MIXED EFFECTS MODELING AND CORRELATION STRUCTURE SELECTION FOR HIGH DIMENSIONAL CORRELATED DATA

BY

PENG WANG

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Statistics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2011

Urbana, Illinois

Doctoral Committee:

Associate Professor Annie Qu, Chair
Professor Xuming He
Professor Douglas Simpson
Assistant Professor Xiaofeng Shao

Abstract

Longitudinal data arise frequently in many studies where measurements are obtained from a subject repeatedly over time. Consequently, measurements within a subject are correlated. We address two rather important but challenging issues in this thesis: mixed-effect modeling with unspecified random effects and correlation structure selection for high-dimensional data.

In longitudinal studies, mixed-effects models are important for addressing subject-specific effects. However, most existing approaches assume normal distributions for the random effects, which could affect the bias and efficiency of the fixed-effects estimators. Even in the cases where the estimation of the fixed effects is robust against a misspecified distribution of the random effects, the inference based on the random effects could be invalid. We propose a new approach to estimate fixed and random effects using conditional quadratic inference functions. The new approach does not require any specification of the likelihood functions. It can also accommodate serial correlation between observations within the same cluster, in addition to mixed-effects modeling. Other advantages include not requiring the estimation of the unknown variance components associated with the random effects, or the nuisance parameters associated with the working correlations. Real data examples and simulations are used to compare the new approach with the penalized quasi-likelihood approach, and SAS the GLIMMIX and nonlinear mixed effects model (NLMIXED) procedures.

Model selection of correlation structure for non-normal correlated data is very challenging when the cluster size increases with the sample size, because of the high dimensional correlation parameters involved and lack of the likelihood function for non-normal correlated data. However, identifying the correct correlation structure can improve estimation efficiency and the power of tests for correlated data. We propose to approximate the inverse of the empirical correlation matrix using a linear combination of candidate basis matrices, and select the correlation structure by identifying non-zero coefficients of the basis matrices. This is carried out by minimizing penalized estimating functions, which balances the complexity and informativeness of modeling for the correlation matrix. The new approach does not require estimating each entry of the correlation matrix, nor the specification of the likelihood function, and can effectively handle non-normal

correlated data. Asymptotic theory on model selection consistency and oracle properties are established in the framework of diverging cluster size of correlated data, where the derivation of the asymptotic results is challenging. Our numerical studies indicate that even when the cluster size is very large, the correlation structure can be identified effectively for both normal responses and binary responses.

To All Who Believe in Me.

Acknowledgments

First of all, I would like to express my most gratitude to my advisor Professor Annie Qu. The dissertation would not have been possible without her guidance and support. I am thankful for her stimulated discussions, inspirations, and suggestions that help to shape my technical, computational, writing and other research skills. She read and helped with numerous revisions of this dissertation. I greatly appreciate her encouragement and patience, which carried me through difficult times, in both my research and life.

I owe my thanks to Professor Xuming He and Professor Xiaofeng Shao for taking time to help with my presentations, and for being my committee members. I also would like to thank Professor Douglas Simpson, for being my committee member and for his service as the department chair. The valuable feedback of the committee members helped to improve the dissertation substantially.

My thanks also go to Christopher J. Vecoli who attentively proof read my dissertation with great patience. He has offered helpful advice to improve the writing of this dissertation and my academic writing skill in general. My thanks are also given to Professor Jianhui Zhou of Unviersity of Virginia for helpful discussions.

Many thanks to the faculty and staff members of the Department of Statistics in University of Illinois for creating a supportive environment for my study. I also thank my fellow students in the department, particularly Na Cui, Yang Feng, Bin Li, Yunwen Yang, Jing Xia, Xianyang Zhang, Chunxiao Zhou, and Zhi He for their support, encouragement and friendship.

Finally, I want to thank my mother and my wife, and all my beloved friends, for their continuous and unconditional love, support and faith in me.

Table of Contents

List of Tables . . . . . **viii**

List of Figures . . . . . **ix**

List of Abbreviations . . . . . **x**

Chapter 1 Introduction **1**
1.1 Introduction to Longitudinal Data 1
1.2 A Review of Methods on Marginal Approach and the Mixed-Effects Approach 3
1.2.1 Review of Methods on Marginal Approach 3
1.2.2 Review of Methods on Mixed-Effects Approach 5
1.3 The New Proposed Approaches 6

Chapter 2 Mixed-Effects Modelling **8**
2.1 Introduction 8
2.2 Quadratic Inference Functions for Mixed-Effects Models 10
2.2.1 Quadratic Inference Functions for Fixed Models 11
2.2.2 Implementation 16
2.3 Asymptotic Properties 17
2.4 Simulation 18
2.5 Real-data Example 22
2.5.1 Periodontal Example for Binary Data 22
2.5.2 Epileptic Seizure Data Example for Poisson Data 23
2.6 Discussion 24
2.7 Proofs of Theorems and Lemmas 25
2.7.1 Notation 25
2.7.2 Regularity Conditions and Assumptions 26
2.7.3 Proofs of Lemmas and Theorem 1 26
2.7.4 Conditions and Proof of Consistency of Random-effect Estimator 30

Chapter 3 Correlation Structure Selection for High Dimensional Correlated Data **35**
3.1 Introduction 35
3.2 General Framework 37
3.3 Selection of Correlation Structure 38
3.4 Asymptotic Properties 40
3.5 Implementation 43
3.5.1 Examples of Basis Matrices 43
3.5.2 Tuning Parameter Selection 44
3.6 Simulations 45
3.7 Data Example on Air Pollution 48
3.8 Discussion 50
3.9 Proof of Lemmas and Theorems 51

3.9.1 Lemma 1 51
3.9.2 Lemma 2 53
3.9.3 Proof of Theorem 1 54
3.9.4 Proof of Theorem 2 55
3.9.5 Proof of Theorem 3 56

References 59

Vita 62

List of Tables

2.1 MSE and the standard errors of MSE (provided in the lower right corner) for the estimator of the intercept $\beta_0 = -0.3$ for binary responses when $\rho = 0.4$ from 200 simulations. 20
2.2 MSE and the standard errors of MSE (provided in the lower right corner) for the estimator of the slope $\beta_1 = 0.3$ for binary responses when $\rho = 0.4$ from 200 simulations. 20
2.3 MSE and the standard errors of MSE (provided in the lower right corner) for the estimator of the intercept $\beta_0 = -0.3$ for binary responses when $\rho = 0.7$ from 200 simulations. 21
2.4 MSE and the standard errors of MSE (provided in the lower right corner) for the estimator of the slope $\beta_1 = 0.3$ for binary responses when $\rho = 0.7$ from 200 simulations. 21
2.5 Mean and the standard errors of mean (provided in the lower right corner) of the variance component estimator for random intercepts out of 200 simulations for binary response, when $N = 100$ and $\rho = 0.7$. The true variance of the random intercept is 0.015. 32
2.6 Mean and the standard errors of mean (provided in the lower right corner) of the variance component estimator for random intercepts out of 200 simulations for binary response, when $N = 100$ and $\rho = 0.7$. The true variance of the random intercept is 0.015. 32
2.7 Comparison of mixed QIF and the other approaches for non-surgical periodontal treatment data. 34

3.1 Percentages of correctly identified signals and non-signals using the GIC criterion with correlation $\rho = 0.5$ for normal responses, sample size $n = 200$ 45
3.2 Percentages of correctly identified signals and non-signals using the GIC criterion with correlation $\rho = 0.7$ for normal responses, sample size $n = 200$ 46
3.3 Percentages of correctly identified signals and non-signals using the GIC criterion with correlation $\rho = 0.6$ for binary response, sample size $n = 200$ 46
3.4 Percentages of correctly identified signals and non-signals using the GIC criterion with correlation $\rho = 0.6$ for binary response, sample size $n = 300$ 47
3.5 Comparison of the GEE estimators, standard errors and Z-values using different working correlation structures for air pollution data. 49

List of Figures

2.1 Histogram of the estimates of the random slopes from the binary data sets with $N = 100$ and $\rho = 0.7$. The true correlation structure of the data set is AR(1), and the estimates are obtained by the mixed-QIF method with AR(1) working correlation. The solid line in the graph provides the random-effects density function generated from the true Beta distribution. 33
2.2 Histogram of the estimates of the random slopes from the binary data sets with $N = 100$ and $\rho = 0.7$. The true correlation structure of the data set is AR(1), and the estimates are obtained by the PQL approach. 33

List of Abbreviations

AIC	Akaike Information Criteria.
AR-1	Autoregressive Order One.
BIC	Bayesian Information Criteria.
CGEE2	Conditional Second Order Generalized Estimating Equation.
CWD	Coordinate-wise Decent.
EXCH	Exchangeable.
GEE	Generalized Estimating Equation.
GCV	Generalized Cross Validation.
GIC	Generalized Information Criteria.
GLMM	Generalized Linear Mixed Model.
IND	Independent.
LASSO	Least Absolute Shrinkage and Selection Operator.
MLE	Maximum Likelihood Estimate.
MSE	Mean Square Error.
PWGLS	Penalized Weighted Generalized Least Square.
PQL	Penalized Quasi-likelihood.
QIF	Quadratic Inference Function.
SCAD	Smoothly Clipped Absolute Deviation.

Chapter 1

Introduction

1.1 Introduction to Longitudinal Data

Longitudinal data involving repeated measurements over time is widely used in many research areas, such as psychology, sociology, biology, environmental and medical science. In such areas, longitudinal studies allow one to explore the association between the responses and the relevant covariates over a certain period of time and the dynamic changes of the treatment effects over time. Moreover, since longitudinal study allows each subject to be used as its own control, the treatment effects can be detected more accurately and effectively as the heterogeneity variations among subjects are reduced to a minimum.

We first provide three examples of longitudinal data. The first two examples are analyzed in Chapter 1 and the third example is analyzed in Chapter 2.

Example 1: Dental data to study the effects of non-surgical periodontal effects. This is a randomized clinical study carried out in the University of Washington dental clinic. The purpose of the study is to evaluate the effect of non-surgical periodontal treatments on tooth loss over time (Stoner, 2000). There are 722 patients with chronic periodontal diseases, and each patient is intended to have a 7-year follow up in the study. The tooth loss of the patients during 7-year period is recorded, along with many explanatory variables including a history of non-surgical periodontal treatments, gender, age, and other covariates corresponding to the health condition of the teeth. Here the response variable is the tooth loss for each patient, which is a binary indicator. We are interested in incorporating subject-specific variation for the model in order to better evaluate the non-surgical treatment effect on tooth loss over time.

Example 2: Epileptic seizure count data. This is a double-blind randomized clinical trial to compare a new anti-epileptic drug with a placebo for reducing epileptic seizure occurrence (Thall and Vail, 1990, p. 664). The new drug and the placebo are randomly assigned to the 59 patients in the study, and each patient is followed for four 2-week periods. The epileptic seizure counts, treatment, the patient's age and the baseline seizure counts (the number of seizure counts in the 8-week period before receiving the drug or the placebo), are collected. The response variable of epileptic seizure counts is considered to follow a Poisson distribution.

In addition to assessing the effect of the observed covariates, we are interested in addressing unobserved subject-specific variation among the patients.

Example 3: Effect of air pollution on asthmatic status. This longitudinal study conducted in Ontario, Canada consists of observations from 39 asthmatic patients during a period of 21 days (Fu, 2003). The environmental researchers are interested in investigating the impact of air pollution on patients' asthmatic status. In the 3-week period, the daily mean temperature and a number of air quality measurements are collected. The response is the patient's daily asthmatic status, which is a binary variable. The sample size of 39 is relatively small compared to the cluster size of 21. The standard approaches for the generalized linear model such as the maximum likelihood and the generalized estimating equation with unspecified correlation structure do not have converged solutions, as the dimension of the nuisance parameters being estimated is large. For this data example, it is important to identify the correct correlation structure to obtain a more efficient estimator.

The challenges of analyzing these longitudinal data sets are that we need to take into account the correlation between the repeated measurements. This is especially challenging if the outcomes are not normal, as in these data examples. Ignoring the correlation or misspecifying the correlation structure may lead to inefficient estimators for the regression parameters. Moreover, the variance estimation of the regression parameter estimators are inconsistent (Liang and Zeger, 1986; Qu et al., 2000). However, the existing literature on the model selection of correlation structure is limited, especially for the high-dimensional setting where the number of repeated measurements diverges with the sample size. The available work mainly focuses on the estimation of the covariance matrix rather than the selection of correlation structures. In Chapter 3 of this thesis, we propose a new method to identity the correlation structure for longitudinal data when the cluster size diverges.

The two major approaches for longitudinal data are marginal models, and random effects models deal with the subject-specific variation. In the marginal approach, all the regression parameters are "fixed," and the within-subject correlation is modeled through the residual error. The marginal model is applicable if only the inference the population average is of our interest. The mixed effects model, on the other hand, could also account for the heterogeneity variation among the subjects. The random effects model is suitable when the subject-specific effect is of interest as it can provide a richer interpretation from the individual level. However, it imposes additional challenges as the random effects are not necessarily normally distributed, and the likelihood function might be intractable for non-normal response when the serial correlation is considered in the model.

1.2 A Review of Methods on Marginal Approach and the Mixed-Effects Approach

1.2.1 Review of Methods on Marginal Approach

Liang and Zeger (1986) proposed the following generalized estimation equation (GEE) approach for longitudinal data. The GEE extends the quasi-likelihood approach for longitudinal data, and is defined as

$$\sum_{i=1}^{N} \dot{\boldsymbol{\mu}}_i' \mathbf{V}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) = 0, \tag{1.1}$$

where $\mathbf{y}_i$ is the response vector and $\boldsymbol{\mu}_i$ is the mean vector of the response variable for the ith subject, $\mathbf{V}_i = \mathbf{A}_i^{1/2} \mathbf{R} \mathbf{A}_i^{1/2}$, $\mathbf{A}_i$ is a diagonal marginal variance matrix and $\mathbf{R}$ is a working correlation matrix which involves a limited number of correlation parameters. The regression parameters $\boldsymbol{\beta}$ can be estimated consistently even if the correlation structure is unspecified. It also provides the following robust sandwich estimator for the variance of the regression parameter estimator,

$$(\sum_{i=1}^{N} \dot{\boldsymbol{\mu}}_i' \mathbf{V}_i^{-1} \dot{\boldsymbol{\mu}}_i')^{-1} \{\sum_{i=1}^{N} \dot{\boldsymbol{\mu}}_i' \mathbf{V}_i^{-1} \mathrm{cov}(\mathbf{y}_i) \mathbf{V}_i^{-1} \dot{\boldsymbol{\mu}}_i'\} (\sum_{i=1}^{N} \dot{\boldsymbol{\mu}}_i' \mathbf{V}_i^{-1} \dot{\boldsymbol{\mu}}_i')^{-1}.$$

If the correlation structure is correctly specified, the GEE estimator of $\boldsymbol{\beta}$ is efficient. However, when the correlation structure is misspecified, the GEE estimators could be inefficient (Liang and Zeger, 1986).

To improve the efficiency of the GEE estimator, Qu et al. (2000) proposed the quadratic inference function (QIF) approach, where the inverse of the correlation matrix is approximated by a linear combination of basis matrices through

$$\mathbf{R}^{-1} = \sum_{j=1}^{m} a_i \mathbf{M}_i. \tag{1.2}$$

The linear representation of (1.2) can replace the inverse of working correlation matrix $\mathbf{R}$ in the quasi-likelihood equation (1.1), and the GEE can be approximated as

$$\sum_{i=1}^{N} \dot{\boldsymbol{\mu}}_i' \mathbf{A}_i^{-1/2} \left(\sum_{j=1}^{m} a_j \mathbf{M}_j \right) \mathbf{A}_i^{-1/2} (\mathbf{y}_i - \boldsymbol{\mu}_i) = 0. \tag{1.3}$$

Instead of estimating the nuisance parameters $\mathbf{a} = (a_1, \ldots, a_m)$ they define an extended score derived from

(1.3),

$$\mathbf{G}_N(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N}\mathbf{g}_i(\boldsymbol{\beta}) = \frac{1}{N}\begin{pmatrix} \sum_{\mathrm{i}=1}^{\mathrm{N}} \dot{\boldsymbol{\mu}}_{\mathrm{i}}'\mathbf{A}_{\mathrm{i}}^{-1/2}\mathbf{M}_1\mathbf{A}_{\mathrm{i}}^{-1/2}\left(\mathbf{y}_{\mathrm{i}} - \boldsymbol{\mu}_{\mathrm{i}}\right) \\ \vdots \\ \sum_{\mathrm{i}=1}^{\mathrm{N}} \dot{\boldsymbol{\mu}}_{\mathrm{i}}'\mathbf{A}_{\mathrm{i}}^{-1/2}\mathbf{M}_{\mathrm{m}}\mathbf{A}_{\mathrm{i}}^{-1/2}\left(\mathbf{y}_{\mathrm{i}} - \boldsymbol{\mu}_{\mathrm{i}}\right) \end{pmatrix}. \tag{1.4}$$

It is obvious that (1.3) is a linear combination of the above extended score. When the number of basis matrices $m > 1$, there are more equations than the dimension of parameters, therefore, the generalized method of moments (Hansen, 1982) can be applied to estimate the regression parameter $\boldsymbol{\beta}$ by minimizing the following quadratic inference function

$$Q_N(\boldsymbol{\beta}) = N\mathbf{G}_N'\mathbf{C}_N^{-1}\mathbf{G}_N.$$

The advantage of the QIF approach is that the correlation parameters are not required to be estimated in order to obtain the estimator of $\boldsymbol{\beta}$. The QIF estimator $\hat{\boldsymbol{\beta}}$ has typical $\sqrt{N}$-consistency and asymptotic normality, with the asymptotic variance $(D'\Sigma D)^{-1}/N$, where $D = \partial\mathbf{g}_i(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$, and $\Sigma = E\{\mathbf{g}_i(\boldsymbol{\beta})\mathbf{g}_i(\boldsymbol{\beta})\}$. The asymptotic variance of the QIF estimator is optimal among the estimators for solving the same class of estimating functions. This implies that the QIF estimator is more efficient than the GEE estimators under the same misspecfied working correlation structure. If the working correlation structure is correctly specified, both the GEE and the QIF approaches are efficient. In practice, since the true correlation structure is typically not known, the efficiency gain under the misspecification makes the QIF estimator more attractive than the GEE estimator.

In addition, the QIF estimators have other advantages comparing with the GEE approach: (i) The QIF provides a statistical inference function under the same model assumption as the GEE, and this can be applied for model checking and testing. In the estimating equation approach, the validity of the first moment condition is crucial. However, it is difficult to develop a goodness-of-fit test through the GEE approach because the GEE does not have an objective function. (ii) The QIF is more robust against outliers than the GEE approach. Qu and Song (2004) show that the QIF has a bounded influence function, while the influence function of GEE is not bounded. (iii) The QIF is analog to minus twice the log-likelihood, which implies that the model selection criteria similar to the AIC and BIC can be developed (Wang and Qu, 2009). (iv) This also leads to the development of the penalized QIF approach, where a penalty function, such as SCAD, can be added to the QIF (Xue et al., 2010). The advantage of the penalization approach is that the relevant covariates can be identified simultaneously.

1.2.2 Review of Methods on Mixed-Effects Approach

Another major approach for handling longitudinal data is the mixed-effects model if the subject variation is also one of the interests. The conditional mean of the response can be formulated as

$$E(\mathbf{y}_i|\mathbf{X}_i, \mathbf{Z}_i, \mathbf{b}_i) = g(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i), \tag{1.5}$$

where g is the inverse of the link function, and $\mathbf{Z}_i$ are the covariates associated with the random effects parameters $\mathbf{b}_i$.

Wedderburn (1974) proposed the quasi-likelihood approach when the likelihood function is not available, as this approach only requires the first two moments. The quasi-likelihood function is defined as

$$L = \frac{1}{\sqrt{(2\pi)^q|\mathbf{D}|}} \int_{R^q} \exp\left\{-\frac{1}{2\phi}\sum_{i=1}^{N} d_i(\mathbf{y}_i, \boldsymbol{\mu}_i^{\mathbf{b}}) - \frac{1}{2}\mathbf{b}'\mathbf{D}^{-1}\mathbf{b}\right\} d\mathbf{b}, \tag{1.6}$$

where $\boldsymbol{\mu}_i^{\mathbf{b}} = E(\mathbf{y}_i|\mathbf{b})$, $\mathbf{D}$ is the variance component matrix for the random effects b, the weighted deviance function is

$$d_i(\mathbf{y}, \mathbf{u}) = -2\int_{\mathbf{y}}^{\mathbf{u}} \frac{\mathbf{y} - \mathbf{u}}{a_i v(\mathbf{u})} d\mathbf{u},$$

where a_i is a known weight, ϕ is a dispersion parameter and $v(\mathbf{u})$ is a variance function.

However, the quasi-likelihood approach requires high-dimensional integration, and (1.6) does not have a closed form if the correlation of longitudinal data is taken into account. Breslow and Clayton (1993) apply a Laplace approximation to the quasi-likelihood to develop the penalized quasi-likelihood (PQL) approach. The PQL is defined as

$$\text{PQL} = -\frac{1}{2\phi}\sum_{i=1}^{N}\sum_{j=1}^{T} d_{ij}(y_{ij}, \mu_{ij}^b) - \frac{1}{2}\mathbf{b}'\mathbf{D}^{-1}\mathbf{b}, \tag{1.7}$$

where $-\frac{1}{2}\mathbf{b}'\mathbf{D}^{-1}\mathbf{b}$ can also be treated as a penalty function in (1.7) (McCulloch and Searle, 2001). the drawbacks of the PQL approach are that the serial correlation is still ignored. Moreover, the PQL estimators are inconsistent for non-normal response (Booth and Hobert, 1999).

Vonesh et al. (2002) extended the second order GEE approach (Prentice and Zhao, 1991) to the generalized mixed-effects model. They incorporate the estimating equations for the conditional mean and the conditional variance in order to improve the efficiency of the fixed-effects and random-effects estimators. The advantage of their approach is that it leads to consistent estimators without specifying the likelihood function. However, the CGEE2 is not able to incorporate the serial correlation either. In addition, both the normality assumption for the random effects and the estimation of the variance components are required.

Jiang (1999) proposed the penalized generalized weighted least square (PGWLS) approach, which does not require the normality assumption of the random effects. This is achieved by adding a constraint $\mathbf{P_A b} = 0$ to ensure that the fixed effects and the random effects are identifiable through a Lagrange multiplier. The Lagrange function is constructed as,

$$l_q = -\frac{1}{2\phi}\sum_{i=1}^{N} d_i(\mathbf{y}_i, \boldsymbol{\mu}_i^{\mathbf{b}}) - \frac{1}{2}\lambda|\mathbf{P_A b}|^2. \tag{1.8}$$

However, the PGWLS does not incorporate the serial correlation either in their model, since the first part in (1.8) does not have a closed form if there is serial correlation present.

In general, the current mixed-effects model approaches can not incorporate the serial correlation, or require normality assumption on the random effects. However, these could be rather restrictive. First, it is not convincing that the random effects must be normally distributed. The normality assumption is required mainly for convenience. More importantly, misspecifying the random effects distribution could lead to inconsistent estimators of the fixed effects for binary responses, as is indicated in our simulation studies in Chapter 2. Furthermore, random effects cannot account for all the within-subject correlation, and modeling serial correlation for the residuals is necessary to improve the efficiency of the estimators. It is important to develop a new mixed-effects models approach which can incorporate serial correlation without any distributional assumption for the random effects.

1.3 The New Proposed Approaches

Chapter 2 of this thesis focuses on the new approach for the mixed-effects model, which incorporates correlation from both the random effects variation and the serial correlation. In this approach, we develop conditional extended scores for fixed and random effects parameters, and construct the objective function to incorporate the correlation structure. Because the objective function only involves the first two conditional moments, the likelihood function does not need to be specified. Moreover, we do not involve any integration in the estimation procedures, and this makes it computationally more feasible when the dimension of random effects is high. In addition, we do not impose any distributional assumption on the random effects as is the PQL and the CGEE2 approaches, and the new approach does not require us to estimate any variance component or nuisance parameters for the correlation.

We derive the asymptotic theory of $\sqrt{N}$-consistency and asymptotic normality of the fixed-effects estimators. The derivation of the asymptotic properties does not depend on the distributional assumption of the random effects, nor the consistency of the random effects estimator. These assumptions are typically

required in other existing approaches such as the CGEE2 and PQL. Instead, we only require that the expectation of the conditional extended scores converges to 0 in probability. We show that in order to prove the consistency of the random effects estimator, the correlation structure of the observations within a subject is required to have a mixingale condition (Andrews, 1988).

Our numerical studies show that when the serial correlation is introduced into the true model, the new approach greatly outperforms the most commonly used methods, such as the PQL, SAS GLIMMIXED and SAS NLMIXED. The data examples 1 and 2 are also analyzed to illustrate the new approach.

In Chapter 3 of this thesis, we develop a new approach to identify the correlation structure of longitudinal data when the cluster size diverges with the sample size. To the author's best knowledge, no existing work has addressed this problem for non-normal longitudinal data for diverging cluster size. Most current literature focuses on the estimation of the covariance matrix or the inverse of the covariance matrix rather than the selection of correlation structure. In our approach, we approximate the inverse of the correlation matrix with a linear combination of group basis matrices. We use the Euclidean distance between two estimation functions based on the empirical correlation matrix and a model-based approximation to assess the adequacy of the approximated model. By adding a penalty function, the problem can be transformed into a penalized least square problem.

The penalty function and the selection criteria are rather different compared to Zhou and Qu (2011). More importantly, we allow the cluster size to diverge as sample size increases. Note that the generalization is not trivial here because of the challenges in deriving the asymptotic properties and the computational complexity involved. We prove that under typical regularity conditions, the estimator of the coefficients corresponding to the basis matrices enjoys the oracle property (Fan and Li, 2001). In our approach, the estimating equations based on the empirical correlation are not independent, and this makes the model selection problem here quite different from a typical penalized least square approach.

Chapter 2

Mixed-Effects Modelling

2.1 Introduction

Longitudinal data arise frequently in many studies where measurements are obtained from a subject repeatedly over time. Consequently, measurements within a subject are correlated. Major statistical models such as marginal models and mixed-effects models have been developed for longitudinal data. Marginal models are applicable when the inference of the population average is of interest. One widely used marginal model is generalized estimating equations (GEE) (Liang and Zeger, 1986), which only requires the first two moments of the distribution. The advantage of the GEE is that it provides consistent estimators for regression parameters regardless of whether the working correlation is correctly specified or not.

In contrast, mixed-effects models are able to incorporate random-effect variation and have a richer interpretation when the subject-specific effect is one of the interests. However, in most of the current mixed-model literature (Breslow and Clayton, 1993; Laird and Ware, 1982; Jiang, 1999), the cluster correlations are assumed to be induced by the random effects only; that is, conditional on the random effects, the observations within a cluster are assumed to be independent. Although the generalized linear mixed-model approach is capable of incorporating serial correlation using numerical integration to maximize the likelihood or pseudo-likelihood, it is infeasible in practice when the dimension of random effects is high and the random effects are also correlated.

The major drawback to subject-specific approaches is that the random effects are assumed to follow an explicit distribution, and typically a normal random-effect distribution is assumed. Neuhaus et al. (1992) show that when the distributions of the random effects are misspecified, the estimators of the fixed effects could be inconsistent for binary data. Even in cases where the estimation of the fixed effects appears robust with a misspecified distribution assumption for random effects, the predicted distribution of the random effects may be invalid (Zhang et al., 2008). This could be critical if the prediction of the random effects is one of the main interests.

When the likelihood function has an explicit form and is tractable, the maximum likelihood approach

is applicable. However, for non-Gaussian outcomes, the likelihood function of the generalized linear mixed model (GLMM) (Breslow and Clayton, 1993; McCulloch, 1997; McCulloch and Neuhaus, 2001) often involves high-dimensional integrations and may be intractable. Numerical integration such as the Gaussian-Hermite quadrature (Liu and Pierce, 1994) and the Monte Carlo EM algorithm (McCulloch, 1997) could be computationally intensive, and infeasible in practice when the number of random effects is large (Zhang et al., 2008).

The penalized quasi-likelihood (PQL) (Breslow and Clayton, 1993) and conditional second-order generalized estimating equations (CGEE2) (Vonesh et al., 2002) are alternative approaches when the likelihood does not have a specific form. The PQL requires the estimation of the unknown variance components and typically ignores serial correlation from repeated measurements. In addition, the PQL is known to produce inconsistent estimators of the generalized linear mixed-model parameters (Booth and Hobert, 1999). The CGEE2 extends the second-order GEE (Prentice and Zhao, 1991) to generalized linear mixed models. It obtains estimators of fixed and random effects by solving estimating equations associated with the conditional mean and covariance matrix.

However, the CGEE2 requires estimation of the nuisance parameters associated with the working correlations and the variance components associated with the random effects. Jiang and Zhang (2001) propose a two-step estimation procedure which only requires the base statistics associated with the random effects, instead of the full distribution of the response. However, the covariance matrix of the base statistics must be known in order to improve the efficiency of the estimators in the second step. Furthermore, the normality distribution for the random effects is assumed for all of the above methods. This could be restrictive if the normality assumptions fails.

Jiang (1999) proposed a conditional inference approach which relaxes the normality assumptions for the random effects. However, conditional on random effects, the serial correlation of the responses is not taken into consideration. Molenberghs and Verbeke (2005, Chapter 22) indicate that it is important to model serial correlation for the random effects model. In general, for non-normal random effects models, incorporating correlation information is still quite challenging as existing approaches do not provide a feasible solution to handle serial correlation theoretically and computationally.

In this chapter, we develop a mixed-effect estimating equation approach which incorporates both random-effect variation and serial correlations simultaneously from repeated measurements. We do this through constructing conditional extended scores associated with the fixed and random effects by incorporating correlation structures. The proposed approach can be applied to both Gaussian and categorical data. We estimate the fixed and random effects using conditional extended scores which only involve the first and

second conditional moments. Therefore, the specification of the likelihood is not required. Furthermore, our approach does not involve intractable integrations and computationally it is feasible to implement for the GLMM. In addition, there are no distribution assumptions such as normality for random effects; instead, we allow the dimension of the random-effect parameters to increase as the sample size increases. The proposed approach enables one to incorporate multiple sources of variation from random effects and serial correlations. Moreover, it does not require the estimation of unknown variance components as in the PQL and CGEE2, and it does not require the estimation of the nuisance parameters associated with the working correlations as in the CGEE2.

We also establish root-N consistency and asymptotic normality for the fixed-effect parameter estimators. Existing approaches such as the CGEE2 and PQL require that the random-effect estimator be consistent with the true random effect; however, in practice, this assumption is difficult to verify. Our asymptotic results for fixed-effect estimators do not require such assumptions. We only require that the expectation of the estimating function conditional on the estimated random effects converges to 0 in probability. This assumption is more general than the consistency of the random-effect estimator. On the other hand, if the consistency of the random-effect estimator holds, then the asymptotic variance of the fixed-effect estimator has a closed form. To establish the consistency of the random-effect estimator, we only require mild conditions of mixingale (Andrews, 1988) on correlated observations.

This chapter is organized as follows: Section 2.2 proposes the conditional mixed-effects model using the quadratic inference function. Section 2.3 provides asymptotic properties for the proposed estimators. Section 2.4 illustrates simulation results for binary responses. Section 2.5 demonstrates real data examples by comparing the conditional mixed effects approach to the PQL, the generalized linear mixed model using SAS GLIMMIX, and the maximum likelihood approach using SAS NLMIXED. Discussion and concluding remarks are provided in Section 2.6. We provide the proofs of the asymptotic properties in Section 2.7.

2.2 Quadratic Inference Functions for Mixed-Effects Models

In this section, we will first give a brief description of the quadratic inference function for fixed-effects models. Then we will demonstrate how to incorporate both random and fixed effects for longitudinal data where there are multiple sources of variation.

2.2.1 Quadratic Inference Functions for Fixed Models

Consider the marginal model

$$E(\mathbf{y}_i) = \mathbf{g}(\mathbf{X}_i\boldsymbol{\beta}), \quad i = 1, \ldots, N,$$

where $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT})'$, $\mathbf{g}$ is a known function, and $\mathbf{X}_i$ is a known $T \times p$ matrix associated with a p-dimensional vector of fixed effects $\boldsymbol{\beta}$.

Liang and Zeger (1986) proposed the generalized estimating equation (GEE) which extends the quasi-likelihood equation

$$\sum_{i=1}^{N} \dot{\boldsymbol{\mu}}_i' \mathbf{V}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) = 0$$

by assuming $\mathbf{V}_i = \mathbf{A}_i^{1/2}\mathbf{R}\mathbf{A}_i^{1/2}$, where $\mathbf{A}_i$ is a diagonal marginal variance matrix and R is a working correlation matrix which involves correlation parameters. Note that $\mathbf{R}$ could also be an identity matrix if the working correlation is independent. We define $\boldsymbol{\mu}_i = E(\mathbf{y}_i)$, $\mathbf{V}_i = \text{var}(\mathbf{y}_i)$, and $\dot{\boldsymbol{\mu}}_i$ is the first derivative of $\boldsymbol{\mu}_i$ with respect to $\boldsymbol{\beta}$. The GEE estimators are consistent and asymptotically normal even if the working correlation matrix is misspecified. However, the estimator of the regression parameters is not efficient under the misspecification of the working correlation.

Qu et al. (2000) proposed the quadratic inference function to improve efficiency for longitudinal data. Their approach only requires the first two moments of the distribution. In addition, it takes correlation into account without estimating the nuisance parameters associated with the correlation structure. The main idea of their approach is to assume that the inverse of the working correlation $\mathbf{R}^{-1}$ is approximated by a class of linear combinations of known matrices $\mathbf{M}_1, \ldots, \mathbf{M}_m$, that is $\mathbf{R}^{-1} \approx \sum_{j=1}^{m} a_j\mathbf{M}_j$, where $\mathbf{M}_1$ is usually an identity matrix. Then the GEE can be approximated by

$$\sum_{i=1}^{N} \dot{\boldsymbol{\mu}}_i' \mathbf{A}_i^{-1/2} \left(\sum_{j=1}^{m} a_j \mathbf{M}_j \right) \mathbf{A}_i^{-1/2} (\mathbf{y}_i - \boldsymbol{\mu}_i) = 0.$$

Qu et al. (2000) defined the extended scores to be

$$\mathbf{G}_N(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{g}_i(\boldsymbol{\beta}) = \frac{1}{N} \begin{pmatrix} \sum_{\mathrm{i}=1}^{\mathrm{N}} \dot{\boldsymbol{\mu}}_\mathrm{i}' \mathbf{A}_\mathrm{i}^{-1/2} \mathbf{M}_1 \mathbf{A}_\mathrm{i}^{-1/2} (\mathbf{y}_\mathrm{i} - \boldsymbol{\mu}_\mathrm{i}) \\ \vdots \\ \sum_{\mathrm{i}=1}^{\mathrm{N}} \dot{\boldsymbol{\mu}}_\mathrm{i}' \mathbf{A}_\mathrm{i}^{-1/2} \mathbf{M}_\mathrm{m} \mathbf{A}_\mathrm{i}^{-1/2} (\mathbf{y}_\mathrm{i} - \boldsymbol{\mu}_\mathrm{i}) \end{pmatrix}, \tag{2.1}$$

where $\mathbf{G}_N$ is a mp-dimensional vector. Note that the GEE is a linear combination of extended scores $\mathbf{G}_N$.

The extended scores in (3.3) contain more estimating equations than unknown parameters if $m > 1$,

where the working correlation matrix $\mathbf{R}$ is not an identity matrix. Qu et al. (2000) adopted the idea of the generalized method of moments (Hansen, 1982) and proposed the quadratic inference function (QIF) to optimally combine the estimating equations in (3.3), and estimate the parameters $\boldsymbol{\beta}$ defined in (3.3). The QIF is defined as

$$Q_N(\boldsymbol{\beta}) = N\mathbf{G}_N'\mathbf{C}_N^{-1}\mathbf{G}_N,$$

where $\mathbf{C}_N$ is a $mp \times mp$ matrix and can be estimated consistently by $\mathbf{C}_N = \frac{1}{N}\sum_{i=1}^{N}\mathbf{g}_i(\boldsymbol{\beta})\mathbf{g}_i(\boldsymbol{\beta})'$. Here N must be greater than mp to ensure the invertibility of the variance matrix $\mathbf{C}_N$.

If the longitudinal data is unbalanced, we apply the transformation matrix to each cluster as follows. We create the largest cluster with a size T which contains time points for all possible measurements, and assume that fully observed clusters contain T observations. We define the $T \times T_i$ transformation matrix Λ_i for the ith cluster by removing the columns of the identity matrix corresponding to the missing observations. We define $\mathbf{y}_i^* = \boldsymbol{\Lambda}_i\mathbf{y}_i$, $\boldsymbol{\mu}_i^*(\tilde{\boldsymbol{\beta}}) = \boldsymbol{\Lambda}_i\boldsymbol{\mu}_i(\tilde{\boldsymbol{\beta}})$, $\dot{\boldsymbol{\mu}}_i^*(\tilde{\boldsymbol{\beta}}) = \boldsymbol{\Lambda}_i\dot{\boldsymbol{\mu}}_i(\tilde{\boldsymbol{\beta}})$, and $\mathbf{A}_i^* = \boldsymbol{\Lambda}_i\mathbf{A}_i$, where components in $\mathbf{y}_i^*$ are the same as in $\mathbf{y}_i$ for non-missing responses but are 0 for the missing responses; and similarly define $\boldsymbol{\mu}_i^*$ and $\dot{\boldsymbol{\mu}}_i^*$. The marginal variance for $\mathbf{A}_i^*$ is 0 for the missing observations. In the following methodology development, we assume that the cluster size is equal for notational convenience.

Generalized linear mixed models (GLMM) extend the linear mixed model (Laird and Ware, 1982) for non-normal longitudinal data via a specific link function. For a link function $\mathbf{g}$, the conditional mean $E(\mathbf{y}|\mathbf{b}) = \boldsymbol{\mu}^{\mathbf{b}}$ is a function of the linear predictor $\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}$ with $\mathbf{g}(\boldsymbol{\mu}^{\mathbf{b}}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}$, where $\mathbf{Z}$ is the covariate associated with random effect $\mathbf{b}$.

In the GLMM, if the conditional likelihood of $\mathbf{y}$ given $\mathbf{b}$ is unknown, we can apply the quasi-likelihood (Wedderburn, 1974) which only requires the first two moments. The integrated quasi-likelihood is defined by

$$L = \frac{1}{\sqrt{(2\pi)^q|\mathbf{D}|}}\int_{R^q}\exp\left\{-\frac{1}{2\phi}\sum_{i=1}^{N}d_i(\mathbf{y}_i,\boldsymbol{\mu}_i^{\mathbf{b}}) - \frac{1}{2}\mathbf{b}'\mathbf{D}^{-1}\mathbf{b}\right\}d\mathbf{b}, \tag{2.2}$$

where $\boldsymbol{\mu}_i^{\mathbf{b}} = E(\mathbf{y}_i|\mathbf{b})$, $\mathbf{D}$ is the variance component matrix for random effect b, the weighted deviance function

$$d_i(\mathbf{y},\mathbf{u}) = -2\int_{\mathbf{y}}^{\mathbf{u}}\frac{\mathbf{y}-\mathbf{u}}{a_v(\mathbf{u})}d\mathbf{u}$$

with a_i as a known weight, ϕ is a dispersion parameter and $v(\mathbf{u})$ is a variance function. However, the quasi-likelihood (2.2) does not have a closed form and it is not tractable if $\mathbf{y}$ is not normal. (Breslow and Clayton, 1993) proposed the penalized quasi-likelihood which applies a Laplace approximation to solve the integrated quasi-likelihood for GLMM.

The penalized quasi-likelihood can be written as

$$\text{PQL} = -\frac{1}{2\phi}\sum_{i=1}^{N}\sum_{j=1}^{T} d_{ij}(y_{ij}, \mu_{ij}^b) - \frac{1}{2}\mathbf{b}'\mathbf{D}^{-1}\mathbf{b}, \tag{2.3}$$

where $-\frac{1}{2}\mathbf{b}'\mathbf{D}^{-1}\mathbf{b}$ can also be treated as a penalty to the log-quasi-likelihood (McCulloch and Searle, 2001). The corresponding two sets of quasi-score equations are derived by taking the derivatives of PQL with respect to fixed effects $\boldsymbol{\beta}$ and random effects $\mathbf{b}$ as follows:

$$\sum_{i=1}^{N}(\frac{\partial \boldsymbol{\mu}_i^{\mathbf{b}}}{\partial \boldsymbol{\beta}})'(\mathbf{W}_i^{\mathbf{b}})^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i^{\mathbf{b}}) = 0 \tag{2.4}$$

and

$$\sum_{i=1}^{N}(\frac{\partial \boldsymbol{\mu}_i^{\mathbf{b}}}{\partial \mathbf{b}})'(\mathbf{W}_i^{\mathbf{b}})^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i^{\mathbf{b}}) - \boldsymbol{D}^{-1}\mathbf{b} = 0, \tag{2.5}$$

where $\mathbf{W}_i^b = \text{var}(\mathbf{y}_i|\mathbf{b})$, the covariance matrix $\mathbf{D} = \mathbf{D}(\boldsymbol{\theta})$, and $\boldsymbol{\theta}$ is an unknown vector of the variance components. Breslow and Clayton (1993) applied Green (1987) Fisher scoring algorithm iteratively to solve equations (2.4) and (2.5).

The penalized quasi-likelihood requires the normality assumption for random effects and the estimation of variance components. It also ignores serial correlation from repeated measurements. In the following, we propose a conditional inference function for the GLMM which does not require variance components estimation or a normality assumption for the random effects, and yet is still able to incorporate serial correlation.

In the GLMM, the conditional quasi-log-likelihood of $\mathbf{y}$ with conditional mean $\boldsymbol{\mu}_i^{\mathbf{b}}$ is

$$l_q = -\frac{1}{2\phi}\sum_{i=1}^{N} d_i(\mathbf{y}_i, \boldsymbol{\mu}_i^{\mathbf{b}}). \tag{2.6}$$

Note that (2.6) is the first term of PQL in (2.3) regardless of the distribution of random effects. If we impose constraint $\mathbf{P_A b} = 0$ associated with the random effects, then we can ensure that the fixed effect $\boldsymbol{\beta}$ and random effect $\mathbf{b}$ are identifiable without requiring the distribution assumption of random effects. Here $\mathbf{P_A}$ is a known orthogonal matrix, and can be constructed as follows. For any vector space $\mathbf{V}$ and matrix $\mathbf{M}$, $\mathbf{B}(\mathbf{V}) = \{\mathbf{B} : \mathbf{B}$ is a matrix whose columns constitute bases for $\mathbf{V}\}$ and $\mathcal{N}(\mathbf{M})$ is the null space of $\mathbf{M}$, e.g., $\mathcal{N}(\mathbf{M}) = \{\mathbf{v} : \mathbf{Mv} = 0\}$. Then the matrix of $\mathbf{P_A} = \mathbf{A}(\mathbf{A}'\mathbf{A})^{-\mathbf{1}}\mathbf{A}'$ is a projection matrix where $\mathbf{A} \in \mathcal{B}(\mathcal{N}\{(\mathbf{I} - \mathbf{P_X})\mathbf{Z}\})$. For example, for a linear mixed-effects model with a random intercept and covariate x_{ij}, $E(\mu_{ij}^b) = \alpha_0 + \alpha_1 x_{ij} + b_i$. It can be shown that $\mathbf{P_A} = \frac{1}{N}\mathbf{1}_N\mathbf{1}_N'$, where $\mathbf{1}_N$ is a $N \times 1$ vector with entries

1, therefore the constraint associated with the random effects is $\frac{1}{N}\sum b_i = 0$ (Jiang, 1999).

The estimators of fixed effects and random effects can be obtained by maximizing (2.6) subject to $\mathbf{P_A b} = 0$. In order to achieve this, the penalized generalized weighted least square (PGWLS)(Jiang, 1999) can be utilized to build a Lagrange function

$$l_q = -\frac{1}{2\phi}\sum_{i=1}^{N} d_i(\mathbf{y}_i, \boldsymbol{\mu}_i^{\mathbf{b}}) - \frac{1}{2}\lambda|\mathbf{P_A b}|^2, \tag{2.7}$$

where λ is a Lagrange multiplier. The last term of (2.7) can be viewed as a penalizer and (2.7) can be viewed as a penalized quasi-log-likelihood (Jiang, 1999).

The quasi-score equations corresponding to $\boldsymbol{\beta}$ and $\mathbf{b}_i$ can be derived from (2.6) and (2.7) as

$$\sum_{i=1}^{N}(\frac{\partial \boldsymbol{\mu}_i^{\mathbf{b}}}{\partial \boldsymbol{\beta}})'(\mathbf{W}_i^{\mathbf{b}})^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i^{\mathbf{b}}) = 0 \tag{2.8}$$

and

$$\left(\begin{array}{rcl} \mathbf{h}_1 & = & (\frac{\partial \boldsymbol{\mu}_1^{\mathbf{b}_1}}{\partial \mathbf{b}_1})'(\mathbf{W}_1^{\mathbf{b}})^{-1}(\mathbf{y}_1 - \boldsymbol{\mu}_1^{\mathbf{b}_1}) - \lambda\frac{\partial \mathbf{P_A b}}{\partial \mathbf{b}_1}\mathbf{P_A b} = 0 \\ & & \vdots \\ \mathbf{h}_N & = & (\frac{\partial \boldsymbol{\mu}_N^{\mathbf{b}_N}}{\partial \mathbf{b}_N})'(\mathbf{W}_N^{\mathbf{b}})^{-1}(\mathbf{y}_N - \boldsymbol{\mu}_N^{\mathbf{b}_N}) - \lambda\frac{\partial \mathbf{P_A b}}{\partial \mathbf{b}_N}\mathbf{P_A b} = 0 \end{array}\right). \tag{2.9}$$

In the PQL approach, the correlation structure conditional on the random effects is assumed to have an independent structure, therefore $\mathbf{W}_i^b$ is diagonal. In contrast to the PQL, we assume that conditional on $\mathbf{b}_i$, the measurements within the ith cluster can be correlated. Thus $\mathbf{W}_i^b = \mathrm{var}(\mathbf{y}_i|\mathbf{b}_i)$ is not necessarily a diagonal matrix.

Suppose the working correlation $\mathbf{R}$ in $\mathbf{W}_i^{\mathbf{b}} = \mathbf{A}_i^{\frac{1}{2}}\mathbf{R}\mathbf{A}_i^{\frac{1}{2}}$ has a linear approximation of several basis matrices, that is, $\mathbf{R}^{-1} \approx \sum_{j=1}^{m} a_j\mathbf{M}_j$, where $\mathbf{A}_i = \mathrm{diag}\{\mathrm{var}(\mathbf{y}_{i1}|\mathbf{b}), \ldots, \mathrm{var}(\mathbf{y}_{iT}|\mathbf{b})\}$. Based on (2.8), we define the conditional extended scores associated with the fixed effects, namely fixed-effects extended scores, as

$$\mathbf{G}_N^f = \frac{1}{N}\sum_{i=1}^{N}\mathbf{g}_i^f(\boldsymbol{\beta}) = \frac{1}{N}\left(\begin{array}{c} \sum_{\mathrm{i}=1}^{\mathrm{N}}(\frac{\partial \boldsymbol{\mu}_{\mathrm{i}}^{\mathrm{b}}}{\partial \boldsymbol{\beta}})'\mathbf{A}_{\mathrm{i}}^{-1/2}\mathbf{M}_1\mathbf{A}_{\mathrm{i}}^{-1/2}\left(\mathbf{y}_{\mathrm{i}} - \boldsymbol{\mu}_{\mathrm{i}}^{\mathrm{b}}\right) \\ \vdots \\ \sum_{\mathrm{i}=1}^{\mathrm{N}}(\frac{\partial \boldsymbol{\mu}_{\mathrm{i}}^{\mathbf{b}}}{\partial \boldsymbol{\beta}})'\mathbf{A}_{\mathrm{i}}^{-1/2}\mathbf{M}_{\mathrm{m}}\mathbf{A}_{\mathrm{i}}^{-1/2}\left(\mathbf{y}_{\mathrm{i}} - \boldsymbol{\mu}_{\mathrm{i}}^{\mathrm{b}}\right) \end{array}\right). \tag{2.10}$$

In addition, the quasi-score equations $\mathbf{h}_i$ in (2.9) can be represented as a linear combination of the elements in

$$\mathbf{g}_i^r = \left(\begin{array}{c} (\frac{\partial \boldsymbol{\mu}_{\mathrm{i}}^{\mathbf{b}_{\mathrm{i}}}}{\partial \mathbf{b}_{\mathrm{i}}})'\mathbf{A}_{\mathrm{i}}^{-1/2}\mathbf{M}_1\mathbf{A}_{\mathrm{i}}^{-1/2}(\mathbf{y}_{\mathrm{i}} - \boldsymbol{\mu}_{\mathrm{i}}^{\mathrm{b}}) \\ \vdots \\ (\frac{\partial \boldsymbol{\mu}_{\mathrm{i}}^{\mathbf{b}_{\mathrm{i}}}}{\partial \mathbf{b}_{\mathrm{i}}})'\mathbf{A}_{\mathrm{i}}^{-1/2}\mathbf{M}_{\mathrm{m}}\mathbf{A}_{\mathrm{i}}^{-1/2}(\mathbf{y}_{\mathrm{i}} - \boldsymbol{\mu}_{\mathrm{i}}^{\mathrm{b}}) \end{array}\right) \tag{2.11}$$

and the penalty term $\lambda\frac{\partial \mathbf{P_A b}}{\partial \mathbf{b}_i}\mathbf{P_A b}$. For example, in a simple random intercept model $\mathbf{P_A b} = (\sum_{i=1}^{N} b_i/N, \ldots, \sum_{i=1}^{N} b_i/N)'$ and $\frac{\partial \mathbf{P_A b}}{\partial \mathbf{b}_i}\mathbf{P_A b} = \sum_{i=1}^{N} b_i/N$.

In the PGWLS approach (Jiang, 1999), the only constraint imposed on the random effect is $\mathbf{P_A b} = 0$ for the purpose of identifiability. However, this constraint is not sufficient to ensure algorithm convergence on estimating the fixed and random effects. The convergence problem becomes more serious when there are more random effects involved in the model, and this is also confirmed by our numerical studies where PGWLS fails to converge and is extremely sensitive to the initial values of the estimators. To solve the convergence problem, we propose an alternative approach by including a penalty term $\lambda \mathbf{b}_i$ in addition to the extended score $\mathbf{g}_i^r$ for the random effects. This allows one to control the variance of fixed and random effects estimators effectively and ensures that the algorithm converges.

In most cases, correlation for the random-effects model in (2.11) is not as critical as the correlation for the fixed-effects modeling in (2.10). Therefore we just include the first term of (2.11). That is, $\mathbf{g}_i^r$ can be modified as $\mathbf{g}_i^{r*} = (\frac{\partial \boldsymbol{\mu}_i^{\mathbf{b}_i}}{\partial \mathbf{b}_i})'\mathbf{A}_i^{-1/2}\mathbf{M}_1\mathbf{A}_i^{-1/2}(\mathbf{y}_i - \boldsymbol{\mu}_i^{\mathbf{b}})$. We define the combined random-effects extended scores for $\mathbf{b} = (\mathbf{b}_1', \ldots, \mathbf{b}_N')'$ as

$$\mathbf{G}_N^r = \{(\mathbf{g}_1^{r*})', \lambda\mathbf{b}_1', \ldots, (\mathbf{g}_N^{r*})', \lambda\mathbf{b}_N', \lambda(\mathbf{P_A b})'\}'. \tag{2.12}$$

Note that we only need one $\lambda\mathbf{P_A b}$ for the combined extended scores since $\lambda\frac{\partial \mathbf{P_A b}}{\partial \mathbf{b}_i}\mathbf{P_A b}, i = 1 \ldots, N$, are linear combinations of $\lambda\mathbf{P_A b}$.

In addition to the penalty difference, the proposed approach also differs from the PGWLS by incorporating within-cluster serial correlation without involving estimation of the correlation parameters. Incorporating correlation can improve estimation efficiency significantly for the fixed-effect parameters when there is strong serial correlation present in longitudinal data.

The estimators of the fixed-effects and random-effects can be obtained by solving (2.10) and (2.12) iteratively. However, both (2.10) and (2.12) are over-identified in the sense that there are more equations than the dimension of the fixed-effect and random-effect parameters. We apply the quadratic inference function to achieve the estimation of the fixed-effect and random-effect parameters iteratively. That is, for given random-effects b, the fixed-effect estimators are obtained by minimizing

$$N(\mathbf{G}_N^f)'(\mathbf{C}_N^f)^{-1}(\mathbf{G}_N^f), \tag{2.13}$$

where $\mathbf{C}_N^f = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{g}_i^f)(\mathbf{g}_i^f)'$. In addition, for given fixed-effects $\boldsymbol{\beta}$, the random-effect estimators are obtained by minimizing

$$(\mathbf{G}_N^r)'(\mathbf{G}_N^r). \tag{2.14}$$

Note that in (2.14) the identity matrix is used as the weighting matrix for $\mathbf{g}_i^{r*}$, since the random-effect parameter is subject-specific and there are no replicates across different clusters.

2.2.2 Implementation

For a fixed λ, the iterative estimating procedure is summarized as follows.

1. Start with an initial vector $\hat{\boldsymbol{\beta}}$ obtained from the generalized linear model assuming independent correlation structure and set the initial $\mathbf{b} = 0$.
2. Replace $\boldsymbol{\beta}$ in (2.12) with $\hat{\boldsymbol{\beta}}$ and obtain the random-effects estimator $\hat{\mathbf{b}}$ by minimizing (2.14).
3. Replace $\mathbf{b}$ with $\hat{\mathbf{b}}$ in (2.10). Then update $\hat{\boldsymbol{\beta}}$ by minimizing (2.13).

Iterate step 2 and step 3 until the convergence criterion

$$|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}| + |\hat{\mathbf{b}} - \mathbf{b}| < \varepsilon$$

is reached, where ε is a small positive number and is typically chosen as 10^{-6}.

In the above algorithm, it is important to select a Lagrange multiplier λ. Once λ is obtained, both fixed and random-effects parameters can be estimated through minimizing (2.13) and (2.14), and they are functions of λ. Notice that λ also plays a role similar to the tuning parameter in the penalty term for model selection. We can choose $\lambda = \log(N)$ which works quite effectively in our numerical studies. Alternatively, we can use the consistent BIC-type model selection criterion to search λ, where the objective function

$$N(\mathbf{G}_N^f)'(\mathbf{C}_N^f)^{-1}(\mathbf{G}_N^f) + (\log N)(\mathbf{P}_\mathbf{A}\mathbf{b})'\boldsymbol{\Sigma}_\mathbf{b}^{-1}(\mathbf{P}_\mathbf{A}\mathbf{b}) \tag{2.15}$$

reaches minimum. Here $\boldsymbol{\Sigma}_\mathbf{b}$ is the covariance matrix of $\mathbf{P}_\mathbf{A}\mathbf{b}$. The first part of (2.15) is analog to minus twice the log-likelihood and the second part is analog to the BIC-type of penalty term for the number of regression parameters. This selection criterion can be interpreted so as to balance the minimum quadratic inference function for the fixed-effects, and the variability of the constrained random-effects

Note that our approach differs from (Jiang, 1999) in that the objective function in (2.13) is different, and is able to incorporate serial correlation of the repeated measurements. In addition, the penalty term for the proposed conditional inference approach is very different from the PGWLS, since the PGWLS only penalizes constraints of the random effects parameter to ensure identifiability, while our approach also penalizes the random effects with large variances. Computationally, the new algorithm is much more stable and fast since the variance of random effects can be regularized.

2.3 Asymptotic Properties

In this section, we investigate the asymptotic properties of the mixed-effects QIF estimators. We define the true random effects of the whole population as $\mathbf{b}_{0,\infty} = (\mathbf{b}_{01}, \ldots, \mathbf{b}_{0N}, \ldots)$, where $\mathbf{b}_{0i}$ are the true random effects for subject i. Here $\mathbf{b}_{0,\infty}$ are treated as fixed parameters rather than random variables. Moreover, for a fixed sample size N, denote the true realization of the random effects as $\mathbf{b}_0^N = (\mathbf{b}_{01}, \ldots, \mathbf{b}_{0N})$, since only N random effects among $\mathbf{b}_{0,\infty}$ are involved. When the sample size $N \to \infty$, $\mathbf{b}_0^N$ is equivalent to $\mathbf{b}_{0,\infty}$. In the following, we would write $\mathbf{b}_0 = \mathbf{b}_0^N$ for notational convenience, and let $\hat{\mathbf{b}} = (\hat{\mathbf{b}}_1, \ldots, \hat{\mathbf{b}}_N)$ be the estimator of the random effects $\mathbf{b}_0^N$.

We denote $\boldsymbol{\beta}_0$ as the true parameter of the fixed effects, $\hat{\boldsymbol{\beta}}_1$ as the estimator of the fixed-effects obtained by minimizing the QIF conditional on the random effects estimator $\hat{\mathbf{b}}$, and $\hat{\boldsymbol{\beta}}_0$ as the fixed-effect estimator obtained by minimizing the QIF conditional on the true random-effects $\mathbf{b}_0$. The proofs of the following Lemmas and Theorem 1 are provided in Section 2.7.

Lemma 1. *Under the regularity conditions provided in A.2, there exists a minimizer $\hat{\boldsymbol{\beta}}_1$ of $Q(\boldsymbol{\beta}|\hat{\mathbf{b}})$ for some fixed λ, such that $\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_0 = O_p(N^{-1/2})$.*

Consistency of $\hat{\boldsymbol{\beta}}_1$ follows immediately from Lemma 1. Here we do not assume the consistency of the random effects estimator $\hat{\mathbf{b}}$ in order to ensure the consistency of the fixed-effect estimator $\hat{\boldsymbol{\beta}}_1$. Instead we only require Condition 6 in A.2, that is, $E[E\{\mathbf{g}(\boldsymbol{\beta}_0|\hat{\mathbf{b}})\}]$ converges in probability to $E\{\mathbf{g}(\boldsymbol{\beta}_0|\mathbf{b}_0)\}$, which is equivalent to 0 by Condition 3 in A.2. This is a rather weaker condition than the consistency of $\hat{\mathbf{b}}$, since the consistency of $\hat{\mathbf{b}}$ implies Condition 6; however, Condition 6 does not necessarily guarantee the consistency of $\hat{\mathbf{b}}$.

Lemma 2 and Lemma 3 provide the consistency and asymptotic distribution of $\hat{\boldsymbol{\beta}}_0$.

Lemma 2. *The estimator of fixed-effect $\hat{\boldsymbol{\beta}}_0$ is consistent, and has a rate of root-N convergence as $N \to \infty$, that is, $\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0 = O_p(N^{-1/2})$.*

Lemma 3. *Under the regularity conditions provided in A.2., $\sqrt{N}(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0) \xrightarrow{d} N(0, \boldsymbol{\Omega}_0)$ as $N \to \infty$, where $\ddot{Q}^{-1}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_0|\mathbf{b}_0) \to_p \boldsymbol{\Omega}_0$ and $\ddot{Q}^{-1}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_0|\mathbf{b}_0)$ is the inverse of the second derivative of the conditional quadratic inference function with respect to $\boldsymbol{\beta}$ at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_0$ and $\mathbf{b} = \mathbf{b}_0$.*

Theorem 1. *Under the regularity conditions provided in A.2., for a fixed λ, $\hat{\boldsymbol{\beta}}_1$ has the following asymptotic properties as $N \to \infty$.*

I. (Consistency) $\hat{\boldsymbol{\beta}}_1 \to_p \boldsymbol{\beta}_0$.

II. (Asymptotic Normality) $\sqrt{N}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_0) \xrightarrow{d} N(0, \boldsymbol{\Omega}_1)$*, where* $\boldsymbol{\Omega}_1$ *is provided in (A-7). If each component of* $\hat{\mathbf{b}}$ *is consistent i.e.* $\hat{\mathbf{b}}_i \to_p \mathbf{b}_{0i}$ *for* $i = 1, \ldots, N$*, then* $\ddot{Q}^{-1}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_1|\hat{\mathbf{b}}) \to_p \boldsymbol{\Omega}_1$*.*

Note that $\boldsymbol{\Omega}_1$ does not have a closed form if $\hat{\mathbf{b}}$ is not a consistent estimator of $\mathbf{b}_0$, although the variance estimator of $\boldsymbol{\Omega}_1$ can be obtained through bootstrap sampling.

Theorem 1 also indicates that if a consistent estimator of the random effects $\hat{\mathbf{b}}$ is obtained, the covariance matrix $\boldsymbol{\Omega}_1$ can be approximated by $\ddot{Q}^{-1}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_1|\hat{\mathbf{b}})$, since it converges to $\boldsymbol{\Omega}_0$ as the sample size of clusters N and the cluster size n both go to infinity. In addition, the weighting matrix $\mathbf{C}_N^f$ defined in (2.13) is optimal in the estimation of $\boldsymbol{\beta}_0$ due to the asymptotic efficiency property of the generalized method of moments (Hansen, 1982). Therefore $\hat{\boldsymbol{\beta}}_0$ is efficient as the covariance matrix of $\sqrt{N}(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0)$ reaches the minimum in the sense of Loewner ordering. Since $\hat{\boldsymbol{\beta}}_1$ is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_0$, as the asymptotic variance of $\hat{\boldsymbol{\beta}}_1$ converges to the asymptotic variance of $\hat{\boldsymbol{\beta}}_0$ if the random effect is consistently estimated, therefore the estimator $\hat{\boldsymbol{\beta}}_1$ also achieves the same asymptotic efficiency as $\hat{\boldsymbol{\beta}}_0$.

In A.4., we provide regularity conditions and a proof to achieve a consistent estimator of random-effect $\mathbf{b}_0$ for correlated data when the cluster size n goes to infinity. This type of condition is satisfied for correlated response such as autoregressive, stationary Gaussian, M-dependent and other sequences with a decaying correlation structure satisfying the mixingale condition.

However, if the estimator of the random effects $\hat{\mathbf{b}}$ is not consistent with $\mathbf{b}_0$, we show in A.3 that the asymptotic efficiency of $\hat{\boldsymbol{\beta}}_1$ will be affected since $\boldsymbol{\Sigma}^*$ in (A-6) cannot be guaranteed to simplify as $\boldsymbol{\Sigma} = \text{var}(\sqrt{N}\mathbf{G}_N)$, and consequently the asymptotic variance $\boldsymbol{\Omega}_1$ for the fixed-effect estimator might not necessarily reach the minimum asymptotic variance $\boldsymbol{\Omega}_0$.

2.4 Simulation

To evaluate the performance of the QIF method for GLMMs, we fit the model with the mixed-effects QIF method in Section 2.2 with three types of working correlations: independent, exchangeable and AR-1. We compare our approach to the PQL method using the R package lme4, the SAS GLIMMIX (Schabenberger, 2005) and SAS NLIMIXED. The GLIMMIX procedure fits the generalized linear mixed model based on linear approximation for the nonlinear function, and the NLIMXED procedure approximates the likelihood function using Laplace approximation. Note that the PQL, GLIMMIX and NLMIIXED all assume that the random effects follow the normal assumption. We cannot compare the PGWLS numerically because of its non- convergence problem.

We conduct simulation studies for binary responses here, since the random effects model for binary data

is the most challenging when the random effects do not follow a normal distribution. We consider the random effects model with both a random intercept and a random slope. The conditional correlated binary responses are generated using the logistic regression model

$$\text{logit}(\boldsymbol{\mu}_i^{\mathbf{b}}) = \beta_0 + b_{0i} + \mathbf{x}_i(\beta_1 + b_{1i}),\ \texttt{corr}(\mathbf{y}_i|\mathbf{x}_i, b_{0i}, b_{1i}) = \mathbf{R}, i = 1, \ldots, N,$$

where $\beta_0 = -0.3$, $\beta_1 = 0.3$, the covariate x_i's are generated from uniform (0.5, 1.5), and the sample size N is chosen to be either 50 or 100. The correlated data is generated with unequal cluster sizes, that is, the half of samples have cluster sizes either 4 or 5. The true correlation structures are independent, exchangeable or AR-1 with the correlation coefficient $\rho = 0.4$ or $\rho = 0.7$. Both the random intercept b_{0i} and the random slope b_{1i} are generated from a bimodal distribution with the probability density function $f(b/0.3 + 0.5)$ from a rescaled Beta(0.5, 0.5) distribution. We use R package mtvBinaryEP to generate the multivariate correlated binary data.

Table 2.1 - 2.4 provide the MSEs of the estimators and their standard errors for the fixed-effect β_0 and β_1 based on 200 simulations under two different correlations $\rho = 0.4$ and $\rho = 0.7$, with two different sample sizes. Results from the GLIMMIX procedure under exchangeable correlation structure are not presented here since the GLIMMIX does not converge in most cases.

When the true correlation structure is independent, the performances of the mixed-effect QIF, the PQL and the NLMIXED procedure are comparable, while there are serious convergence problems for the SAS GLIMMIX under independent structure with more than 90% of non-convergence rate. When the true correlation structure is either AR-1 or exchangeable, the MSEs of the mixed-effects QIF method are smaller than those obtained from the PQL, GLIMMIX and NLMIXED approaches, even under the misspecified working correlation structure. In general, the efficiency improvement becomes more significant as the correlation ρ increases. For example, when the sample size $N = 100$ and the true correlation structure is AR-1 with $\rho = 0.7$, the mixed-effect QIF estimator for the slope parameter β_1 using the true correlation structure has the lowest MSE of 0.0494, compared to the MSEs of 0.0979 and 0.0732 under the misspecified independent and exchangeable working correlation structures. The MSEs of the estimators from the PQL, the GLIMMIX under independent and AR-1 working structures, and the NLMIXED are 0.4354, 0.3509, 0.0755 and 0.3971, respectively. Note that the non-convergence rate is 174 out of 200 for the GLIMMIX under AR-1 working structure; therefore the MSE of 0.0755 is also questionable.

We also observe that the MSEs of the mixed-effect QIF estimators with correctly specified working correlation structure are smaller than those obtained with misspecified working correlation structures. This

Table 2.1: MSE and the standard errors of MSE (provided in the lower right corner) for the estimator of the intercept $\beta_0 = -0.3$ for binary responses when $\rho = 0.4$ from 200 simulations.

Method	$N = 50$ True correlation			$N = 100$ True correlation		
	Independent	Exchangeable	AR-1	Independent	Exchangeable	AR-1
QIF (ind)	$0.2180_{0.0220}$	$0.2568_{0.0260}$	$0.2496_{0.0208}$	$0.1211_{0.0115}$	$0.1173_{0.0106}$	$0.1389_{0.0144}$
QIF (exch)	$0.2345_{0.0243}$	$0.2492_{0.0287}$	$0.2742_{0.0224}$	$0.1236_{0.0118}$	$0.1058_{0.0095}$	$0.1361_{0.0141}$
QIF (AR-1)	$0.2359_{0.0251}$	$0.2657_{0.0271}$	$0.2443_{0.0226}$	$0.1244_{0.0118}$	$0.1118_{0.0101}$	$0.1269_{0.0135}$
PQL	$0.2242_{0.0226}$	$0.5211_{0.0537}$	$0.4933_{0.0459}$	$0.1258_{0.0118}$	$0.2762_{0.0277}$	$0.2215_{0.0237}$
GLIMMIX (Ind)	$0.1580^{\mathbf{1}}_{0.0145}$	$0.3870^{\mathbf{2}}_{0.0423}$	$0.4092^{\mathbf{3}}_{0.0357}$	$0.1476^{\mathbf{9}}_{0.0167}$	$0.2040^{\mathbf{10}}_{0.0187}$	$0.1616^{\mathbf{11}}_{0.0182}$
GLIMMIX (AR-1)	$0.1979^{\mathbf{4}}_{0.0188}$	$0.3959^{\mathbf{5}}_{0.0429}$	$0.2628^{\mathbf{6}}_{0.0202}$	$0.1189^{\mathbf{12}}_{0.0167}$	$0.2020^{\mathbf{13}}_{0.0193}$	$0.1224^{\mathbf{14}}_{0.0133}$
NLMIXED	$0.2193^{\mathbf{7}}_{0.0222}$	$0.4452_{0.0490}$	$0.4458^{\mathbf{8}}_{0.0383}$	$0.1212^{\mathbf{15}}_{0.0115}$	$0.2631_{0.0255}$	$0.1904^{\mathbf{16}}_{0.0200}$

Table 2.2: MSE and the standard errors of MSE (provided in the lower right corner) for the estimator of the slope $\beta_1 = 0.3$ for binary responses when $\rho = 0.4$ from 200 simulations.

Method	$N = 50$ True correlation			$N = 100$ True correlation		
	Independent	Exchangeable	AR-1	Independent	Exchangeable	AR-1
QIF (ind)	$0.1870_{0.0184}$	$0.2165_{0.0192}$	$0.2279_{0.0195}$	$0.1077_{0.0106}$	$0.0962_{0.0082}$	$0.1199_{0.0110}$
QIF (exch)	$0.1991_{0.0201}$	$0.2028_{0.0210}$	$0.2437_{0.0203}$	$0.1087_{0.0107}$	$0.0878_{0.0077}$	$0.1169_{0.0102}$
QIF (AR-1)	$0.2021_{0.0210}$	$0.2182_{0.0204}$	$0.2180_{0.0195}$	$0.1102_{0.0108}$	$0.0938_{0.0078}$	$0.1050_{0.0089}$
PQL	$0.1915_{0.0191}$	$0.4150_{0.0383}$	$0.4438_{0.0394}$	$0.1108_{0.0109}$	$0.2324_{0.0193}$	$0.1882_{0.0167}$
GLIMMIX (Ind)	$0.0915^{\mathbf{1}}_{0.0084}$	$0.3328^{\mathbf{2}}_{0.0325}$	$0.3872^{\mathbf{3}}_{0.0335}$	$0.1453^{\mathbf{9}}_{0.0160}$	$0.1805^{\mathbf{10}}_{0.0147}$	$0.1444^{\mathbf{11}}_{0.0134}$
GLIMMIX (AR-1)	$0.1195^{\mathbf{4}}_{0.0129}$	$0.3456^{\mathbf{5}}_{0.0336}$	$0.3281^{\mathbf{6}}_{0.0217}$	$0.1307^{\mathbf{12}}_{0.0161}$	$0.1737^{\mathbf{13}}_{0.0147}$	$0.1194^{\mathbf{14}}_{0.0110}$
NLMIXED	$0.1877^{\mathbf{7}}_{0.0186}$	$0.3857_{0.0368}$	$0.4236^{\mathbf{8}}_{0.0374}$	$0.1079^{\mathbf{15}}_{0.0107}$	$0.2415_{0.0209}$	$0.1722^{\mathbf{16}}_{0.0159}$

Note: Number of non-convergence outcomes from GLIMMIX procedures are tabulated as follows: **1.** 188; **2.** 60; **3.** 86; **4.** 183; **5.** 69; **6.** 166; **7.** 16; **8.** 1; **9.** 180; **10.** 35; **11.** 92; **12.** 180; **13.** 45; **14.** 176;**15.** 7; **16.** 2.

is especially true when the correlation parameter is as high as $\rho = 0.7$. On comparing the MSEs from two different sample sizes, the ratios between the MSEs of the mixed-effect QIF estimators under the true and misspecified correlation structures increase as the sample size increases, in general. This indicates that correctly specifying the correlation structure is important for achieving high efficiency when the sample size becomes larger. In addition, as the sample size increases, the efficiency of the mixed-effect QIF estimator also improves as expected.

We also provide the means of the variance component estimators for the random effects and their standard errors in Table 2.5 - 2.6 for $N = 100$ and $\rho = 0.7$. Overall, the mixed-effect QIF approach provides more accurate variance component estimators compared to other approaches. Note that the PQL, GLIMMIX and

Table 2.3: MSE and the standard errors of MSE (provided in the lower right corner) for the estimator of the intercept $\beta_0 = -0.3$ for binary responses when $\rho = 0.7$ from 200 simulations.

Method	$N = 50$ True correlation			$N = 100$ True correlation		
	Independent	Exchangeable	AR-1	Independent	Exchangeable	AR-1
QIF (ind)	$0.2436_{0.0260}$	$0.3253_{0.0347}$	$0.2358_{0.0227}$	$0.1464_{0.0142}$	$0.1373_{0.0148}$	$0.1298_{0.0143}$
QIF (exch)	$0.2541_{0.0250}$	$0.2391_{0.0226}$	$0.2189_{0.0209}$	$0.1494_{0.0144}$	$0.0762_{0.0082}$	$0.1080_{0.0101}$
QIF (AR-1)	$0.2579_{0.0267}$	$0.2664_{0.0265}$	$0.1706_{0.0180}$	$0.1499_{0.0145}$	$0.0842_{0.0091}$	$0.0802_{0.0078}$
PQL	$0.2591_{0.0282}$	$4.4578_{0.6305}$	$1.3758_{0.2319}$	$0.1517_{0.0146}$	$1.8556_{0.2864}$	$0.6628_{0.0906}$
GLIMMIX (Ind)	$0.0483^{\mathbf{1}}_{0.0062}$	$2.7467_{0.3529}$	$0.7870^{\mathbf{2}}_{0.0945}$	$0.1604^{\mathbf{7}}_{0.0138}$	$1.0159_{0.1059}$	$0.3912^{\mathbf{8}}_{0.0366}$
GLIMMIX (AR-1)	$0.1911^{\mathbf{3}}_{0.0302}$	$2.9716^{\mathbf{4}}_{0.3709}$	$0.3500^{\mathbf{5}}_{0.0431}$	$0.1713^{\mathbf{9}}_{0.0179}$	$1.0616^{\mathbf{10}}_{0.1151}$	$0.1226^{\mathbf{11}}_{0.0108}$
NLMIXED	$0.2494^{\mathbf{6}}_{0.0267}$	$3.0178_{0.4163}$	$0.9779_{0.1023}$	$0.1505^{\mathbf{12}}_{0.0148}$	$1.1474_{0.1091}$	$0.5126_{0.0480}$

Table 2.4: MSE and the standard errors of MSE (provided in the lower right corner) for the estimator of the slope $\beta_1 = 0.3$ for binary responses when $\rho = 0.7$ from 200 simulations.

Method	$N = 50$ True correlation			$N = 100$ True correlation		
	Independent	Exchangeable	AR-1	Independent	Exchangeable	AR-1
QIF (ind)	$0.2250_{0.0236}$	$0.2206_{0.0227}$	$0.1676_{0.0145}$	$0.1414_{0.0141}$	$0.1072_{0.0110}$	$0.0979_{0.0098}$
QIF (exch)	$0.2312_{0.0221}$	$0.1477_{0.0150}$	$0.1476_{0.0122}$	$0.1425_{0.0142}$	$0.0534_{0.0056}$	$0.0732_{0.0069}$
QIF (AR-1)	$0.2372_{0.0239}$	$0.1660_{0.0160}$	$0.1131_{0.0105}$	$0.1447_{0.0143}$	$0.0578_{0.0067}$	$0.0494_{0.0044}$
PQL	$0.2337_{0.0248}$	$2.5490_{0.2675}$	$0.8298_{0.0995}$	$0.1451_{0.0145}$	$1.1949_{0.1148}$	$0.4354_{0.0459}$
GLIMMIX (Ind)	$0.0279^{\mathbf{1}}_{0.0024}$	$3.0933_{0.4540}$	$0.6635^{\mathbf{2}}_{0.0762}$	$0.1445^{\mathbf{7}}_{0.0124}$	$1.2468_{0.1235}$	$0.3509^{\mathbf{8}}_{0.0352}$
GLIMMIX (AR-1)	$0.1770^{\mathbf{3}}_{0.0278}$	$3.3627^{\mathbf{4}}_{0.4691}$	$0.2164^{\mathbf{5}}_{0.0225}$	$0.1656^{\mathbf{9}}_{0.0189}$	$1.2860^{\mathbf{10}}_{0.1331}$	$0.0755^{\mathbf{11}}_{0.0071}$
NLMIXED	$0.2293^{\mathbf{6}}_{0.0241}$	$2.1011_{0.1954}$	$0.6844_{0.0668}$	$0.1448^{\mathbf{12}}_{0.0148}$	$0.9600_{0.0857}$	$0.3971_{0.0398}$

Note: Number of non-convergence outcomes from GLIMMIX procedures are tabulated as follows: **1.** 189; **2.** 16; **3.** 185; **4.** 9; **5.** 170; **6.** 15; **7.** 173; **8.** 7; **9.** 174; **10.** 1; **11.** 174; **12.** 7.

NLMIXED approaches provide much larger variance component estimators than the true variance and the mixed-effect QIF estimators. The discrepancy becomes much larger if the true structure is not independent. For example, when the true structure is exchangeable, the variance estimators of the random slopes from the PQL, the GLIMMIX procedure with independent structure and the NLMIXED procedures are 12.45, 4.85 and 8.51 respectively, while the true variance of the random slopes is 0.015 and the mixed-effect QIF variance estimator is 0.013.

To visualize the estimated distribution for the random effects, Figure 1 provides a probability histogram of a total of 20,000 random slope estimators from 200 simulated data sets when the true correlation structure is AR-1 with $\rho = 0.7$, and the sample size $N = 100$. The random-effect estimators are obtained based on the mixed-effect QIF approach with correct specified correlation structure. The solid line in the graph provides

the random-effects density function generated from the true Beta distribution. Figure 2.1 indicates that the estimated random effects resemble the rescaled Beta(0.5, 0.5) distribution reasonably well for the non-boundary region. In addition, we also provide the probability histogram of the random effects estimators in the PQL approach in Figure 2. The histogram shows that the PQL estimators of the random effects follow a normal-curve pattern. This is likely bacause the PQL imposes a normality assumption on the random effects distribution.

2.5 Real-data Example

2.5.1 Periodontal Example for Binary Data

In this section, we apply the mixed-effect QIF method to an observational periodontal disease study to determine if nonsurgical periodontal treatment is effective for the prevention of tooth loss (Stoner, 2000). The data are from 722 subjects with chronic periodontal diseases with 7-year follow up. To take into account the lag time between periodontal treatment and tooth loss, a history of non-surgical periodontal treatments is defined as an indicator for three years prior to a particular time-point of interest (Nonsurg). The association between the binary response of whether there is any tooth extraction, and the history of non-surgical periodontal treatment, is modeled. Other covariate factors include patients' characteristics such as gender (Gender), age (Age), number of teeth (Teeth), number of diseased sites (Sites), mean pocket depth of diseased sites (Pddis) and mean pocket depth of all sites (Pdall) at the initial visit. In addition, the number of non-periodontal dental treatments (Dent), the number of non-periodontal preventive procedures (Prev) and the number of surgical treatments (Surg) over the 3-year baseline period were considered in the model as well. We fit a random intercept model to incorporate the heterogeneity variation among patients. Specifically, we have the following logistic model:

$$\begin{aligned}\texttt{logit}(\mu_{ij}^b) &= \beta_0 + b_i + \beta_1 \texttt{Gender}_{ij} + \beta_2 \texttt{Age}_{ij} + \beta_3 \texttt{Teeth}_{ij} + \beta_4 \texttt{Sites}_{ij} \\ &\quad + \beta_5 \texttt{Pddis}_{ij} + \beta_6 \texttt{Pdall}_{ij} + \beta_7 \texttt{Surg}_{ij} + \beta_8 \texttt{Dent}_{ij} + \beta_9 \texttt{Prev}_{ij} + \beta_{10} \texttt{Nonsurg}_{ij}.\end{aligned}$$

Among the 722 patients, 558 (77%) of them had at least 5 years of follow-up information. Since patients have different numbers of years of follow-up information, the data has an unbalanced cluster design. We fit the mixed-effect QIF methods with three different types of working correlation structures, and the PQL method using the logistic model above, for the 558 patients with balanced data only. The estimators, standard errors and z-values are summarized in Table 3.7.

Table 3.7 indicates that the mixed-effect QIF with independent and exchangeable working correlation structures and the NLMIXED procedure produce very similar results, except for Surg and Pddis. However, there are significant differences between the estimators from the mixed-effect QIF with AR-1 working correlation and the PQL and NLMIXED estimators. For example, the PQL estimator and the NLMIXED estimator for Pdall effect are 0.6425 and 0.4832, respectively, and neither of these are significant. On the other hand, the mixed-effect QIF estimator for Pdall with AR-1 working correlation structure is 0.7792, which is significant. The GLIMMIX procedure does not converge under the exchangeable working correlation structure. The GLIMMIX assuming independent or AR-1 structure produces results similar to the mixed-effect QIF approaches with AR-1 working correlation, except for the effects of Age and Teeth. In general, the standard errors of the mixed-effect QIF estimators are smaller than the standard errors from the other approaches, except for Surg, Dent, and Prev effects.

2.5.2 Epileptic Seizure Data Example for Poisson Data

We also apply (Thall and Vail, 1990, p. 664) epileptic seizure data to illustrate the conditional mixed-effect approach for Poisson response. The epileptic seizure data consists of 59 epilepsy patients, 31 of whom received a new anti-epileptic drug and 28 of whom received a placebo. For each patient, baseline data are recorded including the patient's age and the number of epileptic seizures during an 8-week interval prior to receiving treatment. The responses are the number of epileptic seizures occurring in the 2-week period before each of four clinic visits. Clearly, measurements within each patient are correlated.

Let y_{ij} be the seizure count for patient i at the jth visit ($i = 1, \ldots, 59$, $j = 1, \ldots, 4$). To incorporate random effects among patients for the count data, the log link function is specified as

$$\log(\mu_{ij}^b/O_i) = \beta_0 - \log(O_i) + b_i + \beta_1 x_{i1} + \beta_2 T_i + \beta_3 x_{i2} + \beta_4 \text{Visit}_{ij}/10,$$

where $\mu_{ij}^b = E(y_{ij}|b_i)$, and $O_i = \sqrt{m_i}\text{s.e.}(\mathbf{y})$ is used to adjust for the cluster size m_i and the variance of the count data, which should have no effect on fixed-effects estimation for covariates. Here the cluster size $m_i = 4$ is the same for all subjects. The covariates in the model include $x_{i1} = \log(\text{base}_i/4)$ which is the logarithm of 1/4 of the number of baseline seizures, $x_{i2} = \log(\text{age}_i)$ which is the logarithm of the ith patient's age, T_i is a treatment indicator variable defined to be 1 for the new drug and 0 for the placebo, and visit_{ij} is a time-dependent covariate for four visits where $\text{visit}_{ij} = -3, -1, 1, 3$ for $j = 1, 2, 3$ and 4. In addition, for the count data, the conditional variance is the same as the conditional mean, that is $\text{var}(y_{ij}|\mathbf{b}_i) = \mu_{ij}^{\mathbf{b}}$.

We compare the mixed-effect QIF approach with three different working correlation structures as well

as the PQL method, the GLIMMIX approach and the NLMIXED procedure. The results are summarized in Table 3.8. Table 3.8 shows that the treatment effect obtained from the mixed-effect QIF approach is not significant, but is significant in the PQL and NLMIXED approaches. On the other hand, the mixed-effect QIF provides a highly significant age effect, while the other approaches produce non-significant results. The GLIMMIX with exchangeable working correlation does not converge here, so it is not presented. The GLIMMIX estimators under other working structures are similar to those obtained from the PQL and the NLMIXED procedure except for treatment. For this data example, the mixed-effect QIF under AR-1 and exchangeable structures provides smaller standard errors of fixed effects compared to the other approaches.

2.6 Discussion

Mixed-effects models are extremely useful for longitudinal data when subject-specific variation is one of our main interests. However, most of the random-effects approaches, including Gaussian-Hermite quadrature, are based on exact likelihood functions which could be difficult to approximate numerically for large dimensions of random effects. The penalized quasi-likelihood and the conditional second-order generalized estimating equations all rely on the normality assumption for random effects, which can be restrictive in practice.

We propose a new approach for generalized linear mixed models, which allows one to estimate the fixed and random effects simultaneously. The main difference between our method and existing generalized linear mixed-effects models is that we do not have parametric model assumptions for the random effects. Our method is based on conditional extended scores which only involve the first two moment conditions, and the random-effect parameters are estimated through minimizing the penalized quadratic inference function. Therefore, the specification of the likelihood function is not required for estimation. The main advantage of the proposed method is that it is able to incorporate both serial correlation from repeated measurements and heterogeneous variation from individuals. In addition, the distribution assumption for random effects is not required.

Moreover, the proposed method does not involve unknown variance components estimation as in the PQL, nor the estimation of the nuisance parameters associated with the working correlations as in the CGEE2. We provide consistency and asymptotic normality for the fixed-effects estimator. The derivation of the asymptotic property does not rely on the specific distribution of the random effects, as in other existing random-effects methods.

In addition, when serial correlation is introduced into the data, the mixed-effect QIF performs better than the PQL, SAS GLIMMIX and NLMIXED in general. This improvement proves to be particularly significant

for binary response data in our simulation. We also note that the GLIMMIX procedure tends to have a convergence problem for binary data. Finally, in our simulation, even if the dimension of the random-effects parameters increases as the sample size increases, the computation is fairly fast and efficient.

2.7 Proofs of Theorems and Lemmas

2.7.1 Notation

We denote the estimate of the random effects as $\hat{\mathbf{b}}$, and let $Q(\boldsymbol{\beta}|\mathbf{b}_0)$ be the quadratic inference function defined in (4) conditional on the true random effects $\mathbf{b}_0$,

$$\dot{Q}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_0|\mathbf{b}_0) = \frac{\partial}{\partial \boldsymbol{\beta}} Q(\boldsymbol{\beta}|\mathbf{b}_0)\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_0},$$

and $\dot{Q}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_0|\hat{\mathbf{b}})$, $\dot{Q}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_1|\mathbf{b}_0)$, and $\dot{Q}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_1|\hat{\mathbf{b}})$ can be defined similarly. In addition, let

$$\ddot{Q}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_0|\mathbf{b}_0) = \frac{\partial^2}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}} Q(\boldsymbol{\beta}|\mathbf{b}_0)\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_0},$$

$$\ddot{Q}_{\boldsymbol{\beta}\mathbf{b}}(\hat{\boldsymbol{\beta}}_0|\mathbf{b}_0) = \frac{\partial^2}{\partial \boldsymbol{\beta}\partial \mathbf{b}} Q(\boldsymbol{\beta}|\mathbf{b}_0)\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_0,\ \mathbf{b}=\mathbf{b}_0},$$

and $\ddot{Q}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_1|\mathbf{b}_0)$ and $\ddot{Q}_{\boldsymbol{\beta}\mathbf{b}}(\hat{\boldsymbol{\beta}}_1|\mathbf{b}_0)$ are defined similarly. Let $\mathbf{G}_N(\boldsymbol{\beta}|\mathbf{b}) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{g}_i(\boldsymbol{\beta}|\mathbf{b}_i)$. We can define

$$\dot{\mathbf{G}}_{N,\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_1|\mathbf{b}_0) = \frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{G}_{N,\boldsymbol{\beta}}(\boldsymbol{\beta}|\mathbf{b}_0)\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_1},$$

$$\dot{\mathbf{G}}_{N,\mathbf{b}}(\hat{\boldsymbol{\beta}}_1|\mathbf{b}_0) = \frac{\partial}{\partial \mathbf{b}} \mathbf{G}_{N,b}(\hat{\boldsymbol{\beta}}_1|\mathbf{b}_0)\Big|_{\mathbf{b}=\mathbf{b}_0},$$

$$\text{and } \dot{\mathbf{G}}_{N,\mathbf{b}}(\hat{\boldsymbol{\beta}}_1|\mathbf{b}_0) = \frac{\partial}{\partial \mathbf{b}} \mathbf{G}_{N,b}(\hat{\boldsymbol{\beta}}_1|\mathbf{b}_0)\Big|_{\mathbf{b}=\mathbf{b}_0}.$$

The other second derivatives associated with the different parameters are defined in the same fashion. Let

$$\hat{\boldsymbol{\beta}}_0 = \mathtt{arg}\min Q(\boldsymbol{\beta}|\mathbf{b});\ \hat{\boldsymbol{\beta}}_1 = \mathtt{arg}\min Q(\boldsymbol{\beta}|\hat{\mathbf{b}}).$$

Both $\hat{\boldsymbol{\beta}}_0$ and $\hat{\boldsymbol{\beta}}_1$ are in S, that is,

$$\dot{Q}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_0|\mathbf{b}_0) = 0, \quad \dot{Q}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_1|\hat{\mathbf{b}}) = 0.$$

Also let $\mathbf{A}_N(\boldsymbol{\beta}|\mathbf{b})$ be the weighting matrix such that

$$\mathbf{C}_N^{-1}(\boldsymbol{\beta}|\mathbf{b}) = \mathbf{A}_N(\boldsymbol{\beta}|\mathbf{b})'\mathbf{A}_N(\boldsymbol{\beta}|\mathbf{b})$$

and $Q(\boldsymbol{\beta}|\mathbf{b}) = |\mathbf{A}_N(\boldsymbol{\beta}|\mathbf{b})\mathbf{G}_N(\boldsymbol{\beta}|\mathbf{b})|^2$.

2.7.2 Regularity Conditions and Assumptions

Here we prove consistency and asymptotic normality for the fixed-effect estimator under the following assumptions.

1. Define n_i as the cluster size for subject i, let $n = \min(n_i)$, then $n_i = O_p(n)$ uniformly for $i = 1, \ldots, N$.
2. The parameter space S is compact.
3. Conditional on the true random effects $\mathbf{b}_0$, the parameter $\boldsymbol{\beta}$ is identifiable; that is, there is a unique $\boldsymbol{\beta}_0 \in S$ which satisfies $E\{\mathbf{g}(\boldsymbol{\beta}_0|\mathbf{b}_0)\} = 0$.
4. The derivative of the score function with respect to the random effects $\dot{\mathbf{g}}_{i,\mathbf{b}}(\hat{\boldsymbol{\beta}}|\mathbf{b}_0)$ is uniformly bounded in probability, i.e. $\dot{\mathbf{g}}_{i,\mathbf{b}}(\hat{\boldsymbol{\beta}}|\mathbf{b}_0) = O_p(1)$.
5. We require that $E[\mathbf{g}(\boldsymbol{\beta}|\mathbf{b})]$ be continuous and differentiable in both $\boldsymbol{\beta}$ and $\mathbf{b}$.
6. The expectation of $\mathbf{g}_i(\boldsymbol{\beta}_0|\hat{\mathbf{b}})$, the estimating functions conditional on the estimated random effects, converges to 0 in probability, i.e.

$$E[E\{\mathbf{g}_i(\boldsymbol{\beta}_0|\hat{\mathbf{b}})\}] \xrightarrow{p} 0 \quad \texttt{as } N \to \infty.$$

7. The weighting matrix $\mathbf{C}_N(\boldsymbol{\beta}|\mathbf{b})$ converges almost surely to a constant matrix $\mathbf{C}_0(\boldsymbol{\beta}|\mathbf{b})$, while $\mathbf{A}_N(\boldsymbol{\beta}|\mathbf{b})$ converges almost surely to a constant matrix $\mathbf{A}_0(\boldsymbol{\beta}|\mathbf{b})$ where $\mathbf{C}_0^{-1}(\boldsymbol{\beta}|\mathbf{b}) = \mathbf{A}_0(\boldsymbol{\beta}|\mathbf{b})\mathbf{A}_0(\boldsymbol{\beta}|\mathbf{b})'$.

2.7.3 Proofs of Lemmas and Theorem 1

Proof of Lemma 1. Define $B_N(r, \boldsymbol{\beta}_0) = \{\boldsymbol{\beta} | \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| < r/\sqrt{N}\}$ for a fixed constant r. Then by Taylor expansion, we have

$$\sup_{\boldsymbol{\beta} \in B_N(r,\boldsymbol{\beta}_0)} |\sqrt{N}\{\dot{Q}_{\boldsymbol{\beta}}(\boldsymbol{\beta}|\hat{\mathbf{b}}) - \dot{Q}_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0|\hat{\mathbf{b}})\}| = \sup_{\boldsymbol{\beta} \in B_N(r,\boldsymbol{\beta}_0)} |\sqrt{N}\ddot{Q}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}_0|\hat{\mathbf{b}})(\boldsymbol{\beta} - \boldsymbol{\beta}_0)| + o_p(1).$$

Since $\dot{Q}_{\boldsymbol{\beta}}(\boldsymbol{\beta}|\hat{\mathbf{b}}) = \dot{Q}_{\boldsymbol{\beta}}(\boldsymbol{\beta}|\hat{\mathbf{b}}) - \dot{Q}_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0|\hat{\mathbf{b}}) + \dot{Q}_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0|\hat{\mathbf{b}})$, we have

$$\sup_{\boldsymbol{\beta}\in B_N(r,\boldsymbol{\beta}_0)} |\sqrt{N}\dot{Q}_{\boldsymbol{\beta}}(\boldsymbol{\beta}|\hat{b}) - \sqrt{N}\ddot{Q}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}_0|\hat{\mathbf{b}})(\boldsymbol{\beta}-\boldsymbol{\beta}_0) - \sqrt{N}\dot{Q}_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0|\hat{\mathbf{b}})| = o_p(1). \tag{2.16}$$

Further, when $\boldsymbol{\beta}$ is on the boundary of $B_N(r,\boldsymbol{\beta}_0)$, i.e. $\boldsymbol{\beta} \in \{\boldsymbol{\beta}|\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\| = r/\sqrt{N}\}$,

$$N(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\ddot{Q}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}_0|\hat{b})(\boldsymbol{\beta}-\boldsymbol{\beta}_0) = O(r^2) > 0$$

since $\ddot{Q}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}_0|\hat{\mathbf{b}})$ is positive-definite and uniformly bounded.

In addition, by the weak law of large numbers and Condition 3, $\sqrt{N}\dot{Q}_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0|\mathbf{b}_0) = O_p(1)$, since

$$\sqrt{N}\dot{Q}_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0|\hat{\mathbf{b}}) = \sqrt{N}\dot{\mathbf{G}}_{N,\boldsymbol{\beta}}(\boldsymbol{\beta}_0|\hat{\mathbf{b}})\mathbf{C}_N^{-1}(\hat{\mathbf{b}})\mathbf{G}_{N,\boldsymbol{\beta}}(\boldsymbol{\beta}_0|\hat{\mathbf{b}}) + o_p(1).$$

It can be concluded by Conditions 4 and 6 that

$$\sqrt{N}\{\dot{Q}_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0|\hat{\mathbf{b}}) - \dot{Q}_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0|\mathbf{b}_0)\} = O_p(1).$$

Hence it follows from the above that

$$\sqrt{N}\dot{Q}_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0|\hat{\mathbf{b}}) = \sqrt{N}\{\dot{Q}_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0|\hat{\mathbf{b}}) - \dot{Q}_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0|\mathbf{b}_0)\} + \sqrt{N}\dot{Q}_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0|\mathbf{b}_0) = O_p(1),$$

which leads to $N(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\dot{Q}_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0|\hat{\mathbf{b}}) = O_p(r)$. Therefore for any $\epsilon > 0$, there exists an M, such that when $r > M$,

$$P\{N(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\ddot{Q}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}_0|\hat{\mathbf{b}})(\boldsymbol{\beta}-\boldsymbol{\beta}_0) + N(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\dot{Q}_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0|\hat{\mathbf{b}}) > 0\} > 1-\epsilon \tag{2.17}$$

for all $\boldsymbol{\beta}$ on the boundary of $B_N(r,\boldsymbol{\beta}_0)$. Therefore (2.17) certainly holds for all $\boldsymbol{\beta} \notin B_N(r,\boldsymbol{\beta}_0)$.

It follows from (2.17) and (3.10) that

$$(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\dot{Q}_{\boldsymbol{\beta}}(\boldsymbol{\beta}|\hat{\mathbf{b}}) > 0 \tag{2.18}$$

for $\boldsymbol{\beta} \notin B_N(r,\boldsymbol{\beta}_0)$ and some sufficiently large but finite r. Since the left-hand side of (2.18) is continuous for $\boldsymbol{\beta}$, by theorem (6.3.4) of Ortega and Rheinboldt (1973, p. 163), there must be a solution in $B_N(r,\boldsymbol{\beta}_0)$

satisfying

$$\dot{Q}_{\boldsymbol{\beta}}(\boldsymbol{\beta}|\hat{\mathbf{b}}) = 0.$$

□

Proof of Lemma 2. Since $\hat{\boldsymbol{\beta}}_0 = \texttt{arg}\min Q(\boldsymbol{\beta}|\mathbf{b}_0)$,

$$|Q(\hat{\boldsymbol{\beta}}_0|\mathbf{b}_0)|^2 < |Q(\boldsymbol{\beta}_0|\mathbf{b}_0)|^2.$$

That is,

$$|\mathbf{A}_N(\hat{\boldsymbol{\beta}}_0|\mathbf{b}_0)\frac{1}{N}\sum_{i=1}^{N}\mathbf{g}_i(\hat{\boldsymbol{\beta}}_0|\mathbf{b}_0)|^2 < |\mathbf{A}_N(\boldsymbol{\beta}_0|\mathbf{b}_0)\frac{1}{N}\sum_{i=1}^{N}\mathbf{g}_i(\boldsymbol{\beta}_0|\mathbf{b}_0)|^2.$$

By the law of large numbers, we know that the right side of the above converges to 0 as $E[\mathbf{g}(\boldsymbol{\beta}_0|\mathbf{b}_0)] = 0$. Further, by Assumption 8, the uniform law of large numbers and the continuity mapping theorem, we can prove that

$$|\mathbf{A}_N(\hat{\boldsymbol{\beta}}_0|\mathbf{b}_0)\frac{1}{N}\sum_{i=1}^{N}\mathbf{g}_i(\hat{\boldsymbol{\beta}}_0) - \mathbf{A}_0(\hat{\boldsymbol{\beta}}_0|\mathbf{b}_0)E[\mathbf{g}(\hat{\boldsymbol{\beta}}_0|\mathbf{b}_0)]| \to_{a.s.} 0.$$

It follows that

$$|A_0(\hat{\boldsymbol{\beta}}_0|\mathbf{b}_0)E[\mathbf{g}(\hat{\boldsymbol{\beta}}_0|\mathbf{b}_0)]|^2 \to_{a.s.} 0.$$

Hence $\hat{\boldsymbol{\beta}}_0$ converges to $\boldsymbol{\beta}_0$ almost surely. □

Proof of Lemma 3. Since $\dot{Q}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_0|\mathbf{b}_0) = \dot{\mathbf{G}}_{N,\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_0|\mathbf{b}_0)'\mathbf{C}_N^{-1}(\hat{\boldsymbol{\beta}}_0|\mathbf{b}_0)\mathbf{G}_N(\hat{\boldsymbol{\beta}}_0|\mathbf{b}_0)$, by Taylor's Expansion,

$$\begin{aligned} 0 &= \dot{Q}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_0|\mathbf{b}_0) = \dot{Q}_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0|\mathbf{b}_0) + \ddot{Q}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\tilde{\boldsymbol{\beta}}|\mathbf{b}_0)(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0) \\ &= \dot{\mathbf{G}}_{N,\boldsymbol{\beta}}(\boldsymbol{\beta}_0|\mathbf{b}_0)'\mathbf{C}_N^{-1}(\boldsymbol{\beta}_0|\mathbf{b}_0)\mathbf{G}_N(\boldsymbol{\beta}_0|\mathbf{b}_0) + \ddot{Q}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\tilde{\boldsymbol{\beta}}|\mathbf{b}_0)(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0), \end{aligned}$$

where $\tilde{\boldsymbol{\beta}}$ is between $\hat{\boldsymbol{\beta}}_0$ and $\boldsymbol{\beta}_0$. Then we have

$$\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0 = -\ddot{Q}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1}(\tilde{\boldsymbol{\beta}}|\mathbf{b}_0)\dot{\mathbf{G}}_{N,\boldsymbol{\beta}}(\boldsymbol{\beta}_0|\mathbf{b}_0)'\mathbf{C}_N^{-1}(\boldsymbol{\beta}_0|\mathbf{b}_0)\mathbf{G}_N(\boldsymbol{\beta}_0|\mathbf{b}_0).$$

Since $\hat{\boldsymbol{\beta}}_0 \to_{a.s.} \boldsymbol{\beta}_0$, it follows immediately that

$$\tilde{\boldsymbol{\beta}} \to_{a.s.} \boldsymbol{\beta}_0, \texttt{and } \dot{\mathbf{G}}_{N,\boldsymbol{\beta}}(\tilde{\boldsymbol{\beta}}|\mathbf{b}_0) \to_{a.s.} \mathbf{d}_0.$$

By the Central Limit Theorem and Assumption 3, $\sqrt{N}\mathbf{G}_N(\boldsymbol{\beta}_0|\mathbf{b}_0) \xrightarrow{d} N(0,\boldsymbol{\Sigma})$ and $\mathbf{C}_N(\boldsymbol{\beta}_0|\mathbf{b}_0) \to_p \boldsymbol{\Sigma} = N\boldsymbol{\Sigma}_N$. Therefore $\sqrt{N}(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0)$ converges to a normal distribution of mean 0 with asymptotic covariance matrix

$$\begin{aligned}
&\mathsf{cov}(\sqrt{N}(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0)) \\
&= \ddot{Q}^{-1}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}_0|\mathbf{b}_0)\dot{\mathbf{G}}'_{N,\boldsymbol{\beta}}(\boldsymbol{\beta}_0|\mathbf{b}_0)\mathbf{C}_N^{-1}(\boldsymbol{\beta}_0|\mathbf{b}_0)\boldsymbol{\Sigma}\mathbf{C}_N^{-1}(\boldsymbol{\beta}_0|\mathbf{b}_0)\dot{\mathbf{G}}_{N,\boldsymbol{\beta}}(\boldsymbol{\beta}_0|\mathbf{b}_0)\ddot{Q}^{-1}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}_0|\mathbf{b}_0) \\
&\to (\mathbf{d}_0'\boldsymbol{\Sigma}^{-1}\mathbf{d}_0)^{-1}\mathbf{d}_0'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\mathbf{d}_0(\mathbf{d}_0'\boldsymbol{\Sigma}^{-1}\mathbf{d}_0)^{-1} = (\mathbf{d}_0\boldsymbol{\Sigma}\mathbf{d}_0)^{-1} = \boldsymbol{\Omega}_0. \qquad (2.19)
\end{aligned}$$

This is because $\ddot{Q}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}_0|\mathbf{b}_0) = \dot{\mathbf{G}}_{N,\boldsymbol{\beta}}(\boldsymbol{\beta}_0|\mathbf{b}_0)'\mathbf{C}_N^{-1}(\boldsymbol{\beta}_0|\mathbf{b}_0)\dot{\mathbf{G}}_{N,\beta}(\boldsymbol{\beta}_0|\mathbf{b}_0) + o_p(1)$. Hence it follows immediately that $\ddot{Q}^{-1}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_0|\mathbf{b}_0) \to_{a.s.} \boldsymbol{\Omega}_0$. □

Proof of Theorem 1. Consistency of $\hat{\boldsymbol{\beta}}_1$ follows immediately from Lemma 1. By Lemma 3, $\sqrt{N}(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0)$ also converges to the normal distribution. Furthermore,

$$\begin{aligned}
&\sqrt{N}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_0) = \sqrt{N}(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0) + \sqrt{N}(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0) \\
&= \sqrt{N}\ddot{Q}^{-1}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\tilde{\boldsymbol{\beta}}|\mathbf{b}_0)\dot{Q}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_1|\mathbf{b}_0) - \sqrt{N}\ddot{Q}^{-1}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}_0|\mathbf{b}_0)\dot{Q}_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0|\mathbf{b}_0) + o_p(1) \qquad (2.20) \\
&= \sqrt{N}\ddot{Q}^{-1}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}_0|\mathbf{b}_0)\dot{\mathbf{G}}_{N,\boldsymbol{\beta}}(\boldsymbol{\beta}_0|\mathbf{b}_0)\mathbf{C}_N^{-1}(\mathbf{b}_0)1/N\sum_{i=1}^{N}[\mathbf{g}_i(\hat{\boldsymbol{\beta}}_1|\mathbf{b}_0) - \mathbf{g}_i(\boldsymbol{\beta}_0|\mathbf{b}_0)] + o_p(1).
\end{aligned}$$

Define $\boldsymbol{\Sigma}^*$ as

$$\boldsymbol{\Sigma}^* = \lim_{N\to\infty} E[N\{\mathbf{G}_N(\hat{\boldsymbol{\beta}}_1|\mathbf{b}_0) - \mathbf{G}_N(\boldsymbol{\beta}_0|\mathbf{b}_0)\}\{\mathbf{G}_N(\hat{\boldsymbol{\beta}}_1|\mathbf{b}_0) - \mathbf{G}_N(\boldsymbol{\beta}_0|\mathbf{b}_0)\}']. \qquad (2.21)$$

Hence, the asymptotic variance of $\sqrt{N}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_0)$ can be written as

$$\boldsymbol{\Omega}_1 = (\mathbf{d}_0'\boldsymbol{\Sigma}^{-1}\mathbf{d}_0)^{-1}\mathbf{d}_0'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^*\boldsymbol{\Sigma}^{-1}\mathbf{d}_0(\mathbf{d}_0'\boldsymbol{\Sigma}^{-1}\mathbf{d}_0)^{-1}. \qquad (2.22)$$

When the estimate of the random effects is consistent, i.e. $\hat{\mathbf{b}} \to_p \mathbf{b}_0$ as $n \to \infty$, it can be shown that

$$\sqrt{N}\dot{\mathbf{G}}_{N,\mathbf{b}}(\hat{\boldsymbol{\beta}}_1|\tilde{\mathbf{b}})(\hat{\mathbf{b}} - \mathbf{b}_0) = O_p(1)o_p(1) = o_p(1).$$

Therefore,

$$\begin{aligned}\sqrt{N}\ddot{Q}_{\boldsymbol{\beta}\mathbf{b}}(\hat{\boldsymbol{\beta}}_1|\tilde{\mathbf{b}})(\hat{\mathbf{b}}-\mathbf{b}_0) &= \sqrt{N}\{\dot{\mathbf{G}}_{N,\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_1|\tilde{\mathbf{b}})\}'\mathbf{C}_N^{-1}(\tilde{\mathbf{b}})\dot{\mathbf{G}}_{N,\mathbf{b}}(\hat{\boldsymbol{\beta}}_1|\tilde{\mathbf{b}})(\hat{\mathbf{b}}-\mathbf{b}_0)+o_p(1)\\ &= \sqrt{N}\{\dot{\mathbf{G}}_{N,\boldsymbol{\beta}}(\boldsymbol{\beta}_0|\mathbf{b}_0)\}'\mathbf{C}_N^{-1}(\mathbf{b}_0)\dot{\mathbf{G}}_{N,\mathbf{b}}(\hat{\boldsymbol{\beta}}_1|\tilde{\mathbf{b}})(\hat{\mathbf{b}}-\mathbf{b})+o_p(1)\\ &= o_p(1).\end{aligned}$$

Then by Taylor expansion, we have

$$\sqrt{N}\{\dot{Q}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_1|\hat{\mathbf{b}})-\dot{Q}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_1|\mathbf{b}_0)\}=\sqrt{N}\ddot{Q}_{\boldsymbol{\beta}b}(\hat{\boldsymbol{\beta}}_1|\tilde{\mathbf{b}})(\hat{\mathbf{b}}-\mathbf{b}_0)=o_p(1).$$

It follows immediately from $\dot{Q}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_1|\hat{\mathbf{b}})=0$ that

$$\sqrt{N}\dot{Q}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_1|\mathbf{b}_0)=\sqrt{N}\dot{\mathbf{G}}_{N,\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_1|\mathbf{b}_0)'\mathbf{C}_N^{-1}(\mathbf{b}_0)\mathbf{G}_N(\hat{\boldsymbol{\beta}}_1|\mathbf{b}_0)+o_p(1)=o_p(1). \tag{2.23}$$

Then by (2.20) and (2.23), we can conclude that

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_0)=\sqrt{N}(\hat{\boldsymbol{\beta}}_0-\boldsymbol{\beta}_0)+o_p(1).$$

Hence it follows from (2.19) that

$$\boldsymbol{\Omega}_1=\ddot{Q}^{-1}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}_0|\mathbf{b}_0)+o_p(1),$$

which can be approximated by $\ddot{Q}^{-1}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_1|\hat{\mathbf{b}})$ since $\ddot{Q}^{-1}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_1|\hat{\mathbf{b}})\xrightarrow{p}\ddot{Q}^{-1}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}_0|\mathbf{b}_0)$. □

2.7.4 Conditions and Proof of Consistency of Random-effect Estimator

We estimate $b_{0,i}$ by solving

$$\mathbf{g}_i^*(\hat{\boldsymbol{\beta}}_1|\hat{\mathbf{b}}_i)=\dot{\boldsymbol{\mu}}_{i,b}(\hat{\boldsymbol{\beta}}_1|\hat{\mathbf{b}}_i)(\mathbf{y}_i-\boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}}_1|\hat{\mathbf{b}}_i))=0.$$

Therefore, by Taylor expansion we have

$$\hat{\mathbf{b}}_i-\mathbf{b}_{0i}=\{\dot{\mathbf{g}}^*_{i,\mathbf{b}_i}(\hat{\boldsymbol{\beta}}_1|\tilde{\mathbf{b}}_i)\}^{-1}\sum_{j=1}^{n_i}\dot{\mu}_{ij,b}(\hat{\boldsymbol{\beta}}_1|\mathbf{b}_0)(y_{ij}-\mu_{ij}(\hat{\boldsymbol{\beta}}_1|\mathbf{b}_0)).$$

Since $\hat{\boldsymbol{\beta}}_1 \xrightarrow{p} \boldsymbol{\beta}_0$, then

$$\hat{\mathbf{b}}_i - \mathbf{b}_{0i} \rightarrow \{\dot{\mathbf{g}}^*_{i,\mathbf{b}_i}(\boldsymbol{\beta}_0|\tilde{\mathbf{b}}_i)\}^{-1} \sum_{j=1}^{n_i} \dot{\mu}_{ij,b}(\boldsymbol{\beta}_0|\mathbf{b}_0)(y_{ij} - \mu_{ij}(\boldsymbol{\beta}_0|\mathbf{b}_0)).$$

Since $\{\dot{\mathbf{g}}^*_{i,\mathbf{b}_i}(\boldsymbol{\beta}_0|\tilde{\mathbf{b}}_i)\}^{-1}$ is bounded in probability, therefore if the law of large numbers holds for the sequence $\dot{\mu}_{i1,b}(\boldsymbol{\beta}_0|\mathbf{b}_0)\{y_{i1} - \mu_{i1}(\boldsymbol{\beta}_0|\mathbf{b}_0)\}, \ldots, \dot{\mu}_{in_i,b}(\boldsymbol{\beta}_0|\mathbf{b}_0)\{y_{in_i} - \mu_{in_i}(\boldsymbol{\beta}_0|\mathbf{b}_0)\}$, we can conclude that

$$\hat{\mathbf{b}}_i - \mathbf{b}_{0i} = O_p(n_i^{-1/2}).$$

That is, $\hat{\mathbf{b}}$ is a consistent estimator of $\mathbf{b}_0$. This is because $E\{\dot{\mu}_{ij,b}(\boldsymbol{\beta}_0|\mathbf{b}_0)(y_{ij} - \mu_{ij}(\boldsymbol{\beta}_0|\mathbf{b}_0))\} = 0$.

Let $Z_{ij} = \dot{\mu}_{ij,b}(\boldsymbol{\beta}_0|\mathbf{b}_0)\{y_{ij} - \mu_{ij}(\boldsymbol{\beta}_0|\mathbf{b}_0)\}$. From Andrews (1988), if the sequence of random variables satisfies the L_1 mixingale conditions:

(a) $\|E(Z_{ij}|Z_{i,j-m}\|_1 \leq c_j\psi_m$, and

(b) $\|Z_i - E(Z_{ij}|Z_{i,j+m})\|_1 \leq c_j\psi_{m+1}$,

where $\{c_j : i \geq 1\}$ and $\{\psi_m : m \geq 0\}$ are some non-negative constants and $\psi_m \rightarrow 0$ as $m \rightarrow \infty$, and if $\overline{\lim}_{n\rightarrow\infty} \frac{1}{n} \sum_{j=1}^{n} c_j < \infty$ or $\{c_j\}$ can be given by $\{||Z_{ij}||_1\}$, we have the law of large numbers for the dependent sequence $\bar{Z}_i = 1/n_i \sum_{j=1}^{n_i} Z_{ij} \rightarrow_p 0$. Such conditions can be satisfied for sequences such as autoregressive, stationary Gaussian, or M-dependent and other sequences with decaying α mixing numbers.

Table 2.5: Mean and the standard errors of mean (provided in the lower right corner) of the variance component estimator for random intercepts out of 200 simulations for binary response, when $N = 100$ and $\rho = 0.7$. The true variance of the random intercept is 0.015.

Method	$N = 100$ True correlation		
	Independent	Exchangeable	AR-1
QIF (ind)	$0.0051_{0.0000}$	$0.0153_{0.0000}$	$0.0133_{0.0000}$
QIF (exch)	$0.0051_{0.0000}$	$0.0154_{0.0000}$	$0.0133_{0.0000}$
QIF (AR-1)	$0.0051_{0.0000}$	$0.0154_{0.0000}$	$0.0134_{0.0000}$
PQL	$0.4602_{0.0706}$	$52.6074_{5.8626}$	$12.4455_{1.0486}$
GLIMMIX (Ind)	$0.0812^{\mathbf{1}}_{0.0043}$	$11.2695_{0.2750}$	$4.8532^{\mathbf{2}}_{0.0979}$
GLIMMIX (AR-1)	$0.1622^{\mathbf{3}}_{0.0084}$	$11.4377^{\mathbf{4}}_{0.3306}$	$0.5640^{\mathbf{5}}_{0.0313}$
NLMIXED	$0.0515^{\mathbf{6}}_{0.0067}$	$25.4358_{0.6391}$	$8.5088_{0.2236}$

Table 2.6: Mean and the standard errors of mean (provided in the lower right corner) of the variance component estimator for random intercepts out of 200 simulations for binary response, when $N = 100$ and $\rho = 0.7$. The true variance of the random intercept is 0.015.

Method	$N = 100$ True correlation		
	Independent	Exchangeable	AR-1
QIF (ind)	$0.0073_{0.0000}$	$0.0262_{0.0000}$	$0.0212_{0.0000}$
QIF (exch)	$0.0073_{0.0000}$	$0.0263_{0.0000}$	$0.0213_{0.0000}$
QIF (AR-1)	$0.0073_{0.0000}$	$0.0263_{0.0000}$	$0.0213_{0.0000}$
PQL	$0.3573_{0.0592}$	$23.6498_{2.0657}$	$3.0542_{0.3737}$
GLIMMIX (Ind)	$0.0763^{\mathbf{1}}_{0.0049}$	$8.9704_{0.3339}$	$2.6071^{\mathbf{2}}_{0.1082}$
GLIMMIX (AR-1)	$0.0932^{\mathbf{3}}_{0.0063}$	$9.1258^{\mathbf{4}}_{0.3857}$	$0.2159^{\mathbf{5}}_{0.0136}$
NLMIXED	$0.0335^{\mathbf{6}}_{0.0060}$	$3.2856_{0.1699}$	$1.4055_{0.1175}$

Note: Number of non-convergence outcomes from GLIMMIX procedures are tabulated as follows: **1.** 173; **2.** 7; **3.** 174; **4.** 1; **5.** 174; **6.** 7.

Figure 2.1: Histogram of the estimates of the random slopes from the binary data sets with $N = 100$ and $\rho = 0.7$. The true correlation structure of the data set is AR(1), and the estimates are obtained by the mixed-QIF method with AR(1) working correlation. The solid line in the graph provides the random-effects density function generated from the true Beta distribution.

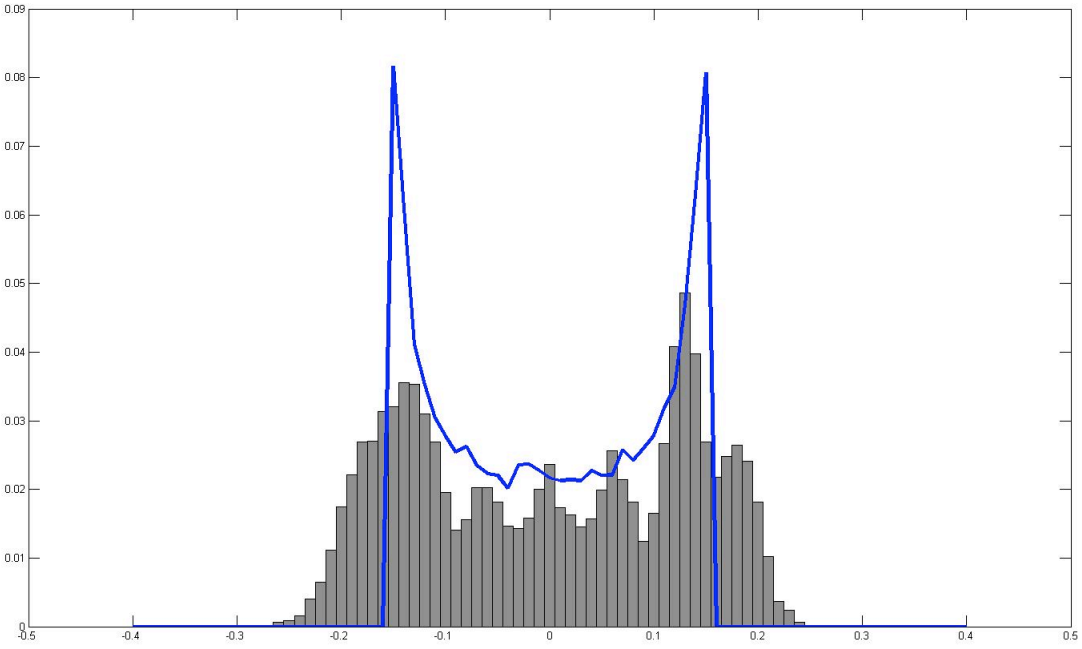

Figure 2.2: Histogram of the estimates of the random slopes from the binary data sets with $N = 100$ and $\rho = 0.7$. The true correlation structure of the data set is AR(1), and the estimates are obtained by the PQL approach.

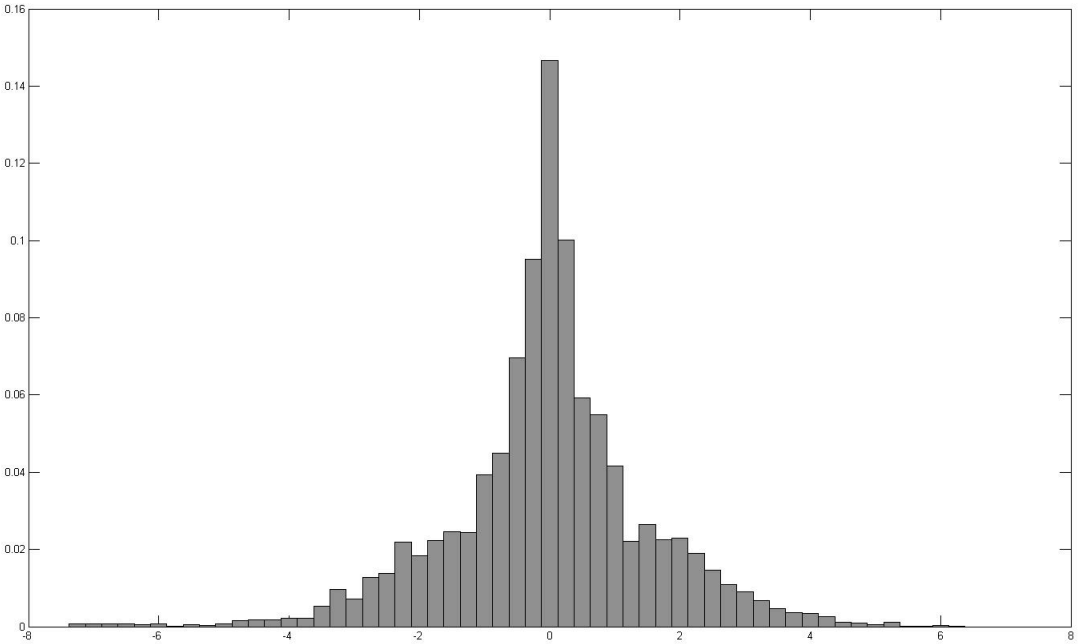

Table 2.7: Comparison of mixed QIF and the other approaches for non-surgical periodontal treatment data.

Covariates	QIF_{ind}	QIF_{CS}	$QIF_{AR\text{-}1}$	PQL	$GLIMMIX_{ind}$	$GLIMMIX_{AR\text{-}1}$	NLMIXED
Intercept	-7.1549	-7.4769	-8.1300	-8.0824	-9.3476	-9.5602	-6.8281
s.e.	1.3630	1.3476	1.3637	1.6265	1.7433	1.7804	1.5600
z-value	-5.2492	-5.5482	-5.9615	-4.9691	-5.3620	-5.3697	-4.3770
Gender	0.2257	0.2138	0.2409	0.2383	0.2317	0.2387	0.2526
s.e.	0.1522	0.1530	0.1589	0.1720	0.1766	0.1802	0.1588
z-value	1.4828	1.3974	1.5155	1.3865	1.3120	1.3246	1.5907
Age	0.0168	0.0175	0.0152	0.0202	0.0279	0.0291	0.0173
s.e.	0.0105	0.0104	0.0108	0.0123	0.0128	0.0131	0.0114
z-value	1.5948	1.6781	1.4072	1.6427	2.1772	2.2305	1.5109
Teeth	-0.0334	-0.0325	-0.0177	-0.0353	-0.0388	-0.0406	-0.0440
s.e.	0.0246	0.0241	0.0242	0.0271	0.2767	0.0282	0.0254
z-value	-1.3591	-1.3518	-0.7295	-1.3010	-1.4051	-1.4385	-1.7323
Sites	0.0024	0.0025	-0.0042	-0.0005	-0.0042	-0.0048	0.0032
s.e.	0.0097	0.0099	0.0090	0.0102	0.0105	0.0107	0.0098
z-value	0.2468	0.2555	-0.4684	-0.0524	-0.4029	-0.4440	0.3301
Pddis	0.2689	0.3469	0.1948	0.2719	0.2944	0.2899	0.2587
s.e.	0.1864	0.1790	0.1904	0.2293	0.2370	0.2418	0.2124
z-value	1.4428	1.9377	1.0232	1.1866	1.2422	1.1989	1.2180
Pdall	0.4644	0.3960	0.7792	0.6425	0.8465	0.8885	0.4832
s.e.	0.3880	0.3946	0.3626	0.4200	0.4329	0.4423	0.4020
z-value	1.1968	1.0035	2.1489	1.5292	1.9554	2.0088	1.2020
Surg	-0.1377	0.0039	-0.1636	-0.0932	-0.1020	0.1304	-0.1087
s.e.	0.2741	0.2336	0.2863	0.2901	0.2019	0.2034	0.2790
z-value	-0.5024	0.0168	-0.5716	-0.3213	-0.5052	0.6411	-0.3896
Dent	0.1074	0.1132	0.1158	0.1205	0.1353	0.1365	0.1172
s.e.	0.0083	0.0082	0.0086	0.0080	0.0061	0.0061	0.0084
z-value	12.9164	13.8498	13.4963	15.0844	22.3636	22.4433	13.9126
Prev	0.0404	0.0271	0.0169	0.0353	0.0381	0.0395	0.0363
s.e.	0.1349	0.1353	0.1398	0.1500	0.0988	0.0990	0.1378
z-value	0.2992	0.2004	0.1207	0.2420	0.3856	0.3988	0.2636
Nonsurg	-0.2360	-0.2037	-0.2149	-0.2207	-0.1995	-0.2041	-0.2266
s.e.	0.1500	0.1504	0.1577	0.1767	0.1839	0.1876	0.1632
z-value	-1.5732	-1.3548	-1.3624	-1.2500	-1.0848	-1.0880	-1.3885

Note: QIF_{ind}, QIF_{CS} and $QIF_{AR\text{-}1}$ are the mixed-effect QIF methods with independent, exchangeable and AR-1 working correlation structures.
Pddis: mean pocket depth of disease sites; Pdall: mean pocket depth of all sites; Surg: number of surgical treatments; Dent: number of non-periodontal dental treatments; Prev: number of non-periodontal preventive procedures; Nonsurg: indicator for nonsurgical periodontal treatments during the three years prior to the time-point of interest.
Neither GLIMMIX with AR-1 working correlation nor GLIMMIX with exchangeable working correlation can converge for this data-set.

Chapter 3

Correlation Structure Selection for High Dimensional Correlated Data

3.1 Introduction

In longitudinal data analysis, it is important to identify the correct correlation structure since it can improve the estimation efficiency for regression parameter if the true information of the correlation structure can be utilized (Liang and Zeger, 1986; Qu et al., 2000). In addition, incorporating the correct correlation structure can also reduce the bias of parameter estimation in nonparametric modeling (Wang, 2003), and increases statistical power for hypothesis testing.

However, model selection for correlation structure remains a challenging problem due to the involvement of higher order of moment estimations. The problem is especially challenging when the cluster size m is large, and the associated dimension of the correlation parameters increases at a rate of m^2. This makes it practically infeasiblel to use the empirical estimation of the correlation matrix directly, even if it might be close to the true correlation structure.

The existing work mainly focuses on the estimation of the covariance matrix rather than the selection of correlation structure. Huang et al. (2006, 2007) proposed estimation and covariance selection based on Cholesky decomposition. Fan et al. (2008) developed the factor modeling approach, and El Karoui (2008) introduced the spectrum random matrix approach to estimate the high dimension sparse covariance matrix. However, these methods are mainly applicable for continuous outcomes. Other recent developments for large covariance matrices estimation include the nested LASSO approach (Levina et al., 2008), the banding and tapering approach (Bickel and Levina, 2008b), the thresholding approaches (Bickel and Levina, 2008a; Rothman et al., 2009), and the multivariate linear regression (Yuan, 2010) and the penalized normal likelihood function approach (Rothman et al., 2008) for estimating the inverse of the covariance matrix. In addition, Fisher and Sun (2010) proposed an improved Stein-type shrinkage estimator for the high-dimensional multivariate normal covariance matrix. These approaches are still not capable of handling discrete responses. However, discrete outcomes occur frequently in longitudinal data studies. In addition, selection of the true correlation structure itself is scientifically important in longitudinal studies.

In this paper, we approximate the inverse of the empirical correlation matrix using a linear combination of candidate basis matrices. The possible candidate basis matrices could contain most the common correlation structures or linear combinations of different correlation structures, which are based on prior knowledge of possible correlation structures. We select the correlation structure by identifying groups of basis matrices with non-zero coefficients, where each group of basis matrices represents a specified correlation structure. This is carried out by minimizing an objective function which involves two parts. The first part is the Euclidean norm between two estimation equations, one based on the empirical information of the correlation matrix, the other based on the model approximation by a linear combination of candidate correlation structures. The second part is a penalty function which penalizes a model involving too many basis matrices for the correlating modeling. The correlation structure is selected by group-wise (Yuan and Lin, 2006) SCAD penalty function (Fan and Li, 2001). That is, the specific correlation structure is identified, if and only if, the whole group of basis matrices is selected.

Through identifying non-zero coefficients of the basis matrices with certain structures, one can avoid the estimation of each individual entry of the correlation matrix. Note that the dimension of the correlation parameters could be very high if the cluster size is large, without assuming certain structures of the correlation matrix. One of advantages is that the dimension of the parameters involved in the estimation is greatly reduced. Another main advantage of our approach is that it does not require the specification of the likelihood function. Therefore it can be applied to non-normal correlated responses, such as binary and count data. To the authors' best knowledge, there is very limited literature discussing model selection for correlation structure in non-continuous outcome data. In contrast to Zhou and Qu (2011) model selection for correlation structure where the cluster size is fixed, here we allow the cluster size to diverge as the sample size increases. Note that the asymptotic theory and numerical implementation for model selection of correlation structure for diverging cluster size is very different and much more challenging compared to the fixed cluster size case.

We show that the proposed selection procedure tends to identify the true correlation structure with probability tending to 1 when the sample size is sufficiently large. Therefore, a positive-definite correlation matrix can be ensured asymptotically. In addition, our approach enjoys the oracle property, such that the estimated non-zero coefficients of the basis matrices have an asymptotic normal distribution. The derivation of the asymptotic results is challenging when the cluster size diverges. Our simulation results also confirm that even when the cluster size is quite large, the proposed method works effectively to identify the correct correlation structures for both continuous and discrete outcomes.

The chapter is organized as follows. Section 3.2 describes the general framework of basis matrices representations for various correlation structures. Section 3.3 proposes model selection by minimizing the

penalized estimating equations. Section 3.4 develops asymptotic properties of model selection consistency, and the oracle property when the number of group basis matrices diverges as the sample size increases. Section 3.5 focus on implementation of the proposed selection procedure. Section 3.6 provides simulation results for both continuous and binary cases. Section 3.7 provides a data example on air pollution for illustration. We provide the final conclusion and discussion in Section 3.8. The technical proofs of the lemmas and theorems on the asymptotic properties of the proposed estimators are provided in the last section.

3.2 General Framework

In the longitudinal data framework, the marginal mean model is

$$E(y_i) = \mu(X_i\beta), \quad i = 1, \ldots, n, \tag{3.1}$$

where $y_i = (y_{i1}, \ldots, y_{im})'$ is the response variable which is repeatedly measured over time $t = 1, \ldots, m$; X_i is a known $m \times \dim(\beta)$ covariate matrix associated with a parameter vector β; and $\mu(\cdot)$ is the inverse of the link function between the response variable and predictor variables.

To estimate the regression parameter β in (3.1), Wedderburn (1974) proposed the quasi-likelihood function for correlated data

$$\sum_{i=1}^{n} \dot{\mu_i}^T V_i^{-1}(y_i - \mu_i) = 0, \tag{3.2}$$

where V_i is the covariance matrix of y_i and $\mu_i = \mu(X_i\beta)$. In practice, V_i is often unknown, but can be substituted by the empirical covariance. However, the empirical estimator of V_i could be unstable, especially when the sample size is relatively small compared to the cluster size. Liang and Zeger (1986) introduced the generalized estimation equation approach which extends (3.2) by assuming $V_i = A_i^{1/2} R A_i^{1/2}$, where A_i is a diagonal matrix with the marginal variance of y_i as the diagonal component, and R is a working correlation matrix.

To improve the efficiency of the GEE estimator under the misspecified working correlation structure, Qu et al. (2000) introduced the quadratic inference function approach. The key idea is that the inverse of the working correlation matrix R^{-1} can be approximated by a linear combination of basis matrices. That is, $R^{-1} \approx \sum_{j=1}^{t} a_j M_j$, where M_1 is usually an identity matrix and k is the number of basis matrices. Then,

the GEE can be approximated by

$$\sum_{i=1}^{n} \dot{\mu}_i^T A_i^{-1/2} \left(\sum_{j=1}^{k} a_j M_j \right) A_i^{-1/2} (y_i - \mu_i) = 0.$$

Qu et al. (2000) defined the extended scores to be

$$\bar{g}_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} g_i(\beta) = \frac{1}{n} \begin{pmatrix} \sum_{\mathrm{i=1}}^{\mathrm{n}} \dot{\mu}_{\mathrm{i}}^{\mathrm{T}} \mathrm{A}_{\mathrm{i}}^{-1/2} \mathrm{M}_1 \mathrm{A}_{\mathrm{i}}^{-1/2} (\mathrm{y_i} - \mu_{\mathrm{i}}) \\ \vdots \\ \sum_{\mathrm{i=1}}^{\mathrm{n}} \dot{\mu}_{\mathrm{i}}^{\mathrm{T}} \mathrm{A}_{\mathrm{i}}^{-1/2} \mathrm{M}_{\mathrm{k}} \mathrm{A}_{\mathrm{i}}^{-1/2} (\mathrm{y_i} - \mu_{\mathrm{i}}) \end{pmatrix},$$

where $\bar{g}_n$ is a $k \dim(\beta)$-dimensional vector. Note that the GEE is a linear combination of extended scores $\bar{g}_n$. The quadratic inference function is defined as

$$n \bar{g}_n^T C_n^{-1} \bar{g}_n,$$

where C_n is a $k \dim(\beta) \times k \dim(\beta)$ matrix and can be estimated consistently by $\hat{C}_n = \frac{1}{n} \sum_{i=1}^{n} g_i(\beta) g_i(\beta)^T$.

Besides for the regression parameter estimation, the quadratic inference function can be utilized in the model selection for correlation structure. However, here the parameters associated with correlation structures will no longer be treated as nuisance parameters anymore. For a fixed cluster size, Zhou and Qu (2011) proposed that the basis matrices can be divided into different groups, that is, R^{-1} can be represented as follows:

$$R^{-1} \approx \sum_{j=1}^{J_m} \sum_{b=1}^{B_j} \alpha_{jb} M_{jb} \tag{3.3}$$

where M_{jb} is the bth basis matrix in the jth group, and α_{jb} is the coefficient of M_{jb}. Here B_j is the number of basis matrices in the jth group, and J_m is the number of groups of basis matrices, which depends on the cluster size m. We denote $p_m = \sum_{j=1}^{J_m} B_j$ as the total number of basis matrices included. In the following section, we simplify $\mathbf{G}_j = (M_{j1}, \ldots M_{jB_j})$ and $\boldsymbol{\alpha}_j = (\alpha_{j1}, \ldots, \alpha_{jB_j})$, and $\boldsymbol{\alpha}_j \mathbf{G}_j = \sum_{b=1}^{B_j} \alpha_{jb} M_{jb}$.

3.3 Selection of Correlation Structure

In this section, we will illustrate how to perform model selection for correlation structure. The key idea is to estimate the coefficient $\boldsymbol{\alpha}_j$ by minimizing the discrepancy between the estimation function from the empirical correlation matrix estimator, and the estimating function based on the approximation of R^{-1} by

the candidate basis matrices. The discrepancy between the two estimating functions for the ith cluster is

$$S_i = \dot{\mu}_i^T(\hat{\beta})A_i^{-1/2}\{\tilde{R}^{-1} - \sum_{j=1}^{J_m}\boldsymbol{\alpha}_j\mathbf{G}_j\}A_i^{-1/2}(y_i - \mu_i(\hat{\beta})),$$

where $\hat{\beta}$ is the estimator of the regression parameter β through the generalized estimating equations assuming independent structure, and $\tilde{R}$ is the empirical correlation matrix computed from the residual $y - \mu(\hat{\beta})$. In order to ensure that the empirical correlation $\tilde{R}$ is invertible, we require the sample size n to be larger than $m + \dim(\beta)$. If an approximation to R^{-1} by groups of candidate matrices is sufficient, we would expect the Euclidean norm of $S = (S_1, \ldots, S_n)^T$ be sufficiently small.

In general, it is not desirable to include more basis matrices to achieve a better approximation for R^{-1} since a more complex correlation structure does not necessarily lead more efficient estimator for the regression parameters. This is especially of the case if the cluster size diverges as the sample size increase since estimating correlation matrix with diverging dimension could bring more variability for the regression parameter estimation.

We formulate a penalized Euclidean norm of S as an objective function, where a correlation matrix model involving too many basis matrices is penalized. More specifically, the coefficients of the basis matrices $\boldsymbol{\alpha}_j$ are estimated by minimizing the following objective function

$$\sum_{i=1}^{n} S_i^T S_i + n\dim(\beta)\sum_{j=2}^{J_m} p_\lambda(||\boldsymbol{\alpha}_j||), \tag{3.4}$$

where $p_\lambda(\cdot)$ is the SCAD penalty function, λ is the tuning parameter and $||\boldsymbol{\alpha}_j||$ is the L_2-norm of the coefficients for the j-th group of basis matrices $\mathbf{G}_j$. By imposing the L_2-norm of the coefficients $\boldsymbol{\alpha}_j$'s, the basis matrices within the same group are selected simultaneously. This penalty is rather different from Zhou and Qu (2011) where L_1-norm of the coefficients $\boldsymbol{\alpha}_j$'s is imposed. The new penalty performs better numerically. Note that the first group of basis matrices is not penalized, since it typically contains the block identity matrices which should always be included for the independent structure, the null model.

To minimize the objective function (3.4), we define

$$\begin{aligned}
U_i &= \dot{\mu}_i^T(\hat{\beta})A_i^{-1/2}\tilde{R}^{-1}A_i^{-1/2}\{y_i - \mu_i(\hat{\beta})\}, \quad i = 1, \ldots, n, \\
V_{i,jb} &= \dot{\mu}_i^T(\hat{\beta})A_i^{-1/2}M_{jb}A_i^{-1/2}\{y_i - \mu_i(\hat{\beta})\}, \quad j = 1, \ldots, J_m, b = 1, \ldots B_j.
\end{aligned}$$

Let V_{ij} be a $\dim(\beta) \times B_j$ matrix $V_{ij} = (V_{i,j1}, \ldots, V_{i,jB_j})$ and V_i be a $\dim(\beta) \times p_m$ matrix $V_i = (V_{i1}, \ldots, V_{iJ_m})$

, then the objective function in (3.4) can be represented as

$$Q(\boldsymbol{\alpha}) = \sum_{i=1}^{n} ||U_i - \sum_{j=1}^{J_m} V_{ij}\boldsymbol{\alpha}_j||^2 + n\dim(\beta)\sum_{j=2}^{J_m} p_\lambda(||\boldsymbol{\alpha}_j||). \tag{3.5}$$

The advantage of the above reformulation is that the model selection for the correlation structure is transformed to a penalized linear regression problem, where estimator of the coefficients of $\boldsymbol{\alpha}_j$'s are obtained by minimizing (3.5). We apply one-step local approximation to the SCAD penalty (Zou and Li, 2008), and (3.5) becomes

$$\sum_{i=1}^{n} ||U_i - \sum_{j=1}^{J_m} V_{ij}\boldsymbol{\alpha}_j||^2 + n\dim(\beta)\sum_{j=2}^{J_m} p'_\lambda(||\hat{\boldsymbol{\alpha}}_j^{(0)}||)||\boldsymbol{\alpha}_j||, \tag{3.6}$$

where $\hat{\boldsymbol{\alpha}}_j^{(0)}$ is the initial estimate of $\boldsymbol{\alpha}_j$, and can be obtained by the unpenalized least square estimator. Therefore minimizing (3.6) is equivalent to minimizing the adaptive group LASSO (Yuan and Lin, 2006) objective function, with component-specific tuning parameter

$$p'_\lambda(||\hat{\boldsymbol{\alpha}}_j^{(0)}||) = \lambda\{I(||\hat{\boldsymbol{\alpha}}_j^{(0)}|| \le \lambda) + \frac{(a\lambda - ||\hat{\boldsymbol{\alpha}}_j^{(0)}||)_+}{(a-1)\lambda} I(||\hat{\boldsymbol{\alpha}}_j^{(0)}|| \ge \lambda)\},$$

where $a > 2$ is another unknown parameter besides λ, and is usually chosen as 3.7 in practice (Fan and Li, 2001). Therefore, the objective function (3.6) can be solved by applying the coordinate-wise descent (CWD) algorithm, as in Yuan and Lin (2006).

3.4 Asymptotic Properties

In this section, we will provide the asymptotic properties of the estimators for the coefficients associated with the basis matrices, which include the model selection consistency and the oracle property. We first list the regularity conditions required to establish the asymptotic properties. In the following, we define α^j and α_0^j as the jth component of $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}_0$.

Regularity Conditions:

1. Let $a_n = \max_{1\le j\le p_m}\{p'_{\lambda_n}(|\alpha_0^j|), \alpha_0^j \ne 0\}$, and $b_n = \max_{1\le j\le p_m}\{p''_{\lambda_n}(|\alpha_0^j|),\ \alpha_0^j \ne 0\}$. The following conditions are associated with the penalty functions:

   a. $a_n = O(n^{-1/2})$;

   b. $b_n \to 0$ as $n \to \infty$;

c. $\liminf_{n\to\infty} \liminf_{\theta\to 0^+} p'_{\lambda_n}(\theta)/\lambda_n > 0$;

d. There are constants c_1 and c_2, such that when $\theta_1, \theta_2 > c_1\lambda_n$, $|p''_{\lambda_n}(\theta_1) - p''_{\lambda_n}(\theta_2))| \leq c_2|\theta_1 - \theta_2|$.

2. The empirical estimator of the correlation matrix $\tilde{R}$ is $\sqrt{n}$-consistent for each component $R(i,j)$, that is,

$$\sqrt{n}|\tilde{R}(i,j) - R(i,j)| = O_p(1),\ 1 \leq i \leq m, 1 \leq j \leq m.$$

3. The eigenvalues of the matrices $V_i^T V_i, i = 1, \ldots, n$ are bounded away from zero and infinity with probability tending to 1, that is, for any $\epsilon > 0$, there exist constants l_1 and l_2 such that

$$P(0 < l_1 < \lambda_{\min}\{V_i^T V_i\} \leq \lambda_{\max}\{V_i^T V_i\} < l_2 < \infty) > 1 - \epsilon.$$

4. The L_1 norm of the basis matrices is bounded, i.e., there is a constant K such that

$$||M_{jb}||_1 < K, \quad 1 \leq j \leq J_m,\ b = 1, \ldots B_j,$$

where the L_1 norm of a $m \times m$ matrix M is defined as

$$||M||_1 = \sum_{i=1}^{m}\sum_{j=1}^{m} |M(i,j)|.$$

Note Condition 4 ensures that basis matrix involved for model selection is sparse through decomposition since the number of basis matrices is allowed to increase as the cluster size increases.

Based on regularity Condition 2, if the correlation matrix R can be consistently estimated by the empirical correlation matrix $\tilde{R}$, then the coefficients $\boldsymbol{\alpha}_j$ associated with the group basis matrices can be estimated consistently under the assumption if the groups of basis matrices $\mathbf{G}_j$'s can capture the information of R^{-1} sufficiently well. Moreover, we show in Lemma 1 provided in Section 3.9 that once $\tilde{R}$ is a consistent estimator for R, $\tilde{R}^{-1}$ is also a consistent estimator for R^{-1}. By solving (3.5), we can approximate the inverse of the empirical correlation $\tilde{R}^{-1}$ by $\sum \hat{\boldsymbol{\alpha}}_j \mathbf{G}_j$ sufficiently well. We establish the following theorem for the consistency of $\hat{\boldsymbol{\alpha}}_j$.

Theorem 1. *Suppose the regularity conditions 1-4 are satisfied, if* $p_m^2/n \to 0$ *as* $n \to \infty$*, then there is a local minimizer* $\hat{\boldsymbol{\alpha}}$ *for minimizing* $Q(\boldsymbol{\alpha})$ *in* (3.5)*, such that* $||\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0|| = O_p\{\sqrt{p_m}(n^{-1/2} + a_n)\}$*, where* a_n *is given in Condition 1 and* $\boldsymbol{\alpha}_0 = (\boldsymbol{\alpha}_{01}, \ldots, \boldsymbol{\alpha}_{0J_m})$ *is the true coefficient vector associated with all the basis matrices.*

It follows from Theorem 1 that as long as $a_n = O_p(n^{-1/2})$, there is a $\sqrt{n/p_m}$-consistent estimator of $\boldsymbol{\alpha}$. For the SCAD penalty, $a_n = 0$ when n is large, therefore the SCAD estimator is consistent. The technical proof provided in the Section 3.9 is quite different from Fan and Peng (2004) since U_i here is no longer independent as all the U_i's contain common information of $\tilde{R}^{-1}$, while Fan and Peng assume that the score functions derived from the known likelihood function are independent for different subjects in the covariate model selection setting. However, the rate of convergence in Theorem 1 is still the same as in Fan and Peng (2004).

In the model selection framework, the parameter vector can be partitioned into two parts $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^T, \boldsymbol{\alpha}_2^T)^T$, where $\boldsymbol{\alpha}_1$ contains the non-zero coefficients of the basis matrices which capture the true structure of R^{-1}, while $\boldsymbol{\alpha}_2$ contains the zero coefficients which are irrelevant for modeling R^{-1}. We define the true parameter vector $\boldsymbol{\alpha}_0 = (\boldsymbol{\alpha}_{01}^T, \boldsymbol{\alpha}_{02}^T)^T$. Let $\hat{\boldsymbol{\alpha}}_1$ and $\hat{\boldsymbol{\alpha}}_2$ be the estimators of $\boldsymbol{\alpha}_{01}$ and $\boldsymbol{\alpha}_{02}$ respectively, we then establish the following oracle properties of $\hat{\boldsymbol{\alpha}}$.

Theorem 2. *Given all the regularity conditions are satisfied, if* $\lambda_n \to 0$*,* $\sqrt{n/p_m}\lambda_n \to \infty$ *and* $p_m^2/n \to 0$*, then with probability tending to 1, for any given constant* C*, and any* $\boldsymbol{\alpha}_1$ *satisfying* $||\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_{01}|| = O_p(\sqrt{p_m/n})$*,*

$$Q(\hat{\boldsymbol{\alpha}}_1, 0) = \min_{||\boldsymbol{\alpha}_2|| \le C(p_m/n)^{1/2}} Q(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2).$$

Theorem 2 indicates that the minimizer of $Q(\boldsymbol{\alpha})$ must occur when $\hat{\boldsymbol{\alpha}}_2 = 0$, which implies that part (i) in Theorem 3 holds.

Let $\nabla P_{\lambda_n}(\boldsymbol{\alpha}_0)$ be the first derivative vector of the penalty function and $\nabla^2 P_{\lambda_n}(\boldsymbol{\alpha}_0)$ be a diagonal matrix with the second derivatives of the penalty function as the diagonal components, that is,

$$\nabla P_{\lambda_n}(\boldsymbol{\alpha}_0) = \{p'_{\lambda_n}(\alpha_1^1), \ldots, p'_{\lambda_n}(\alpha_0^{p_m})\}^T,\ \nabla^2 P_{\lambda_n}(\boldsymbol{\alpha}_0) = \texttt{diag}\{p''_{\lambda_n}(\alpha_0^1), \ldots, p''_{\lambda_n}(\alpha_0^{p_m})\},$$

and $K_m = \mathcal{C}_m \Sigma_m \mathcal{C}_m^T$, where $\mathcal{C}_m = \{\text{vec}(C_{11})^T, \ldots, \text{vec}(C_{J_m B_{J_m}})\}^T$, C_{jb} is a constant matrix associated with the basis matrix M_{jb}, and Σ_m is the asymptotic variance of $\sqrt{n}\text{vec}\{R^{-2}(\tilde{R} - R)\}$.

Theorem 3. *Suppose all the regularity conditions are satisfied, if* $\lambda_n \to 0$*,* $\sqrt{n/p_m}\lambda_n \to \infty$ *and* $p_m^2/n \to 0$ *as* $n \to \infty$*, then with probability tending to 1, we establish the following oracle properties:*

(i) *(Sparsity)* $\hat{\boldsymbol{\alpha}}_2 = 0$.

(ii) *(Asymptotic normality)*

$$\sqrt{n}A_m K_{m,11}^{-1/2}\{I_{n,11} + \frac{1}{n}\nabla^2 P_{\lambda_n}(\boldsymbol{\alpha}_{01})\}(\hat{\boldsymbol{\alpha}}_{01} - \boldsymbol{\alpha}_{01}) + \frac{1}{\sqrt{n}}A_m K_{m,11}^{-1/2}\nabla P_{\lambda_n}(\boldsymbol{\alpha}_{01}) \xrightarrow{d} N(0, G),$$

where A_m *is any given matrix with a dimension of* $q \times p_m$ *which satisfies* $A_m^T A_m \to G$, $I_{n,11}$ *is the identity matrix and* $K_{m,11}$ *is a submatrix of* K_m *associated with* $\boldsymbol{\alpha}_1$.

The technical proofs of Lemma 1-2, and Theorems 1-3 are provided in the Section 3.9. The above asymptotic properties guarantee that the estimators of the coefficients achieve oracle properties: the non-zero coefficients are identified correctly with probability tending to 1, and the corresponding estimators are consistent and asymptotically normal as if the true correlation structure is known. The SCAD penalty also ensures that the estimators of the coefficients are not shrunk at all if the true coefficients are sufficiently far away from 0.

3.5 Implementation

3.5.1 Examples of Basis Matrices

Since selecting candidate basis matrices plays a critical role for the correlation structure model selection, in this section, we provide several examples for potential candidate basis matrices. These examples are also provided in Zhou and Qu (2011).

Example 1: If the correlation matrix R has an AR(1) structure with the correlation parameter ρ for any two adjacent longitudinal measurements, then the inverse of the correlation matrix can be represented as

$$R^{-1} = \alpha_{11} I_m + \alpha_{21} M_{2,1} + \alpha_{22} M_{2,2}.$$

Here I_m is the identity matrix in group $\mathbf{G}_1$; $M_{2,1}$ and $M_{2,2}$ are two basis matrices in group $\mathbf{G}_2$, where $M_{2,1}$ has 1 on the subdiagonal, and 0 elsewhere, and $M_{2,2}$ has 1 on the $(1,1)$ and (m,m) components, and 0 elsewhere. The corresponding coefficients for the candidate matrices are $\alpha_{11} = (1+\rho^2)/(1-\rho^2)$ and $\boldsymbol{\alpha}_2 = (\alpha_{21}, \alpha_{22}) = (-\rho/(1-\rho^2), -\rho^2/(1-\rho^2))$, respectively.

Example 2: If R is exchangeable with the correlation parameter ρ, we have

$$R^{-1} = \alpha_{11} I_m + \alpha_{31} M_{3,1}.$$

The second basis matrix $M_{3,1}$ has 0 on its main diagonal, and 1 elsewhere. The corresponding coefficients of the basis matrices are $\alpha_{11} = -\{(m-2)\rho+1\}/\{(m-1)\rho^2-(m-2)\rho-1\}$ and $\alpha_{31} = \rho/\{(m-1)\rho^2-(m-2)\rho-1\}$, respectively.

Example 3: The correlation matrix R has a block diagonal matrix structure. This could be quite common for correlated spatial data. Suppose there are total d blocks in a block diagonal matrix, where each block has block size m_j $(j = 1, \ldots, d)$, and the total cluster size $m = \sum_{j=1}^{d} m_j$. If each block can be represented as either independent, exchangeable, or AR(1) structure, then $\mathbf{G}_1$ has the identity matrix I_m, and $d-1$ matrices such that I_{m_j} $(j = 1, \ldots, d-1)$ contains the identity matrix for the jth block, and 0 on the other blocks. In addition, the remaining groups of basis matrices can be selected as follows. For any jth block with AR(1) structure, the group basis matrices contain two basis matrices $M_{2,1}$ and $M_{2,2}$ as provided in Example 1 with $m_j \times m_j$ dimension for the jth block, and 0 matrices for the other blocks. For any block with exchangeable structure, the group basis matrices contain a basis matrix $M_{3,1}$ as in Example 2 for the corresponding block and 0 matrices for the other blocks. The coefficients of these groups of basis matrices can be calculated similarly as in Examples 1 and 2.

In practice, the correlation structure can be as simple as one of the examples above, but can also be a combination of several simple correlation structures. In addition, other choices of basis matrices could be created using 0 or 1 components based on prior information. In this paper, we assume that all possible basis matrices candidates are included before proceeding the model selection. Therefore the correct correlation structure can be selected through identifying non-zero coefficients.

3.5.2 Tuning Parameter Selection

In this section, we discuss the selection of the tuning parameter λ for the group SCAD penalty. Traditional tuning parameter selection criteria include generalized cross-validation (GCV), AIC and BIC. However, none of these criteria work well for selecting tuning parameter. The simulation results (not provided here) show that both GCV and AIC tend to overfit the model and select null basis matrices more frequently, especially in the simulation for binary case. On the other hand, the BIC criterion tends to underfit the model. This is possibly due to the fact that none of the U_i's in the objective functions are independent assumed in other model selection approaches, and the dimension of the cluster size and the number of basis matrices diverge.

We introduce a rather different criterion for tuning parameter selection and propose a generalized information criterion (GIC) to select λ by minimizing

$$BIC_T(\lambda) = nr \log \frac{\eta_{\max}(\hat{R}^{-1}\tilde{R}^2\hat{R}^{-1})}{\eta_{\min}(\hat{R}^{-1}\tilde{R}^2\hat{R}^{-1})} + \log(n)k(\lambda), \tag{3.7}$$

where $\tilde{R}$ is the empirical correlation matrix and $\hat{R}^{-1}$ is the estimated inverse of the correlation matrix based on $\hat{R}^{-1} = \hat{\boldsymbol{\alpha}}_1\mathbf{G}_1 + \cdots + \hat{\boldsymbol{\alpha}}_{J_m}\mathbf{G}_{J_m}$, and the estimated coefficients $\hat{\boldsymbol{\alpha}}_1, \ldots, \hat{\boldsymbol{\alpha}}_{J_m}$ are obtained by minimizing (3.5) given a tuning parameter λ, and $k(\lambda)$ is the number of non-zero components among $\hat{\boldsymbol{\alpha}}_1, \ldots, \hat{\boldsymbol{\alpha}}_{J_m}$. The largest and the smallest eigenvalues of $\hat{R}^{-1}\tilde{R}^2\hat{R}^{-1}$ are denoted by $\eta_{\max}(\cdot)$ and $\eta_{\min}(\cdot)$ respectively. The tuning parameter λ is obtained such that the GIC (3.7) is minimized, where the first term of (3.7) measures the discrepancy between the empirical correlation and the selected correlation structure from the candidate basis matrices. If most of the correlation information in the empirical $\hat{R}$ is selected through correlation structure $\tilde{R}$, then $\hat{R}^{-1}\tilde{R}$ is close to the identity matrix, and the log of the ratio of the largest and the smallest eigenvalues of $\hat{R}^{-1}\tilde{R}^2\hat{R}^{-1}$ is close to 0.

The main difference between our tuning parameter selection and the usual BIC criteria is that there is an additional tuning parameter $r > 0$ in the first part of (3.7). This is analog to the generalized information criterion (Nishii, 1984; Zhang et al., 2010) in the linear model. The GIC criterion allows the tuning of r to provide additional control of its influence on the choice of λ. Note that the larger the r is, the more the first part of (3.7) dominates; and a smaller tuning parameter λ leads more groups of basis matrices being selected.

In order to determine the tuning parameter r, we perform a grid search of r for our simulation settings for both continuous and binary responses. From our empirical observation, the values of r associated with the best model selection performance are usually close to m/n, the ratio of cluster size and sample size. In addition, the choice of $r = m/n$ along with the criterion in (3.7) consistently outperforms the traditional GCV, AIC and BIC selection criteria. Note that the choice of r in Zhou and Qu (2011) is fixed, but here we provide various choices of r to assess the performance of model selection for correlation structures in the following simulation section.

3.6 Simulations

Table 3.1: Percentages of correctly identified signals and non-signals using the GIC criterion with correlation $\rho = 0.5$ for normal responses, sample size $n = 200$

						% of fits		
Cluster Size	r	Signals			Ave. of Non-signals	Correct	Under	Over
$m = 25$	0.125	100	100	100	100	1	0	0
$m = 50$	0.250	98	96	100	99.9	0.94	0.05	0.01
$m = 75$	0.375	89	92	94	99.5	0.76	0.20	0.04
$m = 100$	0.500	51	56	73	99.8	0.37	0.6	0.03

In this section, we evaluate the performance of our method in selecting the correlation structure through

Table 3.2: Percentages of correctly identified signals and non-signals using the GIC criterion with correlation $\rho = 0.7$ for normal responses, sample size $n = 200$

						% of fits		
Cluster Size	r	Signals			Ave. of Non-signals	Correct	Under	Over
$m = 25$	0.125	100	100	100	100	1	0	0
$m = 50$	0.250	99	100	100	99.9	0.98	0.01	0.01
$m = 75$	0.375	96	97	97	99.9	0.92	0.04	0.04
$m = 100$	0.500	97	96	98	98	0.72	0.06	0.22

Table 3.3: Percentages of correctly identified signals and non-signals using the GIC criterion with correlation $\rho = 0.6$ for binary response, sample size $n = 200$

						% of fits		
Cluster Size	r	Signals			Ave. of Non-signals	Correct	Under	Over
$m = 25$	0.125	100	100	100	99.4	0.96	0	0.04
$m = 50$	0.250	96	96	96	99.5	0.87	0.09	0.04
$m = 75$	0.375	89	94	94	97.9	0.61	0.19	0.2
$m = 100$	0.500	77	78	79	97.6	0.3	0.44	0.26

simulation studies for normal responses and binary responses.

The correlation matrix R is assumed to have a block diagonal structure, where each 5×5-dimensional block has a correlation structure either as AR(1), exchangeable or independent. We evaluate the model selection performance when the number of blocks d diverges as the sample size increases. The candidate basis matrices to represent the combination of these correlation structures are divided into several groups. As illustrated in Example 3 of Section 5.1, Group $\mathbf{G}_1$ contains the identity matrix I_{5d} and $d-1$ matrices with block identity matrices I_5 on the first, second, ..., $(d-1)$th blocks respectively, and 0 matrix on the other blocks. Group $\mathbf{G}_2$ contains two matrices with $M_{2,1}$ and $M_{2,2}$ in Example 1 of Section 5.1 for the first block and 0 matrix for the other blocks, corresponding to the AR(1) structure for block 1. Group $\mathbf{G}_3$ contains one matrix with $M_{3,1}$ for the first block and 0 matrix for the other blocks, corresponding to the exchangeable structure for block 1. The other groups of basis matrices are defined similarly as above, except with different block locations. In summary, the above group basis matrices represent the independence, AR(1) and exchangeable structures for each block. There are total $4d$ basis matrices, and $2d+1$ groups of basis matrices.

In our simulations, we let $d = 5, 10, 15$ and 20 represent the number of blocks. The corresponding cluster sizes are $m = 25, 50, 75$ and 100. The sample size n is 200 for continuous outcomes, and 200 and 300 for binary responses. The tuning parameter is chosen based on GIC criterion in (3.7) with $r = m/n$.

For the normal response, we generate the data from the following longitudinal model,

$$Y_i = \beta_0 + X_{1i}\beta_1 + X_{2i}\beta_2 + X_{3i}\beta_3 + \epsilon_i,$$

Table 3.4: Percentages of correctly identified signals and non-signals using the GIC criterion with correlation $\rho = 0.6$ for binary response, sample size $n = 300$

Cluster Size	r	Signals			Ave. of Non-signals	% of fits Correct	Under	Over
$m = 25$	0.833	100	100	100	99.9	0.99	0	0.01
$m = 50$	0.167	99	100	100	99.9	0.98	0.01	0.01
$m = 75$	0.250	94	98	97	99.2	0.82	0.08	0.10
$m = 100$	0.333	89	94	91	98.4	0.66	0.19	0.15

where Y_i is the response vector, $X_{ki}, k = 1, 2, 3$ are the covariates generated from the standard normal distribution, and ϵ_i is the error term following the multivariate normal distribution $N(0, R)$. Here the first two blocks of correlation matrix R have AR(1) correlation structures, and the third block is specified as exchangeable structure. The remaining $d-3$ blocks are independent. The covariates $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^T = (2, 1, 1, 1)^T$.

Table 3.1 and Table 3.2 summarize the simulation results from 100 simulations for continuous outcomes, with correlation coefficient $\rho = 0.5$ and $\rho = 0.7$ respectively; and provide the percentages of correct selection for each set of correlated blocks, and the overall percentage of not selecting for blocks which are independent. In addition, the percentages of correct-fitting, under-fitting and over-fitting are provided to assess the performance of the entire model selection method.

When the number of blocks $d = 5$ and the cluster size is 25, the correlation structures are identified 100% correctly in both cases. As the number of blocks increases, Our simulations show that the percentage of correctly identifying each correlated block decreases gradually. Consequently, the percentage of correct-fitting decreases, and the percentages of over-fitting and under-fitting increase. For example, with $\rho = 0.7$, the percentages of correct-fitting for the entire model are 98%, 92% and 72% when $d = 10, 15$ and 20 respectively. Note that when the correlation is strong such as $\rho = 0.7$, the percentages of correctly identifying the first three blocks with AR(1) and exchangeable correlation structures are reasonably high (above 95%), and the overall percentage of correctly identifying independent blocks (with no signal) is 98% even when $d = 20$ and cluster size $m = 100$. However, with a weaker correlation such as $\rho = 0.5$, the percentage of correct-fitting for the entire model drops to 76% when $d = 15$ and $m = 75$. In general, blocks with the exchangeable correlation structure can be identified more correctly than those with the AR(1) structure.

For the binary response, the responses are generated from the logistic regression model

$$\texttt{logit}\{E(Y_i)\} = \beta_0 + X_{1i}\beta_1 + X_{2i}\beta_2 + X_{3i}\beta_3,$$

where Y_i is the response vector, and $X_{ki}(k = 1, 2, 3)$ are the covariates, generated from a normal distribution

$N(0, 0.01)$. The correlated binary responses are generated using the R-package "mvtBinaryEP" with the first two blocks as exchangeable, and the third block as AR(1) correlation structures. The correlation parameter is set as 0.6. The remaining blocks have independent structure. The covariates $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^T = (0.2, 1, -1, -1)^T$.

Table 3.3 and Table 3.4 summarize the simulation results for the model selection with sample size $n = 200$ and $n = 300$ respectively. Similar to normal outcomes, when $d = 5$ and $m = 25$, the proposed model selection method can identify all the correlated block structures 100% correctly, and the percentage of correct-fitting for the entire model is 96% for $n = 200$ and 99% for $n = 300$. When the number of blocks and the cluster size increase, the percentages of correct-fitting for the entire model decrease. Specifically, when $d = 15$ and $m = 75$, the percentages of correct-fitting are 61% and 82% for $n = 200$ and $n = 300$, respectively. Our simulations also indicate that when the sample size $n = 300$, the proposed model selection strategy is able to correctly identify more than 89% for each specified correlated block structure, and over 98% for all the independent blocks even when $d = 20$ and $m = 100$.

3.7 Data Example on Air Pollution

We apply our method to a longitudinal study investigating the impact of air pollution on asthmatic patients. The study is based on observations from 39 asthmatic patients during a period of 21 consecutive days in Windsor, Ontario, Canada in 1992 (Fu, 2003). The response is the patient's daily asthmatic status, which is coded 1 if there is the presence of difficulties in breathing and 0 otherwise, which is measured by the forced expiratory volume (FEV) of each patient daily. The covariates are collected through air pollution indicators measured by several pollutants, and daily mean temperature (Meantemp) measured on each of 21 days. The pollutants measured in the study include nitrogen oxide (NO), nitrogen dioxide (NO2), mixture of NO and NO2 (NOX), total reduced sulphur (TRS), ozone level (OZ), carbon monoxide (CO), and coefficient of haze (COH). The following logistic model is fit for the binary outcomes:

$$\texttt{logit}(\mu_i) = \beta_1 \text{Meantemp} + \beta_2 \text{NO} + \beta_3 \text{NO2} + \beta_4 \text{NOX} + \beta_5 \text{TRS} + \beta_6 \text{OZ} + \beta_7 \text{CO} + \beta_8 \text{COH}. \tag{3.8}$$

All the covariates are centered and standardized before fitting the model, therefore the intercept is not included in the model.

To identify the correct correlation structure for this data, we first obtain an initial estimator of regression parameters by the GEE with independent working structure, and then calculate the empirical correlation matrix from the residuals using the initial estimators. We estimate the coefficients of the basis matrices by

Table 3.5: Comparison of the GEE estimators, standard errors and Z-values using different working correlation structures for air pollution data.

Effects	Independent	GIC	GCV
Meantemp	-0.2494	-0.1009	0.0660
s.e.	0.2563	0.0908	0.0892
z-value	-0.9733	-1.1112	0.7403
NO	0.2860	0.0553	-0.1362
s.e.	0.3419	0.1170	0.1178
z-value	0.8365	0.4724	-1.1555
NO2	-0.0105	-0.0335	0.0133
s.e.	0.0728	0.0235	0.0179
z-value	-0.1447	-1.4218	0.7425
NOX	-0.2717	-0.0728	0.0700
s.e.	0.1904	0.0676	0.0679
z-value	-1.4268	-1.0778	1.0298
TRS	-0.1784	-0.0037	-0.0063
s.e.	0.0947	0.0413	0.0340
z-value	-1.8836	-0.0892	-0.1856
OZ	0.1266	0.1190	0.1082
s.e.	0.1023	0.0341	0.0290
z-value	1.2384	3.4897	3.7244
CO	-0.0122	0.0200	-0.0504
s.e.	0.1504	0.0547	0.0487
z-value	-0.0810	0.3661	-1.0347
COH	0.1191	-0.0223	-0.1092
s.e.	0.0853	0.0289	0.0251
z-value	1.3967	-0.7740	-4.3530

minimizing the penalized objective function in (3.6), and select the tuning parameter based on the proposed GIC criterion (3.7) and the traditional GCV, AIC and BIC criteria. The candidate matrices include the identity matrix, M_{21} and M_{22} in Example 1 to represent the AR(1) structure, M_{31} in Example 2 for the exchangeable correlation structure, and some additional basis matrices to represent a mixture structure of AR(1) and exchangeable structures. In addition, we also include the candidate matrices to represent the sub-block structures illustrated in Example 3, where each week is considered as a sub-block and there are a total of three sub-blocks for 21 days.

Based on the GIC tuning parameter selection criterion (3.7) which chooses $r = 21/39$, the exchangeable working correlation structure is selected. The AIC, BIC and GCV tuning parameter selection criteria produce the same correlation structure, which selects all the basis matrices except the one with the exchangeable structure for the third block. We compare the estimators of the regression parameters for the logistic model (3.8), and their standard errors obtained by the GEE with the exchangeable structure selected by the GIC criterion, the complex structure selected by the AIC, BIC and GCV criteria, and the independent

working structure. In Table 3.5, the standard errors for all the regression parameter estimator obtained by the independent working correlation are the largest, and the standard errors obtained by the complex structure are the smallest in general, although the standard errors are comparable between the exchangeable correlation structure and the more complex structure. Note that since the sample size is relatively small compared to the cluster size, tuning parameter selection based on the AIC, BIC and GCV tends to select more basis matrices, and therefore leads to smaller standard errors.

Note that the GEE method with "unstructured" correlation structure requiring the estimation of each entry of the correlation matrix does not converge for this data set, as it cannot handle the cluster size of 21. This implies that using the empirical correlation structure may not be feasible when the cluster size is relatively large compared to the sample size. This example indicates that using the correlation structure selected by the proposed method (either based on GIC or GCV for tuning parameter selection) is better than either using independent working correlation or using the "unstructured" correlation for the purpose of the regression parameters estimation.

3.8 Discussion

In this paper, we propose a new approach to identify the correlation structure for longitudinal data when the cluster size increases as the sample size increases. The new approach does not require estimating each entry of the correlation matrix as in existing approaches. Instead, we approximate the empirical correlation structure with a linear combination of groups of candidate basis matrices, and estimate the coefficients via minimizing an objective function measuring the difference between the empirical estimating functions and the model approximated estimating functions. In addition, we penalize a approximated model if it contains too many basis matrices.

The advantages of the proposed approach include not requiring the specification of the likelihood function, and therefore it is applicable for non-normal correlated data. We also show that the proposed correlation structure selection possesses the consistency property for selecting the true model, and also holds an oracle property asymptotically. In our setting, we allow the cluster size and the number of group basis matrices to diverge as the sample size increases. The theoretical derivation for correlation structure model selection is nontrivial for diverging cluster size.

Our simulation studies show that even when the cluster size is quite large, the correlation structure can be identified effectively for both normal responses and binary responses through the new selection procedure. In our data example, we show that we are able to select a rather simple correlation structure

for the binary correlated data when the cluster size is large, while the GEE method with an unspecified correlation structure fails to produce converged estimators of the regression parameters. This indicates that it is important to select correlation structure which is sufficiently close to the true structure, rather than to apply naive independent structure or the empirical correlation structure.

3.9 Proof of Lemmas and Theorems

3.9.1 Lemma 1

Define $\tilde{e}_i$ as the standardized residual from $\hat{\beta}$ such that

$$\tilde{R} = \frac{1}{n}\sum_{i=1}^{n} \tilde{e}_i \tilde{e}_i^T.$$

Let Σ_m be the covariance matrix of $\text{vec}\{R^{-2}(\tilde{e}_i\tilde{e}_i^T - R)\}$, then we have the following lemma.

Lemma 1. *Under conditions 1 and 4, and if* $m^2/\sqrt{n} \to 0$

$$\sqrt{n} F_m vec(\tilde{R}^{-1} - R^{-1}) \xrightarrow{d} N(0, H),$$

where F_m *is a* $q \times m^2$ *matrix such that* $F_m \Sigma_m F_m^T \to H$*, and* H *is a* $q \times q$ *nonnegative symmetric matrix.*

Proof. By the Mean Value Theorem, we have

$$\sqrt{n} F_m \text{vec}(\tilde{R}^{-1} - R^{-1}) = -F_m \text{vec}\{D\sqrt{n}(\tilde{R} - R)\},$$

where

$$D = \int_0^1 (R + t\mathcal{H})^{-2} dt,$$

and $\mathcal{H} = \tilde{R} - R = O_p(n^{-1/2})$. It follows from the above that

$$\begin{aligned}
\text{vec}(|D - R^{-2}|) &\leq \int_0^1 \text{vec}(|(R + t\mathcal{H})^{-2} - R^{-2}|) dt \\
&\leq \int_0^1 \text{vec}(2t|\mathcal{H}R^{-1}| + t^2\mathcal{H}^2) dt \\
&\leq \text{vec}(2|\mathcal{H}R^{-1}| + \mathcal{H}^2) \\
&= O_p(mn^{-1/2}) + O_p(mn^{-1}) = O_p(mn^{-1/2}),
\end{aligned}$$

where $|M|$ is the the matrix with the absolute values of the components of M.

Therefore, by simple matrix calculation and the condition $F_m \Sigma_m F_m^T \to H$,

$$\begin{aligned}\sqrt{n}F_m \mathrm{vec}(\tilde{R}^{-1} - R^{-1}) &= -F_m \mathrm{vec}\{D\sqrt{n}(\tilde{R} - R)\} \\ &= -F_m \mathrm{vec}\{R^{-2}\sqrt{n}(\tilde{R} - R)\} + O_p(mn^{-1/2}).\end{aligned}$$

Then it follows from the condition $m = o_p(n^{1/2})$ that

$$\sqrt{n}F_m \mathrm{vec}(\tilde{R}^{-1} - R^{-1}) = -F_m \mathrm{vec}\{R^{-2}\sqrt{n}(\tilde{R} - R)\} + o_p(1).$$

Next we can use the Lindberg-Feller Central Limit Theorem to prove the asymptotic normality. Let

$$Y_{ni} = F_m \mathrm{vec}\{R^{-2}\frac{1}{\sqrt{n}}(\tilde{e}_i\tilde{e}_i^T - R)\}.$$

Then for any $\varepsilon > 0$,

$$\begin{aligned}\sum_{i=1}^{n} E||Y_{ni}||^2 1(||Y_{ni}|| > \varepsilon) &= nE||Y_{n1}^2||1(||Y_{n1}|| > \varepsilon) \\ &\leq n(E||Y_{n1}||^4)^{1/2}\{P(||Y_{n1}|| > \varepsilon)\}^{1/2}.\end{aligned}$$

Then by condition $F_m \Sigma_m F_m \to H$ and $\tilde{e}_i\tilde{e}_i^T - R = O_p(1)$, we have

$$\begin{aligned}P(||Y_{n1}|| > \varepsilon) &\leq \frac{E||F_m \mathrm{vec}\{R^{-2}(\tilde{e}_i\tilde{e}_i^T - R)\}||^2}{n\varepsilon} \\ &\leq \frac{||F_m \Sigma_m F_m||^2}{n\varepsilon} = O(n^{-1}),\end{aligned}$$

and

$$\begin{aligned}E||Y_{n1}||^4 &= \frac{1}{n^2}E||F_m \mathrm{vec}\{R^{-2}(\tilde{e}_i\tilde{e}_i^T - R)\}||^4 \\ &\leq \frac{1}{n^2}\lambda_{\max}(F_m F_m^T)\lambda_{\max}(R^{-4})E||(\tilde{e}_i\tilde{e}_i^T - R)^2||^2 \\ &= O(m^4/n^2).\end{aligned}$$

Therefore

$$\sum_{i=1}^{n} E||Y_{ni}||^2 1(||Y_{ni}|| > \varepsilon) = O(n\frac{m^2}{n}\frac{1}{\sqrt{n}}) = o(1).$$

The Lindberg condition is satisfied, so the Lindberg-Feller Central Limit Theorem can be applied. Suppose Σ_m is the asymptotic variance of $\sqrt{n}\text{vec}\{R^{-2}(\tilde{R}-R)\}$, we require the condition $F_m\Sigma_m A_m^T \to H$, where H is a finite dimension constant matrix, so that the aysmptotic variance of $\sum_{i=1}^n Y_{ni}$ is H. □

3.9.2 Lemma 2

Lemma 2. *Each element of the vector* $\frac{1}{\sqrt{n}}\sum_{i=1}^n (U_i - V_i\boldsymbol{\alpha}_0)^T V_i$ *is asymptotic normal, i.e.*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(U_i - V_i\boldsymbol{\alpha}_0)^T V_{i,jb} \xrightarrow{d} N(0, vec(C_{jb})\Sigma_m vec(C_{jb})),$$

where C_{jb} *is a constant matrix of size* $m \times m$ *only associated with the basis matrix* M_{jb}.

Proof. For each fixed j, we have

$$\begin{aligned}
(U_i - V_i\boldsymbol{\alpha}_0)^T V_{i,jb} &= Tr\{{V_{i,jb}}^T(U_i - V_i\boldsymbol{\alpha}_0)\} \\
&= Tr[\dot{\mu}_i^T(\hat{\beta})A_i(\hat{\beta})^{-1/2}M_{jb}A_i(\hat{\beta})^{-1/2}\{y_i - \mu_i(\hat{\beta})\}\{y_i - \mu_i(\hat{\beta})\}^T \\
&\quad A_i(\hat{\beta})^{-1/2}(\tilde{R}^{-1} - R^{-1})A_i(\hat{\beta})^{-1/2}\dot{\mu}_i(\hat{\beta})] \\
&= Tr[(\tilde{R}^{-1} - R^{-1})A_i(\hat{\beta})^{-1/2}\dot{\mu}_i(\hat{\beta})\dot{\mu}_i^T(\hat{\beta})A_i(\hat{\beta})^{-1/2}M_{jb} \\
&\quad A_i(\hat{\beta})^{-1/2}\{y_i - \mu_i(\hat{\beta})\}\{y_i - \mu_i(\hat{\beta})\}^T A_i(\hat{\beta})^{-1/2}].
\end{aligned}$$

It follows from the above that

$$\begin{aligned}
(U_i - V_i\boldsymbol{\alpha}_0)^T V_{i,jb} &= \texttt{Tr}\{(\tilde{R}^{-1} - R^{-1})T_{ij}(\hat{\beta})\} \\
&= [\text{vec}\{T_{ij}(\hat{\beta})\}]^T\sqrt{n}\text{vec}(\tilde{R}^{-1} - R^{-1}),
\end{aligned}$$

where

$$\begin{aligned}
T_{i,jb}(\hat{\beta}) &= A_i(\hat{\beta})^{-1/2}\dot{\mu}_i(\hat{\beta})\dot{\mu}_i^T(\hat{\beta})A_i(\hat{\beta})^{-1/2}M_{jb}A_i(\hat{\beta})^{-1/2} \\
&\quad \{y_i - u_i(\hat{\beta})\}\{y_i - \mu_i(\hat{\beta})\}^T A_i(\hat{\beta})^{-1/2}. \qquad (3.9)
\end{aligned}$$

It can be seen that $T_{i,jb}(\hat{\beta})$ are i.i.d. and are only associated with M_{jb}. By condition 4, it can be concluded

that $T_{i,jb}(\hat{\beta}) = O_p(1)$. From (3.9), we have

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(U_i - V_i\boldsymbol{\alpha}_0)^T V_{i,jb} &= \mathtt{Tr}\{n^{1/2}(\tilde{R}^{-1} - R^{-1})\frac{1}{n}\sum_{i=1}^{n} T_{ij}(\hat{\beta})\} \\ &= \left[\mathrm{vec}\{\sum_{i=1}^{N} T_{ij}(\hat{\beta})\}\right]^T \sqrt{n}\mathrm{vec}(\tilde{R}^{-1} - R^{-1}).\end{aligned}$$

By the weak law of large numbers, for any β, $\frac{1}{n}\sum_{i=1}^{n} T_{i,jb}(\beta)$ converges to some function $C_{jb}(\beta)$. Because the GEE estimator is consistent $\hat{\beta}$ is consistent with β_0, it follows from the continuous mapping theorem that

$$\frac{1}{n}\sum_{i=1}^{n} T_{i,jb}(\hat{\beta}) \to C_{jb},$$

where C_{jb} is a constant matrix only associated with the basis matrix M_{jb}. By Condition 1, Condition 4 and $n^{1/2}(\tilde{R}^{-1}(\hat{\beta}) - R^{-1}) = O_p(1)$, it can be concluded that $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(U_i - V_i\boldsymbol{\alpha}_0)^T V_{i,jb} = O_p(1)$. It follows from Lemma 1 that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(U_i - V_i\boldsymbol{\alpha}_0)^T V_{i,jb} \xrightarrow{d} N(0, \mathrm{vec}(C_{jb})\Sigma_m\mathrm{vec}(C_{jb})^T),$$

where Σ_m is defined in Lemma 1.

□

3.9.3 Proof of Theorem 1

Define

$$L(\boldsymbol{\alpha}) = \sum_i ||U_i - V_i\boldsymbol{\alpha}||^2,$$

and s_m as the number of groups of basis matrices with non-zero coefficients.

Proof. Let $\delta_n = \sqrt{p_m}(n^{-1/2} + a_n)$, then

$$\begin{aligned}Q(\boldsymbol{\alpha}_0) - Q(\boldsymbol{\alpha}_0 + \delta_n u) &= L(\boldsymbol{\alpha}_0) - L(\boldsymbol{\alpha}_0 + \delta_n u) \\ &\quad -n\dim(\beta)\sum_{j=1}^{s_m}\{p_{\lambda_n}(||\boldsymbol{\alpha}_{0j} + \delta_n u||) - p_{\lambda_n}(||\boldsymbol{\alpha}_{0j}||)\} \\ &= I + II.\end{aligned}$$

By Taylor's expansion, I can be extended as

$$\begin{aligned} I &= 2\delta_n[\sum_{i=1}^{n}(U_i - V_i\boldsymbol{\alpha}_0)^T V_i]u - \delta_n^2 u^T[\sum_{i=1}^{n} V_i^T V_i]u \\ &= I_1 + I_2. \end{aligned}$$

By Lemma 1 and the Cauchy-Schwarz inequality,

$$|I_1| \leq 2|\delta_n| O_p(\sqrt{np_m})||u|| = O_p(\delta_n^2 n)||u||.$$

By condition 3, with probability tending to 1

$$I_2 = -\delta_n^2 u^T[\sum_{i=1}^{n} V_i^T V_i]u \leq -nl_1\delta_n^2||u||^2.$$

Then we can apply Taylor expansion to II to have

$$\begin{aligned} II &= \sum_{j=1}^{s_m}\sum_{b=1}^{B_j}[n\delta_n p'_{\lambda_n}(|\alpha_{0jb}|)\mathtt{sgn}(\alpha_{0jb})u_j - n\delta_n^2 p''_{\lambda_n}(\alpha_{0jb})u_j^2\{1+o(1)\}] \\ &= II_1 + II_2. \end{aligned}$$

The upper bound of terms II_1 and II_2 can be derived,

$$\begin{aligned} |II_1| &\leq \sum_{j=1}^{s_m}\sum_{b=1}^{B_j}|n\delta_n p'_{\lambda_n}(|\alpha_{0jb}|)\mathtt{sgn}(\alpha_{0jb})u_j| \leq \sqrt{s_m}n\delta_n a_n||u|| \leq n\delta_n^2||u||, \\ II_2 &\leq \sum_{j=1}^{s_m}\sum_{b=1}^{B_j} n\delta_n^2 p''_{\lambda_n}(\alpha_{0jb})u_j^2\{1+o(1)\} \leq 2\max_{1\leq j\leq p_m} p''_{\lambda_n}(|\alpha_0^j|)n\delta_n^2||u||^2. \end{aligned}$$

By the regularity condition 1.b on the penalty function, $II_2 \to 0$. Therefore, allowing $||u||$ to be large enough, I_1, II_1 and II_2 are dominated by I_2 which is negative. It follows that

$$p\{\inf_{||u||=c} Q(\boldsymbol{\alpha}_0 + \delta_n u) > Q(\boldsymbol{\alpha}_0)\} \geq 1 - \varepsilon. \tag{3.10}$$

The consistency follows from (3.10). □

3.9.4 Proof of Theorem 2

Proof. Consider an $\boldsymbol{\alpha}$ such that $||\boldsymbol{\alpha}_2|| < C\sqrt{p_m/n}$,

$$\begin{aligned}\frac{\partial Q(\boldsymbol{\alpha})}{\partial \alpha_{jb}} &= -\sum_{i=1}^{n} 2(U_i - V_i\boldsymbol{\alpha}_0)^T V_{i,jb} + \sum_{i=1}^{n} 2(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^T V_i^T V_{i,jb} + n\dim(\beta) p'_{\lambda_n}(|\alpha_{jb}|)\mathtt{sgn}(\alpha_{jb}) \\ &= I + II + III.\end{aligned}$$

By Lemma 1, $I = O_p(\sqrt{n})$. And further $I = O_p(\sqrt{np_m})$ because $np_m \geq n$. For II, By condition 3, we have

$$||V_i^T V_{i,jb}|| = O_p(1), \tag{3.11}$$

and by (3.11) and the convergence rate obtained in Theorem 1,

$$II < nO_p(\sqrt{p_m/n}) = O_p(\sqrt{np_m}).$$

Therefore, for $j = s_m + 1, \ldots, J_m$ and $b = 1, \ldots, B_j$

$$\begin{aligned}\frac{\partial Q(\boldsymbol{\alpha})}{\partial \alpha_{jb}} &= O_p(\sqrt{np_m}) + n\dim(\beta) p'_{\lambda_n}(|\alpha_{jb}|)\mathtt{sgn}(\alpha_{jb}) \\ &= n\lambda_n\{p'_{\lambda_n}(|\alpha_{jb}|)\mathtt{sgn}(\alpha_{jb})/\lambda_n + O_p(\sqrt{p_m/n})/\lambda_n\}.\end{aligned}$$

From the condition $\sqrt{p_m/n}/\lambda_n \to 0$, and $\liminf_{n\to\infty} \liminf_{\theta\to 0^+} p'_{\lambda_n}(\theta)/\lambda_n > 0$, the sign of α_{jb} determines the sign of $\partial Q(\boldsymbol{\alpha})/\partial\alpha_{jb}$, i. e.

$$\begin{aligned}&\frac{\partial Q(\boldsymbol{\alpha})}{\partial \alpha_{jb}} < 0, \quad \mathtt{for} \ -c\sqrt{p_m/n} \leq \alpha_{jb} < 0, \\ \text{and} \quad &\frac{\partial Q(\boldsymbol{\alpha})}{\partial \alpha_{jb}} > 0, \quad \mathtt{for} \ c\sqrt{p_m/n} \geq \alpha_{jb} > 0\end{aligned}$$

for $j = s_m + 1, \ldots, J_m$ and $b = 1, \ldots, B_j$. Hence, Theorem 2 follows from the above. □

3.9.5 Proof of Theorem 3

The first derivative of the objective function with respect to $\boldsymbol{\alpha}_{01}$ is zero at $\hat{\boldsymbol{\alpha}}_{01}$, i.e. $\nabla Q(\hat{\boldsymbol{\alpha}}_{01}) = 0$. Let V_i^1 be the submatrix of V_i that corresponds with the non-zero coefficients $\boldsymbol{\alpha}_{01}$. By applying Taylor's expansion

of $\nabla Q(\hat{\boldsymbol{\alpha}}_{01})$ on point $\boldsymbol{\alpha}_{01}$, We haev

$$\begin{aligned}&\frac{1}{n}\left\{\sum_{i=1}^{n}(V_i^1)^T V_i^1 - \nabla^2 P_{\lambda_n}(\boldsymbol{\alpha}_{01}^*)\right\}(\hat{\boldsymbol{\alpha}}_{01}-\boldsymbol{\alpha}_{01}) + \frac{1}{n}\nabla P_{\lambda_n}(\boldsymbol{\alpha}_{01}) \\ &= -\frac{2}{n}\left\{\sum_{i=1}^{n}(U_i - V_i^1\boldsymbol{\alpha}_{01})^T V_i^1\right\},\end{aligned} \tag{3.12}$$

where $\boldsymbol{\alpha}_{01}^*$ is between $\boldsymbol{\alpha}_{01}$ and $\hat{\boldsymbol{\alpha}}_{01}$. First by regularity condition 1.d on the penalty function and the consistency result obtained in part I,

$$\Lambda_i\left[\frac{1}{n}\{\nabla^2 P_{\lambda_n}(\boldsymbol{\alpha}_{01}^*) - \nabla^2 P_{\lambda_n}(\boldsymbol{\alpha}_0)\}\right] = O_p(\sqrt{p_m/n}),$$

where $\Lambda_i(M)$ is the ith eigen value of the symmetric matrix M. Therefore, it can be concluded that

$$\frac{1}{n}\{\nabla^2 P_{\lambda_n}(\boldsymbol{\alpha}_{01}^*) - \nabla^2 P_{\lambda_n}(\boldsymbol{\alpha}_0)\}(\hat{\boldsymbol{\alpha}}_{01}-\boldsymbol{\alpha}_{01}) = O_p(p_m/n) = o_p(1/\sqrt{n}). \tag{3.13}$$

And then by (3.12) and (3.13), we have

$$\begin{aligned}&\frac{1}{n}\left\{\sum_{i=1}^{n}(V_i^1)^T V_i^1 + \nabla^2 P_{\lambda_n}(\boldsymbol{\alpha}_{01})\right\}(\hat{\boldsymbol{\alpha}}_{01}-\boldsymbol{\alpha}_{01}) + \frac{1}{n}\nabla P_{\lambda_n}(\boldsymbol{\alpha}_{01}) \\ &= -\frac{2}{n}\left\{\sum_{i=1}^{n}(U_i - V_i^1\boldsymbol{\alpha}_{01})^T V_i^1\right\} + o_p(1/\sqrt{n}).\end{aligned} \tag{3.14}$$

Define

$$I_{n,11} = \frac{1}{n}\sum_{i=1}^{n}(V_i^1)^T V_i^1,$$

$$L_{n,11} = -\frac{2}{n}\sum_{i=1}^{n}(U_i - V_i^1\boldsymbol{\alpha}_{01})^T V_i^1.$$

Let

$$K_m = \mathcal{C}_m \Sigma_m \mathcal{C}_m^T,$$

where $\mathcal{C}_m = \{\mathrm{vec}(C_1)^T, \ldots, \mathrm{vec}(C_m)\}^T$ and Σ_m and $C_j,\ j = 1, \ldots, m$ are defined in Lemma 2. Let $K_{m,11}$,

$\mathcal{C}_{m,11}$ and $\Sigma_{m,11}$ be the parts of K_m, $\mathcal{C}_m$ and Σ_m associated with $\boldsymbol{\alpha}_{01}$. By (3.14), we have

$$\begin{aligned}
&\sqrt{n}A_m K_{m,11}^{-1/2}\{I_{n,11} + \frac{1}{n}\nabla^2 P_{\lambda_n}(\boldsymbol{\alpha}_{01})\}(\hat{\boldsymbol{\alpha}}_{01} - \boldsymbol{\alpha}_{01}) + \frac{1}{\sqrt{n}}A_m K_{m,11}^{-1/2}\nabla P_{\lambda_n}(\boldsymbol{\alpha}_{01}) \\
&= \sqrt{n}A_m K_{m,11}^{-1/2} L_{n,11} + o_p(1).
\end{aligned}$$

By the condition $A_m A_m^T \to G$, $A_m K_{m,11}^{-1/2}\mathcal{C}_{m,11}\Sigma_{m,11}\mathcal{C}_{m,11}^T K_{m,11}^{-1/2} A_m^T = A_m A_m^T \to G$. Therefore, it follows from Lemma 2 that $\sqrt{n}A_m K_{m,11}^{-1/2} L_{n,11}$ is asymptotic normal with the variance

$$cov(\sqrt{n}A_m K_{m,11}^{-1/2} L_{n,11}) = A_m A_m^T \to G.$$

References

Andrews, D. (1988). Laws of large numbers for dependent non-identically distributed random variables. *Econometric Theory 4*(3), 458–467.

Bickel, P. and E. Levina (2008a). Covariance regularization by thresholding. *The Annals of Statistics 36*(6), 2577–2604.

Bickel, P. and E. Levina (2008b). Regularized estimation of large covariance matrices. *The Annals of Statistics 36*(1), 199–227.

Booth, J. and J. Hobert (1999). Maximizing generalized linear mixed model likelihoods with an automated monte carlo em algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61*(1), 265–285.

Breslow, N. and D. Clayton (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association 88*(421), 9–25.

El Karoui, N. (2008). Spectrum estimation for large dimensional covariance matrices using random matrix theory. *The Annals of Statistics 36*(6), 2757–2790.

Fan, J., Y. Fan, and J. Lv (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics 147*(1), 186–197.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

Fan, J. and H. Peng (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics 32*(3), 928–961.

Fisher, T. and X. Sun (2010). Improved stein-type shrinkage estimators for the high-dimensional multivariate normal covariance matrix. *Computational Statistics and Data Analysis 55*(5), 1909–1918.

Fu, W. (2003). Penalized estimating equations. *Biometrics 59*(1), 126–132.

Green, P. (1987). Penalized likelihood for general semi-parametric regression models. *International Statistical Review/Revue Internationale de Statistique 55*(3), 245–259.

Hansen, L. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society 50*(4), 1029–1054.

Huang, J., L. Liu, and N. Liu (2007). Estimation of large covariance matrices of longitudinal data with basis function approximations. *Journal of Computational and Graphical Statistics 16*(1), 189–209.

Huang, J., N. Liu, M. Pourahmadi, and L. Liu (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika 93*(1), 85–98.

Jiang, J. (1999). Conditional inference about generalized linear mixed models. *The Annals of Statistics 27*(6), 1974–2007.

Jiang, J. and W. Zhang (2001). Robust estimation in generalised linear mixed models. *Biometrika 88*(3), 753–765.

Laird, N. and J. Ware (1982). Random-effects models for longitudinal data. *Biometrics 38*(4), 963–974.

Levina, E., A. Rothman, and J. Zhu (2008). Sparse estimation of large covariance matrices via a nested lasso penalty. *The Annals of Applied Statistics 2*(1), 245–263.

Liang, K. and S. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika 73*(1), 13–22.

Liu, Q. and D. Pierce (1994). A note on gausshermite quadrature. *Biometrika 81*(3), 624–629.

McCulloch, C. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American statistical Association 92*(437), 162–170.

McCulloch, C. and J. Neuhaus (2001). *Generalized linear mixed models*. Wiley Online Library.

Molenberghs, G. and G. Verbeke (2005). *Models for discrete longitudinal data*. New York: Springer Verlag.

Neuhaus, J., W. Hauck, and J. Kalbfleisch (1992). The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika 79*(4), 755–762.

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics 12*(2), 758–765.

Ortega, J. M. and W. C. Rheinboldt (1973). *Iterative solution of nonlinear equations in several variables*. Academic Press.

Prentice, R. and L. Zhao (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics 47*(3), 825–839.

Qu, A., B. Lindsay, and B. Li (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika 87*(4), 823–836.

Qu, A. and P. Song (2004). Assessing robustness of generalised estimating equations and quadratic inference functions. *Biometrika 91*(2), 447–459.

Rothman, A., P. Bickel, E. Levina, and J. Zhu (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics 2*, 494–515.

Rothman, A., E. Levina, and J. Zhu (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association 104*(485), 177–186.

Schabenberger, O. (2005). Introducing the glimmix procedure for generalized linear mixed models. In *SUGI*, Volume 30, pp. 196–226.

Stoner, J. A. (2000). *Analysis of clustered data: A combined estimating equation approach*. Ph. D. thesis, University of Washington.

Thall, P. and S. Vail (1990). Some covariance models for longitudinal count data with overdispersion. *Biometrics 46*(3), 657–671.

Vonesh, E., H. Wang, L. Nie, and D. Majumdar (2002). Conditional second-order generalized estimating equations for generalized linear and nonlinear mixed-effects models. *Journal of the American Statistical Association 97*(457), 271–283.

Wang, L. and A. Qu (2009). Consistent model selection and data-driven smooth tests for longitudinal data in the estimating equations approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(1), 177–190.

Wang, N. (2003). Marginal nonparametric kernel regression accounting for within-subject correlation. *Biometrika 90*(1), 43–52.

Wedderburn, R. (1974). Quasi-likelihood functions, generalized linear models, and the gaussnewton method. *Biometrika 61*(3), 439–447.

Xue, L., A. Qu, and J. Zhou (2010). Consistent model selection for marginal generalized additive model for correlated data. *Journal of the American Statistical Association 105*(492), 1518–1530.

Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *The Journal of Machine Learning Research 11*, 2261–2286.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68*(1), 49–67.

Zhang, P., P. Song, A. Qu, and T. Greene (2008). Efficient estimation for patient-specific rates of disease progression using nonnormal linear mixed models. *Biometrics 64*(1), 29–38.

Zhang, Y., R. Li, and C. Tsai (2010). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association 105*(489), 312–323.

Zhou, J. and A. Qu (2011). Model selection of correlation structure for longitudinal data.

Zou, H. and R. Li (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics 36*(4), 1509–1533.

Vita

Peng Wang was born in Shenyang, the capital city of Liaoning Province in China on September 11, 1982. He received his Bachelor of Science degree from Peking University, Beijing, China in 2005. He received his Master of Science in statistics from Oregon State University, Corvallis, Oregon in 2007. Upon the completion of his Ph.D in statistics from the University of Illinois at Urbana-Champaign, Peng Wang will become an Assistant Professor of statistics in the Department of Mathematics and Statistics at Bowling Green State University.