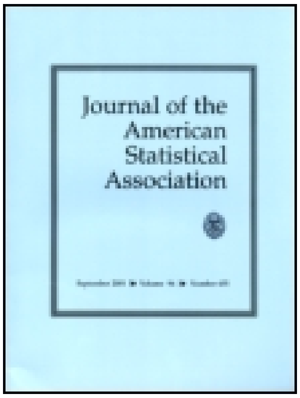

Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://amstat.tandfonline.com/loi/uasa20

Informative Estimation and Selection of Correlation Structure for Longitudinal Data

Jianhui Zhou [a] & Annie Qu [b]

[a] Department of Statistics, University of Virginia, Charlottesville, VA, 22904

[b] Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL, 61820

Accepted author version posted online: 14 May 2012. Version of record first published: 24 Jul 2012

To cite this article: Jianhui Zhou & Annie Qu (2012): Informative Estimation and Selection of Correlation Structure for Longitudinal Data, Journal of the American Statistical Association, 107:498, 701-710

To link to this article: http://dx.doi.org/10.1080/01621459.2012.682534

Informative Estimation and Selection of Correlation Structure for Longitudinal Data

Jianhui ZHOU and Annie QU

Identifying an informative correlation structure is important in improving estimation efficiency for longitudinal data. We approximate the empirical estimator of the correlation matrix by groups of known basis matrices that represent different correlation structures, and transform the correlation structure selection problem to a covariate selection problem. To address both the complexity and the informativeness of the correlation matrix, we minimize an objective function that consists of two parts: the difference between the empirical information and a model approximation of the correlation matrix, and a penalty that penalizes models with too many basis matrices. The unique feature of the proposed estimation and selection of correlation structure is that it does not require the specification of the likelihood function, and therefore it is applicable for discrete longitudinal data. We carry out the proposed method through a groupwise penalty strategy, which is able to identify more complex structures. The proposed method possesses the oracle property and selects the true correlation structure consistently. In addition, the estimator of the correlation parameters follows a normal distribution asymptotically. Simulation studies and a data example confirm that the proposed method works effectively in estimating and selecting the true structure in finite samples, and it enables improvement in estimation efficiency by selecting the true structures.

KEY WORDS: Correlation structure; Longitudinal data; Oracle property; Quadratic inference function.

1. INTRODUCTION

For longitudinal data it is essential to estimate and select an informative correlation structure since correctly modeling correlation structure will increase the efficiency of the regression parameter estimator, increase statistical power for hypothesis testing, and reduce the bias of the estimator in nonparametric modeling for longitudinal data (Wang 2003; Lin et al. 2004; Wang, Carroll, and Lin 2005). In addition, estimation of the correlation itself can provide additional information on the association among observations measured over time for longitudinal studies.

Although the empirical estimator of the correlation structure might be the closest to the true correlation, it is often not practical to use it directly since it involves high-dimensional correlation parameter estimation when the cluster size is large. In addition, the estimation of correlation parameters could be unstable if the sample size is relatively small compared with the cluster size. In our simulation provided in Section 5, it is rather surprising that the regression parameter estimator using unstructured correlation has a much lower efficiency than the estimator assuming independence structure, even when the cluster size is moderate.

Estimation and model selection of correlation structure remain a challenging problem since a higher order of moments is likely involved compared with model selection of covariates. Existing work mainly focuses on the estimation of the covariance matrix rather than on the selection of correlation structure, including the Cholesky decomposition approach (Huang et al. 2006; Huang, Liu, and Liu 2007), the factor modeling approach (Fan, Fan, and Lv 2008), and the spectrum random matrix approach (El Karoui 2008). These approaches are mainly suitable for continuous data. Other estimation approaches for high-dimensional covariance matrices include the nested least absolute shrinkage and selection operator (LASSO) approach (Levina, Rothman, and Zhu 2008) and thresholding approaches (Bickel and Levina 2008; Rothman, Levina, and Zhu 2009). These regularized covariance estimation approaches mainly focus on distinguishing nonzero components from zero components, but do not address the selection of correlation structure in general.

We propose an alternative strategy that approximates the empirical estimator of the correlation matrix by a linear combination of candidate basis matrices that contain either 0 or 1 as components. The linear combination of basis matrices can represent common correlation structures as well as mixtures of several correlation structures. We minimize the Euclidean norm of the difference between two estimating functions based on the empirical correlation information and the model-based approximation, in conjunction with a groupwise penalty on the basis matrices in the model approximation. Through the penalization, we can capture correlation information from longitudinal data sufficiently well, yet not be burdened by the high dimension of nuisance parameter estimation if it contains little information for the correlation structure.

The advantage of the proposed approach is that it allows the flexibility of modeling the correlation without requiring the estimation of each entry of the correlation matrix individually. Another advantage is that the specification of the likelihood function is not required, and therefore it is applicable for nonnormal responses that occur frequently for longitudinal data. More importantly, it is not restricted by large cluster size, since the number of basis matrices needed to represent a structured correlation matrix usually is not associated with the dimension of the correlation matrix. In addition, groupwise basis matrices model selection has the advantage of selecting and estimating a

Jianhui Zhou is Associate Professor, Department of Statistics, University of Virginia, Charlottesville, VA 22904 (E-mail: *jz9p@virginia.edu*). Annie Qu is Professor, Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820 (E-mail: *anniequ@illinois.edu*). Zhou's research was supported by the National Science Foundation (DMS-0906665). Qu's research was supported by the National Science Foundation (DMS-0906660). The authors are grateful to the reviewers, the associate editor, and the coeditors for their insightful comments and suggestions that have significantly improved the article.

© 2012 American Statistical Association
Journal of the American Statistical Association
June 2012, Vol. 107, No. 498, Theory and Methods
DOI: 10.1080/01621459.2012.682534

group of correlation parameters associated with the same correlation structure simultaneously, which enables one to select and estimate the true correlation structure sufficiently well. The proposed approach also ensures a positive definite correlation matrix asymptotically.

In theory, we show that the correlation structure can be selected consistently, assuming that the candidate basis matrices are from a sufficiently rich class to represent the true structure. Furthermore, we show that the proposed estimator of the correlation parameters possesses the oracle property (Fan and Li 2001) and follows an asymptotic normal distribution as if the true structure were known in advance.

The primary focus of this article is on the selection of correlation structure. Once the true correlation structure is selected, the efficiency of regression parameter estimation can be improved in the generalized linear model setting for longitudinal data using existing approaches such as the generalized estimating equation (GEE) (Liang and Zeger 1986) and the quadratic inference function (QIF) method by Qu, Lindsay, and Li (2000). The rest of the article is organized as follows. In Section 2, we describe the basis matrices representation of correlation structures. In Section 3, the proposed method for estimating and selecting correlation structure for longitudinal data is presented. Section 4 provides the asymptotic properties of the proposed estimator. In Section 5, we illustrate the performance of the proposed method through simulation studies with both Gaussian and binary responses. HIV data example is analyzed using the proposed method in Section 6. Section 7 provides a concluding discussion. Finally, the proofs are provided in the Appendix.

2. NOTATIONS AND MATRIX REPRESENTATION

For longitudinal data, the response variable y_{ij} and the p-dimensional covariate $\mathbf{x}_{ij}$ are measured at time t_{ij}, where the subject $i = 1, \ldots, n$ and the time points $j = 1, \ldots, m_i$. We first assume balanced data with $m_i = m$ for all i, and present the implementation of the proposed method for unbalanced data in Section 3.3.

Let $\mu_{ij} = E\{y_{ij}\} = \mu\{\mathbf{x}_{ij}^T \boldsymbol{\beta}\}$, where $\mu(\cdot)$ is a known inverse link function and $\boldsymbol{\beta}$ is a p-dimensional parameter vector. The quasi-likelihood equation (Wedderburn 1974) for estimating $\boldsymbol{\beta}$ is $\sum_{i=1}^n \dot{\boldsymbol{\mu}}_i^T \mathbf{V}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) = 0$, where $\mathbf{y}_i = (y_{i1}, \ldots, y_{im})^T$, $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{im})^T$, $\dot{\boldsymbol{\mu}}_i = \partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta}$, and $\mathbf{V}_i = \text{var}(\mathbf{y}_i)$. In practice, $\mathbf{V}_i$ is often unknown, and the empirical estimator of $\mathbf{V}_i$ based on the sample variance could be unstable, especially when the sample size is relatively small compared with a large number of variance components. Liang and Zeger (1986) introduced GEE to simplify $\mathbf{V}_i$ by assuming $\mathbf{V}_i = \mathbf{A}_i^{1/2} \mathbf{R} \mathbf{A}_i^{1/2}$, where $\mathbf{A}_i$ is the diagonal marginal variance matrix and $\mathbf{R}$ is a working correlation matrix.

The QIF method introduced by Qu, Lindsay, and Li (2000) assumes that $\mathbf{R}^{-1}$ can be approximated by a linear combination of several basis matrices, $\mathbf{I}_m$, $\mathbf{B}_1, \ldots, \mathbf{B}_J$, where $\mathbf{I}_m$ is the identity matrix and $\mathbf{B}_i$'s are symmetric matrices. The GEE can be approximated by a linear combination of elements in the estimating functions $\bar{\mathbf{g}}_n = n^{-1} \sum \mathbf{g}_i$, where $\mathbf{g}_i = (\dot{\boldsymbol{\mu}}_i^T \mathbf{A}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i), \dot{\boldsymbol{\mu}}_i^T \mathbf{A}_i^{-1/2} \mathbf{B}_1 \mathbf{A}_i^{-1/2}(\mathbf{y}_i - \boldsymbol{\mu}_i), \ldots, \dot{\boldsymbol{\mu}}_i^T \mathbf{A}_i^{-1/2} \mathbf{B}_J \mathbf{A}_i^{-1/2}(\mathbf{y}_i - \boldsymbol{\mu}_i))^T$. However, the dimension of $\bar{\mathbf{g}}_n$, $(J+1)p$, is greater than the number of unknown parameters p, and therefore $\boldsymbol{\beta}$ cannot be estimated by equating $\bar{\mathbf{g}}_n$ exactly to zero. Instead, the QIF method minimizes the quadratic distance function (Hansen 1982): $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \bar{\mathbf{g}}_n^T \boldsymbol{\Omega}^{-1} \bar{\mathbf{g}}_n$, where $\boldsymbol{\Omega} = \text{var}(\mathbf{g}_i)$ can be estimated by $\bar{\mathbf{W}}_n = n^{-1} \sum_{i=1}^n \mathbf{g}_i \mathbf{g}_i^T$. The quadratic function $Q_n(\boldsymbol{\beta}) = n \bar{\mathbf{g}}_n^T \bar{\mathbf{W}}_n^{-1} \bar{\mathbf{g}}_n$ is called the QIF since it provides an inference function for the regression parameters. The QIF method does not estimate the basis matrices coefficients, but can still improve the efficiency of regression parameter estimation.

In contrast to Qu, Lindsay, and Li's (2000) focus on the estimation of regression parameters, the primary interest here is to identify the correct correlation structure. In our method, the inverse of the correlation matrix is linearly represented by groups of basis matrices, where each group $\mathbf{M}_j$ represents a certain correlation structure. That is,

$$\mathbf{R}_{m\times m}^{-1} \approx \mathbf{M}_1 \boldsymbol{\alpha}_1 + \mathbf{M}_2 \boldsymbol{\alpha}_2 + \cdots + \mathbf{M}_J \boldsymbol{\alpha}_J, \qquad (1)$$

where m is the cluster size, $\mathbf{M}_j = \{\mathbf{M}_{j,1}, \mathbf{M}_{j,2}, \ldots, \mathbf{M}_{j,d_j}\}$ consists of a group of basis matrices, $\boldsymbol{\alpha}_j = (\alpha_{j,1}, \alpha_{j,2}, \ldots, \alpha_{j,d_j})^T$ are the corresponding coefficients, and each group has $\mathbf{M}_j \boldsymbol{\alpha}_j = \sum_{k=1}^{d_j} \alpha_{j,k} \mathbf{M}_{j,k}$. The basis matrices in $\mathbf{M}_j$ can be specified according to the candidate structures from prior information, such as exchangeable, first-order autoregressive (AR(1)), or blockwise structures. The illustration of these basis matrices will be provided in Section 5.

In the cases where the prior information for correlation structure is unknown, we can use a linear representation of a simple and complete set of basis matrices that contains 1 for (i, j) and (j, i) entries and 0 elsewhere. Therefore, any correlation matrix can be represented by a linear combination of the complete set of basis matrices. Alternatively, the basis matrices can be selected from the spectral decomposition of the empirical correlation matrix. These types of basis specification do not require prior information for the correlation matrix. However, their main disadvantage is that they do not provide much information about the correlation structure, but rather serve the purpose of estimation of the correlation matrix. Therefore, the matrix representation by these types of basis matrices is not of our particular interest in this article.

3. ESTIMATION AND SELECTION OF CORRELATION STRUCTURE

We propose to estimate and select the correlation structure for longitudinal data by approximating the empirical correlation estimate with prespecified candidate basis matrices. The correlation structure is identified through selecting the groups of nonzero coefficients associated with the basis matrices.

We transform the problem of selecting correlation structure to the problem of identifying nonzero coefficients $\boldsymbol{\alpha}_j$ through the representation of $\mathbf{R}^{-1}$ in (1). If a group of candidate basis matrices in $\mathbf{M}_j$ represents the true correlation structure sufficiently well, the associated coefficients $\boldsymbol{\alpha}_j$ will be nonzero and can be identified through model selection. Therefore, the correlation structure can be selected correctly by selecting the corresponding group(s) of basis matrices. This is in contrast to Qu, Lindsay, and Li (2000), where the coefficients of the basis matrices are considered as nuisance parameters and are not estimated. Here, the estimation of the coefficients of the basis matrices is essential for the purpose of correlation structure estimation and selection.

3.1 Selection of Basis Matrices Groups

The selection of the correlation structure $\mathbf{R}$ is performed through identifying nonzero vectors of $\boldsymbol{\alpha}_j$. To estimate $\boldsymbol{\alpha}_j$, we minimize the discrepancy between the estimating function using the empirical correlation matrix estimate $\tilde{\mathbf{R}}^{-1}$ and the estimating function using the basis matrices representation of $\mathbf{R}^{-1}$. The discrepancy between the two estimating functions for the ith cluster is measured by

$$\mathbf{S}_i = \dot{\boldsymbol{\mu}}_i^T(\tilde{\boldsymbol{\beta}})\mathbf{A}_i^{-\frac{1}{2}}\{\tilde{\mathbf{R}}^{-1} - \mathbf{M}_1\boldsymbol{\alpha}_1 - \cdots - \mathbf{M}_J\boldsymbol{\alpha}_J\}\mathbf{A}_i^{-\frac{1}{2}}(\mathbf{y}_i - \boldsymbol{\mu}_i(\tilde{\boldsymbol{\beta}})), \tag{2}$$

where $\tilde{\boldsymbol{\beta}}$ is the estimator of $\boldsymbol{\beta}$ via the GEE with independence structure $\mathbf{R}$, and $\tilde{\mathbf{R}}$ is the sample correlation estimator based on $\tilde{\boldsymbol{\beta}}$.

The Euclidean norm of $\mathbf{S} = (\mathbf{S}_1^T, \ldots, \mathbf{S}_n^T)^T$ should be sufficiently small if $\mathbf{R}^{-1}$ is approximated sufficiently well by the selected groups of basis matrices. However, it is important to balance model sufficiency and model complexity. We can always include more basis matrices to achieve an exact $\mathbf{R}^{-1}$, but that may likely lead to overfitting the model for correlation structure when the cluster size increases. Therefore, we propose to select the structure by minimizing the Euclidean norm of $\mathbf{S}$ and also by penalizing models involving too many basis matrices.

The coefficient vectors $\boldsymbol{\alpha}_j$ are estimated by minimizing the objective function

$$\sum_{i=1}^{n} \mathbf{S}_i^T\mathbf{S}_i + np\sum_{j=2}^{J} p_\lambda(||\boldsymbol{\alpha}_j||_1), \tag{3}$$

where $p_\lambda(\cdot)$ is the smoothly clipped absolute deviation (SCAD) penalty function (Fan and Li 2001), λ is a tuning parameter, $||\cdot||_1$ denotes the L_1 norm, and the positive value involved in $p_\lambda(\cdot)$ is typically chosen as 3.7. The selection of the tuning parameter λ will be discussed in more detail in Section 3.2. Note that the coefficient $\boldsymbol{\alpha}_1$ associated with the identity matrix $\mathbf{M}_{1,1}$ is not penalized in (3), since the identity matrix should always be included as a basis matrix for any correlation structure. For this reason, the first group basis matrix is $\mathbf{M}_1 = \{\mathbf{M}_{1,1}\}$ throughout this article. To minimize (3), we define

$$\begin{aligned}
\mathbf{U}_i &= \dot{\boldsymbol{\mu}}_i^T(\tilde{\boldsymbol{\beta}})\mathbf{A}_i^{-\frac{1}{2}}\tilde{\mathbf{R}}^{-1}\mathbf{A}_i^{-\frac{1}{2}}(\mathbf{y}_i - \boldsymbol{\mu}_i(\tilde{\boldsymbol{\beta}})), && i = 1, \ldots, n,\\
\mathbf{V}_{ij,k} &= \dot{\boldsymbol{\mu}}_i^T(\tilde{\boldsymbol{\beta}})\mathbf{A}_i^{-\frac{1}{2}}\mathbf{M}_{j,k}\mathbf{A}_i^{-\frac{1}{2}}(\mathbf{y}_i - \boldsymbol{\mu}_i(\tilde{\boldsymbol{\beta}})), && j = 1, \ldots, J,\\
& && k = 1, \ldots, d_j,
\end{aligned}$$

and $\mathbf{V}_{ij} = (\mathbf{V}_{ij,1}, \ldots, \mathbf{V}_{ij,d_j})$, with each matrix in $\mathbf{M}_j = \{\mathbf{M}_{j,1}, \mathbf{M}_{j,2}, \ldots, \mathbf{M}_{j,d_j}\}$ corresponding to a column of $\mathbf{V}_{ij}$. Applying the one-step local linear approximation to the SCAD function (Zou and Li 2008), we can achieve an approximate solution to (3) by minimizing

$$\sum_{i=1}^{n}\left|\left|\mathbf{U}_i - \sum_{j=1}^{J}\mathbf{V}_{ij}\boldsymbol{\alpha}_j\right|\right|_2^2 + np\sum_{j=2}^{J} p'_\lambda\left(\left|\left|\hat{\boldsymbol{\alpha}}_j^{(0)}\right|\right|_1\right)||\boldsymbol{\alpha}_j||_1, \tag{4}$$

where $\hat{\boldsymbol{\alpha}}_j^{(0)}$ is an initial estimate of $\boldsymbol{\alpha}_j$, and can be obtained by the least squares estimator.

Here, $\boldsymbol{\alpha}_j$'s are estimated in the transformed regression problem of $\mathbf{U}_i$ on $\mathbf{V}_{ij}$. To obtain the sparse estimators of $\boldsymbol{\alpha}_j$'s for correlation structure selection, we adopt the SCAD penalty in our method. Due to the use of L_1 norm and local linear approximation, coefficients in the same group are penalized using the same weight, and, most importantly, the objective function (4) is indeed an adaptive LASSO objective function (Zou 2006). Therefore, the minimizers of the objective function (4) can be obtained by the efficient least angle regression (LARS) algorithm of Efron et al. (2004).

3.2 Tuning Parameter Selection

Tuning parameter selection is critical to achieving better model selection performance in finite samples. Wang, Li, and Tsai (2007) proposed the Bayesian information criterion (BIC) to select the tuning parameter of the SCAD penalty in penalized least squares for consistent model selection, and Zhang, Li, and Tsai (2010) proposed the generalized information criterion (GIC-type) for tuning parameter selection in a nonconcave penalized likelihood approach and studied its consistency and asymptotic loss efficiency. We propose a GIC-type criterion to select λ in the SCAD penalty function in our framework:

$$\mathrm{GIC}(\lambda) = nr\log\frac{\eta_{\max}(\hat{\mathbf{R}}^{-1}\tilde{\mathbf{R}}^2\hat{\mathbf{R}}^{-1})}{\eta_{\min}(\hat{\mathbf{R}}^{-1}\tilde{\mathbf{R}}^2\hat{\mathbf{R}}^{-1})} + \log(n)k(\lambda), \tag{5}$$

where $\tilde{\mathbf{R}}$ is the empirical estimator as in (2) and $\hat{\mathbf{R}}^{-1} = \mathbf{M}_1\hat{\boldsymbol{\alpha}}_1 + \cdots + \mathbf{M}_J\hat{\boldsymbol{\alpha}}_J$. Here, the vectors $\hat{\boldsymbol{\alpha}}_j$ are estimated using the tuning parameter value λ, $\eta_{\max}(\cdot)$ and $\eta_{\min}(\cdot)$ denote the largest and smallest eigenvalues of the matrix $\hat{\mathbf{R}}^{-1}\tilde{\mathbf{R}}^2\hat{\mathbf{R}}^{-1}$, $k(\lambda)$ is the number of nonzero estimates in $\hat{\alpha}_{j,k}$, and $r > 0$ controls the sensitivity of the criterion to the discrepancy between the empirical and the selected correlation structures, where a larger r leads to a smaller λ and the procedure tends to overselect. We can choose an optimal r through cross-validation to achieve a high percentage of overall correct fit. Our unreported numerical studies show that the performance of the proposed procedure is quite robust against the value of r in the range from 0.1 to 0.5. In this article, we use $r = 0.25$ in Sections 5 and 6.

The rationale behind the proposed GIC-type criterion is the fact that if the selected groups of basis matrices capture most of the information of $\mathbf{R}^{-1}$, the largest and smallest eigenvalues of $\hat{\mathbf{R}}^{-1}\tilde{\mathbf{R}}^2\hat{\mathbf{R}}^{-1}$ should be very close to 1 since $\hat{\mathbf{R}}^{-1}\tilde{\mathbf{R}}$ is close to the identity matrix. Therefore, the tuning parameter is selected to minimize GIC(λ). We also explore generalized cross-validation (GCV), the Akaike's information criterion (AIC), BIC, and residual information criterion (RIC) to select λ by replacing the likelihood function with the QIF as in Wang and Qu (2009). Our simulation results, not provided here, indicate that the GCV and the AIC have similar performance; while the BIC's and RIC's performances are quite similar, and both are slightly better than the GCV and the AIC. However, none of them perform as well as the proposed GIC criterion. This is probably due to the additional parameter r in (5), which can adjust the size of λ to incorporate the ratio of noise versus signal levels from the data.

3.3 Implementation With Unbalanced Data Due to Missingness

The above method is presented with balanced data, that is, $m_i = m$. In practice, longitudinal data may not be measured with the same cluster size, and could be unbalanced due to missingness or experimental constraints. To configure the proposed method for unbalanced data, we apply the transformation matrix to each cluster. We create the largest cluster with a size m, which contains time points for all possible measurements, and assume that fully observed clusters contain m observations. We define the $m \times m_i$ transformation matrix $\mathbf{T}_i$ for the ith cluster by removing the columns of the identity matrix, where the removed columns correspond to the missing observations.

We define $\mathbf{y}_i^* = \mathbf{T}_i\mathbf{y}_i$, $\boldsymbol{\mu}_i^*(\tilde{\boldsymbol{\beta}}) = \mathbf{T}_i\boldsymbol{\mu}_i(\tilde{\boldsymbol{\beta}})$, $\dot{\boldsymbol{\mu}}_i^*(\tilde{\boldsymbol{\beta}}) = \mathbf{T}_i\dot{\boldsymbol{\mu}}_i(\tilde{\boldsymbol{\beta}})$, and $\mathbf{A}_i^* = \mathbf{T}_i\mathbf{A}_i\mathbf{T}_i^T$, where components in $\mathbf{y}_i^*$ are the same as in $\mathbf{y}_i$ for nonmissing responses but are 0 for the missing responses, and similarly for $\boldsymbol{\mu}_i^*$ and $\dot{\boldsymbol{\mu}}_i^*$. Note that the pseudo-marginal variance values in $\mathbf{A}_i^*$ for the missing observations have no effects on the values of $\mathbf{U}_i^*$ and $\mathbf{V}_{ij,k}^*$ since the 0 values specified in $\dot{\boldsymbol{\mu}}_i^*$ and $\mathbf{y}_i^* - \boldsymbol{\mu}_i^*(\tilde{\boldsymbol{\beta}})$ corresponding to the missing observations ensure that the missing observations do not contribute to the objective function. Therefore, we can specify the variance to be 0 in $\mathbf{A}_i^*$ for the missing observations just for convenience.

For the empirical estimate of $\tilde{\mathbf{R}}^*$, we use the sample correlation matrix estimated from fully observed clusters if the number of fully observed clusters is sufficiently large. Otherwise, the empirical estimator could be obtained based on the method by Qu et al. (2010). Numerical studies in Section 5 show that the proposed method is quite effective even if only 30% of the clusters are fully observed and 50% of the observations in the other 70% of the clusters are missing.

We replace $\mathbf{U}_i$ and $\mathbf{V}_{ij,k}$ with $\mathbf{U}_i^*$ and $\mathbf{V}_{ij,k}^*$ to formulate the objective function (4), where $\mathbf{U}_i^*$ and $\mathbf{V}_{ij,k}^*$ for each cluster of the unbalanced data are computed using the same formulas as for $\mathbf{U}_i$ and $\mathbf{V}_{ij,k}$ in Section 3.1, but based on $\mathbf{y}_i^*$, $\boldsymbol{\mu}_i^*(\tilde{\boldsymbol{\beta}})$, $\dot{\boldsymbol{\mu}}_i^*(\tilde{\boldsymbol{\beta}})$, $\mathbf{A}_i^*$, and $\tilde{\mathbf{R}}^*$ instead. Similar to the balanced data case, the groups of basis matrices can be selected through identifying nonzero coefficients $\boldsymbol{\alpha}_j$ in the objection function with $\mathbf{U}_i^*$ and $\mathbf{V}_{ij,k}^*$.

4. ASYMPTOTIC PROPERTY

The proposed regularization approach achieves the sparse estimator. That is, if the true parameter $\boldsymbol{\alpha}_j$ is 0, it is estimated to be exactly 0, with probability tending to 1 as n increases. In addition, the SCAD penalty employed in (4) performs better than the adaptive group LASSO in general, since the formulation of SCAD allows almost no penalty if the true parameter is far from 0; however, the adaptive group LASSO penalizes all parameters. Moreover, the sparsity of the estimator is achieved groupwise, which enables us to identify a specific correlation structure correctly through identifying groups of nonzero coefficients.

Moreover, we allow the basis matrices to be misspecified up to a small amount in the asymptotic study. Specifically, the parameter vector corresponding to the specified basis matrices, $\boldsymbol{\alpha}^0 = (\boldsymbol{\alpha}_1^{0T}, \ldots, \boldsymbol{\alpha}_J^{0T})^T$, is partitioned into $\boldsymbol{\alpha}^0 = (\boldsymbol{\alpha}_{\mathrm{I}}^{0T}, \boldsymbol{\alpha}_{\mathrm{II}}^{0T})^T$, where $\boldsymbol{\alpha}_{\mathrm{I}}^0 = \mathbf{0}$ corresponds to the coefficients of the basis matrices that are irrelevant to the structure of $\mathbf{R}$ and $\boldsymbol{\alpha}_{\mathrm{II}}^0 \neq \mathbf{0}$ consists of the coefficients of the basis matrices related to the correlation structure of $\mathbf{R}$. In addition, we assume that there is a misspecified part with $\boldsymbol{\alpha}_{\mathrm{III}}^0 \neq \mathbf{0}$ consisting of the coefficients of basis matrices related to the correlation structure of $\mathbf{R}$, but not specified in the basis matrix representation (1). Let $\hat{\boldsymbol{\alpha}} = (\hat{\boldsymbol{\alpha}}_{\mathrm{I}}^T, \hat{\boldsymbol{\alpha}}_{\mathrm{II}}^T)^T$ be the estimator of $\boldsymbol{\alpha}^0$ by minimizing (4) using the basis matrices corresponding to $\boldsymbol{\alpha}_{\mathrm{I}}^{0T}$ and $\boldsymbol{\alpha}_{\mathrm{II}}^{0T}$ only. Denote the tuning parameter in the SCAD function by λ_n here. The subscript n is imposed since the tuning parameter depends on the number of clusters n.

The following conditions are assumed in order to achieve the asymptotic properties:

C_1: $E(||\mathbf{y}_i||_2^4) < \infty$;
C_2: $\lambda_n \to 0$ and $n^{\frac{1}{2}}\lambda_n \to \infty$.

Theorem 1. Let $\mathbf{V}_i = (\mathbf{V}_{i1}, \ldots, \mathbf{V}_{iJ})$ be the matrix of covariates in (4) for the ith cluster. Assume that the correlation matrix is possibly misspecified at the rate of $\boldsymbol{\alpha}_{\mathrm{III}}^0 = o_p(n^{-\frac{1}{2}})$. Given that condition C_1 is satisfied, we have

(a) $n^{-1}\sum_{i=1}^n \mathbf{V}_i^T\mathbf{V}_i \xrightarrow{P} \boldsymbol{\Sigma}$, for some covariance matrix $\boldsymbol{\Sigma}$.
(b) $n^{-\frac{1}{2}}\sum_{i=1}^n (\mathbf{U}_i - \mathbf{V}_i\boldsymbol{\alpha}^0)^T\mathbf{V}_i \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Sigma}^*)$, where $\boldsymbol{\Sigma}^*$ is provided in the Appendix.

Following Theorem 1, we show in Theorem 2 that, given the rate of misspecification, the oracle property holds for our approach; that is, the correct structure for $\mathbf{R}$ is selected consistently and the estimator of the correlation parameters involved in $\mathbf{R}^{-1}$ has the same asymptotic distribution as if the true structure is known in advance.

Theorem 2. Assume that the correlation matrix is possibly misspecified at the rate of $\boldsymbol{\alpha}_{\mathrm{III}}^0 = o_p(n^{-\frac{1}{2}})$. Given the conditions C_1 and C_2, we provide the sparsity property of $\hat{\boldsymbol{\alpha}}_{\mathrm{I}}$ and the asymptotic distribution of $\hat{\boldsymbol{\alpha}}_{\mathrm{II}}$ as follows:

(a) The structure of $\mathbf{R}$ can be identified correctly with probability tending to 1, that is, $\hat{\boldsymbol{\alpha}}_{\mathrm{I}} = \mathbf{0}$.
(b) The estimator of the nonzero coefficients associated with $\mathbf{R}^{-1}$ is asymptotically normal, that is, $\sqrt{n}(\hat{\boldsymbol{\alpha}}_{\mathrm{II}} - \boldsymbol{\alpha}_{\mathrm{II}}^0) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Theta})$, where the asymptotic covariance matrix $\boldsymbol{\Theta}$ is provided in the Appendix.

The proofs of Theorems 1 and 2 are given in the Appendix. By Theorem 2, the basis matrices associated with the true structure can be selected consistently. Consequently, the regression parameter estimator by the QIF method using the selected basis matrices is efficient within the class of mean zero moment conditions (Hansen 1982). The efficiency gain in regression parameter estimation is also confirmed in the following simulation studies and data example.

5. SIMULATION STUDIES

We provide simulation studies to illustrate the performance of the proposed estimation and selection method for correlation structure in finite samples, and compare the efficiency of regression parameter estimation under different correlation structures for both balanced and unbalanced data, and also under misspecification of basis matrices. Here, we allow the response variables to be both normal and binary.

5.1 Study 1: Binary Responses

We evaluate the correlation structure selection for binary responses. The datasets are generated from the model $y_{ij} \sim \text{Binomial}(1, \mu_{ij})$ for $i = 1, \ldots, n$ and $j = 1, \ldots, 9$, where $\text{logit}(\mu_{ij}) = 0.5 + \mathbf{x}_{ij}^T \boldsymbol{\beta}$ and $\mathbf{x}_{ij} \in \mathbb{R}^3$ from $N(\mathbf{0}, \mathbf{I}_3)$. This is balanced data with cluster size 9. We let $\boldsymbol{\beta} = (0.01, 0.01, 0.01)^T$ such that μ_{ij} are all close, to facilitate the generation of correlated binary responses. The correlated binary responses are generated using the R package "mvtBinaryEP" with three different correlation structures of $\mathbf{R} = (r_{ij})$: (1) $\mathbf{R} = \mathbf{R}_1$ is AR(1) with $r_{ij} = 0.7^{|i-j|}$, (2) $\mathbf{R} = \mathbf{R}_2$ is exchangeable with $r_{ij} = 0.6$ for $i \neq j$, and (3) $\mathbf{R} = \mathbf{R}_3$ is blockwise exchangeable with block sizes 4 and 5 and correlation parameters $\rho_1 = 0.8$ and $\rho_2 = 0.7$ for each block. Observations in different blocks are independent.

The unbalanced datasets are created from the balanced ones by keeping 30% of the clusters with the fully observed nine measurements, and for the remaining 70% clusters, each observation has a probability of 0.5 to be missing. Given the initial estimator $\tilde{\boldsymbol{\beta}}$ assuming independence structure, the empirical estimator $\tilde{\mathbf{R}}^*$ is computed based on the 30% fully observed clusters.

We specify the basis matrix groups $\mathbf{M}_1 = \{\mathbf{M}_{1,1}\}$, $\mathbf{M}_2 = \{\mathbf{M}_{2,1}, \mathbf{M}_{2,2}\}$, $\mathbf{M}_3 = \{\mathbf{M}_{3,1}\}$, and $\mathbf{M}_4 = \{\mathbf{M}_{4,1}, \mathbf{M}_{4,2}, \mathbf{M}_{4,3}\}$, and the corresponding coefficients are $\alpha_{i,j}$ in vectors $\boldsymbol{\alpha}_i$ for $i = 1, \ldots, 4$ accordingly. Here, $\mathbf{M}_{1,1}$ is the identity matrix, $\mathbf{M}_2$ contains the other two basis matrices for AR(1) with one matrix of 1 on the subdiagonal and 0 elsewhere and another matrix with 1 on two corner components of the diagonal, $\mathbf{M}_3$ contains the other basis matrix for exchangeable structure with 1 on the off-diagonal and 0 elsewhere (Qu, Lindsay, and Li 2000), and $\mathbf{M}_4$ contains three block-diagonal basis matrices to represent $\mathbf{R}_3$ besides $\mathbf{M}_{1,1}$. The correlation structure can be selected using the sparsity property of the estimator of $\boldsymbol{\alpha}_i$ proposed in Section 3.1. For example, the AR(1) structure $\mathbf{R}_1$ will be selected if the coefficients $\hat{\boldsymbol{\alpha}}_1$ and $\hat{\boldsymbol{\alpha}}_2$ associated with groups 1 and 2 basis matrices are nonzero, and $\hat{\boldsymbol{\alpha}}_3$ and $\hat{\boldsymbol{\alpha}}_4$ associated with groups 3 and 4 are zero.

We calculate the percentages of the basis matrices being correctly selected (C), underselected (U), and overselected (O) out of 100 generated datasets for each correlation structure. Table 1 summarizes the results with sample sizes n =100 and 300, and shows that the proposed method selects the correct correlation structure effectively for balanced datasets. Since almost half of the observations are missing in the unbalanced datasets, the proposed method performs less effectively for the unbalanced data with sample size $n = 100$, but shows a consistent model selection trend when the sample size increases, such as when $n = 300$.

In addition, the percentages of positive definite correlation matrix $\hat{\mathbf{R}}$ in (5) obtained by minimizing (4) are also reported in Table 1, which shows that more than 98% of the time the estimated correlation matrix is positive definite with various sample sizes and correlation structures. We also study the sensitivity of the basis matrices selection to the initial value $\tilde{\boldsymbol{\beta}}$. In the unreported study, we replace each component of $\tilde{\boldsymbol{\beta}}$ by a randomly sampled value from its corresponding 95% confidence interval. We observe that the proposed approach achieves very similar results as in Table 1.

Table 1. Study 1: Selection of the correlation structures and positive definiteness of the estimated correlation matrix with binary responses. The columns C, U, and O are the percentages of correct selection, underselection, and overselection of the basis matrices out of 100 simulation datasets. The column P is the percentage of the estimated correlation matrix being positive definite

		Equal cluster size				Unequal cluster size			
R	n	C	U	O	P	C	U	O	P
$\mathbf{R}_1$	100	0.99	0.00	0.01	1.00	0.30	0.60	0.10	0.98
	300	0.97	0.00	0.03	1.00	0.89	0.00	0.11	1.00
$\mathbf{R}_2$	100	0.98	0.00	0.02	1.00	0.42	0.26	0.32	0.98
	300	0.97	0.00	0.03	1.00	0.80	0.00	0.20	1.00
$\mathbf{R}_3$	100	0.53	0.46	0.01	1.00	0.20	0.79	0.01	1.00
	200	0.87	0.08	0.05	1.00	0.70	0.23	0.07	1.00

5.2 Study 2: Normal Responses With Many Basis Matrices

To investigate the performance of the proposed method with a larger cluster size and a larger number of basis matrices, we increase the cluster size to 25 and the number of basis matrices to 20. The normal responses are generated from the model $y_{ij} = 2 + \mathbf{x}_{ij}^T \boldsymbol{\beta} + \epsilon_{ij}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, 25$, where $\boldsymbol{\beta} = (1, 1, 1)^T$, $\mathbf{x}_{ij} \in \mathbb{R}^3$ from $N(\mathbf{0}, \mathbf{I}_3)$, and $\boldsymbol{\epsilon}_i = (\epsilon_{i,1}, \ldots, \epsilon_{i,25})^T$ from $N(\mathbf{0}, \mathbf{R}_4)$. The correlation structure $\mathbf{R}_4$ is block diagonal with total five blocks and each block size is 5, where the first block is AR(1) with $\rho_1 = 0.7$, the third block is exchangeable with $\rho_2 = 0.8$, and the other blocks have independence structure. Unbalanced data are created using the same missing data scheme as described in Study 1, with 50% of the clusters having missing observations and the missing probability for each observation being 0.4. Since each block is either independence, AR(1), or exchangeable, the overall possible combinations of three different structures for five blocks is $3^5 = 243$, which might not be practical to try one at a time with the BIC or AIC criterion. In contrast, the proposed approach is able to select the blockwise structure through identifying nonzero coefficients in a single model, and therefore is much more efficient than traditional model selection approaches.

We choose 20 basis matrices from the block-diagonal matrices with block size 5, and divide them into 11 groups. Group $\mathbf{G}_1$ contains the identity matrix $\mathbf{I}_{25}$ and four matrices with $\mathbf{I}_5$ on the first, second, third, and fourth blocks, respectively, and 0 matrix on the other blocks. Group $\mathbf{G}_2$ contains two matrices with $\mathbf{M}_{2,1}$ and $\mathbf{M}_{2,2}$ as in Study 1 for the first block and 0 matrix for the other blocks, which corresponds to the AR(1) structure for block 1. Group $\mathbf{G}_3$ contains one matrix with $\mathbf{M}_{3,1}$ for the first block and 0 matrix for the other blocks, which corresponds to the exchangeable structure for block 1. The other groups are defined similarly using the same $\mathbf{M}_{2,1}$, $\mathbf{M}_{2,2}$, and $\mathbf{M}_{3,1}$, but on different block locations. In summary, $\mathbf{G}_1$ represents the independence structures; $\mathbf{G}_2$ and $\mathbf{G}_3$ represent the AR(1) and exchangeable structures for block 1; similarly, $\mathbf{G}_4$ and $\mathbf{G}_5$ for block 2; $\mathbf{G}_6$ and $\mathbf{G}_7$ for block 3; $\mathbf{G}_8$ and $\mathbf{G}_9$ for block 4; and $\mathbf{G}_{10}$ and $\mathbf{G}_{11}$ for block 5.

To identify the correlation structure $\mathbf{R}_4$ correctly, the group parameter estimates should be $\hat{\boldsymbol{\alpha}}_i = \mathbf{0}$ for $i \neq 1, 2, 7$ and $\hat{\boldsymbol{\alpha}}_i \neq \mathbf{0}$

Table 2. Study 2: Selection of correlation structures with a large number of basis matrices with normal response. The frequency of $\hat{\boldsymbol{\alpha}}_i = \mathbf{0}$ out of 100 simulation datasets. Structure $\mathbf{R}_4$ corresponds to vectors $\boldsymbol{\alpha}_1 \neq \mathbf{0}$, $\boldsymbol{\alpha}_2 \neq \mathbf{0}$, and $\boldsymbol{\alpha}_7 \neq \mathbf{0}$. The column P is the percentage of the estimated correlation matrix being positive definite

			i											
R	Balance	n	1	2	3	4	5	6	7	8	9	10	11	P
$\mathbf{R}_4$	Equal	100	0	0	99	100	100	100	0	100	100	100	99	1.00
		200	0	0	100	100	100	100	0	100	100	100	100	1.00
	Unequal	100	0	9	98	100	98	91	5	98	99	99	100	0.97
		200	0	0	100	99	100	100	0	100	100	100	100	1.00

for $i = 1, 2, 7$. The frequencies of $\hat{\boldsymbol{\alpha}}_i = \mathbf{0}$ $(i = 1, \ldots, 11)$ and the percentages of the estimated correlation matrix being positive definite are summarized in Table 2 out of 100 datasets with sample sizes $n = 100$ and 200. The true correlation parameters $\boldsymbol{\alpha}_i$, $i = 1, 2, 7$, and the mean and standard deviation of $\hat{\boldsymbol{\alpha}}_i$ are given in Table 3. Tables 2 and 3 show that the proposed method performs well in correlation structure selection and parameter estimation for both balanced and unbalanced data, and the performance in model selection consistency, bias, and standard deviation of the estimator improves as the sample size increases, especially for unbalanced data.

5.3 Study 3: Efficiency Improvement in Regression Parameter Estimation

We evaluate the efficiency improvement for regression parameter estimation using the correlation structure selected by the proposed method. We generate 100 datasets of equal cluster size 5 with the relatively simple structure $\mathbf{R}_1$ in Study 1, and of cluster size 25 with the complex structure $\mathbf{R}_4$ in Study 2. Both are generated from the normal response model in Study 2, and the basis matrices in Study 1 and Study 2 are used for $\mathbf{R}_1$ and $\mathbf{R}_4$, respectively. The efficiencies of the three regression parameter estimators are compared for the GEE estimators with the independence and unstructured working correlation structures, and the QIF estimator using groups of basis matrices selected by the proposed method. The relative efficiency, defined as the ratio of the variances of the regression parameter estimators between the independence and the selected structures, and between the unstructured and the selected structures, is summarized in Table 4 with various sample sizes. Table 4 indicates that with the selected structure, the estimation efficiency of the regression parameter is significantly improved, and this holds especially for small and moderate sample sizes. Surprisingly, the efficiency of the unstructured GEE estimator is the worst of all, with the variance of the estimator of β_3 1500 times more than that using the selected structure for $\mathbf{R}_1$ when $n = 50$. Even when n increases to 200, the unstructured estimator performs notably worse than the estimators based on the selected correlation structures, and even worse than the independence structure.

5.4 Study 4: Efficiency Improvement and Correlation Estimation Bias Under Basis Matrices Misspecification

We conduct this study to show the efficiency improvement for estimating $\boldsymbol{\beta}$ and the bias for estimating the correlation matrix when the inverse of the correlation matrix is not exactly a linear combination of the basis matrices. We generate data using the model in Study 2 with $m = 9$ and a new correlation structure $\mathbf{R}_5$, the hybrid structure of AR(1) and exchangeable with parameters $\rho_1 = 0.8$ and $\rho_2 = 0.9$, respectively. For this hybrid structure, each component of $\mathbf{R}$ is $r_{ij} = (\rho_1^{|i-j|} + \rho_2)/2$, $i \neq j$, if the two random processes with AR(1) and exchangeable correlations are independent and have the same variance. In the normal response case, it is equivalent to saying that the random error can be decomposed into two sources, $\epsilon_{ij} = \epsilon'_{ij} + \epsilon''_{ij}$, where ϵ'_{ij} and ϵ''_{ij} are independent of each other with the same variance, ϵ'_{ij} has the AR(1) structure with the correlation parameter ρ_1, and ϵ''_{ij} has the exchangeable structure with the correlation parameter ρ_2.

Since the inverse of a hybrid correlation structure does not have a specific structure, more than 20 basis matrices

Table 3. Study 2: Estimation of correlation parameters using a large number of basis matrices with normal response. The mean and standard deviation of the estimators of the nonzero correlation parameters in $\mathbf{R}_4$ from 100 simulation datasets

			$\boldsymbol{\alpha}_1$			$\boldsymbol{\alpha}_2$		$\boldsymbol{\alpha}_7$
R	Balance	n	$\alpha_{1,1}$	$\alpha_{1,2}$	$\alpha_{1,4}$	$\alpha_{2,1}$	$\alpha_{2,2}$	$\alpha_{7,1}$
$\mathbf{R}_4$		True	1.000	1.922	3.048	−0.961	−1.373	−0.952
	Equal	100	1.011 (0.093)	1.927 (0.477)	3.207 (0.661)	−0.945 (0.300)	−1.384 (0.247)	−0.995 (0.166)
		200	1.006 (0.052)	1.928 (0.290)	3.050 (0.446)	−0.951 (0.195)	−1.385 (0.153)	−0.955 (0.112)
	Unequal	100	1.425 (0.266)	3.212 (1.675)	5.316 (1.920)	−1.545 (0.954)	−2.156 (0.968)	−1.484 (0.524)
		200	1.130 (0.098)	2.359 (0.422)	3.809 (0.721)	−1.149 (0.297)	−1.650 (0.237)	−1.161 (0.183)

Table 4. Study 3: Efficiency study with balanced and normal response. Ratio of variances of regression parameter estimators between independence and selected, and unstructured and selected correlation structures

R	n	Ratio of variance	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
$\mathbf{R}_1$	50	Independence/selected	0.965	2.161	2.614	2.429
		Unstructured/selected	3.074	131.581	660.804	1528.858
	100	Independence/selected	0.918	2.257	2.526	1.978
		Unstructured/selected	14.801	3.274	2.603	1.887
	200	Independence/selected	0.901	3.977	2.503	2.157
		Unstructured/selected	0.980	1.220	1.316	1.706
$\mathbf{R}_4$	100	Independence/selected	1.183	1.204	0.986	1.690
		Unstructured/selected	153.778	20.948	634.152	129.319
	200	Independence/selected	1.238	1.583	1.676	1.893
		Unstructured/selected	2.067	1.581	4.130	3.357

are needed to fully represent the inverse of the correlation matrix. In this study, we select seven of them, including the three matrices in $\mathbf{M}_1$ and $\mathbf{M}_2$ and four matrices in $\mathbf{M}_5 = \{\mathbf{M}_{5,1}, \mathbf{M}_{5,2}, \mathbf{M}_{5,3}, \mathbf{M}_{5,4}\}$, where $\mathbf{M}_{5,1}$ has 1 on the entries $(1, 2)$, $(2, 1)$, $(m - 1, m)$, $(m, m - 1)$, and 0 elsewhere; $\mathbf{M}_{5,2}$ has 1 on the two second-main off-diagonals and 0 elsewhere; $\mathbf{M}_{5,3}$ has 1 on the entries $(1, 3)$, $(3, 1)$, $(m - 2, m)$, $(m, m - 2)$, and 0 elsewhere; and $\mathbf{M}_{5,4}$ has 1 on the corners $(1, m)$ and $(m, 1)$ and 0 elsewhere. Although these seven basis matrices do not fully represent $\mathbf{R}_5^{-1}$, they provide a good approximation to $\mathbf{R}_5^{-1}$. The irrelevant basis matrix $\mathbf{M}_3$ in Study 1 is also included, so that the basis misspecification not only omits relevant matrices but also contains irrelevant matrices.

For 100 generated datasets, we report the efficiency improvement for estimating $\boldsymbol{\beta}$ using the selected matrices in Table 5, and the difference between $\hat{\mathbf{R}}_5$ and $\mathbf{R}_5$ measured by $||\hat{\mathbf{R}}_5 - \mathbf{R}||_{\mathrm{F}}/m$ in Table 6, where $||\mathbf{A}||_{\mathrm{F}}$ is the Frobenius norm defined by the square root of $tr(\mathbf{A}^T\mathbf{A})$. Table 5 shows that under the misspecified basis matrices, the proposed method can still improve the estimation efficiency significantly for $\hat{\boldsymbol{\beta}}$, compared with the independence or the unstructured working correlation structure. In addition, Table 6 indicates that the Frobenius norm shows less bias in correlation matrix estimation using the proposed method than in the initial estimate. The efficiency improvement for regression parameter estimation and the bias reduction for correlation parameter estimation are more significant with a smaller sample size.

6. EXAMPLE: HIV DATA

We apply the proposed method to HIV (human immunodeficiency virus) AIDS (acquired immunodeficiency syndrome) data (Huang, Wu, and Zhou 2002; Fan and Li 2004; Qu and Li 2006; Fan, Huang, and Li 2007) for illustration. In this dataset, there are 283 homosexual males who were HIV positive between 1984 and 1991. Each patient had his first visit after HIV infection, and his CD4 (cluster of differentiation 4) counts were measured repeatedly about every 6 months. Due to missing data, the number of repeated measurements of CD4 varies from a minimum of 1 to a maximum of 14. It is known that HIV destroys CD4 cells, and therefore, it is important to monitor progression of the disease through CD4 counts over time. The response variable here is the CD4 percentage over time, and is considered to be approximately normal. Four covariates were also collected: smoking status as a binary measurement (x_{i1}), standardized patient age (x_{i2}), standardized CD4 cell percentage before infection (x_{i3}), and measurement time (t_{ij}). We adopt the following model, suggested by Fan and Li (2004), for the marginal mean:

$$y_{ij} = \alpha_0(t_{ij}) + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i2}^2 + \beta_5 x_{i3}^2 + \beta_6 x_{i1} x_{i2} + \beta_7 x_{i1} x_{i3} + \beta_8 x_{i2} x_{i3} + \epsilon_{ij}, \qquad (6)$$

where the varying intercept coefficient $\alpha_0(t_{ij})$ is estimated by B-splines.

We study the correlation structure of the first six observations from each subject since these measurements were followed more regularly in the 6-month interval in the earlier phase of the study. There are total 244 subjects remaining in our subset, where there are 32 with two measurements, 20 with three measurements, 17 with four measurements, 25 with five measurements, and 150 with six measurements. To select the correlation structure for this unbalanced data, we obtain the initial estimator of the correlation matrix using the 150 patients with six measurements, and implement the procedure proposed in Section 3.3. Model

Table 5. Study 4: Efficiency improvement with balanced and normal response under basis matrices misspecification. The true correlation structure is the hybrid structure of AR(1) and exchangeable

R	n	Ratio of variance	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
$\mathbf{R}_5$	50	Independence/selected	0.721	3.908	5.706	3.607
		Unstructured/selected	892.943	275.668	462.136	529.959
	100	Independence/selected	0.834	3.994	5.792	5.254
		Unstructured/selected	86.097	186.235	408.396	113.139
	200	Independence/selected	0.956	5.871	3.183	5.117
		Unstructured/selected	1.497	2.853	1.721	2.709

Table 6. Study 4: Bias reduction with balanced and normal response under basis matrices misspecification. $||\hat{\mathbf{R}} - \mathbf{R}||_F$ is the Frobenius norm of the difference between the estimated correlation matrix and the true matrix. $||\tilde{\mathbf{R}} - \mathbf{R}||_F$ is the Frobenius norm of the difference between the initial estimated correlation matrix and the true matrix. The mean and the standard deviation (in parenthesis) from 100 datasets are reported

$\mathbf{R}$	n	$\|\|\hat{\mathbf{R}} - \mathbf{R}\|\|_F/m$	$\|\|\tilde{\mathbf{R}} - \mathbf{R}\|\|_F/m$
$\mathbf{R}_5$	50	0.418 (0.200)	1.199 (0.467)
	100	0.325 (0.137)	0.723 (0.214)
	200	0.191 (0.096)	0.454 (0.115)

(6) is first fitted using the GEE with independence correlation structure to obtain the initial regression parameter estimate. The empirical correlation matrix estimated by the sample correlation of the fitted residuals from 150 patients is $\tilde{\mathbf{R}}^*_{\mathrm{HIV}}$,

$$\tilde{\mathbf{R}}^*_{\mathrm{HIV}}$$

$$\begin{pmatrix} 1.00 & 0.71 & 0.59 & 0.52 & 0.47 & 0.43 \\ 0.71 & 1.00 & 0.73 & 0.69 & 0.56 & 0.52 \\ 0.59 & 0.73 & 1.00 & 0.77 & 0.70 & 0.58 \\ 0.52 & 0.69 & 0.77 & 1.00 & 0.77 & 0.68 \\ 0.47 & 0.56 & 0.70 & 0.77 & 1.00 & 0.76 \\ 0.43 & 0.52 & 0.58 & 0.68 & 0.76 & 1.00 \end{pmatrix}$$

$$\hat{\mathbf{R}}_{\mathrm{HIV}}$$

$$\times \begin{pmatrix} 1.00 & 0.76 & 0.63 & 0.56 & 0.48 & 0.41 \\ 0.76 & 1.00 & 0.75 & 0.68 & 0.58 & 0.48 \\ 0.63 & 0.75 & 1.00 & 0.76 & 0.68 & 0.56 \\ 0.56 & 0.68 & 0.76 & 1.00 & 0.75 & 0.63 \\ 0.48 & 0.58 & 0.68 & 0.75 & 1.00 & 0.76 \\ 0.41 & 0.48 & 0.56 & 0.63 & 0.76 & 1.00 \end{pmatrix}.$$

We consider three candidate structures, namely AR(1), exchangeable, and the hybrid structure of AR(1) and exchangeable, to select the structure for $\tilde{\mathbf{R}}^*_{\mathrm{HIV}}$. For the hybrid structure, we include $\mathbf{M}_5 = \{\mathbf{M}_{5,1}, \mathbf{M}_{5,2}, \mathbf{M}_{5,3}, \mathbf{M}_{5,4}\}$ in Study 4 of Section 5 to approximate the inverse of this structure. We apply the proposed approach with the basis matrices in groups $\mathbf{M}_1$, $\mathbf{M}_2$, $\mathbf{M}_3$, and $\mathbf{M}_5$ of Section 5. The final result shows that group $\mathbf{M}_3$ is not selected, indicating that the correlation structure for this data can be well approximated by a hybrid correlation structure of AR(1) and exchangeable. After taking the inverse of the estimated linear combination of the selected basis matrices and standardizing for diagonal elements through dividing the ith row and column by the square root of the corresponding diagonal element, the estimated correlation matrix is $\hat{\mathbf{R}}_{\mathrm{HIV}}$. Note that the standardization performed in $\hat{\mathbf{R}}_{\mathrm{HIV}}$ is only to ensure 1 on the diagonal. Before standardization, the diagonal entries were (1.023, 0.984, 0.944, 0.944, 0.984, 1.023), which are close to 1. The negligible deviance from 1 is mainly due to the estimation error in the $\boldsymbol{\alpha}_i$'s.

Note that neither the exchangeable nor the AR(1) structure alone is a good approximation to the empirical $\tilde{\mathbf{R}}^*_{\mathrm{HIV}}$ due to different off-diagonal entries and a much slower decay of correlations than the AR(1) structure. Based on the selected structure $\hat{\mathbf{R}}_{\mathrm{HIV}}$, we conclude that the correlation structure for this data contains a hybrid structure of AR(1) and exchangeable, with correlation parameters around 0.7 and 0.8 in the AR(1) and exchangeable structures, respectively.

To illustrate the efficiency gain in estimating β_i's through the selected hybrid structure, we compare the regression parameter estimated by the QIF with hybrid structure basis matrices to the GEE estimators with working independence, exchangeable, AR(1), and unstructured correlation structures, respectively. The regression parameter estimates and the standard errors are reported in Table 7, showing that the standard errors by QIF using the selected structure are the smallest among these estimators.

7. DISCUSSION

A new approach is proposed to estimate and select correlation structures simultaneously for longitudinal data. It is able to capture the major correlation structure in the process of model selection and balance model complexity and informativeness. One of the advantages of the proposed approach is that it is able to identify complicated structures that contain a mixture of common structures. This is different from other correlation structure selection approaches such as comparing Akaike's information criteria (Pan 2001) or correlation information criteria (Hin and Wang 2009). Comparing different correlation structures one by one is impractical if the number of candidate structures increases dramatically.

Even if the inverse of the true correlation/covariance matrix does not belong to the space spanned by the selected basis matrices, as long as the candidate basis matrices are in the class of bases with a good approximation for the true structure, the bias is negligible. Although our approach might not obtain the most

Table 7. The estimated regression parameters for the HIV data example. QIF is the quadratic inference function estimator with correlation structure selected by the developed method. The other four estimators are the GEE estimators with working independence, exchangeable, AR(1), and unstructured correlation structures, respectively. The estimated standard errors are reported inside the parentheses

	QIF	GEE.indep	GEE.exch	GEE.AR(1)	GEE.unstr
Smoking	0.92 (0.83)	0.59 (1.11)	0.72 (1.11)	0.81 (1.12)	0.92 (1.33)
Age	0.27 (0.54)	0.02 (0.80)	0.02 (0.79)	0.23 (0.78)	0.16 (0.82)
PreCD4	4.04 (0.50)	3.29 (0.67)	3.61 (0.69)	3.41 (0.68)	3.74 (0.93)
Age^2	−0.43 (0.25)	0.03 (0.39)	0.01 (0.39)	−0.09 (0.38)	−0.21 (0.40)
PreCD4^2	0.66 (0.21)	0.40 (0.31)	0.42 (0.31)	0.39 (0.35)	0.41 (0.56)
Smoking × Age	−2.85 (0.83)	−1.90 (1.18)	−1.78 (1.17)	−1.83 (1.16)	−1.76 (1.39)
Smoking × PreCD4	1.86 (0.92)	0.01 (1.34)	−0.66 (1.33)	−0.47 (1.33)	−1.07 (1.64)
Age × PreCD4	−0.86 (0.34)	−0.01 (0.47)	0.09 (0.48)	−0.07 (0.52)	0.45 (0.60)

accurate working correlation structure, our simulation shows that it can still improve the efficiency of regression parameter estimation compared with an independence structure or an unspecified structure.

In addition, the proposed approach also possesses the oracle property of selecting the true correlation structure and estimating the correlation parameters consistently. Although numerically we cannot guarantee that the linear combination of basis matrices is always positive definite, the oracle property ensures that the estimated correlation matrix is positive definite if the sample size is sufficiently large, as shown in our simulation studies. Note, however, that once the basis matrices are selected to approximate the true correlation structure, the estimation of the regression parameter by minimizing the QIF does not require positive definiteness of the estimated correlation matrix. This is because the construction of the QIF has the advantage of not relying on the inverse of the correlation matrix. Therefore, positive definiteness of the estimated correlation matrix is not crucial for the purpose of increasing the estimation efficiency of the regression parameters.

APPENDIX

Proof of Theorem 1. Let $\mathbf{U}_i$, $\mathbf{V}_{ij}$, and $\mathbf{A}_i$ be $\mathbf{U}_i(\tilde{\boldsymbol{\beta}})$, $\mathbf{V}_{ij}(\tilde{\boldsymbol{\beta}})$, and $\mathbf{A}_i(\tilde{\boldsymbol{\beta}})$ here to indicate their dependence on $\tilde{\boldsymbol{\beta}}$. Under regularity conditions, $n^{\frac{1}{2}}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})$ is asymptotically normal. Given C_1 and the asymptotical normality of $\tilde{\boldsymbol{\beta}}$, the proof of Theorem 1(a) is straightforward and thus is omitted.

Assuming C_1, it can be shown that, for some covariance matrix $\boldsymbol{\Sigma}_1$,

$$\mathrm{vec}\{n^{\frac{1}{2}}[\tilde{\mathbf{R}}^{-1}(\tilde{\boldsymbol{\beta}})-\mathbf{R}^{-1}]\}\xrightarrow{D}N(\mathbf{0},\boldsymbol{\Sigma}_1).\tag{A.1}$$

For fixed j and k, we have

$$\begin{aligned}(\mathbf{U}_i-\mathbf{V}_i\boldsymbol{\alpha}^0)^T\mathbf{V}_{ij,k} &= \mathbf{U}_i^T\mathbf{V}_{ij,k}-(\mathbf{V}_i\boldsymbol{\alpha}^0)^T\mathbf{V}_{ij,k}\\ &= \mathrm{tr}\Big(\mathbf{V}_{ij,k}\mathbf{U}_i^T\Big)-\mathrm{tr}\Big(\mathbf{V}_{ij,k}(\mathbf{V}_i\boldsymbol{\alpha}^0)^T\Big)\\ &= \mathrm{tr}\Big\{\dot{\boldsymbol{\mu}}_i^T(\tilde{\boldsymbol{\beta}})\mathbf{A}_i(\tilde{\boldsymbol{\beta}})^{-\frac{1}{2}}\mathbf{M}_{j,k}\mathbf{A}_i(\tilde{\boldsymbol{\beta}})^{-\frac{1}{2}}(\mathbf{y}_i-\boldsymbol{\mu}_i(\tilde{\boldsymbol{\beta}}))\\ &\quad\times(\mathbf{y}_i-\boldsymbol{\mu}_i(\tilde{\boldsymbol{\beta}}))^T\mathbf{A}_i(\tilde{\boldsymbol{\beta}})^{-\frac{1}{2}}\\ &\quad\times\Big[\tilde{\mathbf{R}}^{-1}(\tilde{\boldsymbol{\beta}})-\mathbf{R}^{-1}+o_p(n^{-\frac{1}{2}})\Big]\mathbf{A}_i(\tilde{\boldsymbol{\beta}})^{-\frac{1}{2}}\dot{\boldsymbol{\mu}}_i(\tilde{\boldsymbol{\beta}})\Big\}\\ &= \mathrm{tr}\Big\{[\tilde{\mathbf{R}}^{-1}(\tilde{\boldsymbol{\beta}})-\mathbf{R}^{-1}+o_p(n^{-\frac{1}{2}})]\mathbf{A}_i(\tilde{\boldsymbol{\beta}})^{-\frac{1}{2}}\dot{\boldsymbol{\mu}}_i(\tilde{\boldsymbol{\beta}})\dot{\boldsymbol{\mu}}_i^T(\tilde{\boldsymbol{\beta}})\\ &\quad\times\mathbf{A}_i(\tilde{\boldsymbol{\beta}})^{-\frac{1}{2}}\mathbf{M}_{j,k}\mathbf{A}_i(\tilde{\boldsymbol{\beta}})^{-\frac{1}{2}}\\ &\quad\times(\mathbf{y}_i-\boldsymbol{\mu}_i(\tilde{\boldsymbol{\beta}}))(\mathbf{y}_i-\boldsymbol{\mu}_i(\tilde{\boldsymbol{\beta}}))^T\mathbf{A}_i(\tilde{\boldsymbol{\beta}})^{-\frac{1}{2}}\Big\}.\end{aligned}$$

Let $\mathbf{Q}_{ij,k}(\tilde{\boldsymbol{\beta}})=\mathbf{A}_i(\tilde{\boldsymbol{\beta}})^{-\frac{1}{2}}\dot{\boldsymbol{\mu}}_i(\tilde{\boldsymbol{\beta}})\dot{\boldsymbol{\mu}}_i^T(\tilde{\boldsymbol{\beta}})\mathbf{A}_i(\tilde{\boldsymbol{\beta}})^{-\frac{1}{2}}\mathbf{M}_{j,k}\mathbf{A}_i(\tilde{\boldsymbol{\beta}})^{-\frac{1}{2}}(\mathbf{y}_i-\boldsymbol{\mu}_i(\tilde{\boldsymbol{\beta}}))\times(\mathbf{y}_i-\boldsymbol{\mu}_i(\tilde{\boldsymbol{\beta}}))^T\mathbf{A}_i(\tilde{\boldsymbol{\beta}})^{-\frac{1}{2}}$,

$$\sum_{i=1}^n(\mathbf{U}_i-\mathbf{V}_i\boldsymbol{\alpha}^0)^T\mathbf{V}_{ij,k}=\mathrm{tr}\Big\{\Big[\tilde{\mathbf{R}}^{-1}(\tilde{\boldsymbol{\beta}})-\mathbf{R}^{-1}+o_p(n^{-\frac{1}{2}})\Big]\Big[\sum_{i=1}^n\mathbf{Q}_{ij,k}(\tilde{\boldsymbol{\beta}})\Big]\Big\}.$$

Thus, we have $n^{-\frac{1}{2}}\sum_{i=1}^n(\mathbf{U}_i-\mathbf{V}_i\boldsymbol{\alpha}^0)^T\mathbf{V}_{ij,k}$ as

$$\mathrm{tr}\Big\{n^{\frac{1}{2}}\Big[\tilde{\mathbf{R}}^{-1}(\tilde{\boldsymbol{\beta}})-\mathbf{R}^{-1}+o_p(n^{-\frac{1}{2}})\Big]\Big[n^{-1}\sum_{i=1}^n\mathbf{Q}_{ij,k}(\tilde{\boldsymbol{\beta}})\Big]\Big\}.$$

For fixed j and k and the true $\boldsymbol{\beta}$, we know that $n^{-1}\sum_{i=1}^n\mathbf{Q}_{ij,k}(\boldsymbol{\beta})\xrightarrow{P}\mathbf{C}_{jk}$ for some constant matrix $\mathbf{C}_{jk}$. Since $\tilde{\boldsymbol{\beta}}$ is $\sqrt{n}$-consistent with $\boldsymbol{\beta}$, we have

$$n^{-1}\sum_{i=1}^n\mathbf{Q}_{ij,k}(\tilde{\boldsymbol{\beta}})\xrightarrow{P}\mathbf{C}_{jk}$$

for $j=1,\dots,J$ and $k=1,\dots,d_j$.

By Slutsky's theorem and (A.1), for any constants a_{jk}, we have the following:

$$\begin{aligned}&n^{-\frac{1}{2}}\sum_{j=1}^J\sum_{k=1}^{d_j}a_{jk}\sum_{i=1}^n(\mathbf{U}_i-\mathbf{V}_i\boldsymbol{\alpha}^0)^T\mathbf{V}_{ij,k}\\ &\quad\xrightarrow{D}\mathrm{tr}\Big\{\Big[n^{\frac{1}{2}}\Big(\tilde{\mathbf{R}}^{-1}(\tilde{\boldsymbol{\beta}})-\ \mathbf{R}^{-1}+o_p\Big(n^{-\frac{1}{2}}\Big)\Big)\Big]\sum_{j=1}^J\sum_{k=1}^{d_j}a_{jk}\mathbf{C}_{jk}\Big\}.\end{aligned}$$

By the Cramer–Wold theorem, we have

$$\begin{aligned}&n^{-\frac{1}{2}}\sum_{i=1}^n\Big(\mathbf{U}_i-\mathbf{V}_i\boldsymbol{\alpha}^0\Big)^T\mathbf{V}_i\\ &\quad\xrightarrow{D}\Big(\mathrm{tr}\Big\{\Big[n^{\frac{1}{2}}(\tilde{\mathbf{R}}^{-1}(\tilde{\boldsymbol{\beta}})-\mathbf{R}^{-1}+o_p\Big(n^{-\frac{1}{2}}\Big)\Big)\Big]\mathbf{C}_{11}\Big\},\dots,\\ &\qquad\mathrm{tr}\Big\{\Big[n^{\frac{1}{2}}\Big(\tilde{\mathbf{R}}^{-1}(\tilde{\boldsymbol{\beta}})-\mathbf{R}^{-1}+o_p(n^{-\frac{1}{2}})\Big)\Big]\mathbf{C}_{Jd_J}\Big\}\Big).\end{aligned}$$

The above trace part can be written as

$$\begin{aligned}&\Big(\mathrm{tr}\Big\{\Big[n^{\frac{1}{2}}\Big(\tilde{\mathbf{R}}^{-1}(\tilde{\boldsymbol{\beta}})-\mathbf{R}^{-1}+o_p\Big(n^{-\frac{1}{2}}\Big)\Big)\Big]\mathbf{C}_{11}\Big\},\dots,\\ &\mathrm{tr}\Big\{\Big[n^{\frac{1}{2}}\Big(\tilde{\mathbf{R}}^{-1}(\tilde{\boldsymbol{\beta}})-\mathbf{R}^{-1}+o_p\Big(n^{-\frac{1}{2}}\Big)\Big)\Big]\mathbf{C}_{Jd_J}\Big\}\Big)\\ &=\Big(\mathrm{vec}\Big\{n^{\frac{1}{2}}\Big[\tilde{\mathbf{R}}^{-1}(\tilde{\boldsymbol{\beta}})-\mathbf{R}^{-1}+o_p\Big(n^{-\frac{1}{2}}\Big)\Big]\Big\}\Big)^T(\mathrm{vec}(\mathbf{C}_{11}),\dots,\mathrm{vec}(\mathbf{C}_{Jd_J}))\\ &=\Big(\mathrm{vec}\Big\{n^{\frac{1}{2}}\Big[\tilde{\mathbf{R}}^{-1}(\tilde{\boldsymbol{\beta}})-\mathbf{R}^{-1}+o_p\Big(n^{-\frac{1}{2}}\Big)\Big]\Big\}\Big)^T\mathbf{C},\end{aligned}$$

where $\mathbf{C}=(\mathrm{vec}(\mathbf{C}_{11}),\dots,\mathrm{vec}(\mathbf{C}_{Jd_J}))$. By (A.1), we have $n^{-\frac{1}{2}}\sum_{i=1}^n(\mathbf{U}_i-\mathbf{V}_i\boldsymbol{\alpha}^0)^T\mathbf{V}_i\xrightarrow{D}N(\mathbf{0},\boldsymbol{\Sigma}^*)$, and $\boldsymbol{\Sigma}^*=\mathbf{C}^T\boldsymbol{\Sigma}_1\mathbf{C}$. Theorem 1(b) is proved.

Proof of Theorem 2. We first establish the convergence rate of $\hat{\boldsymbol{\alpha}}^{(0)}$ to $\boldsymbol{\alpha}^0$. Define $L_n(\boldsymbol{\alpha})=\sum_{i=1}^n||\mathbf{U}_i-\sum_{j=1}^J\mathbf{V}_{ij}\boldsymbol{\alpha}_j||_2^2$. Given that $\hat{\boldsymbol{\alpha}}^{(0)}$ minimizes $L_n(\boldsymbol{\alpha})$, we have

$$\begin{aligned}L_n\Big(\hat{\boldsymbol{\alpha}}^{(0)}\Big)-L_n(\boldsymbol{\alpha}^0)&=\Big(\hat{\boldsymbol{\alpha}}^{(0)}-\boldsymbol{\alpha}^0\Big)^T\Big[\sum_{i=1}^n\mathbf{V}_i^T\mathbf{V}_i\Big]\Big(\hat{\boldsymbol{\alpha}}^{(0)}-\boldsymbol{\alpha}^0\Big)\\ &\quad-2\Big[\sum_{i=1}^n(\mathbf{U}_i-\mathbf{V}_i\boldsymbol{\alpha}^0)^T\mathbf{V}_i\Big]\Big(\hat{\boldsymbol{\alpha}}^{(0)}-\boldsymbol{\alpha}^0\Big)\le 0.\end{aligned}$$

Let ξ_1 be the smallest eigenvalue of $\boldsymbol{\Sigma}$ in Theorem 1(a). By Theorem 1(b), we have

$$\begin{aligned}\xi_1 n\Big|\Big|\hat{\boldsymbol{\alpha}}^{(0)}-\boldsymbol{\alpha}^0\Big|\Big|_2^2&\le\Big(\hat{\boldsymbol{\alpha}}^{(0)}-\boldsymbol{\alpha}^0\Big)^T\Big[\sum_{i=1}^n\mathbf{V}_i^T\mathbf{V}_i\Big]\Big(\hat{\boldsymbol{\alpha}}^{(0)}-\boldsymbol{\alpha}^0\Big)\\ &\le 2\Big[\sum_{i=1}^n(\mathbf{U}_i-\mathbf{V}_i\boldsymbol{\alpha}^0)^T\mathbf{V}_i\Big]\Big(\hat{\boldsymbol{\alpha}}^{(0)}-\boldsymbol{\alpha}^0\Big)\\ &\le 2\Big|\Big|\sum_{i=1}^n(\mathbf{U}_i-\mathbf{V}_i\boldsymbol{\alpha}^0)^T\mathbf{V}_i\Big|\Big|_2\Big|\Big|\hat{\boldsymbol{\alpha}}^{(0)}-\boldsymbol{\alpha}^0\Big|\Big|_2\\ &=O_p(n^{\frac{1}{2}})\Big|\Big|\hat{\boldsymbol{\alpha}}^{(0)}-\boldsymbol{\alpha}^0\Big|\Big|_2,\end{aligned}$$

with probability tending to 1. Thus, we establish the convergence rate of

$$\Big|\Big|\hat{\boldsymbol{\alpha}}^{(0)}-\boldsymbol{\alpha}^0\Big|\Big|_2=O_p\Big(n^{-\frac{1}{2}}\Big).\tag{A.2}$$

Consequently, $\hat{\alpha}_{j,k}^{(0)}=O_p(1)$ for $\alpha_{j,k}^0\neq 0$ and $\hat{\alpha}_{j,k}^{(0)}=O_p(n^{-\frac{1}{2}})$ for $\alpha_{j,k}^0=0$. This also implies that if n is large enough and C_2 holds, we have $p'_{\lambda_n}(||\hat{\boldsymbol{\alpha}}_j^{(0)}||_1)=0$ for $\boldsymbol{\alpha}_j^0\neq\mathbf{0}$ and $p'_{\lambda_n}(||\hat{\boldsymbol{\alpha}}_j^{(0)}||_1)=\lambda_n$ for $\boldsymbol{\alpha}_j^0=\mathbf{0}$. Thus, letting $T_n(\boldsymbol{\alpha})$ be the objective function in (4), we have $T_n(\hat{\boldsymbol{\alpha}})-T_n(\boldsymbol{\alpha}^0)=(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^0)^T[\sum_{i=1}^n\mathbf{V}_i^T\mathbf{V}_i](\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^0)-2[\sum_{i=1}^n(\mathbf{U}_i-\mathbf{V}_i\boldsymbol{\alpha}^0)^T\mathbf{V}_i](\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^0)+np\sum_{\boldsymbol{\alpha}_j^0=0}\lambda_n||\hat{\boldsymbol{\alpha}}_j||_1\le 0.$ Since

$np\sum_{\boldsymbol{\alpha}_j^0=0}\lambda_n||\hat{\boldsymbol{\alpha}}_j||_1 \geq 0$, similar to obtaining (A.2), we have

$$||\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^0||_2 = O_p\left(n^{-\frac{1}{2}}\right). \tag{A.3}$$

Since $p'_{\lambda_n}(||\hat{\boldsymbol{\alpha}}_j^{(0)}||_1)=0$ for $\boldsymbol{\alpha}_j^0 \neq \mathbf{0}$ and $p'_{\lambda_n}(||\hat{\boldsymbol{\alpha}}_j^{(0)}||_1)=\lambda_n$ for $\boldsymbol{\alpha}_j^0=\mathbf{0}$ for large enough n, the objective function $T_n(\boldsymbol{\alpha})$ becomes $T_n(\boldsymbol{\alpha})=\sum_{i=1}^n||\mathbf{U}_i-\mathbf{V}_i\boldsymbol{\alpha}||_2^2+np\lambda_n||\boldsymbol{\alpha}_{\mathrm{I}}||_1$. Taking the partial derivative of $T_n(\boldsymbol{\alpha})$ with respect to $\alpha_{j,k}$ for $\alpha_{j,k}\in\boldsymbol{\alpha}_{\mathrm{I}}$, we have $\frac{\partial T_n(\boldsymbol{\alpha})}{\partial\alpha_{j,k}}|_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}}=-\sum_{i=1}^n 2(\mathbf{U}_i-\mathbf{V}_i\hat{\boldsymbol{\alpha}})^T\mathbf{V}_{ij,k}+np\lambda_n\mathrm{Sign}(\hat{\alpha}_{j,k})=-\sum_{i=1}^n 2(\mathbf{U}_i-\mathbf{V}_i\boldsymbol{\alpha}^0)^T\mathbf{V}_{ij,k}+\sum_{i=1}^n 2(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^0)^T\mathbf{V}_i^T\mathbf{V}_{ij,k}+np\lambda_n\mathrm{Sign}(\hat{\alpha}_{j,k})$, where $\mathbf{V}_{ij,k}$ is the column of V_i corresponding to $\alpha_{j,k}$. Given Theorem 1(b), we have $\sum_{i=1}^n 2(\mathbf{U}_i-\mathbf{V}_i\boldsymbol{\alpha}^0)^T\mathbf{V}_{ij,k}=O_p(n^{\frac{1}{2}})$. Theorem 1(a) and (A.3) imply that $\sum_{i=1}^n 2(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^0)^T\mathbf{V}_i^T\mathbf{V}_{ij,k}=O_p(n^{\frac{1}{2}})$. Suppose $\hat{\alpha}_{j,k}\neq 0$ for $\alpha_{j,k}^0\in\boldsymbol{\alpha}_{\mathrm{I}}^0$. Since $\hat{\boldsymbol{\alpha}}$ minimizes $T_n(\boldsymbol{\alpha})$, we have $\frac{\partial T_n(\boldsymbol{\alpha})}{\partial\alpha_{j,k}}|_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}}=0$, indicating $np\lambda_n\mathrm{Sign}(\hat{\alpha}_{j,k})=\sum_{i=1}^n 2(\mathbf{U}_i-\mathbf{V}_i\boldsymbol{\alpha}^0)^T\mathbf{V}_{ij,k}-\sum_{i=1}^n 2(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^0)^T\mathbf{V}_i^T\mathbf{V}_{ij,k}$. However, the condition C_2 ensures that

$$\frac{\sum_{i=1}^n 2(\mathbf{U}_i-\mathbf{V}_i\boldsymbol{\alpha}^0)^T\mathbf{V}_{ij,k}-\sum_{i=1}^n 2(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^0)^T\mathbf{V}_i^T\mathbf{V}_{ij,k}}{np\lambda_n\mathrm{Sign}(\hat{\alpha}_{j,k})}=o_p(1).$$

Thus, $P\{\hat{\alpha}_{j,k}\neq 0\}\leq P\{np\lambda_n\mathrm{Sign}(\hat{\alpha}_{j,k})=\sum_{i=1}^n 2((\mathbf{U}_i-\mathbf{V}_i\boldsymbol{\alpha}^0)^T\mathbf{V}_{ij,k}-\sum_{i=1}^n 2(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^0)^T\mathbf{V}_i^T\mathbf{V}_{ij,k}\}\to 0$. Therefore, with probability tending to 1, we have $\hat{\alpha}_{j,k}=0$ for $\alpha_{j,k}^0\in\boldsymbol{\alpha}_{\mathrm{I}}^0$, which indicates that the structure of $\mathbf{R}^{-1}$ is identified correctly with probability tending to 1. Theorem 2(a) is proved.

Next, we show the asymptotic normality of $\hat{\boldsymbol{\alpha}}_{\mathrm{II}}$. We define $R_n(\boldsymbol{\alpha}_{\mathrm{II}})=\sum_{i=1}^n||\mathbf{U}_i-\mathbf{V}_{i,\mathrm{II}}\boldsymbol{\alpha}_{\mathrm{II}}||_2^2$, where $\boldsymbol{\alpha}_{\mathrm{II}}$ is the second partitioning part of $\boldsymbol{\alpha}=(\boldsymbol{\alpha}_{\mathrm{I}}^T,\boldsymbol{\alpha}_{\mathrm{II}}^T)^T$, and $\mathbf{V}_{i,\mathrm{II}}$ contains the columns of $\mathbf{V}_i$ corresponding to the parameters in $\boldsymbol{\alpha}_{\mathrm{II}}$. Since $\hat{\boldsymbol{\alpha}}$ minimizes $T_n(\boldsymbol{\alpha})$ and $p'_{\lambda_n}(||\hat{\boldsymbol{\alpha}}_j^{(0)}||_1)=0$ for $\boldsymbol{\alpha}_j^0\neq\mathbf{0}$, and $p'_{\lambda_n}(||\hat{\boldsymbol{\alpha}}_j^{(0)}||_1)=\lambda_n$ for $\boldsymbol{\alpha}_j^0=\mathbf{0}$, and $\hat{\boldsymbol{\alpha}}_{\mathrm{I}}=\mathbf{0}$ with probability tending to 1 from the proof of Theorem 2(a), we know that $\hat{\boldsymbol{\alpha}}_{\mathrm{II}}$ minimizes the objective function $R_n(\boldsymbol{\alpha}_{\mathrm{II}})$ and $\nabla R_n(\hat{\boldsymbol{\alpha}}_{\mathrm{II}})=\mathbf{0}$, with probability tending to 1. By Taylor expansion, we have

$$\nabla R_n(\hat{\boldsymbol{\alpha}}_{\mathrm{II}})=\nabla R_n\left(\boldsymbol{\alpha}_{\mathrm{II}}^0\right)+\nabla^2 R_n\left(\boldsymbol{\alpha}_{\mathrm{II}}^*\right)\left(\hat{\boldsymbol{\alpha}}_{\mathrm{II}}-\boldsymbol{\alpha}_{\mathrm{II}}^0\right),$$

for some vector $\boldsymbol{\alpha}_{\mathrm{II}}^*$. Since $\nabla R_n(\hat{\boldsymbol{\alpha}}_{\mathrm{II}})=0$ with probability tending to 1, we have

$$\begin{aligned}\hat{\boldsymbol{\alpha}}_{\mathrm{II}}-\boldsymbol{\alpha}_{\mathrm{II}}^0&=-[\nabla^2 R_n(\boldsymbol{\alpha}_{\mathrm{II}}^*)]^{-1}\nabla R_n\left(\boldsymbol{\alpha}_{\mathrm{II}}^0\right)\\&=\left(\sum_{i=1}^n\mathbf{V}_{i,\mathrm{II}}^T\mathbf{V}_{i,\mathrm{II}}\right)^{-1}\left[\sum_{i=1}^n\mathbf{V}_{i,\mathrm{II}}^T\left(\mathbf{U}_i-\mathbf{V}_{i,\mathrm{II}}\boldsymbol{\alpha}_{\mathrm{II}}^0\right)\right].\end{aligned}$$

By Theorem 1, we have $\sqrt{n}(\hat{\boldsymbol{\alpha}}_{\mathrm{II}}-\boldsymbol{\alpha}_{\mathrm{II}}^0)\sim N(\mathbf{0},\boldsymbol{\Sigma}_{\mathrm{II}}^{-1}\boldsymbol{\Sigma}_{\mathrm{II}}^*\boldsymbol{\Sigma}_{\mathrm{II}}^{-1})$, where $\boldsymbol{\Sigma}_{\mathrm{II}}$ and $\boldsymbol{\Sigma}_{\mathrm{II}}^*$ are the submatrices of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^*$, corresponding to the columns of $\mathbf{V}_{i,\mathrm{II}}$. Theorem 2(b) is proved.

[Received February 2011. Revised January 2012.]

REFERENCES

Bickel, P., and Levina, E. (2008), "Covariance Regularization by Thresholding," *The Annals of Statistics*, 36, 2577–2604. [701]

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression" (with discussion), *The Annals of Statistics*, 32, 407–499. [703]

El Karoui, N. (2008), "Operator Norm Consistent Estimation of Large-Dimensional Sparse Covariance Matrices," *The Annals of Statistics*, 36, 2717–2756. [701]

Fan, J., Fan, Y., and Lv, J. (2008), "High-Dimensional Covariance Matrix Estimation Using a Factor Model," *Econometrics*, 147, 186–197. [701]

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. [707]

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. [702,703]

——— (2004), "New Estimation and Model Selection Procedures for Semiparametric Modeling in Longitudinal Data Analysis," *Journal of the American Statistical Association*, 99, 710–723. [707]

Hansen, L. (1982), "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054. [702,704]

Hin, L. Y., and Wang, Y. G. (2009), "Working-Correlation-Structure Identification in Generalized Estimating Equations," *Statistics in Medicine*, 28, 642–658. [708]

Huang, J. Z., Liu, L., and Liu, N. (2007), "Estimation of Large Covariance Matrices of Longitudinal Data With Basis Function Approximations," *Journal of Computational and Graphical Statistics*, 16, 189–209. [701]

Huang, J. Z., Liu, N., Pourahmadi, M., and Liu, L. (2006), "Covariance Selection and Estimation via Penalised Normal Likelihood," *Biometrika*, 93, 85–98. [701]

Huang, J. Z., Wu, C. O., and Zhou, L. (2002), "Varying-Coefficient Models and Basis Function Approximations for the Analysis of Repeated Measurements," *Biometrika*, 89, 111–128. [707]

Levina, E., Rothman, A. J., and Zhu, J. (2008), "Sparse Estimation of Large Covariance Matrices via Nested Lasso Penalty," *The Annals of Applied Statistics*, 2, 245–263. [701]

Liang, K. Y., and Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalised Linear Models," *Biometrika*, 73, 13–22. [702]

Lin, X., Wang, N., Welsh, A., and Carroll, R. J. (2004), "Equivalent Kernels of Smoothing Splines in Nonparametric Regression for Clustered Data," *Biometrika*, 91, 177–193. [701]

Pan, W. (2001), "Akaike's Information Criterion in Generalized Estimating Equations," *Biometrics*, 57, 120–125. [708]

Qu, A., and Li, R. (2006), "Quadratic Inference Functions for Varying Coefficient Models With Longitudinal Data," *Biometrics*, 62, 379–391. [707]

Qu, A., Lindsay, B. G., and Li, B. (2000), "Improving Generalised Estimating Equations Using Quadratic Inference Functions," *Biometrika*, 87, 823–836. [702,705]

Qu, A., Lindsay, B. G., and Lu, L. (2010), "Highly Efficient Aggregate Unbiased Estimating Functions Approach for Correlated Data with Missing at Random," *Journal of the American Statistical Association*, 105, 194–204. [704]

Rothman, A. J., Levina, E., and Zhu, J. (2009), "Generalized Thresholding of Large Covariance Matrices," *Journal of the American Statistical Association*, 104, 177–186. [701]

Wang, H., Li, R., and Tsai, C.-L. (2007), "Tuning Parameter Selectors for the Smoothly Clipped Absolute Deviation Method," *Biometrika*, 94, 553–568. [703]

Wang, L., and Qu, A. (2009), "Consistent Model Selection and Data-Driven Tests for Longitudinal Data in the Estimating Equation Approach," *Journal of the Royal Statistical Society,* Series B, 71, 177–190. [703]

Wang, N. (2003), "Marginal Nonparametric Kernel Regression Accounting for Within-Subject Correlation," *Biometrika*, 90, 43–52. [701]

Wang, N., Carroll, R. J., and Lin, X. (2005), "Efficient Semiparametric Marginal Estimation for Longitudinal/Clustered Data," *Journal of the American Statistical Association*, 100, 147–157. [701]

Wedderburn, R. W. M. (1974), "Quasi-Likelihood Functions, Generalised Linear Models and the Gauss-Newton Method," *Biometrika*, 61, 439–488. [702]

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. [703]

Zou, H. (2006), "The Adaptive LASSO and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [703]

Zou, H., and Li, R. (2008), "One-Step Sparse Estimates in Nonconcave Penalized Likelihood Models (with discussion)," *The Annals of Statistics*, 36, 1509–1533. [703]