (2006) approach and the GEE estimators with three types of working correlation structures, IND, EX and AR-1 based on 100 simulations. The proposed method is denoted by "Main" if the main moment conditions involving the identity basis matrix are preselected, or denoted by "No-main" if no prior moment conditions are selected. We choose the number of moment conditions t based on the BIC-type of criterion in (3.15). For Doran and Schmidt's approach, denoted by "DaS", t is chosen such that the sum of the t largest eigenvalues of the covariance matrix contains 95% of the sum of all eigenvalues.

Tables 3.1 and 3.2 provide the mean squared errors of the parameter estimator (mse($\hat{\beta}$)) and the average number of selected eigenvalues (ave($\hat{t}$)). To illustrate estimation efficiency, we define mse($\hat{\beta}$) $= \sum_{i=1}^{100} \|\hat{\beta}^{(i)} - \beta\|^2/(100 \times p)$, where $\hat{\beta}^{(i)}$ is the estimator from the ith simulation, β is the true parameter, and $\|\cdot\|$ denotes the Euclidean-norm. In addition, Figure 1 also provides the mean squared errors of the estimators based on the second correlation structure setting.

Our simulations show that the proposed method utilizing the main moment conditions is superior to other approaches, such as the GEE under misspecified correlation structures and Doran and

Schmidt's approach, in terms of the mean squared errors. Specifically, Table 3.1 indicates that the mean squared errors of the proposed method's estimators are closer to those of the oracle estimator as the sample size increases. For example, when $n = 500$, the mean squared errors of the proposed method and the oracle estimator are the same, while Doran and Schmidt's approach is not able to fully recover the efficiency of estimation. When the correlation structure is more complicated, Table 3.2 and Figure 3.1 show that the mean squared errors of the proposed method's estimators decrease as the cluster size increases, and the proposed method utilizing the main moment conditions performs considerably better than other approaches when the sample size is relatively small compared to the cluster size. For the GEE approach, the relatively low efficiency of the estimator can be explained in that the GEE approach is inefficient under misspecified working correlation structures.

For the proposed method, the number of selected principle components t tends to increase regardless of the correlation structure when the sample sizes increases. For example, when the correlation structure is more complicated, such as in Table 3.2, t tends to increase and the estimation efficiency improves as the cluster size increases. On the other hand, under a simpler correlation structure such as in Table 3.1, t tends to decrease when the cluster size increases. However, Doran and Schmidt's approach tends to select a smaller t even when a more complicated correlation structure is imposed. In general, their 95% fixed criterion selects only a few principle components, which are unable to retrieve sufficient information from the data.

3.4.2 Correlated Binary Response

We also conduct simulation studies with correlated binary responses. We generate the covariate $x_{ij}^{(1)} = \left(\frac{m-j}{2m}\right)^2 + N\left(0, \frac{1}{m}\right)$ and $x_{ij}^{(2)} = \left(\frac{j}{2m}\right)^3 + N\left(0, \frac{1}{m}\right)$, and the correlated binary response variable from a marginal logit model

$$\text{logit}(\mu_{ij}) = X_{ij}^T \beta, \qquad \text{for } i = 1, ..., n \text{ and } j = 1, ..., m,$$

where $X_{ij} = \left(x_{ij}^{(1)}, x_{ij}^{(2)}\right)^T$ and $\beta = (\beta_1, \beta_2)^T = (1, -1)^T$. The R package *mvtBinaryEP* is implemented to generate the correlated binary responses with three-block exchangeable correlation matrices. The dimensions for each block are $\frac{m}{5} \times \frac{m}{5}$, $\frac{3m}{5} \times \frac{3m}{5}$ and $\frac{m}{5} \times \frac{m}{5}$ respectively, and the

correlation coefficients are $\rho = (0.8, 0.4, 0.7)$.

We apply the proposed method, Doran and Schmidt's (2006) method and the GEE for the 100 simulated data sets. To investigate how various cluster sizes and sample sizes influence principle component selection and parameter estimation, we choose the same settings for cluster sizes, sample sizes and basis matrices as in Section 3.4.1. The results in Table 3.3 and Figure 1 confirm that the proposed method utilizing the main moment conditions outperforms the other methods in terms of the mean squared errors. When the sample size increases, Table 3.3 shows that the mean squared errors of the proposed method are closer to those of the oracle estimators. In addition, Figure 1 indicates that the proposed method provides more efficient estimation when the cluster size increases; however, this does not hold for the GEE and Doran and Schmidt's approaches even when the sample size increases to 500.

Based on the 95% selection criterion of Doran and Schmidt's approach, the average number of principle components t is close to 3 regardless of the sample size and cluster size. The proposed method, on the other hand, tends to select a larger number of principle components when the sample size increases. This is more sensible, as the selected number of principle components should vary depending on the given data.

3.4.3 Dynamic Panel Data Models

We generate the covariance stationary data (Bond and Windmeijer, 2002) based on the dynamic panel model, which contains one exogenous variable $x_{ij} = \frac{j}{m} + N\Big(0, \frac{1}{m}\Big)$,

$$y_{ij} = \rho y_{ij-1} + \beta x_{ij} + u_{ij}, \quad \text{for } i = 1, ..., n \text{ and } j = 1, ..., m,$$

where $u_{ij} = \eta_i + \varepsilon_{ij}$, $\beta = 1$, $\eta_i \sim N(0, 1)$, $\varepsilon_{ij} \sim N(0, 1)$, $y_{i0} = \frac{\eta_i}{1-\rho} + v_i$, and $v_i \sim N\Big(0, \frac{1}{1-\rho^2}\Big)$. We simulate 100 data sets with two correlation coefficients, $\rho = 0.4$ and $\rho = 0.7$. The valid moment conditions are generated based on (3.1), (3.4), (3.5), (3.7) and (3.10) following assumptions (***A1***)-(***A7***). For example, if the sample size n is 100 and the cluster size m is 10, the number of valid moment conditions is 172, which exceeds the sample size, and the GMM estimator using all moment conditions is infeasible.

We compare the proposed method with Doran and Schmidt's (2006) approach. Consider two

sample sizes ($n = 100$ and $n = 500$), and two cluster sizes ($m = 7$ and $m = 10$). Table 3.4 provides the mean, standard error, bias and average number of selected eigenvalues. It shows that the standard errors and the estimation bias of the proposed method decrease when the number of moment conditions and the sample size increase. On the other hand, the standard errors and the estimation bias of Doran and Schmidt's approach increase as the cluster size increases. In addition, the efficiency of Doran and Schmidt's estimation decreases as the correlation coefficient increases. This is possibly due to the fact that Doran and Schmidt's selection criterion tends to select a smaller number of moment conditions when the correlation ρ increases. In contrast, the number of principle components selected by the proposed method varies for different settings.

3.4.4 Fortune 500 Data Example

We compare the proposed method with Doran and Schmidt's (2006) approach and the generalized method of moments (GMM) with all available moment conditions for Fortune 500 data between 2000 and 2010. The 136 largest US corporations were ranked in the Global 500 in 2010. Out of the 136 companies, 105 companies were ranked over 10 consecutive years in Fortune 500 data. Therefore the sample size is 105. For this data, we apply the log-linear model based on the stationary conditional employee demand equation (Arellano and Bond, 1991; Blundell and Bond, 1998). The response variable of interest is the number of employees (Employees) from each firm and the exogenous variable is Revenue. The dynamic panel model is formulated as follows:

$$\log(\text{Employees})_{ij} = \rho \log(\text{Employees})_{ij-1} + \beta \log(\text{Revenue})_{ij} + u_{ij}, \tag{3.16}$$

where $u_{ij} = \eta_i + \varepsilon_{ij}$, $\log(\text{Employees})_{ij}$ is the log of the Employees and $\log(\text{Revenue})_{ij}$ is the log of the Revenue for the firm i in year j for $i = 1, ..., 105$ and $j = 1, ..., 10$.

We generate moment conditions based on assumptions ($\boldsymbol{A1}$)-($\boldsymbol{A4}$) and ($\boldsymbol{A6}$)-($\boldsymbol{A7}$) corresponding to η_i, ε_{ij}, $\log(\text{Employees})_{i0}$ and $\log(\text{Revenue})_{ij}$. Assumption ($\boldsymbol{A5}$) is not valid here because the log of Employees for the initial year cannot be zero. We exam the model in (3.16) for two time periods: one is for the short-term (2007-2010) and the other is for the long-term (2001-2010). Based on the formulations of (3.1), (3.4), (3.5) and (3.10), we can generate maximum 27 and 162 linear moment conditions for the short-term and long-term models. Note that the regular GMM

approach based on all available moment conditions is not feasible for the long-term model, since the number of moment conditions 162 exceeds the sample size of 105.

We implement the proposed method and compare it with Doran and Schmidt's (2006) approach and the GMM using all available moment conditions. Table 3.5 provides the parameter estimators, the standard errors (s.e.) of the estimators, and the number of selected principal components ($\hat{t}$) out of the total valid moment conditions (k) for the proposed method and Doran and Schmidt's (2006) approach. The standard errors of the proposed method are larger than the ones for the regular GMM estimators for the short-term model. This indicates that when the sample size is sufficiently large compared to the number of moment conditions, the GMM estimators using all available moment conditions are more efficient than the ones of the proposed method. However, the estimators obtained by the proposed method and the GMM approach are comparable for the short-term model, where the correlation coefficients on the lagged dependent variable are between 0.17 and 0.2, and the response variable of the Employees number and exogenous variable Revenue are positively associated.

For the Doran and Schmidt approach, their fixed selection criterion only allows selection of one moment condition for the short-term model, which is not sufficient to provide valid estimation. For the long-term model, it selects two moment conditions. Although they provide estimations for ρ and β, these estimators are not sensible since the correlation estimator $\hat{\rho} = -11.049$ is out of range, and $\hat{\beta} = 5.526$ is very different from the GMM estimator for the short-term model or the proposed estimators for either the short-term or long-term models. In addition, the standard errors of Doran and Schmidt estimators are extremely large, indicating that their estimations are unstable. In contrast, the estimations of the proposed method are sensible in selecting 4 and 10 linear combinations of moment conditions for the short-term and long-term models, and the estimators of ρ and β are in ranges consistent with smaller standard errors.

3.5 Discussion

The generalized method of moments provides consistent and efficient estimators when valid moment conditions are available. However, the GMM estimator is found to be extremely unstable when the dimension of moment conditions is larger than the sample size, due to the singularity problem of

the weighting matrix. Most existing methods are applicable for large sample sizes, but the cluster sizes are small.

The new approach combines all available moment conditions through principle components analysis for the weighting matrix, in contrast to existing approaches which select a subset of valid moment conditions. The BIC-type of criterion we propose is able to identify the optimal number of principal components consistently. Moreover, the proposed procedure enables one to include a set of important moment conditions, in addition to selecting the optimal linear combinations of the remaining moment conditions. Through the orthogonalization and spectral decomposition of the moment conditions, the new approach allows one to reduce the dimensionality of valid moment conditions, while retaining the important information of the moment conditions. Our numerical studies indicate that the proposed method outperforms existing methods in the sense of reducing bias and improving the efficiency of the estimation.

Note that the problem we study here is also related to low rank approximation for the large dimensional matrix, which has wide applications such as in data compression, large-dimensional matrix operations, recommendation systems and machine learning. The proposed method provides an objective criterion for selecting the optimal number of principle components to achieve efficiency in the moment selection problem.

3.6 Proofs of Theorem and Lemma

Proof of Lemma 3.1

Suppose that the largest eigenvalue λ_1 of the population covariance matrix of the orthogonalized moment conditions G_2^* is bounded, and $\lambda_j = O_p(1/nw)$ for any $j > t_0$. This condition ensures that the eigenvalues are sufficiently small if they are not selected as principal components. It also guarantees that the sum of the eigenvalues selected from a finite number of principal components is bounded.

By spectral decomposition, the sample covariance matrix of the orthogonalized moment conditions G_2^* is decomposed as $C_2^* = \sum_{j=1}^{M} \lambda_j e_j e_j^T$, where $M = min(n, w)$ and $e_j = (e_{1j}, \ldots, e_{wj})^T$ is the jth eigenvector corresponding to the jth largest eigenvalue of C_2^*. Since every component

of the eigenvector e_j for C_2^* is uniformly bounded, there exist constants K_2 and K_3 such that $0 < K_2 < |e_{ij}| < K_3 < \infty$ for $i = 1, \ldots, w$. It follows that

$$\|C_2^* - \tilde{C}(t_0)\| = \Big\| \sum_{j=t_0+1}^{M} \lambda_j e_j e_j^T \Big\| \leq K_3 \sqrt{\sum_{j=t_0+1}^{M} \lambda_j^2} \leq K_3 \sum_{j=t_0+1}^{M} \lambda_j = O_p\{1/max(n,w)\}.$$

Proof of Theorem 3.1

We need to show that $\lim_{n,w\to\infty} P\big\{\mathrm{J}(\hat{t}) < \mathrm{J}(t_0)\big\} = 0$ for all $\hat{t} \neq t_0$ and $\hat{t}$ is a finite integer. Since we have

$$\mathrm{J}(t_0) - \mathrm{J}(\hat{t}) = \frac{\mathrm{tr}\big\{\tilde{C}(\hat{t}) - \tilde{C}(t_0)\big\}}{\mathrm{tr}(C_2^*)} + (t_0 - \hat{t})\frac{log(nw)}{nw},$$

it is sufficient to prove that $P\big[\mathrm{tr}\{\tilde{C}(t_0) - \tilde{C}(\hat{t})\} - \mathrm{tr}(C_2^*)(t_0 - \hat{t})\frac{log(nw)}{nw} < 0\big] \to 0$ as $n, w \to \infty$.

First, we consider $\hat{t} < t_0$. Note that $\frac{1}{w}\mathrm{tr}(C_2^*) = O_p(1)$, because it is bounded by $K_2^2 \sum_{j=1}^{M} \lambda_j < \frac{1}{w}\mathrm{tr}(C_2^*) < K_3^2 \sum_{j=1}^{M} \lambda_j$. Since the eigenvector of C_2^* is bounded, it follows that

$$\begin{aligned}
&\frac{1}{w}\mathrm{tr}\big\{\tilde{C}(t_0) - \tilde{C}(\hat{t})\big\} - \frac{1}{w}\mathrm{tr}(C_2^*)(t_0 - \hat{t})\frac{log(nw)}{nw} \\
&= \frac{1}{w}\mathrm{tr}\bigg(\sum_{j=\hat{t}+1}^{t_0} \lambda_j e_j e_j^T \bigg) - \frac{1}{w}\mathrm{tr}(C_2^*)(t_0 - \hat{t})\frac{log(nw)}{nw} \\
&\geq K_2^2 \sum_{j=\hat{t}+1}^{t_0} \lambda_j - \frac{1}{w}\mathrm{tr}(C_2^*)(t_0 - \hat{t})\frac{log(nw)}{nw} \\
&\to K_2^2 \sum_{j=\hat{t}+1}^{t_0} \lambda_j > 0 \text{ as } n, w \to \infty.
\end{aligned}$$

Therefore, $\lim_{n,w\to\infty} P\big\{\mathrm{J}(\hat{t}) - \mathrm{J}(t_0) < 0\big\} = 0$ holds.

Second, we consider $\hat{t} > t_0$. We have

$$P\big\{n\mathrm{J}(\hat{t}) - n\mathrm{J}(t_0) < 0\big\} = P\Big[n\mathrm{tr}\big\{\tilde{C}(\hat{t}) - \tilde{C}(t_0)\big\} > \frac{1}{w}\mathrm{tr}(C_2^*)(\hat{t} - t_0)log(nw)\Big],$$

and we obtain

$$n\mathrm{tr}\big\{\tilde{C}(\hat{t}) - \tilde{C}(t_0)\big\} = n\mathrm{tr}\Big(\sum_{j=t_0+1}^{\hat{t}} \lambda_j e_j e_j^T\Big) > K_2^2 nw \sum_{j=t_0+1}^{\hat{t}} \lambda_j = K_2^2 nw O_p(1/nw) = O_p(1).$$

On the other hand, $\frac{1}{w}\mathrm{tr}(C_2^*)(\hat{t} - t_0)log(nw) = O\big\{log(nw)\big\}$. Consequently, this ensures that

$$P\big\{\mathrm{J}(\hat{t}) - \mathrm{J}(t_0) < 0\big\} = P\Big[\mathrm{tr}\big\{\tilde{C}(\hat{t}) - \tilde{C}(t_0)\big\} > \mathrm{tr}(C_2^*)(\hat{t} - t_0)\frac{log(nw)}{nw}\Big] \to 0 \text{ as } n, w \to \infty.$$

Table 3.1: Comparison of the proposed method, Doran and Schmidt's approach and the GEE for continuous cases. The true correlation structure is a 3-block diagonal correlation matrix with dimensions of $\frac{m}{5} \times \frac{m}{5}$, $\frac{3m}{5} \times \frac{3m}{5}$ and $\frac{m}{5} \times \frac{m}{5}$: each block is an exchangeable structure with $\rho = 0.6$, and k is the number of moment conditions.

		$n = 50$		$n = 100$		$n = 500$	
m (k)	Method	mse($\hat{\beta}$)	ave($\hat{t}$)	mse($\hat{\beta}$)	ave($\hat{t}$)	mse($\hat{\beta}$)	ave($\hat{t}$)
25 (52)	Main	0.0089	4.78	0.0042	5.98	0.0007	18.20
	No-main	0.0122	3.18	0.0046	3.71	0.0008	5.37
	DaS	0.0127	2.98	0.0071	3.00	0.0013	3.00
	IND	0.0123	-	0.0074	-	0.0013	-
	EX	0.0112	-	0.0059	-	0.0011	-
	AR-1	0.0133	-	0.0068	-	0.0012	-
	Oracle	0.0077	-	0.0036	-	0.0007	-
50 (102)	Main	0.0079	3.93	0.0045	4.96	0.0006	16.22
	No-main	0.0114	3.07	0.0055	3.49	0.0007	4.50
	DaS	0.0121	2.99	0.0073	3.00	0.0013	3.00
	IND	0.0124	-	0.0073	-	0.0012	-
	EX	0.0128	-	0.0074	-	0.0013	-
	AR-1	0.0188	-	0.0111	-	0.0018	-
	Oracle	0.0066	-	0.0035	-	0.0006	-
100 (202)	Main	0.0073	3.55	0.0043	4.07	0.0006	14.90
	No-main	0.0126	3.02	0.0061	3.38	0.0006	4.11
	DaS	0.0126	2.99	0.0078	3.00	0.0014	3.00
	IND	0.0121	-	0.0076	-	0.0013	-
	EX	0.0671	-	0.0191	-	0.0042	-
	AR-1	0.0308	-	0.0145	-	0.0028	-
	Oracle	0.0067	-	0.0034	-	0.0006	-

Table 3.2: Comparison of the proposed method, Doran and Schmidt's approach and the GEE for continuous cases. The true correlation structure is a 3-block diagonal correlation matrix with dimensions of $\frac{3m}{5} \times \frac{3m}{5}$, $\frac{m}{5} \times \frac{m}{5}$ and $\frac{m}{5} \times \frac{m}{5}$: two exchangeable structures and one AR-1 structure with $\rho = (0.7, 0.6, 0.8)$, and k is the number of moment conditions.

		$n = 50$		$n = 100$		$n = 500$	
m (k)	Method	mse($\hat{\beta}$)	ave($\hat{t}$)	mse($\hat{\beta}$)	ave($\hat{t}$)	mse($\hat{\beta}$)	ave($\hat{t}$)
	Main	0.0083	8.29	0.0036	10.99	0.0007	17.65
	No-main	0.0157	3.25	0.0048	3.80	0.0007	5.15
	DaS	0.0282	2.50	0.0129	2.56	0.0024	2.64
25 (52)	IND	0.0198	-	0.0098	-	0.0017	-
	EX	0.0177	-	0.0086	-	0.0015	-
	AR-1	0.0173	-	0.0085	-	0.0014	-
	Main	0.0064	11.17	0.0031	14.40	0.0005	29.45
	No-main	0.0133	3.65	0.0048	4.28	0.0005	5.65
	DaS	0.0279	2.14	0.0146	2.18	0.0027	2.01
50 (102)	IND	0.0205	-	0.0105	-	0.0019	-
	EX	0.0194	-	0.0094	-	0.0017	-
	AR-1	0.0239	-	0.0101	-	0.0017	-
	Main	0.0039	15.83	0.0025	22.80	0.0003	37.75
	No-main	0.0111	4.34	0.0046	5.01	0.0003	6.38
	DaS	0.0261	2.00	0.0139	2.00	0.0029	2.00
100 (202)	IND	0.0173	-	0.0095	-	0.0018	-
	EX	0.0265	-	0.0141	-	0.0021	-
	AR-1	0.0273	-	0.0116	-	0.0033	-

Table 3.3: Comparison of the proposed method, Doran and Schmidt's approach and the GEE for binary cases. The true correlation structure is a 3-block diagonal correlation matrix with dimensions of $\frac{m}{5} \times \frac{m}{5}$, $\frac{3m}{5} \times \frac{3m}{5}$ and $\frac{m}{5} \times \frac{m}{5}$: each block has an exchangeable structure with $\rho = (0.8, 0.4, 0.7)$, and k is the number of moment conditions.

		$n = 50$		$n = 100$		$n = 500$	
m (k)	Method	mse($\hat{\beta}$)	ave($\hat{t}$)	mse($\hat{\beta}$)	ave($\hat{t}$)	mse($\hat{\beta}$)	ave($\hat{t}$)
25 (52)	Main	0.1979	8.20	0.1147	10.20	0.0127	21.87
	No-main	0.2033	4.20	0.1225	4.86	0.0128	10.93
	DaS	0.2674	3.00	0.1479	2.99	0.0224	3.00
	IND	0.2724	-	0.1398	-	0.0201	-
	EX	0.4604	-	0.1565	-	0.0266	-
	AR-1	0.2702	-	0.1390	-	0.0237	-
	Oracle	0.1353	-	0.1059	-	0.0114	-
50 (102)	Main	0.1684	8.47	0.0662	10.90	0.0108	25.06
	No-main	0.1999	4.11	0.0899	4.38	0.0127	9.90
	DaS	0.3249	2.97	0.1635	3.00	0.0235	3.00
	IND	0.3223	-	0.1468	-	0.0209	-
	EX	0.4757	-	0.2898	-	0.0419	-
	AR-1	0.3526	-	0.1622	-	0.0281	-
	Oracle	0.1380	-	0.0656	-	0.0105	-
100 (202)	Main	0.1396	9.60	0.0460	12.07	0.0088	28.43
	No-main	0.1627	4.07	0.0819	4.07	0.0119	7.10
	DaS	0.2615	2.93	0.1559	3.00	0.0269	3.00
	IND	0.2419	-	0.1446	-	0.0236	-
	EX	0.7511	-	0.3589	-	0.0574	-
	AR-1	0.5940	-	0.2375	-	0.0435	-
	Oracle	0.1153	-	0.0430	-	0.0087	-

Table 3.4: Comparison of the proposed method and Doran and Schmidt's approach for dynamic panel data, and k is the number of moment conditions.

			$n = 100$				$n = 500$			
			No-main		DaS		No-main		DaS	
ρ	m (k)		$\hat{\rho}$	$\hat{\beta}$	$\hat{\rho}$	$\hat{\beta}$	$\hat{\rho}$	$\hat{\beta}$	$\hat{\rho}$	$\hat{\beta}$
0.4	7 (88)	Mean	0.386	1.043	0.361	1.056	0.399	1.002	0.387	1.017
		Standard error	0.075	0.166	0.108	0.171	0.026	0.060	0.043	0.068
		Bias	0.060	0.128	0.089	0.138	0.021	0.048	0.038	0.057
		ave($\hat{t}$)	31.12		17.26		44.30		20.88	
	10 (172)	Mean	0.389	1.039	0.313	1.133	0.401	0.998	0.384	1.024
		Standard error	0.064	0.142	0.115	0.199	0.019	0.046	0.049	0.073
		Bias	0.052	0.122	0.104	0.186	0.015	0.036	0.041	0.059
		ave($\hat{t}$)	49.44		25.74		72.70		33.60	
0.7	7(88)	Mean	0.669	1.060	0.603	1.157	0.704	0.991	0.671	1.044
		Standard error	0.101	0.245	0.200	0.437	0.026	0.075	0.066	0.129
		Bias	0.088	0.186	0.163	0.329	0.021	0.062	0.058	0.110
		ave($\hat{t}$)	23.98		10.34		37.90		10.90	
	10 (172)	Mean	0.674	1.042	0.532	1.360	0.700	1.000	0.649	1.089
		Standard error	0.085	0.198	0.165	0.448	0.023	0.053	0.084	0.193
		Bias	0.069	0.150	0.184	0.442	0.019	0.040	0.074	0.161
		ave($\hat{t}$)	37.18		14.78		64.76		16.90	

Table 3.5: For the Fortune 500 data set, comparison of the estimator and its standard error (s.e.) by the proposed method, Doran and Schmidt's approach, and the generalized method of moments (GMM) with all moment conditions. k is the number of moment conditions, $\hat{t}$ is the number of selected moment conditions, and '–' is not estimable.

Methods	Parameters	2007-2010	2001-2010
GMM	$\hat{\rho}$ (s.e.)	0.172 (0.005)	–
	$\hat{\beta}$ (s.e.)	0.370 (0.002)	–
No-main	$\hat{\rho}$ (s.e.)	0.202 (0.014)	0.397 (0.007)
	$\hat{\beta}$ (s.e.)	0.365 (0.006)	0.278 (0.003)
	$\hat{t}$ (k)	4 (27)	10 (162)
DaS	$\hat{\rho}$ (s.e.)	–	-11.049 (8.952)
	$\hat{\beta}$ (s.e.)	–	5.526 (4.105)
	$\hat{t}$ (k)	1 (27)	2 (162)

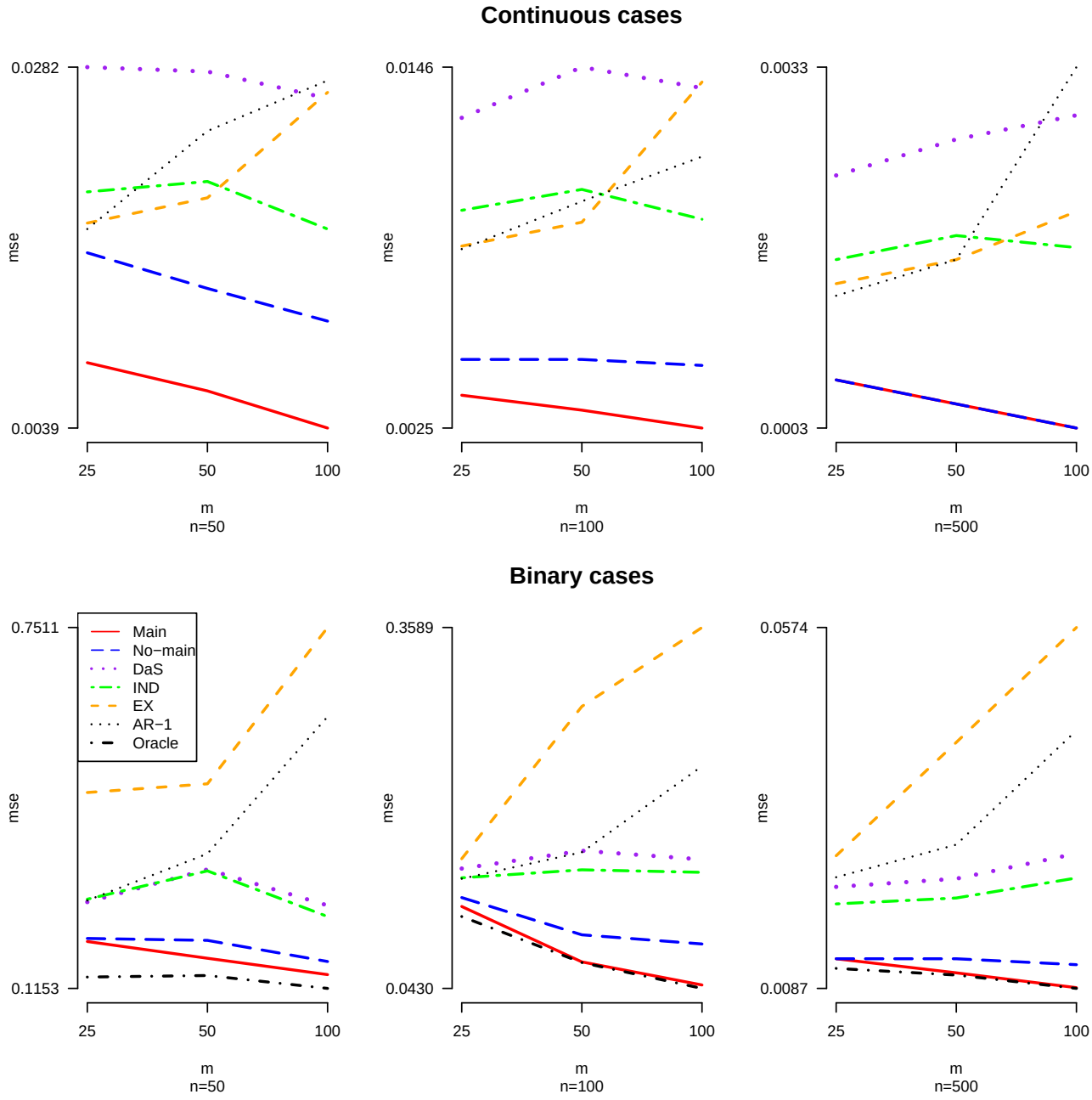

Figure 3.1: MSE comparison of the proposed method, Doran and Schmidt's approach and the GEE. The first row is the mean squared errors of estimators for continuous cases based on the results in Table 2, and the second row is the mean squared errors of estimators for binary cases based on the results in Table 3.

References

Ahn, S. C., Lee, Y. H., and Schmidt, P. (2001). GMM estimation of linear panel data models with time-varying individual effects. *Journal of Econometrics* **101**, 219-255.

Ahn, S. C. and Schmidt, P. (1997). Efficient estimation of dynamic panel data models: Alternative assumptions and simplified estimation. *Journal of Econometrics* **76**, 309-322.

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

Anderson, T. W. and Hsiao, C. (1981). Estimation of dynamic models with error components. *Journal of the American Statistical Association* **76**, 598-606.

Andrews, D. W. K. and Lu, B. (2001). Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data models. *Journal of Econometrics* **101**, 123-164.

Antoniadis, A. (1997). Wavelets in statistics: A review (with discussion). *Journal of the Italian Statistical Association* **6**, 97-144.

Arellano, M. and Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies* **58**, 277-297.

Arellano, M. and Bover, O. (1995). Another look at the instrumental variables estimation of error-component models. *Journal of Econometrics* **68**, 29-52.

Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. *The Annals of Statistics* **36**, 2577-2604.

Bond, S. (2002). Dynamic panel data models: A guide to micro data methods and practice. *Portuguese Economic Journal* **1**, 141-162.

Bond, S. and Windmeijer, F. (2002). Finite sample inference for GMM estimators in linear panel data models. *10th International Conference on Panel Data*, C6-3.

Blundell, R. and Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics* **87**, 115-144.

Cai, T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association* **106**, 672-684.

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

Danaher, P., Wang, P., and Witten, D. (2012). The joint graphical lasso for inverse covariance estimation across multiple classes. Arxiv preprint arXiv:1111.0324.

Donald, S. G., Imbens, G. W., and Newey, W. K. (2009). Choosing instrumental variables in conditional moment restriction models. *Journal of Econometrics* **152**, 28-36.

Donald, S. G., and Newey, W. K. (2001). Choosing the number of instruments. *Econometrica* **69**, 1161-1191.

Doran, H. E. and Schmidt, P. (2006). GMM estimators with improved finite sample properties using principal components of the weighting matrix, with an application to the dynamic panel data model. *Journal of Econometrics* **133**, 387-409.

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* (Fan, J., Liu, J. S. and Lin, X., eds), 49-72. World Scientific Publishing.

Fan, J., Fan, Y., and Lv, J. (2008). High-dimensional covariance matrix estimation using a factor model. *Journal of Econometrics* **147**, 186-197.

Fan, J., Huang, T., and Li, R. (2007). Analysis of longitudinal data with semiparametric estimation of covariance function. *Journal of the American Statistical Association* **102**, 632-641.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **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* (M. Sanz-Sole, J. Soria, J. L. Varona and J. Verdera, eds.) **3**, 595-622. European Mathematical Society.

Fan, J., Liao, Y., and Mincheva, M. (2011). High-dimensional covariance matrix estimation in approximate factor models. *The Annals of Statistics* **39**, 3320-3358.

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

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

Fan, J. and Wu, Y. (2008). Semiparametric estimation of covariance matrices for longitudinal data. *Journal of the American Statistical Association* **103**, 1520-1533.

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

Friedman, J., Hastie, T., and Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432-441.

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

Gallant, A. R. and Tauchen, G. (1996). Which moments to match? *Econometric Theory* **12**, 657-681.

Gao, X., Pu, Q., Wu, Y., and Xu, H. (2012). Tuning parameter selection for penalized likelihood estimation of gaussian graphical model. *Statistica 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. *The Annals of Statistics* **34**, 1493-1517.

Han, C., Orea, L., and Schmidt, P. (2005). Estimation of a panel data model with parametric temporal variation in individual effects. *Journal of Econometrics* **126**, 241-267.

Han, C. and Phillips, P. C. B. (2006). GMM with many moment conditions. *Econometrica* **74**, 147-192.

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

Huang, J., Liu, N., Pourahmadi, M., and Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika* **93**, 85-98.

Hunter, D. R. and Li, R. (2005). Variable selection using MM algorithms. *The Annals of Statistics* **33**, 1617-1642.

James, G. and Hastie, T. (2001). Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society B* **63**, 533-550.

James, G. and Sugar, C. A. (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association* **98**, 397-408.

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

Johnson, B., Lin, D. Y., and Zeng, D. (2008). Penalized estimating functions and variable selection in semiparametric regression models. *Journal of the American Statistical Association* **103**, 672-680.

Lai, T., Small, D. Y., and Liu, J. (2008). Statistical inference in dynamic panel data models. *Journal of Statistical Planning and Inference* **138**, 2763-2776.

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.

Luo, X. (2011). High-dimensional low rank and sparse covariance matrix estimation via convex minimization. Arxiv preprint arXiv:1111.1133.

Magdon-Ismail, M. and Purnell, J. T. (2011). Approximating the covariance matrix of GMMs with low-rank perturbations. *Int. J. Data Mining and Bioinformatics* **3**, 1-15.

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

Meinshausen, M. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics* **34**, 1436-1462.

Newey, W. K. and Smith, R. J. (2004). Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica* **72**, 219-255.

Okui, R. (2009). The optimal choice of moments in dynamic panel data models. *Journal of Econometrics* **151**, 1-16.

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

Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009). Partial correlation estimation by joint sparse regression model. *Journal of the American Statistical Association* **104**, 735-746.

Pourahmadi, M. (2007). Cholesky decompositions and estimation of a covariance matrix: Orthogonality of variance-correlation parameters. *Biometrika* **94**, 1006-1013.

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

Rothman, A. J., Levina, E., and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association* **104**, 177-186.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **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. *Journal of the Royal Statistical Society B* **58**, 267-288.

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

Wang, H., Li, B., and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society B* **71**, 671-683.

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

Wang, L. and Qu, A. (2009). Consistent model selection and data-driven smooth tests for longitudinal data in the estimating equations approach. *Journal of the Royal Statistical Society 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.

Witten, D., Friedman, J., and Simon, N. (2011). New insights and faster computations for the graphical lasso. *Journal of Computational and Graphical Statistics* **20**, 892-900.

Wu, W. B. and Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika* **90**, 831-844.

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. *Journal of the American Statistical Association* **105**, 1518-1530.

Yao, F., Müller, H. G., and Wang, J. L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* **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. *Journal of the American Statistical Association* **105**, 312-323.

Zhou, J. and Qu, A. (2012). Informative estimation and selection of correlation structure for longitudinal data. *Journal of the American Statistical Association* **107**, 701-710.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418-1429.

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