Electronic Journal of Statistics
Vol. 9 (2015) 1315–1334
ISSN: 1935-7524
DOI: 10.1214/15-EJS1036

Efficient estimation for longitudinal data by combining large-dimensional moment conditions

Hyunkeun Cho

Department of Statistics
Western Michigan University, Kalamazoo, MI 49008
e-mail: hyunkeun.cho@wmich.edu

and

Annie Qu

Department of Statistics
University of Illinois at Urbana-Champaign, Champaign, IL 61820
e-mail: anniequ@illinois.edu

Abstract: The quadratic inference function approach is able to provide a consistent and efficient estimator if valid moment conditions are available. However, the QIF estimator is unstable when the dimension of moment conditions is large compared to the sample size, due to the singularity problem for the estimated weighting matrix. We propose a new estimation procedure which combines all valid moment conditions optimally via the spectral decomposition of the weighting matrix. In theory, we show that the proposed method yields a consistent and efficient estimator which follows an asymptotic normal distribution. In addition, Monte Carlo studies indicate that the proposed method performs well in the sense of reducing bias and improving estimation efficiency. A real data example of Fortune 500 companies is used to compare the performance of the new method with existing methods.

MSC 2010 subject classifications: 62H25.
Keywords and phrases: Generalized method of moments, moment selection, principal components, quadratic inference function, singularity matrix.

Received September 2014.

Contents

1. Introduction

Longitudinal data arise frequently in many studies where repeated measurements from a subject are correlated. The correlated nature of longitudinal data makes it difficult to specify the full likelihood function for non-normal responses. [10] proposed the generalized estimating equation (GEE) for correlated data, which only requires the first two moments and a working correlation matrix of errors to account for correlations. Although the GEE provides a consistent estimator regardless of whether the working correlation is correctly specified or not, the estimator can be inefficient under misspecified correlation structures. [13] developed the quadratic inference function (QIF) based on the generalized method of moments [8] to achieve better estimation efficiency.

For correlated data with large-dimensional cluster size, it is important to account for the true correlation information since it can reduce the bias and increase the efficiency of estimation. For example, the QIF utilizing a full set of basis matrices allows one to select a flexible correlation structure, and can increase the efficiency of the estimator significantly [15]. However, this generates many moment conditions for large sized clusters since the dimension of moment conditions depends on the number of basis matrices, which relies on the cluster size. This could be problematic in estimating the inverse of the sample covariance matrix of moment conditions, which is an optimal weighting matrix for the QIF estimator and plays a crucial role in achieving an efficient QIF estimator. First, the sample covariance matrix might not be full rank when there are more moment conditions than the sample size. Second, even if the sample covariance matrix is invertible, the estimation of its inverse could be biased with high variation. Therefore the QIF could perform poorly due to infeasible or imprecise estimation of the optimal weighing matrix.

In the generalized method of moments literature, it has been shown that over-identified moment conditions may cause poor performance in finite sample estimation [9, 11]. In this paper, we are motivated by a problem in longitudinal data where a dimension of moment conditions is relatively large compared to the sample size, or the moment conditions are highly correlated. The singularity or near singularity of the weighting matrix makes the QIF estimator infeasible or unstable. In order to solve this problem, the subset moment selection method has been developed for large-dimensional moment conditions. [7, 1, 12] propose to eliminate the least informative moment conditions to reduce the overall number of moment conditions. However, this requires prior information on the

moment conditions. [6, 2, 3] utilized penalized objective functions to select informative moment conditions. However, the underlying assumption is that the most of the moment conditions are not informative. [5, 4] propose selecting moment conditions based on the criterion of minimizing the mean square error of the estimator. However, their criterion requires inverting the sample covariance matrix, which could be infeasible when the dimension of moment conditions exceeds the sample size. Moreover, most moment selection approaches result in efficiency loss for parameter estimation, since the information from unselected moment conditions is not utilized.

We propose a new estimation procedure which combines all valid moment conditions using principle components analysis. We apply a spectral decomposition of the covariance matrix for the moment conditions and select an optimal number of linear combinations of the moment conditions through a new objective function based on a Bayesian information type of criterion [14]. This allows one to reduce the dimensionality of valid moment conditions, while retaining most of the information from all moment conditions. The proposed method performs well in the sense of reducing bias and improving the efficiency of QIF estimation, and is especially effective when the dimension of moment conditions is high compared to the sample size. Furthermore, it is capable of incorporating a set of preselected moment conditions, in conjunction with selecting the optimal linear combinations of remaining moment conditions. This has the advantage of preventing any information loss from moment conditions which surely should be included for estimation.

In theory, we show that the proposed criterion is able to select the number of principal components consistently, when the sample size goes to infinity. The QIF estimator using the selected linear combinations of moment conditions is consistent and asymptotically normal. In addition, the proposed approach yields an efficient estimator in the sense that its asymptotic variance matrix reaches the minimum. Our numerical studies also confirm that a subset moment selection approach, or replacing an identity matrix as the weighting matrix approach result in less accurate and efficient estimation compared to the proposed estimator.

The paper is organized as follows. Section 2 provides the background of the quadratic inference function and motivation of the problem. Section 3 introduces an efficient estimation approach which combines all valid moment conditions optimally and provides asymptotic properties of the proposed estimator. Section 4 illustrates our method with simulation studies and an application to real data. The final section gives concluding remarks and discussion. All proofs of the lemmas and theories are provided in the Appendix.

2. Quadratic inference function

2.1. Notation and preliminaries

Let the response variable for the ith subject be $\mathbf{y}_i = (y_{i1}, \ldots, y_{im_i})'$, where $\mathbf{y}_i$'s are independent identically distributed for $i = 1, \ldots, n$, n is the sample size and

m_i is the cluster size. To simplify the notation, we first set $m_i = m$ for all i, and the unbalanced data case will be discussed in more detail in Section 3.3. The corresponding covariate for the ith subject is $\mathbf{x}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{im})'$, which is $m \times p$-dimensional. For the generalized linear model, the marginal mean of y_{ij} is represented as $\mu_{ij} = E(y_{ij}|\mathbf{x}_{ij}) = \mu(\mathbf{x}_{ij}'\boldsymbol{\beta})$, where $\mu(\cdot)$ is an inverse link function and $\boldsymbol{\beta}$ is a p-dimensional parameter vector. [10] proposed the generalized estimating equation (GEE) as a marginal model approach for estimating $\boldsymbol{\beta}$ by solving

$$\sum_{i=1}^{n} \dot{\boldsymbol{\mu}}_i' \mathbf{A}_i^{-1/2} \mathbf{R}^{-1} \mathbf{A}_i^{-1/2} (\mathbf{y}_i - \boldsymbol{\mu}_i) = 0,$$

where $\dot{\boldsymbol{\mu}}_i = (\partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta})$, $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{im})'$, $\mathbf{A}_i$ is the diagonal marginal variance matrix of $\mathbf{y}_i$ and $\mathbf{R}$ is a common working correlation matrix for all subjects.

[13] approximate the inverse of the working correlation using a linear combination of basis matrices,

$$\mathbf{R}^{-1} \approx a_0 \mathbf{I} + \sum_{j=1}^{q} a_j \mathbf{B}_j, \tag{2.1}$$

where $\mathbf{I}$ is an identity matrix, $\mathbf{B}_1, \ldots, \mathbf{B}_q$ are basis matrices with 0 and 1 components and a_j's are unknown coefficients. Consequently, the GEE can be approximated as a linear combination of the elements in the following moment conditions

$$\mathbf{G}_n(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{g}_i(\boldsymbol{\beta}), \tag{2.2}$$

where

$$\mathbf{g}_i = \begin{pmatrix} \dot{\boldsymbol{\mu}}_i' \mathbf{A}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) \\ \dot{\boldsymbol{\mu}}_i' \mathbf{A}_i^{-1/2} \mathbf{B}_1 \mathbf{A}_i^{-1/2} (\mathbf{y}_i - \boldsymbol{\mu}_i) \\ \vdots \\ \dot{\boldsymbol{\mu}}_i' \mathbf{A}_i^{-1/2} \mathbf{B}_q \mathbf{A}_i^{-1/2} (\mathbf{y}_i - \boldsymbol{\mu}_i) \end{pmatrix}. \tag{2.3}$$

Note that $\mathbf{g}_i$ in (2.2) does not involve the nuisance parameters $a_0, \ldots, a_q$ associated with the linear weights in (2.1). However, it is impossible to set each estimating equation in (2.2) to zero simultaneously in solving $\boldsymbol{\beta}$, as the dimension of the moment conditions exceeds the dimension of parameters.

[13] proposed obtaining an estimator of $\boldsymbol{\beta}$ by minimizing the quadratic inference function,

$$Q_n(\boldsymbol{\beta}) = n \mathbf{G}_n(\boldsymbol{\beta})' \mathbf{V}(\boldsymbol{\beta})^{-1} \mathbf{G}_n(\boldsymbol{\beta}), \tag{2.4}$$

where $\mathbf{V}(\boldsymbol{\beta})^{-1} = \big[E\{\mathbf{g}_i(\boldsymbol{\beta}) \mathbf{g}_i(\boldsymbol{\beta})'\}\big]^{-1}$ is a weighting matrix and $\mathbf{V}(\boldsymbol{\beta})$ is estimated consistently by a sample covariance matrix $\mathbf{C}_n(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{g}_i(\boldsymbol{\beta}) \mathbf{g}_i(\boldsymbol{\beta})'$. Similar to the generalized method of moments, the QIF estimator utilizing $\mathbf{C}_n(\boldsymbol{\beta})$ is optimal in the sense that the asymptotic variance matrix of the QIF

estimator reaches the minimum among all estimators solved by the same linear class of the moment conditions given in (2.3).

2.2. QIF with large-dimensional moment conditions

For high-dimensional clustered data, utilizing accurate correlation structures for correlated measurements is essential for improving the efficiency of regression parameter estimators and reducing the bias of the estimator. Although the GEE approach requires only a few nuisance parameters to specify a common working correlation structure, this 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. The QIF approach is able to improve the efficiency of parameter estimation by representing the correlation structure as pre-specified basis matrices.

The pre-specified basis matrices are useful to approximate the working correlation matrix $\mathbf{R}$ if the inverse of the correlation structure has a linear representation in (2.1). For example, if $\mathbf{R}$ corresponds to an exchangeable structure, then $\mathbf{R}^{-1} = a_0\mathbf{I} + a_1\mathbf{B}_1$, where a_0 and a_1 are coefficients associated with the exchangeable correlation parameter, and $\mathbf{B}_1$ is a symmetric matrix with 0 on the diagonal and 1 elsewhere. If $\mathbf{R}$ is the first-order autoregressive (AR1), then $\mathbf{R}^{-1} = a_0\mathbf{I} + a_2\mathbf{B}_2 + a_3\mathbf{B}_3$, where a_0, a_2 and a_3 are coefficients associated with the AR1 correlation parameters, $\mathbf{B}_2$ is a symmetric matrix with 1 on the sub-diagonal entries and 0 elsewhere, and $\mathbf{B}_3$ is a symmetric matrix with 1 in elements $(1,1)$ and (m,m). However, this kind of representation requires prior information for working correlation matrices.

Suppose the prior information for correlation structure is unknown. We can use a linear representation of a complete set of basis matrices with 1 for the (i,j) and (j,i) entries and 0 elsewhere, which can handle any form of the correlation matrix. Alternatively, the basis matrices $\mathbf{B}_j$'s can also be obtained through an eigenvector decomposition, $\mathbf{R}^{-1} \approx a_0\mathbf{I} + \sum_{j=1}^{m} a_j\mathbf{B}_j$, where $\mathbf{B}_j = \mathbf{e}_j\mathbf{e}_j'$ is the jth basis matrix and $\mathbf{e}_j$ is the eigenvector corresponding to the jth largest eigenvalue of the sample correlation matrix for $\mathbf{y}_i$. However, this will lead to the generation of many moment conditions when the cluster size is large and prior information is not provided.

If the number of moment conditions is much larger than the number of parameters, some moment conditions could be either less informative or highly correlated. This could lead to a large variability in estimating the weighting matrix $\mathbf{C}_n(\boldsymbol{\beta})^{-1}$ and result in an unstable QIF estimator in finite samples. Moreover, if the number of moment conditions is greater than the sample size, the sample covariance matrix $\mathbf{C}_n(\boldsymbol{\beta})$ in (2.3) is singular and therefore the QIF estimator is infeasible. To solve the singularity problem caused by highly over-identified moment conditions, [7, 1, 12] proposed to select a subset of moment conditions for parameter estimation. However, a subset selection approach may lose efficiency in parameter estimation. In the following section, we propose a new method combining all valid moment conditions optimally which is capable of achieving high efficiency in estimation.

3. A new estimation procedure

3.1. Methodology

We first decompose a moment condition vector $\mathbf{G}_n$ into two sets of moment conditions $\mathbf{G}_n = (\mathbf{G}'_{n1}, \mathbf{G}'_{n2})'$, where $\mathbf{G}_{n1}$ could be an s-dimensional preselected moment condition vector, and $\mathbf{G}_{n2}$ are the remaining moment conditions. The preselected moment conditions are the ones which should definitely be included in the estimation. For example, in modeling an unspecified correlation structure, the first set of moment conditions in (2.3) involving the identity basis matrix should be preselected, since the moment conditions generated from any type of correlation structure always contain the one with an identity basis matrix. It is well-known that estimation efficiency can be achieved under the true correlation information. Thus, there might be a loss of estimation efficiency if the moment conditions generated from a misspecified correlation structure are selected. Note that the dimension of preselected moment conditions s is finite, and smaller than the sample size n, to avoid the singularity problem discussed in Section 2.2.

In the following development, we retain the first set of preselected moment conditions $\mathbf{G}_{n1}$ and extract important information from the remaining large-dimensional moment conditions $\mathbf{G}_{n2}$ for parameter estimation. We first orthogonalize $\mathbf{G}_{n2}$ against $\mathbf{G}_{n1}$ to distinguish the contributions of the two sets of moment conditions for estimation, where the orthogonalized moment conditions $\mathbf{G}^o_{n2}$ are obtained by $\mathbf{G}^o_{n2} = \mathbf{G}_{n2} - \mathbf{V}_{21}\mathbf{V}_{11}^{-1}\mathbf{G}_{n1}$, with $\mathbf{V}_{21} = \mathrm{cov}(\mathbf{G}_{n1}, \mathbf{G}_{n2})$ and $\mathbf{V}_{11} = \mathrm{cov}(\mathbf{G}_{n1})$. Through orthogonalization, the two moment conditions $\mathbf{G}_{n1}$ and $\mathbf{G}^o_{n2}$ are no longer correlated, e.g., $\mathrm{cov}(\mathbf{G}_{n1}, \mathbf{G}^o_{n2}) = 0$. In the second step, we reduce the dimension of $\mathbf{G}^o_{n2}$ through spectral decomposition to extract most of the information from $\mathbf{G}^o_{n2}$. Specifically, we convert $\mathbf{G}^o_{n2}$ into linearly uncorrelated moment conditions. It follows that the sample covariance matrix $\mathbf{V}^o_2$ for $\mathbf{G}^o_{n2}$ can be represented as a spectral decomposition $\mathbf{V}^o_2 = \sum_{j=1}^{r} \lambda_j \mathbf{e}_j \mathbf{e}'_j$, where $\mathbf{e}_j$ is the jth eigenvector of $\mathbf{V}^o_2$ corresponding to the jth largest eigenvalue λ_j and $r = p(q+1) - s$. Equivalently, the jth principal component is $\mathbf{e}'_j \mathbf{G}^o_{n2}$, a linear combination of $\mathbf{G}^o_{n2}$.

To reduce the dimensionality of the moment conditions $\mathbf{G}_{n2}$, we select the first t principal components and obtain t orthogonal linear combinations of $\mathbf{G}^o_{n2}$. That is, the reduced moment conditions $\mathbf{G}^*_n$ incorporating the first set of moment conditions $\mathbf{G}_{n1}$ and t principal components are:

$$\mathbf{G}^*_n = \begin{pmatrix} \mathbf{I}_s & \mathbf{0} \\ \mathbf{0} & \mathbf{U} \end{pmatrix} \begin{pmatrix} \mathbf{I}_s & \mathbf{0} \\ -\mathbf{V}_{21}\mathbf{V}_{11}^{-1} & \mathbf{I}_r \end{pmatrix} \begin{pmatrix} \mathbf{G}_{n1} \\ \mathbf{G}_{n2} \end{pmatrix} = \mathbf{T}_2 \mathbf{T}_1 \mathbf{G}_n, \tag{3.1}$$

where $\mathbf{U}$ is the matrix containing t eigenvectors $(\mathbf{e}_1, \ldots, \mathbf{e}_t)'$, and $\mathbf{I}_s$ and $\mathbf{I}_r$ are identity matrices with $s \times s$ and $r \times r$ dimensions. Consequently, the QIF estimator $\hat{\boldsymbol{\beta}}$ based on $\mathbf{G}^*_n$ is obtained via minimizing

$$Q^*_n(\boldsymbol{\beta}) = n\mathbf{G}^*_n(\boldsymbol{\beta})' \mathbf{V}^*_n(\boldsymbol{\beta})^{-1} \mathbf{G}^*_n(\boldsymbol{\beta}), \tag{3.2}$$

where $\mathbf{V}^*_n(\boldsymbol{\beta})$ is the sample covariance matrix of $\mathbf{G}^*_n(\boldsymbol{\beta})$. Note that the objective function in (3.2) can be expressed with the full moment conditions $\mathbf{G}_n$

as $Q_n^* = n\mathbf{G}_n'\mathbf{T}_1'\mathbf{T}_2'\mathbf{V}_n^{*^{-1}}\mathbf{T}_2\mathbf{T}_1\mathbf{G}_n$, which utilizes all moment conditions with a different weighting matrix $\mathbf{T}_1'\mathbf{T}_2'\mathbf{V}_n^{*^{-1}}\mathbf{T}_2\mathbf{T}_1$ (denoted by $\mathbf{V}_n^{-1}$) to capture important information, but with much lower dimension of the sample covariance matrix $\mathbf{V}_n^*$ relative to the sample size. Our method is still applicable if there are no preselected moment conditions G_{n1} if we do not have prior knowledge of a subset of moment conditions to be included for estimation. This will be discussed with an example in our simulation studies.

One important step is to select t such that most of the information from the moment conditions $\mathbf{G}_{n2}$ can be captured. We propose a Bayesian information type of criterion to select the number of principal components t through minimizing the objective function

$$\mathrm{J}(t) = \frac{\mathrm{tr}\{\mathbf{V}_2^o - \tilde{\mathbf{V}}(t)\}}{\mathrm{tr}(\mathbf{V}_2^o)} + t\frac{\log(nr)}{nr}, \tag{3.3}$$

where $\tilde{\mathbf{V}}(t) = \sum_{j=1}^{t}\lambda_j\mathbf{e}_j\mathbf{e}_j'$ and $\mathrm{tr}\{\mathbf{X}\}$ is the trace of a square matrix $\mathbf{X}$. Note that the first term in (3.3) measures the difference between the sample covariance matrices of the moment conditions $\mathbf{G}_{n2}^o$ and t-selected linear combinations of moment conditions. The second term of (3.3) is a penalty function of both n and r to ensure an appropriate convergence rate for a consistent selection of the number of principal components. This penalty term also guarantees that the number of selected principal components is always smaller than the sample size.

The advantage of the proposed approach is that it does not require inversion of the sample covariance matrix $\mathbf{V}_2^o$. This is quite critical when the dimension of moment conditions is high relative to the sample size, and the inversion of the high-dimensional covariance matrix is infeasible. Note that the proposed approach is very different from [5, 4] which require the inverse of the sample covariance matrix to minimize the mean square errors.

3.2. Asymptotic properties

In this section, we provide the asymptotic properties of the proposed estimator, when the number of moment conditions and the sample size both increase. We denote $\boldsymbol{\beta}_0$ as the true parameter, t_0 as the optimal number of selected principal components, $\mathbf{f}_i(\boldsymbol{\beta}) = E\{\mathbf{g}_i(\boldsymbol{\beta})\}$ and $\mathbf{F}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{f}_i(\boldsymbol{\beta})$, where $\mathbf{F}_n(\boldsymbol{\beta}) = \{F_{n1}(\boldsymbol{\beta}), \ldots, F_{nk}(\boldsymbol{\beta})\}'$. The following regularity conditions are required in order to establish the asymptotic properties:

(C-1) There exists a $\boldsymbol{\beta}_0$ such that $\mathbf{f}_i(\boldsymbol{\beta}) = 0$ for all i if and only if $\boldsymbol{\beta} = \boldsymbol{\beta}_0$.
(C-2) $\mathbf{g}_i(\boldsymbol{\beta})$ is continuously differentiable with respect to $\boldsymbol{\beta}$.
(C-3) The parameter space $\mathbf{S}$ is compact and $\boldsymbol{\beta}$ is an interior point of $\mathbf{S}$.
(C-4) $\{G_{nj}(\boldsymbol{\beta}) - F_{nj}(\boldsymbol{\beta})\} \xrightarrow{p} 0$ uniformly for all $\boldsymbol{\beta}$ and j, where $G_{nj}(\boldsymbol{\beta})$ is the jth element in $\mathbf{G}_n(\boldsymbol{\beta})$.
(C-5) Eigenvalues of the variance matrix of $\mathbf{g}_i(\boldsymbol{\beta})$ are uniformly bounded by positive constants and $\lambda_j = O_p(1/nr)$ for any $j > t_0$.

Condition (C-1) states that the population moment conditions exist, and the mean-zero assumption for the estimating function $\mathbf{g}_i(\boldsymbol{\beta})$ enables one to identify the true parameter $\boldsymbol{\beta}_0$. Condition (C-2) is required for the minimization of $Q_n^*(\boldsymbol{\beta})$ in (3.2), where the parameter space is closed and bounded under (C-3). Condition (C-4) ensures that $\mathbf{g}_i(\boldsymbol{\beta})$ satisfies a uniform weak law of large numbers so that the difference between the average sample moments and population moments converges in probability to zero. Condition (C-5) indicates that the asymptotic covariance matrix of our estimator exists and the eigenvalues λ_js are sufficiently small, if they are not selected as one of the principal components.

We first investigate whether the minimizing criterion $\mathrm{J}(t)$ in (3.3) leads to consistent estimation of the covariance matrix. The following lemmas provide the asymptotic rate of convergence for the estimated covariance matrix through a consistent selection of t_0 principal components.

Lemma 1. *If the condition (C-5) holds, there exists t_0 such that $\|\mathbf{V}_2^o - \tilde{\mathbf{V}}(t_0)\| = O_p(n^{-1})$, where $\|\mathbf{X}\|$ is defined as $\sqrt{tr(\mathbf{X}'\mathbf{X})/ij}$ and $i \times j$ is the dimension of matrix $\mathbf{X}$.*

Lemma 1 indicates that the discrepancy (in matrix norms) between the estimated covariance matrix $\tilde{\mathbf{V}}(t_0)$ and the covariance matrix $\mathbf{V}_2^o$ of $\mathbf{G}_{n2}^o$ converges to 0 as $n \to \infty$. The following lemma shows that the number of principal components can be consistently selected based on the criterion $\mathrm{J}(t)$ in (3.3) when the sample size goes to infinity.

Lemma 2. *Under the condition (C-5), there exists a minimizer $\hat{t}$ of $J(t)$ in (3.3) such that $lim_{n\to\infty}$ $Prob[\hat{t} = t_0] = 1$.*

Note that the choice of a penalty function plays an important role in selecting the number of principal components consistently. Here the penalty term in (3.3) vanishes at an appropriate rate such that the number of linear combinations of moment conditions is consistently selected with probability tending to 1. The above lemmas ensure that the proposed criterion $\mathrm{J}(t)$ results in consistent estimation of the covariance matrix $\mathbf{V}_2^o$ for moment conditions $\mathbf{G}_{n2}^o$. The following theorem provides the asymptotic normality and efficiency of the estimator $\hat{\boldsymbol{\beta}}$.

Theorem 1. *If regularity conditions (C-1)–(C-5) hold, there exists a minimizer $\hat{\boldsymbol{\beta}}$ of $Q_n^*(\boldsymbol{\beta})$ in (3.2) which has the following asymptotic properties as $n \to \infty$.*

I. (Consistency) $\hat{\boldsymbol{\beta}} \xrightarrow{\mathrm{p}} \boldsymbol{\beta}_0$.

II. (Asymptotic Normality) $\left(\dot{\mathbf{G}}_n' \mathbf{V}_n^{-1} \mathbf{V} \mathbf{V}_n^{-1} \dot{\mathbf{G}}_n\right)^{-1/2} \left(\dot{\mathbf{G}}_n' \mathbf{V}_n^{-1} \dot{\mathbf{G}}_n\right) \cdot \sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) \xrightarrow{\mathrm{d}} N\left(0, \mathbf{I}\right)$, *where* $\dot{\mathbf{G}}_n = \partial \mathbf{G}_n(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$ *and* $\mathbf{V} = var\{\mathbf{g}_i(\boldsymbol{\beta})\}$.

Theorem 1 indicates that the estimator $\hat{\boldsymbol{\beta}}$ is consistent and asymptotically normal. This implies that asymptotically there is no efficiency loss if the number of principle components t_0 is selected based on the proposed criterion in (3.3), since the sandwich form of the estimated covariance matrix $\left(\dot{\mathbf{G}}_n' \mathbf{V}_n^{-1} \dot{\mathbf{G}}_n\right)^{-1} \left(\dot{\mathbf{G}}_n' \mathbf{V}_n^{-1} \mathbf{V} \mathbf{V}_n^{-1} \dot{\mathbf{G}}_n\right) \left(\dot{\mathbf{G}}_n' \mathbf{V}_n^{-1} \dot{\mathbf{G}}_n\right)^{-1}$ for the estimator $\hat{\boldsymbol{\beta}}$ converges to the asymptotic variance matrix $\left(\dot{\mathbf{F}}' \mathbf{V}^{-1} \dot{\mathbf{F}}\right)^{-1}$, where $\dot{\mathbf{F}} = E\{\partial \mathbf{g}_i(\boldsymbol{\beta})/\partial \boldsymbol{\beta}\}$. Consequently,

the new weighting matrix $\mathbf{V}_n^*(\boldsymbol{\beta})^{-1}$ in $Q_n^*(\boldsymbol{\beta})$ enables one to combine all valid moment conditions optimally without loss of efficiency. Furthermore, the following theorem illustrates that the estimator based on any subset of moment conditions is less efficient than the one utilizing all moment conditions. In the following, we denote $\hat{\boldsymbol{\beta}}_A$ as the estimator based on all moment conditions, and $\hat{\boldsymbol{\beta}}_S$ as the estimator using a subset of moment conditions.

Theorem 2. *Under (C-1)–(C-5), the estimator* $\hat{\boldsymbol{\beta}}_A$ *is more efficient than the estimator* $\hat{\boldsymbol{\beta}}_S$*, that is,* $var(\mathbf{a}'\hat{\boldsymbol{\beta}}_A) \leq var(\mathbf{a}'\hat{\boldsymbol{\beta}}_S)$ *for any constant vector* $\mathbf{a}$.

The above theorem shows that higher estimation efficiency can be achieved by combining all valid moment conditions optimally. The proofs of the lemmas and theorems are provided in the Appendix.

3.3. Implementation with unbalanced data

In longitudinal studies, unbalanced are quite common as cluster size m_i for the ith subject varies due to missing. If the measurements from unbalanced data are regarded as cluster data without considering the order of lag time, then the marginal mean of response $\boldsymbol{\mu}_i$ is a m_i-dimensional vector and basis matrices $\mathbf{B}_j$'s are $m_i \times m_i$ matrices for $i = 1, \ldots, n$. On the other hand, when the lag time between measurements is considered, we provide a strategy to implement the proposed method for unbalanced data using a transformation matrix for each subject.

Let $\mathbf{M}_i$ be a $m \times m_i$ transformation matrix of the ith subject where $m = \max(m_1, \ldots, m_n)$. The matrix $\mathbf{M}_i$'s are generated by deleting the columns of the $m \times m$ identity matrix corresponding to the missing measurements for the ith subject. We transform the unbalanced data to artificial balanced data using $\boldsymbol{\mu}_i^\tau = \mathbf{M}_i\boldsymbol{\mu}_i$, $\mathbf{y}_i^\tau = \mathbf{M}_i\mathbf{y}_i$, and $\mathbf{A}_i^\tau = \mathbf{M}_i\mathbf{A}_i\mathbf{M}_i'$. We then replace $\mathbf{g}_i$ in (2.3) by $\mathbf{g}_i^\tau = \{\dot{\boldsymbol{\mu}}_i^{\tau'}\mathbf{A}_i^{\tau^{-1}}(\mathbf{y}_i^\tau - \boldsymbol{\mu}_i^\tau), \dot{\boldsymbol{\mu}}_i^{\tau'}\mathbf{A}_i^{\tau^{-1/2}}\mathbf{B}_1\mathbf{A}_i^{\tau^{-1/2}}(\mathbf{y}_i^\tau - \boldsymbol{\mu}_i^\tau), \ldots, \dot{\boldsymbol{\mu}}_i^{\tau'}\mathbf{A}_i^{\tau^{-1/2}}\mathbf{B}_q\mathbf{A}_i^{\tau^{-1/2}}(\mathbf{y}_i^\tau - \boldsymbol{\mu}_i^\tau)\}'$. The QIF estimator with unbalanced data is obtained based on the transformed extended score vector. Note that the estimator holds the aforementioned properties if the data is missing completely at random [15].

4. Numerical studies

4.1. Continuous responses

We generate the correlated continuous response variable from a marginal model

$$y_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + \varepsilon_{ij}, \quad \text{for } i = 1, \ldots, n \text{ and } j = 1, \ldots, m,$$

where $\mathbf{x}_{ij} = \big(x_{ij}^{(1)}, x_{ij}^{(2)}\big)'$, $x_{ij}^{(1)} = \frac{j}{m} + N\big(0, \frac{1}{m}\big)$, $x_{ij}^{(2)} = \big(\frac{m-j}{m}\big)^2 + N\big(0, \frac{1}{m}\big)$, $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{im})' \sim N(0, \mathbf{R})$ and $\boldsymbol{\beta} = (\beta_1, \beta_2)' = (1, 1)'$. The repeated responses are generated with a cluster size of $m = 25$, 50 or 100; and the sample size

ranges from $n = 50$ to 500. We design a simulation setting based on a three-block diagonal correlation structure $\mathbf{R}$, 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 $\frac{m}{5} \times \frac{m}{5}$ AR1 structure with correlation 0.6, and the third block has an $\frac{m}{5} \times \frac{m}{5}$ exchangeable structure with correlation 0.8.

The basis matrices are obtained via an eigenvector decomposition, $\mathbf{R}^{-1} \approx a_0\mathbf{I} + \sum_{j=1}^{m} a_j\mathbf{B}_j$, where $\mathbf{B}_j = \mathbf{e}_j\mathbf{e}_j'$ and $\mathbf{e}_j$ is the eigenvector corresponding to the jth largest eigenvalue of the sample correlation matrix of $\mathbf{y}_i$. There are a total of $m + 1$ basis matrices. When $n = 50$, the number of moment conditions $2(m+1)$ 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 due to the singularity problem if the inverse of the sample covariance matrix of the moment conditions is used for the weighting matrix in (2.4).

We compare the performance of the proposed method to the GEE estimators under two types of working correlation structures: exchangeable correlation structure (GEE_{EX}) and AR1 correlation structure (GEE_{AR1}) based on 200 simulations. Here we suppose that the first set of moment conditions containing the identity basis matrix are preselected moment conditions (QIF_{PC1}). To illustrate the importance of utilizing all valid moment conditions using a consistent weighting matrix, we perform QIF parameter estimations based on all valid moment conditions with the identity weighting matrix (QIF_{I}), and a subset of moment conditions (QIF_{Sub}). In addition, we compare all these estimators with the GEE estimator using the true correlation structure, denoted as the oracle estimator. In practice, the oracle estimator cannot be achieved since the true correlation structure is unknown.

To illustrate estimation efficiency, we define the mean squared error $\text{mse}(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{200} \|\hat{\boldsymbol{\beta}}^{(i)} - \boldsymbol{\beta}_0\|^2/(200 \times p)$, where $\hat{\boldsymbol{\beta}}^{(i)}$ is the estimator from the ith simulation, $\boldsymbol{\beta}_0$ is the true parameter, and $\|\cdot\|$ denotes the Euclidean-norm. Figure 1 provides the mean squared errors of the estimators corresponding to various cluster sizes and sample sizes. In addition, Table 1 provides the means and standard errors of the estimators, and the ratio of the mean squared error obtained from other approaches to the mean squared error from the proposed method (QIF_{PC1}).

Our simulations show that the proposed method is superior compared to the GEE under exchangeable and AR1 correlation structures, the QIF using the identity weighting matrix, and a subset of moment conditions in terms of the standard errors and the mean squared errors of the estimators. Specifically, Figure 1 indicates that the mean squared errors of the proposed method's estimators decrease and are closer to those of the oracle estimator as the cluster size increases when the sample size is 50, while the mean squared errors of the GEE approach increase. The relatively low efficiency of the GEE estimator can be explained in that the GEE is inefficient under the misspecified working correlation structure. In contrast, the proposed method is able to improve the finite sample performance of the QIF estimator with a small loss of efficiency.

When the sample size increases to 500, the mean squared errors of the pro-

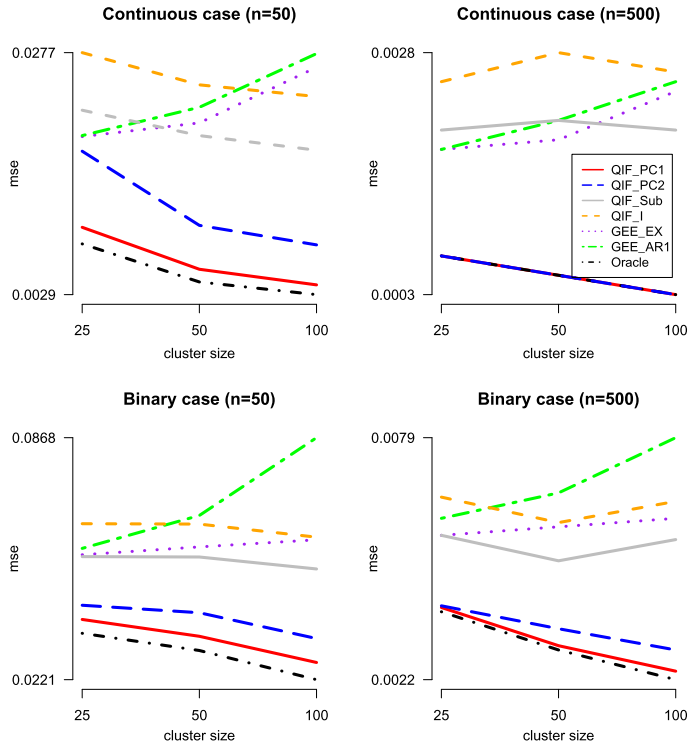

FIG 1. *The first row is the mean squared errors of estimators for continuous responses and the second row is the mean squared errors of estimators for binary responses.*

posed method and the oracle estimator are the same regardless of the cluster size. On the other hand, the QIF using a subset of moment conditions is not able to recover estimation efficiency, as a subset selection approach fails to capture information from the remaining unselected moment conditions. For the QIF approach using the identity weighting matrix, the weighting matrix is not optimal and therefore has a clear loss in efficiency. The GEE estimators with exchangeable or AR1 working correlations also have poor performance with more than eight times the mean squared errors of the proposed method when the cluster size $m = 100$. We also apply the proposed approach assuming that there are no preselected moment conditions where G_{n1} in (3.1) is an empty set (QIF_{PC2}). When the sample size is small, the resulting estimator is not as efficient as QIF_{PC1}, although it still outperforms the existing methods. However, estimation efficiency of QIF_{PC2} can be achieved with a large sample size $n = 500$.

We further investigate whether the BIC criterion selects the optimal number of principal components in a finite sample. Figure 2 illustrates the mean squared errors of the QIF approach using the reduced moment condition conditions $\mathbf{G}_n^*$ generated by t principal components when t varies between 0 and 30 for $n = 50$ and $m = 25, 50$ and 100. Figure 2 shows that the minimum of the MSE of the

TABLE 1

For continuous responses, mean and standard errors (se) of estimators, and the ratio of the mean squared error (mse) for other approaches to the mse for the proposed method

m	Method	$n = 50$			$n = 500$		
		mean$(\hat{\beta}_1)_{\text{se}}$	mean$(\hat{\beta}_2)_{\text{se}}$	ratio	mean$(\hat{\beta}_1)_{\text{se}}$	mean$(\hat{\beta}_2)_{\text{se}}$	ratio
25	QIF_{PC1}	$1.002_{0.083}$	$0.993_{0.109}$	1.00	$0.997_{0.026}$	$0.997_{0.027}$	1.00
	QIF_{PC2}	$1.004_{0.100}$	$0.995_{0.138}$	1.80	$0.997_{0.026}$	$0.996_{0.028}$	1.00
	QIF_{Sub}	$0.993_{0.089}$	$0.996_{0.189}$	2.22	$0.998_{0.028}$	$1.003_{0.056}$	2.86
	QIF_{I}	$0.992_{0.096}$	$0.996_{0.216}$	2.82	$0.998_{0.030}$	$1.003_{0.064}$	3.57
	GEE_{EX}	$0.991_{0.112}$	$0.993_{0.160}$	1.95	$0.998_{0.034}$	$1.003_{0.049}$	2.57
	GEE_{AR1}	$0.999_{0.105}$	$0.979_{0.164}$	1.96	$0.998_{0.033}$	$1.003_{0.049}$	2.58
	Oracle	$0.995_{0.078}$	$0.995_{0.101}$	0.83	$0.998_{0.025}$	$0.998_{0.027}$	1.00
50	QIF_{PC1}	$1.005_{0.067}$	$1.006_{0.078}$	1.00	$1.000_{0.022}$	$0.997_{0.023}$	1.00
	QIF_{PC2}	$1.003_{0.092}$	$1.006_{0.100}$	1.89	$0.999_{0.022}$	$0.998_{0.024}$	1.00
	QIF_{Sub}	$0.997_{0.073}$	$1.004_{0.182}$	3.49	$0.996_{0.027}$	$0.998_{0.059}$	4.19
	QIF_{I}	$0.998_{0.079}$	$1.004_{0.207}$	4.44	$0.996_{0.029}$	$0.999_{0.068}$	5.58
	GEE_{EX}	$0.994_{0.123}$	$1.000_{0.161}$	3.72	$0.995_{0.042}$	$0.997_{0.044}$	3.83
	GEE_{AR1}	$1.004_{0.132}$	$1.004_{0.170}$	4.02	$0.997_{0.039}$	$1.002_{0.052}$	4.20
	Oracle	$1.001_{0.058}$	$1.006_{0.070}$	0.76	$0.998_{0.021}$	$0.997_{0.023}$	1.00
100	QIF_{PC1}	$0.997_{0.060}$	$0.993_{0.061}$	1.00	$0.999_{0.020}$	$1.000_{0.017}$	1.00
	QIF_{PC2}	$0.995_{0.063}$	$0.990_{0.081}$	2.08	$1.000_{0.021}$	$1.000_{0.018}$	1.00
	QIF_{Sub}	$0.993_{0.071}$	$0.966_{0.171}$	4.54	$1.000_{0.026}$	$0.994_{0.057}$	6.67
	QIF_{I}	$0.994_{0.081}$	$0.962_{0.196}$	5.95	$1.000_{0.029}$	$0.993_{0.065}$	8.66
	GEE_{EX}	$1.001_{0.198}$	$0.974_{0.112}$	6.72	$1.003_{0.057}$	$0.998_{0.038}$	8.00
	GEE_{AR1}	$0.974_{0.148}$	$0.991_{0.181}$	7.08	$1.002_{0.042}$	$0.999_{0.055}$	8.33
	Oracle	$0.995_{0.054}$	$0.994_{0.054}$	0.74	$1.000_{0.019}$	$1.000_{0.017}$	1.00

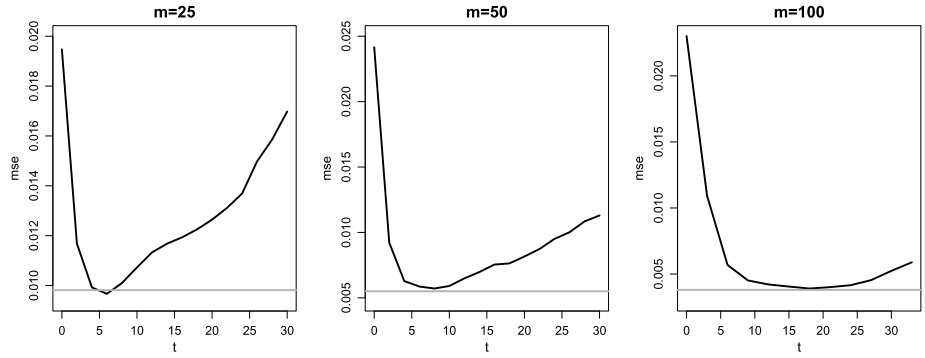

FIG 2. *The mean squared errors of the QIF based on the reduced moment condition vector generated by t principal components with n = 50. The gray line denotes the mean squared errors of the proposed approach based on the BIC criterion.*

QIF estimator is slightly below the one selected by the BIC when $m = 25$. However, when the cluster size m increases to 50 and 100, the MSE of the QIF estimator based on the BIC reaches the minimum. This indicates that the BIC criterion is quite effective in selecting the number of moment conditions if the cluster size and sample size are moderately large.

TABLE 2
For binary responses, mean and standard errors (se) of estimators, and the ratio of the mean squared error (mse) for other approaches to the mse for the proposed method

m	Method	$n = 50$			$n = 500$		
		$\text{mean}(\hat{\beta}_1)_{\text{se}}$	$\text{mean}(\hat{\beta}_2)_{\text{se}}$	ratio	$\text{mean}(\hat{\beta}_1)_{\text{se}}$	$\text{mean}(\hat{\beta}_2)_{\text{se}}$	ratio
25	QIF_{PC1}	$0.490_{0.208}$	$-0.513_{0.175}$	1.00	$0.494_{0.064}$	$-0.496_{0.060}$	1.00
	QIF_{PC2}	$0.488_{0.215}$	$-0.510_{0.181}$	1.10	$0.492_{0.064}$	$-0.495_{0.061}$	1.01
	QIF_{Sub}	$0.529_{0.252}$	$-0.505_{0.214}$	1.44	$0.507_{0.078}$	$-0.508_{0.070}$	1.44
	QIF_{I}	$0.538_{0.270}$	$-0.508_{0.231}$	1.67	$0.508_{0.084}$	$-0.510_{0.075}$	1.67
	GEE_{EX}	$0.529_{0.252}$	$-0.507_{0.216}$	1.45	$0.506_{0.078}$	$-0.508_{0.070}$	1.44
	GEE_{AR1}	$0.508_{0.259}$	$-0.496_{0.218}$	1.50	$0.504_{0.076}$	$-0.509_{0.077}$	1.54
	Oracle	$0.513_{0.200}$	$-0.507_{0.169}$	0.90	$0.498_{0.064}$	$-0.501_{0.059}$	0.97
50	QIF_{PC1}	$0.525_{0.196}$	$-0.481_{0.167}$	1.00	$0.494_{0.055}$	$-0.495_{0.053}$	1.00
	QIF_{PC2}	$0.503_{0.203}$	$-0.486_{0.178}$	1.19	$0.495_{0.059}$	$-0.498_{0.054}$	1.15
	QIF_{Sub}	$0.537_{0.245}$	$-0.492_{0.220}$	1.63	$0.507_{0.072}$	$-0.500_{0.069}$	1.67
	QIF_{I}	$0.542_{0.265}$	$-0.496_{0.236}$	1.89	$0.508_{0.078}$	$-0.501_{0.074}$	1.97
	GEE_{EX}	$0.537_{0.253}$	$-0.493_{0.224}$	1.71	$0.507_{0.072}$	$-0.500_{0.069}$	1.93
	GEE_{AR1}	$0.541_{0.256}$	$-0.519_{0.254}$	1.96	$0.501_{0.081}$	$-0.501_{0.081}$	2.20
	Oracle	$0.525_{0.183}$	$-0.484_{0.159}$	0.89	$0.502_{0.053}$	$-0.500_{0.053}$	0.97
100	QIF_{PC1}	$0.510_{0.164}$	$-0.494_{0.159}$	1.00	$0.499_{0.049}$	$-0.498_{0.047}$	1.00
	QIF_{PC2}	$0.514_{0.190}$	$-0.494_{0.166}$	1.24	$0.497_{0.055}$	$-0.497_{0.050}$	1.21
	QIF_{Sub}	$0.519_{0.230}$	$-0.505_{0.224}$	1.94	$0.498_{0.078}$	$-0.499_{0.069}$	2.29
	QIF_{I}	$0.523_{0.248}$	$-0.507_{0.241}$	2.25	$0.498_{0.084}$	$-0.498_{0.075}$	2.67
	GEE_{EX}	$0.517_{0.239}$	$-0.508_{0.243}$	2.23	$0.498_{0.080}$	$-0.499_{0.072}$	2.51
	GEE_{AR1}	$0.515_{0.300}$	$-0.496_{0.290}$	3.25	$0.494_{0.095}$	$-0.499_{0.082}$	3.30
	Oracle	$0.507_{0.149}$	$-0.488_{0.148}$	0.82	$0.499_{0.049}$	$-0.499_{0.046}$	0.92

4.2. Binary responses

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

$$\text{logit}(\mu_{ij}) = \mathbf{x}_{ij}'\boldsymbol{\beta}, \qquad \text{for } i = 1, \ldots, n \text{ and } j = 1, \ldots, m,$$

where $\mathbf{x}_{ij} = \left(x_{ij}^{(1)}, x_{ij}^{(2)}\right)'$ and $\boldsymbol{\beta} = (\beta_1, \beta_2)' = (0.5, -0.5)'$. We choose the sample sizes to be 50 and 500, and the cluster sizes to be 25, 50 and 100 from 200 simulations, respectively. 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 $\rho = (0.8, 0.5, 0.7)$.

Similar to the continuous case, we compare the proposed method with the QIF based on the identity weighting matrix, a subset of moment conditions, and the GEE approach with two working correlation matrices. Simulation results with various cluster sizes and sample sizes are reported in Table 2 and Figure 1, which confirms that the proposed method combining all moment conditions outperforms the other methods.

When the sample size increases, Table 2 indicates that the ratio of the mean squared error of the oracle estimators to the mean squared error for the proposed

approach is closer to 1. Moreover, Figure 1 shows that the proposed method provides more efficient estimation as the cluster size increases. However, this does not hold for the GEE method and the QIF approaches based on the identity weighting matrix or a subset of moment conditions even when the sample size reaches 500. The simulation results from binary responses are quite comparable to those reported for the continuous responses.

4.3. Fortune 500 data example

We apply Fortune 500 data between 2000 and 2010 to illustrate the proposed approach. The 136 largest US corporations were ranked among the Global 500 in 2010, and 105 of these companies have been ranked over 11 consecutive years in the Fortune 500 data. Therefore we choose the sample size as 105 with an equal cluster size of 11. For this data, we apply a log-linear model based on the employee demand equation, where the response variable is the number of employees (Employees) from each firm, and corresponding covariates of interest are the revenue and the assets. The log-linear model is formulated as follows:

$$\log(\text{Employees})_{ij} = \beta_0 + \beta_1 \log(\text{Revenue})_{ij} + \beta_2 \log(\text{Assets})_{ij} + \varepsilon_{ij}$$

for $i = 1, \ldots, 105$ and $j = 1, \ldots, 11$, where $\log(\text{Employees})_{ij}$, $\log(\text{Revenue})_{ij}$ and $\log(\text{Assets})_{ij}$ are the log of the employees, the revenue and the assets for the firm i at the jth year respectively.

Through an eigenvector decomposition of the sample correlation matrix for the response, a total of 12 basis matrices are generated as $R^{-1} \approx a_0 I + \sum_{j=1}^{11} a_j B_j$, where $B_j = e_j e_j'$ and e_j is the eigenvector corresponding to the jth largest eigenvalue of its sample correlation matrix. Therefore a total of 36 valid moment conditions are constructed for parameter estimation. We implement the proposed method, and compare it with the QIF using the identity weighting matrix and a subset of moment conditions, and the GEE method with exchangeable and AR1 working correlation structures. Note that the sample covariance matrix from all available moment conditions is not applicable for the weighting matrix, since it is nearly singular due to high collinearity among some of the moment conditions.

Table 3 provides the parameter estimators, the standard errors of the corresponding estimators, Z test statistics and the p-values. In general, the estimators obtained by the proposed method are the most sensible compared to other approaches. Specifically, the coefficients of log(Revenue) and log(Assets) from the proposed method are all positive, implying that the response variable of the number of employees and the predictive variables of revenue and assets are positively associated, with the corresponding p-values all less than 0.001. On the other hand, the p-value of the log(Revenue) using the GEE under the AR1 working correlation is insignificant (p-value=0.252). The QIF using the identity weighting matrix and a subset of moment conditions produces more extreme coefficient estimators for log(Revenue) and intercept compared to the other approaches, and negative coefficients of log(Assets) with insignificant p-values.

TABLE 3
For the Fortune 500 data, comparison of estimated coefficients along with the standard errors (se), Z test statistics and p-values

Covariate	Method	Estimator	se	Z	p-value
intercept	QIF_{PC1}	−3.221	1.166	−2.76	0.006
	QIF_{Sub}	−6.466	2.702	−2.39	0.017
	QIF_{I}	−6.466	0.003	−1992.02	< 0.001
	GEE_{EX}	−1.996	0.873	−2.29	0.022
	GEE_{AR1}	−3.130	1.230	−2.55	0.011
log(Revenue)	QIF_{PC1}	0.244	0.071	3.40	< 0.001
	QIF_{Sub}	0.777	0.158	4.90	< 0.001
	QIF_{I}	0.786	0.074	10.56	< 0.001
	GEE_{EX}	0.148	0.059	2.50	0.012
	GEE_{AR1}	0.070	0.061	1.14	0.252
log(Assets)	QIF_{PC1}	0.352	0.055	6.35	< 0.001
	QIF_{Sub}	−0.049	0.088	−0.56	0.577
	QIF_{I}	−0.058	0.074	−0.78	0.434
	GEE_{EX}	0.391	0.057	6.84	< 0.001
	GEE_{AR1}	0.515	0.071	7.22	< 0.001

This data example confirms that the proposed method utilizing all available moment conditions with a consistent weighting matrix provides better interpretable estimations.

5. Discussion

We propose an efficient and stable QIF estimation procedure through combining all available moment conditions when the dimension of moment conditions is large compared to the sample size. The proposed procedure utilizes a set of preselected moment conditions in addition to optimal linear combinations of remaining moment conditions through principle component analysis. The new approach allows one to reduce the dimensionality of moment conditions, while retaining most of the important information from all valid moment conditions. This is very different from existing approaches which obtain information from a subset of moment conditions only. The performance of the QIF approach relies on selecting the number of principal components accurately, which is essential for estimation efficiency. We provide a new objective function based on the Bayesian information type of criterion. This selects the optimal number of principle components consistently and leads to desirable asymptotic properties such as consistency, asymptotic normality and efficiency for the proposed estimator. We also try other selection criteria such as the AIC, the corrected AIC and the corrected RIC. However, our simulation studies, which are not provided here, indicate that the BIC provides better performance than the other selection criteria in that the BIC selects more accurate t_0 and leads to more efficient parameter estimation. This is because the other criteria tend to over-select the number of principal components, which leads to a loss of estimation efficiency.

In recent years, estimating the inverse of the high-dimensional covariance matrix has become increasingly important due to the rise of big data. The proposed method can also be applied in choosing an appropriate rank for a singular or nearly singular covariance matrix. This is quite useful in low-rank approximation for high-dimensional matrix problems, which has wide applications such as in data compression, large-dimensional matrix operations, recommender systems and machine learning.

Acknowledgements

The authors are very grateful to the Editor, an associate editor and a referee for their insightful comments and suggestions that have improved the manuscript significantly. The research was supported by a National Science Foundation Grant (DMS-1308227).

Appendix

Proof of Lemma 1. By spectral decomposition, the sample covariance matrix of the orthogonalized moment conditions $\mathbf{G}_{n2}^o$ is decomposed as $\mathbf{V}_2^o = \sum_{j=1}^r \lambda_j \mathbf{e}_j \mathbf{e}_j'$, where $\mathbf{e}_j = (e_{1j}, \ldots, e_{rj})'$ is the jth eigenvector corresponding to the jth largest eigenvalue of $\mathbf{V}_2^o$. Since every component of the eigenvector $\mathbf{e}_j$ for $\mathbf{V}_2^o$ is uniformly bounded, there exist constants K_1 and K_2 such that $0 < K_1 < |e_{ij}| < K_2 < \infty$ for $i = 1, \ldots, r$. It follows from (C-5) that

$$\|\mathbf{V}_2^o - \tilde{\mathbf{V}}(t_0)\| = \Big\| \sum_{j=t_0+1}^{r} \lambda_j \mathbf{e}_j \mathbf{e}_j' \Big\| \le K_2 \sqrt{\sum_{j=t_0+1}^{r} \lambda_j^2} \le K_2 \sum_{j=t_0+1}^{r} \lambda_j = O_p(1/n).$$

□

Proof of Lemma 2. We need to show that $\lim_{n\to\infty} P\big\{\mathrm{J}(\hat{t}) < \mathrm{J}(t_0)\big\} = 0$ for all $\hat{t} \neq t_0$ and $\hat{t}$ is a finite integer. Since we have

$$\mathrm{J}(t_0) - \mathrm{J}(\hat{t}) = \frac{\mathrm{tr}\big\{\tilde{\mathbf{V}}(\hat{t}) - \tilde{\mathbf{V}}(t_0)\big\}}{\mathrm{tr}(\mathbf{V}_2^o)} + (t_0 - \hat{t}) \frac{\log(nr)}{nr},$$

it is sufficient to prove that

$$P\big[\mathrm{tr}\{\tilde{\mathbf{V}}(t_0) - \tilde{\mathbf{V}}(\hat{t})\} - \mathrm{tr}(\mathbf{V}_2^o)(t_0 - \hat{t}) \frac{\log(nr)}{nr} < 0\big] \to 0 \text{ as } n \to \infty.$$

First, we consider $\hat{t} < t_0$. Note that $\frac{1}{r}\mathrm{tr}(\mathbf{V}_2^o) = O_p(1)$, because it is bounded by $K_1^2 \sum_{j=1}^r \lambda_j < \frac{1}{r}\mathrm{tr}(\mathbf{V}_2^o) < K_2^2 \sum_{j=1}^r \lambda_j$. Since the eigenvector of $\mathbf{V}_2^o$ is bounded, it follows that

$$\begin{aligned}
&\frac{1}{r}\mathrm{tr}\big\{\tilde{\mathbf{V}}(t_0)-\tilde{\mathbf{V}}(\hat{t})\big\}-\frac{1}{r}\mathrm{tr}(\mathbf{V}_2^o)(t_0-\hat{t})\frac{\log(nr)}{nr}\\
&=\frac{1}{r}\mathrm{tr}\Bigg(\sum_{j=\hat{t}+1}^{t_0}\lambda_j\mathbf{e}_j\mathbf{e}_j'\Bigg)-\frac{1}{r}\mathrm{tr}(\mathbf{V}_2^o)(t_0-\hat{t})\frac{\log(nr)}{nr}\\
&\geq K_1^2\sum_{j=\hat{t}+1}^{t_0}\lambda_j-\frac{1}{r}\mathrm{tr}(\mathbf{V}_2^o)(t_0-\hat{t})\frac{\log(nr)}{nr}\to K_1^2\sum_{j=\hat{t}+1}^{t_0}\lambda_j>0,
\end{aligned}$$

as $n\to\infty$. Therefore, $\lim_{n\to\infty}P\big\{\mathrm{J}(\hat{t})-\mathrm{J}(t_0)<0\big\}=0$ holds.
Second, we consider $\hat{t}>t_0$. We have

$$P\big\{n\mathrm{J}(\hat{t})-n\mathrm{J}(t_0)<0\big\}=P\Big[n\mathrm{tr}\big\{\tilde{\mathbf{V}}(\hat{t})-\tilde{\mathbf{V}}(t_0)\big\}>\frac{1}{r}\mathrm{tr}(\mathbf{V}_2^o)(\hat{t}-t_0)\log(nr)\Big],$$

and we obtain

$$\begin{aligned}
n\mathrm{tr}\big\{\tilde{\mathbf{V}}(\hat{t})-\tilde{\mathbf{V}}(t_0)\big\} &= n\mathrm{tr}\Bigg(\sum_{j=t_0+1}^{\hat{t}}\lambda_j\mathbf{e}_j\mathbf{e}_j'\Bigg)\\
&> K_1^2nr\sum_{j=t_0+1}^{\hat{t}}\lambda_j=K_1^2nrO_p(1/nr)=O_p(1).
\end{aligned}$$

On the other hand, $\frac{1}{r}\mathrm{tr}(\mathbf{V}_2^o)(\hat{t}-t_0)\log(nr)=O\big\{\log(nr)\big\}$.
Consequently, this ensures that as $n\to\infty$,

$$P\big\{\mathrm{J}(\hat{t})-\mathrm{J}(t_0)<0\big\}=P\Big[\mathrm{tr}\big\{\tilde{\mathbf{V}}(\hat{t})-\tilde{\mathbf{V}}(t_0)\big\}>\mathrm{tr}(\mathbf{V}_2^o)(\hat{t}-t_0)\frac{\log(nr)}{nr}\Big]\to 0.\ \square$$

Proof of Theorem 1. The estimator $\hat{\boldsymbol{\beta}}$ is defined as $\hat{\boldsymbol{\beta}}=\mathrm{argmin}\ Q_n^*(\boldsymbol{\beta})$

$$=\mathrm{argmin}\ n\mathbf{G}_n^*(\boldsymbol{\beta})'\mathbf{V}_n^{*^{-1}}\mathbf{G}_n^*(\boldsymbol{\beta})=\mathrm{argmin}\ n\mathbf{G}_n(\boldsymbol{\beta})'\mathbf{V}_n^{-1}\mathbf{G}_n(\boldsymbol{\beta}),\tag{5.1}$$

where $\mathbf{G}_n^*=\mathbf{T}_2\mathbf{T}_1\mathbf{G}_n$ and $\mathbf{V}_n^{-1}=\mathbf{T}_1'\mathbf{T}_2'\mathbf{V}_n^{*^{-1}}\mathbf{T}_2\mathbf{T}_1$. Lemmas 1 and 2 confirm that the discrepancy between $\mathbf{V}_n$ and the sample covariance matrix of $\mathbf{g}_i$ $\mathbf{C}_n$ converges to 0 in probability. In addition, $\mathbf{C}_n$ converges almost surely to $\mathbf{V}$ by the weak law of large numbers. Thus, it immediately follows from (C-4) that there exists $\tilde{Q}_n(\boldsymbol{\beta})=\mathbf{F}_n(\boldsymbol{\beta})'\mathbf{V}^{-1}\mathbf{F}_n(\boldsymbol{\beta})$ such that

$$Q_n^*(\boldsymbol{\beta})-\tilde{Q}_n(\boldsymbol{\beta})\xrightarrow{\mathrm{p}}0\ \text{uniformly for all}\ \boldsymbol{\beta}.\tag{5.2}$$

We now deduce from (C-1) that $\tilde{Q}_n(\boldsymbol{\beta})=0$ if and only if $\boldsymbol{\beta}=\boldsymbol{\beta}_0$, and $\tilde{Q}_n(\boldsymbol{\beta})>0$ otherwise. Consequently, this ensures that

$$\boldsymbol{\beta}_0=\mathrm{argmin}\ \tilde{Q}_n(\boldsymbol{\beta}).\tag{5.3}$$

Hence, (5.1), (5.2) and (5.3) imply that $\hat{\boldsymbol{\beta}} \xrightarrow{\mathrm{p}} \boldsymbol{\beta}_0$.

Next we prove the asymptotic normality of the estimator $\hat{\boldsymbol{\beta}}$. By Taylor expansion, we have

$$\mathbf{G}_n(\hat{\boldsymbol{\beta}}) = \mathbf{G}_n(\boldsymbol{\beta}_0) + \dot{\mathbf{G}}_n(\tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0), \tag{5.4}$$

where $\tilde{\boldsymbol{\beta}}$ lies between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$. By multiplying the equation (5.4) by $\dot{\mathbf{G}}_n(\hat{\boldsymbol{\beta}})'\mathbf{V}_n^{-1}$, it immediately follows that

$$\dot{\mathbf{G}}_n(\hat{\boldsymbol{\beta}})'\mathbf{V}_n^{-1}\mathbf{G}_n(\hat{\boldsymbol{\beta}}) = \dot{\mathbf{G}}_n(\hat{\boldsymbol{\beta}})'\mathbf{V}_n^{-1}\mathbf{G}_n(\boldsymbol{\beta}_0) + \dot{\mathbf{G}}_n(\hat{\boldsymbol{\beta}})'\mathbf{V}_n^{-1}\dot{\mathbf{G}}_n(\tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0). \tag{5.5}$$

Since the left hand side $\dot{\mathbf{G}}_n(\hat{\boldsymbol{\beta}})'\mathbf{V}_n^{-1}\mathbf{G}_n(\hat{\boldsymbol{\beta}}) = \dot{Q}_n(\hat{\boldsymbol{\beta}}) = 0$ in (5.5), we rearrange the equation as follows:

$$\begin{aligned} & \{\dot{\mathbf{G}}_n(\hat{\boldsymbol{\beta}})'\mathbf{V}_n^{-1}\mathbf{V}\mathbf{V}_n^{-1}\dot{\mathbf{G}}_n(\hat{\boldsymbol{\beta}})\}^{-1/2}\{\dot{\mathbf{G}}_n(\hat{\boldsymbol{\beta}})'\mathbf{V}_n^{-1}\dot{\mathbf{G}}_n(\tilde{\boldsymbol{\beta}})\} \cdot \sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\ = \quad & -\{\dot{\mathbf{G}}_n(\hat{\boldsymbol{\beta}})'\mathbf{V}_n^{-1}\mathbf{V}\mathbf{V}_n^{-1}\dot{\mathbf{G}}_n(\hat{\boldsymbol{\beta}})\}^{-1/2}\dot{\mathbf{G}}_n(\hat{\boldsymbol{\beta}})'\mathbf{V}_n^{-1}\mathbf{V}^{1/2} \cdot \sqrt{n}\mathbf{V}^{-1/2}\mathbf{G}_n(\boldsymbol{\beta}_0). \end{aligned}$$

Note that we replace $\dot{\mathbf{G}}_n(\tilde{\boldsymbol{\beta}})$ on the left side with $\dot{\mathbf{G}}_n(\hat{\boldsymbol{\beta}})$, since $\tilde{\boldsymbol{\beta}}$ converges to $\boldsymbol{\beta}_0$ in probability. Hence it follows from $\sqrt{n}\mathbf{V}^{-1/2}\mathbf{G}_n(\boldsymbol{\beta}_0) \xrightarrow{\mathrm{d}} N(0, \mathbf{I})$ by the central limit theorem that

$$\{\dot{\mathbf{G}}_n(\hat{\boldsymbol{\beta}})'\mathbf{V}_n^{-1}\mathbf{V}\mathbf{V}_n^{-1}\dot{\mathbf{G}}_n(\hat{\boldsymbol{\beta}})\}^{-1/2}\{\dot{\mathbf{G}}_n(\hat{\boldsymbol{\beta}})'\mathbf{V}_n^{-1}\dot{\mathbf{G}}_n(\hat{\boldsymbol{\beta}})\} \cdot \sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{\mathrm{d}} N(0, \mathbf{I}).$$

□

Proof of Theorem 2. Let $\mathbf{G}_n(\boldsymbol{\beta}) = \{\mathbf{G}_{n1}(\boldsymbol{\beta})', \mathbf{G}_{n2}(\boldsymbol{\beta})'\}'$, which contains two sets of moment conditions, and denote $\hat{\boldsymbol{\beta}}_S$ as the estimator using any arbitrary set $\mathbf{G}_{n1}(\boldsymbol{\beta})$ only. We orthogonalize $\mathbf{G}_{n2}(\boldsymbol{\beta})$ from $\mathbf{G}_{n1}(\boldsymbol{\beta})$ as

$$\mathbf{G}_{n2}^o(\boldsymbol{\beta}) = \mathbf{G}_{n2}(\boldsymbol{\beta}) - \mathbf{V}_{21}\mathbf{V}_{11}^{-1}\mathbf{G}_{n1}(\boldsymbol{\beta}),$$

where $\mathbf{V}_{21} = \mathrm{cov}\{\mathbf{G}_{n1}(\boldsymbol{\beta}), \mathbf{G}_{n2}(\boldsymbol{\beta})\}$ and $\mathbf{V}_{11} = \mathrm{cov}\{\mathbf{G}_{n1}(\boldsymbol{\beta})\}$. Through the orthogonalization, $\mathrm{cov}\{\mathbf{G}_{n1}(\boldsymbol{\beta}), \mathbf{G}_{n2}^o(\boldsymbol{\beta})\} = 0$. We denote $\mathbf{G}_n^o(\boldsymbol{\beta}) = \{\mathbf{G}_{n1}(\boldsymbol{\beta})', \mathbf{G}_{n2}^o(\boldsymbol{\beta})'\}'$. The estimator $\hat{\boldsymbol{\beta}}_A$ is obtained by minimizing $\mathbf{G}_n(\boldsymbol{\beta})'\mathbf{V}^{-1}\mathbf{G}_n(\boldsymbol{\beta})$, which is equivalent to $\mathbf{G}_n^o(\boldsymbol{\beta})'\mathbf{V}^{o-1}\mathbf{G}_n^o(\boldsymbol{\beta})$, where $\mathbf{V} = \mathrm{cov}\{\mathbf{G}_n(\boldsymbol{\beta})\}$ and $\mathbf{V}^o = \mathrm{cov}\{\mathbf{G}_n^o(\boldsymbol{\beta})\}$.

The information matrix of $\hat{\boldsymbol{\beta}}_A$ is proportional to

$$\begin{aligned} \dot{\mathbf{G}}_n(\hat{\boldsymbol{\beta}}_A)'\mathbf{V}^{-1}\dot{\mathbf{G}}_n(\hat{\boldsymbol{\beta}}_A) &= \mathbf{G}_n^o(\hat{\boldsymbol{\beta}}_A)'\mathbf{V}^{o-1}\mathbf{G}_n^o(\hat{\boldsymbol{\beta}}_A) \\ &= \begin{pmatrix} \dot{\mathbf{G}}_{n1}(\hat{\boldsymbol{\beta}}_A) \\ \dot{\mathbf{G}}_{n2}^o(\hat{\boldsymbol{\beta}}_A) \end{pmatrix}' \begin{pmatrix} \mathbf{V}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_{22}^{o-1} \end{pmatrix} \begin{pmatrix} \dot{\mathbf{G}}_{n1}(\hat{\boldsymbol{\beta}}_A) \\ \dot{\mathbf{G}}_{n2}^o(\hat{\boldsymbol{\beta}}_A) \end{pmatrix} \\ &= \dot{\mathbf{G}}_{n1}(\hat{\boldsymbol{\beta}}_A)'\mathbf{V}_{11}^{-1}\dot{\mathbf{G}}_{n1}(\hat{\boldsymbol{\beta}}_A) + \dot{\mathbf{G}}_{n2}^o(\hat{\boldsymbol{\beta}}_A)'\mathbf{V}_{22}^{o-1}\dot{\mathbf{G}}_{n2}^o(\hat{\boldsymbol{\beta}}_A), \end{aligned}$$

where $\mathbf{V}_{22}^o = \mathrm{cov}\{\mathbf{G}_{n2}^o(\boldsymbol{\beta})\}$. Under $\hat{\boldsymbol{\beta}}_A \xrightarrow{\mathrm{p}} \boldsymbol{\beta}_0$, it follows that

$$\dot{\mathbf{G}}_n(\hat{\boldsymbol{\beta}}_A)'\mathbf{V}^{-1}\dot{\mathbf{G}}_n(\hat{\boldsymbol{\beta}}_A) \to \dot{\mathbf{G}}_{n1}(\boldsymbol{\beta}_0)'\mathbf{V}_{11}^{-1}\dot{\mathbf{G}}_{n1}(\boldsymbol{\beta}_0) + \dot{\mathbf{G}}_{n2}^o(\boldsymbol{\beta}_0)'\mathbf{V}_{22}^{o-1}\dot{\mathbf{G}}_{n2}^o(\boldsymbol{\beta}_0). \tag{5.6}$$

On the other hand, we obtain $\hat{\boldsymbol{\beta}}_S$ by minimizing $\mathbf{G}_{n1}(\boldsymbol{\beta})'\mathbf{V}_{11}^{-1}\mathbf{G}_{n1}(\boldsymbol{\beta})$, which utilizes the first set of moment conditions. Since the estimator $\hat{\boldsymbol{\beta}}_S$ converges to $\boldsymbol{\beta}_0$, we have

$$\dot{\mathbf{G}}_{n1}(\hat{\boldsymbol{\beta}}_S)'\mathbf{V}_{11}^{-1}\dot{\mathbf{G}}_{n1}(\hat{\boldsymbol{\beta}}_S) \to \dot{\mathbf{G}}_{n1}(\boldsymbol{\beta}_0)'\mathbf{V}_{11}^{-1}\dot{\mathbf{G}}_{n1}(\boldsymbol{\beta}_0). \tag{5.7}$$

Considering that $\mathbf{V}_{22}^o$ in (5.6) is a non-negative definite weighting matrix, it consequently follows from (5.6) and (5.7) that

$$\dot{\mathbf{G}}_n(\boldsymbol{\beta}_0)'\mathbf{V}^{-1}\dot{\mathbf{G}}_n(\boldsymbol{\beta}_0) \geq \dot{\mathbf{G}}_{n1}(\boldsymbol{\beta}_0)'\mathbf{V}_{11}^{-1}\dot{\mathbf{G}}_{n1}(\boldsymbol{\beta}_0)$$

in the sense of Loewner ordering. Therefore, the efficiency of the estimator $\hat{\boldsymbol{\beta}}_A$ is improved by utilizing all moment conditions instead of using some of the moment conditions. □

References

[1] 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. MR1805875

[2] Belloni, A., Chen, D., Chernozhukov, V. and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* **80**, 2369–2429. MR3001131

[3] Caner, M. and Fan, Q. (2015). Hybrid generalized empirical likelihood estimators: Instrument selection with adaptive lasso. *Journal of Econometrics* **187**, 256–274. MR3347306

[4] 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. MR2562761

[5] Donald, S. G. and Newey, W. K. (2001). Choosing the number of instruments. *Econometrica* **69**, 1161–1191. MR1848779

[6] Fan, J. and Liao, Y. (2011). Ultra high dimensional variable selection with endogenous covariates. In manuscript.

[7] Gallant, A. R. and Tauchen, G. (1996). Which moments to match? *Econometric Theory* **12**, 657–681. MR1422547

[8] Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50**, 1029–1054. MR0666123

[9] Imbens, G. W. (2002). Generalized method of moments and empirical likelihood. *Journal of Business and Economic Statistics* **20**, 493–506. MR1973800

[10] Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 12–22. MR0836430

[11] Newey, W. K. and Smith, R. J. (2004). Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica* **72**, 219–255. MR2031017

[12] Okui, R. (2009). The optimal choice of moments in dynamic panel data models. *Journal of Econometrics* **151**, 1–16. MR2538268

[13] Qu, A., Lindsay, B. G. and Li, B. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika* **87**, 823–836. MR1813977

[14] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464. MR0468014

[15] Zhou, J. and Qu, A. (2012). Informative estimation and selection of correlation structure for longitudinal data. *Journal of the American Statistics Association* **107**, 701–710. MR2980078