AN ABSTRACT OF THE DISSERTATION OF

Guei-Feng Tsai for the degree of Doctor of Philosophy in Statistics presented on August 29, 2005.

Title: Semiparametric Marginal and Mixed Models for Longitudinal Data

Redacted for Privacy

Abstract approved: _____

Annie Qu

This thesis consists of three papers which investigate marginal models, nonparametric approaches, generalized mixed effects models and variance components estimation in longitudinal data analysis.

In the first paper, a new marginal approach is introduced for high-dimensional cell-cycle microarray data with no replicates. There are two kinds of correlation for cell-cycle microarray data. Measurements within a gene are correlated and measurements between genes are also correlated since some genes may be biologically related. The proposed procedure combines a classifying method, quadratic inference function method and nonparametric techniques for complex high dimensional cell cycle microarray data. The gene classifying method is first applied to identify genes with similar cell cycle patterns into the same class. Then we use genes within the same group as pseudo-replicates to fit a nonparametric model. The quadratic inference function is applied to incorporate within-gene correlations. An asymptotic chi-squared test is also applied to test whether certain genes have cell cycles

phenomena. Simulations and an example of cell-cycle microarray data are illustrated.

The second paper proposes a new approach for generalized linear mixed models in longitudinal data analysis. This new approach is an extension of the quadratic inference function (Qu et al., 2000) for generalized linear mixed models. Two conditional extended scores are constructed for estimating fixed effects and random effects. This new approach involves only the first and second conditional moments. It does not require the specification of a likelihood function and also takes serial correlations of errors into account. In addition, the estimation of unknown variance components associated with random effects or nuisance parameters associated with working correlations are not required. Furthermore, it does not require the normality assumption for random effects.

In the third paper, we develop a new approach to estimate variance components using the second-order quadratic inference function. This is an extension of the quadratic inference function for variance components estimation in linear mixed models. The new approach does not require the specification of a likelihood function. In addition, we propose a chi-squared test to test whether the variance components of interest are significant. This chi-squared test can also be used for testing whether the serial correlation is significant. Simulations and a real data example are provided as illustration.

Semiparametric Marginal and Mixed Models for Longitudinal Data

by

Guei-Feng Tsai

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented August 29, 2005
Commencement June 2006

Doctor of Philosophy dissertation of Guei-Feng Tsai presented on August 29, 2005

APPROVED:

Redacted for Privacy

Major Professor, representing Statistics

Redacted for Privacy

Chair of the Department of Statistics

Redacted for Privacy

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Redacted for Privacy

Guei-Feng Tsai, Author

ACKNOWLEDGMENTS

My primary thanks go to my major professor Dr. Annie Qu. Without her, this dissertation would not have been possible. Her patience and encouragement carried me through difficult times. Throughout my doctoral work she encouraged me to develop independent thinking and research skills. Her inspiration and guidance helped me to shape my research skills. She continually stimulated my analytical thinking and greatly assisted me with scientific writing. She is not only my major professor, but also an excellent mentor for my life and career.

I am grateful to Dr. David S. Birkes, who provided constant help and suggestions in various aspects of my research. He is one whom I can always count on to discuss any details of problems. I would also like to thank Dr. Robert T. Smythe as his careful reading and comments improved the quality of the thesis greatly. I am also grateful to Professor Daniel W. Schafer for helping me as Graduate Advisors, during my initial terms at the department, and who posed valuable questions for the thesis. In addition, I am grateful to my committee members, Dr. Paul A. Murtaugh from the Statistics Department and Dr. Logen R. Logendran from the Industrial and Manufacturing Engineering Department. Many thanks to the faculty and staff for creating a supportive environment.

My thanks also go to my proof-reader, Christopher J. Vecoli, for his patience and support on technical writing. He provided me with valuable advice to improve my English. I also thank Dr. Chi-Chow Liao who encouraged

me to pursue a Ph.D. degree and provided crucial advice and caring support through the course of my study.

As for my families, I want to thank my parents who offered me love, support and faith through the course of my study and my life. Many thanks are due to my parents-in-law for providing me advice and financial support. I owe my loving thanks to my husband Chi-Shen Chen, and my daughters Anita Chen and Emily Chen. They provided me the inspiration and moral support for the study. Without their encouragement and understanding it would have been impossible for me to finish this work.

CONTRIBUTION OF AUTHORS

Professor Annie Qu proposed the original questions that motivated this research. Professor Qu was also involved in the solutions and editing of the manuscript.

Professor Annie Qu contributed helpful suggestions and review as co-author on Chapter Two and Chapter Three.

Professor Annie Qu and Professor David S. Birkes contributed helpful suggestions and reviews as co-authors on Chapter Four.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

LIST OF TABLES

Semiparametric Marginal and Mixed Models for Longitudinal Data

1. INTRODUCTION

Longitudinal data arise frequently in biomedical research, social sciences and environmental studies, where measurements are obtained from a subject repeatedly over time. Therefore, the observations within a subject are correlated. In order to draw valid inferences, this correlation must be taken into account. In longitudinal data, there are three different sources of random variation associated with random effects of subjects, serial correlations of errors and measurement errors (Diggle et al., 2002, section 5.2). Random effects explain the variation between subjects. Serial correlations of errors arise due to time course measurements. In this thesis, we will not consider the variation from measurement errors. Major statistical models such as marginal models and mixed effects models have been proposed for the analysis of longitudinal data. General theory, methods and applications of marginal models and mixed effects models for longitudinal data can be found in Verbeke and Molenberghs (2000) and Diggle et. al (2002).

Marginal models are applicable when the inference on the population average is of primary interest. Generalized estimating equation (GEE) is one widely used marginal model, which extended the quasi-likelihood approach to correlated data. The main advantage of GEE is that it requires only the first

two moments of the distribution and uses a working correlation matrix to avoid the specification of correlation between observations within a subject. The GEE provides consistent estimators for regression parameters even when the working correlation is misspecified. However, misspecification of the working correlation may lead to inefficient estimators of regression parameters.

Several methods have been proposed to improve the generalized estimating equation. Prentice and Zhao (1991) proposed an ad hoc method (GEE1) which introduces additional estimating equations for the working correlation parameter α. The estimators of regression parameters are obtained by solving the estimating equations associated with the regression parameters and the working covariance matrix. Although GEE1 is more efficient than GEE for the nuisance parameter α, it may be less robust than GEE for regression parameters when the working correlation is misspecified (Qu, 1998). Qu et al. (2000) proposed the quadratic inference function (QIF) which requires the first two moments of observations. The main advantage of QIF is that it does not involve the estimation of nuisance parameters α. It has been shown that the QIF estimators are consistent and more efficient than the GEE estimators when the working correlation is misspecified (Qu et al., 2000).

Although the marginal models are easily implemented, mixed effects models are more appropriate while the subject variation is of interest. In addition, the variation from random effects could be combined additively with serial correlation. Thus, the same marginal model may have different interpretations, as it may lead to different random effects models if different decom-

positions of variations are applied. In contrast, mixed effects models are able to incorporate several sources of variation.

If data are continuous and normal, maximum likelihood estimation along with the Bayesian approach can be used to estimate fixed effects and random effects. Nevertheless, in longitudinal studies, non-normal data arise frequently. A generalized linear mixed model (GLMM) is widely used for non-normal data while random effects are considered. In GLMM, the likelihood function often involves high dimensional integration. Gaussian-Hermite quadrature (Liu and Pierce, 1994) and Monte Carlo EM algorithm (McCulloch, 1997) can be applied to approximate the integration over the distributions of random effects. However, these methods require the specific distribution of random effects. Penalized quasi-likelihood (PQL) (Breslow and Clayton,1993) is an alternative to the numerical integrations when the likelihood function is not tractable. In PQL, the regression parameters estimation is performed by solving a set of conditional estimating equations. However, the estimating equations involve unknown variance components. In addition, it requires the normality assumption for random effects.

This thesis investigates both marginal models and mixed-effects models in longitudinal data analysis. In Chapter 2, we apply the marginal models for high dimensional cell cycle microarray data with no replicates. The cell-cycle microarray data can be considered as longitudinal data because gene expression levels are measured over multiple time points during the progression of cell cycle. The challenge of analyzing cell-cycle gene expression data is that the data is highly dimensional and inter-correlated. In this chapter, we

propose a new approach which utilizes a data classifying method, quadratic inference function method and nonparametric techniques. The following is our strategy: we first classify genes into the same class with similar cell cycle patterns which also include a class with no cell cycle phenomena at all. Then, we fit a nonparametric model using genes within the same group as pseudo-replicates for inference. In order to incorporate correlations of longitudinal measurements, the quadratic inference function is applied. This allows us to perform chi-squared tests for testing whether the coefficients are time varying in order to determine whether certain genes have cell-cycle phenomena or not.

Chapter 3 introduces a new approach for generalized linear mixed models in longitudinal data analysis. This new approach is an extension of the quadratic inference function for generalized linear mixed models. Two conditional extended scores are constructed for estimating fixed effects and random effects. The new approach requires only the first and second conditional moments. It does not require the specification of a likelihood function and also takes serial correlations into account. In addition, the estimation of unknown variance components or nuisance parameters associated with working correlations are not involved. Furthermore, it does not require the normality assumption of random effects.

In Chapter 4, we propose the second-order extended scores for the estimation of variance components. This is an extension of the quadratic inference function for variance components in linear mixed models. We first estimate the fixed effects using the first-order quadratic inference function which involves the first and second moments. Then we estimate the variance

components using the the first-order QIF estimators and the second-order quadratic inference function which involves the third and fourth moments. The advantage of the new approach is that the first-order quadratic inference function does not involve the unknown variance components. In this paper, we incorporate the serial correlation into our model. In addition, we propose an asymptotic chi-squared test to test the variance components of interest. This chi-squared test also allows us to test whether the serial correlation is significant.

The final chapter provides the conclusions and directions for future work.

2. SEMI-NONPARAMETRIC MODELS AND INFERENCE FOR HIGH DIMENSIONAL MICROARRAY DATA

Guei-Feng C. Tsai

Annie Qu

Department of Statistics, Oregon State University

2.1. Abstract

We develop a new approach to analyze high dimensional cell-cycle microarray data with no replicates. There are two kinds of correlations for cell-cycle microarray data. Measurements are correlated within a gene, and measurements are also correlated between genes since some genes may be biologically related. The proposed procedure combines a classification method, the quadratic inference function method and nonparametric techniques for complex high dimensional data. We first perform a gene classifying analysis to classify genes into classes with similar cell-cycle patterns, including a class with no cell-cycle phenomena at all. We use genes within the same group as pseudo-replicates to build nonparametric models and inference functions. In order to incorporate correlation of longitudinal measurements, the quadratic inference function method is also applied. This approach allows us to perform chi-squared tests for testing whether the coefficients are time varying or not. This also allows us to determine whether certain genes regulate cell cycles. A real data example on cell-cycle microarray data as well as simulations are illustrated.

2.2. Introduction

Microarray has become a powerful tool in molecular biology because it can measure gene expression levels for thousands of genes simultaneously. This allows scientists to discover the latent genetic causes of many diseases. Regulation of cell cycles plays an important role in the normal development of multicellular organisms, because diseases such as cancer are a consequence of uncontrolled cell growth. Therefore, researchers are interested in finding the genes that are involved in the regulation of cell cycles. In this paper we are particularly interested in problems arising from cell-cycle microarray data, where gene expression levels are measured over multiple time points during the progression of the cell cycle.

Cell-cycle microarray studies have been primarily performed on budding yeast (*Saccharomyces cerevisiae*) since it can be controlled easily in a lab setting. There are 6178 genes in a budding yeast cell of which 104 genes were identified to be cell-cycle regulated through traditional lab methods. Cho et al. (1998) and Spellman et al. (1998) employed DNA microarrays to measure expression levels of genes over time where cell cycles might occur in yeast cells.

Clustering is a commonly used method for grouping genes with similar expression patterns. For example, there are the K-means method (Tavazoie et al., 1999) and the self-organizing maps approach (Tomayo et al., 1999). Although these methods allow the visual detection of expression patterns, they do not provide numerical evidence to determine genes whose expression levels are cell-cycle regulated.

Other approaches have been derived for identifying genes whose mRNA levels are regulated during the cell cycle. Spellman et al. (1998) applied Fourier transformation to obtain numerical scores (CDC score) for each gene and selected a threshold CDC score based on the 104 known genes to identify cell-cycle regulated genes. Zhao et al. (2001) developed a single pulse model (SPM) which assumed that cell-cycle regulated genes peak at one invariant time and deflate at subsequent time during the cell cycle. However, these two approaches may be too simple for biological systems, since gene expressions in their approach are treated with evenly spaced peaks and valleys (Li et al., 2002).

Li et al. (2002) applied a principal component analysis along with a nested model to identify CDC15 genes with cyclic expression patterns. However, this approach becomes inappropriate for other genes if the cyclic pattern cannot be detected in the first three basis curves found in principal component analysis. Lian and Li (2004) applied a shape-invariant model with a cubic B-spline function based on known periodic genes (guide genes) and employed the false discovery rate procedure (Benjamini and Hochberg, 1995) to identify genes whose expression patterns are similar to the guide genes. Nevertheless, this approach cannot be applied to other datasets where pre-biological knowledge, such as concerning guide genes, is unknown.

The challenge of analyzing cell-cycle gene expression data is that the data is highly dimensional and inter-correlated. In fact, there exist two kinds of correlations in cell-cycle microarray data. Measurements that are obtained within a gene are certainly correlated since they are repeatedly measured over

time. Measurements between genes might also be correlated since some genes are biologically related.

We first perform data grouping analysis to classify genes with similar cell cycle patterns into the same class, or into a class with no cell cycle phenomena at all. This enables us to use genes within the same class as pseudo-replicates. Secondly, in order to investigate whether expression levels change during the cell cycle we develop marginal varying coefficient models since they provide flexible statistical models and take both time and covariate effects into considerations. In addition, we apply the quadratic inference function method to incorporate the within-gene correlations as they are measured over time. This approach also allows us to test whether genes have cell-cycle phenomena or not.

This paper will proceed as follows. In the next section we describe the data grouping technique to classify genes into groups with similar expression patterns in cell-cycle microarray data. Nonparametric varying coefficient modeling is discussed in Section 2.4. Section 2.5 discusses the quadratic inference function proposed by Qu et al. (2000). Simulation results are illustrated in section 2.6. Section 2.7 demonstrates a real cell-cycle microarray data example using our approach. A brief discussion is provided in Section 2.8.

2.3. Gene classifying analysis in cell-cycle microarray data

A gene is a template to form a protein which is the basic unit of genetic function, so genes play an essential role in living organisms. Traditional methods in molecular biology usually examine one gene in one experiment, which

means that interactions among genes are barely studied. Microarray scans a whole genome on a single or few chips so that scientists can have an entire perspective of the interactions among thousands of genes simultaneously.

The primary goal of cell-cycle microarray experiments is to investigate whether gene expressions change over time. Clustering is a common approach to group genes with similar expression level patterns. This allows us to visualize data; however, it cannot provide quantitative justification to identify cell-cycle regulated genes.

Cell-cycle microarray data is difficult for statistical modeling because there are large numbers of parameters associated with genes, yet there are no replications; and furthermore, cell-cycle patterns also vary for different groups of genes. For that reason, we first classify genes with similar expression patterns into the same group so that each group has homogeneous patterns.

The gene grouping process begins with primary (target) genes selection. One can refer to pre-screening results and interests or previous knowledge for selection of primary genes. For example, in yeast cells, gene YKL185W has been identified to be a cell-cycle regulated gene. We might find other genes whose expression profiles over time are similar to YKL185W so these genes can be grouped into the same class in order to perform further statistical analysis.

To quantify the distance between a primary gene and other gene expressions, we use two kinds of measurements, short time-series (STS) distance and squared Euclidean distance. The STS distances measure whether the shapes of two expression curves are similar or not. The squared Euclidean distances measure the differences of amplitude between two gene expression levels.

The STS distance was first introduced by Möller-Levet et al. (2003), which measures similarity of shape or slope for time-series data. Specifically, the STS distance between two time-series $u = [u_1, \ldots, u_{n_t}], v = [v_1, \ldots, v_{n_t}]$ is defined as

$$d^2_{STS} = \sum_{k=0}^{n_t-1} \left(\frac{v_{k+1} - v_k}{t_{k+1} - t_k} - \frac{u_{k+1} - u_k}{t_{k+1} - t_k} \right)^2, \tag{2.1}$$

where t_k represents successive time points, n_t is the number of time points, $v_0 = 0$, $x_0 = 0$ and $t_0 = 0$.

Although we can use STS distances to find genes with similar expression patterns, STS measurements cannot distinguish between genes whose amplitudes of expression levels are different. Therefore, we also use Euclidean distances to distinguish the differences of amplitude between two gene expression levels. The squared Euclidean distance between two time-series $u = [u_1, \ldots, u_{n_t}], v = [v_1, \ldots, v_{n_t}]$ is defined as

$$d^2_E = \sum_{k=1}^{n_t} (v_k - u_k)^2, \tag{2.2}$$

where n_t is the number of time points.

We calculate the STS distances and squared Euclidean distances between expression levels of a primary gene and other genes. In addition, standardization of distances are applied to adjust the different scales between STS distances and squared Euclidean distances. Finally, we combine STS and squared Euclidean distances which are sums of these two.

Consider a primary gene $\mathbf{z_1}$ and other genes $\mathbf{z_2}, ..., \mathbf{z_n}$ with the expression levels $\{z_{11}, ..., z_{1t}\}$ and $\{z_{21}, ..., z_{2t}\}$, $\cdots$, $\{z_{n1}, \cdots, z_{nt}\}$, respectively. We are interested in finding genes whose expression levels are similar to gene $\mathbf{z_1}$

and allocating them into the group $\mathbf{z_1}$. Table 2.1 summarizes the distance matrix of STS distances and squared Euclidean distances between gene $\mathbf{z_1}$ and other genes $\mathbf{z_2}, ..., \mathbf{z_n}$. Define S_{i1} $(i = 1, ..., n)$ and E_{i1} as STS distances and squared Euclidean distances between gene $\mathbf{z_i}$ $(i = 2, ..., n)$ and $\mathbf{z_1}$, which can be obtained by (2.1) and (2.2), respectively. Note that S_{11} and E_{11} are zeros, and S_{i1}^N and E_{i1}^N indicate the standardized STS distances and squared Euclidean distances between genes $\mathbf{z_i}$ and $\mathbf{z_1}$, respectively. We denote D_{i1}^o as the overall distance, that is $D_{i1}^o = S_{11}^N + E_{11}^N$. The final step of the grouping process is to assign h genes, including the primary gene, into group $\mathbf{z_1}$ where these h genes have the smallest overall distances. After this classifying step, genes with similar expression patterns and amplitudes will be categorized in the same group.

To illustrate this grouping strategy, we apply our classifying procedure to the yeast CDC15 dataset. Two primary genes are selected, including YKL185W and YAL001C. We assign 20 genes into each group to ensure that the genes within that group have the most similar expression levels. We will discuss the choice of number of genes within a group in Section 2.5.

By visualization Figure 2.3 shows that genes in the YKL185W group might be cell-cycle regulated, and Figure 2.4 shows that genes in the YAL001C group may not have the cell cycle phenomenon. To quantify our findings, we need to develop a statistical model and test whether gene expressions have cyclic patterns. The above grouping techniques will provide pseudo-replicates to perform further statistical analysis and modeling.

2.4. Nonparametric time-varying coefficient models

Once genes have been categorized into the same class, we consider genes within the same class as pseudo-replicates and fit a nonparametric time-varying coefficient model in order to identify patterns of gene expression.

Nonparametric approaches have been extensively studied when parametric assumptions are not valid or a meaningful parametric model cannot be determined. In particular, nonparametric time-varying coefficient models not only provide flexible statistical models but also take both time and covariates into account.

Varying coefficient models were introduced by Hastie and Tibshirani (1993) which allow regression coefficients to vary as a function of other factors. Zeger and Digger (1994) and Moyeed and Digger (1994) considered a semiparametric model which allowed the intercept to be time varying. Hoover et al. (1998) further extended it to a linear time-varying coefficient model for longitudinal data.

Suppose $Y_i(t)$, $i = 1, ..., N$, is the response of the ith subject obtained at time t, and $X_i(t)$ is the vector of covariates for the ith subject at time t. A time-varying coefficient model can be presented as

$$Y_i(t) = X_i(t)\beta(t) + \varepsilon_i(t), \tag{2.3}$$

where $\beta(t) = (\beta_0(t), \ldots, \beta_u(t))'$ are smooth functions of t and $\varepsilon_i(t)$ is a mean zero random process.

In order to obtain a flexible varying coefficient model, we consider a q degree polynomial truncated spline basis function with knots which can be

written (Ruppert and Carroll, 1997) as

$$\beta_u(t) = \beta_{u0} + \beta_{u1}t + \cdots + \beta_{uq}t^q + \sum_{k=1}^{K_u} \beta_{u(q+k)}(t - \kappa_k)_+^q, \qquad (2.4)$$

where $q \geq 1$ is an integer, K_u is the number of knots, and $\kappa_1 < \cdots < \kappa_{K_u}$ are fixed knots.

In cell-cycle microarray data, since there are no covariates, we fit an intercept-only model with a time-varying coefficient,

$$Y_i(t) = \gamma_0 + \gamma_1 t + \cdots + \gamma_q t^q + \sum_{k=1}^{K} \gamma_{q+k}(t - \kappa_k)_+^q + \varepsilon_i(t), \qquad (2.5)$$

where $\gamma = (\gamma_0, ..., \gamma_q, ..., \gamma_{q+K})$ and $p = q + K + 1$ is the number of parameters.

Huang et al. (2002) fitted a varying coefficient model using weighted least squares. However, the correlation structures for repeated measurements are not incorporated into the models, so their approach may not be efficient. Qu and Li (2005) proposed a new approach which is able to incorporate correlation into the model. Their approach is also applicable to non-normal responses under the generalized linear model framework. We will apply Qu and Li's approach for cell-cycle microarray data. Specifically, in our marginal nonparametric varying coefficient models setting, quadratic inference functions are applied to estimate parameters in order to incorporate correlations of repeated measurements.

2.5. Quadratic inference functions

In this section, we first describe generalized estimating equations and then the quadratic inference function, as both methods take correlations into account.

2.5.1. Generalized estimating equations

Generalized estimating equation (GEE) (Liang and Zeger, 1986) is a widely used model to account for the correlations within a cluster in longitudinal data analysis. In particular, GEE extends quasi-likelihood equations $\sum_{i=1}^{N}(\dot{\mu}_i)'V_i^{-1}(Y_i - \mu_i) = 0$ when $V_i = \text{var}(Y_i)$ is unknown. Here we define $E(Y_i) = \mu_i$, and $\dot{\mu}_i$ is the derivative of μ_i with respect to the parameter γ.

More specifically, the generalized estimating equations are

$$S_\gamma = \sum_{i=1}^{N}(\dot{\mu}_i)'(A_i^{-1/2}R^{-1}A_i^{-1/2})(Y_i - \mu_i) = 0, \tag{2.6}$$

where A_i is a diagonal matrix of variance components of Y_i, and R is the working correlation $W(\alpha)$ which can be specified by a small number of nuisance parameters α.

The estimators $\hat{\gamma}$ obtained from GEE in (2.6) are consistent and asymptotically normally distributed even if the working correlation matrix is misspecified. However, misspecification of the working correlation can cause inefficient estimators of regression parameters.

2.5.2. Quadratic inference functions

Qu et al. (2000) proposed the quadratic inference function (QIF) to improve generalized estimating equations. The quadratic inference function estimators are consistent and more efficient than GEE when the working correlation is misspecified.

Suppose the inverse of the correlation, R^{-1}, can be approximated by a class of known matrices, $R^{-1} \approx \sum_{j=1}^{m} a_j M_j$, where $M_1, \ldots, M_m$ are known

matrices and $a_1, \ldots, a_m$ are unknown constants. Then (2.6) is equivalent to

$$\sum_{i=1}^{N} \dot{\mu_i}' A_i^{-1/2} \left(\sum_{j}^{m} a_j M_j \right) A_i^{-1/2} (Y_i - \mu_i) = 0. \tag{2.7}$$

Note that the left-hand side of (2.7) is a linear combination of $\bar{g}_N$, where

$$\bar{g}_N(\gamma) = \frac{1}{N} \sum_{i=1}^{N} g_i(\gamma) = \frac{1}{N} \begin{pmatrix} \sum_{i=1}^{N} \dot{\mu_i}' A_i^{-1/2} M_1 A_i^{-1/2} (Y_i - \mu_i) \\ \vdots \\ \sum_{i=1}^{N} \dot{\mu_i}' A_i^{-1/2} M_m A_i^{-1/2} (Y_i - \mu_i) \end{pmatrix}. \tag{2.8}$$

Clearly, the dimension of $\bar{g}_N$ in (2.8) is mp and is overidentified as it contains more estimating equations than unknown parameters. Hansen (1982) proposed the generalized method of moments, which can be applied to estimate parameters γ defined in (2.8). That is, $\hat{\gamma}$ is obtained by

$$\min_{\gamma} \bar{g}_N(\gamma) \Sigma_N^{-1} \bar{g}_N(\gamma),$$

where Σ_N is the variance matrix of $\bar{g}_N$. However Σ_N is typically unknown, but can be consistently estimated by $\frac{1}{N} \sum_{i=1}^{N} g_i g_i'$. Qu et. al (2000) defined the quadratic inference function to be

$$Q_N(\gamma) = N \bar{g}_N \bar{C}_N^{-1} \bar{g}_N, \tag{2.9}$$

where $\bar{C}_N = \frac{1}{N} \sum_{i=1}^{N} \mathbf{g_i}(\gamma) \mathbf{g_i}(\gamma)'$. Note that N must be greater than mp to ensure the invertability of the variance matrix $\bar{C}_N$ (Qu and Birkes, 2004). The invertability of variance matrix of estimating equations plays an important role in selecting criteria for determining the number of genes into the same class.

Qu and Li (2005) proposed a new estimating approach which adapted the penalized spline (Ruppert and Carroll, 1997). More specifically, Qu and Li (2005) estimate parameters by minimizing the sum of the quadratic inference function and penalty, where the penalty part is used to control overfitting. For the varying coefficient model in (??) estimate parameters $\gamma = (\gamma_0, ..., \gamma_q, ..., \gamma_{q+K})$ in (2.5) by minimizing

$$Q_N(\gamma) + \alpha \sum_{w=q+1}^{q+K} \gamma_w^2, \tag{2.10}$$

where α is a smoothing parameter. Note that if α is small, it is undersmoothed since it has less shrinkage on a polynomial fit and has more weight on selected knots. Alternatively, if α is large, it is oversmoothed.

The goodness-of-fit test can be applied to determine the number of knots and the degree of the polynomial basis function in (2.4) based on the quadratic inference (Qu and Li, 2005). In addition, the quadratic inference function approach also provides an asymptotic chi-squared test for testing whether some coefficients are time-varying or time invariant. For example, suppose that a group of genes is of interest and we apply a varying coefficient penalized spline model in (2.5) for the mean μ_i. The null hypothesis for testing is

$$H_0 : \gamma_w = 0, w = 1, ..., q + K,$$

which implies that under the null hypothesis, the intercept coefficient is constant over time. The alternative hypothesis is

$$H_1 : \gamma_w \text{ is not all equal to zero}, w = 1, ..., q + K,$$

which means the intercept coefficient might be time-varying. Then a test statistic based on quadratic inference function can be obtained as

$$T = Q(\tilde{\gamma}) - Q(\hat{\gamma}), \tag{2.11}$$

where $\tilde{\gamma} = (\tilde{\gamma}_0, 0, ..., 0)$ and $\hat{\gamma}$ are estimators by minimizing (2.9) under H_0 and H_1, respectively (Qu and Li 2005). The test statistic T also follows a chi-squared distribution with degrees of freedom equal to $q + K$.

2.6. Simulation

In this section we study the test size and power of chi-squared hypothesis tests in (2.11) under a varying coefficient models setting for a finite sample.

The simulation demonstrates the effect of sample size on test size using our approach. First, we calculate the test size under the null hypothesis, where the coefficient in (2.5) is constant over time. We also choose 5 different sample sizes $(50, 100, 150, 200, 250)$ and each subject has 61 repeated measurements at time points t $(t = 0, 1, ..., 60)$. The response variable $y_i(t)$ is generated from a constant intercept 2 and Gaussian random errors with mean zero and variance 1, $y_i(t) = 2 + \varepsilon_i(t)$.

We first fit a 3-degree polynomial spline with 5 equally spaced knots $(10, 20, ..., 50)$. Then we estimate the parameters using the quadratic inference function approach with AR(1) working correlation. Finally, we perform a chi-squared test and calculate test statistics in (2.11). Note that the choice of

the degree of polynomial spline and the number of knots does not affect our results significantly in this simulation.

Table 2.2 shows different test sizes under different sample sizes with 1000 simulations. Table 2.2 shows that the test size is 0.037 when sample size is 50, a more conservative test. Evidently, as sample size increases, the test size also increases. The test size is 0.0495 when the sample size is 200.

Secondly, we examine the power function when a cyclic pattern is presented. We generate samples from a cyclic function with different amplitudes which has the form:

$$f(t) = 2 - \theta \cos\{\frac{\pi}{15}(t - 25)\}, \tag{2.12}$$

where θ is an amplitude parameter indicating the strength of the cyclic pattern. Note that (2.12) is cyclic with period 30, phase shift 25 and vertical shift 2. For example, Figure 2.1 shows the cyclic pattern when θ is 0.5 in (2.12). In each simulation, a simple random sample of 100 subjects is generated according to the model

$$Y_i(t) = 2 - f(t) + \varepsilon_i(t) \qquad i = 1, ..., 100, \quad t = 0, 1, ..., 60, \tag{2.13}$$

where ε_i is a Gaussian process with mean zero and variance 1.

For each simulated dataset, we calculate the quadratic function estimators using the 3-degree polynomial spline with 5 equally spaced knots $(10, 20, ..., 50)$. Then, we perform a chi-squared test and obtain the power. Figure 2.2 illustrates the power curve for different θ, $\theta = (-1, -0.08, ..., 0, 0.02, ..., 1)$. The power function represents the probability of rejecting the null hypothesis as a function of θ. As expected, as the abso-

lute value of θ increases, the probability of rejecting the null hypothesis also increases. The power is close to 0.05 when θ is close to 0, and is close to 1 at $\theta = 0.13$.

2.7. Applications to cell cycle microarray data

In this Section, we apply our approach to a CDC15-synchronized cell cycle microarray dataset (Spellman et al., 1998). They used DNA microarray techniques to scan mRNA levels of yeast genes by the arrest of a cdc15 temperature-sensitive mutant. These data are available at http://cellcycle-www.stanford.edu/.

In the CDC15 dataset, there are a total of 6178 genes and each gene is supposed to have 24 repeated expression measurements at time $t = (10, 30, 50, 70, 80, 90, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200, 210, 220, 230, 240, 250, 270, 290)$ during the progression of the experiment. Note that there is uneven time spacing, and some gene expressions have missing values. We use the complete dataset which contains 4381 genes with all 24 repeated measurements in our analysis. Our approach includes a data classifying procedure, nonparametric time-varying coefficient techniques and the quadratic inference function strategy.

For illustration, two primary genes, namely YKL185W and YAL001C, are selected from the CDC15 dataset. Note that YKL185W is a cell-cycle regulated gene identified using traditional lab methods.

First we obtain the standardized STS and squared Euclidean distances between the chosen genes and others. Then we pick the first 20 genes, in-

cluding the chosen gene, with smaller overall distances among them. Figures 2.3 and 2.4 indicate that grouping technique works well in the identification of similar patterns associated with YKL185W and YAL001C. We call them group YLK185W and group YAL001C. Figure 2.3 shows that the genes among group YKL185W may have cell-cycle phenomena, by visualization. In addition, Figure 2.4 suggests that group YAL001C may not have the cell-cycle feature.

We model the marginal mean of expression levels for group YKL185W and YAL001C with intercept only as time varying coefficient because no additional covariate is available in the CDC15 dataset. We choose the basis function for the intercept as a truncated 3-degree polynomial spline $B_u = (1, t, t^2, t^3, (t-50)_+^3, (t-100)_+^3, (t-150)_+^3, (t-200)_+^3, (t-250)_+^3)$, with a total of 9 parameters including 5 equal spaced knots. Note that the number of knots to be selected can be determined by the goodness-of-fit test (Qu and Li, 2005).

To incorporate the correlations, we apply the quadratic inference function method to estimate the parameters. Because the observation times are not equally spaced, we assume compound symmetric working correlation which indicates that there are a total of $9 \times 2 = 18$ estimating equations. Figures 2.5 and 2.6 provide fitted curves over time for groups YKL185W and YAL001C which indicate that the polynomial spline curves work well. This can also be tested using the goodness-of-fit test given by Qu and Li (2005). Figure 2.5 shows that the expression levels for group YKL185W change over time and may have cell-cycle phenomena.

To quantify our findings, a hypothesis test described in Section 2.5 is applied to test whether genes in the same group are cell-cycle regulated. Under the null hypothesis, we set the parameters associated with time to be zeros, so the intercept is constant over time. For groups YKL185W and YAL001C, the quadratic inference functions under H_1 are 0.1315 and 5.4711, respectively. The quadratic inference functions under H_0 are 17.3417 and 17.8766, respectively. The p-values of the chi-squared tests (2.11) with degrees of freedom 8 for groups YKL185W and YAL001C are 0.0280 and 0.1340, respectively.

Note that the p-value of the hypothesis test for group YAL001C is not significant, which indicates that genes among group YAL001C do not have cell-cycle phenomena. In addition, p-value for group YKL185W is statistically significant which shows quantitatively that genes in YKL185W may have cell-cycle phenomena and are worth further examination by scientists. We can run a similar analysis as above for other groups with similar gene expression patterns.

2.8. Discussion

We propose a new approach to analyze cell-cycle microarray data using data classifying analysis, varying coefficient modeling and quadratic inference functions. Our data classifying technique successfully classifies genes with similar cell-cycle patterns into the same class, or into a class with no cell-cycle phenomena at all. Furthermore, it produces pseudo-replicates for statistical modeling and inference.

In our model setting, varying coefficient models provide more flexibility than parametric models and also take both time and covariates effects into account. We also apply the quadratic inference function for estimating parameters, which enables us to incorporate the correlations but does not require the estimation of the nuisance parameters associated with the correlation. Our approach also provides statistical evidence to identify whether genes are cell cycle regulated based on asymptotic properties of the quadratic inference function.

From the simulation results, our approach performs well for large sample sizes and reasonably well for smaller or moderate sample sizes. In addition, our approach successfully identifies genes whose expression levels are time-varying. In our example, we are able to identify cell-cycle regulated genes even though we use rather small pseudo-replicates within a group. Our approach is more conservative for small samples according to the simulation results.

In this paper, we can only distinguish whether expression levels are time-varying or time-invariant. If time-invariant, then there are no cell-cycle phenomena. However, if time-varying, we can only conclude that there may be cell-cycle phenomena. Finally, our approach can be easily implemented using periodic functions instead of polynomial functions. We will investigate this further in the future.

2.9. References

[1] Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc.* B 57, 289-300.

[2] Cho, R., Campbell, M., Winzeler, E., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T., Gabrielian, A., Landsman, D., Lockhart, D., & Davis, R. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell* 2, 65-73.

[3] Hastie, T. & Tibshirani, R. (1993). Varying-coefficient models (with discussion). *J.R. Statist. Soc.* B 55, 757-796.

[4] Huang, J. Z., Wu, C. O. & Zhou, L. (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika* 89, 111-128.

[5] Hoover, D. R. , Rice, J. A. , Wu, C. O. & Yang, L.-P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* 85 , 809-822.

[6] Li, K. C., Yan, M. & Yuan, S. (2002) A simple statistical model for depicting the CDC15-synchronized yeast cell-cycle regulated gene expression data. *Statistica Sinca* 12, 141-158.

[7] Liang, K. Y. & Zeger, S. L. (1986). Longitudinal data analysis using general linear models. *Biometrika* 73, 13-22.

[8] Möller-Levet, C., Klawonn, F., Cho, K.-H. & Wolkenhauer, O. (2003). Fuzzy clustering of short time series and unevenly distributed sampling points. *IDA*, 330-340.

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

[10] Qu, A. & Birkes, D. (2004). Robust estimation when there are more equations than unknown parameters for longitudinal data. *In Progress.*

[11] Qu, A. & Li, R. (2005). Nonparametric modeling and inference function for longitudinal data. *Biometrics, to appear.*

[12] Ruppert, D. & Carroll, R. J. (1997). Penalized regression splines. *Technical Report*, Department of OR & IE, Cornell University, Ithaca, NY.

[13] Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. & Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell* 9, 3273-3297.

[14] Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. & Golub, T. (1999). Interpreting patterns of gene expression with self-organizing maps. *Proc. Natl. Acad. Sci.* 96, 2907-2912.

[15] Tavazoie, S., Hughes, J., Campbell, M., Cho, R. & Church, G. (1999). Systematic determination of genetic network architecture. *Nature Genetics* 22, 281-285.

[16] Wedderburn (1974). Quasi-likelihood functions, generalised linear models, and the Gauss-Newton method. *Biometrika* 61, 439-447.

[17] Zhao, L. P., Prentice, R. & Breeden, L. (2001). Statistical modeling of large microarray data sets to identify stimulus-response profiles. *Proc. Natl. Acad. Sci.* 98, 5631-5636.

TABLE 2.1. Distance matrix of gene $\mathbf{z_1}$

Gene	STS	Euclidean	Standardized STS	Standardized Euclidean	Overall
$\mathbf{z_1}$	S_{11}	E_{11}	S_{11}^N	E_{11}^N	D_{11}^o
$\mathbf{z_2}$	S_{21}	E_{21}	S_{21}^N	E_{21}^N	D_{21}^o
$\mathbf{z_3}$	S_{31}	E_{31}	S_{31}^N	E_{31}^N	D_{31}^o
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{z_n}$	S_{n1}	E_{n1}	S_{n1}^N	E_{n1}^N	D_{n1}^o

TABLE 2.2. Test sizes of asymptotic chi-squared tests

sample size	size
50	0.037
100	0.043
150	0.055
200	0.057
250	0.0495

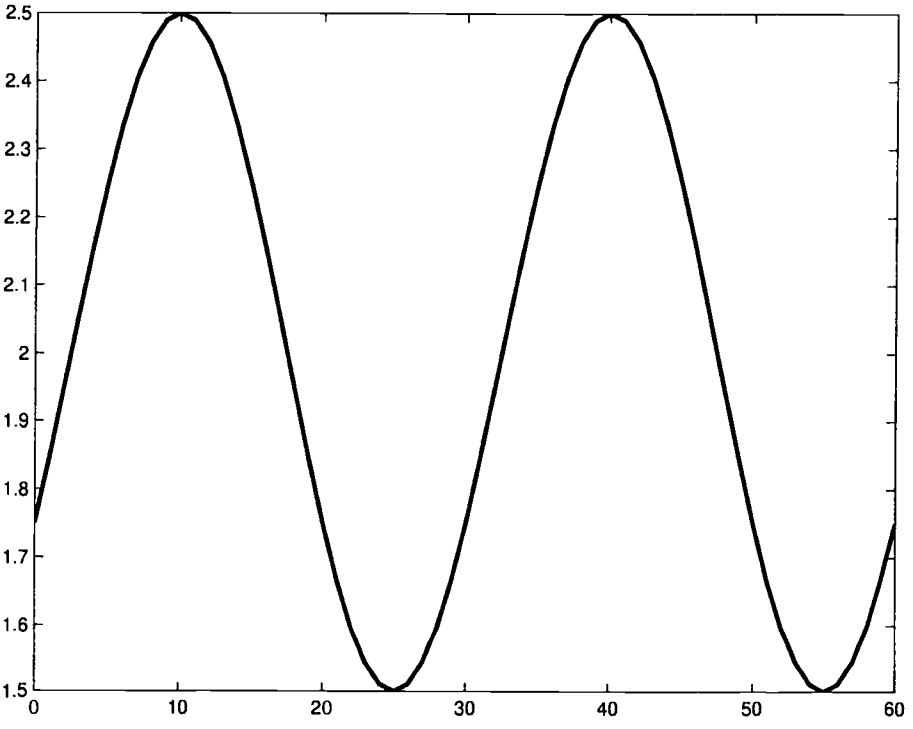

FIGURE 2.1. Cyclic pattern when $\theta = 0.5$

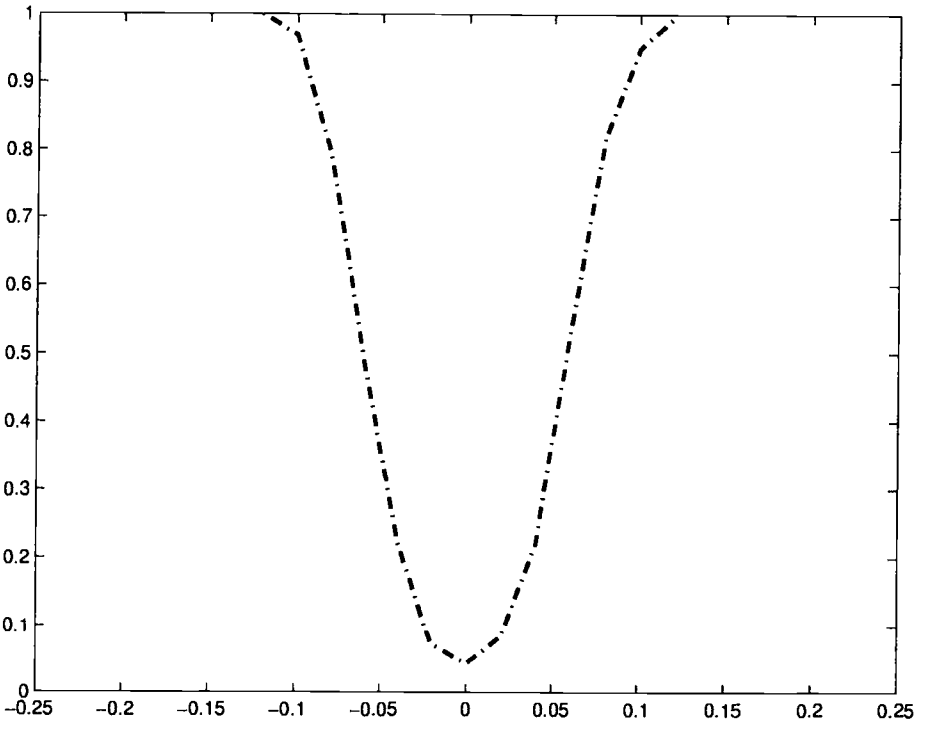

FIGURE 2.2. Power function curve based on different amplitudes θ

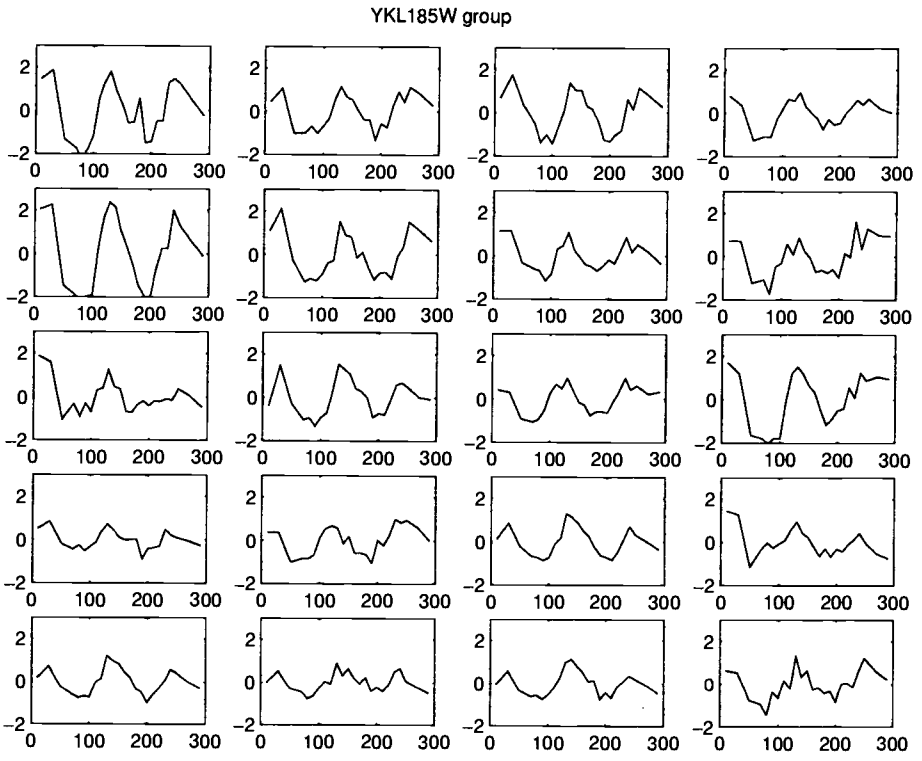

FIGURE 2.3. Grouping results for gene YKL185W: 20 genes with smaller overall distances

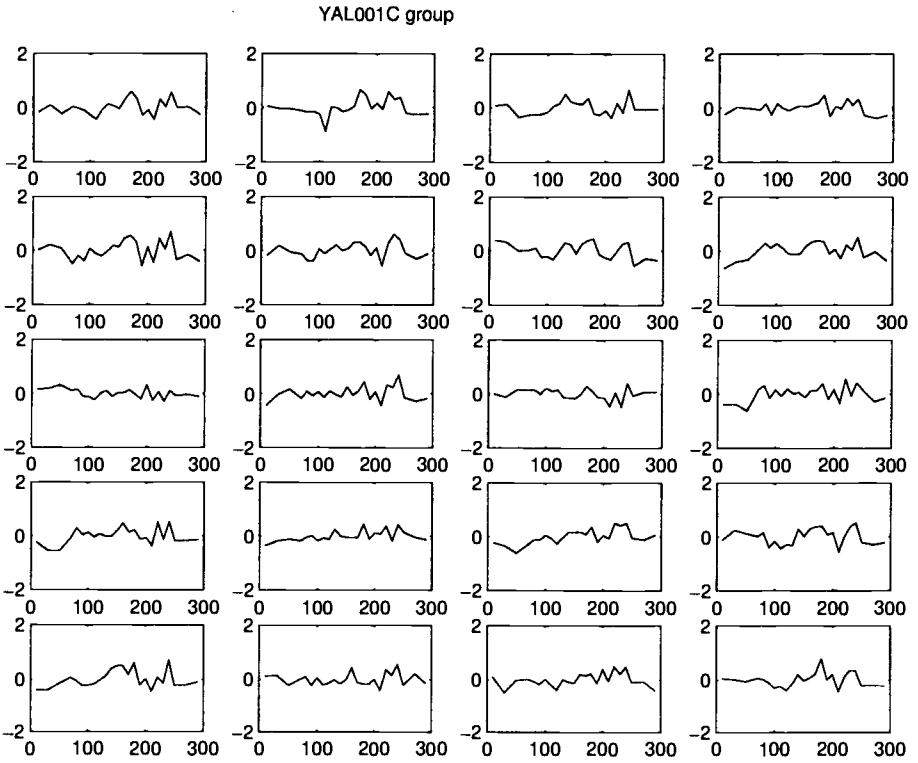

FIGURE 2.4. Grouping results for gene YAL001C: 20 genes with smaller overall distances

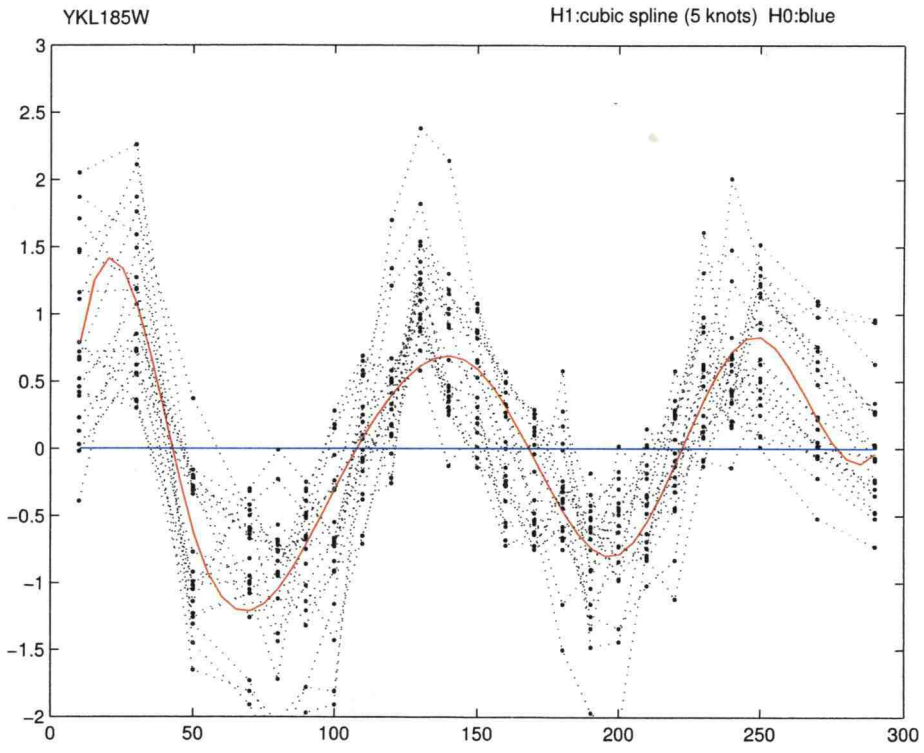

FIGURE 2.5. Fitted curves for gene YKL185W: (a) red curve indicates a 3-degree polynomial varying coefficient model fitted (b) blue straight line indicates a model fitted with only constant intercept, the parameters with time t are set to be zeros

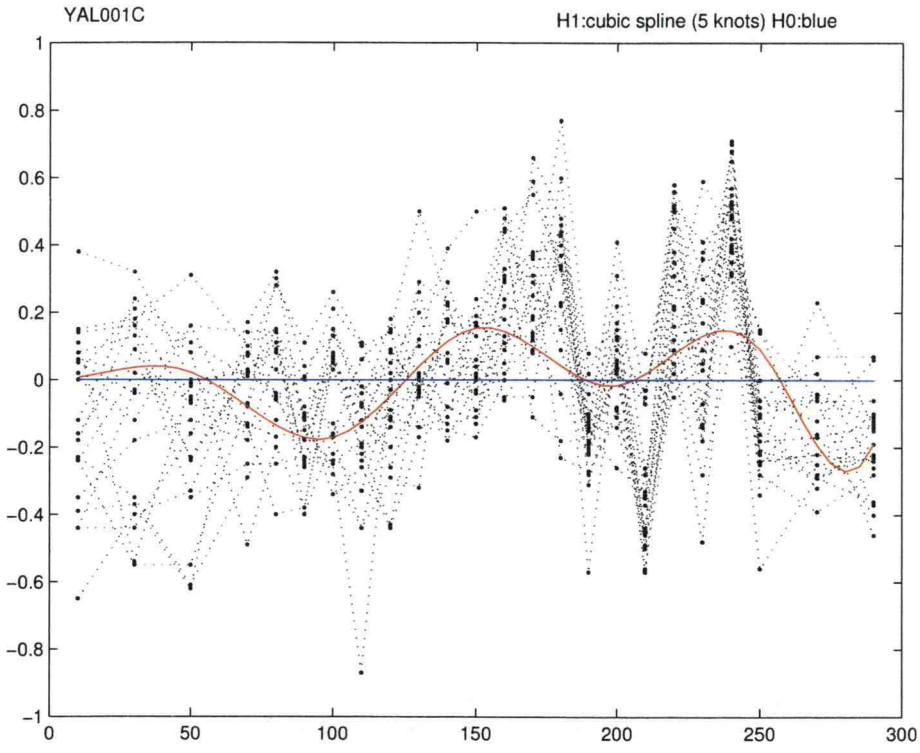

FIGURE 2.6. Fitted curves for gene YAL001C: (a) red curve indicates a 3-degree polynomial varying coefficient model fitted (b) blue straight line indicates a model fitted with only constant intercept, the parameters with time t are set to be zeros

3. CONDITIONAL EXTENDED SCORES FOR GENERALIZED LINEAR MIXED MODELS IN LONGITUDINAL DATA ANALYSIS

Guei-Feng (Cindy) Tsai

Annie Qu

Department of Statistics, Oregon State University

3.1. Abstract

In longitudinal studies, mixed effects models are useful to address subject-specific behavior. However, random effects are mostly assumed to follow a normal distribution and the variance components of random effects are challenging to estimate without likelihood functions. We propose a new approach to estimate fixed and random effects using conditional quadratic inference functions. The new approach does not require the specification of likelihood functions and also takes serial correlations of errors into account. The main advantage of this approach is that it does not require the estimation of unknown variance components associated with random effects or nuisance parameters associated with working correlations. A real data example and simulations are illustrated using our approach.

3.2. 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. There are three different sources of random variation in longitudinal data associated with random effects of subjects, serial correlations of errors and measurement errors (Diggle et al., 2002, section 5.2). Random effects explain the variation between subjects. Serial correlations of errors arise due to time course measurements. We will not consider the variation from measurement errors in this paper.

Major statistical models such as marginal models and mixed 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 GEE provides consistent estimators for regression parameters. However, the GEE estimators might be inefficient when the working correlation is misspecified. Furthermore, variation from random effects are combined additively with serial correlations in marginal models. Therefore, the same marginal model may have different interpretations, as it may lead to different random effects models if different decompositions of variations are applied.

In contrast, mixed models are able to incorporate several sources of variation and are useful when subject specific variation is of concern. However, in most of the current mixed model literature (Laird and Ware, 1982, Breslow and Clayton, 1993, Jiang, 1999), the variations are assumed to be induced by random effects only. That is, conditional on the random effects, the observations within a cluster are assumed to be independent. In this paper we will incorporate both sources of variations.

General theory and methods of linear mixed models (LMM) can be found in Searle et al. (1992), Verbeke and Molenberghs (2000), Diggle et al. (2002), and Demidenko (2004). When data follow a multivariate Gaussian distribution, estimation can be obtained by the weighted least-squares or maximum likelihood approaches since likelihood functions have explicit forms. However, in practice, non-Gaussian outcomes often arise in longitudinal studies. A generalized linear mixed model (GLMM) extends a generalized linear model and a linear mixed model for non-Gaussian data using random effects models. Extensive theory and applications of the GLMM are provided in Breslow and Clayton (1993), McCullagh (1997), Jiang (1999), McCulloch and Searle (2001), Diggle et al. (2002), Demidenko (2004).

In GLMM, likelihood functions often involve high-dimensional integrations and may be intractable. Numerical integration such as Gaussian-Hermite quadrature (Liu and Pierce, 1994) and Monte Carlo EM algorithm (McCulloch, 1997) are proposed to approximate the integration over the random effects distributions. However, these methods require full likelihood specification and could be computationally intensive and difficult to use when the number of random effects is large.

The penalized quasi-likelihood (PQL) (Breslow and Clayton, 1993) and the 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 applies the Laplace approximation using Taylor expansions. However, it requires the estimation of unknown variance components. Vonesh et al.'s CGEE2 extends the second-order GEE (Prentice and

Zhao, 1991) in GLMM. They obtain estimators of fixed and random effects by solving estimating equations associated with the conditional mean and covariance matrix. Although the CGEE2 accommodates variations from both random effects and serial correlations, it still requires estimation of the nuisance parameters associated with the working correlations and variance components associated with random effects. In addition, the normality assumption for random effects is required for both PQL and CGEE2.

In this paper, we propose conditional extended scores associated with fixed and random effects. This is an extension of the quadratic inference function method (Qu et al., 2000) for generalized linear mixed models. The proposed approach can be applied for both Gaussian and categorical data. We estimate fixed and random effects using conditional extended scores which only involve first and second conditional moments. Therefore, the specification of the likelihood is not required. Furthermore, it does not involve intractable integrations and computationally it is feasible to implement for GLMM. In addition, there are no distribution assumptions such as normality for random effects. The proposed approach enables us to incorporate the variations from serial correlations and random effects. In addition, it does not require the estimation of unknown variance components as in PQL and CGEE2 and also it does not require the nuisance parameters estimation associated with the working correlations as in CGEE2.

This paper is organized as follows. Section 3.3 describes the framework of mixed models in longitudinal data including linear or non-linear models. Section 3.4 discusses extended scores and quadratic inference functions for

marginal models as well as conditional extended scores for mixed effects models. In section 3.5, simulation results are illustrated. Section 3.6 demonstrates a real data example using our approach. A brief discussion is provided in section 3.7.

3.3. Mixed Models in Longitudinal Data

In this section we review the framework for linear and nonlinear mixed models in longitudinal data analysis.

3.3.1. Linear Mixed Model for Gaussian Longitudinal Data

Consider a classic random effects model for longitudinal data (Laird and Ware, 1982),

$$y_i = X_i\beta + U_i b_i + e_i, \quad i = 1, \ldots, N, \tag{3.1}$$

where $y_i = (y_{i1}, \ldots, y_{iT})^t$, X_i is a known $T \times p$ matrix associated with a p-dimensional vector of fixed effects β and U_i is a known $T \times q$ matrix associated with a q-dimensional vector of random effects b_i. It is assumed that each cluster has an equal number of measurements T. The b_i and e_i are independent of each other and are normally distributed as $N(0, \mathrm{D})$ and $N(0, \mathrm{W}_i)$. It is also assumed that conditioning on b_i, y_i is normally independent distributed with mean $E(y_i|b_i) = \mu_i^{b_i} = X_i\beta + U_i b_i$ and variance $\mathrm{var}(y_i|b_i) = \mathrm{W}_i$. Laird and Ware (1982) called this conditional model a 'conditional-independent model'; conditioning on b_i, the observations within a cluster are independent (i.e., W_i

is a $T \times T$ diagonal matrix). Equation (3.1) could be combined into a single equation as follows. Let

$$Y = X\beta + Zb + e, \tag{3.2}$$

where $Y^t = (y_1^t, \ldots, y_N^t)$, $X^t = (X_1^t, \ldots, X_N^t)$, $Z^t = \text{diag}(U_1^t, \ldots, U_N^t)$ and $b^t = (b_1^t, \ldots, b_N^t)$. Given b, the expected value of Y equals $X\beta + Zb$. Let θ be a vector of unknown variance components parameters corresponding to D and W. The joint log-likelihood of (β, θ, b) is

$$\log L(\beta, \theta, b) = -\frac{1}{2}\log|W| - \frac{1}{2}(Y - X\beta - Zb)^t W^{-1}(Y - X\beta - Zb) \\ -\frac{1}{2}\log|D| - \frac{1}{2}b^t D^{-1} b, \tag{3.3}$$

where $W = \text{diag}(\text{W}_1, \ldots, \text{W}_N)$ with $\text{W}_1 = \ldots = \text{W}_N = \text{W}$ and $D = \text{diag}(\text{D}_1, \ldots, \text{D}_N)$ with $\text{D}_1 = \ldots = \text{D}_N = \text{D}$. Specifically, the score equation corresponding to β is

$$\frac{\partial \log L}{\partial \beta} = X^t W^{-1}(Y - X\beta - Zb) = 0, \tag{3.4}$$

and the score equation with respect to b is

$$\frac{\partial \log L}{\partial b} = Z^t W^{-1}(Y - X\beta - Zb) - D^{-1}b = 0. \tag{3.5}$$

Combining (3.4) and (3.5) we have Henderson's (1984) mixed-model equation

$$\begin{bmatrix} X^t W^{-1} X & X^t W^{-1} Z \\ Z^t W^{-1} X & Z^t W^{-1} Z + D^{-1} \end{bmatrix} \begin{pmatrix} \beta \\ b \end{pmatrix} = \begin{bmatrix} X^t W^{-1} Y \\ Z^t W^{-1} Y \end{bmatrix}. \tag{3.6}$$

Denote the estimators of β and b to be $\hat{\beta}$ and $\hat{b}$. We can estimate variance components as follows. Let V be the marginal variance-covariance of Y, that is $V = \text{var}(Y) = Z^t D Z + W$. The marginal profile likelihood of θ is

$$\log L(\theta) = -\frac{1}{2}\log|V| - \frac{1}{2}(Y - X\hat{\beta})^t V^{-1}(Y - X\hat{\beta}). \tag{3.7}$$

Since

$$(Y - X\hat{\beta})^t V^{-1}(Y - X\hat{\beta}) = (Y - X\hat{\beta} - Z\hat{b})^t W^{-1}(Y - X\hat{\beta} - Z\hat{b}) + \hat{b}^t D^{-1}\hat{b}$$

and

$$|V| = |W||D||Z^t W^{-1} Z + D^{-1}|,$$

then (3.7) can be represented as

$$\begin{aligned} \log L(\theta) = &-\frac{1}{2}\log|W| - \frac{1}{2}(Y - X\hat{\beta} - Z\hat{b})^t W^{-1}(Y - X\hat{\beta} - Z\hat{b}) \\ &-\frac{1}{2}\log|D| - \frac{1}{2}\hat{b}^t D^{-1}\hat{b} - \frac{1}{2}\log|Z^t W^{-1} Z + D^{-1}|. \end{aligned} \tag{3.8}$$

Hence we can estimate variance components by maximizing (3.8). Note that (3.8) involves two diagonal matrices W and D but (3.7) involves V. The advantage of using (3.8) instead of (3.7) is that the inverses of W and D in (3.8) are easier to compute. The fixed effects and random effects can be obtained through (3.6) and (3.8) iteratively. We briefly summarize the procedure as follows:

1. Assume $W = I$ and $D = I$ and obtain initial estimators of β and b which satisfy (3.6).
$$\begin{bmatrix} X^t X & X^t Z \\ Z^t X & Z^t Z + I \end{bmatrix} \begin{pmatrix} \beta \\ b \end{pmatrix} = \begin{bmatrix} X^t Y \\ Z^t Y \end{bmatrix}.$$

2. First, compute estimator $\hat{\theta}$ by maximizing (3.8). Then use $\hat{\theta}$ to obtain $D(\hat{\theta})$ and $W(\hat{\theta})$ because D and W depend on θ.

3. Replace $D(\theta)$ and $W(\theta)$ with $D(\hat{\theta})$ and $W(\hat{\theta})$ in (3.6), then update $\hat{\beta}$ and $\hat{b}$.

4. Return to 2 and iterate between 2 and 3 until convergence is reached.

More details can be found in Pawitan (2001, section 17.5).

3.3.2. Mixed Models in non-Gaussian Longitudinal Data

Generalized linear mixed models (GLMM) extend the linear mixed model for non-normal longitudinal data via a specific link function. For a link function g, the conditional mean $E(Y|b) = \mu^b$ is a function of the linear predictor $X\beta + Zb$ with $g(\mu^b) = X\beta + Zb$. The choice of link function depends on the outcome of the data. For example, the identity link $g(\mu_{ij}^b) = \mu_{ij}^b$ is for normal data, the logistic link $g(\mu_{ij}^b) = \log\{\mu_{ij}^b/(1-\mu_{ij}^b)\}$ or probit link is for binary data and the log link $g(\mu_{ij}^b) = \log \mu_{ij}^b$ is for Poisson data.

The quasi-likelihood (Wedderburn, 1974) is useful if the first two moments are known but the specific distribution is unknown. In GLMM, if the conditional likelihood of Y given b is unknown, we can apply a quasi-likelihood. Suppose the normal assumption of random effects holds, then the integrated quasi-likelihood function can be used to estimate the parameters. The integrated quasi-likelihood is defined by

$$L = \frac{1}{\sqrt{(2\pi)^q |D|}} \int_{R^q} \exp\{-\frac{1}{2\phi}\sum_{i=1}^{N} d_i(y_i, \mu_i^b) - \frac{1}{2} b^t D^{-1} b\} db, \qquad (3.9)$$

where

$$d_i(y,u) = -2\int_y^u \frac{y-u}{a_i v(u)} du$$

is the weighted deviance, a_i is a known weight, ϕ is a dispersion parameter and $v(u)$ is a variance function. Again, the quasi-likelihood (3.9) does not

have a closed form if Y is not normal. Breslow and Clayton (1993) proposed the penalized quasi-likelihood (PQL) which applies a Laplace approximation to solve the integrated quasi-likelihood for GLMM.

The penalized quasi-likelihood can be written as

$$\text{PQL} = -\frac{1}{2\phi}\sum_{i=1}^{N}\sum_{j=1}^{T} d_{ij}(y_{ij}, \mu_{ij}^b) - \frac{1}{2}b^t D^{-1} b, \tag{3.10}$$

where $-\frac{1}{2}b^t D^{-1} b$ can be also treated as a penalty to the log-quasi-likelihood (McCulloch and Searle, 2001). Two quasi-score equations can be derived by taking derivatives of PQL with respect to fixed effects β and random effects b as

$$\sum_{i=1}^{N}(\frac{\partial \mu_i^b}{\partial \beta})^t (\mathrm{W}_i^b)^{-1}(y_i - \mu_i^b) = 0 \tag{3.11}$$

and

$$\sum_{i=1}^{N}(\frac{\partial \mu_i^b}{\partial b})^t (\mathrm{W}_i^b)^{-1}(y_i - \mu_i^b) - D^{-1} b = 0, \tag{3.12}$$

where $\mathrm{W}_i^b = \mathrm{var}(y_i|b)$, and the covariance matrix $D = D(\theta)$, where θ is an unknown vector of variance components.

Breslow and Clayton (1993) applied Green's Fisher scoring (1987) algorithm using a working vector Y^* to obtain solutions of (3.11) and (3.12). The working vector $Y^* = (Y_1^*, \cdots, Y_N^*)^t$ is used to transform a non-linear model to a linear model as follows,

$$y_{ij}^* = g(\mu_{ij}^b) + \dot{g}(\mu_{ij}^b)(y_{ij} - \mu_{ij}^b) = x_{ij}\beta + z_{ij}b + \dot{g}(\mu_{ij}^b)(y_{ij} - \mu_{ij}^b),$$

where $\dot{g}(\mu_{ij}^b)$ is the first derivatives of g with respect to μ_i^b. Clearly, $E(y_{ij}^*) = E\{g(\mu_{ij}^b)\} = x_{ij}\beta$, since $\dot{g}(\mu_{ij}^b)(y_{ij} - \mu_{ij}^b)$ and b have mean zero. Thus, if we

treat $\dot{g}(\mu_{ij}^b)(y_{ij} - \mu_{ij}^b)$ as an error term of the working vector Y^*, then Y^* follows a linear model

$$Y^* = X\beta + Zb + \varepsilon,$$

where $b \sim N(0, D)$ and $\varepsilon \sim N(0, W^*)$, $W^* = \text{diag}(W_1^*, \ldots, W_N^*)$ and $W_i^* = \text{diag}(w_{i1}^*, \ldots, w_{iT}^*)$ with $w_{ij}^* = \{\dot{g}(\mu_{ij}^b)\}^2 \text{var}(y_i|b_i)$. Note that W_i^* does not depend on θ.

Following Henderson's mixed-model equation of (3.6), we have

$$\begin{bmatrix} X^t W^{*-1} X & X^t W^{*-1} Z \\ Z^t W^{*-1} X & Z^t W^{*-1} Z + D^{-1} \end{bmatrix} \begin{pmatrix} \beta \\ b \end{pmatrix} = \begin{bmatrix} X^t W^{*-1} Y^* \\ Z^t W^{*-1} Y^* \end{bmatrix}. \tag{3.13}$$

The only difference between (3.13) and (3.6) is that (3.13) uses a working vector Y^* instead of a true data vector Y and W^* instead of W.

Similar to (3.6), the covariance matrix $D(\theta)$ is involved in (3.13). Breslow and Clayton (1993) proposed a modified REML estimator for variance components θ using the working vector Y^* and estimators of β and b. This procedure is similar to the one described in Section 3.3. The only difference is that in the GLMM, the working vector Y^* should be updated at the end of each iteration. More details can be found in Breslow and Clayton (1993).

3.4. Quadratic inference functions for GLMM

Although the penalized quasi-likelihood is useful for GLMM, it still requires the estimation of variance components and the normality assumption of random effects. In this paper, we extend the quadratic inference function

(QIF) (Qu et al., 2000) for GLMM, which does not require variance components estimation and normality assumption for random effects.

In this section, we will first give a brief description of the quadratic inference function for marginal models. Then we demonstrate the further extension of the QIF for GLMM in longitudinal data when subject specific effects are of interest.

3.4.1. Quadratic inference functions for marginal models

In a marginal approach, Liang and Zeger (1986) proposed the generalized estimating equation (GEE) which extends the quasi-score equation

$$\sum_{i=1}^{N} \dot{\mu_i}^t V_i^{-1}(y_i - \mu_i) = 0$$

by assuming $V_i = A_i^{1/2} R A_i^{1/2}$, where A_i is a diagonal marginal variance matrix and R is a working correlation which involves nuisance parameters. Here $\mu_i = E(y_i)$, $V_i = \text{var}(y_i)$ and $\dot{\mu_i}$ is the first derivative of μ_i with respect to β. The GEE estimators are consistent and asymptotically normal even if the working correlation matrix is misspecified. However, the estimator of regression parameters is not efficient under the misspecification of the working correlation.

Qu et al. (2000) proposed the quadratic inference functions (QIF) for longitudinal data. Their approach is based on the method of moments which only requires the first two moments of the distribution. In addition, it takes correlation into account without the estimation of nuisance parameters associated with working correlations. The main idea of their approach is to

assume that the inverse of the working correlation, R^{-1}, is approximated by a class of linear combination of known matrices $M_1, \ldots, M_m$, that is $R^{-1} \approx \sum_{j=1}^{m} a_j M_j$, where M_1 is usually an identity matrix. Suppose the marginal mean can be modeled as $\mu_i = E(y_i) = g(X_i\beta)$, where β is a vector of fixed effects. Then the GEE can be approximated by

$$\sum_{i=1}^{N} \dot{\mu_i}^t A_i^{-1/2} \left(\sum_{j}^{m} a_j M_j \right) A_i^{-1/2} (y_i - \mu_i) = 0. \tag{3.14}$$

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

$$\bar{g}_N(\beta) = \frac{1}{N} \sum_{i=1}^{N} g_i(\beta) = \frac{1}{N} \begin{pmatrix} \sum_{i=1}^{N} \dot{\mu_i}^t A_i^{-1/2} M_1 A_i^{-1/2} (y_i - \mu_i) \\ \vdots \\ \sum_{i=1}^{N} \dot{\mu_i}^t A_i^{-1/2} M_m A_i^{-1/2} (y_i - \mu_i) \end{pmatrix}, \tag{3.15}$$

where $\bar{g}_N$ is a mp-dimensional vector. Note that the left-hand side of the GEE is a linear combination of extended scores $\bar{g}_N$.

The extended scores in (3.15) contain more estimating equations than unknown parameters if $m > 1$. Qu et al. (2000) adapted the idea of the generalized method of moments (Hansen, 1982) and proposed the quadratic inference function (QIF) to estimate the parameters β defined in (3.15). The QIF is defined as

$$Q_N(\beta) = N \bar{g}_N^t \bar{C}_N^{-1} \bar{g}_N, \tag{3.16}$$

where Q_N is a $mp \times mp$ matrix and $\bar{C}_N = \frac{1}{N} \sum_{i=1}^{N} g_i(\beta) g_i(\beta)^t$. Here N must be greater than mp to ensure the invertibility of the variance matrix $\bar{C}_N$ (Qu and Birkes, 2004).

The quadratic inference function also provides an asymptotic chi-squared test (Qu et al., 2000). Suppose that the parameter vector β can be

partitioned into (δ, γ), where δ is the p_1-dimensional interest parameter vector and γ is the p_2-dimensional nuisance parameter vector where $p_1 + p_2 = p$. To test a null hypothesis $H_0 : \delta = \delta_0$ against an alternative hypothesis $H_1 : \delta \neq \delta_0$, we obtain a test statistic $T^* = Q(\delta_0, \tilde{\gamma}) - Q(\hat{\delta}, \hat{\gamma})$, where $\tilde{\gamma}$ and $(\hat{\delta}, \hat{\gamma})$ are estimators by minimizing (3.16) under the null and alternative hypothesis. In addition, the test statistic T^* asymptotically follows a chi-squared distribution with degrees of freedom p_1.

3.4.2. Quadratic inference functions for GLMM

Consider a GLMM as in section 3.3.2 with conditional mean μ_i^b. The conditional quasi-likelihood of Y is

$$l_q = -\frac{1}{2\phi} \sum_{i=1}^{N} d_i(y_i, \mu_i^b), \tag{3.17}$$

where $\mu_i^b = E(y_i|b)$. Note that (3.17) is the first term of PQL as in (3.10) regardless of the distribution of random effects. A constraint $P_A b = 0$ associated with the random effects is required to ensure the identifiability of fixed effects β and random effects b, where P_A is a known orthogonal matrix (Jiang, 1999). Thus, the estimators of fixed effects and random effects are obtained by maximizing (3.17) subject to $P_A b = 0$. In order to do that, we can build a Lagrange function

$$l_q = -\frac{1}{2\phi} \sum_{i=1}^{N} d_i(y_i, \mu_i^b) - \frac{1}{2}\lambda |P_A b|^2, \tag{3.18}$$

where λ is a Lagrange multiplier.

The details about the specification of P_A can be found in Jiang (1999). Define $\mathcal{B}(M) = \{B : B$ is a matrix whose columns constitute a base for M$\}$

and $\mathcal{N}(M)$ is the null space of M. The matrix of $P_A = A(A^tA)^{-1}A'$ is a projection matrix where $A \in \mathcal{B}(\mathcal{N}((I-P_X)Z))$. For example, assume a GLMM model with random intercept only as $E(\mu_{ij}^b) = \alpha_0 + \alpha_1 x_{ij} + b_i$. For this random intercept model, it has been shown that $P_A = (\frac{1}{N}, \ldots, \frac{1}{N})^t$ and the constraint associated with the random effects is $\frac{1}{N}\sum b_i = 0$ (Jiang, 1999).

The quasi-score equations corresponding to β and b_i can be derived from (3.17) as

$$\sum_{i=1}^{N} (\frac{\partial \mu_i^b}{\partial \beta})^t (\mathrm{W}_i^b)^{-1} (y_i - \mu_i^b) = 0. \tag{3.19}$$

and

$$\begin{pmatrix} h_1 = (\frac{\partial \mu_1^{b_1}}{\partial b_1})^t (\mathrm{W}_1^b)^{-1} (y_1 - \mu_1^{b_1}) - \lambda \frac{\partial P_A b}{\partial b_1} = 0 \\ \vdots \\ h_N = (\frac{\partial \mu_i^{b_N}}{\partial b_N})^t (\mathrm{W}_N^b)^{-1} (y_N - \mu_N^{b_N}) - \lambda \frac{\partial P_A b}{\partial b_N} = 0 \end{pmatrix}. \tag{3.20}$$

In section 2, W_i^b is assumed to be a diagonal matrix. However, here we assume that given b_i, measurements within i cluster may not be independent. Thus $\mathrm{W}_i^b = \mathrm{var}(y_i|b_i)$ is not necessarily a diagonal matrix.

Suppose a working correlation R for $W_i = A_i^{\frac{1}{2}} R^{-1} A_i^{\frac{1}{2}}$, where $R^{-1} \approx \sum_{j=1}^{m} a_j M_j$ and $A_i = \mathrm{diag}\{\mathrm{var}(y_{i1}|b), \ldots, \mathrm{var}(y_{iT}|b)\}$ as shown in section 3.4.1. Based on (3.19), we define the conditional extended scores associated with the fixed effects, namely fixed-effects extended scores, as

$$\bar{g}_N^f = \frac{1}{N}\sum_{i=1}^{N} g_i^f(\beta) = \frac{1}{N} \begin{pmatrix} \sum_{\mathrm{i}=1}^{\mathrm{N}} (\frac{\partial \mu_\mathrm{i}^\mathrm{b}}{\partial \beta})^\mathrm{t} \mathrm{A}_\mathrm{i}^{-1/2} \mathrm{M}_1 \mathrm{A}_\mathrm{i}^{-1/2} \left(\mathrm{y}_\mathrm{i} - \mu_\mathrm{i}^\mathrm{b}\right) \\ \vdots \\ \sum_{\mathrm{i}=1}^{\mathrm{N}} (\frac{\partial \mu_\mathrm{i}^\mathrm{b}}{\partial \beta})^\mathrm{t} \mathrm{A}_\mathrm{i}^{-1/2} \mathrm{M}_\mathrm{m} \mathrm{A}_\mathrm{i}^{-1/2} \left(\mathrm{y}_\mathrm{i} - \mu_\mathrm{i}^\mathrm{b}\right). \end{pmatrix} \tag{3.21}$$

In addition, the quasi-score equations h_i in (3.20) can be represented as a linear combination of

$$g_i^r = \begin{pmatrix} (\frac{\partial \mu_i^{b_i}}{\partial b_i})^t A_i^{-1/2} M_1 A_i^{-1/2} (y_i - \mu_i^b) \\ \vdots \\ (\frac{\partial \mu_i^{b_i}}{\partial b_i})^t A_i^{-1/2} M_m A_i^{-1/2} (y_i - \mu_i^b) \\ \frac{\partial P_A b}{\partial b_i} \end{pmatrix}. \tag{3.22}$$

Therefore, we define the random-effects extended scores for b as

$$\begin{pmatrix} g_1^r \\ \vdots \\ g_N^r \end{pmatrix}. \tag{3.23}$$

Note that extended scores (3.21) and (3.23) are able to incorporate the within-subject correlation without involving the nuisance parameters associated with correlation and the penalized parameter λ. In addition, specification of the distribution for random effects is also not required.

The estimators of fixed effects and random effects can be obtained by solving (3.21) and (3.23). However, both (3.21) and (3.23) are overidentified, so we adapt the QIF idea to obtain the fix effects and random effects estimators. That is, for given random effects b_i, we estimate the fixed effects by minimizing

$$N(\bar{g}_N^f)^t (\bar{C}_N^f)^{-1} (\bar{g}_N^f), \tag{3.24}$$

where $\bar{C}_N^f = \frac{1}{N} \sum_{i=1}^N (g_i^f)(g_i^f)^t$. In addition, for given fixed effects β, we estimate the random effects by minimizing

$$\sum_{i=1}^N (g_i^r)^t (g_i^r). \tag{3.25}$$

Note that in (3.25) the identity matrix is used as the weighting matrix for g_i^r, which may lead to inefficient estimators of random effects. However, the efficiency of random effects estimation is not of concern.

The iterative estimating procedure is summarized as follows.

1. Start with an initial vector $\hat{\beta}$.

2. Replace β in (3.23) with $\hat{\beta}$ and obtain the estimators of random effects by minimizing (3.25).

3. Replace b with $\hat{b}$ in (3.21). Then update $\hat{\beta}$ by minimizing (3.24) and return to S2.

Iterate between 2 and 3 until convergence is reached.

3.5. Simulation results of fixed effect estimates

To evaluate the performance of the QIF method for GLMMs, we conduct two simulation studies. In the first simulation study, 500 conditional independent binary datasets are generated. For each dataset, 50 clusters are generated and each cluster has 13 repeated measurements at time points $t = -3, -2.5, \ldots, 3$. The conditional independent binary outcomes y_{ij} $(i = 1, \ldots, 50 \;\; j = 1, , \ldots, 13)$ are generated from

$$\text{logit}(\mu_{ij}^b) = \alpha_0 + \alpha_1 t_j + \alpha_2 x_i + \alpha_3 x_i t_j + b_i,$$

where $x_i = 1$ for $i = 1, \ldots, 25$ and $x_i = 0$ for $i = 26, \ldots, 50$. The regression coefficients are fixed at $\alpha_0 = -2.5$, $\alpha_1 = 1$, $\alpha_2 = -1$ and $\alpha_3 = -0.5$. The

random intercepts b_i are generated from a normal distribution with mean 0 and variance 1.

We apply both mixed QIF and marginal QIF methods along with PQL and GEE to these datasets. The AR-1 working correlation is assumed for (3.21) and (3.23), marginal QIF and GEE. The inverse of AR-1 working structure can be represented as a linear combination of three known matrices, M_0, M_1 and M_2. Here M_0 is an identity matrix, M_1 is a matrix with 1 on the two main off-diagonals and 0 elsewhere, and M_2 is 1 on $(1,1)$ and (T,T) where T is the number of repeated measurements for each subjects. Here, we use a simplified the AR-1 correlation structure with only M_0 and M_1. In addition, a constraint $\frac{1}{50}\sum b_i = 0$ is assumed for estimating random effects.

Table 3.1 provides the mean values, standard errors and bias of the fixed effects estimators based on 500 simulations using mixed QIF, PQL, marginal QIF and GEE, denoted as $\text{QIF}_{\text{mixed}}$, PQL, QIF, and GEE, respectively. First, the difference between the mixed models and marginal models for the α_0 estimator is much larger than α_1, α_2 and α_3. The mixed models perform better for α_0. This is because the mixed models take the random intercept into account. Noticeably, the mixed and marginal models yield similar estimators of α_1, α_2 and α_3. As expected the PQL performs best. This is because conditional independent is assumed for PQL. Although in our approach the correlation structure is misspecified, it still yields similar results.

In the second simulation study, we generate data with both random effects and serial correlation. There are 500 Poisson correlated datasets gen-

erated from the following model

$$y_{ij}|b_i = w_{ij} \quad i = 1, \ldots, 50 \quad j = 1, \ldots, 8,$$

where w_{ij} are conditionally (on b_i) correlated Poisson random variables with mean $\mu_{ij}^b = \exp(\alpha_0 + b_i + \alpha_1 t + \alpha_2 G_i)$ and covariance $\text{cov}(y_{ij}, y_{ij'}) = 0.7$ for $j \neq j'$. The covariates are $t = (.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0)$ for all subjects i, and G_i are treatment effects where $G_i = 0$ for $i \leq 25$ and $G_i = 1$ for $i > 25$. The regression coefficients are fixed at $\alpha_0 = 1.5$, $\alpha_1 = -1$, $\alpha_2 = 1$. The strategy of generating data is as follows. First, we generate b_i independently from the normal distribution with mean zero and variance one. After we calculate the conditional mean of y_{ij} for each subject i, we generate the Poisson correlated data as in Park and Shin (1998), and Madsen and Dalthorp (2005).

We also apply the mixed QIF and PQL along with the marginal QIF and GEE to the second simulated data sets. The log link function is assumed. A constraint $\frac{1}{50}\sum b_i = 0$ is also assumed for estimating random effects. Results of this simulation are shown in Table 3.2 including point estimation, standard errors and bias. The difference between the mixed models and marginal models for estimators of α_0 is larger than α_1 and α_2. Mixed models perform better than marginal models for α_0. The estimator of α_0 obtained from the GEE has the largest bias. In addition, the estimator of α_0 obtained from the mixed QIF has less bias than the other methods. That is because the serial correlations are considered in the mixed QIF. In addition, the estimators of the regression parameters α_1 and α_2 are very similar for the four methods.

3.6. Applications

In this section, we apply Thall and Vail's (1990, page 664) epileptic seizure dataset to illustrate the quadratic inference function approach for random effects models. The epileptic seizure data consist of 59 epilepsy patients, 31 of whom received an new anti-epileptic drug and 28 of whom received a placebo. For each patient, baseline data were 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 j visit ($i = 1, \ldots, 59$, $j = 1, \ldots, 4$). The log link function is used for the random effects model, that is

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

where $\mu_{ij}^b = E(y_{ij}|b_i)$. The covariates in the model include $x_{i1} = \log(\text{base}_i/4)$ which is the logarithm of $\frac{1}{4}$ of the number of baseline seizures, $x_{i2} = \log(\text{age}_i)$ which is the logarithm of the ith patient's age, T_i which is a treatment indicator variable defined to be 1 for the new drug and 0 for the placebo, and visit_{ij} which is a time-dependent covariate for four visits where $\text{visit}_{i1} = -3$, $\text{visit}_{i2} = -1$, $\text{visit}_{i3} = 1$ and $\text{visit}_{i4} = 3$. In addition, for the count data, the conditional variance is the same as the conditional mean, that is $\mu_{ij}^b = \text{var}(y_{ij}|b_i)$.

We compare the mixed model approaches (mixed QIF, and PQL) to two marginal model approaches (the GEE and marginal QIF). For estimating fixed effects, the mixed QIF estimators are obtained by minimizing (3.24) and

(3.25). The PQL fits the data with no serial correlations assumed. The PQL and GEE estimators are obtained from the available Splus procedures "glme" and "gee". These procedures can be found in the Splus library. To obtain GEE estimator, we use an AR1 working correlation structure. The marginal QIF estimator is obtained by minimizing (3.16) with a simplified AR-1 working structure.

Table 3.3 provides the fixed-effects estimators using mixed QIF, PQL, marginal QIF and GEE. Noticeably, there is a discrepancy of intercept estimators between the mixed models and the marginal models. The intercept estimators obtained from the mixed QIF and PQL are less than half of the estimators obtained from marginal QIF and GEE. In addition, the treatment-effect, age-effect and visit-effect estimators obtained from the mixed QIF and PQL are also quite different from those obtained from the marginal QIF and GEE. The mixed models and marginal models yield similar estimators of baseline effect. In addition, all four approaches yield strong evidence that the effect of the baseline is significant ($p < 0.01$). In addition, the marginal models also yield strong evidence that the effect of intercept is significant. However, after accounting for the random intercept, there is no evidence that the effect of intercept is significant using both PQL and mixed QIF.

3.7. Discussion

Mixed effects models are useful for longitudinal data when subject specific variation is of interest. Many approaches have been proposed for estimating fixed effects and random effects for generalized linear mixed effects models.

These include Gaussian-Hermite quadrature based on exact likelihoods, penalized quasi-likelihood and the conditional second-order generalized estimating equations based on the penalized quasi-likelihood. However, the normality assumption is required for these approaches.

We propose a new approach for generalized linear mixed models. The new approach estimates the fixed and random effects using two conditional extended scores which only involve the first two moments. Thus, the specification of the likelihood function is not required. In addition, it does not require the distribution assumption for random effects and is able to incorporate both serial correlation and random effects. Moreover, it does not involve unknown variance components as in PQL, or the nuisance parameters associated with working correlations as in CGEE2.

Two simulation results show that the PQL and mixed QIF yield lower bias of estimators than marginal approaches when data are from random effects models. In addition, when the serial correlation is introduced into the data, the mixed QIF performs better than the others.

Although our approach does not require the estimation of variance components for estimating fixed and random effects, it may perform slowly when the dimension of the random effects is high. We will continue to investigate the performance of our approach when the dimension of the random effects is high.

3.8. References

[1] Breslow, N. E. & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistics Association*, 88, 9-25.

[2] Demidenko, E. (2004). *Mixed Models: Theory and Applications.* John Wiley, New York.

[3] Diggle P. J., Heagerty, P., Liang, K.-Y. & Zeger S. L. (2002). *Analysis of longitudinal data.* Oxford: Oxford University Press.

[4] Jiang, J. (1999). Conditional inference about generalized linear mixed models. *The Annals of Statistics*, 27, 19742007.

[5] Laird, N. M. & Ware, J. H. (1982). Random-effects model for longitudinal data. *Biometrics*, 38, 963-974.

[6] Liang, K. Y. & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.

[7] Lin, X. & Breslow, N. L. (1996). Biased correction in generalized linear mixed effects models with multiple components of dispersion. *Journal of the American Statistics Association*, 91, 1007-1016.

[8] Liu, Q. & Pierce, D. A. (1994). A note on Gauss-Hermite quadrature. *Biometrika*, 81, 624-629.

[9] Madsen, L. & Dalthorp, D. 2005. Simulating correlated count data. *Environmental and Ecological Statistics*, In press.

[10] McCullogh, C. E. (1997). Maximum likelihood algorithm for generalized linear mixed models. *Journal of the American Statistics Association*, 92, 162-170.

[11] McCulloch, C. E. & Searle, S. R. (2001). *Generalized, Linear and Mixed Models.* John Wiley, New York.

[12] Newey, W.K. & McFadden, D. (1994). Large Sample Estimation and Hypothesis Testing. *In Handbook of Econometrics*, Vol. 4, ed. by R.F. Engle and D.L. McFadden, 2111-2245. Elsevier, North Holland.

[13] Park, C.G. & Shin, D.W. (1998). An algorithm for generating correlated random variables in a class of infinitely divisible distributions. *Journal of Statistical Computation and Simulation*, 61, 127-139.

[14] Pawitan, Y. (2001). *In all likelihood: Statistical Modelling and Inference using likelihood.* Oxford University Press.

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

[16] Qu, A. & Birkes, D. (2004). Robust estimation when there are more equations than unknown parameters for longitudinal data. *Manuscript.*

[17] Searle, S. R., Casella, G. & McCulloch, C.E. (1992). *Variance Components.* John Wiley, New York.

[18] Thall, P. F. & Vail, S. C. (1990). Some covariance models for longitudinal count data with over-dispersion. *Biometrics*, 46, 657671.

[19] Verbeke, G. & Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data.* Springer-Verlag, New York.

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

[21] Zeger, S. L. & Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42, 121-130.

[22] Zeger, S. L., Liang, K. Y. & Alberts, P.S. (1988). Models for longitudinal data: A generalized estimating equations approach. *Biometrics*, 44, 1049-1060.

TABLE 3.1. Mean Values, Standard Errors and Bias of Fixed Effects Estimators for Conditional Independent Binary Data

Parameter	QIF_{mixed}	PQL	QIF	GEE
α_0	-2.35	−2.47	-2.22	−2.21
bias	.15	.03	.27	.29
s.e.	.36	.21	.31	.32
α_1	.97	.97	.90	.90
bias	-.03	-.03	-.10	-.10
s.e.	.16	.10	.12	.13
α_2	-.91	-.98	.95	.96
bias	.09	.02	.05	.04
s.e.	.34	.32	.51	.50
α_3	-.44	-.47	-.41	-.40
bias	.06	.03	.09	.10
s.e.	.16	.14	.22	.20

TABLE 3.2. Mean Values and Standard Errors of Fixed Effects Estimators for Conditional Correlated Count Data

Parameter	QIF_{mixed}	PQL	QIF	GEE
α_0	1.46	1.33	1.70	1.83
bias	-.04	-.17	.20	.33
s.e.	.43	.24	.28	.29
α_1	-1.01	-1.00	-1.01	-1.06
bias	-.01	0	-.01	-.06
s.e.	.08	.06	.07	.15
α_2	1.03	1.10	1.08	1.00
bias	-.03	-.11	-.15	-.17
s.e.	.47	.39	.35	.33

TABLE 3.3. Comparision of mixed QIF, PQL, marginal QIF and GEE Methods for Thall and Vail's (1990) epilepsy data analysis

Effects	QIF_{mixed}	PQL	QIF	GEE
Intercept	-1.11	-1.23	-2.35	-2.73
s.e.	1.32	1.23	.89	1.04
t-value	-0.84	-1	**-2.64**	**-2.63**
Base	.90	1.02	1.18	1.25
s.e.	.14	.10	.10	.16
t-value	**6.43**	**10.20**	**11.80**	**7.81**
Trt	-.31	-.27	-.09	-.02
s.e.	.23	.15	.14	.19
t-value	-1.35	-1.08	-0.64	-0.11
Age	.37	.37	0.59	0.66
s.e.	.43	.35	.24	.29
t-value	0.86	1.06	2.46	2.28
Visit/10	-.21	-.29	-0.27	-0.32
s.e.	.28	.14	.13	.17
t-value	-0.75	-2.07	-2.08	-1.88

4. VARIANCE COMPONENTS ESTIMATION AND BIAS CORRECTION FOR LINEAR MIXED MODELS IN LONGITUDINAL DATA

Guei-Feng (Cindy) Tsai

Annie Qu

David Birkes

Department of Statistics, Oregon State University

4.1. Abstract

In this paper we develop a new approach for the estimation of variance components using the second-order extended scores. This is an extension of the quadratic inference function for variance components in linear mixed models. The new approach does not require the specification of a likelihood function and incorporates both random effects and serial correlations. In addition, we propose an asymptotic chi-squared test to test the variance components of interest. This chi-squared test can also be used for testing whether the serial correlation is significant. Simulations and a real data example are provided as illustration.

4.2. Introduction

In longitudinal studies, subjects are measured repeatedly over time. Therefore, measurements within a subject are likely to be correlated. Laird and Ware (1982) proposed random effects models for longitudinal data. In

their model, conditional on random effects, the observations within a subject are assumed to be independent. However, in practice, serial correlations may often arise within subjects (Verbeke et al., 1998; Diggle et al., 2002).

In mixed effects models, the estimation of variance components is one of our interests. The maximum likelihood estimation (MLE) (Hartley and Rao, 1967) and residual maximum likelihood estimation (REMLE) (Patterson and Thompson, 1971) are widely used for estimating variance components when likelihood functions are known. The main difference between MLE and REMLE is that REML estimators are obtained by transforming a linear likelihood function which takes into account the loss of degrees of freedom for estimating fixed effects. Although the MLE and REMLE are useful for estimating variance components in linear mixed effects models, they require the specification of a likelihood function.

The generalized method of moments (GMM) (Hansen, 1982) uses the moment conditions to estimate unknown parameters, which does not require the specification of a likelihood function. The GMM estimator is consistent and asymptotic normal (Hansen, 1982). Qu et al. (2000) proposed the quadratic inference function (QIF) which extends the GMM for estimating regression parameters in the generalized estimating equations (GEE) (Liang and Zeger, 1986) setting. The QIF estimation is a marginal approach which requires only the first two moments. It is also able to incorporate the within-subject correlation without involving the nuisance parameters associated with the working correlation.

In this paper we propose using the second-order extended scores for estimating variance components. This is an extension of the quadratic inference function for variance components in linear mixed models. The following is our strategy. First, we estimate the fixed effects using the first-order quadratic inference function which only involves the first and second moments. Therefore, the likelihood specification is not required. Then we estimate the variance components using the second-order quadratic inference function which involves the third and fourth moments, and the regression estimators obtained from the first-order quadratic inference function. The advantage of the new approach is that the first-order quadratic inference function does not involve unknown variance components. Following Qu et al. (2000), we propose an asymptotical chi-squared test based on the second-order quadratic inference function for testing whether variance components of interest are significant.

In addition, we incorporate serial correlations into our models. We first approximate the covariance of errors by a linear representation of the basis matrix using Taylor expansion. Then we estimate the variance components associated with random effects and serial correlations by minimizing the second-order quadratic inference function. To test serial correlation, Chi et al. (1989) proposed a score test for testing whether the AR-1 autocorrelation is significant. However, in their approach, the specification of a likelihood function is required. The asymptotic chi-squared test based on the second-order quadratic inference function can also be applied for testing whether the serial correlation is significant.

This paper is organized as follows. Section 4.3 proposes the second-order quadratic inference functions for estimating variance components in linear mixed effects models. In section 4.4, asymptotic properties of the second-order QIF estimators are provided. Section 4.5 discusses a bias correction procedure for the second-order quadratic inference function. Section 4.6 proposes an asymptotic chi-squared test for testing variance components of interest. Simulation results are illustrated in section 4.7. Section 4.8 demonstrates a real data example using our approach. A brief discussion is provided in section 4.9.

4.3. Variance Components Estimation

In this section, we introduce a second-order quadratic inference function for variance components estimation. This approach does not require the specification of a likelihood function.

Let $y_i = (y_{i1}, \ldots, y_{iT})'$, $i = 1, \ldots, N$, be the sequence of repeated measurements on subject i. For simplicity, it is assumed that each cluster has an equal number of measurements T. Consider a linear mixed effects model

$$y_i = X_i\beta + U_{i1}b_{i1} + \ldots + U_{ir}b_{ir} + e_i \tag{4.1}$$

where β is a p-dimensional vector of fixed effects, b_{il} is a q_l-dimensional vector of random effects for $l = 1, \ldots, r$, X_i and U_{il} are $T \times p$ and $T \times q_l$ -dimensional known covariate matrices, and e_i is a T-dimensional vector of residual components. Assume that b_{il} and e_i follow normal distributions $N(0, \sigma_l^2 I)$ and $N(0, \sigma_e^2 I)$. It is also assumed that b_{i1}, ..., b_{ir} and e_1, ..., e_N are also inde-

pendent of each other. In this section, we assume that covariance of e_i is a diagonal covariance matrix with each diagonal element σ_e^2. We will discuss a model with non-diagonal covariance of e_i in section 4.6.

The marginal model of y_i follows a normal distribution with mean $X_i\beta$ and covariance

$$V_i = \sigma_1^2 U_{i1} U_{i1}' + \ldots + \sigma_r^2 U_{ir} U_{ir}' + \sigma_e^2 I. \tag{4.2}$$

Let $\psi = (\sigma_1^2, \ldots, \sigma_r^2, \sigma_e^2)$ be a vector of unknown variance component parameters.

Liang and Zeger (1986) proposed the generalized estimating equation (GEE) to incorporate the correlation within a subject in longitudinal analysis. The GEE provides consistent estimators for regression parameters even if the working correlation is misspecified. However, misspecification of the working correlation may lead to inefficient estimators of regression parameters. Qu et al. (2000) proposed the quadratic inference function (QIF) for estimating the fixed effects vector β. Their approach requires only the first two moments of the distribution. The basic idea is to suppose the inverse of the working correlation can be approximately represented as a linear combination of a class of m known basis matrices $M_1, \ldots, M_m$ as $R^{-1} \approx \sum_{l=1}^{m} a_l M_l$. Then the generalized estimating equation can be represented as

$$\sum_{i=1}^{N} \dot{\mu}_i' A_i^{-1/2} \left(\sum_{j}^{m} a_j M_j \right) A_i^{-1/2} (y_i - \mu_i) = 0, \tag{4.3}$$

where $\mu_i = E(y_i)$, $\dot{\mu}_i = \frac{\partial}{\partial \beta} \mu_i$ and A_i is a diagonal marginal variance matrix. Note that the left-hand side of the GEE is a linear combination of the extended

scores in the mp-dimensional vector

$$\bar{g}_N(\beta) = \frac{1}{N}\sum_{i=1}^{N} g_i(\beta) = \frac{1}{N}\begin{pmatrix} \sum_{i=1}^{N} \dot{\mu}_i' A_i^{-1/2} M_1 A_i^{-1/2} (y_i - \mu_i) \\ \vdots \\ \sum_{i=1}^{N} \dot{\mu}_i' A_i^{-1/2} M_m A_i^{-1/2} (y_i - \mu_i) \end{pmatrix}. \quad (4.4)$$

Because the dimension of the vector in (4.4) exceeds the dimension of the regression parameters, the quadratic inference function is used to obtain the estimator of β. The QIF is defined as

$$Q_N(\beta) = N\bar{g}_N' \bar{C}_N^{-1} \bar{g}_N, \quad (4.5)$$

where $\bar{C}_N = \frac{1}{N}\sum_{i=1}^{N} g_i(\beta) g_i(\beta)'$. The QIF estimator is the value of β that minimizes $Q_N(\beta)$. The advantage of the quadratic inference function approach is that it does not involve the nuisance parameters associated with the working correlation.

In the following we will illustrate how the idea of QIF can be applied for variance components estimation. Suppose V_i is a variance-covariance matrix for subject i, and w_i is a vector containing the components of V_i as

$$w_i = w_i(\psi) = (w_{i11}, w_{i1T}, \ldots, w_{i22}, \ldots, w_{iTT})'.$$

Let k_{il} be a vector containing the components of $U_{il}U_{il}'$ in (4.2), and let $k_{i(r+1)}$ be a vector with the entries of the matrix I. Then the estimation of variance components can be transformed to a linear estimation problem. Using (4.2), we have a linear form of w_i,

$$w_i = \sum_{l=1}^{r+1} \psi_l k_{il}, \quad (4.6)$$

where ψ_j is the element of ψ. Let $s_i = (s_{i11}, s_{i12}, \ldots, s_{i22}, \ldots, s_{iTT})'$ where

$$s_{ijk} = (y_{ij} - \mu_{ij})(y_{ik} - \mu_{ik}).$$

Note that $E(s_i) = w_i$ if μ_i is known. Following Qu et al. (2000), we define the second-order extended scores associated with ψ as

$$\bar{g}_N^v(\psi) = \frac{1}{N}\sum_{i=1}^{N} g_i^v = \frac{1}{N}\begin{pmatrix} \sum_{i=1}^{N} \dot{w}_i' B_i^{-1/2} M_1^* B_i^{-1/2} (s_i - w_i) \\ \vdots \\ \sum_{i=1}^{N} \dot{w}_i' B_i^{-1/2} M_m^* B_i^{-1/2} (s_i - w_i) \end{pmatrix}, \tag{4.7}$$

where $\dot{w}_i = \frac{\partial}{\partial \psi} w_i$, B_i is a diagonal variance matrix of s_i and M_j^* are known basis matrices associated with the working correlation. There are more equations than unknown parameters in (4.7). We can estimate ψ using the second-order quadratic inference function. That is, estimate ψ by minimizing

$$Q_N^v(\psi) = N\bar{g}_N^v(\psi)'\bar{C}_v^{-1}\bar{g}_N^v(\psi), \tag{4.8}$$

where $\bar{C}_v = (1/N)\sum_{i=1}^{N} g_i^v(\psi) g_i^v(\psi)'$.

If the mean vector μ_i is known, setting (4.7) to zero gives an unbiased estimating equation. However, the mean vector μ_i for each individual is usually unknown. But we could replace μ_i with its estimator $\hat{\mu}_i = X_i\hat{\beta}$ where $\hat{\beta}$ is obtained by minimizing (4.5). For large sample sizes, the second-order QIF estimators are asymptotically consistent. However, for small samples, a bias correction procedure might be required. Details about the bias correction are provided in section 4.5.

This estimating equation approach for variance components can be applied to linear or nonlinear regression as long as (4.2) is known. In this paper,

we apply the second-order quadratic inference function only for linear mixed models. Further work on the nonlinear case will be investigated in the future.

4.4. Asymptotic properties of the second-order QIF estimators

In this section, we will establish the asymptotic properties of the second-order QIF estimator $\hat{\psi}$ which minimizes (4.8). First we require the following theorems and regularity conditions to establish the asymptotic properties of the second-order QIF estimator $\hat{\psi}$.

The regularity conditions are

(i) the weighting matrix $\bar{C}_N = N^{-1}\sum g_i g_i'$ converges almost surely to a constant matrix C_0

(ii) the regression parameter β is identified, that is, there exists a unique $\beta_0 \in \Theta$ satisfying $E\{g(\beta_0)\} = 0$, where Θ is the parameter space

(iii) the parameter space Θ is compact and β_0 is an interior point of Θ

(iv) $E\{\bar{g}(\beta)\}$ is continuous in β.

THEOREM 4.4.1 (Qu, 1998) *If the above regularity conditions are satisfied, then the first-order QIF estimator $\hat{\beta}$ obtained by minimizing* (4.5) *is consistent.*

THEOREM 4.4.2 (Qu, 1998) *If the regularity conditions of Theorem 4.4.1 hold and the first derivative of $\bar{g}_N$ exists and is continuous, and $\frac{\partial}{\partial\beta}\bar{g}_N(\hat{\beta}) \xrightarrow{p} J_0$, where $J_0 = E[\frac{\partial}{\partial\beta}\bar{g}_N(\beta_0)]$ when $\hat{\beta} \xrightarrow{p} \beta_0$, then the first-order QIF estimator $\hat{\beta}$ minimizing* (4.5) *is asymptotically normal and efficient. That is*

$\sqrt{N}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, (J_0' C_0^{-1} J_0)^{-1})$.

The following assumptions $1-5$ are required to establish the consistency property for $\hat{\psi}$.

Assumption 1: The parameter space Ψ and Θ are compact. The true parameters ψ_0 and β_0 are interior points of Ψ and Θ.

Assumption 2: There is a unique (ψ_0, β_0) which satisfies $\mathrm{E}[g^v(\psi_0, \beta_0)] = 0$.

Assumption 3: The weighting matrix $\bar{C}_v$ converges almost surely to a constant matrix D_0.

Assumption 4: We require that $\mathrm{E}[g^v(\psi, \beta)]$ is continuous in ψ and β.

Assumption 5: $\mathrm{E}[\sup_{\beta \in B(\beta_0), \psi \in \Psi} |g^v(\psi, \beta)|] < \infty$, where $B(\beta_0)$ is a neighborhood around β_0.

THEOREM 4.4.3 (Consistency) *Under assumptions 1-5, if the estimator $\hat{\beta}$ converges in probability to β_0, then the estimator $\hat{\psi}(\hat{\beta})$ by minimizing* (4.8) *exists and $\hat{\psi} \xrightarrow{p} \psi_0$.*

The proof of Theorem 4.4.3 is provided in Appendix A.

Next, we establish the asymptotic normality of the second-order QIF estimator. The following additional assumptions $6-10$ are required for establishing the asymptotic normality property for $\hat{\psi}$.

Assumption 6: $Q_N^v(\psi, \beta)$ is continuous and twice differentiable in a neighborhood B of (ψ, β), where $B = B_\psi \times B_\beta$.

Assumption 7: The first derivative of $\bar{g}_N^v$ exists and is continuous, and $\frac{\partial}{\partial \psi} \bar{g}_N^v(\hat{\psi}, \hat{\beta})$ converges in probability to K_0 when $\hat{\psi}$ and $\hat{\beta}$ converge in probability to ψ_0 and β_0.

Assumption 8: Suppose $U_N(\hat{\psi}, \hat{\beta}) = \frac{\partial}{\partial \psi} Q_N^v(\hat{\psi}, \hat{\beta})$. We require that the covariance matrix $\text{cov}(U_N(\hat{\psi}, \hat{\beta}), \hat{\beta} - \beta_0)$ converges almost surely to a constant matrix $\Omega_{\psi\beta}$ and $\Omega_{\beta\psi} = \Omega'_{\psi\beta}$.

Assumption 9: We require that $S_N(\hat{\psi}, \hat{\beta}) = \frac{\partial}{\partial \psi \psi'} Q_N^v(\psi, \beta)$ exists and is continuous, and converges in probability to $S_0 = S(\psi_0, \beta_0)$ when $\hat{\psi}$ and $\hat{\beta}$ converge in probability to ψ_0 and β_0.

Assumption 10: We require that $H_N(\hat{\psi}, \hat{\beta}) = \frac{\partial}{\partial \psi \partial \beta'} Q_N^v(\hat{\psi}, \hat{\beta})$ exists and is continuous, and converges in probability to $H_0 = H(\psi_0, \beta_0)$ when $\hat{\psi}$ and $\hat{\beta}$ converge in probability to ψ_0 and β_0.

THEOREM 4.4.4 (Normality) *If $\hat{\beta} \xrightarrow{p} \beta_0$ and $\hat{\psi} \xrightarrow{p} \psi_0$, under assumptions 1-10, the second-order QIF estimator $\hat{\psi}$ obtained by minimizing* (4.8) *is asymptotically normal. That is*

$$\sqrt{N}(\hat{\psi} - \psi_0) \xrightarrow{d} N(0, S_0^{-1}(\Omega_{\psi\psi} + \Omega_{\psi\beta} H_0 + H_0 \Omega_{\beta\psi} + H_0 \Omega_{\beta\beta} H_0) S_0^{-1}),$$

where $\Omega_{\psi\psi} = (K_0' D_0^{-1} K_0)^{-1}$ and $\Omega_{\beta\beta} = (J_0' C_0^{-1} J_0)^{-1}$.

The proof of Theorem 4.4.4 is also given in Appendix A.

4.5. Bias Correction

Based on the large sample properties in section 4.4, the second-order QIF provides asymptotically consistent estimators for variance components. However, for small samples, bias of the estimators may occur. In this section, we illustrate a bias correction procedure for finite samples.

Suppose that the marginal mean of y_i is $\mu_i = X_i\beta$. Let $g_i = T_i'(y_i - \mu_i)$ in (4.4), where $T_i = (A_i^{-1/2}M_1A_i^{-1/2}X_i, \ldots, A_i^{-1/2}M_mA_i^{-1/2}X_i)$. In addition, let $z_i = T_i'y_i$, $K_i = T_i'X_i$, then we have $g_i = z_i - K_i\beta$. The QIF estimator of β is obtained by minimizing (4.5) which is equivalent to solving the estimating equation (Qu and Birkes, 2004)

$$\dot{\bar{g}}_N'\bar{C}_N^{-1}\bar{g}_N = 0,$$

where $\bar{g}_N = \frac{1}{N}\sum_{i=1}^{N} T_i'(y_i - X_i\beta)$ and $\dot{\bar{g}}_N' = \frac{\partial}{\partial\beta}\bar{g}_N = -\frac{1}{N}\sum T_i'X_i$. It follows that

$$\hat{\beta} = -(\dot{\bar{g}}_N'\bar{C}_N^{-1}\dot{\bar{g}}_N)^{-1}\dot{\bar{g}}_N'\bar{C}_N^{-1}\bar{z}_N,$$

where $\bar{z}_N = \frac{1}{N}\sum z_i$.

Suppose that T_i are the same for all subjects, then we will show that the correction term is based on $\mathrm{E}(\hat{s}_i - w_i)$ and is approximately equal to $\mathrm{var}(\hat{\mu})$. First, it can be shown that

$$\begin{aligned}
\hat{s}_i &= (y_i - \hat{\mu}_i)(y_i - \hat{\mu}_i)' \\
&= \{y_i - \mu_i - (\hat{\mu}_i - \mu_i)\}\{y_i - \mu_i - (\hat{\mu}_i - \mu_i)\}' \\
&= (y_i - \mu_i)(y_i - \mu_i)' - (y_i - \mu_i)(\hat{\mu}_i - \mu_i)' - (\hat{\mu}_i - \mu_i)(y_i - \mu_i)' + (\hat{\mu}_i - \mu_i)(\hat{\mu}_i - \mu_i)',
\end{aligned}$$

and

$$\mathrm{E}(\hat{s}_i) = w_i - \mathrm{cov}(y_i, \hat{\mu}_i) - \mathrm{cov}(\hat{\mu}_i, y_i) + \mathrm{E}[(\hat{\mu}_i - \mu_i)(\hat{\mu}_i - \mu_i)'].$$

Assume that $\hat{\mu}_i$ is an approximate unbiased estimator of μ_i. It follows that

$$\mathrm{E}(\hat{s}_i) = w_i - \mathrm{cov}(y_i, \hat{\mu}_i) - \mathrm{cov}(\hat{\mu}_i, y_i) + \mathrm{var}(\hat{\mu}_i).$$

If we can show that $\mathrm{cov}(\hat{\mu}_i, y_i) \approx \mathrm{var}(\hat{\mu}_i)$, then we have

$$\mathrm{E}(\hat{s}_i) \approx w_i - \mathrm{var}(\hat{\mu}_i).$$

Therefore, $\mathrm{var}(\hat{\mu}_i)$ can be the correction term for the empirical covariance $\hat{s}_i$. The proof of $\mathrm{cov}(\hat{\mu}_i, y_i) \approx \mathrm{var}(\hat{\mu}_i)$ is given in Appendix B. Note that this bias correction procedure is applied only for continuous data when the covariate matrices are the same for all subjects. We will further investigate this bias correction procedure for more complicated cases.

4.6. Asymptotic Chi-squared Test

In this section, we propose an asymptotic chi-squared test for testing the variance components of interest.

Qu et al.(2000) proposed an asymptotic chi-squared test for regression parameters based on the quadratic inference function. The second-order QIF approach also provides an asymptotic chi-squared test for testing variance components of interest. Suppose the vector ψ is partitioned into (ψ_1, ψ_2), where ψ_1 is a vector of variance components of interest with dimension q_1, and ψ_2 is a vector of nuisance variance components with dimension $q + 1 - q_1$. Suppose that the null hypothesis is $H_0 : \psi_1 = 0$ and the alternative hypothesis is H_1 : at least one element of ψ_1 is greater than zero. A test statistic T^* based on the second-order quadratic inference function can be obtained as

$$T^* = Q_N^v(0, \tilde{\psi}_2) - Q_N^v(\hat{\psi}_1, \hat{\psi}_2), \tag{4.9}$$

where $\hat{\psi}_2$ and $\tilde{\psi}_1, \tilde{\psi}_2$ are the estimators which minimize (4.8) under H_0 and H_1, respectively. Note that the alternative hypothesis is a one-sided test on the boundary of the parameter space. Therefore, T^* does not asymptotically follow a chi-squared distribution with degrees of freedom q_1. Consider a simple case where the dimension of q_1 is one. The asymptotic null distribution of T^* follows a mixture chi-squared distribution as $0.5\chi_0^2 + 0.5\chi_1^2$ (Stram and Lee, 1994), where the χ_0^2 is zero with probability 1.

In addition, this chi-squared test can be used for testing whether the serial correlation is significant. Suppose that W has an AR-1 structure

$$\sigma_e^2 \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{T-1} \\ \rho & 1 & \rho & \cdots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \cdots & 1 \end{pmatrix}. \tag{4.10}$$

Under the constraints $\sigma_e^2 > 0$ and $|\rho| < 1$, W is positive definite. In order to get a linear representation of W as shown in section 4.3, we apply the first-order Taylor expansion at a prior value (σ_0^2, ρ_0) (Witkovsky, 1996). The covariance matrix W can be approximately represented as

$$W(\sigma_e^2, \rho) \approx W_0^* + (\sigma_e^2 - \sigma_0^2)W_1^* + (\rho - \rho_0)W_2^*, \tag{4.11}$$

where $W_0^* = W(\sigma_0^2, \rho_0)$, $W_1^* = \frac{\partial}{\partial \sigma_e^2} W(\sigma_e^2, \rho)\Big|_{\sigma_0^2, \rho_0}$ and $W_2^* = \frac{\partial}{\partial \rho} W(\sigma_e^2, \rho)\Big|_{\sigma_0^2, \rho_0}$. Since $W_0^* = \sigma_0^2 W_1^*$, it follows that $W(\sigma_e^2, \rho) \approx \sigma_e^2 W_1^* + \rho^* W_2^*$, where $\rho^* = \rho - \rho_0$. Following section 4.3, the estimation of covariance components can also be

transformed to a linear estimation problem as

$$w_i = \sum_{l=1}^{r+2} \psi_l k_{il}, \tag{4.12}$$

where $\psi_{r+1} = \sigma_e^2$, $\psi_{r+2} = \rho^*$, $k_{i(r+1)}$ is a vector with components of variance-covariance matrix W_1^* and $k_{i(r+2)}$ is a vector with components of variance-covariance matrix W_2^*. The asymptotic chi-squared test can be used for testing whether the serial correlation ρ is significant. Specifically, for $\sigma_0^2 = 1$ and $\rho_0 = 0$, we have $\psi_(r+1) = \sigma_e^2$, $\psi_(r+2) = \rho$, $W_1^* = I$ and W_2^* is a matrix with 1 on the two main off-diagonals and 0 elsewhere.

Assume that the null hypothesis is $H_0 : \rho = 0$ and the alternative hypothesis is $H_1 : \rho \neq 0$. The test statistic

$$T^* = Q_N^v(\tilde{\sigma}_1^2, \ldots, \tilde{\sigma}_r^2, \tilde{\sigma}_e^2, 0) - Q_N^v(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_r^2, \hat{\sigma}_e^2, \hat{\rho}),$$

where $\tilde{\sigma}_1^2, \ldots, \tilde{\sigma}_r^2, \tilde{\sigma}_e^2$ and $\hat{\sigma}_1^2, \ldots, \hat{\sigma}_r^2, \hat{\sigma}_e^2, \hat{\rho}$ are estimators obtained by minimizing (4.8) under H_0 and H_1. The test statistic T^* asymptotically follows a chi-squared distribution with 1 degree of freedom.

4.7. Simulation

In this section several simulation studies are conducted for evaluating the performance of our approach under the linear mixed model setting.

The first simulation shows the effect of sample size on test size for testing variance components of interest using our approach. We choose 5 different sample sizes $50, 100, 150, 200, 250$. For each subject i, there are 10 repeated

measurements generated at time points $t = (1, 2, \ldots, 10)$. The response variable y_{ij} is generated from a linear mixed model

$$y_{ij} = \alpha_0 + \alpha_1 t + b_i + e_{ij}, \tag{4.13}$$

where $\alpha_0 = 2$, $\alpha_1 = -1$, and b_i and e_{ij} are normal distributed with mean zero and variance one. Note that this simulation setup will be applied for the remaining simulations.

We fit a linear mixed model with random intercept only to these simulated datasets. In this simulation, the serial correlation is not considered. We first obtain the fixed effects estimators by minimizing (4.5) with the AR-1 working correlation. Then we apply the bias correction along with the second-order QIF to obtain the estimators of variance components. In addition, the asymptotic chi-squared test is applied for testing whether the random intercept is significant. Simulation results are given in Table 4.1 including the test sizes under different sample sizes with 1000 simulations. Table 4.1 shows that the test size is .034 when sample size is 50, a more conservative test. Evidently, as sample size increases, the test size also increases. The test size is .044 when the sample size is 200. However, as the sample size increases to 250, the test size does not increase.

Secondly, we examine the power function of the asymptotic chi-squared test for $\sigma_b^2 = (0, .01, \ldots, .09)$. We generate 1000 random samples from (4.13), and each random sample contains 200 subjects. A linear mixed model with random intercept is used to fit these datasets. We obtain the power of the test for different σ_b^2 using our approach. Figure 4.1 illustrates the power curve for different σ_b^2. As expected, as the value of σ_b^2 increases, the probability of

rejecting the null hypothesis also increases. In addition, the power is close to 0.05 when σ_b^2 is close to 0, and is close to 1 at $\sigma_b^2 = .06$.

In the third simulation, we study the effect of sample size on test size for testing whether the serial correlation is significant based on 5 different sample sizes $50, 100, 150, 200, 250$. The response y_{ij} is generated from (4.13). We fit a linear mixed model with random intercept to these datasets. However, in our fitting, the covariance of e_i is assumed to have the form of (4.10). We first obtain the fixed effects estimators by minimizing (4.5) with the AR-1 working correlation. Then we apply Taylor's expansion on W at $\sigma^2 = 1$ and $\rho = 0$ as shown in section 4.6. It follows that the covariance of the marginal distribution y_i can be written as $z_i z_i' \sigma_b^2 + \sigma_e^2 I + \rho W_2^*$. The estimators of variance components σ_b^2, σ_e^2 and ρ are obtained by minimizing (4.8). In addition, the asymptotic chi-squared test is used for testing whether the serial correlation is significant. Table 4.2 provides the test sizes under different sample sizes with 1000 simulations. As shown in Table 4.2 , the test size is 0.077 when the sample size is 50. Apparently, as the sample size increases, the test size is closer to 0.05.

In the fourth simulation, we examine the performance of variance components estimation and the power function of the asymptotic chi-squared test for different serial correlations $\rho = (-.8, .6, \ldots, -.2, 0, .2, \ldots, .8)$. Data are generated from (4.13) with serial correlation for $i = 1, \ldots, 250$. We fit these simulated datasets using our approach by applying Taylor's expansion on W at $\sigma^2 = 1$ and $\rho = 0$. Figure 4.2 illustrates the power curve for different serial correlation from 1000 simulations. Evidently, as the absolute value of ρ

increases, the probability of rejecting the null hypothesis also increases. The power is close to 0.05 when ρ is close 0, and is approximately close to 1 at $\rho = \pm 0.15$.

Table 4.3 provides the estimation of variance components when different serial correlations arise. First, the bias of the estimator $\hat{\sigma}_b^2$ is .02 when the serial correlation is -.8. As the serial correlation increases and is close to zero, the difference between the estimator and the true parameter is smaller and is close to zero. However, the bias is .40 when the serial correlation is extremely large, say .8. This can also be seen in the estimation of σ_e^2. The bias of $\hat{\sigma_e^2}$ is .64 and .67 when the serial correlation is -.8 and .8. This may be caused by using $\rho = 0$ as the prior, which is far from the true value of ρ (Witkovsky, 1996). Therefore, we use $\sigma_0^2 = 1$ and $\rho_0 = .75$ as a prior for $\rho = .8$. The estimators and standard errors (the entry in parentheses) of ${\sigma_b}^2$, ${\sigma_e}^2$ and ρ are .98 (.16), .99 (.06) and .79 (.01), respectively. Therefore, a good Taylor approximation may require accurate prior information of σ_0 and ρ_0.

4.8. Application

As an illustration, we apply the second-order quadratic inference function to the pig bodyweight dataset which can be found in Diggle et al. (2002). This dataset contains pig weights measured on the same day of each week over a 9-week period. There are a total of 48 pigs. We fit this dataset with a linear random intercept model as

$$y_i = \alpha_0 + \alpha_1 t + b_i + e_i, \tag{4.14}$$

where y_i is a 9-dimensional bodyweight vector of pig i, $t = (1, \ldots, 9)'$, and b_i is the random intercept with mean zero and variance σ_b^2. We also assume that the random error e_i follows a normal distribution with mean 0 and variance σ_e^2. Note that serial correlation is not considered in this model.

We first apply the first-order quadratic inference function to obtain the fixed effects estimator with the AR-1 working correlation. Since the covariate matrices are the same for all pigs, we apply the bias correction as shown in section 4.5. The second-order QIF is used to estimate σ_e^2 and σ_b^2. In addition, the asymptotic chi-squared test is also applied to test whether the random intercept is significant. Furthermore, we compare our approach to the MLE and REMLE. The estimators of MLE and REMLE are obtained using the SAS proc mixed procedure. Table 4.4 provides the estimators for fixed effects and variance components using the three approaches. The QIF_v denotes the second-order QIF approach. First, the values of the second-order QIF under $H_0 : \sigma_b^2 = 0$ and $H_1 : \sigma_b^2 > 0$ are 3.71 and 26.26. The p-value of the asymptotic chi-squared test is less than 0.01 which is the same as those obtained from the MLE and REMLE. Thus, there is strong evidence that the random intercept is significant. In addition, our approach provides similar estimators to those obtained from MLE and REML for σ_e^2 and σ_b^2.

We also fit this dataset with a model which includes random effects (random slope) and serial correlation as

$$y_i = \alpha_0 + (\alpha_1 + b_{1i})t + e_i,$$

where b_i is the normal distributed with mean zero and variance σ_b^2. In this model, we assume that e_i follows an AR-1 process. The Taylor's expansion

method is applied at $\sigma_0^2 = 7$ and $\rho_0 = .95$ to obtain the estimators of σ_b^2, σ_e^2 and ρ. The asymptotic chi-squared test is also applied to test whether the serial correlation is significant. We also compare our approach to the MLE and REMLE. Table 4.5 shows the estimators of σ_b^2, σ_e^2 and ρ using the second-order QIF, MLE and REMLE. Noticeably, there is a discrepancy of σ_e^2 estimators between the second-order QIF and likelihood approaches (MLE and REMLE). The estimators of σ_b^2 and ρ are very similar for the three approaches. In addition, the values of the second-order QIF under H_0 and H_1 are 16.74 and 3.41. The p-values of the asymptotic chi-squared test with degrees of freedom 1 for ρ are less than .01, which is also same as those obtained from MLE and REMLE. Therefore, all three approaches yield strong evidence that the serial correlation is significant in the second model.

4.9. Discussion

In this paper, we propose a new approach for estimating variance components in linear mixed models. This approach is an extension of the quadratic inference function. We estimate the variance components using the second-order extended scores along with the fixed effects estimator obtained from the first-order quadratic inference function.

In addition, for small samples, we propose a bias correction procedure for the estimation of variance components using the second-order QIF. The advantage of our approach is that the second-order QIF provides an asymptotic chi-squared test for testing random effects of interest. In addition, in our models, we also incorporate both random effects and serial correlations. To

estimate the serial correlation, the covariance matrix of errors is approximated by a linear form of basis matrices using Taylor's expansion. The asymptotic chi-squared test can also be applied for testing the presence of serial correlation.

The simulation results indicate that our approach performs well on the estimation of variance components for large sample sizes. In addition, the test size of our asymptotical chi-squared test for serial correlation is close to 0.05 as the sample size increases. Although the simulation results show that the test size for testing random effects is more conservative, more simulation studies might be needed to investigate this issue.

In this paper, we apply our approach only for continuous longitudinal data. However, categorical data often arise in longitudinal studies. We will further extend our approach to categorical longitudinal data. In addition, the bias correction procedure is only applicable when all subjects have the same covariates. Further work on the bias correction procedure when subjects have different covariates will be needed. Moreover, the simulation results suggest that the estimation of the second-order QIF might rely on the empirical estimation of σ_0^2 and ρ_0. We will also further investigate the effect of the prior value on the estimation of variance components.

4.10. References

[1] Chi, E. M. & Reinsel, G. C. (1989). Models for longitudinal data with random effects and AR(1) errors. *Journal of the American Statistics Association*, 84, 452-459.

[2] Diggle P. J., Heagerty, P., Liang, K.-Y. & Zeger S. L. (2002). *Analysis of longitudinal data.* Oxford: Oxford University Press.

[3] Hartley, H. O. & Rao, J. N. K. (1967). Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika*, 54, 93-108.

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

[5] Laird, N. M. & Ware, J. H. (1982). Random-effects model for longitudinal data. *Biometrics*, 38, 963-974.

[6] Patterson, H. D. & Thompson, R. (1971). Recovery of interblock information when block sizes are unequal. *Biometrika*, 58, 545-554.

[7] Qu, A. (1998) *Adaptive Generalized Estimating Equations.* Thesis.

[8] Qu, A. & Birkes, D. (2004), Robust estimation when there are more equations than unknown parameters for longitudinal data. *Manuscript.*

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

[10] Stram, D. O. & Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, 50, 1171-1177.

[11] Verbeke, G., Lesaffre, E. & Brant, L. (1998). The detection of residual serial correlation in linear mixed models. *Statistics in Medicine* 17, 1391-1402.

[12] Verbeke, G. & Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data.* Springer-Verlag, New York.

[13] Witkovsky, v. (1996). On variance-covariance components estimation in linear models with AR(1) disturbances. *Acta Math. Univ. Commenianae*, LXV, 1, 129-139.

4.11. Appendix A

Proof of Theorem 4.4.3: First, the estimator $\hat{\psi}$ by minimizing (4.8) exists since the estimating equation $\bar{g}_N^v$ is continuous on a compact space Ψ under Assumption 1 and Assumption 4.

In practice, minimizing (4.8) is equivalent to minimizing

$$\bar{g}_N^v(\psi,\beta)'\bar{C}_v^{-1}\bar{g}_N^v(\psi,\beta).$$

Here we denote $Q_N^v = g_N^{v\prime}\bar{C}_v^{-1}g_N^v$, where $g_N^v = \bar{g}_N^v(\psi,\beta)$. Following Assumption 3, we have $\bar{C}_v$ converges to D_0. Let $Q_0^v = g_0^{v\prime}D_0g_0^v$, where $g_0^v = E\{g^v(\psi,\beta)\}$. First we will show that the $Q_N(\psi,\beta)$ converges uniformly in probability to $Q_0(\psi,\beta)$ for all β in the neighborhood of β_0. This can be shown using the same arguments from Newey and McFadden's (1994) Theorem 2.6. First Q_0 is continuous from Assumption 4. Secondly, $\sup_{\beta\in B(\beta_0)\psi\in\Psi}|\bar{g}_N^v - g_0^v| \xrightarrow{p} 0$ by the uniform law of large numbers or Glivenko-Cantelli theorem (Billingsley, 1995). Finally, by the triangle and Cauchy-Schwartz inequalities (Newey and McFadden, 1994, Theorem 2.6), we have

$$\begin{aligned}
&|Q_N^v(\psi,\beta) - Q_0^v(\psi,\beta)| \\
&\leq |(\bar{g}_N^v - g_0^v)'\bar{C}_v(\bar{g}_N^v - g_0^v)| + |g_0^{v\prime}(\bar{C}_v + \bar{C}_v)'(\bar{g}_N^v - g_0^v)| + |g_0^{v\prime}(\bar{C}_v - D_0)'g_0^v| \\
&\leq |\bar{g}_N^v - g_0^v|^2|\bar{C}_v| + 2|g_0^v||\bar{g}_N^v - g_0^v||\bar{C}_v| + |g_0^v|^2|\bar{C}_v - D_0|, \qquad (4.15)
\end{aligned}$$

where $|.|^2$ is the inner product of a vector. Because $\sup_{\beta\in B(\beta_0)}\sup_{\psi\in\Psi}|\bar{g}_N^v - g_0^v| \xrightarrow{p} 0$ and $\bar{C}_v \xrightarrow{p} D_0$, we have $\sup|Q_N^v(\psi,\beta) - Q_0^v(\psi,\beta)| \xrightarrow{p} 0$. Therefore, $Q_N^v(\psi,\beta)$ converges uniformly in probability to $Q_0^v(\psi,\beta)$.

Then we can use the argument from Newey and McFadden's (1994) Theorem 2.1, along with some modification. The following is our strategy:

Suppose that U is any neighborhood of ψ_0. If we can show that $Q_0^v(\hat{\psi}, \beta_0) < \inf_{\psi \in U^c} Q_0^v(\psi, \beta_0)$, then it is impossible that $\hat{\psi}$ remains in the outside of U, then eventually $\hat{\psi} \in U$.

First, we can rewrite $Q_0^v(\hat{\psi}, \beta_0)$ as

$$\begin{aligned} & Q_0^v(\hat{\psi}, \beta_0) \\ & = \{Q_0^v(\hat{\psi}, \beta_0) - Q_0^v(\hat{\psi}, \hat{\beta})\} + \{Q_0^v(\hat{\psi}, \hat{\beta}) - Q_N^v(\hat{\psi}, \hat{\beta})\} + \\ & \quad \{Q_N^v(\hat{\psi}, \hat{\beta}) - Q_N^v(\psi_0, \hat{\beta})\} + \{Q_N^v(\psi_0, \hat{\beta}) - Q_0^v(\psi_0, \hat{\beta})\} + Q_0^v(\psi_0, \hat{\beta}). \quad (4.16) \end{aligned}$$

First, by Theorem 4.4.1, $\hat{\beta}$ converges in probability to β_0. Therefore, by the Slutzky theorem, we have

$$|Q_0^v(\hat{\psi}, \beta_0) - Q_0^v(\hat{\psi}, \hat{\beta})| < \epsilon/4.$$

Secondly, by Assumption 3 along with Assumption 4 and the uniform convergence result earlier, we have

$$|Q_0^v(\hat{\psi}, \hat{\beta}) - Q_N^v(\hat{\psi}, \hat{\beta})| < \epsilon/4,$$

and

$$|Q_N^v(\psi_0, \hat{\beta}) - Q_0^v(\psi_0, \hat{\beta})| < \epsilon/4.$$

Third, we have

$$Q_N^v(\hat{\psi}, \hat{\beta}) - Q_N^v(\psi_0, \hat{\beta}) < 0,$$

since $\hat{\psi} = \text{argmin}_{\psi} Q_N^v(\psi, \hat{\beta})$. Last, we have

$$Q_0^v(\psi_0, \hat{\beta}) < \epsilon/4,$$

by the Slutzky theorem and Assumption 2, $Q_0^v(\psi_0, \hat{\beta}) \xrightarrow{p} 0$.

It follows that $Q_0^v(\hat{\psi}, \beta_0) < \epsilon$, with probability one. Thus, if we choose $\epsilon = \inf_{\psi \in U^c} Q_0^v(\psi, \beta_0)$, it follows that $Q_0^v(\hat{\psi}, \beta_0) < \inf_{\psi \in U^c} Q_0^v(\psi, \beta_0)$ with probability 1.

Proof of Theorem 4.4.4: Since $\hat{\psi}$ is obtained by minimizing (4.8), $\hat{\psi}$ satisfies

$$\dot{Q}_N^v(\hat{\psi}, \hat{\beta}) = 0, \tag{4.17}$$

where $\dot{Q}_N^v$ is the first derivative of the quadratic inference function Q_N^v with respect to ψ, $\dot{Q}_N^v = \frac{\partial}{\partial \psi} Q_N^v$. By the first order Taylor's expansion, we have

$$0 = \dot{Q}_N^v(\hat{\psi}, \hat{\beta}) = \dot{Q}_N^v(\psi_0, \beta_0) + \frac{\partial \dot{Q}_N^v(\tilde{\psi}, \tilde{\beta})}{\partial \psi}(\hat{\psi} - \psi_0) + \frac{\partial \dot{Q}_N^v(\tilde{\psi}, \tilde{\beta})}{\partial \beta}(\hat{\beta} - \beta_0), \tag{4.18}$$

where $(\tilde{\psi}, \tilde{\beta})$ is an interior point of parameter space $\Phi \times \Theta$ and is between $(\hat{\psi}, \hat{\beta})$ and (ψ_0, β_0). Thus, we have

$$\begin{aligned} \hat{\psi} - \psi_0 &= -\left(\frac{\partial}{\partial \psi} \dot{Q}_N^v(\tilde{\psi}, \tilde{\beta}) \right)^{-1} \left(\dot{Q}_N^v(\psi_0, \beta_0) + \frac{\partial}{\partial \beta} \dot{Q}_N^v(\tilde{\psi}, \tilde{\beta})(\hat{\beta} - \beta_0) \right) \\ &= -S_N^{-1}(\tilde{\psi}, \tilde{\beta}) \begin{pmatrix} 1 & H_N(\tilde{\psi}, \tilde{\beta}) \end{pmatrix} \begin{pmatrix} U_N(\psi_0, \beta_0) \\ \hat{\beta} - \beta_0 \end{pmatrix}. \end{aligned} \tag{4.19}$$

Using the Central Limit Theorem, it can be shown that

$$\sqrt{N} \bar{g}_N^v(\psi_0, \beta_0) \xrightarrow{d} N(0, D_0).$$

By Assumption 6, $\frac{\partial}{\partial \psi} \bar{g}_N^v(\tilde{\psi}, \tilde{\beta})$ converges in probability to K_0 since $(\tilde{\psi}, \tilde{\beta})$ is between $(\hat{\psi}, \hat{\beta})$ and (ψ_0, β_0), where $(\hat{\psi}, \hat{\beta})$ converges to (ψ_0, β_0) in probability by Theorem 4.4.3. Therefore, it follows that $\sqrt{N} U_N(\psi_0, \beta_0) \xrightarrow{d} N(0, (K_0' D_0^{-1} K_0)^{-1})$.

Finally, by Theorem 4.4.2 and Assumption 8-10, it follows that

$$\sqrt{N}(\hat{\psi} - \psi_0) \xrightarrow{d} N(0, S_0^{-1}(\Omega_{\psi\psi} + \Omega_{\psi\mu}H_0 + H_0\Omega_{\mu\psi} + H_0\Omega_{\mu\mu}H_0)S_0^{-1}),$$

where $\Omega_{\psi\psi} = (K_0'D_0^{-1}K_0)^{-1}$ and $\Omega_{\beta\beta} = (J_0'C_0^{-1}J_0)^{-1}$, as was to be shown.

4.12. Appendix B: Bias Correction

Suppose that $\text{cov}(g_i) = C_0$ and $\bar{C}_N \approx C_0$. The variance of $(\hat{\mu}_i)$ can be written as

$$\begin{aligned}
\text{var}(\hat{\mu}_i) &\approx X_i(\dot{\bar{g}}_N' C_0^{-1} \dot{\bar{g}}_N)^{-1} \dot{\bar{g}}_N' C_0^{-1} \text{var}(\bar{z}_N) C^{-1} \dot{\bar{g}}_N (\dot{\bar{g}}_N' C^{-1} \dot{\bar{g}}_N)^{-1} X_i' \\
&\approx X_i(\dot{\bar{g}}_N' C_0^{-1} \dot{\bar{g}}_N)^{-1} \dot{\bar{g}}_N' C_0^{-1} C_0 C_0^{-1} \dot{\bar{g}}_N (\dot{\bar{g}}_N' C_0^{-1} \dot{\bar{g}}_N)^{-1} X_i' \qquad (4.20) \\
&\approx X_i(\dot{\bar{g}}_N' C_0^{-1} \dot{\bar{g}})^{-1} X_i'.
\end{aligned}$$

Next, we will obtain $\text{cov}(\hat{\mu}_i, y_i)$. First, we have $\text{cov}(y_i, y_j) = 0$ for $i \neq j$. Since $T_i = T_j$ for all $i \neq j$, it can be shown that $\text{cov}(\bar{g}_N) = N^{-1} T_i \text{cov}(y_i) T_i'$, and $\dot{\bar{g}}_N = \dot{g}_i = -T_i' X_i$ for all subject i.

It follows that the covariance of z_i and $\hat{\mu}_i$ is

$$\begin{aligned}
\text{cov}(\hat{\mu}_i, y_i) &\approx \text{cov}(X_i(\dot{\bar{g}}_N' C_0^{-1} \dot{\bar{g}}_N)^{-1} \dot{\bar{g}}' C_0^{-1} (N^{-1} \sum T_i' y_i), y_i) \\
&\approx X_i(\dot{\bar{g}}_N' C_0^{-1} \dot{\bar{g}}_N)^{-1} \dot{\bar{g}}' C_0^{-1} \text{cov}(y_i, N^{-1} \sum T_i' y_i) \\
&\approx X_i(\dot{\bar{g}}_N' C_0^{-1} \dot{\bar{g}}_N)^{-1} (X_i' T_i)(N^{-1} T_i \text{cov}(y_i) T_i')^{-1} (N^{-1} T_i' \text{cov}(y_i)) \\
&\approx X_i(\dot{\bar{g}}_N' C_0^{-1} \dot{\bar{g}}_N)^{-1} X_i'.
\end{aligned} \qquad (4.21)$$

By (4.20) and (4.21), we have $\text{cov}(y_i, \hat{\mu}_i) \approx \text{var}(\hat{\mu}_i)$, as to be shown.

TABLE 4.1. Test Sizes of $H_0 : \sigma_b^2 = 0$ for Different Sample Sizes

Sample Size	50	100	150	200	250
Size	.034	.037	.043	.044	.043

TABLE 4.2. Test Size of $H_0 : \rho = 0$ under Different Sample Sizes

Sample Size	50	100	150	200	250
Size	.077	.063	.064	.053	.051

TABLE 4.3. Estimation of variance components for different serial correlations using prior ($\sigma_0^2 = 1$, $\rho_0 = 0$)

ρ	-0.8	-0.6	-0.4	-0.2	0	0.2	0.4	0.6	0.8
$\hat{\sigma_b}^2$	.96	.99	1.00	.98	.97	.99	1.04	1.16	1.40
s.e.	.13	.13	.11	.11	.11	.11	.13	.15	.19
$\hat{\sigma}_e^2$	.36	.69	.91	.98	1.00	.98	.88	.64	.33
s.e.	.02	.04	.04	.03	.03	.04	.04	.04	.02
$\hat{\rho}$	-.17	-.30	-.32	-.20	-.00	.18	.27	.24	.14
s.e.	.02	.03	.03	.03	.03	.03	.03	.02	.02

TABLE 4.4. Estimation of fixed effects and variance components for pig body-weight data with random intercept

	QIF_v	MLE	REML
$\hat{\alpha_0}$	18.72	19.36	19.36
s.e.	.35	.60	.60
$\hat{\alpha_1}$	6.23	6.21	6.21
s.e.	.13	.04	.04
σ_b^2	15.42	15.14	14.82
s.e.	2.95	3.22	3.12
p-value of σ_b^2	< 0.01	< 0.01	< 0.01
σ_e^2	4.44	4.39	4.38
s.e.	.61	.32	.32

TABLE 4.5. Estimation of fixed effects and variance components for pig body-weight data with serial correlation and random slope

	QIF_v	MLE	REMLE
σ_b^2	.35	.39	.39
s.e.	.23	.10	.10
σ_e^2	4.74	6.20	6.33
s.e.	1.23	1.23	1.27
ρ	.87	.84	.84
s.e.	.05	.03	.03
p-value of ρ	$< .01$	$< .01$	$< .01$

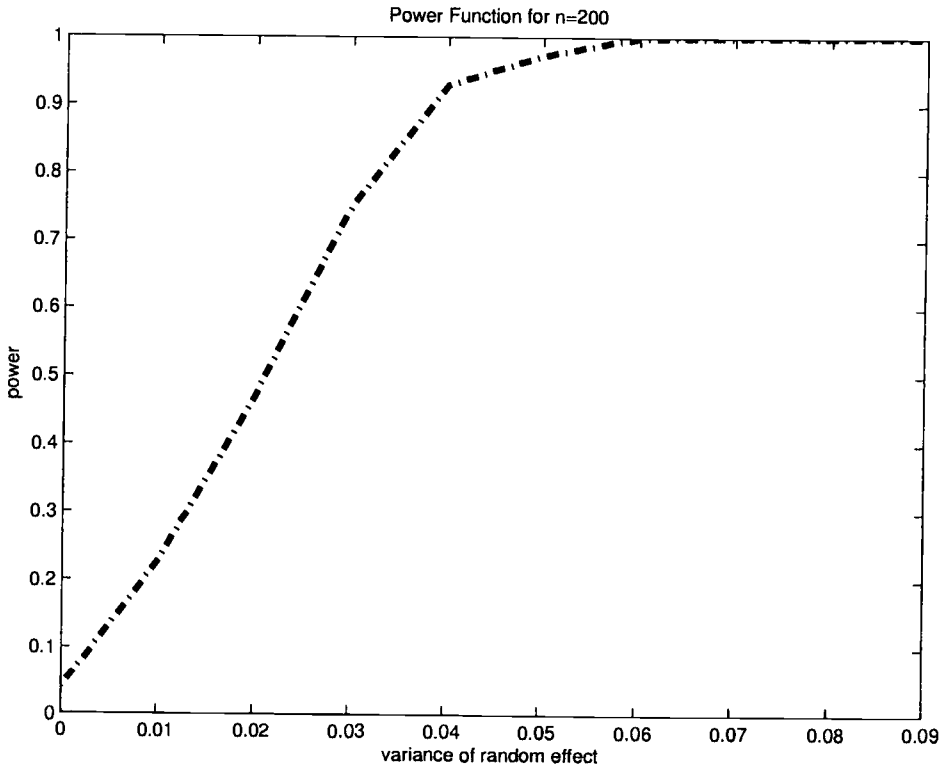

FIGURE 4.1. Power function of σ_b^2 when sample size is 200

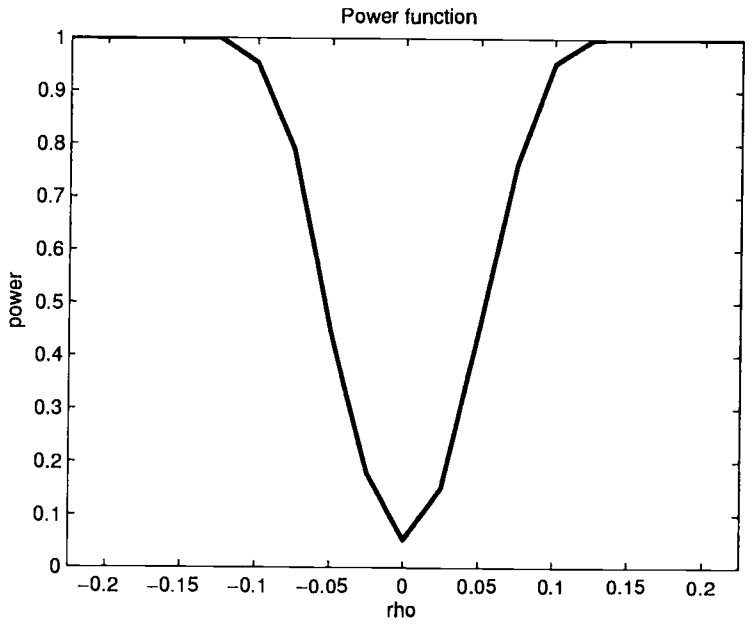

FIGURE 4.2. Power function of ρ when sample size is 250

5. CONCLUSION

This thesis investigates both marginal models and mixed-effects models in longitudinal data analysis. These investigations include marginal models for high dimensional cell cycle microarray data with no replicates, the estimation of fixed effects and random effects for generalized linear mixed models in longitudinal data analysis, and the estimation of variance components for linear mixed models.

In Chapter 2, we develop a new approach to analyze cell-cycle microarray data using a data classifying method, varying coefficient modeling and the quadratic inference function. Our classifying technique successfully identifies genes with similar cell cycle patterns into the same class, or into a class with no cell cycle phenomena at all. It also produces pseudo-replicates for statistical modeling and inference. In our model setting, varying coefficient models provide more flexibility than the classical parametric or semiparametric models and also take both time and covariates effects into account. We also apply the quadratic inference function for estimating parameters in varying coefficient models which not only incorporates the correlations but also does not require the estimation of the nuisance parameters associated with the correlation. Our approach also provides statistical numerical evidence based on the asymptotical properties of quadratic inference function. From the results of simulations, our approach performs well for large sample sizes and reasonably

well for smaller or moderate sample sizes. In addition, our approach successfully identifies genes whose expression levels are time varying. In our example, it is able to identify cell-cycle regulated genes even though we use rather small samples within a group. Our approach is more conservative in small sample size according to the simulation results. However, in small samples we prefer a conservative test.

Chapter 3 proposes a new approach for generalized linear mixed models. The new approach estimates fixed and random effects using two conditional extended scores which involve only the first two moments. Thus, the specification of a likelihood function is not required. In addition, it does not require the distribution assumption for random effects and is able to incorporate both serial correlation and random effects. Moreover, it does not involve unknown variance components as in PQL, or the nuisance parameters associated with working correlations as in CGEE2. Simulations for a few generalized linear mixed models indicate the following: First, the PQL and mixed QIF yield lower bias of estimators than marginal approaches when data are from random effects models. Secondly, if serial correlation is introduced into the data, the mixed QIF performs better than the others.

In Chapter 4, we develop a new approach for the estimation of variance components in linear mixed models. This approach is an extension of quadratic inference functions. We estimate variance components using the second-order extended scores along with the fixed effects estimators. The fixed effects estimator is obtained from the first-order quadratic inference function. In addition, we propose a bias correction procedure for the second-order QIF.

In our models, we also incorporate both random effects and serial correlations. In addition, an asymptotic chi-squared test based on the second-order QIF is proposed for testing random effects of interest and the presence of serial correlation. Simulations for a few linear mixed models indicate the following. First, the second-order QIF performs well on the estimation of variance components for large sample size. Secondly, the test size of our asymptotic chi-squared test for serial correlation is close to 0.05 as the sample size increases. Thirdly, the simulation results show that the test size for testing random effects is more conservative. However, in practice, we prefer conservative tests.

Further work suggested is listed as follows. First, for cell cycle microarray data analysis, our approach can only tell that whether expression levels are time varying or time invariant. If it is time invariant, it does not have cell cycle phenomena. However, if it is time varying, we can only conclude that it may have cell cycle phenomena. Nevertheless, our approach can be still easily implemented using periodic functions instead of polynomial functions. We will investigate it further in the future.

Secondly, for generalized linear mixed models in longitudinal data analysis, although our approach does not require the estimation of variance components for estimating fixed and random effects, it may perform slowly when the dimension of the random effects is high. We will continue to investigate the performance of our approach when the dimension of the random effects is high.

Finally, for the estimation of variance components in linear mixed model, we apply our approach only for continuous longitudinal data. However,

categorical data often arise in longitudinal studies. We will further extend our approach to categorical longitudinal data. In addition, the bias correlation is only applicable when all subjects have the same covariates. We will further investigate the bias correlation procedure if subjects have different covariates. Moreover, from the simulation results, the estimation of the second-order QIF might rely on the empirical estimation of σ_0^2 and ρ_0. Further investigation on the effect of the prior value for the variance components estimation using the second-order QIF will be needed.

BIBLIOGRAPHY

[1] Benjamini, Y. & Hochberg, Y.(1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc.* B 57, 289-300.

[2] Breslow, N. E. & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistics Association*, 88, 9-25.

[3] Chi, E. M. & Reinsel, G. C. (1989). Models for longitudinal data with random effects and AR(1) errors. *Journal of the American Statistics Association*, 84, 452-459.

[4] Cho, R., Campbell, M., Winzeler, E., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T., Gabrielian, A., Landsman, D., Lockhart, D. & Davis, R. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell* 2, 65-73.

[5] Demidenko, E. (2004). *Mixed Models: Theory and Applications.* John Wiley, New York.

[6] Diggle P. J., Heagerty, P., Liang, K.-Y. & Zeger S. L. (2002). *Analysis of longitudinal data.* Oxford: Oxford University Press.

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

[8] Hartley, H. O. & Rao, J. N. K. (1967). Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika*, 54, 93-108.

[9] Hastie, T. & Tibshirani, R. (1993). Varying-coefficient models (with discussion). *J.R. Statist. Soc.* B 55, 757-796.

[10] Huang, J. Z., Wu, C. O. & Zhou, L. (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika* 89, 111-128.

[11] Hoover, D. R. , Rice, J. A. , Wu, C. O. & Yang, L.-P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* 85 , 809-822.

[12] Jiang, J. (1999). Conditional inference about generalized linear mixed models. *The Annals of Statistics*, 27, 1974-2007.

[13] Laird, N. M. & Ware, J. H. (1982). Random-effects model for longitudinal data. *Biometrics*, 38, 963-974.

[14] Li, K. C., Yan, M. & Yuan, S. (2002). A simple statistical model for depicting the CDC15-synchronized yeast cell-cycle regulated gene expression data. *Statistica Sinca* 12, 141-158.

[15] Liang, K. Y. & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.

[16] Lin, X. & Breslow, N. L. (1996). Bias correction in generalized linear mixed effects models with multiple components of dispersion. *Journal of the American Statistics Association*, 91, 1007-1016.

[17] Liu, Q. & Pierce, D. A. (1994). A note on Gauss-Hermite quadrature. *Biometrika*, 81, 624-629.

[18] Madsen, L. & Dalthorp, D. 2005. Simulating correlated count data. *Environmental and Ecological Statistics*, In press.

[19] McCulloch, C. E. (1997). Maximum likelihood algorithm for generalized linear mixed models. *Journal of the American Statistics Association*, 92, 162-170.

[20] McCulloch, C. E. & Searle, S. R. (2001). *Generalized, Linear and Mixed Models.* John Wiley, New York.

[21] Möller-Levet, C., Klawonn, F., Cho, K.-H. & Wolkenhauer, O. (2003). Fuzzy clustering of short time series and unevenly distributed sampling points. *IDA*, 330-340.

[22] Newey, W.K. & McFadden, D. (1994). Large Sample Estimation and Hypothesis Testing. *In Handbook of Econometrics*, Vol. 4, ed. by R.F. Engle and D.L. McFadden, 2111-2245. Elsevier, North Holland.

[23] Park, C.G. & Shin, D.W. (1998). An algorithm for generating correlated random variables in a class of infinitely divisible distributions. *Journal of Statistical Computation and Simulation*, 61, 127-139.

[24] Patterson, H. D. & Thompson, R. (1971). Recovery of interblock information when block sizes are unequal. *Biometrika*, 58, 545-554.

[25] Pawitan, Y. (2001). *In all likelihood: Statistical Modelling and Inference using likelihood.* Oxford University Press.

[26] Qu, A. (1998) *Adaptive Generalized Estimating Equations.* Thesis.

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

[28] Qu, A. & Birkes, D. (2004), Robust estimation when there are more equations than unknown parameters for longitudinal data. *Manuscript.*

[29] Qu, A. & Li, R. (2005). Nonparametric modeling and inference function for longitudinal data. *Biometrics, to appear.*

[30] Ruppert, D. & Carroll, R.J. (1997). Penalized regression splines. *Technical Report*, Department of OR & IE, Cornell University, Ithaca, NY.

[31] Searle, S. R., Casella, G. & McCulloch, C.E. (1992). *Variance Components.* John Wiley, New York.

[32] Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. & Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell* 9, 3273-3297.

[33] Stram, D. O. & Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, 50, 1171-1177.

[34] Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. & Golub, T. (1999). Interpreting patterns of gene expression with self-organizing maps. *Proc. Natl. Acad. Sci.* 96, 2907-2912.

[35] Tavazoie, S., Hughes, J., Campbell, M., Cho, R. & Church, G. (1999). Systematic determination of genetic network architecture. *Nature Genetics* 22, 281-285.

[36] Thall, P. F. & Vail, S. C. (1990). Some covariance models for longitudinal count data with over-dispersion. *Biometrics*, 46, 657671.

[37] Verbeke, G. & Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data.* Springer-Verlag, New York.

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

[39] Wedderburn (1974). Quasi-likelihood functions, generalised linear models, and the Gauss-Newton method. *Biometrika* 61, 439-447.

[40] Witkovsky, v. (1996). On variance-covariance components estimation in linear models with AR(1) disturbances. *Acta Math. Univ. Commenianae*, LXV, 1, 129-139.

[41] Zeger, S. L. & Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42, 121-130.

[42] Zeger, S. L., Liang, K. Y. & P.S. Alberts (1988). Models for longitudinal data: A generalized estimating equations approach. *Biometrics*, 44, 1049-1060.

[43] Zhao, L. P., Prentice, R. & Breeden, L. (2001). Statistical modeling of large microarray data sets to identify stimulus-response profiles. *Proc. Natl. Acad. Sci.* 98, 5631-5636.