Abstract

YANG, HONGMEI. Variable Selection Procedures for Generalized Linear Mixed Models in Longitudinal Data Analysis. (Under the direction of Dr. Daowen Zhang and Dr. Hao Helen Zhang.)

Model selection is important for longitudinal data analysis. But up to date little work has been done on variable selection for generalized linear mixed models (GLMMs). In this dissertation, we propose and study a class of variable selection methods. For GLMMs with low-dimensional random effects, full likelihood (FL) approach is proposed for simultaneous model selection and parameter estimation. For GLMMs with high-dimensional random effects, penalized quasi-likelihood (PQL) approach is developed. By this approach, model selection in GLMMs is able to proceed in the framework of linear mixed models. Since the PQL approach produces biased parameter estimates for sparse binary longitudinal data, two-stage penalized quasi-likelihood approach (TPQL) is proposed to take care of the bias issue in PQL. In other words, we use the PQL approach to do model selection at the first stage, and standard estimation techniques such as the maximum likelihood method or the restricted maximum likelihood method to do parameter estimation at the second stage. Marginal approach based on approximate marginal likelihood (AML) for binary data with special link functions is also developed. Robust standard error estimators of the fitted parameters are derived based on a sandwich formula. A bias correction is proposed to improve the estimation accuracy of the PQL approach for binary data.

The sampling performance of the proposed procedures is evaluated through exten-

sive simulation and their applications to real data analysis. In terms of model selection, all of them perform closely. As for parameter estimation, FL, AML, and TPQL yield similar results. Compared with FL, the other procedures substantially reduce computational load.

Variable Selection Procedures for Generalized Linear Mixed Models in Longitudinal Data Analysis

by

Hongmei Yang

a dissertation submitted to the graduate faculty of
north carolina state university
in partial fulfillment of the
requirements for the degree of
doctor of philosophy

statistics

raleigh, north carolina
August 2007

approved by:

Dr. Daowen Zhang (Chair)	Dr. Hao Helen Zhang (Co-Chair)
Dr. Dennis Boos	Dr. Marie Davidian

Dedication

To my parents Fazu Yang and Defang Liang, my husband Shian, my babies Sophie, Rick and Keefer, my sister Hongling and my brother Qinghua.

Biography

Hongmei was born in Jiangling, Hubei, People's Republic of China in December of 1975. She got B.S in Mathematics Education at Central China Normal University in 1997. She worked as a Math Teacher at a private school in Haikou for two years. In 2000 she joined an international company Beijing office as a Manager Assistant.

After leaving campus for couple of years, she decided that it was important to return to school and learn cutting-edge knowledge. In 2001 she came to United States for the advanced education in Statistics. She got M.S in Statistics at University of Toledo in 2003. After that, she joined the department of Statistics at NC State University to continue her studies in Statistics. She is expected to receive a Ph.D degree in Statistics from NC State University in 2007.

Acknowledgements

I would like to thank my advisor, Dr. Daowen Zhang, for being an excellent mentor. Thanks to him, I have learned many research methods, gained a wide range of communication skills and scientific attitudes, as well as the virtues of patience and integrity. In each weekly meeting, he brings with him an insightful vision, a sharp mind, and a dedicated attitude. He is not only a great teacher, but a wonderful advisor. Without his help and advising, I can not imagine completing my dissertation. I will be in his debt forever.

I would like to give my deepest appreciation to my co-advisor, Dr. Hao Helen Zhang, for her great ideas, wisdom, endless knowledge, excellent guidance and great patience throughout my work. She is very supportive and considerate, a great person to work with. I feel very lucky to have her to be one of my advisors.

My thanks go to my committee members, Dr. Dennis Boos and Dr. Marie Davidian, for their helpful comments and great teaching during my years at NCSU.

Study at this department was a great experience to me. I would like to thank all of the faculty for their teaching and inspiration. In particular, Dr. Swallow deserves a special thanks for his understanding and support during my difficult time.

Thank you to all my fellow graduate students, especially Liqiu Jiang, Lan Lan, Shufang Liu, Lihua Tang, Jiezhun Gu, Justin Shows, Eun Hye Lee, Cristina, Matthew, Amy Nail, John, Joe Boyer, for all your help and the great experience at NCSU.

Special thanks goes to my parents, Fazu Yang and Defang Liang. They are the greatest parents in the world. Without their unreserved support and love, my success is impossible.

I would also like to thank my old sister and old brother, Hongling and Qinghua. Whatever happens, I know they are always beside me, support me and help me. My niece Wang Yang deserves special thanks—You are the best, and I am proud of you! My thanks also goes to my mother-in-law, Manyu Lin, and my sister-in-law, for your

support and encouragement.

Last but not least I would like to give my gratitude to my husband, Shian, for his love and caring; and to my lovely babies, Sophie, Rick and Keefer — you bring endless joy to my life!

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction and Background

1.1 Longitudinal Data Analysis

In longitudinal studies, data are commonly correlated, because measurements are often taken repeatedly over time on the same subject. For example, in the Multicenter AIDS Cohort Study (MACS) (Kaslow *et al.*, 1987), the number of measurements on each subject has a range of 1 to 14, and the average number of observations per unit is 6. This correlation in the data has to be taken into account for a valid inference. For longitudinal data with continuous responses, such as the CD4 counts in the MACS data, the common practice is to use linear mixed models, which take account of the correlation between observations by using random effects. In linear mixed models, fixed effects and random effects are linearly modelled.

However, in longitudinal studies, the outcomes of interest are often discrete or can not be adequately described by a normal distribution. For example, in many medical studies, the researchers' primary interest is the presence of diseases, so the response will be binary as YES or NO. An example of this type of data is the longitudinal study on respiratory infection in Indonesian children, where 275 preschool children were examined every three months for respiratory infection ($0 \equiv no$; $1 \equiv yes$), and they were followed for up to six consecutive quarters.

For longitudinal data with non-normal response, the conventional techniques are generalized estimation equations (GEE) for marginal modeling and generalized linear mixed models (GLMM) for subject-specific modelling.

1.1.1 Generalized Estimation Equations (GEEs)

One of the common perspectives in analyzing longitudinal data is population-averaged. This approach models the mean response across the population of subjects at each time point as a function of covariates. The associated fitting method is GEE. The implementation of GEE usually requires the specification of a correct mean response model and a working covariance matrix model. Multivariate normal distribution is fully represented by the mean and the covariance matrix. With the mean response model and the covariance model specified, it is enough to describe the distribution of a continuous response with normal distribution. However, this does not apply to non-normally distributed response. With only a specified mean response model and a covariance model, it is not possible to appeal to the principle of likelihood for estimation and inference.

The GEE fitting method is not a likelihood method. As a reason, it is difficult to derive quantities like AIC or BIC for comparison of different assumptions about correlation matrix. The performance of GEE is to some extent dependent on the validity of the assumption about the correlation matrix. If the assumption about the working correlation matrix is incorrect, efficiency may be lost, but, under reasonably general conditions, consistency can be retained. However, for longitudinal data with missing values (dropouts), GEE can be applied under the assumption that the dropouts

are completely at random, otherwise the consistency of the estimating equation is lost. Robins *et al.* (1995) extended the GEE method to longitudinal data with random dropouts, which preserved the property of consistent inference on the population mean response without requiring correct specification of variance structure. Unfortunately, this extension requires that the dropout probabilities could be estimated consistently for each subject given their observed measurement history and any relevant covariate, which may be too difficult in practice.

Though GEE is not a likelihood based method, it is possible to use it for model selection with shrinkage penalties, which are commonly likelihood based. Fu (2003) developed the penalized GEE by applying the Bridge penalty model to the GEE in longitudinal studies. They overcome the lack of a joint likelihood in GEE by using the penalized estimation equations.

In consideration of practical limits of GEE mentioned above, we propose three likelihood-based model selection approaches in the analysis of high-dimensional, non-normal longitudinal data. However, we also develop a penalized GEE version of model selection approach for a special GLMM for binary longitudinal data. The details follow this chapter.

1.1.2 Generalized Linear Mixed Models (GLMMs)

The subject-specific approach is another common perspective in longitudinal data analysis. This approach models individual unit trajectory instead of the mean response by incorporating random effects, because in biomedical studies, researchers may be more interested in the individual trend of the response variable as a function of time. As a

consequence, for continuous normally distributed response, we use linear mixed models; otherwise we use GLMMs. Generalized linear models (McCullagh and Nelder, 1989; Nelder and Wedderburn, 1972) are widely used in data sets where observations are assumed to be independent and response variables are discrete, such as binary and count data. However, as we mentioned earlier, in longitudinal studies, observations are commonly correlated due to multiple measurements on each subject. The correlation between observations has to be taken into account to yield a valid inference. As an extension of generalized linear models, GLMMs model covariate effects linearly under monotonic continuous link functions, and account for correlation and over-dispersion by incorporating random effects to the linear predictors.

Specifically, suppose a sample consists of m subjects with observations on subject i being (Y_i, X_i, Z_i), where $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})^T$ are the outcomes, and $X_i = (X_{i1}, X_{i2}, \ldots, X_{in_i})^T$ and $Z_i = (Z_{i1}, Z_{i2}, \ldots, Z_{in_i})^T$ are covariates associated with $d \times 1$ fixed effects β and $q \times 1$ random effects b_i respectively. Given b_i, the responses Y_{ij} on subject i are assumed to be conditionally independent with means $E(Y_{ij}|b_i) = \mu_{ij}^b$ and variances $Var(Y_{ij}|b_i) = \phi w_{ij}^{-1} \nu(\mu_{ij}^b)$. Here $\nu(\cdot)$ is a variance function which depends on the specified conditional distribution of Y_{ij}, w_{ij} is a prior weight (e.g. a binomial denominator), and ϕ is a scale parameter that may or may not be known. The conditional means μ_{ij}^b are related to the linear predictors $\eta_{ij}^b = X_{ij}^T \beta + Z_{ij}^T b_i$ by a link function g: $g(\mu_{ij}^b) = \eta_{ij}^b$, where $g(\cdot)$ is monotonic differentiable.

Define $Y = (Y_1^T, \ldots, Y_m^T$. Similarly define X, μ^b, $g(\mu^b)$ and w. Let

$$Z = \begin{pmatrix} Z_1 & 0 & \cdots & 0 \\ 0 & Z_2 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & Z_m \end{pmatrix}. \tag{1.1}$$

Then a GLMM can be written in matrix notation as

$$g(\mu^b) = X\beta + Zb, \tag{1.2}$$

where $b = (b_1^T, \ldots, b_m^T)^T$. The random effects b_i are assumed to be independent normal variables with mean zero and covariance $D(\theta)$, where D is a positive definite matrix depending on a parameter vector θ. In the examples we consider, the dispersion parameter ϕ is fixed at unity. In the cases where ϕ is unknown, it may be estimated together with θ as a parameter in the covariance matrix of the marginal distribution of outcome variables.

The family of linear mixed models with normal distribution is a special type of GLMMs, with the link function g being the identity link.

Under the assumption of normality about b_i $(i = 1, 2, \ldots, m)$, the integrated log quasi-likelihood function of β and θ is

$$e^{\ell(\beta,\theta;Y)} \propto |G|^{-\frac{1}{2}} \int \exp\left\{ -\frac{1}{2\phi} \sum_{i=1}^{m} \sum_{j=1}^{n_i} d_{ij}(Y_{ij}; \mu_{ij}^b) - \frac{1}{2} b^T G^{-1} b \right\} db, \tag{1.3}$$

where $G = \text{diag}\{D, \ldots, D\}$ and

$$d_{ij}(Y_{ij}; \mu_{ij}^b) = -2 \int_{Y_{ij}}^{\mu_{ij}^b} \frac{w_{ij}(Y_{ij} - u)}{\nu(u)} du$$

represent the conditional deviance function of (β, θ) given b. If the conditional distribution of the observations given b belongs to a linear exponential family, $\ell(\beta, \theta; Y)$ represents the true log-likelihood from the data.

The evaluation of the quasi-likelihood $\ell(\beta, \theta; Y)$ involves integration over the distribution of the random effects. Because the random effects enter the model nonlinearly, the integration is often complicated and even intractable. As a consequence, the full likelihood of the data may not have a closed form, and the full likelihood inference may not be feasible. For GLMMs with low-dimensional random effects such as single source of variation (Laird 1978), the exact likelihood analysis can be implemented by using standard numerical methods such as Gaussian quadrature. However, when the dimension of random effects is high, an exact likelihood analysis is not possible due to the intractable integration. Consequently, many approximate inference procedures have been proposed to avoid the numerical difficulty.

Schall (1991) proposed to approximate the link function $g(\cdot)$ by the first-order Taylor expansion at the conditional mean, so that the working vector could be obtained (McCullagh and Nelder, 1989). The use of the working vector reduces the problem to a working linear mixed model. However, the approximation based on the first-order expansion may be very poor, especially for sparse data. Liu and Pierce (1993) and Solomon and Cox (1992) used the Laplace approximation to the integration (Barndorff-Nielsen and Cox, 1989) involved in the evaluation of the likelihood.

Breslow and Clayton (1993) used modified Laplace approximation and proposed the penalized quasi-likelihood (PQL) approach, and suggested the estimation in GLMMs can be easily implemented by repeated calls of linear mixed model theories. However, the PQL approach will produce biased estimators for binary longitudinal data. So Breslow and Lin (1995) considered bias correction for parameter estimators from the PQL approach in GLMMs with a single source of variation. Further, Lin and Breslow (1996) developed bias correction factors to variance components and coefficients for GLMMs with multiple sources of variation. Lin and Zhang (1999) cast the estimation and inference in generalized additive mixed models (GAMMs) through double penalized quasi-likelihood (DPQL). By representing a GAMM as a working GLMM, they estimated smoothing parameters and variance components simultaneously by treating the smoothing parameters as an extra part of variance components. Zhang (2004) further applied DPQL to GLMMs with varying coefficients, and suggested that the varying coefficients and random effects could be estimated in the framework of linear mixed models.

Active research on this issue also includes Bayesian methods. Nelder (1972) pointed out that there was a strong connection between the random effects and Bayesian regression models. Zeger and Karim (1991) exploited the relationship through a Monte Carlo method, the Gibbs sampler (Gelfand and Smith, 1990; Geman and Geman, 1984), to overcome the difficulties involved in the integration. McCulloch (1997) proposed maximum likelihood algorithms in GLMMs by constructing a Monte Carlo version of the EM algorithm using importance sampling ideas.

Other types of approximations are available in Jiang (1999), Lee and Nelder (1996).

1.2 Classical Variable Selection Methods

In practice, many possible explanatory variables may be included in models at the initial stage to reduce possible model biases, and covariates are usually high-dimensional. However, many of these covariates may not have any contribution to the explanation of the response. For example, in the study of respiratory infection, there were seven covariates in consideration: age, xerophthalmia status, cosine and sine terms for the annual cycle, sex, height and the presence of stunting, but only a few of them may have effect on respiratory infection. Hence model selection is critical for an accurate analysis and model interpretation. By removing the non-informative variables, we can improve the predictability of models and parsimoniously describe the relationship between the outcome and predictive variables. In the literature, many variable selection methods were proposed. But most of them are for linear models with independent data. Below is a general review of classical model selection methods, followed by a summary of modern approaches.

1.2.1 Best Subset Variable Selection

One of the traditional variable selection methods is the best subset variable selection. There are four types of subset selection methods: exhaustive search, forward selection, backward elimination and stepwise selection.

- Exhaustive search

  As its name implies, this method searches all possible subsets and selects the one with the best evaluation criterion. It is the only technique that is guaranteed to find the best subset using a given criterion, and is therefore

ideal when the number of predictive variables d is small. However, when the number of covariates d is large, exhaustive search is very time-consuming and not practically useful.

- Forward selection

  Forward selection starts with an empty set, adds one predictor each time, and finds the subset with the best criterion value. When covariates are statistically independent, forward selection performs very well for linear models (Miller 1990). But the condition of independence between predictive variables is too restrictive in reality. Furthermore, in practice, the relationships between dependent variables and covariates are often not linear.

- Backward elimination

  Backward elimination starts with the full model, and eliminates one variable at a time. When predictive variables are statistically dependent, backward elimination is preferred to forward selection. However, it will be very sensitive to small changes on the design matrix when the ratio of the total number of observations to the number of predictive variables is small.

- Stepwise selection

  Stepwise selection allows movements in either direction, dropping or adding variables at different steps. The process is one of alternations between choosing the least significant variable to drop and then re-considering all variables including those variables dropped previously (except the one that was dropped most recently) for re-introduction into the model. Problems associated with this method include being time-consuming, unable to handle collinearity and producing biased estimators.

Though best subset variable selection methods are practically convenient, they ignore the stochastic error inherent in the model selection process, and the inference based on them is hard to understand. Another drawback of these methods is the lack of stability (Breiman, 1996).

1.3 Modern Variable Selection Approaches

1.3.1 Shrinkage Penalty Model Selection

An important family of modern model selection methods is based on shrinkage penalties, where penalty functions are added to the residual sum of squares or subtracted from the log-likelihood, and minimization or maximization of penalized functions with respect to coefficients will yield penalized likelihood estimators. Under these methods, non-informative variables are removed from models by shrinking their coefficients all the way to zero, and important variables will be kept in the model with little or no shrinkage. Compared to other variable selection methods, shrinkage penalty methods have the advantage of selecting variables and estimating these coefficients simultaneously.

Well-known methods from this family are ridge regression (Hoerl and Kennard, 1970a, b) with L_2 penalty $p_\lambda(|\omega|) = \lambda|\omega|^2$, bridge regression (Frank and Friedman, 1993) with L_q penalty $p_\lambda(|\omega|) = \lambda|\omega|^q$ for $q \geq 0$, and LASSO (Tibshirani, 1996 and 1997) with L_1 penalty $p_\lambda(|\omega|) = \lambda|\omega|$. Knight and Fu (2000) studied the asymptotic properties for lasso-type estimators. Efron *et al.* (2004) proposed to do model selection by the least angle regression (LARS), and showed that a modification of the LARS

algorithm can implement the LASSO and the forward stage-wise regression.

Fan and Li (2001) developed a new penalty function named smoothly clipped absolute deviation (SCAD) penalty for variable selection in linear models. They argued that, a good penalty function should result in an unbiased (to avoid unnecessary bias for non-zero parameters), sparse (non-informative variable coefficients are estimated as zeros) and continuous (to keep models stable) estimator. Above penalty functions such as ridge, bridge and LASSO do not satisfy the mathematical conditions for unbiasedness, sparsity and continuity at the same time. The SCAD penalty is symmetric, continuous on $(0, \infty)$ and singular at the origin, meeting all of the preceding properties. In the framework of linear or generalized linear models, Fan and Li (2001) showed that under certain regular conditions, the SCAD penalized estimators perform as well as the oracle procedure; in other words, zero coefficients are estimated as zero with probability tending to 1, and nonzero coefficients are estimated as well as if the corrected model were known. Fan and Li (2002) extended the SCAD penalty to the Cox proportional hazards model, and applied it to semi-parametric modelling for longitudinal data (2004), using the independent working correlation matrix.

1.3.2 Smoothly Clipped Absolute Deviation Penalty

A SCAD penalty function can be expressed as:

$$p_\lambda(|\omega|) = \lambda \begin{cases} |\omega| & |\omega| \leq \lambda, \\ -\frac{1}{2(a-1)\lambda}(|\omega|^2 - 2a\lambda|\omega| + \lambda^2) & \lambda < |\omega| \leq a\lambda, \\ \frac{1}{2}(a+1)\lambda & |\omega| > a\lambda, \end{cases} \tag{1.4}$$

where λ and a are two tuning parameters to be determined. Figure 1.1 gives a graph

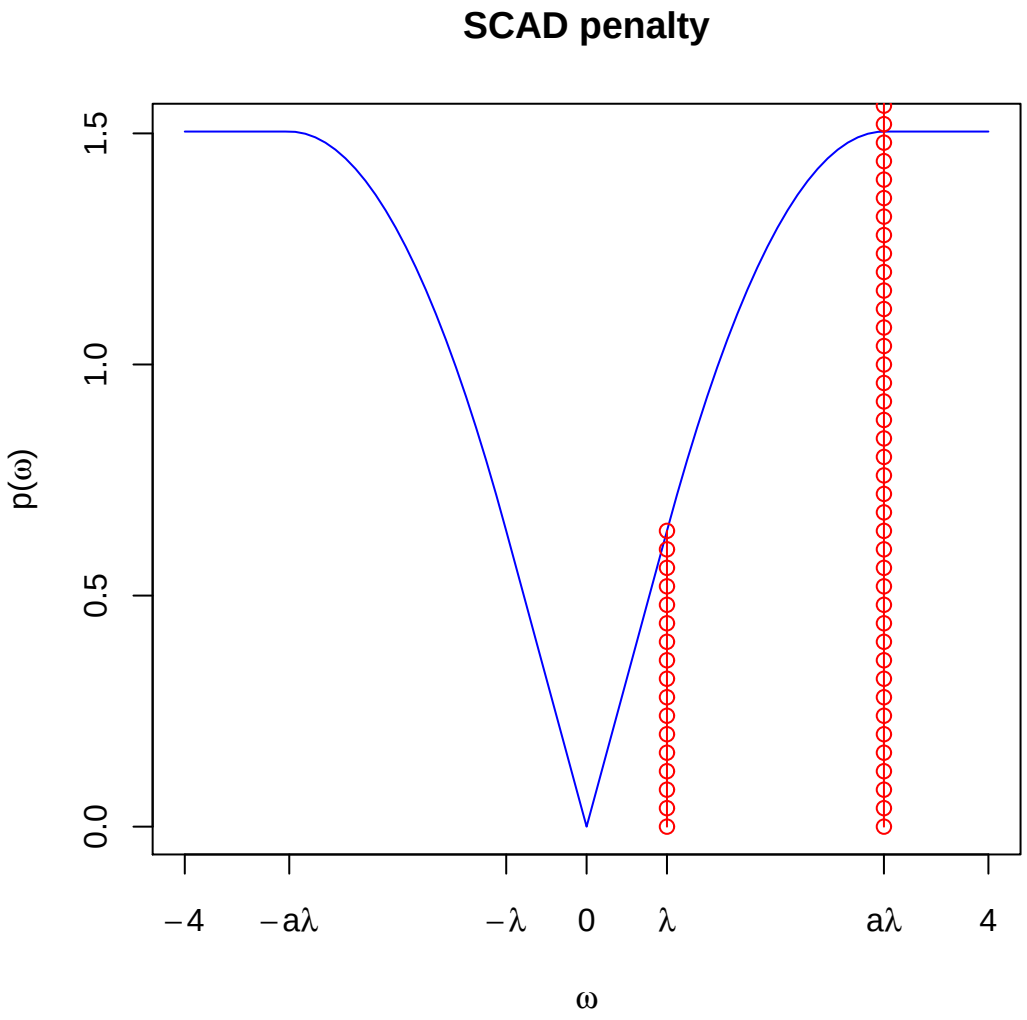

Figure 1.1: SCAD penalty with $a = 3.7$ and $\lambda = 0.8$

representation of the SCAD penalty with $a = 3.7$ and $\lambda = 0.8$. One can see that the SCAD penalty is differentiable except at the origin and can be regarded as a quadratic

spline function with knots of λ and aλ for $\omega \geq 0$. In fact, the SCAD penalty has the features of symmetry, continuity on $(0, \infty)$ and singularity at the origin, and it satisfies all the requirements for unbiasedness, sparsity and continuity.

The first derivative of the SCAD penalty function $p_\lambda(|\omega|)$ with respect to $|\omega|$ is

$$p'_\lambda(|\omega|) = \begin{cases} \lambda & |\omega| \leq \lambda, \\ \frac{a\lambda - |\omega|}{a-1} & \lambda < |\omega| \leq a\lambda, \\ 0 & |\omega| > a\lambda. \end{cases} \tag{1.5}$$

Its graph is depicted in Figure 1.2.

In knowledge of the good properties of the SCAD penalty, we will use it for model selection and parameter estimation in the context of GLMMs.

There are two tuning parameters, λ and a, involved in the SCAD penalty function. In practice, they can be searched over two-dimensional grids based on criteria such as cross-validation (CV) or generalized cross-validation (GCV). These procedures can be computationally expensive. Fan and Li (2001) demonstrated that $a = 3.7$ works well in practice for most models. Thereof, we set $a = 3.7$ throughout our study. We will choose the other tuning parameter λ in general by BIC (Bayesian information criteria, Schwarz 1978), though GCV and REML are also developed.

1.3.3 Other Approaches

Other modern methods for model selection include the Bayesian approaches. Equipped with a natural measurement of uncertainty with the posterior probability, Bayesian variable selection methods generally choose models with the highest posterior proba-

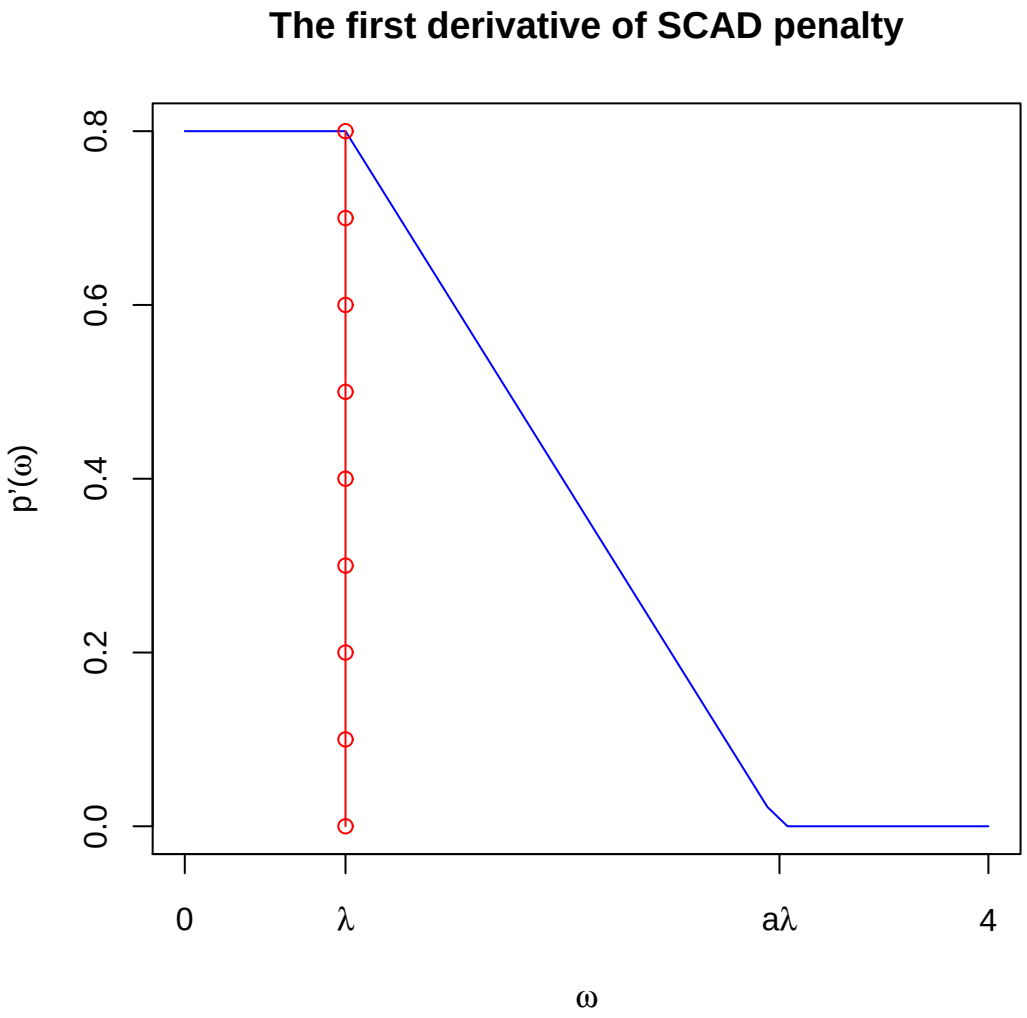

Figure 1.2: The first derivative of SCAD penalty with $a = 3.7$ and $\lambda = 0.8$ for $\omega > 0$

bilities. The choice of prior is vital to the performance of a Bayesian procedure, so lots of Bayesian research on model selection is focused on how to specify good priors. See Mitchell and Beauchamp (1988), Yuan and Lin (2005), and so on.

Luo, Stefanski and Boos (2006) provided an interesting variable selection method by adding additional noise variables. Wu, Boos and Stefanski (2006) proposed to add a number of pseudo-variables to the real data set and to do variable selection through monitoring the falsely selected pseudo-variables. These procedures have shown effective performance in various situations.

1.4 Proposal of Work

Model selection is important for longitudinal data analysis, but very challenging due to data correlation and model complexity. Therefore, very few methods have been developed for correlated data.

In this dissertation, we extend the shrinkage penalty approach using the SCAD penalty to variable selection in GLMMs for longitudinal data. First we develop a unified algorithm based on full likelihood (FL). Gaussian quadrature is used for the evaluation of the likelihood function. Then we use a local quadratic approximation algorithm to iteratively estimate coefficients and variance components. Due to the intensive computation involved in the FL approach, a penalized quasi-likelihood (PQL) approach is proposed for simultaneous model selection and parameter estimation, especially for non-sparse longitudinal data such as count data and binomial data with moderate to large binomial denominators. What makes the PQL method attractive is the gain in computational efficiency by using the Laplace approximation to avoid the

cumbersome integration. Furthermore, by the PQL method, a GLMM can be represented by a working linear mixed model, so that linear mixed model theories can be used to estimate parameters and make inferences. For the binary correlated data, it is known that the PQL method often produces biased parameter estimators (Breslow and Lin, 1995; Lin and Breslow, 1996). We then propose to correct the bias of the estimated variance components and model coefficients. Third, a two-stage penalized quasi-likelihood (TPQL) method is developed to take care of the bias issue of the PQL approach. Essentially, we use the PQL method to do model selection at the first stage, and apply standard estimation techniques of ML or REML based on the selected variables to do parameter estimation at the second stage. In the last, for binary data with special types of link functions, the approximate marginal likelihood (AML) approach is proposed from a marginal point of view. In particular, for this type of data, based on the approximate linear relationship between coefficients of the GLMM and those of its corresponding marginal model, we do variable selection based on the marginal model, which is approximately equivalent to variable selection in the original model, and then use ML or REML for parameter estimation. The two-stage feature makes the TPQL and AML gain both computational efficiency and estimation accuracy. Robust estimators of standard errors are derived using the sandwich formula and tested through simulations for both PQL and FL. Numerical results are presented for sparse binary data, which is a very extreme case. In general, if one procedure works well for binary data, it often performs even better for other types of non-Gaussian data.

We did extensive simulation studies to evaluate the performance of our proposed procedures, and illustrated them with real data analysis. For each scenario we consider, we calculate the corresponding coefficient of determination R^2, which represents the

proportion of variation explained by the true model, to describe the difficulty levels of the numerical examples. We show that, even under very challenging cases (where R^2's are very small), the proposed procedures perform quite well in terms of model selection.

The following chapters are organized as follows. In Chapter 2, we introduce the FL approach, and use local quadratic approximation to iteratively solve for both coefficients and variance components. In Chapter 3, we propose the PQL approach for model selection in GLMMs, and describe the bias correction. Then we show that the bias problem of the PQL approach can be alternatively treated by a different approach: two-stage PQL (TPQL). Chapter 4 deals with variable selection in generalized linear mixed models using the marginal approach. Simulation results are summarized in Chapter 5. In Chapter 6 we present the extension of the PQL to model selection in generalized semi-parametric mixed models (GSMMs). Final discussion is given in Chapter 7. The proof of equivalence between two methods for the estimation of coefficients by the PQL and some tables are included in the Appendixes.

Chapter 2

Full Likelihood Approach

2.1 Motivation

Because we do model selection in GLMMs (1.2) using the shrinkage penalty method, the penalized log-likelihood ℓ_p can be obtained by subtracting the penalty function from the quasi-likelihood $\ell(\beta, \theta; Y)$ (defined in 1.3),

$$\ell_p(\beta, \theta; Y) = \ell(\beta, \theta; Y) - n \sum_{j=1}^{d} p_\lambda(|\beta_j|), \tag{2.1}$$

where $p_\lambda(\cdot)$ is the SCAD penalty function defined in (1.4) and λ is the tuning parameter. For convenience we use $\gamma = (\beta^T, \theta^T)^T$ to denote the parameter vector.

Notice that we do not impose any penalty on the variance component θ.

It is difficult to obtain an exact solution for parameter estimation in (2.1), because the evaluation of $\ell_p(\gamma; Y)$ involves integration, which is often intractable. However, when the involved random effects are low-dimensional, such as random intercept only, the integration can be approximated well by the standard numerical method of Gaussian quadrature. In other words, within the domain of the integration, we can use a weighted sum of the joint density at specified points as an approximation of the definite integral in the form of $\ell(\gamma; Y)$. An c-point Gaussian quadrature rule can yield an exact result for polynomials of degree $2c - 1$, with a suitable use of the c quadratures and

the corresponding weights.

In this section, we propose the use of the full likelihood (FL) approach to simultaneously do variable selection and parameter estimation in GLMMs, and introduce a unified algorithm for the joint estimation of coefficients and variance components. The involvements of integration and variance components bring more challenges to model selection and parameter estimation in GLMMs.

2.2 Model Formulation

As aforementioned, for GLMMs with low-dimensional random effects, Gaussian quadrature can be applied for the evaluation of the quasi-likelihood. Specifically, equation (1.3) can be written in the form of $c|D|^{-m/2}\int \exp\left\{-k(b)-\frac{1}{2}b^TG^{-1}b\right\}db$, with

$$k(b)=\frac{1}{2\phi}\sum_{i=1}^{m}\sum_{j=1}^{n_i}d_{ij}(Y_{ij};\mu_{ij}^b). \tag{2.2}$$

Let $k(b_i)=\frac{1}{2\phi}\sum_{j=1}^{n_i}d_{ij}(Y_{ij};\mu_{ij}^b)$. Then $k(b)=\sum_{i=1}^{m}k(b_i)$, and the log quasi-likelihood in equation (1.3) is equivalent to

$$e^{\ell(\gamma;Y)}=c\prod_{i=1}^{m}l(\gamma;Y_i)=c\prod_{i=1}^{m}\left\{|D|^{-1/2}\int\exp\left[-k(b_i)-\frac{1}{2}b_i^TD^{-1}b_i\right]db_i\right\}. \tag{2.3}$$

For subject i, its likelihood $l(\gamma;Y_i)$ can be approximated using Gaussian quadrature as

$$l(\gamma;Y_i)\approx\sum_{i_q=1}^{c_q}\cdots\sum_{i_1=1}^{c_1}\omega_{i_1\cdots i_q}\exp\left\{-k(D^{\frac{1}{2}}\zeta_{i_1\cdots i_q})\right\}, \tag{2.4}$$

where $\omega_{i_1\cdots i_q}$ and $q \times 1$ dimensional $\zeta_{i_1\cdots i_q}$ are chosen weights and quadrature points respectively for any function $f(\zeta)$ where ζ is $q \times 1$ and multivariate standard normally distributed.

The maximization of the approximation function of the penalized likelihood $\ell_p(\gamma; Y)$ is challenging, because it not only involves variance component θ, but the local behavior of the penalty function around the origin is irregular. However, the penalty function can be locally approximated by a quadratic function (Fan and Li, 2001), which will facilitate the computation. Specifically, given an initial value $\widehat{\gamma}^{(0)} = \left((\widehat{\beta}^{(0)})^T, (\widehat{\theta}^{(0)})^T\right)^T$ which is close to the maximizer of (2.1), if $|\widehat{\beta}_j^{(0)}| > \xi$ with ξ being a predetermined value, then $[p_\lambda(|\beta_j|)]'$ can be approximated by

$$[p_\lambda(|\beta_j|)]' = p'_{\lambda_j}(|\beta_j|)sgn(\beta_j) \approx \{p'_\lambda(|\widehat{\beta}_j^{(0)}|)/|\widehat{\beta}_j^{(0)}|\}\beta_j.$$

Taking integration of the above equation w.r.t β_j, the SCAD penalty can be approximated by a local quadratic function

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\widehat{\beta}_j^{(0)}|) + \frac{1}{2}\frac{p'_\lambda(|\widehat{\beta}_j^{(0)}|)}{|\widehat{\beta}_j^{(0)}|}(\beta_j^2 - (\widehat{\beta}_j^{(0)})^2) \quad \text{for } \beta_j \approx \widehat{\beta}_j^{(0)}. \tag{2.5}$$

The quasi-likelihood $\ell(\gamma; Y)$ is smooth, and its first and second derivative with respect to γ are continuous. So $\ell_p(\gamma; Y)$ can also be approximated by the following quadratic function up to a constant,

$$\begin{aligned} \ell_p(\gamma; Y) \;\approx\; & \ell(\widehat{\gamma}^{(0)}; Y) + \nabla\ell(\widehat{\gamma}^{(0)}; Y)^T(\gamma - \widehat{\gamma}^{(0)}) \\ & + \frac{1}{2}(\gamma - \widehat{\gamma}^{(0)})^T\nabla^2\ell(\widehat{\gamma}^{(0)}; Y)(\gamma - \widehat{\gamma}^{(0)}) - \frac{1}{2}\gamma^T n\Sigma_\lambda(\widehat{\gamma}^{(0)})\gamma, \end{aligned} \tag{2.6}$$

where

$$\nabla \ell(\widehat{\gamma}^{(0)}; Y) = \begin{pmatrix} \frac{\partial \ell(\hat{\gamma}^{(0)}; Y)}{\partial \beta} \\ \frac{\partial \ell(\hat{\gamma}^{(0)}; Y)}{\partial \theta} \end{pmatrix}, \tag{2.7}$$

$$\nabla^2 \ell(\widehat{\gamma}^{(0)}) = \begin{pmatrix} \frac{\partial^2 \ell(\hat{\gamma}^{(0)}; Y)}{\partial \beta \partial \beta^T} & \frac{\partial^2 \ell(\hat{\gamma}^{(0)}; Y)}{\partial \beta \partial \theta^T} \\ \frac{\partial^2 \ell(\hat{\gamma}^{(0)}; Y)}{\partial \theta \partial \beta^T} & \frac{\partial^2 \ell(\hat{\gamma}^{(0)}; Y)}{\partial \theta \partial \theta^T} \end{pmatrix}, \tag{2.8}$$

and

$$n\Sigma_\lambda(\widehat{\gamma}^{(0)}) = \begin{pmatrix} \frac{np'_\lambda(|\hat{\beta}_1^{(0)}|)}{|\hat{\beta}_1^{(0)}|} & 0 & \cdots & 0 & 0 \\ 0 & \frac{np'_\lambda(|\hat{\beta}_2^{(0)}|)}{|\hat{\beta}_2^{(0)}|} & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & \frac{np'_\lambda(|\hat{\beta}_d^{(0)}|)}{|\hat{\beta}_d^{(0)}|} & 0 \\ 0 & \cdots & 0 & 0 & 0 \end{pmatrix}. \tag{2.9}$$

With this local quadratic approximation, given λ, the maximization of (2.1) with respect to γ can be regarded as an iterative ridge regression algorithm, so that γ can be updated as,

$$\widehat{\gamma}^{(1)} = \widehat{\gamma}^{(0)} + \left\{ -\nabla^2 \ell(\widehat{\gamma}^{(0)}; Y) + n\Sigma_\lambda(\widehat{\gamma}^{(0)}) \right\}^{-1} \left\{ \nabla \ell(\widehat{\gamma}^{(0)}; Y) - n\Sigma_\lambda(\widehat{\gamma}^{(0)}) \widehat{\gamma}^{(0)} \right\}. \tag{2.10}$$

Using this iterative ridge regression, with a good initial value $\widehat{\gamma}^{(0)}$, parameter vector γ,

which contains both coefficient β and variance component θ, can be iteratively solved. In practice, $\widehat{\gamma}^{(0)}$ can be obtained by fitting a full model. Now regard $\widehat{\gamma}^{(k-1)}$ as a good initial value at kth step. For the components of $\widehat{\gamma}^{(k-1)}$, use (1.4) to decide their corresponding penalty, form $n\Sigma_\lambda(\widehat{\gamma}^{(k-1)})$, and use (2.10) to get $\widehat{\gamma}^{(k)}$. If the maximum absolute difference of $\widehat{\gamma}^{(k)}$ from $\widehat{\gamma}^{(k-1)}$ is less than a predetermined value, say 10^{-4}, the iteration will stop and $\widehat{\gamma}^{(k)}$ will be claimed to be the penalized maximum likelihood estimator of γ. Hence, model selection and parameter estimation can be done simultaneously for GLMMs.

As Fan and Li (2001) argued in the linear model setting, with a good initial value $\widehat{\gamma}_0$, the estimator obtained from the above iterative algorithm with a few iterations can be regraded as a one-step estimator, which is as efficient as the fully iterative method.

2.3 Standard Error Formula

With the iterative ridge regression algorithm (2.10), we can estimate parameters and select variables at the same time. Therefore, the standard errors of parameter estimators can be obtained directly by using the conventional technique of the sandwich formula in the likelihood setting. For the estimated parameter $\widehat{\gamma}_1$, which consists of non-vanishing coefficient estimator $\widehat{\beta}_{imp}$ and variance component estimator $\widehat{\theta}$, their covariance estimator by the sandwich formula is

$$\begin{aligned} \widehat{cov}(\widehat{\gamma}_1) &= \left\{-\nabla^2\ell(\widehat{\gamma}_1;Y)+n\Sigma_\lambda(\widehat{\gamma}_1)\right\}^{-1}\widehat{cov}\{\nabla\ell(\widehat{\gamma}_1;Y)\} \\ &\quad \left\{-\nabla^2\ell(\widehat{\gamma}_1;Y)+n\Sigma_\lambda(\widehat{\gamma}_1)\right\}^{-1}. \end{aligned} \tag{2.11}$$

We use $-\nabla^2\ell(\widehat{\gamma}_1; Y)$ to estimate $cov\{\nabla\ell(\widehat{\gamma}_1; Y)\}$.

The performance of the sandwich formula is evaluated through extensive simulations. The results showed this formula works very well for the variance estimator of $\widehat{\beta}_{imp}$, and has a good performance for the standard error estimator of $\widehat{\theta}$ when θ is not very large. Detailed results can be found in Chapter 5.

2.4 Selection of Tuning Parameter λ

The selection of tuning parameters is vital to ensure good performance of the proposed procedures. There are two tuning parameters involved in the SCAD penalty function: a and λ. For a, as aforementioned, Fan and Li (2001) showed by Bayesian approach that 3.7 works well in many data settings, so we set $a = 3.7$ throughout our study.

One traditional criterion for selecting λ is the data-driven method GCV. Due to the complicated structure of GLMMs, it is not easy to derive the GCV in general. For the FL approach, we use BIC to tune λ over a one-dimension grid search, considering that BIC has the tendency to seek the simplest model consistent with the data.

The information criterion BIC, derived from Bayes' theorem by Schwarz (1978), is defined as

$$BIC = -2\ell + d_1 \log n,$$

where ℓ is the log-likelihood at the SCAD penalized estimator $\widehat{\gamma}$, d_1 is the number of important explanatory variables after model selection, and n is the total number of observations. For the FL method, the full log-likelihood $\ell(\widehat{\gamma}; Y)$ is used for BIC

calculation.

λ is the chosen value with the smallest BIC.

2.5 Computational Algorithm of FL

The practical implementation of the FL method involves the following steps:

1. data standardization

   Given data, we first standardize covariates X by their sample means and sample standard deviations.

2. initiation

   Given data, we use the procedure *nlmixed* in *SAS* to get the initial values $\widehat{\beta}^{(0)}$ and $\widehat{\theta}^{(0)}$ for β and θ by fitting a full model.

3. Outer loop

   (a) Set a range for λ, and grid λ into s levels $\lambda_1, \lambda_2, \ldots, \lambda_s$.

   (b) For each λ_i, go to the inner loop to jointly estimate β and θ, compute $BIC(\lambda_i)$, select λ_i that yields smallest BIC.

   (c) Identify the final estimators $\widehat{\beta}$ and $\widehat{\theta}$ which are associated with the selected λ_i.

   (d) Adjust $\widehat{\beta}$ to its original scale.

4. Inner loop

(a) For the $(k+1)$th iteration, let $\widehat{\gamma}^{(k)}$ be the estimator from the kth iteration. We update $\gamma = (\beta^T, \theta^T)^T$ by

$$\widehat{\gamma}^{(k+1)} = \widehat{\gamma}^{(k)} + \left\{-\nabla^2 \ell(\widehat{\gamma}^{(k)}; Y) + n\Sigma_\lambda(\widehat{\gamma}^{(k)})\right\}^{-1} \left\{\nabla \ell(\widehat{\gamma}^{(k)}; Y) - n\Sigma_\lambda(\widehat{\gamma}^{(k)})\widehat{\gamma}^{(k)}\right\}.$$

(b) Compute $\max|\widehat{\gamma}^{(k+1)} - \widehat{\gamma}^{(k)}|$. Compare this value to a predetermined value δ, say $\delta = 10^{-4}$. If it is no larger than δ, then stop the inner loop and return to the outer loop 3.

(c) Repeat above steps until $\widehat{\gamma}$ converges.

2.6 Summary

In this chapter, a unified algorithm was proposed for simultaneous model selection and parameter estimation in GLMMs with low-dimensional random effects. This algorithm is based on the full likelihood penalized with SCAD penalty. A standard error estimator for the estimated parameter was established through the sandwich formula. A criterion was presented for the selection of the tuning parameter. The computational algorithm of the proposed method was elaborated.

Its finite sampling performance will be studies in Chapter 5.

Chapter 3

Penalized Quasi-Likelihood Approach

3.1 Motivation

When the number of random effects is large, the penalized ML estimator of the penalized quasi-likelihood (2.1) is unattainable due to the intractable integration. The Gaussian quadrature works well only for some types of distributions, and may yield poor approximation when the dimension of integrands increases. Therefore, the application of the FL approach is restrictive.

In view of the complicated and often intractable integration, Brewslow and Clayton (1993) proposed to approximate the log quasi-likelihood $\ell(\beta, \theta; Y)$ by the Laplace method, and suggested that parameters could be estimated in the framework of working linear mixed models. We will extend this approximation procedure to model selection in GLMMs, and propose the penalized quasi-likelihood (PQL) approach under the profile pseudo likelihood, which depends on an estimator of θ.

3.2 Penalized Quasi-Likelihood Approach

3.2.1 Double Penalized Quasi-Likelihood

Brewslow and Clayton (1993) and Lin and Zhang (1999) demonstrated that, if Laplace's method is used for integral approximation, and the dependence of the weight matrix

$$W = diag\{w_{ij}/[\phi\nu(\mu_{ij}^b)(g'(\mu_{ij}^b))^2]\} \tag{3.1}$$

on the conditional mean μ_{ij}^b is ignored, it can be shown that, for given θ, $(\widehat{\beta}(\theta), \widehat{b}(\theta))$ jointly maximize the double penalized quasi-likelihood (DPQL)

$$\tilde{\ell}_{dpql}(\beta, \theta; Y) \;\; = \;\; \tilde{\ell}_p(\beta, \theta; Y) - n\sum_{j=1}^{d} p_\lambda(|\beta_j|), \tag{3.2}$$

with

$$\tilde{\ell}_p(\beta, \theta; Y) = -\frac{1}{2\phi}\sum_{i=1}^{m}\sum_{j=1}^{n_i} d_{ij}(Y_{ij}; \mu_{ij}^b) - \frac{1}{2}b^T G^{-1} b. \tag{3.3}$$

$\tilde{\ell}_p$ is the penalized quasi-likelihood for a GLMM, with the penalty for the random effects. With $\tilde{\ell}_p$, it can be shown that model selection in GLMMs can be an estimation problem in working linear mixed models. Detailed development follows this section.

3.2.2 Linear Mixed Model Representation

For the DPQL $\tilde{\ell}_{dpql}$, we approximate it similarly to (2.5) in Chapter 2 with a local quadratic function. Specifically, first we obtain an initial estimator $(\widehat{\beta}^{(0)}, \widehat{\theta}^{(0)})$ by fitting

a full model. Then we fix θ at $\widehat{\theta}^{(0)}$. For given λ, we approximate $\tilde{\ell}_{dpql}$ by the local quadratic function

$$\tilde{\ell}_{dpql}(\beta, \theta; Y) \approx \tilde{\ell}_p(\beta, \theta; Y) - \frac{1}{2}\beta^T n\Sigma_\lambda(\widehat{\beta}^{(0)})\beta. \tag{3.4}$$

Next we maximize $\tilde{\ell}_{dpql}$ with respect to (β, b) by using the Fisher scoring algorithm

$$\begin{pmatrix} X^TWX + n\Sigma_\lambda & X^TWZ \\ Z^TWX & Z^TWZ + G^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ b \end{pmatrix} = \begin{pmatrix} X^TWy \\ Z^TWy \end{pmatrix}. \tag{3.5}$$

Here y is a working vector with an initial value

$$y^{(0)} = X\widehat{\beta}^{(0)} + Z\widehat{b}^{(0)} + \Delta^{(0)}(Y - \mu^{\widehat{b}^{(0)}}), \tag{3.6}$$

Δ =diag$\{g'(\mu_{ij}^b)\}$, and $n\Sigma_\lambda$ =diag$\left\{ \frac{np'_\lambda(|\widehat{\beta}_1^{(0)}|)}{|\widehat{\beta}_1^{(0)}|}, \ldots, \frac{np'_\lambda(|\widehat{\beta}_d^{(0)}|)}{|\widehat{\beta}_d^{(0)}|} \right\}$.

Note the penalty matrix $n\Sigma_\lambda$ is diagonal. We can re-write it as $n\Sigma_\lambda = \begin{pmatrix} 0 & 0 \\ 0 & \Sigma_{22} \end{pmatrix}$, that is, moving all of zero penalty to the left top and non-zero penalty to the right bottom. Accordingly, we divide β and X as $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$ and $X = (X_1, X_2)$, such that β_1 corresponds to the part of coefficients without penalty, β_2 is the remaining part with penalty, and X_1 and X_2 are the covariates associated with β_1 and β_2 respectively.

By this re-ordering, (3.5) can be written as

$$\begin{pmatrix} X_1^T W X_1 & X_1^T W X_2 & X_1^T W Z \\ X_2^T W X_1 & X_2^T W X_2 + \Sigma_{22} & X_2^T W Z \\ Z^T W X_1 & Z^T W X_2 & Z^T W Z + G^{-1} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ b \end{pmatrix} = \begin{pmatrix} X_1^T W y \\ X_2^T W y \\ Z^T W y \end{pmatrix}. \quad (3.7)$$

An examination shows that (3.7) is the normal equation of the BLUPs of β and b under the working linear mixed model

$$y = X_1\beta_1 + X_2\beta_2 + Zb + \epsilon, \quad (3.8)$$

where β_2 and b are random effects, $\beta_2 \sim N(0, \Sigma_{22}^{-1})$, $b \sim N(0, G)$ and $\epsilon \sim N(0, W^{-1})$.

Hence, variable selection and parameter estimation can proceed in the framework of the working linear mixed model (3.8).

3.2.3 Estimation of β

For the working linear mixed model (3.8), the computation for inference is intensive since it involves three random effects and the marginal variance $V^* = X_2\Sigma_{22}^{-1}X_2^T + ZGZ^T + W^{-1}$ has a complicated structure. A feasible but equivalent way to solve for β is based on the linear mixed model representation of GLMMs (1.2) (Brewslow and Clayton, 1993)

$$y = X\beta + Zb + \epsilon, \quad (3.9)$$

with $b \sim N(0, G)$ and $\epsilon \sim N(0, W^{-1})$. The pseudo log-likelihood of the GLMM (1.2) (or the log-likelihood of a linear mixed model (3.9)) is

$$\ell_{pl}(\beta, \theta; y) \;\; = \;\; -\frac{1}{2}\log|V| - \frac{1}{2}(y - X\beta)^T V^{-1}(y - X\beta), \tag{3.10}$$

with $V = ZGZ^T + W^{-1}$.

For given λ and an initial value $(\widehat{\beta}^{(0)}, \widehat{\theta}^{(0)})$, again we approximate the penalized pseudo likelihood

$$\ell_{ppl}(\beta, \theta; y) = \ell_{pl}(\beta, \theta; y) - n\sum_{j=1}^{d} p_\lambda(|\beta_j|) \tag{3.11}$$

with a local quadratic function,

$$\ell_{ppl}(\beta, \theta; y) \approx \ell_{pl}(\beta, \theta; y) - \frac{1}{2}\beta^T n\Sigma_\lambda(\widehat{\beta}^{(0)})\beta, \tag{3.12}$$

with

$$n\Sigma_\lambda(\widehat{\beta}^{(0)}) = diag\left\{\frac{np'_\lambda(|\widehat{\beta}_1^{(0)}|)}{|\widehat{\beta}_1^{(0)}|}, \frac{np'_\lambda(|\widehat{\beta}_2^{(0)}|)}{|\widehat{\beta}_2^{(0)}|}, \ldots, \frac{np'_\lambda(|\widehat{\beta}_d^{(0)}|)}{|\widehat{\beta}_d^{(0)}|}\right\}.$$

The optimization of (3.12) with respect to β yields

$$\widehat{\beta}^{(1)} = \left(X^T V^{-1} X + n\Sigma_\lambda(\widehat{\beta}^{(0)})\right)^{-1} X^T V^{-1} y^{(0)}. \tag{3.13}$$

The random effect b can be jointly predicted using the normal theory of the linear

mixed model (3.9),

$$\widehat{b}^{(1)} = GZ^T V^{-1}(y^{(0)} - X\widehat{\beta}^{(1)}). \tag{3.14}$$

Given a tuning parameter λ and an initial parameter estimator, $\widehat{\beta}$ and $\widehat{b}$ can be iteratively solved from (3.13) and (3.14). In practice, $(\widehat{\beta}^{(0)}, (\widehat{\theta}^{(0)})$ and $\widehat{b}^{(0)}$ can be obtained by fitting a full model. Now regard $\widehat{\beta}^{(k-1)}$ as a good initial value at kth step. For the components of $\widehat{\beta}^{(k-1)}$, use (1.4) to decide their corresponding penalty, form $n\Sigma_\lambda(\widehat{\beta}^{(k-1)})$, and use (3.13) to get $\widehat{\beta}^{(k)}$. If the maximum absolute difference of $\widehat{\beta}^{(k)}$ from $\widehat{\beta}^{(k-1)}$ is less than a predetermined value, say 10^{-4}, the iteration will stop and $\widehat{\beta}^{(k)}$ will be claimed to be the double penalized maximum quasi-likelihood estimator of β. Hence, model selection and parameter estimation can be done simultaneously for GLMMs using DPQL. The proof of the equivalence between (3.13) and (3.5) (or (3.7)) is given in Appendix A.

3.2.4 Estimation of θ

In the framework of linear mixed models, the estimation of variance component θ is conventionally based on the restricted maximum likelihood, which takes into account the lost degree of freedom resulting from the estimation of fixed effects β. The restricted maximum likelihood for the working linear mixed model (3.9) is defined as

$$\ell_{REML}(\beta, \theta; y) = -\frac{1}{2}\log|V| - \frac{1}{2}\log|X^T V^{-1} X| - \frac{1}{2}(y - X\widehat{\beta})^T V^{-1}(y - X\widehat{\beta}). \tag{3.15}$$

Note that the difference between our formula of ℓ_{REML} and the normal definition is that y is a working vector defined as in (3.6), and $\widehat{\beta}$ is an estimator to coefficients of important variables only.

Using the Fisher scoring algorithm, the REML estimator of θ can be iteratively solved using the following equation:

$$\widehat{\theta}^{(1)} \;\; = \;\; \widehat{\theta}^{(0)} + \{I(\widehat{\theta}^{(0)})\}^{-1} S(\widehat{\theta}^{(0)}),$$

where $\widehat{\theta}^{(0)}$ being a starting value, $I(\theta)$ and $S(\theta)$ are the Fisher information matrix and score function for θ respectively,

$$S(\theta_i) = \frac{\partial \ell_{REML}}{\partial \theta_i} = -\frac{1}{2} tr(P\frac{\partial V}{\partial \theta_i}) + \frac{1}{2}(y - X\widehat{\beta})^T V^{-1} \frac{\partial V}{\partial \theta_i} V^{-1}(y - X\widehat{\beta}),$$

$$I_{ij}(\theta) = \frac{\partial^2 \ell_{REML}}{\partial \theta_i \partial \theta_j^T} = \frac{1}{2} tr(P\frac{\partial V}{\partial \theta_i} P \frac{\partial V}{\partial \theta_j}),$$

with

$$P = V^{-1} - V^{-1}X(X^T V^{-1} X)^{-1} X^T V^{-1}.$$

3.2.5 Bias Correction for $\widehat{\theta}$ and $\widehat{\beta}$

In the GLMMs (1.2), the estimators of variance components and regression coefficients obtained by the DPQL may be subject to serious bias when the data are sparse, particularly for binary data. Breslow and Lin (1995) studied the bias associated with

the estimator of θ and suggested a bias correction factor for it. Similarly, we propose to correct the bias of the estimated variance components and regression coefficients in our PQL method.

For the GLMMs (1.2), the corrected estimator of the variance component is

$$\widehat{\theta}_c \;=\; C^{-1}C_p\widehat{\theta}, \tag{3.16}$$

where the correction terms C and C_p are identical to those given in equation (17) of Breslow and Lin (1995), and equation (20) of Lin and Breslow (1996), except that $\widehat{\beta}$ is the DPQL estimator obtained after unimportant variables have been removed.

Note that C is in fact the approximated Fisher information of θ under the true log profile likelihood $\ell^\sharp(\theta) = \log \ell\{\widehat{\beta}(\theta), \theta\}$ for independent data ($\theta = 0$), and C_p is the approximated Fisher information of θ under the approximated log profile likelihood under independence ($\theta = 0$).

For the correlated binary data with the canonical link, the corresponding model is

$$Y_{ij}|b \sim B(1, \mu^b_{ij}), \;\; \log\frac{\mu^b_{ij}}{1-\mu^b_{ij}} = x^T_{ij}\beta + b_i. \tag{3.17}$$

The corresponding correction factors are $C = C_1 - C_2 - C_3$, and $C_p = C_1$ with

$$\begin{aligned}
C_1 &= \sum_{i=1}^{m}\left(-\sum_{j=1}^{n_i}\nu(\mu^0_{ij})\right)^2/2,\\
C_2 &= A^TX(X^TW_0X)^{-1}X^TA/4,\\
C_3 &= \sum_{i=1}^{m}\left(-\sum_{j=1}^{n_i}\left[\nu(\mu^0_{ij})(1-2\mu^0_{ij})^2 - 2\nu^2(\mu^0_{ij})\right]\right)/4,
\end{aligned}$$

where μ_{ij}^0 are the means of outcome variables for independent data ($\theta = 0$). Here W_0 is a diagonal matrix with diagonal elements $w_{ij}\nu(\mu_{ij}^0)/\phi$, A is an $n \times 1$ vector with elements $w_{ij}\nu(\mu_{ij}^0)\nu'(\mu_{ij}^0)/\phi$, and both are evaluated using the DPQL regression coefficient estimators $\widehat{\beta}(\theta)$. The definitions of w_{ij}, $\nu(\cdot)$ and ϕ can be found in section 1.1.2.

After correcting the bias of $\widehat{\theta}$, we will correct $\widehat{\beta}$ by using $\widehat{\theta}_c$ to re-estimate β.

3.2.6 Standard Error Formula

With the iterative algorithm (3.13), the selection of variables and the estimation of the non-vanishing coefficient β_{imp} can proceed simultaneously. Therefore, we can estimate the covariance of $\widehat{\beta}_{imp}$ by the sandwich formula

$$\begin{aligned} \widehat{cov}(\widehat{\beta}_{imp}) &= \left\{-\nabla^2\ell_{pl}(\widehat{\beta}_{imp}(\widehat{\theta}); y) + n\Sigma_\lambda(\widehat{\beta}_{imp})\right\}^{-1} \widehat{cov}\left\{\nabla\ell_{pl}(\widehat{\beta}_{imp}(\widehat{\theta}); y)\right\} \\ &\quad \left\{-\nabla^2\ell_{pl}(\widehat{\beta}_{imp}(\widehat{\theta}); y) + n\Sigma_\lambda(\widehat{\beta}_{imp})\right\}^{-1}, \end{aligned} \tag{3.18}$$

with $-\nabla^2\ell_{pl}(\widehat{\beta}_{imp}(\widehat{\theta}); y) = X_{imp}^T V^{-1} X_{imp}$, and $\widehat{cov}\left\{\nabla\ell_{pl}(\widehat{\beta}_{imp}(\widehat{\theta}); y)\right\} = X_{imp}^T V^{-1} X_{imp}$. Here X_{imp} is the important covariate associated with β_{imp}.

As conventionally, the variance of the variance component θ under linear mixed models is estimated by the inverse of its information matrix

$$\widehat{cov}(\widehat{\theta}) = I(\widehat{\theta})^{-1}. \tag{3.19}$$

Because $\widehat{\theta}_c$ is a linear function of $\widehat{\theta}$ as $C^{-1}C_p\widehat{\theta}$, its variance can be easily obtained by

the Delta method

$$\widehat{cov}(\widehat{\theta}_c) \quad = \quad (C^{-1}C_p)\widehat{cov}(\widehat{\theta})(C^{-1}C_p)^T. \tag{3.20}$$

Simulation studies showed that both the sandwich formula for $\widehat{cov}(\widehat{\beta})$ and the working covariance estimator for $\widehat{\theta}$ have good performance under regular conditions.

3.3 Two-stage Penalized Quasi-Likelihood Approach

Our experience shows that, for non-sparse longitudinal data such as count data and binomial data with moderate to large binomial denominators, the PQL approach has good performance in selecting covariates and estimating regression parameters. However, for sparse binary longitudinal data, the PQL approach tends to produce bias in estimates (Breslow and Lin, 1995; Lin and Breslow, 1996; Zhang and Lin, 1999), though it still works well in terms of variable selection. Therefore, we propose the two-stage penalized quasi-likelihood (TPQL) approach to keep the good property of PQL in model selection, and also to get better parameter estimators. In other words, at the first stage , we use the PQL for model selection; at the second stage, with the selected important variables, we use standard estimation techniques such as ML or REML to estimate the corresponding coefficient β_{imp}.

This procedure is easily applied to other types of data such as count data. The detailed implementation of this procedure can be found in Section 3.5.

3.4 Selection of Tuning Parameter λ

One traditional criterion for selecting λ is the data-driven method GCV. Due to the complicated structure of GLMMs, it is not easy to derive the GCV in general. However, for the PQL approach, because we do variable selection and parameter estimation in the framework of linear mixed models, normal theories can be used to establish a working formula for GCV. In addition, the restricted maximum likelihood (REML) criterion is established under the PQL approach. Alternatively, we can use the BIC to tune the parameter λ. Our simulation studies show that the BIC has the best performance.

For the TPQL approach, because we use the PQL approach to do model selection at the first stage, the criteria used in the TPQL approach are essentially same as those used at the PQL approach.

3.4.1 GCV

Wahba (1990) proposed the below GCV for tuning the nonparametric smoothing parameter λ in the independent data case

$$GCV(\lambda) = \frac{\frac{1}{n}||(I - A(\lambda))Y||^2}{[\frac{1}{n}tr(I - A(\lambda))]^2}, \tag{3.21}$$

where $A(\lambda)$ is a projection matrix.

Wang (1998) extended the GCV formula for correlated longitudinal data. We further develop the GCV formula for the choice of λ for variable selection in GLMMs.

For the model (3.8), let $e = X_2\beta_2 + Zb + \epsilon$, then $\text{Var}(e) = V^* = X_2\Sigma_{22}^{-1}X_2^T +$

$ZGZ^T + W^{-1}$, and hence model (3.8) is equivalent to

$$y = X_1\beta_1 + e. \tag{3.22}$$

We multiply $V^{*-\frac{1}{2}}$ on both sides of equation (3.22), and let $y^* = V^{*-\frac{1}{2}}y$, $X_1^* = V^{*-\frac{1}{2}}X_1$, and $\epsilon^* = V^{*-\frac{1}{2}}e$. Then equation (3.22) can be written as

$$y^* = X_1^*\beta_1 + \epsilon^*, \tag{3.23}$$

where $\epsilon^* \sim N(0, I)$. The GCV score for (3.23) using Wahba's formula is

$$GCV(\lambda) = \frac{\frac{1}{n}\|(I - A_\lambda)y^*\|^2}{[\frac{1}{n}tr(I - A_\lambda)]^2}. \tag{3.24}$$

We define $\widehat{y}^* = A_\lambda y^*$, then we have

$$\begin{aligned}
\|(I - A_\lambda)y^*\|^2 &= [(I - A_\lambda)y^*]^T[(I - A_\lambda)y^*] \\
&= (y^* - \widehat{y}^*)^T(y^* - \widehat{y}^*) \\
&= (V^{*-\frac{1}{2}}y - V^{*-\frac{1}{2}}\widehat{y})^T(V^{*-\frac{1}{2}}y - V^{*-\frac{1}{2}}\widehat{y}) \\
&= (y - \widehat{y})^T V^{*-1}(y - \widehat{y}).
\end{aligned} \tag{3.25}$$

Furthermore,

$$\begin{aligned}
\widehat{y^*} &= X_1^* \widehat{\beta}_1 \\
&= X_1^* (X_1^T V^{*-1} X_1)^{-1} X_1^T V^{*-1} y \\
&= V^{*-\frac{1}{2}} X_1 (X_1^T V^{*-1} X_1)^{-1} X_1^T V^{*-\frac{1}{2}} V^{*-\frac{1}{2}} y \\
&= V^{*-\frac{1}{2}} X_1 (X_1^T V^{*-1} X_1)^{-1} X_1^T V^{*-\frac{1}{2}} y^* \\
&= A_\lambda y^*,
\end{aligned}$$

where

$$A_\lambda = V^{*-\frac{1}{2}} X_1 (X_1^T V^{*-1} X_1)^{-1} X_1^T V^{*-\frac{1}{2}}. \tag{3.26}$$

Obviously, $A_\lambda^2 = A_\lambda$, and

$$tr(A_\lambda) = rank(X_1) = dim(\widehat{\beta}_1) = p^*. \tag{3.27}$$

Because β_1 is the part of coefficients which is not subject to any penalty, p^* is less or equal to the number of important covariates d_1.

By plugging equations (3.25) and (3.27) into the GCV formula (3.24), we have

$$GCV(\lambda) = \frac{n(y - \widehat{y})^T V^{*-1} (y - \widehat{y})}{(n - p^*)^2}. \tag{3.28}$$

3.4.2 REML

Patterson and Thompson (1971) proposed to estimate variance components under the restricted maximum likelihood (REML) for linear mixed models , which reduces the bias of variance component estimators by taking into account the lost degree of freedom resulting from the estimation of fixed effects. In knowledge of this good property, we propose the REML criterion to select the tuning parameter λ.

By its definition, the REML of a linear mixed model (3.8) is

$$\ell_{REML} = -\frac{1}{2}\log|V^*| - \frac{1}{2}\log|X_1^T V^{*-1} X_1| - \frac{1}{2}(y - X_1\widehat{\beta}_1)^T V^{*-1}(y - X_1\widehat{\beta}_1). \quad (3.29)$$

The difference between our formula and the conventional one is that y in (3.29) is not the initial observation vector, but a working vector defined by formula (3.6). At each iteration, we need to update y by formula (3.6), and use the latest version of y after the convergence of β to calculate REML values by equation (3.29).

3.4.3 BIC

The BIC criterion is used to choose the tuning parameter λ for the PQL approach. Under the PQL method, the model selection and further inference are based on the penalized quasi-likelihood $\tilde{\ell}_p$. Hence, the formula of BIC for the PQL approach is

$$BIC = -2\tilde{\ell_p}\left(\widehat{\beta}, \widehat{\theta}; Y\right) + d_1 \log n. \quad (3.30)$$

3.5 Computational Algorithm of PQL and TPQL

The practical implementations of the PQL and TPQL approaches are more complicated than that of the FL approach due to two reasons. First the variance component θ is updated after coefficient β is convergent. Second more criteria are used for tuning λ.

3.5.1 Computational Algorithm of PQL

1. standardization

   Given data, standardize the covariates X by their sample means and sample standard deviations.

2. initiation

   Given data, use procedure *glimmix* in *SAS* to get initial values $\widehat{\beta}^{(0)}$, $\widehat{\theta}^{(0)}$, and $\widehat{b}^{(0)}$ for β, θ and b respectively by fitting a full model.

3. Outer loop

   (a) Set a range for λ, and grid λ into g levels $\lambda_1, \lambda_2, \ldots, \lambda_s$.

   (b) For each λ_i, go to the inner loop for β to jointly estimate β and b, then use $\widehat{\beta}$, $\widehat{b}$ and the working vector y from the inner loop, and $\widehat{\theta}^{(0)}$ to compute $GCV(\lambda_i)$, $REML(\lambda_i)$, and $BIC(\lambda_i)$. The selected λ_i is the one associated with the lowest GCV score, the smallest BIC or the largest REML value.

   (c) Identify the final estimators $\widehat{\beta}$ and $\widehat{b}$ for GCV, REML and BIC respectively according to the above rules.

   (d) With these final estimators, go to the inner loop for θ to get the variance

component estimator $\widehat{\theta}$ and the corrected variance component estimator $\widehat{\theta}_c$ for GCV, REML and BIC respectively.

(e) Use $\widehat{\theta}$ and the selected λ to re-estimate β for GCV, REML and BIC respectively.

(f) Use $\widehat{\theta}_c$ and the selected λ to get the corrected coefficient estimator $\widehat{\beta}_c$ for GCV, REML and BIC respectively.

(g) Adjust $\widehat{\beta}$ and $\widehat{\beta}_c$ to their original scale for GCV, REML and BIC respectively.

4. Inner loop

(a) inner loop for β

i. Use $\widehat{\beta}^{(0)}$, $\widehat{b}^{(0)}$ and $\widehat{\theta}$ to get working vector $y^{(0)}$ by (3.6)

ii. For the $(k+1)$th iteration, set the estimators $\widehat{\beta}^{(k)}$, $\widehat{b}^{(k)}$ and $y^{(k)}$ from the kth iteration, update β, b and working y by

$$\begin{aligned}
\widehat{\beta}^{(k+1)} &= \left\{X^T(V^{(k)})^{-1}X + n\Sigma_\lambda(\widehat{\beta}^{(k)})\right\}^{-1} X^T(V^{(k)})^{-1}y^{(k)},\\
\widehat{b}^{(k+1)} &= GZ^T(V^{(k)})^{-1}\left(y^{(k)} - X\widehat{\beta}^{(k+1)}\right),\\
y^{(k+1)} &= X\widehat{\beta}^{(k+1)} + Z\widehat{b}^{(k+1)} + \Delta^{(k+1)}\left(Y - \mu^{\widehat{b}^{(k+1)}}\right).
\end{aligned}$$

iii. Compute $\max|\widehat{\beta}^{(k+1)} - \widehat{\beta}^{(k)}|$. Compare this value to a predetermined value δ, say $\delta = 10^{-4}$. If it is no larger than δ, then stop the inner loop and return to the outer loop 3.

iv. Repeat above steps until $\widehat{\beta}$ converges.

(b) inner loop for θ

i. For the $(k+1)$th iteration, set the estimator $\widehat{\theta}^{(k)}$ from the kth iteration, use the nonzero components of $\widehat{\beta}$ and $\widehat{b}$ to update θ by

$$\widehat{\theta}^{(k+1)} = \widehat{\theta}^{(k)} + \left\{ I(\widehat{\theta}^{(k)}) \right\}^{-1} S(\widehat{\theta}^{(k)}).$$

ii. Compute $\max |\widehat{\theta}^{(k+1)} - \widehat{\theta}^{(k)}|$. Compare this value to a predetermined value ς, say $\varsigma = 10^{-4}$. If it is no larger than ς, then stop the inner loop and return to the outer loop 3.

iii. Repeat above step until θ converges.

3.5.2 Computational Algorithm of TPQL

The implementation of the TPQL approach is quite similar to that of the PQL method, except that its outer loop is simpler and no inner loop for the variance component θ is required. The same steps of standardization and initiation as the PQL method will not be repeated here. Its outer loop includes the following steps:

1. Set a range for λ, and grid λ into s levels $\lambda_1, \lambda_2, \ldots, \lambda_s$.

2. For each λ_i, go to the inner loop for β to jointly estimate β and b, then use $\widehat{\beta}$, $\widehat{b}$ and the working vector y from the inner loop, and $\widehat{\theta}^{(0)}$ to compute $GCV(\lambda_i)$, $REML(\lambda_i)$, and $BIC(\lambda_i)$. The selected λ_i is the one associated with the lowest GCV score, the smallest BIC or the largest REML value.

3. Identify the estimator $\widehat{\beta}$ for GCV, REML and BIC respectively according to the above rules.

4. For each criterion, compare each component $\widehat{\beta}_j$ of its corresponding coefficient estimator $\widehat{\beta}$ to a small threshold ζ, say $\zeta = 10^{-4}$. If $|\widehat{\beta}_j| \leq \zeta$, then set $\widehat{\beta}_j = 0$.

5. Identify those explanatory variables which correspond to non-zero components of $\widehat{\beta}$. They are the selected important covariates.

6. Use these informative covariates and the standard software (*nlmixed* procedure in *SAS*) to estimate the coefficients of important variables and variance components.

7. The final coefficient estimator $\widehat{\beta}$ consists of some zeros, which are coefficient estimators for non-important variables, and non-zeros, which are from the previous step and are the coefficient estimators for those selected important variables.

8. Adjust $\widehat{\beta}$ to its original scale for GCV, REML and BIC respectively.

3.6 summary

In this chapter, the PQL method of Breslow and Clayton (1993) was extended to model selection problem in GLMMs, especially for GLMMs with high-dimensional random effects, so that normal theories can be used to do variable selection and make inference. A sandwich formula for the covariance of the coefficient estimator and a working covariance formula for the variance component estimator were established. A bias-correction procedure was presented to improve the estimation capability of the PQL method. In addition, the TPQL approach was also developed to take care of the bias issue of the PQL method. Besides BIC, working GCV and REML were also developed for the selection of the tuning parameter. The computational algorithm of the proposed methods were elaborated.

Their finite sampling performance will be studied in Chapter 5.

Chapter 4

Approximate Marginal Likelihood Approach

4.1 Motivation

Zeger *et al.* (1988) studied the relationship between random-effect models and marginal models. The focus of random-effect models is generally on the conditional mean of responses given subject specific effects, while the interest of marginal models is on the marginal mean. For the random-intercept-effect model with random effects $b_i \sim N(0, \theta)$ $(i = 1, \ldots, m)$ for binary data:

$$Y_{ij}|b_i \sim Bin(1, \mu_{ij}^b), \quad g(\mu_{ij}^b) = X_{ij}^T\beta + b_i, \tag{4.1}$$

its approximate marginal model is

$$Y_{ij} \sim Bin(1, \mu_{ij}), \qquad g(\mu_{ij}) = X_{ij}^T\beta_*. \tag{4.2}$$

Zeger *et al.* (1988) demonstrated that β is related to β_* in the following way: if $g = probit = \Phi^{-1}$, then $\beta_* = \beta/\sqrt{\theta + 1}$; if $g = logit$, then $\beta_* \approx \beta/\sqrt{c^2\theta + 1}$ with $c^2 \approx 0.346$ (Also see equation 7.4.7 in Diggle *et al.*, 2002).

The approximate linear relationship between β and β_* suggests that, variable selection in the random effect model (4.1) is approximately equivalent to variable selection in its corresponding marginal model (4.2). The possible advantages of using the latter to do variable selection include less computation, due to the simpler structure of the marginal model. Therefore, we can re-order the data by ignoring the correlation between observations, and do variable selection in the marginal model at the first stage. Then at the second stage we may apply ML or REML to the original data for parameter estimation with those selected variables. Our simulation studies show that, the flexible two-stage design of the AML procedure is not only computationally efficient by avoiding integration involved in the FL method, but produces good parameter estimators.

4.2 Model Selection by Approximate Marginal Model

For the *logit* link, after re-ordering of the data, the pseudo marginal-likelihood (PML) (or the likelihood of the approximate marginal model (4.2)) is

$$\ell_{pml}(\beta_*; Y) = \log \left\{ \prod_{k=1}^{n} \mu_k^{Y_k} (1-\mu_k)^{1-Y_k} \right\} = Y^T X \beta_* - \sum_{k=1}^{n} \log \left\{ 1 + \exp\left(X_k^T \beta_*\right) \right\}, \quad (4.3)$$

where n is the total number of observations.

As before, we apply the local quadratic approximation on the penalized pseudo marginal-likelihood (PPML)

$$\ell_{ppml}(\beta_*; Y) = \ell_{pml}(\beta_*; Y) - n \sum_{j=1}^{d} p_\lambda(|\beta_{j*}|). \quad (4.4)$$

In other words, for given λ, and an initial value $\widehat{\beta}_*^{(0)}$ which is close to the maximizer of $l_{ppml}(\beta_*; Y)$,

$$\begin{aligned}\ell_{ppml}(\beta_*; Y) &\approx \ell_{pml}(\widehat{\beta}_*^{(0)}; Y) + \nabla\ell_{pml}(\widehat{\beta}_*^{(0)}; Y)^T(\beta_* - \widehat{\beta}_*^{(0)}) \\ &\quad + \frac{1}{2}(\beta_* - \widehat{\beta}_*^{(0)})^T\nabla^2\ell_{pml}(\widehat{\beta}_*^{(0)}; Y)(\beta_* - \widehat{\beta}_*^{(0)}) \\ &\quad - \frac{1}{2}\beta_*^T n\Sigma_\lambda(\widehat{\beta}_*^{(0)})\beta_*, \end{aligned} \tag{4.5}$$

where $\nabla\ell(\widehat{\beta}_*^{(0)}; Y) = \frac{\partial\ell(\hat{\beta}_*^{(0)}; Y)}{\partial\beta_*}$, $\nabla^2\ell(\widehat{\beta}_*^{(0)}; Y) = \frac{\partial\ell(\hat{\beta}_*^{(0)}; Y)}{\partial\beta_*\partial\beta_*^T}$, and

$$n\Sigma_\lambda(\widehat{\beta}_*^{(0)}) = diag\left\{\frac{np'_\lambda(|\widehat{\beta}_{1*}^{(0)}|)}{|\widehat{\beta}_{1*}^{(0)}|}, \ldots, \frac{np'_\lambda(|\widehat{\beta}_{d*}^{(0)}|)}{|\widehat{\beta}_{d*}^{(0)}|}\right\}.$$

The maximization on the approximation function to ℓ_{ppml} produces

$$\begin{aligned}\widehat{\beta}_*^{(1)} &= \widehat{\beta}_*^{(0)} + \\ &\left\{-\nabla^2\ell_{pml}(\widehat{\beta}_*^{(0)}; Y) + n\Sigma_\lambda(\widehat{\beta}_*^{(0)})\right\}^{-1}\left\{\nabla\ell_{pml}(\widehat{\beta}_*^{(0)}; Y) - n\Sigma_\lambda(\widehat{\beta}_*^{(0)})\widehat{\beta}_*^{(0)}\right\}. \end{aligned} \tag{4.6}$$

This iterative ridge algorithm allows us to select the important variables in the approximate marginal model (4.2), which are also important explanatory variables in the random effect model (4.1). Then the standard estimation techniques such as ML or REML can be applied to get estimators for the coefficient β_{imp} of important variables and the variance component θ (for example, use *nlmixed* procedure of *SAS*).

4.3 Selection of Tuning Parameter λ

BIC is the criterion for the selection of λ for the AML approach. Because we do model selection from the marginal point of view, the pseudo-marginal likelihood $\ell_{pml}(\widehat{\beta}_*; Y)$ is used for the evaluation of BIC. Specifically,

$$BIC = -2\ell_{pml}(\widehat{\beta}_*; Y) + d_1 \log n. \tag{4.7}$$

4.4 Computational Algorithm of AML

The practical implementation of the AML is similar to that of the TPQL approach. The detailed steps are as follows.

1. standardization

   Given data, standardize the covariates X by their sample means and sample standard deviations.

2. initiation

   Given data, use procedure *genmod* in *SAS* to get an initial value $\widehat{\beta}_*^{(0)}$ for β_* by fitting a full model.

3. Outer loop

   (a) Set a range for λ, and grid λ into s levels $\lambda_1, \lambda_2, \ldots, \lambda_s$.

   (b) For each λ_i, go to the inner loop to estimate β_*, compute $BIC(\lambda_i)$, select λ_i that yields smallest BIC.

   (c) Identify the estimator $\widehat{\beta}_*$ which is associated with the selected λ_i.

(d) Compare each component $\widehat{\beta}_{j*}$ of $\widehat{\beta}_*$ to a small threshold ζ, say $\zeta = 10^{-4}$. If $|\widehat{\beta}_{j*}| \leq \zeta$, then set $\widehat{\beta}_{j*} = 0$.

(e) Identify those explanatory variables which correspond to non-zero components of $\widehat{\beta}_*$. They are selected important covariates.

(f) Use these informative covariates and the standard software (*nlmixed* procedure in *SAS*) to estimate the coefficients of important variables and variance components.

(g) The final coefficient estimator $\widehat{\beta}$ consists of some zeros, which are coefficient estimators for non-important variables, and non-zeros, which are from the previous step and are the coefficient estimators for those selected important variables.

(h) Adjust $\widehat{\beta}$ to its original scale.

4. Inner loop

(a) For the $(k+1)$th iteration, let $\widehat{\beta}_*^{(k)}$ be the estimator from the kth iteration. We update β_* by

$$\begin{aligned}\widehat{\beta}_*^{(k+1)} &= \widehat{\beta}_*^{(k)} + \\ &\left\{-\nabla^2\ell(\widehat{\beta}_*^{(k)}; Y) + n\Sigma_\lambda(\widehat{\beta}_*^{(k)})\right\}^{-1}\left\{\nabla\ell(\widehat{\beta}_*^{(k)}; Y) - n\Sigma_\lambda(\widehat{\beta}_*^{(k)})\widehat{\beta}_*^{(k)}\right\}.\end{aligned}$$

(b) Compute $\max|\widehat{\beta}_*^{(k+1)} - \widehat{\beta}_*^{(k)}|$. Compare this value to a predetermined value δ, say $\delta = 10^{-4}$. If it is no larger than δ, then stop the inner loop and return to the outer loop 3.

(c) Repeat above steps until $\widehat{\beta}_*$ converges.

4.5 Summary

An marginal method for model selection in logistic random effect model was proposed. This method was based on the approximate marginal likelihood. The two-stage design brings computational efficiency, and leads to good parameter estimators.

Its finite sampling performance will be studied in Chapter 5.

Chapter 5

Numerical Results

5.1 Introduction

In the preceding chapters, four variable selection procedures using the SCAD penalty have been proposed for model selection in GLMMs. They are the FL approach, the PQL approach, the TPQL approach and the AML approach. Among them, the FL approach is intended for GLMMs with low-dimensional random effects, while the PQL approach can easily handle GLMMs with multiple-level random effects for discrete responses. The TPQL approach is proposed to deal with the bias issue of the PQL approach in parameter estimation. The AML approach has the feature of doing model selection in GLMMs by marginal approach, and using standard optimization techniques to do parameter estimation.

In this chapter, we illustrate the finite sampling performance of the proposed procedures. In section (5.2), we study the performance of the approaches via simulation. Their utility is demonstrated via their application to the infectious disease data in section (5.3).

5.2 Simulation Studies

5.2.1 Design of Simulations

To evaluate the performance of the proposed procedures, we did simulation studies using the extremely spare case of binary data with a canonical link. We believe that all procedures other than the AML procedure, which works only for correlated binary data with random intercept only, generally work even better for less spare data.

We generate data from the following model

$$logit\left\{Pr(Y_{ij}|b_i)\right\} = X_{ij}^T\beta + Z_{ij}^T b_i, \tag{5.1}$$

where $i = 1, \ldots, m$ with m being the total number of subjects, and $j = 1, \ldots, n_i$ with n_i being the number of observations on each subject. We generate n_i from $Bin(7, 0.7)$, so on average there are 5 observations on each subject. To explore the effect of sample size on the performance of the proposed procedures, in general we consider two sample sizes: $m = 100$ and $m = 200$. The total sample size is around 500 when $m = 100$ and around 1000 when $m = 200$. But we also present the model selection results for $m = 50$, under which, the ratio of the total number of observations to the number of parameters is only around 15:1.

For each subject i, its covariates are $X_{ij} = (X_{1ij}, X_{2ij}, \ldots, X_{16ij})$ with $X_{1ij} = 1$ being the intercept, $(X_{2ij}, \ldots, X_{12ij})$ distributed as multivariate normal with mean 0 and $AR(1)$ covariance structure with the coefficient $\rho = 0.5$. We generate $X_{13ij}, \ldots, X_{16ij}$ independently from the Bernoulli distribution with a probability of success of 0.5.

For our simulation studies, we consider models with random intercept only. The

random effects b_i are generated from $N(0, \theta)$. The performance of the proposed procedures may be affected by the variance of random effects. Intuitively, all of procedures should have better performance when the variation of random effects is smaller. Therefore four values of θ are included in the study: $\theta = 1.96$, $\theta = 1$, $\theta = 0.81$ and $\theta = 0.64$.

We investigate the performance of the proposed procedures under two scenarios: scenario I: $(\beta_2, \beta_5, \beta_6, \beta_7) = (0.9, 0.9, 0.9, 0.9)$, $\beta_1 = 0.5$, and $\beta_k = 0$ $(k \neq 1, 2, 5, 6, 7)$; scenario II: $(\beta_2, \beta_5, \beta_6, \beta_7) = (0.5, 0.5, 0.5, 0.5)$, $\beta_1 = 0.5$, and $\beta_k = 0$ $(k \neq 1, 2, 5, 6, 7)$. The coefficient for the intercept is same for both scenarios. But the signals of important covariates are stronger under scenario I than under scenario II. In total we have eight settings. For each simulation, the Monte Carlo sample size is 100.

The difficulty level of models in each setting is evaluated by R^2, referred as the proportion of variation explained by the true model. In the literature, R^2 is proposed only for linear models, and no formula of R^2 has been developed for GLMMs. We propose two ways to calculate R^2 for GLMMs in the following section. The first method is based on its original definition, and the second method is based on the Kullback-Leibler divergence.

For each generated data set, model selection and parameter estimation are proceeded with four approaches: the full likelihood (FL) approach, the penalized quasi-likelihood (PQL) approach, the two-stage penalized quasi-likelihood (TPQL) approach and the approximate marginal likelihood (AML) approach. The general tuning procedure for λ is BIC. For the approach of PQL, we developed two more tuning procedures: GCV and REML. However, due to the poor performance of GCV and REML in model selection, we present only the results of variable selection and model fitting when using these two criteria.

We evaluate the performance of the proposed procedures in variable selection and model fitting using three criteria: (1) the average model size, denoted by size, (2) the average number of coefficients which are set to 0 correctly, denoted by $Corr.0$, and (3) the average number of coefficients which are set to 0 by mistake, denoted by $Inc.0$. Also we present the frequency of being selected for each variable over 100 runs.

We evaluate the proposed procedures in parameter estimation by reporting (1) the Monte Carlo sample mean of β and θ over 100 runs, denoted by $\widehat{\beta}$ and $\widehat{\theta}$, (2) the median absolute deviation divided by 0.6745, denoted by SD, regarded as an estimate of the Monte Carlo standard error of the parameter estimator, (3) the mean of the 100 estimated standard error, denoted by SE, (4) bias, and (5) 95% coverage probability of confidence interval.

5.2.2 Definition of R^2

First, we consider R^2 defined as the proportion of the variation explained by the covariates:

$$R^2 = 1 - \frac{Var(Y|X)}{Var(Y)}. \tag{5.2}$$

By the conditional variance identity theorem,

$$Var(Y|X) = Var_b \left\{E(Y|X,b)\right\} + E_b \left\{Var(Y|X,b)\right\}. \tag{5.3}$$

For model (5.1), from which we are generating data, $E(Y|X,b) = \mu^b$ and $Var(Y|X,b) = \mu^b(1-\mu^b)$. Using the relationship between moments and variance, we have

$$Var_b\left\{E(Y|X,b)\right\} = E_b[(\mu^b)^2] - [E_b(\mu^b)]^2, \tag{5.4}$$

and

$$E_b\left\{Var(Y|X,b)\right\} = E_b(\mu^b) - E_b[(\mu^b)^2]. \tag{5.5}$$

Therefore,

$$Var(Y|X) = E_b(\mu^b) - [E_b(\mu^b)]^2. \tag{5.6}$$

$E_b(\mu^b)$ will be evaluated using the true model. The involved computation is intensive because of the integration. When the dimension of random effects is not high, the standard numerical technique of Gaussian quadrature can be applied for the values of $E_b(\mu^b)$ and $[E_b(\mu^b)]^2$. Plugging (5.6) into (5.2), R^2 can be evaluated for GLMMs under its original definition.

The value of R^2 depends on the design matrix for each setting. In our simulation studies, the design matrix changes from one simulation run to the next for each setting. Therefore, to get a better idea about the difficulty level of each setting, we use simulations for the sampling values of R^2. In other words, for each generated data set, we use above formula to calculate R^2. The final result we present in this dissertation is the average of R^2 values for 100 data sets generated. For one generated data set, $Var(Y|X)$ can be estimated by the mean of $E_b(\mu^b) - [E_b(\mu^b)]^2$, and the variance of

Y can be estimated by the its empirical standard error. Considering that X is random (changes from simulation to simulation), we fix seed when generating data. In this way it is also easier for interpretation. However, the performance of the proposed procedures is not affected by the seed.

Second, we consider R^2 based on the Kullback-Leibler (KL) divergence. Cameron and Windmeijer (1995) constructed the R^2 measure of goodness of fit for the class of exponential family regression models in the case of independent data. They defined R^2 as the proportionate reduction in uncertainty, measured by the Kullback-Leibler divergence, due to the inclusion of regressors. We extend this definition to GLMMs, where R^2 measures the proportionate reduction in the potentially recoverable informaton achieved by the fitted models. Specifically,

$$R^2_{KL} = 1 - \frac{K(Y, \widehat{\mu})}{K(Y, \widehat{\mu}_0)}, \tag{5.7}$$

where $\widehat{\mu}$ is the fitted mean of the true model (with only important variables included), $\widehat{\mu}_0$ is the fitted mean of the intercept only model, and $K(Y, \widehat{\mu})$ is a measure of the deviation of Y from its estimated mean $\widehat{\mu}$. The Kullback-Leibler (KL) divergence can be defined as

$$K(Y, \mu) \;\; = \;\; 2E_Y \log \left\{ f_Y(Y) / f_\mu(Y) \right\}. \tag{5.8}$$

For exponential family regression models, Hastie (1987) and Vos (1991) showed that

the expectation in (5.8) drops out. So we have

$$\begin{aligned} K(Y,\mu) &= 2\log\{f_Y(Y)/f_\mu(Y)\} \\ &= -2\log\{f_\mu(Y)\}, \end{aligned} \tag{5.9}$$

because $\ell(Y;Y) = \log\{f_Y(Y)\} = 0$.

Hence for GLMMs (1.2),

$$R^2_{KL} = 1 - \frac{\ell_{true}(\widehat{\mu};Y)}{\ell_0(\widehat{\mu}_0;Y)}, \tag{5.10}$$

where $\ell_{true}(\widehat{\mu};Y)$ is the log-likelihood of the true model evaluated at the penalized maximum likelihood estimator, and $\ell_0(\widehat{\mu}_0;Y)$ is the log-likelihood of the intercept only model evaluated at the maximum likelihood estimator.

Same as the evaluation of R^2 based on its original definition, we use simulations to get the sampling value of R^2_{KL}. For each data set we generate, we use formula (5.10) to get R^2_{KL}. The final R^2_{KL} is the average of 100 values.

5.2.3 Simulation Results for R^2

Table 5.1: R^2 at two scenarios: I. $(\beta_2,\beta_5,\beta_6,\beta_7) = (0.9,0.9,0.9,0.9)$, II. $(\beta_2,\beta_5,\beta_6,\beta_7) = (0.5,0.5,0.5,0.5)$. R^2, evaluated by its definition; R^2_{KL}, evaluated by KL divergence.

	Scenario I				*Scenario II*			
	$\theta_1 = 1.96$	$\theta_2 = 1.00$	$\theta_3 = 0.81$	$\theta_4 = 0.64$	$\theta_1 = 1.96$	$\theta_2 = 1.00$	$\theta_3 = 0.81$	$\theta_4 = 0.64$
R^2	0.35	0.39	0.40	0.41	0.16	0.18	0.20	0.21
R^2_{KL}	0.33	0.36	0.36	0.37	0.16	0.18	0.18	0.18

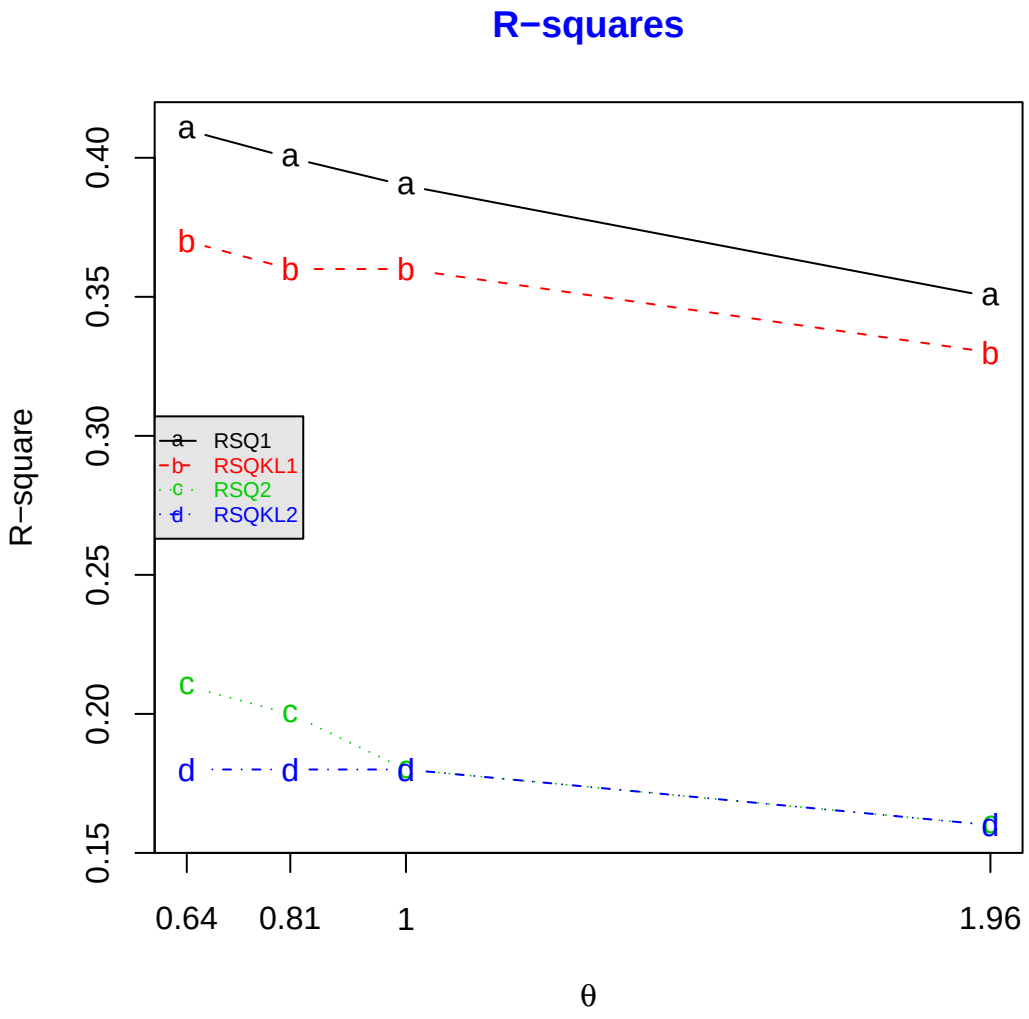

Figure 5.1: R^2 at two scenarios: I. $(\beta_2, \beta_5, \beta_6, \beta_7) = (0.9, 0.9, 0.9, 0.9)$, II. $(\beta_2, \beta_5, \beta_6, \beta_7) = (0.5, 0.5, 0.5, 0.5)$. RSQ1, R^2 by its definition at scenario I; RSQKL1, R^2 by KL divergence at scenario I; RSQ2, R^2 by its definition at scenario II; RSQKL2, R^2 by KL divergence at scenario II.

Table 5.1 presents the R^2 values for different settings. We notice that R^2 for scenario I are much larger than those for scenario II, which means that the settings under scenario II are much more difficult than scenario I. Therefore, we would expect that our procedures have better performance under scenario I.

Also, the R^2 values depend on θ. Larger θ's lead to smaller R^2. Consequently, one would expect a better performance of the proposed procedures under smaller θ's.

Furthermore, a comparison of R^2 to R^2_{KL} shows that R^2 values under two definitions are very close to each other, which suggests that either can be used to judge the degree of difficulty of the considered model.

Figure 5.1 further presents the above patterns.

5.2.4 Simulation Results for Variable Selection

According to our simulation design, the important covariates are X_1, the intercept, X_2, X_5, X_6 and X_7. Since a GLMM almost always contains an intercept, we do not impose any penalty on β_1. So the true model size is 4, and the true number of zero coefficients is 11.

Tables 5.2 and 5.3 give a summary of selection results by BIC when sample sizes are moderate to large ($m = 100$ or $m = 200$). Overall, all of our proposed procedures have very good performance in variable selection. As shown by the values of R^2, the settings we chose are challenging relative to the sample size. In particular, when β is small and θ is large, both R^2 are less than 0.2, which exhibits the difficulty level for variable selection and model inference. However, even under this extreme case, the values of $Corr.0$ all above 10.30 and $Inc.0$ all below 0.8 when $m = 100$ indicate that

Table 5.2: Model selection summary at scenario I: $(\beta_2, \beta_5, \beta_6, \beta_7) = (0.9, 0.9, 0.9, 0.9)$ by BIC. $size$, average model size; $Corr.0$, average number of coefficients which are set to 0 correctly; $Inc.0$, average number of coefficients which are set to 0 by mistake; FL, full likelihood approach; PQL, penalized quasi-likelihood approach; AML, approximate marginal likelihood approach.

m	*method*	$\theta_1 = 1.96$			$\theta_2 = 1.00$			$\theta_3 = 0.81$			$\theta_4 = 0.64$		
		size	$Corr.0$	$Inc.0$	*size*	$Corr.0$	$Inc.0$	*size*	$Corr.0$	$Inc.0$	*size*	$Corr.0$	$Inc.0$
100	FL	4.17	10.82	0.01	4.10	10.87	0.03	4.08	10.89	0.03	4.12	10.87	0.01
		(0.55)	(0.54)	(0.01)	(0.46)	(0.42)	(0.17)	(0.37)	(0.31)	(0.17)	(0.52)	(0.50)	(0.01)
	PQL	4.24	10.72	0.04	4.06	10.90	0.04	4.03	10.93	0.04	4.05	10.94	0.01
		(0.65)	(0.60)	(0.20)	(0.42)	(0.36)	(0.20)	(0.36)	(0.29)	(0.20)	(0.30)	(0.28)	(0.01)
	AML	4.17	10.78	0.05	4.15	10.83	0.02	4.17	10.82	0.01	4.07	10.92	0.01
		(0.66)	(0.61)	(0.22)	(0.48)	(0.45)	(0.14)	(0.49)	(0.48)	(0.01)	(0.32)	(0.31)	(0.01)
200	FL	4.18	10.82	0.00	4.16	10.84	0.00	4.13	10.87	0.00	4.17	10.83	0.00
		(0.52)	(0.52)	(0.00)	(0.42)	(0.42)	(0.00)	(0.39)	(0.39)	(0.00)	(0.43)	(0.43)	(0.00)
	PQL	4.03	10.97	0.00	4.07	10.93	0.00	4.08	10.92	0.00	4.11	10.89	0.00
		(0.17)	(0.17)	(0.00)	(0.29)	(0.29)	(0.00)	(0.27)	(0.27)	(0.00)	(0.34)	(0.34)	(0.00)
	AML	4.12	10.88	0.00	4.09	10.91	0.00	4.13	10.87	0.00	4.09	10.91	0.00
		(0.41)	(0.41)	(0.00)	(0.35)	(0.35)	(0.00)	(0.39)	(0.39)	(0.00)	(0.32)	(0.32)	(0.00)

the proposed procedures work still well in terms of model selection.

From Table 5.2 we also notice that in general, as the sample size increases, the model size, the average number of coefficients which are set to zero correctly ($Corr.0$ in tables) and the average number of coefficients which are set to zero by mistake ($Inc.0$ in tables) are all closer to their true values, which demonstrates the better performance of the proposed procedures with larger sample sizes.

Compared with the values under scenario II, the proposed procedures have larger $Corr.0$ values (closer to 11), and smaller $Inc.0$ values under scenario I. As we expect from the values of R^2 under different settings, these procedures have better performance

Table 5.3: Model selection summary at scenario II: $(\beta_2, \beta_5, \beta_6, \beta_7) = (0.5, 0.5, 0.5, 0.5)$ by BIC. $size$, average model size; $Corr.0$, average number of coefficients which are set to 0 correctly; $Inc.0$, average number of coefficients which are set to 0 by mistake; FL, full likelihood approach; PQL, penalized quasi-likelihood approach; AML, approximate marginal likelihood approach.

m	*method*	$\theta_1 = 1.96$			$\theta_2 = 1.00$			$\theta_3 = 0.81$			$\theta_4 = 0.64$		
		size	$Corr.0$	$Inc.0$	*size*	$Corr.0$	$Inc.0$	*size*	$Corr.0$	$Inc.0$	*size*	$Corr.0$	$Inc.0$
100	FL	3.73	10.49	0.78	4.33	10.30	0.37	4.42	10.26	0.32	4.28	10.33	0.39
		(1.38)	(0.83)	(0.79)	(1.19)	(0.85)	(0.58)	(1.23)	(0.96)	(0.55)	(1.27)	(0.86)	(0.68)
	PQL	3.72	10.68	0.60	4.11	10.58	0.31	4.00	10.64	0.36	4.08	10.59	0.33
		(0.96)	(0.58)	(0.72)	(0.93)	(0.71)	(0.54)	(0.88)	(0.66)	(0.60)	(0.98)	(0.71)	(0.53)
	AML	4.00	10.33	0.67	4.23	10.42	0.35	4.03	10.51	0.46	4.43	10.34	0.23
		(1.10)	(0.75)	(0.76)	(1.09)	(0.76)	(0.64)	(1.21)	(0.77)	(0.74)	(1.04)	(0.84)	(0.53)
200	FL	4.15	10.81	0.04	4.18	10.82	0.00	4.15	10.85	0.00	4.27	10.73	0.00
		(0.48)	(0.44)	(0.20)	(0.41)	(0.41)	(0.00)	(0.41)	(0.41)	(0.00)	(0.56)	(0.56)	(0.00)
	PQL	3.92	10.96	0.12	4.02	10.95	0.03	4.04	10.96	0.00	4.05	10.94	0.01
		(0.39)	(0.20)	(0.32)	(0.28)	(0.22)	(0.17)	(0.20)	(0.20)	(0.00)	(0.26)	(0.24)	(0.01)
	AML	4.14	10.76	0.10	4.18	10.81	0.01	4.12	10.88	0.00	4.16	10.82	0.02
		(0.66)	(0.57)	(0.30)	(0.59)	(0.58)	(0.01)	(0.47)	(0.47)	(0.00)	(0.52)	(0.50)	(0.14)

when the signals are stronger. Under scenario I, except the PQL approach when $m = 200$, as θ increases, the capability of all other procedures in model selection decreases. When $m = 200$, the model selection capability of the PQL approach seems to be positively related to the value of θ, which is hard to explain. Under scenario II, as θ increases, $Corr.0$ decrease, and $Inc.0$ increase, with an exception of $Corr.0$ for the FL approach when $m = 100$. This further confirms the negative relationship of the performance of the proposed procedures to the variance of random effects. The good performance of the proposed procedures in model selection exists whether the signals are strong or weak when the sample sizes are moderate to large, though with stronger

signals, their performance is even better.

Table 5.4: Variable selection frequency at scenario I: $(\beta_2, \beta_5, \beta_6, \beta_7) = (0.9, 0.9, 0.9, 0.9)$ with $m = 100$ by BIC. FL, full likelihood approach; PQL, penalized quasi-likelihood approach; AML, approximate marginal likelihood approach.

β	$\theta_1 = 1.96$			$\theta_2 = 1.00$			$\theta_3 = 0.81$			$\theta_4 = 0.64$		
	FL	PQL	AML	FL	PQL	AML	FL	PQL	AML	FL	PQL	AML
β_1	100	100	100	100	100	100	100	100	100	100	100	100
β_2	100	100	99	100	100	100	100	100	100	100	100	100
β_3	0	0	0	1	1	1	0	0	0	0	0	0
β_4	4	9	7	0	0	2	0	1	4	2	1	1
β_5	100	100	98	99	99	99	99	99	100	100	100	100
β_6	99	96	98	98	98	99	99	98	99	99	99	99
β_7	100	100	100	100	99	100	99	99	100	100	100	100
β_8	2	2	0	0	0	3	1	2	2	2	1	1
β_9	1	3	1	0	0	0	1	0	0	0	0	0
β_{10}	0	3	2	2	2	2	1	2	3	2	2	2
β_{11}	3	1	0	2	2	3	1	1	1	0	1	1
β_{12}	2	2	0	1	0	2	0	0	1	0	0	0
β_{13}	0	2	2	1	1	1	0	0	2	2	0	0
β_{14}	2	3	4	3	2	2	3	1	2	2	0	1
β_{15}	2	2	2	2	2	1	3	0	2	1	0	1
β_{16}	2	1	4	1	0	0	1	0	1	2	1	1

The above patterns and phenomena also exist in the frequency tables 5.4 to 5.7, where the selection frequency is reported for each coefficient β_k $(k = 1, \ldots, d)$ over 100 runs.

Figure 5.2 and Figure 5.3 are the plots of the average number of coefficients which are correctly set to zero for different values of θ, and Figure 5.4 and Figure 5.5 give plots of the average number of coefficients which are set to zero by mistake. All the

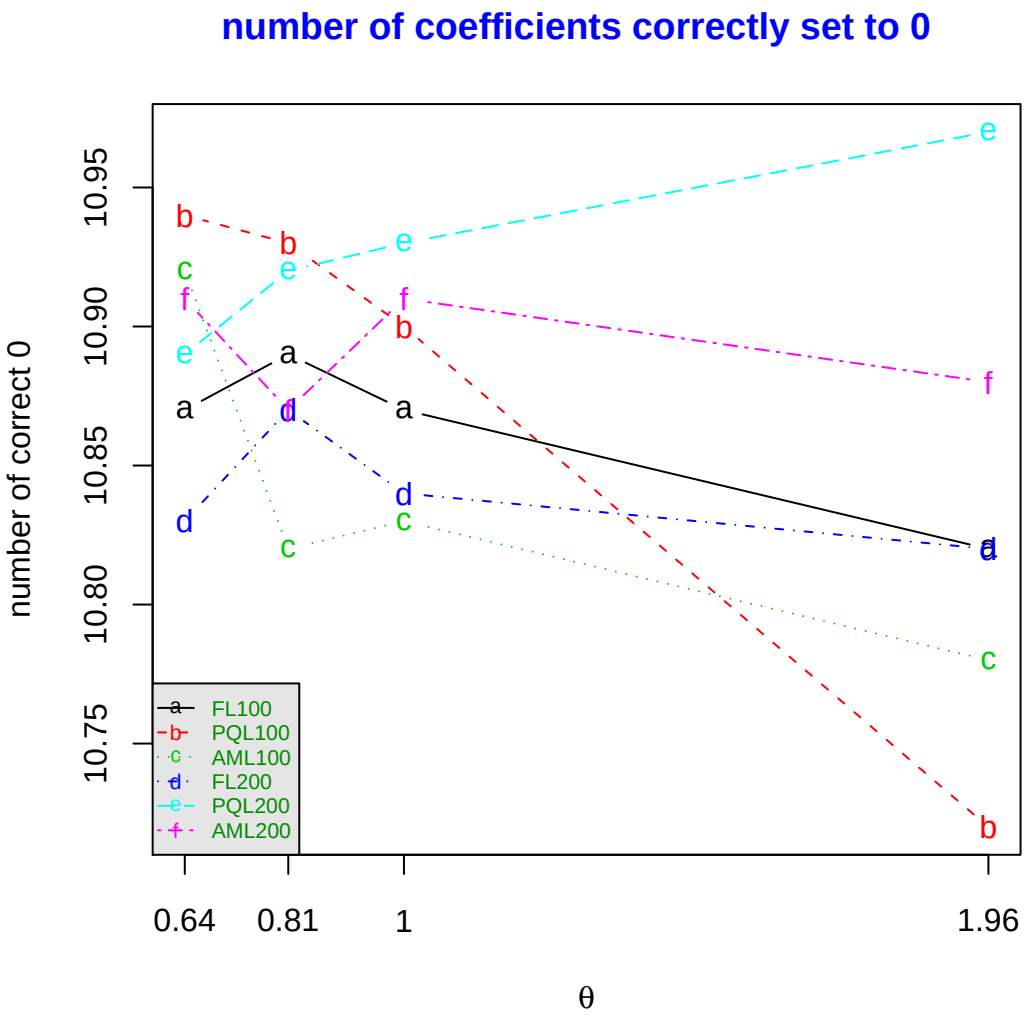

Figure 5.2: Average number of coefficients correctly set to 0 under scenario I: $(\beta_2, \beta_5, \beta_6, \beta_7) = (0.9, 0.9, 0.9, 0.9)$ by BIC. FL100, full likelihood approach with $m = 100$; PQL100, penalized quasi-likelihood approach with $m = 100$; AML100, approximate marginal likelihood approach with $m = 100$; FL200, full likelihood approach with $m = 200$; PQL200, penalized quasi-likelihood approach with $m = 200$; AML200, approximate marginal likelihood approach with $m = 200$

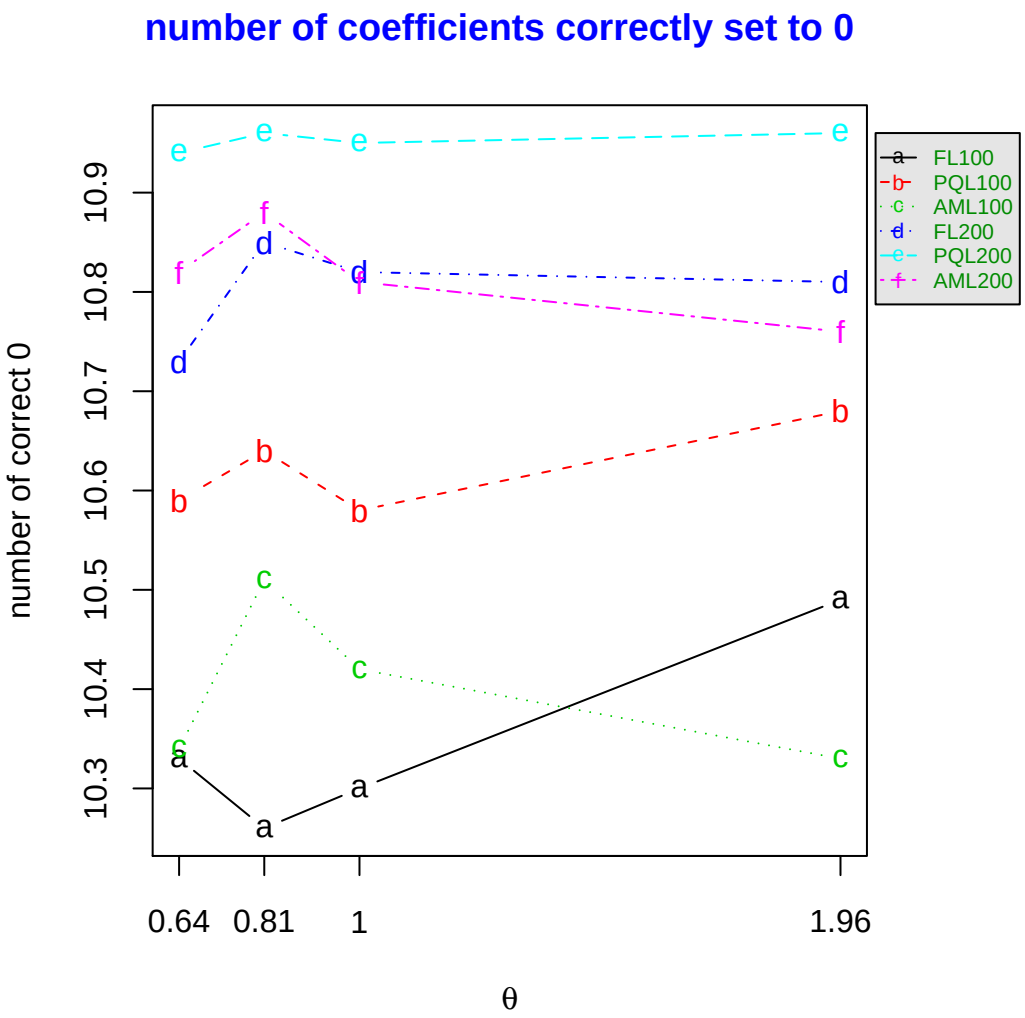

Figure 5.3: Average number of coefficients correctly set to 0 under scenario II: $(\beta_2, \beta_5, \beta_6, \beta_7) = (0.5, 0.5, 0.5, 0.5)$ by BIC. FL100, full likelihood approach with $m = 100$; PQL100, penalized quasi-likelihood approach with $m = 100$; AML100, approximate marginal likelihood approach with $m = 100$; FL200, full likelihood approach with $m = 200$; PQL200, penalized quasi-likelihood approach with $m = 200$; AML200, approximate marginal likelihood approach with $m = 200$

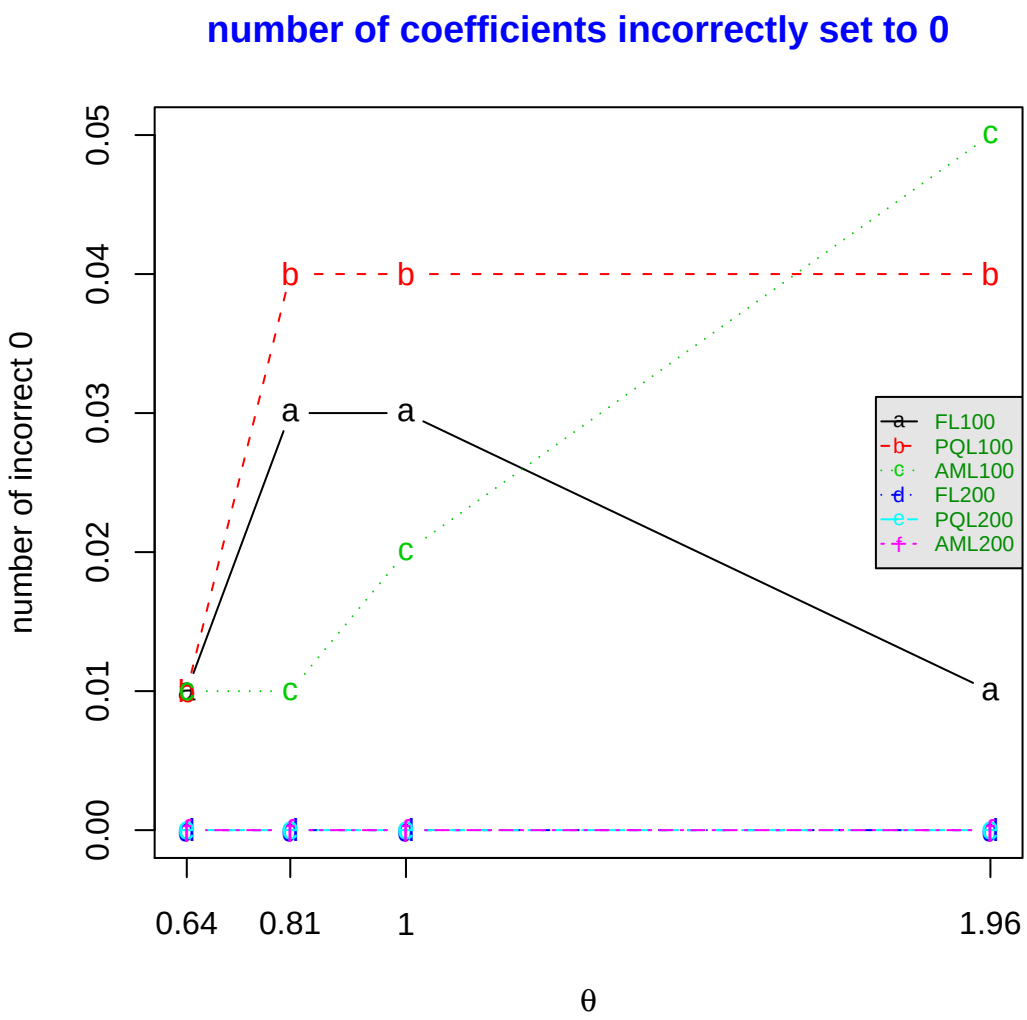

Figure 5.4: Average number of coefficients incorrectly set to 0 under scenario I: $(\beta_2, \beta_5, \beta_6, \beta_7) = (0.9, 0.9, 0.9, 0.9)$ by BIC. FL100, full likelihood approach with $m = 100$; PQL100, penalized quasi-likelihood approach with $m = 100$; AML100, approximate marginal likelihood approach with $m = 100$; FL200, full likelihood approach with $m = 200$; PQL200, penalized quasi-likelihood approach with $m = 200$; AML200, approximate marginal likelihood approach with $m = 200$

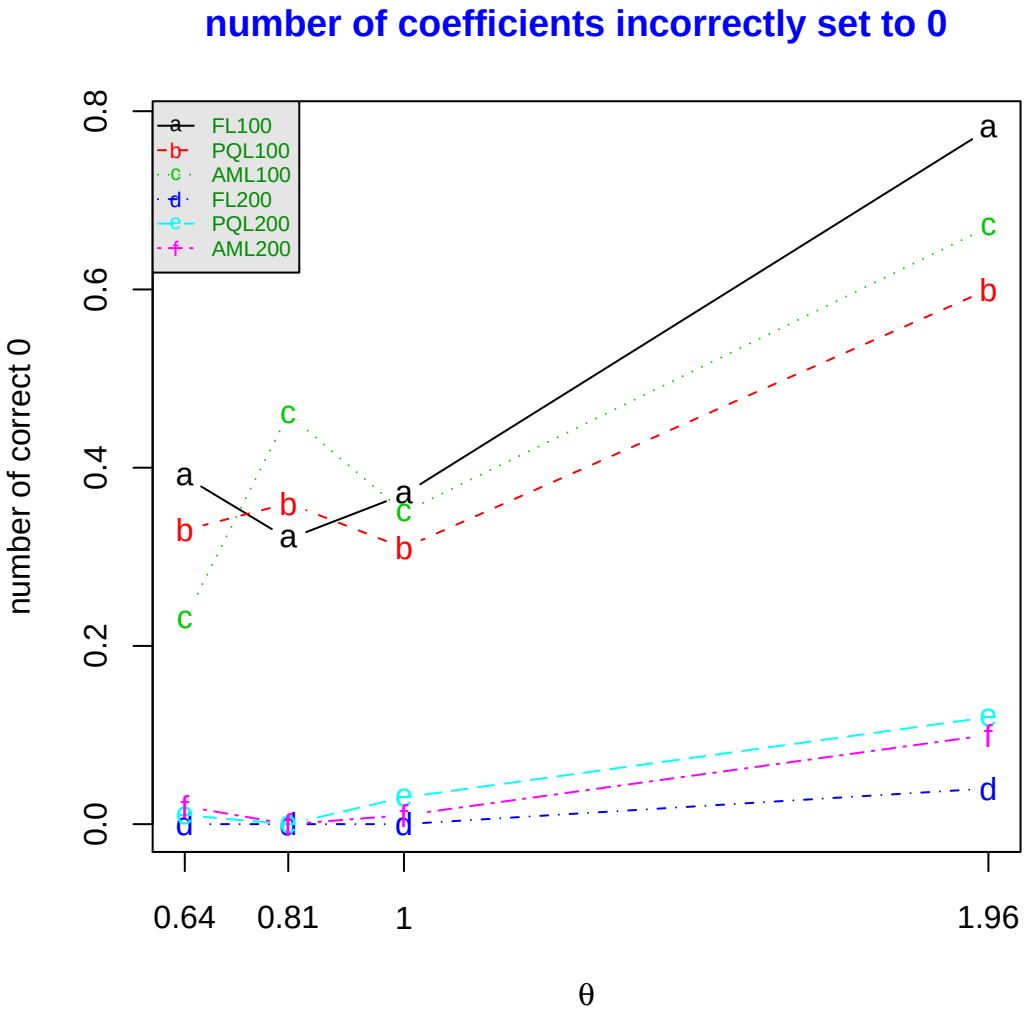

Figure 5.5: Average number of coefficients incorrectly set to 0 under scenario II: $(\beta_2, \beta_5, \beta_6, \beta_7) = (0.5, 0.5, 0.5, 0.5)$ by BIC. FL100, full likelihood approach with $m = 100$; PQL100, penalized quasi-likelihood approach with $m = 100$; AML100, approximate marginal likelihood approach with $m = 100$; FL200, full likelihood approach with $m = 200$; PQL200, penalized quasi-likelihood approach with $m = 200$; AML200, approximate marginal likelihood approach with $m = 200$

Table 5.5: Variable selection frequency at scenario I: $(\beta_2, \beta_5, \beta_6, \beta_7) = (0.9, 0.9, 0.9, 0.9)$ with $m = 200$ by BIC. FL, full likelihood approach; PQL, penalized quasi-likelihood approach; AML, approximate marginal likelihood approach.

β	$\theta_1 = 1.96$			$\theta_2 = 1.00$			$\theta_3 = 0.81$			$\theta_4 = 0.64$		
	FL	PQL	AML	FL	PQL	AML	FL	PQL	AML	FL	PQL	AML
β_1	100	100	100	100	100	100	100	100	100	100	100	100
β_2	100	100	100	100	100	100	100	100	100	100	100	100
β_3	0	0	0	0	0	0	0	0	0	1	0	0
β_4	0	0	0	4	1	1	2	2	3	2	2	2
β_5	100	100	100	100	100	100	100	100	100	100	100	100
β_6	100	100	100	100	100	100	100	100	100	100	100	100
β_7	100	100	100	100	100	100	100	100	100	100	100	100
β_8	2	0	2	0	0	0	0	0	0	0	0	0
β_9	0	0	0	0	0	0	0	0	0	0	0	0
β_{10}	3	1	1	0	0	0	0	0	0	1	1	1
β_{11}	1	0	1	1	1	1	1	1	2	3	2	1
β_{12}	0	0	0	1	1	1	1	1	3	1	1	2
β_{13}	3	0	1	3	1	3	3	1	2	3	3	2
β_{14}	3	0	2	3	1	1	2	2	2	3	1	1
β_{15}	3	1	2	2	2	2	4	1	1	3	1	0
β_{16}	3	1	3	2	0	0	0	0	0	0	0	0

figures clearly show the patterns we mentioned above.

To further investigate the effect of sample size to the performance of the proposed procedures in model selection, we also conducted simulation studies with the total number of subjects being only 50. Table 5.8 and Table 5.9 are the frequency tables for this very small size case. Table 5.10 is the corresponding summary of model selection results. When the signals of the important variables are strong (scenario I) and the variance of random effects is very large or very small, there are numerical problems

Table 5.6: Variable selection frequency at scenario II: $(\beta_2, \beta_5, \beta_6, \beta_7) = (0.5, 0.5, 0.5, 0.5)$ with $m = 100$ by BIC. FL, full likelihood approach; PQL, penalized quasi-likelihood approach; AML, approximate marginal likelihood approach.

β	$\theta_1 = 1.96$			$\theta_2 = 1.00$			$\theta_3 = 0.81$			$\theta_4 = 0.64$		
	FL	PQL	AML	FL	PQL	AML	FL	PQL	AML	FL	PQL	AML
β_1	100	100	100	100	100	100	100	100	100	100	100	100
β_2	83	83	84	94	94	93	93	91	91	94	92	96
β_3	7	5	7	1	4	4	2	3	5	5	5	8
β_4	7	6	9	7	7	5	6	5	5	6	5	5
β_5	82	87	86	97	93	94	95	92	91	88	92	92
β_6	77	82	80	85	91	91	88	89	85	90	89	94
β_7	80	88	83	87	91	87	92	92	87	89	94	95
β_8	5	5	7	7	6	4	7	5	3	4	2	5
β_9	4	4	7	5	5	4	7	5	7	4	4	4
β_{10}	3	2	6	6	5	6	5	2	3	6	3	6
β_{11}	4	2	5	8	4	5	9	6	4	4	5	7
β_{12}	5	3	6	7	2	6	8	3	7	6	6	10
β_{13}	4	1	6	9	2	9	9	1	4	7	3	6
β_{14}	2	0	3	3	1	1	6	1	2	8	1	3
β_{15}	8	2	8	12	5	8	9	3	4	9	4	8
β_{16}	2	2	3	5	1	6	6	2	5	8	3	4

involved in the model selection and parameter estimation process for some generated design matrixes. Therefore, over 100 simulation runs, there are only 72 valid runs when $\theta = 1.96$ and 93 valid runs when $\theta = 0.64$. A comparison of Table 5.10 to Table 5.2 and Table 5.3 shows that sample size does have much effect on the capability of the proposed procedures in model selection. In general, when the number of subjects is smaller, values of $size$ and $Corr.0$, the average number of coefficients which are correctly set to zeros, are smaller, and those of $Inc.0$, the average number of coefficients which are

Table 5.7: Variable selection frequency at scenario II: $(\beta_2, \beta_5, \beta_6, \beta_7) = (0.5, 0.5, 0.5, 0.5)$ with $m = 200$ by BIC. FL, full likelihood approach; PQL, penalized quasi-likelihood approach; AML, approximate marginal likelihood approach.

β	$\theta_1 = 1.96$			$\theta_2 = 1.00$			$\theta_3 = 0.81$			$\theta_4 = 0.64$		
	FL	PQL	AML	FL	PQL	AML	FL	PQL	AML	FL	PQL	AML
β_1	100	100	100	100	100	100	100	100	100	100	100	100
β_2	100	98	99	100	99	100	100	100	100	100	100	100
β_3	1	1	2	0	0	3	0	0	1	2	0	1
β_4	3	0	3	4	2	2	3	2	2	3	1	1
β_5	98	96	97	100	100	100	100	100	100	100	100	100
β_6	99	97	97	100	100	99	100	100	100	100	100	100
β_7	99	97	97	100	98	100	100	100	100	100	99	98
β_8	5	2	3	0	0	2	0	0	1	1	1	2
β_9	0	0	0	0	0	0	1	0	0	1	0	1
β_{10}	1	0	3	0	0	2	0	0	1	1	0	2
β_{11}	2	0	2	1	1	2	1	0	2	0	0	3
β_{12}	1	0	2	1	1	3	1	1	2	1	1	2
β_{13}	0	0	1	2	0	1	3	1	1	4	1	1
β_{14}	1	0	0	1	0	1	1	0	0	4	1	1
β_{15}	4	0	4	6	1	2	5	0	1	7	1	3
β_{16}	1	1	4	3	0	1	0	0	1	3	0	1

set to zeroes by mistake, are larger. Under scenario I, though the sample size is very small, the ranges of 10.13 to 10.83 for $Corr.0$ and 0.11 to 0.49 for $Inc.0$ show that the proposed procedures still work pretty well in model selection. The weak signals (under scenario II) and the small sample size (around 250 for 17 parameters) bring great negative effects to the performance of the proposed procedures, with smallest $Inc.0$ being 1.28.

Figures 5.3 and 5.5 show that the PQL procedure works differently from the FL

Table 5.8: Variable selection frequency at scenario I: $(\beta_2, \beta_5, \beta_6, \beta_7) = (0.9, 0.9, 0.9, 0.9)$ with $m = 50$ by BIC. FL, full likelihood approach; PQL, penalized quasi-likelihood approach; AML, approximate marginal likelihood approach.

β	$\theta_1 = 1.96$			$\theta_2 = 1.00$			$\theta_3 = 0.81$			$\theta_4 = 0.64$		
	FL	PQL	AML	FL	PQL	AML	FL	PQL	AML	FL	PQL	AML
β_1	72	72	72	100	100	100	100	100	100	93	93	93
β_2	68	66	66	98	95	95	100	98	98	93	91	91
β_3	8	3	4	6	3	3	11	5	6	8	3	4
β_4	8	5	5	8	5	5	14	7	6	10	7	8
β_5	66	58	58	97	97	97	97	96	95	91	89	89
β_6	68	61	60	95	93	92	92	90	89	87	87	88
β_7	70	70	69	97	92	94	96	96	95	91	91	91
β_8	4	1	1	6	2	4	9	2	2	5	1	2
β_9	4	0	0	2	1	1	6	2	2	4	3	2
β_{10}	2	1	1	5	3	3	5	2	2	4	1	2
β_{11}	3	0	0	7	2	4	7	3	3	6	3	3
β_{12}	1	1	1	6	1	3	8	2	2	6	3	3
β_{13}	1	1	1	5	1	3	8	1	1	5	1	2
β_{14}	1	0	0	3	0	0	7	1	1	6	1	2
β_{15}	4	0	0	3	1	2	4	0	0	5	0	0
β_{16}	0	0	0	1	0	0	8	2	2	4	1	0

and AML procedures in model selection. To statistically justify this impression, we did McNemar test for the equivalence of the PQL procedure to the FL and AML procedures in selecting the important variables. Table 5.11 gives p-values of McNemar test under scenario II when $m = 50$. With large p-values for large and small θ, we conclude the PQL procedure has equivalent performance to the FL and AML procedures in selecting the important variables when the variation of random effects is large or small. When θ is moderate, say around 1.00, the p-values are smaller than the significance level

Table 5.9: Variable selection frequency at scenario II: $(\beta_2, \beta_5, \beta_6, \beta_7) = (0.5, 0.5, 0.5, 0.5)$ with $m = 50$ by BIC. FL, full likelihood approach; PQL, penalized quasi-likelihood approach; AML, approximate marginal likelihood approach.

β	$\theta_1 = 1.96$			$\theta_2 = 1.00$			$\theta_3 = 0.81$			$\theta_4 = 0.64$		
	FL	PQL	AML	FL	PQL	AML	FL	PQL	AML	FL	PQL	AML
β_1	100	100	100	100	100	100	100	100	100	100	100	100
β_2	55	46	46	65	59	67	66	58	65	65	56	70
β_3	6	4	6	7	4	9	9	5	9	4	3	10
β_4	6	5	5	4	3	7	8	5	7	5	6	10
β_5	58	55	53	65	55	58	62	58	54	56	51	58
β_6	63	61	59	72	68	72	79	70	71	71	68	74
β_7	55	50	48	66	60	58	65	58	61	57	52	63
β_8	3	3	3	2	1	5	4	4	4	1	2	9
β_9	4	5	5	5	2	7	8	4	10	1	2	8
β_{10}	5	5	5	1	0	3	6	4	8	2	1	10
β_{11}	4	5	5	2	0	3	9	6	6	5	3	8
β_{12}	1	2	2	1	1	5	7	4	5	4	4	6
β_{13}	3	2	2	2	0	1	7	0	5	2	0	1
β_{14}	1	1	1	3	1	3	4	1	4	1	2	2
β_{15}	4	1	1	8	2	5	12	5	5	8	3	6
β_{16}	3	3	3	2	2	3	9	3	5	3	0	7

of 0.05, which indicates that the performance of the PQL procedure in selecting the important variables is significantly different from that of the FL and AML procedures for moderate θ.

One-sided paired t-tests using $Corr.0$ and $Inc.0$ were conducted to compare the PQL procedure to the FL and AML procedures when the variation of random effects is around 1. In consideration that *size* is not a clear representation of the selected model size, it is not included in the paired t-tests. With the information

Table 5.10: Model selection summary when $m = 50$ at scenario I & II by BIC. $size$, average model size; $Corr.0$, average number of coefficients which are set to 0 correctly; $Inc.0$, average number of coefficients which are set to 0 by mistake; FL, full likelihood approach; PQL, penalized quasi-likelihood approach; AML, approximate marginal likelihood approach.

	method	$\theta_1 = 1.96$			$\theta_2 = 1.00$			$\theta_3 = 0.81$			$\theta_4 = 0.64$		
		size	*Corr.*0	*Inc.*0	*size*	*Corr.*0	*Inc.*0	*size*	*Corr.*0	*Inc.*0	*size*	*Corr.*0	*Inc.*0
I	FL	4.28	10.50	0.22	4.39	10.48	0.31	4.72	10.13	0.15	4.57	10.32	0.11
		(1.00)	(0.90)	(0.79)	(0.96)	(0.90)	(0.34)	(1.82)	(1.75)	(0.36)	(1.39)	(1.33)	(0.31)
	PQL	3.71	10.83	0.46	3.96	10.81	0.23	4.07	10.73	0.20	4.11	10.74	0.15
		(0.75)	(0.50)	(0.64)	(0.69)	(0.48)	(0.44)	(0.65)	(0.58)	(0.42)	(0.63)	(0.53)	(0.39)
	AML	3.69	10.82	0.49	4.06	10.72	0.22	4.04	10.73	0.23	4.16	10.70	0.14
		(0.78)	(0.51)	(0.71)	(0.81)	(0.63)	(0.50)	(0.68)	(0.58)	(0.51)	(0.69)	(0.62)	(0.38)
II	FL	2.71	10.60	1.69	3.05	10.63	1.32	3.55	10.17	1.28	2.85	10.64	1.51
		(1.30)	(0.76)	(0.90)	(1.17)	(0.63)	(0.89)	(1.79)	(1.21)	(0.91)	(1.17)	(0.70)	(0.78)
	PQL	2.48	10.75	1.92	2.58	10.84	1.58	2.85	10.59	1.56	2.53	10.74	1.73
		(1.40)	(1.36)	(1.14)	(1.02)	(0.44)	(0.92)	(1.44)	(0.78)	(0.95)	(1.24)	(0.58)	(0.89)
	AML	2.44	10.73	1.98	3.06	10.60	1.49	3.19	10.43	1.53	3.42	10.23	1.35
		(1.39)	(1.36)	(1.11)	(1.57)	(1.46)	(1.18)	(1.56)	(1.52)	(1.07)	(1.72)	(1.11)	(0.86)

Table 5.11: P-values of McNemar's test for the equivalence of PQL to FL and PQL to AML in selecting all important variables at scenario II: $(\beta_2, \beta_5, \beta_6, \beta_7) = (0.5, 0.5, 0.5, 0.5)$ with $m = 50$ by BIC. PF, penalized quasi-likelihood approach vs. full likelihood approach; PA, penalized quasi-likelihood approach vs. approximate marginal likelihood approach.

$p - value$	$\theta_1 = 1.96$	$\theta_2 = 1.00$	$\theta_3 = 0.81$	$\theta_4 = 0.64$
PF	1.00	0.03	0.07	0.71
PA	0.13	0.01	0.41	0.20

from figures 5.3 and 5.5, for the one-sided t-test using $Corr.0$, the null hypothesis is $H_0 : Corr.0_1 = Corr.0_2$, and the alternative hypothesis is $H_\alpha : Corr.0_1 > Corr.0_2$. For the one-sided t-test using $Inc.0$, the null hypothesis is $H_0 : Inc.0_1 = Inc.0_2$, and the alternative hypothesis is $H_\alpha : Inc.0_1 > Inc.0_2$. Here 1 denotes the PQL procedure, and 2 denotes the FL or AML procedure. P-values are 0.00 for both $Corr.0$ and $Inc.0$ in the comparison of the PQL procedure to the FL procedure, and 0.05 and 0.22 for $Corr.0$ and $Inc.0$ respectively in the comparison of the PQL procedure to the AML procedure. Therefore, as regard to the performance of the proposed procedures in model selection when the variation of random effects is around 1, our conclusion is the PQL procedure has better performance than the FL and AML procedures in removing those non-important variables from models, but it is more possible for the important variables to be removed from models when using the PQL procedure instead of the FL procedure.

For the PQL procedure, in addition to BIC, we also developed two other tuning procedures for the choice of λ: GCV and REML. However, for the simulation design we chose, they are not good criteria for tuning λ. Table 5.12 reports the model selection results using GCV and REML criteria at scenario I: $(\beta_2, \beta_5, \beta_6, \beta_7) = (0.9, 0.9, 0.9, 0.9)$ when $m = 100$ and $\theta = 1.00$. Due to their low capability in tuning λ, we will not present the summary of model selection results for other settings and the parameter estimators based on them.

5.2.5 Simulation Results for Parameter Estimation

1. Results for β_1

Table 5.12: Model selection summary of PQL using GCV & REML criterion at scenario I: $(\beta_2, \beta_5, \beta_6, \beta_7) = (0.9, 0.9, 0.9, 0.9)$ with $m = 100$ and $\theta = 1.00$

GCV	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	$size = 5.99$
	100	96	19	14	97	97	97	23	$Corr.0 = 8.88$
	β_9	β_{10}	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	$\beta 16$	$Inc.0 = 0.03$
	22	22	18	24	17	18	18	17	
REML	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	$size = 0.00$
	100	0	0	0	0	0	0	0	$Corr.0 = 11.00$
	β_9	β_{10}	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	$\beta 16$	$Inc.0 = 1.00$
	0	0	0	0	0	0	0	0	

Table 5.13 and Table 5.14 present the estimation results for β_1 for the proposed procedures by the BIC tuning criterion. Here we use CPQL to denote the PQL approach after correction of bias in parameter estimation. It is easy to find that the approaches of FL, AML and TPQL have good performance in the estimation of β_1. In fact they work similarly in parameter estimation. As mentioned previously, the PQL approach under-estimates variance components, and often coefficients for binary data. As a reason, the coverage probabilities of the parameters are lower than the nominal level when using the PQL approach. The proposed factor for correcting bias in estimators improve its estimation capability, but not much, which indicates that the bias correction procedure may not have much effect on correcting the bias of coefficient estimators. In fact, the correction factors proposed by Breslow and Lin (1995, 1996) were pointed to the bias in variance component estimators.

For the selected simulation design, under the approaches of FL, AML and TPQL, the value of θ seems not to have much effect on the estimation of β_1. However,

Table 5.13: Inference on β_1 at scenario I: $(\beta_2, \beta_5, \beta_6, \beta_7) = (0.9, 0.9, 0.9, 0.9)$ by BIC. SD, the Monte Carlo median absolute deviation divided by 0.6745; SE, the mean of 100 estimated standard error by sandwich formula; SD_e, the empirical standard error estimate of the sandwich standard error estimate; CP, Monte Carlo coverage probability of 95% confidence interval; FL, full likelihood approach; PQL, penalized quasi-likelihood approach; $CPQL$, PQL after bias correction; AML, approximate marginal likelihood approach; $TPQL$, two-stage penalized quasi-likelihood approach

m	*method*	$\theta_1 = 1.96$					$\theta_2 = 1.00$				
		$\widehat{\beta}_1$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$	$\widehat{\beta}_1$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$
100	FL	0.52	0.21	0.21(0.03)	0.02	0.92	0.53	0.19	0.17(0.03)	0.03	0.94
	PQL	0.45	0.20	0.18(0.02)	−0.05	0.91	0.47	0.18	0.16(0.02)	−0.03	0.93
	$CPQL$	0.46	0.21	0.20(0.03)	−0.04	0.94	0.48	0.18	0.17(0.02)	−0.02	0.93
	AML	0.55	0.23	0.21(0.03)	0.05	0.88	0.53	0.20	0.17(0.03)	0.03	0.94
	$TPQL$	0.53	0.24	0.21(0.02)	0.03	0.90	0.52	0.19	0.17(0.03)	0.02	0.95
200	FL	0.52	0.17	0.15(0.02)	0.02	0.94	0.51	0.14	0.12(0.01)	0.01	0.95
	PQL	0.43	0.11	0.12(0.01)	−0.07	0.92	0.46	0.11	0.11(0.01)	−0.04	0.93
	$CPQL$	0.44	0.12	0.13(0.01)	−0.06	0.94	0.46	0.12	0.12(0.01)	−0.04	0.94
	AML	0.51	0.16	0.14(0.02)	0.01	0.92	0.52	0.13	0.12(0.01)	0.02	0.96
	$TPQL$	0.51	0.16	0.14(0.01)	0.01	0.95	0.51	0.12	0.12(0.01)	0.01	0.96
		$\theta_3 = 0.81$					$\theta_4 = 0.64$				
		$\widehat{\beta}_1$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$	$\widehat{\beta}_1$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$
100	FL	0.51	0.19	0.17(0.02)	0.01	0.96	0.52	0.18	0.16(0.02)	0.02	0.97
	PQL	0.48	0.16	0.15(0.01)	−0.02	0.96	0.48	0.17	0.15(0.01)	−0.02	0.97
	$CPQL$	0.48	0.16	0.16(0.02)	−0.02	0.96	0.49	0.17	0.16(0.01)	−0.01	0.97
	AML	0.53	0.19	0.17(0.02)	0.03	0.94	0.52	0.17	0.16(0.02)	0.02	0.96
	$TPQL$	0.52	0.18	0.17(0.02)	0.02	0.97	0.52	0.18	0.16(0.02)	0.02	0.97
200	FL	0.52	0.13	0.12(0.01)	0.02	0.96	0.52	0.13	0.11(0.01)	0.02	0.96
	PQL	0.47	0.12	0.11(0.01)	−0.03	0.96	0.48	0.11	0.10(0.01)	−0.02	0.94
	$CPQL$	0.47	0.12	0.11(0.01)	−0.03	0.96	0.48	0.11	0.11(0.01)	−0.02	0.95
	AML	0.52	0.13	0.12(0.01)	0.02	0.97	0.52	0.13	0.11(0.01)	0.02	0.97
	$TPQL$	0.51	0.13	0.12(0.01)	0.01	0.96	0.52	0.13	0.11(0.01)	0.02	0.95

the absolute biases of the estimators from PQL and CPQL are positively related to the value of θ. In other words, smaller θ's yield better estimators of β_1. This phenomenon is intuitive and also can be expected from the values of R^2. When random effects have a smaller variance, the observations tend to be less correlated. Hence one can expect the proposed procedures to have better performance.

Table 5.14: Inference on β_1 at scenario II: $(\beta_2, \beta_5, \beta_6, \beta_7) = (0.5, 0.5, 0.5, 0.5)$ by BIC. SD, the Monte Carlo median absolute deviation divided by 0.6745; SE, the mean of 100 estimated standard error by sandwich formula; SD_e, the empirical standard error estimate of the sandwich standard error estimate; CP, Monte Carlo coverage probability of 95% confidence interval; FL, full likelihood approach; PQL, penalized quasi-likelihood approach; $CPQL$, PQL after bias correction; AML, approximate marginal likelihood approach; $TPQL$, two-stage penalized quasi-likelihood approach

m	*method*	$\theta_1 = 1.96$					$\theta_2 = 1.00$				
		$\widehat{\beta}_1$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$	$\widehat{\beta}_1$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$
100	FL	0.53	0.23	0.20(0.03)	0.03	0.92	0.53	0.20	0.17(0.03)	0.03	0.93
	PQL	0.44	0.17	0.17(0.02)	−0.06	0.90	0.47	0.17	0.15(0.02)	−0.03	0.93
	$CPQL$	0.45	0.18	0.18(0.02)	−0.05	0.91	0.47	0.17	0.16(0.02)	−0.03	0.94
	AML	0.53	0.23	0.20(0.03)	0.03	0.87	0.52	0.19	0.17(0.03)	0.02	0.90
	$TPQL$	0.53	0.22	0.20(0.03)	0.03	0.93	0.52	0.18	0.16(0.02)	0.02	0.94
200	FL	0.50	0.14	0.14(0.01)	0.00	0.94	0.51	0.11	0.11(0.01)	0.01	0.92
	PQL	0.42	0.12	0.12(0.01)	−0.08	0.86	0.46	0.09	0.10(0.01)	−0.04	0.92
	$CPQL$	0.43	0.12	0.12(0.01)	−0.07	0.89	0.46	0.09	0.11(0.01)	−0.04	0.95
	AML	0.50	0.15	0.14(0.02)	0.00	0.92	0.51	0.11	0.11(0.01)	0.01	0.93
	$TPQL$	0.50	0.13	0.14(0.01)	0.00	0.92	0.51	0.10	0.11(0.01)	0.01	0.93
		$\theta_3 = 0.81$					$\theta_4 = 0.64$				
		$\widehat{\beta}_1$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$	$\widehat{\beta}_1$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$
100	FL	0.53	0.19	0.16(0.03)	0.03	0.95	0.54	0.16	0.15(0.03)	0.04	0.93
	PQL	0.48	0.16	0.15(0.02)	−0.02	0.94	0.50	0.15	0.14(0.02)	0.00	0.94
	$CPQL$	0.49	0.16	0.15(0.02)	−0.01	0.95	0.50	0.14	0.14(0.02)	0.00	0.94
	AML	0.53	0.20	0.16(0.02)	0.03	0.93	0.54	0.17	0.15(0.03)	0.04	0.92
	$TPQL$	0.53	0.18	0.15(0.02)	0.03	0.95	0.54	0.16	0.15(0.02)	0.04	0.92
200	FL	0.51	0.11	0.11(0.01)	0.01	0.94	0.52	0.09	0.10(0.01)	0.02	0.91
	PQL	0.46	0.09	0.10(0.01)	−0.04	0.93	0.48	0.08	0.09(0.01)	−0.02	0.93
	$CPQL$	0.47	0.09	0.10(0.01)	−0.03	0.94	0.48	0.08	0.10(0.01)	−0.02	0.94
	AML	0.51	0.11	0.11(0.01)	0.01	0.94	0.51	0.09	0.10(0.01)	0.01	0.90
	$TPQL$	0.51	0.11	0.11(0.01)	0.01	0.94	0.52	0.09	0.10(0.01)	0.02	0.92

With a larger sample size, the estimation capability of the approaches of FL, AML and TPQL is better. But under the chosen settings, sample size does not have much effect on the performance of PQL and CPQL in parameter estimation.

For the PQL approach, when the coefficients are small, the biases of coefficient estimators tend to be small. The comparison of Table 5.13 to Table 5.14 shows

that the bias of β_1 is reduced at scenario II for the PQL and CPQL procedures. The intuition is that the performance of the PQL and CPQL procedures should become poor when problems get more challenging. An examination of other estimation tables shows that this phenomenon also exists in the estimation of other coefficients. In fact, under the PQL approach, smaller coefficients are associated with smaller biases. This phenomenon was explained by Lin and Breslow (1995).

The good performance of the proposed sandwich estimators of standard errors for β_1 is also demonstrated in Table 5.13 and Table 5.14. For the procedures of FL, PQL and CPQL, SE are the estimated standard errors by the sandwich formula; for the procedures of AML and TPQL, SE are standard error estimators given by the inverse of the observed Fisher information matrix (reported by *nlmixed* procedure of *SAS*). As shown by the tables, the sandwich formula of $\widehat{\beta}_1$ works very well, with the largest difference of the SE from the SD being only 0.03.

Similar patterns in estimation can be found for other coefficients. Their estimation results are given in Appendix B.

2. Results for θ

Table 5.15 and Table 5.16 contain the inference results for θ under two scenarios. It is easy to see that for the estimation of θ, the procedures of FL, AML and TPQL have patterns similar to those for the coefficients. Sample size has some effect on their performance. The proposed procedures tend to yield better estimators with larger sample sizes.

Table 5.15: Inference on θ at Scenario I: $(\beta_2, \beta_5, \beta_6, \beta_7) = (0.9, 0.9, 0.9, 0.9)$ by BIC. SD, the Monte Carlo median absolute deviation divided by 0.6745; SE, the mean of 100 estimated standard error by sandwich formula; SD_e, the empirical standard error estimate of the sandwich standard error estimate; CP, Monte Carlo coverage probability of 95% confidence interval; FL, full likelihood approach; PQL, penalized quasi-likelihood approach; $CPQL$, PQL after bias correction; AML, approximate marginal likelihood approach; $TPQL$, two-stage penalized quasi-likelihood approach

m	*method*	$\theta_1 = 1.96$					$\theta_2 = 1.00$				
		$\widehat{\theta}$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$	$\widehat{\theta}$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$
100	FL	2.06	0.68	0.81(0.33)	0.10	0.94	0.99	0.45	0.52(0.22)	0.01	0.91
	PQL	1.47	0.45	0.46(0.16)	−0.49	0.75	0.77	0.32	0.34(0.12)	−0.23	0.86
	$CPQL$	1.93	0.59	0.60(0.21)	−0.03	0.94	1.02	0.40	0.45(0.16)	0.02	0.93
	AML	1.98	0.64	0.77(0.32)	0.02	0.94	0.99	0.45	0.52(0.22)	−0.01	0.90
	$TPQL$	2.01	0.63	0.77(0.32)	0.05	0.94	1.00	0.45	0.52(0.23)	0.00	0.91
200	FL	2.00	0.51	0.54(0.19)	0.04	0.92	0.99	0.30	0.35(0.13)	−0.01	0.92
	PQL	1.36	0.31	0.30(0.10)	−0.60	0.45	0.73	0.20	0.23(0.08)	−0.27	0.76
	$CPQL$	1.78	0.39	0.39(0.13)	−0.18	0.85	0.95	0.26	0.30(0.10)	−0.05	0.90
	AML	1.93	0.49	0.51(0.18)	−0.03	0.92	0.99	0.29	0.35(0.13)	−0.01	0.92
	$TPQL$	1.93	0.49	0.51(0.18)	−0.03	0.92	0.99	0.29	0.35(0.13)	−0.01	0.92
		$\theta_3 = 0.81$					$\theta_4 = 0.64$				
		$\widehat{\theta}$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$	$\widehat{\theta}$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$
100	FL	0.80	0.34	0.46(0.16)	0.01	0.93	0.65	0.38	0.42(0.15)	0.01	0.95
	PQL	0.65	0.27	0.32(0.09)	−0.16	0.88	0.54	0.26	0.30(0.09)	−0.10	0.94
	$CPQL$	0.85	0.36	0.42(0.12)	0.04	0.99	0.71	0.34	0.40(0.12)	0.07	1.00
	AML	0.80	0.37	0.47(0.17)	−0.01	0.93	0.65	0.38	0.43(0.16)	0.01	0.95
	$TPQL$	0.80	0.37	0.47(0.17)	−0.01	0.93	0.65	0.38	0.43(0.16)	0.01	0.95
200	FL	0.82	0.29	0.32(0.12)	0.01	0.92	0.65	0.26	0.29(0.11)	0.01	0.92
	PQL	0.61	0.20	0.21(0.07)	−0.20	0.79	0.50	0.18	0.20(0.07)	−0.14	0.81
	$CPQL$	0.80	0.26	0.28(0.10)	−0.01	0.94	0.65	0.24	0.27(0.09)	0.01	0.94
	AML	0.81	0.30	0.32(0.12)	0.00	0.92	0.65	0.28	0.29(0.11)	0.01	0.92
	$TPQL$	0.81	0.29	0.32(0.12)	0.00	0.92	0.65	0.27	0.29(0.11)	0.01	0.92

The estimators of θ from the PQL approach are seriously biased under both scenarios. However, after bias correction, the biases are greatly reduced and the corrected estimators are very close to the true values.

Table 5.15 and Table 5.16 also demonstrate the good performance of our proposed sandwich formula when the true value of θ is not very large. When the variation

Table 5.16: Inference on θ at Scenario II: $(\beta_2, \beta_5, \beta_6, \beta_7) = (0.5, 0.5, 0.5, 0.5)$ by BIC. SD, the Monte Carlo median absolute deviation divided by 0.6745; SE, the mean of 100 estimated standard error by sandwich formula; SD_e, the empirical standard error estimate of the sandwich standard error estimate; CP, Monte Carlo coverage probability of 95% confidence interval; FL, full likelihood approach; PQL, penalized quasi-likelihood approach; $CPQL$, PQL after bias correction; AML, approximate marginal likelihood approach; $TPQL$, two-stage penalized quasi-likelihood approach

m	*method*	$\theta_1 = 1.96$					$\theta_2 = 1.00$				
		$\widehat{\theta}$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$	$\widehat{\theta}$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$
100	FL	2.08	0.54	0.71(0.30)	0.12	0.93	1.05	0.42	0.44(0.18)	0.05	0.90
	PQL	1.45	0.35	0.40(0.15)	−0.51	0.67	0.80	0.26	0.29(0.11)	−0.20	0.82
	$CPQL$	1.83	0.45	0.51(0.19)	−0.13	0.90	1.01	0.34	0.37(0.14)	0.01	0.94
	AML	1.97	0.46	0.66(0.27)	0.01	0.92	1.01	0.38	0.43(0.17)	0.01	0.91
	$TPQL$	2.00	0.52	0.66(0.27)	0.04	0.93	1.02	0.38	0.43(0.17)	0.02	0.91
200	FL	2.07	0.46	0.49(0.18)	0.11	0.95	1.03	0.27	0.30(0.12)	0.03	0.91
	PQL	1.39	0.26	0.27(0.09)	−0.57	0.39	0.76	0.17	0.20(0.07)	−0.24	0.65
	$CPQL$	1.74	0.33	0.34(0.12)	−0.22	0.84	0.95	0.23	0.25(0.09)	−0.05	0.90
	AML	1.97	0.43	0.45(0.17)	0.01	0.92	1.03	0.27	0.30(0.12)	0.03	0.90
	$TPQL$	1.98	0.45	0.45(0.17)	0.02	0.92	1.03	0.26	0.30(0.12)	0.03	0.90
		$\theta_3 = 0.81$					$\theta_4 = 0.64$				
		$\widehat{\theta}$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$	$\widehat{\theta}$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$
100	FL	0.84	0.36	0.39(0.16)	0.03	0.92	0.67	0.30	0.35(0.14)	0.03	0.94
	PQL	0.66	0.24	0.27(0.10)	−0.15	0.86	0.54	0.21	0.25(0.09)	−0.10	0.91
	$CPQL$	0.83	0.31	0.34(0.12)	0.02	0.92	0.69	0.26	0.31(0.11)	0.05	0.96
	AML	0.80	0.34	0.38(0.15)	−0.01	0.92	0.66	0.29	0.35(0.14)	0.02	0.93
	$TPQL$	0.83	0.35	0.39(0.16)	0.02	0.91	0.66	0.30	0.35(0.14)	0.02	0.94
200	FL	0.82	0.24	0.27(0.10)	0.01	0.93	0.66	0.21	0.24(0.09)	0.02	0.92
	PQL	0.62	0.16	0.18(0.06)	−0.19	0.75	0.51	0.14	0.17(0.06)	−0.13	0.84
	$CPQL$	0.78	0.20	0.23(0.08)	−0.03	0.93	0.64	0.19	0.21(0.07)	0.00	0.94
	AML	0.82	0.24	0.27(0.10)	0.01	0.93	0.66	0.20	0.24(0.09)	0.02	0.91
	$TPQL$	0.82	0.24	0.27(0.10)	0.01	0.93	0.66	0.21	0.24(0.09)	0.02	0.91

of random effects is larger, there is larger difference between the empirical standard deviation estimators and the estimated standard errors by the sandwich formula. In all, our sandwich formula estimators of the standard errors are reasonable reflection of the variation of the estimated parameters. The 95% coverage probabilities of the true values using the SE further demonstrate its good performance.

5.3 Real Data Analysis

We illustrate the proposed procedures by applying them to the analysis of the longitudinal data on respiratory infection in Indonesian children (reported by by Zeger and Karim, 1991), where 275 preschool children were followed for six consecutive quarters and they were examined every quarter for the presence of respiratory infection ($1 \equiv$ yes; $0 \equiv$ no). The study consisted of 1200 binary outcomes, and the covariates of interest include: age in months (centered at 36 months); presence of xerophthalmia ($1 \equiv$ yes; $0 \equiv$ no), an ocular manifestation of chronic vitamin A deficiency; cosine and sine terms for the annual cycle; gender ($1 \equiv$ female; $0 \equiv$ male); height for age, as a percent of the National Center for Health Statistics (NCHS) standard (centered at 90%), which indicates longer-term nutritional status; and the stunting status ($1 \equiv$ yes; $0 \equiv$ no). See Zeger and Karim (1991) for a detailed description of the study. One of the objectives of the study was to identify those independent variables which had a strong impact on respiratory infection.

We use Y_i to denote the respiratory infection ($0 \equiv$no; $1 \equiv$yes) for child i, X_{2i} for the age (in months), X_{3i} for the presence/absence of xerophthalmia ($0 \equiv$no; $1 \equiv$yes), X_{4i} for the cosine terms for annual cycle, X_{5i} for the sine terms for annual cycle, X_{6i} for the sex ($0 \equiv$male; $1 \equiv$female), X_{7i} for the height and X_{8i} for the presence of stunting ($0 \equiv$no; $1 \equiv$yes), where $i = 1, \ldots, 275$, $n = \sum_{i=1}^{275} n_i = 1200$. To address the above research question, we consider the following GLMM:

$$\begin{aligned} logit\{Pr(Y_i = 1|b_i)\} &= \beta_1 + X_{2i}\beta_2 + X_{3i}\beta_3 + X_{4i}\beta_4 + X_{5i}\beta_5 + X_{6i}\beta_6 \\ &\quad + X_{7i}\beta_7 + X_{8i}\beta_8 + b_i, \end{aligned} \tag{5.11}$$

where the random effects $b_i \sim N(0, \theta)$.

Table 5.17: Variable Selection and Parameter Estimation for Infectious Disease Data

method	*int*		*age*		*cos*		*height*		θ	
	$\widehat{\beta}$	$se(\widehat{\beta})$	$\widehat{\beta}$	$se(\widehat{\beta})$	$\widehat{\beta}$	$se(\widehat{\beta})$	$\widehat{\beta}$	$se(\widehat{\beta})$	$\widehat{\theta}$	$se(\widehat{\theta})$
FL	−2.76	0.19	−0.03	0.01	−0.55	0.17	−0.05	0.02	0.75	0.37
PQL	−2.51	0.13	−0.03	0.01	−0.52	0.16	−0.05	0.02	0.54	0.28
CPQL	−2.52	0.14	−0.03	0.01	−0.52	0.16	−0.05	0.02	0.69	0.36
AML	−2.76	0.19	−0.03	0.01	−0.55	0.17	−0.05	0.02	0.76	0.37
TPQL	−2.76	0.19	−0.03	0.01	−0.55	0.17	−0.05	0.02	0.76	0.37

Table 5.17 lists the parameter estimators for the important variables and their associated standard error estimators under the proposed procedures. As shown in this table, all of the procedures choose the same important covariates. The FL, AML and TPQL methods produce the same parameter estimators. Usually the PQL approach under-estimates parameters for binary data. In this particular case, the coefficient estimators of age and height using the PQL approach are pretty close to those obtained by other approaches. But the variance component estimator $\widehat{\theta}$ by the PQL approach is much smaller. After bias correction, the variance component estimator is much closer to the estimators from the procedures of FL, AML and TPQL. As our previous simulation studies showed that, the variance component estimator after bias correction is often associated with smaller bias.

The results indicate that respiratory infection is strongly related to age, season and height. The chance of having the respiratory infection decrease with age by approximately 3% per month for children from one to five years. The xerophthalmia coefficient provides no evidence for the effect of vitamin A deficiency on respiratory infection. Sex

does not have effect on the respiratory infection. The odds of disease for a child at low percentage of the NCHS height-for-age standard is bigger than that for a child at at high percentage level.

5.4 Summary

In summary, the proposed procedures have good performance in the analysis of high-dimensional longitudinal data with non-normal responses. We are aware that some factors affecting the proposed procedures are: the variance of subject to subject variation and the sample size.

Based on our numerical studies, our suggestions for the choice of model selection procedures in GLMMs are:

1. For non-sparse data (count data or binomial data with moderate to large denominators), use the PQL or the TPQL approach;

2. For binary data, in generally, use the TPQL approach. If there is only random intercept, the AML approach may also be used.

The procedures are implemented using *SAS* on a standard laptop. For 100 simulation runs, the running time for the FL approach is around 30 hours when $m = 100$ and 72 hours when $m = 200$, but for the approaches of PQL, AML and TPQL it is only around one hour when $m = 100$ and around 3 hours when $m = 200$. The running time suggests that the latter ones are much more computationally efficient than the FL approach. This is what we expect, because they avoid the integration involved in model selection process. *SAS* macro code is available from the author.

Chapter 6

Future Work: Variable Selection Procedure in GSMMs

6.1 Motivation

In longitudinal studies, observations are often correlated and non-normal. A common method for analyzing these types of data is using GLMMs. However, the parametric assumption in GLMMs may not always be appropriate, because the time effect often changes in a complicated way. Therefore, it is hard to model the trend with a simple parametric form. It is hence of potential interest to model the time effect non-parametrically while taking into account the correlation between observations. For this chapter, we propose generalized semi-parametric mixed models (GSMMs) for correlated non-normal responses in longitudinal studies. This model employs parametric fixed effects to represent the covariate effects and an arbitrary smoothing function to model the time effect.

The PQL approach with SCAD penalty is extended to GSMMs for model selection in high-dimensional longitudinal data analysis. Jointly with a linear transformation on the smoothing function and local quadratic approximation on the SCAD penalty, the PQL approach allows the simultaneous model selection and parameter estimation in a GSMM to proceed in the framework of a GLMM.

In the following sections, we will describe GSMMs and develop the PQL approach for model selection in GSMMs.

6.2 Generalized Semi-parametric Mixed Models

For a sample with m subjects and each subject with n_i observations over time, let $(Y_{ij}, X_{ij}, t_{ij}, Z_{ij})$ be one observation for subject i at time point t_{ij}, where Y_{ij} is the response, X_{ij} is a covariate vector associated with a $d \times 1$ vector of fixed effects β, and Z_{ij} is associated with random effects. Also let $f(t)$ be an arbitrary smooth function of the scalar covariate t. Given a vector of $q \times 1$ random effects b_i, Y_{ij} $(j = 1, 2, \ldots, n_i)$ are assumed to be conditionally independent with mean $E(Y_{ij}|b_i) = \mu_{ij}^b$ and variance $Var(Y_{ij}|b_i) = \phi w_{ij}^{-1} \nu(\mu_{ij}^b)$, where $\nu(\cdot)$ is a specified variance function, w_{ij} is a prior weight and ϕ is a scale parameter. The conditional mean μ_{ij}^b is related to the linear predictors and nonparametric function via a monotonic differentiable link function g:

$$g(\mu_{ij}^b) = X_{ij}\beta + f(t_{ij}) + Z_{ij}b_i. \tag{6.1}$$

Further we assume that b_i $(i = 1, 2, \ldots, m)$ are independent normal variables from distribution $N(0, D(\theta))$, where D is a positive definite matrix depending on a parameter vector θ. The scale parameter ϕ may or may not be know. In the examples we consider, the dispersion parameter ϕ is fixed at unity. In the cases where ϕ is unknown, it may be estimated together with θ as a parameter in the covariance matrix of the marginal distribution of outcome variables.

We follow Zhang *et al.* (1998) for matrix notation. Define $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})^T$,

and similarly define X_i, Z_i, μ_i^b $(i = 1, 2, \ldots, m)$. Denote $t^0 = (t_1^0, t_2^0, \ldots, t_r^0)^T$ to be a vector of distinct time points in order t_{ij} $(i = 1, 2, \ldots, m, j = 1, 2, \ldots, n_i)$, $f = (f(t_1^0), f(t_2^0), \ldots, f(t_r^0))^T$, N_i is the incidence matrix for subject i which is defined in a way such that $N_{jl} = 1$ if $t_{ij} = t_l^0$, and $N_{jl} = 0$, otherwise $(j = 1, 2, \ldots, n_i, l = 1, 2, \ldots, r)$. Then GSMM (6.1) can be written as

$$g(\mu_i^b) = X_i\beta + N_i f + Z_i b_i, \tag{6.2}$$

with $E(Y_i|b_i) = \mu_i^b$ and $Var(Y_i|b_i) = \phi w_{ij}^{-1}\nu(\mu_i^b)$.

The integrated quasi-likelihood function for the estimation of β, f and θ is defined as (Breslow and Clayton, 1993)

$$e^{\ell(\beta,f,\theta;Y)} \propto |G|^{-\frac{1}{2}} \int \exp\left\{-\frac{1}{2\phi}\sum_{i=1}^{m}\sum_{j=1}^{n_i} d_{ij}(Y_{ij};\mu_{ij}^b) - \frac{1}{2}b^T G^{-1} b\right\} db, \tag{6.3}$$

where Y, b and G are defined as in Chapter 1, $d_{ij}(Y_{ij};\mu_{ij}^b) = -2\int_{Y_{ij}}^{\mu_{ij}^b} \frac{w_{ij}(Y_{ij}-u)}{\nu(u)} du$ represents the conditional deviance function of $(\beta, f(\cdot))$ given b_i.

The roughness penalty approach is employed for the estimation of the nonparametric function f. Naturally, for GSMM (6.2), its ML estimator $(\widehat{\beta}, \widehat{f})$ can be obtained by maximizing

$$\ell(\beta, f, \theta; Y) - \frac{\lambda_1}{2}\int_{T_1}^{T_2} [f''(t)]^2 dt, \tag{6.4}$$

where λ_1 is a smoothing parameter which controls the trade-off between the goodness-of-fit and the roughness of the estimated $f(\cdot)$.

In consideration that we are doing model selection in GSMMs with the shrinkage penalty approach, our objective function is

$$\ell_d(\beta, f, \theta; Y) = \ell(\beta, f, \theta; Y) - \frac{\lambda_1}{2}\int_{T_1}^{T_2}[f''(t)]^2 dt - n\sum_{j=1}^{d} p_\lambda(|\beta_j|), \tag{6.5}$$

where $p_\lambda(\cdot)$ is the SCAD penalty function defined in (1.4) and λ is a tuning parameter.

The above objective function is termed as double penalized quasi-likelihood (DPQL), in that we penalize the quasi-likelihood by two penalties, one penalty for the nonparametric function, and the other for coefficients. By maximizing the DPQL , one would expect to see the coefficients of non-important variables to be shrunk all-the-way to zeroes, and the remaining coefficients, the nonparametric function and variance components to be solved by equating the score functions of the DPQL to zeroes.

6.3 Generalized Linear Mixed Model Representation

6.3.1 Triple Penalized Quasi-Likelihood

Considering the cumbersome and often intractable integration involved in maximizing the DPQL (6.5), we use the Laplace method to approximate ℓ_d (Breslow and Clayton, 1993). By ignoring the dependence of the weight matrix

$$W = diag\left\{w_{ij}/[\phi\nu(\mu_{ij}^b)(g'(\mu_{ij}^b))^2]\right\}$$

on the conditional means μ_{ij}^b, it can be shown that, given θ, λ_1 and λ, $(\widehat{\beta}, \widehat{f})$ of the resultant approximation to the DPQL (6.5) maximizes following triple penalized quasi-likelihood (TPQL) (Lin and Zhang, 1999)

$$\begin{aligned} \ell_t(\beta, f, \theta; Y) &= -\frac{1}{2\phi}\sum_{i=1}^{m}\sum_{j=1}^{n_i} d_{ij}(Y_{ij}; \mu_{ij}^b) - \frac{1}{2}b^T G^{-1} b - \frac{\lambda_1}{2}\int_{T_1}^{T_2}[f''(t)]^2 dt \\ &\quad -n\sum_{j=1}^{d} p_\lambda(|\beta_j|). \end{aligned} \tag{6.6}$$

In following sections we will show the maximization of the TPQL (6.6) can be the estimation issue in a working GLMM through some transformations and approximations.

6.3.2 Linear Transformation of f

We use natural cubic spline to estimate the infinite dimensional nonparametric function $f(\cdot)$. As shown by Green and Silverman (1994), the TPQL can be written as

$$\ell_t(\beta, f, \theta; Y) = -\frac{1}{2\phi}\sum_{i=1}^{m}\sum_{j=1}^{n_i} d_{ij}(Y_{ij}; \mu_{ij}^b) - \frac{1}{2}b^T G^{-1} b - \frac{\lambda}{2}f^T K f - n\sum_{j=1}^{d} p_\lambda(|\beta_j|), \tag{6.7}$$

where K is the nonnegative definite smoothing matrix as defined in equation (2.3) of Green and Silverman (1994).

Green (1987) showed that f could be one-to-one transformed linearly as

$$f = T\delta + Bc, \tag{6.8}$$

where δ and c are vectors with length 2 and $r-2$ respectively, $T=(1,t^0)$ and 1 is a $r\times 1$ vector of 1. Let L be $r\times(r-2)$ full rank matrix and satisfy $K=LL^T$ and $L^TT=0$, then $B=L(L^TL)^{-1}$, and $f^TKf=c^Tc$. Plugging it in (6.7) we have

$$\ell_t(\beta,f,\theta;Y)=-\frac{1}{2\phi}\sum_{i=1}^{m}\sum_{j=1}^{n_i}d_{ij}(Y_{ij};\mu_{ij}^b)-\frac{1}{2}b^TG^{-1}b-\frac{1}{2\tau}c^Tc-n\sum_{j=1}^{d}p_\lambda(|\beta_j|), \quad (6.9)$$

where $\tau=1/\lambda_1$.

6.3.3 Generalized Linear Mixed Model Representation

The TPQL (6.9) can be approximated by a local quadratic function using the previous approximation approach. Let $(\widehat{\beta}^{(0)},\widehat{\theta}^{(0)})$ be an initial value which is close to the maximizer of the TPQL (6.9). Then the TPQL (6.9) can be approximated by the local quadratic function

$$\begin{aligned}\tilde{\ell}_{tpql}(\beta,f,\theta;Y) &= -\frac{1}{2\phi}\sum_{i=1}^{m}\sum_{j=1}^{n_i}d_{ij}(Y_{ij};\mu_{ij}^b)-\frac{1}{2}b^TG^{-1}b-\frac{1}{2\tau}c^Tc\\ &\quad -\frac{1}{2}\beta^Tn\Sigma_\lambda(\widehat{\beta}^{(0)})\beta, \end{aligned}\quad (6.10)$$

where $n\Sigma_\lambda(\widehat{\beta}^{(0)})$ is defined as in Chapter 3.

As in Chapter 3, we similarly define X and Z, and re-write $n\Sigma_\lambda$ as $n\Sigma_\lambda=\begin{pmatrix}0 & 0\\ 0 & \Sigma_{22}\end{pmatrix}$, and β and X as $\beta=\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix}$ and $X=(X_1,X_2)$. So β_1 corresponds to the part of coefficients without penalty, β_2 is the rest part with penalty, and X_1 and X_2 are the covariates

associated with β_1 and β_2 respectively. As a result, $\frac{1}{2}\beta^T n\Sigma_\lambda(\widehat{\beta}^{(0)})\beta = \frac{1}{2}\beta_2^T\Sigma_{22}\beta_2$.

With this re-ordering, $\tilde{\ell}_{tpql}(\beta, f, \theta; Y)$ can be written as

$$\tilde{\ell}_{tpql}(\beta, f, \theta; Y) = -\frac{1}{2\phi}\sum_{i=1}^{m}\sum_{j=1}^{n_i} d_{ij}(Y_{ij}; \mu_{ij}^b) - \frac{1}{2}b^T G^{-1} b - \frac{1}{2\tau}c^T c - \frac{1}{2}\beta_2^T\Sigma_{22}\beta_2. \quad (6.11)$$

Replacing f in model (6.2) by $T\delta + Bc$ as shown in (6.8), in consideration of the relationship of the modified model (6.2) to (6.11) verse that of equation (1) to equation (6) of Breslow and Clayton (1993) (also Lin and Zhang, 1999), we conclude that, through the PQL approach, the penalized estimators of (β, f) can be solved under the following GLMM representation

$$g(\mu^b) = X_\star\beta_\star + B_\star c + Zb + X_2\beta_2, \quad (6.12)$$

with $\beta_\star = (\beta_1^T, \delta^T)^T$ being fixed effect associated with covariate $X_\star = (X_1, NT)$, c, b and β_2 being random effects associated with covariates $B_\star = NB$, Z and X_2, and distributed as $N(0, \tau I)$, $N(0, G)$ and $N(0, \Sigma_{22}^{-1})$ respectively.

Let Δ =diag$\{g'(\mu_{ij}^b)\}$, and

$$y = X\beta + Nf + Zb + \Delta(Y - \mu^b) \quad (6.13)$$

be defined as the working vector.

With the Fisher scoring algorithm, the maximizer of the TPQL (6.11) satisfies the

equations below

$$\begin{pmatrix} X_\star^T W X_\star & X_\star^T W B_\star & X_\star^T W Z & X_\star^T W X_2 \\ B_\star^T W X_\star & B_\star^T W B_\star + 1/\tau I & B_\star^T W Z & B_\star^T W X_2 \\ Z^T W X_\star & Z^T W B_\star & Z^T W Z + G^{-1} & Z^T W X_2 \\ X_2^T W X_\star & X_2^T W B_\star & X_2^T W Z & X_2^T W X_2 + \Sigma_{22} \end{pmatrix} \begin{pmatrix} \beta_\star \\ c \\ b \\ \beta_2 \end{pmatrix} = \begin{pmatrix} X_\star^T W y \\ B_\star^T W y \\ Z^T W y \\ X_2^T W y \end{pmatrix}. \tag{6.14}$$

System (6.14) is, in fact, the normal equations for the best linear unbiased predictors (BLUP) of $\beta_\star$ and (c, b, β_2) under the linear mixed model

$$y = X_\star \beta_\star + B_\star c + Zb + X_2 \beta_2 + \epsilon, \tag{6.15}$$

where random effects c, b and β_2 are distributed as mentioned above, $\epsilon \sim N(0, W^{-1})$, and c, b, β_2, ϵ are independent to each other.

Therefore, through approximations and transformations, the model selection in a GSMM can proceed in a GLMM representation, which is in fact an estimation problem in the working linear mixed models (6.15).

6.4 Estimation and Inference on Parameter and Non-parametric Function f

6.4.1 Estimation of Fixed Effects

Similar to the approach in Section 3.2.3, the penalized ML estimator of fixed effects can be solved in an equivalent way.

As Lin and Zhang (1999) showed, an examination of the relationship of the function

$$\tilde{\ell}_d(\beta, f, \theta; Y) = -\frac{1}{2\phi}\sum_{i=1}^{m}\sum_{j=1}^{n_i} d_{ij}(Y_{ij}; \mu_{ij}^b) - \frac{1}{2}b^T G^{-1} b - \frac{1}{2\tau}c^T c \tag{6.16}$$

to the modified GSMM (6.2) (f replaced by $T\delta + Bc$) yields β and the natural cubic spline estimator of f can be solved under the GLMM representation

$$g(\mu^b) = \tilde{X}\xi + B_\star c + Zb, \tag{6.17}$$

where $\tilde{X} = (X, NT)$, $\xi = (\beta^T, \delta^T)^T$, (c, b) are random effects with distribution $N(0, \tau I)$ and $N(0, G)$ respectively. The application of the Fisher scoring algorithm to (6.16) can solve (β, f) from equations

$$\begin{pmatrix} \tilde{X}^T W \tilde{X} & \tilde{X}^T W B_\star & \tilde{X}^T W Z \\ B_\star^T W \tilde{X} & B_\star^T W B_\star + 1/\tau I & B_\star^T W Z \\ Z^T W \tilde{X} & Z^T W B_\star & Z^T W Z + G^{-1} \end{pmatrix} \begin{pmatrix} \xi \\ c \\ b \end{pmatrix} = \begin{pmatrix} \tilde{X}^T W y \\ B_\star^T W y \\ Z^T W y \end{pmatrix}. \tag{6.18}$$

System (6.18) are normal equations for ξ and (b, c) under the linear mixed model

$$y = \tilde{X}\xi + B_\star c + Zb + \epsilon. \tag{6.19}$$

The working linear mixed model (6.19) provides a nice framework for inference on parameters.

Subtracting the SCAD penalty function from the log-likelihood of the working linear mixed model (6.19), we have our objective function

$$G(\beta) \;=\; -\frac{1}{2}(y - \tilde{X}\xi)^T\tilde{V}^{-1}(y - \tilde{X}\xi) - n\sum_{j=1}^{d} p_\lambda(|\beta_j|), \tag{6.20}$$

where $\tilde{V} = \tau B_\star B_\star^T + V$ with V defined in Chapter 3.

Let $\widehat{\xi}^{(0)}$ be an initial value which is very close to the maximizer of (6.20). Using the local quadratic approximation, we can approximate the objective function by a quadratic function

$$AG(\beta) \;=\; -\frac{1}{2}(y - \tilde{X}\xi)^T\tilde{V}^{-1}(y - \tilde{X}\xi) - \frac{1}{2}\xi^T n\Sigma_\lambda(\widehat{\xi}^{(0)})\xi, \tag{6.21}$$

where

$$n\Sigma_\lambda(\widehat{\xi}^{(0)}) = \begin{pmatrix} n\Sigma_\lambda(\widehat{\beta}^{(0)}) & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 \end{pmatrix}, \tag{6.22}$$

because we do not penalize the fixed effects of the linear mixed model (6.19) which come from the cubic smoothing spline.

Then the approximated objective function can be written as

$$AG(\beta) \;=\; -\frac{1}{2}(y-\tilde{X}\xi)^T\tilde{V}^{-1}(y-\tilde{X}\xi) - \frac{1}{2}\xi^T n\Sigma_\lambda(\widehat{\xi}^{(0)})\xi. \tag{6.23}$$

Taking derivative of AG with respect to ξ, we have

$$\widehat{\xi}^{(1)} \;=\; \left\{\tilde{X}^T\tilde{V}^{-1}\tilde{X} + n\Sigma_\lambda(\widehat{\xi}^{(0)})\right\}^{-1}\tilde{X}^T\tilde{V}^{-1}y. \tag{6.24}$$

6.4.2 Estimation of Smoothing Parameter and Variance Components

Treating the smoothing parameter as an extra variance component, we can jointly estimate variance components $\vartheta = (\tau, \theta)$ of the linear mixed model (6.19) based on its REML

$$\ell_{REMLd}(\xi, \vartheta; y) = -\frac{1}{2}\log|\tilde{V}| - \frac{1}{2}\log|\tilde{X}^T\tilde{V}^{-1}\tilde{X}| - \frac{1}{2}(y-\tilde{X}\widehat{\xi})^T\tilde{V}^{-1}(y-\tilde{X}\widehat{\xi}). \tag{6.25}$$

With the Fisher scoring algorithm, the REML estimator of ϑ can be iteratively solved by the following equation:

$$\widehat{\vartheta}^{(1)} \;=\; \widehat{\vartheta}^{(0)} + \left\{I(\widehat{\vartheta}^{(0)})\right\}^{-1} S(\widehat{\vartheta}^{(0)}),$$

where $\widehat{\vartheta}^{(0)}$ is a starting value.

Here $I(\vartheta)$ and $S(\vartheta)$ are the fisher information matrix and the score function for ϑ respectively,

$$I_{ij}(\vartheta) = \frac{1}{2} tr\left(\tilde{P}\frac{\partial \tilde{V}}{\partial \vartheta_i}\tilde{P}\frac{\partial \tilde{V}}{\partial \vartheta_j}\right),$$

$$S(\vartheta_i) = -\frac{1}{2} tr(\tilde{P}\frac{\partial \tilde{V}}{\partial \vartheta_i}) + \frac{1}{2}(y - \tilde{X}\widehat{\xi})^T \tilde{V}^{-1}\frac{\partial \tilde{V}}{\partial \vartheta_i}\tilde{V}^{-1}(y - \tilde{X}\widehat{\xi}),$$

with

$$\tilde{P} = \tilde{V}^{-1} - \tilde{V}^{-1}\tilde{X}(\tilde{X}^T\tilde{V}^{-1}\tilde{X})^{-1}\tilde{X}^T\tilde{V}^{-1}.$$

6.5 Summary

We extend the GLMM (1.2) to a more general class of models for overdispersed and correlated data, termed as generalized semi-parametric mixed models (GSMMs). This class of models models the time effect non-parametrically using an arbitrary smoothing spline and the covariates parametrically, while accounting for correlation between observations by incorporating random effects. By extending the PQL approach with SCAD penalty to model selection in GSMMs, we can make model selection and systematic inference on all model components in the framework of working linear mixed models. Treated as an extra variance component, the smoothing parameter can be jointly estimated with the variance components.

In the future, we will focus on the computational algorithm of the PQL method for

model selection in GSMMs and the numerical studies to evaluate the performance of the proposed procedure.

Chapter 7

Discussion

We proposed to use the SCAD penalty for model selection in GLMMs and developed four variable selection procedures for longitudinal data. A unified algorithm for simultaneous model selection and parameter estimation using full likelihood was developed. A linear mixed model representation was derived by the PQL approach, which significantly reduces the computational burden. Linear mixed model theories ensure the joint model selection and parameter estimation. For the FL and PQL methods, a robust standard error estimator was given by a sandwich formula, which was numerically tested. A marginal approach was proposed for model selection in longitudinal data analysis. Standard estimation procedure was used for point estimation. In consideration of the low estimation capability of the PQL method for binary data, the two-stage PQL approach was proposed to keep the good performance of PQL in model selection and reduce the biases of the estimators. The tuning parameter λ was selected by BIC for all procedures. GCV and REML criteria were also developed to tune λ for the PQL approach.

The proposed procedures can be easily extended to longitudinal data with missing values. Because of the shrinkage penalty based feature of our procedures, it is possible to apply them to the case in which the sample size is less than the number of parameters. In addition, the PQL approach is readily extended to generalized semi-parametric

mixed models for high-dimensional data. We have already developed model selection algorithms for this type of models. In the future we will focus on their computational algorithms and the numerical studies.

Bibliography

[1] Aitkin, M. (1999), "A General Maximum Likelihood Analysis of Variance Components in Generalized Linear Models," *Biometrics*, 55, 117-128.

[2] Akaike, H. (1973), "Maximum Likelihood Identification of Gaussian Autoregressive Moving Average Models," *Biometrika*, 60, 255-265.

[3] Akaike, H. (1977), On Entropy Maximization Principle. In P.R. Krishnaiah (Ed.) *Applications of Statistics*, Amsterdam: North-Holland.

[4] Barndorff-Nielsen, O.E., and Cox, D.R. (1989), *Asymptotic Techniques for Use in Statistics*, London: Chapmann and Hall.

[5] Breiman, L. (1996), "Heuristics of Instability and Stabilization in Model Selection," *The Annals of Statistics*, 24, 2350-2383.

[6] Breslow, N.E., and Clayton, D.G. (1993), "Approximate Inference in Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 88, 9-25.

[7] Breslow, N.E., and Lin, X. (1995), "Bias Correction in Generalized Linear Mixed Models With a Single Component of Dispersion," *Biometrika*, 82, 81-91.

[8] Diggle, P.J., Heagerty, P.J., Liang, K., and Zeger, S.L. (2002), *Analysis of Longitudinal Data*, Oxford: University Press.

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

[10] Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348-1360.

[11] Fan, J., and Li, R. (2002), "Variable Selection for Cox's proportional Hazards Model and Frailty Model," *The Annals of Statistics*, 30, 74-99.

[12] Fan, J., and Li, R. (2004), "New Estimation and Model Selection Procedures for Semiparametric Modeling in Longitudinal Data Analysis," *Journal of the American Statistical Association*, 99, 710-723.

[13] Frank, I.E., and Friedman, J.H. (1993), "A Statistical view of Some Chenometric Regression Tools (with discussion)," *Technometrics*, 35, 109-148.

[14] Gelfand, A.E., and Smith, A.F.M (1990), "Sampling Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398-409.

[15] Geman, S., and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.

[16] Green, P.J. (1987), "Penalized Likelihood for General Semi-Parametric Regression Models," *International Statistical Review*, 55, 245-259.

[17] Hastie, T. (1987), "A Closer Look at the Deviance," *The American Statistician*, 41, 16-20.

[18] Hoerl, A.E., and Kennard, R.W. (1970a), "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 12, 55-67.

[19] Hoerl, A.E., and Kennard, R.W. (1970b), "Ridge Regression: Application to Nonorthogonal Problems," *Technometrics*, 12, 69-82.

[20] Jiang, J. (1999), "Conditional Inference about Generalized Linear Mixed Models," *The Annals of Statistics*, 27, 1974-2008.

[21] Kaslow, R.A., Ostrow, D.G., Detels, R., Phair, J.P., Polk, B.F., and Rinaldo, C.R. (1987), "The Multicenter AIDS Cohort Study: Rationale, Organization and Selected Characteristics of the Participants," *American Journal of Epidemiology*, 126, 310-318.

[22] Knight, K. and Fu, W.J. (2000), "Asymptotics for Lasso-type estimators," *The Annals of Statistics*, 28, 1356-1378.

[23] Laird, N.M. (1978), "Empirical Bayes Methods for Two-Way Contingency Tables," *Biometrika*, 65, 581-590.

[24] Lee, Y., and Nelder, J.A. (1996), "Hierarchacal Generalized Linear Models (with discussion)," *Journal of the Royal Statistical Society*, Ser. B, 58, 619-678.

[25] Lin, X., and Breslow, N.E. (1996), "Bias Correction in Generalized Linear Mixed Models With Multiple Components of Dispersion," *Journal of the American Statistical Association*, 91, 1007-1016.

[26] Lin, X., and Zhang, D. (1999), "Inference in Generalized Additive Mixed Models

by Using Smoothing Splines," *Journal of the Royal Statistical Society*, Ser. B, 61, 381-400.

[27] Liu, Q., and Pierce, D.A. (1993), "Heterogeneity in Mantel-Haenszel-type Models," *Biometrika*, 80, 543-556.

[28] Luo, X., Stefanski, L.A., and Boos, D.D. (2006), "Tuning Variable Selection Procedures by Adding Noise," to appear in *Technometrics*.

[29] McCullagh, P., and Nelder, J.A. (1989), *Generalized Linear Models*, Chapman and Hall: London.

[30] McCulloch, C. (1997), "Maximum Likelihood Algorithms for Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 92, 162-190.

[31] Miller, A. (1990), *Subset Selection in Regression*, New York: Chapmann and Hall.

[32] Mitchell, T.J., and Beauchamp, J.J (1988), "Bayesian Variable Selection in Linear Regression: Rejoinder," *Journal of the American Statistical Association*, 83, 1035-1036.

[33] Nelder, J.A. (1972), "Discussion of Paper by Lindley and Smith," *Journal of the Royal Statistical Society*, Ser. B, 24, 1-41.

[34] Nelder, J.A., and Wedderburn, R.W.M. (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society*, Ser. A, 135, 370-384.

[35] Schall, R. (1991), "Estimation in Generalized Linear Models With Random Effects," *Biometrika*, 40, 917-927.

[36] Schwarz, G. (1978), "Estimating the dimension of a model," *The Annals of Statistics*, 6, 461-464.

[37] Solomon, P.J., and Cox, D.R. (1992), "Nonlinear Components of Variance Models," *Biometrika*, 79, 1-11.

[38] Tibshirani, R.J. (1996), "Regression Shrinkage and Selection via the LASSO," *Journal of the Royal Statistical Society*, Ser. B, 58, 267-288.

[39] Tibshirani, R.J. (1997), "The LASSO Method for Variable Selection in the Cox Model," *Statistics in Medicine*, 16, 385-395.

[40] Vos, P.W. (1991), "A Geometric Approach to Detecting Influential Cases," *The Annals of Statistics*, 19, 1570-1581.

[41] Wahba, G. (1990), *Spline Models for Observational Data*, Philadelphia: Society for Industrial and Applied Mathematics.

[42] Wu, Y., Boos, D.D., and Stefanski, L.A. (2006), "Controlling Variable Selection by the Addition of Pseudo Variables," to appear in *Journal of the American Statistical Association*.

[43] Yuan, Ming., and Lin, Yi. (2005), "Efficient Empirical Bayes Variable Selection and Estimation in Linear Models ," *Journal of the American Statistical Association*, 100, 1215-1225.

[44] Zeger, S.L., and Karim, M.R. (1991), "Generalized Linear Models with Random Effects: A Gibbs Sampling Approach," *Journal of the American Statistical Association*, 86, 79-86.

[45] Zeger, S.L., Liang, K., and Albert, P.S. (1988), "Models for Longitudinal Data: A Generalized Estimating Equation Approach," *Biometrics*, 44, 1049-1060.

[46] Zhang, D. (2004), "Generalized Linear Mixed Models with Varying Coefficients for Longitudinal Data," *Biometrics*, 60, 8-15.

Appendices

Appendix A

Proof of Equivalence between System (3.7) and Equation (3.13)

In terms of model selection and point estimation, system (3.7) and equation (3.13) are equivalent.

Proof: From the second equation of (3.5), we have

$$b = (Z^T W Z + D^{-1})^{-1} Z^T W (y - X\beta) \tag{1.1}$$

Plugging 1.1 into the first equation of (3.5), it is easy to show that

$$\widehat{\beta}^{(1)} = (X^T \tilde{V}^{-1} X + n\Sigma_\lambda(\widehat{\beta}^{(0)}))^{-1} X^T \tilde{V}^{-1} y \tag{1.2}$$

where $\tilde{V}^{-1} = W - WZ(Z^T W Z + D^{-1})^{-1} Z^T W$.

Appendix A. Proof of Equivalence between System (3.7) *and Equation* (3.13)

Now we will show $\tilde{V}^{-1} = V^{-1} = (ZDZ^T + W^{-1})^{-1}$.

$$\begin{aligned} V^{-1} &= (ZDZ^T + W^{-1})^{-1} \\ &= (I + WZDZ^T)^{-1}W \\ &= \{I - WZD(I + Z^TWZD)^{-1}Z^T\}W \\ &= \{I - WZ(Z^TWZ + D^{-1})^{-1}Z^T\}W \\ &= W - WZ(Z^TWZ + D^{-1})^{-1}Z^TW \\ &= \tilde{V}^{-1} \end{aligned}$$

By now, we have showed the two linear mixed model representations yield the same point estimator $\widehat{\beta}$ of β. Therefore, given λ and θ, we can use the iteration equation (3.13) for model selection and parameter estimation.

Appendix B

Tables of Simulation Results

Table B.1: Inference on β_2 at Scenario I: $(\beta_2, \beta_5, \beta_6, \beta_7) = (0.9, 0.9, 0.9, 0.9)$ by BIC. SD, the Monte Carlo median absolute deviation divided by 0.6745; SE, the mean of 100 estimated standard error by sandwich formula; SD_e, the empirical standard error estimate of the sandwich standard error estimate; CP, Monte Carlo coverage probability of 95% confidence interval; FL, full likelihood approach; PQL, penalized quasi-likelihood approach; $CPQL$, PQL after bias correction; AML, approximate marginal likelihood approach; $TPQL$, two-stage penalized quasi-likelihood approach

m	*method*	$\theta_1 = 1.96$					$\theta_2 = 1.00$				
		$\hat{\beta}_2$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$	$\hat{\beta}_2$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$
100	FL	0.91	0.15	0.17(0.02)	0.01	0.93	0.91	0.19	0.16(0.02)	0.01	0.95
	PQL	0.77	0.13	0.15(0.01)	−0.13	0.83	0.82	0.16	0.14(0.01)	−0.08	0.89
	$CPQL$	0.80	0.14	0.15(0.01)	−0.10	0.86	0.84	0.17	0.15(0.01)	−0.06	0.91
	AML	0.91	0.16	0.17(0.02)	0.01	0.96	0.91	0.17	0.16(0.02)	0.01	0.95
	$TPQL$	0.90	0.16	0.17(0.02)	0.00	0.95	0.91	0.18	0.16(0.02)	0.01	0.95
200	FL	0.91	0.12	0.12(0.01)	0.01	0.95	0.92	0.11	0.11(0.01)	0.02	0.96
	PQL	0.77	0.10	0.10(0.01)	−0.13	0.75	0.82	0.09	0.10(0.00)	−0.08	0.89
	$CPQL$	0.79	0.10	0.10(0.01)	−0.11	0.81	0.83	0.09	0.10(0.00)	−0.07	0.89
	AML	0.91	0.12	0.12(0.01)	0.01	0.95	0.91	0.11	0.11(0.01)	0.01	0.96
	$TPQL$	0.91	0.12	0.12(0.01)	0.01	0.95	0.91	0.11	0.11(0.01)	0.01	0.96
		$\theta_3 = 0.81$					$\theta_4 = 0.64$				
		$\hat{\beta}_2$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$	$\hat{\beta}_2$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$
100	FL	0.91	0.15	0.16(0.01)	0.01	0.95	0.92	0.18	0.16(0.01)	0.02	0.97
	PQL	0.83	0.15	0.14(0.01)	−0.07	0.90	0.85	0.14	0.14(0.01)	−0.05	0.92
	$CPQL$	0.84	0.14	0.15(0.01)	−0.06	0.92	0.86	0.15	0.15(0.01)	−0.04	0.95
	AML	0.91	0.15	0.16(0.01)	0.01	0.95	0.92	0.18	0.16(0.01)	0.02	0.97
	$TPQL$	0.91	0.16	0.16(0.01)	0.01	0.95	0.92	0.17	0.16(0.01)	0.02	0.97
200	FL	0.92	0.12	0.11(0.01)	0.02	0.95	0.92	0.11	0.11(0.01)	0.02	0.96
	PQL	0.84	0.11	0.10(0.00)	−0.06	0.88	0.85	0.10	0.10(0.00)	−0.05	0.87
	$CPQL$	0.85	0.11	0.10(0.00)	−0.05	0.89	0.86	0.10	0.10(0.00)	−0.04	0.89
	AML	0.92	0.12	0.11(0.01)	0.02	0.95	0.92	0.13	0.11(0.01)	0.02	0.96
	$TPQL$	0.92	0.12	0.11(0.01)	0.02	0.95	0.92	0.13	0.11(0.01)	0.02	0.96

Table B.2: Inference on β_2 at Scenario II: $(\beta_2, \beta_5, \beta_6, \beta_7) = (0.5, 0.5, 0.5, 0.5)$ by BIC. SD, the Monte Carlo median absolute deviation divided by 0.6745; SE, the mean of 100 estimated standard error by sandwich formula; SD_e, the empirical standard error estimate of the sandwich standard error estimate; CP, Monte Carlo coverage probability of 95% confidence interval; FL, full likelihood approach; PQL, penalized quasi-likelihood approach; $CPQL$, PQL after bias correction; AML, approximate marginal likelihood approach; $TPQL$, two-stage penalized quasi-likelihood approach

m	*method*	$\theta_1 = 1.96$					$\theta_2 = 1.00$				
		$\hat{\beta}_2$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$	$\hat{\beta}_2$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$
100	FL	0.54	0.13	0.14(0.02)	0.04	0.90	0.54	0.13	0.13(0.02)	0.04	0.94
	PQL	0.48	0.09	0.13(0.01)	−0.02	0.99	0.49	0.12	0.12(0.01)	−0.01	0.98
	$CPQL$	0.49	0.10	0.13(0.01)	−0.01	0.98	0.50	0.13	0.12(0.01)	0.00	0.97
	AML	0.55	0.14	0.14(0.01)	0.05	0.94	0.55	0.14	0.13(0.01)	0.05	0.97
	$TPQL$	0.56	0.11	0.14(0.01)	0.06	0.94	0.55	0.14	0.13(0.01)	0.05	0.95
200	FL	0.51	0.11	0.10(0.01)	0.01	0.90	0.52	0.13	0.11(0.00)	0.02	0.96
	PQL	0.46	0.11	0.11(0.00)	0.04	0.93	0.44	0.08	0.09(0.00)	−0.06	0.85
	$CPQL$	0.46	0.12	0.12(0.00)	0.04	0.94	0.45	0.09	0.09(0.00)	−0.05	0.88
	AML	0.52	0.13	0.12(0.01)	0.02	0.96	0.51	0.11	0.10(0.01)	0.01	0.91
	$TPQL$	0.51	0.12	0.12(0.01)	0.01	0.96	0.51	0.11	0.10(0.00)	0.01	0.92
		$\theta_3 = 0.81$					$\theta_4 = 0.64$				
		$\hat{\beta}_2$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$	$\hat{\beta}_2$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$
100	FL	0.53	0.12	0.13(0.02)	0.03	0.95	0.54	0.16	0.15(0.01)	0.04	0.93
	PQL	0.49	0.11	0.12(0.01)	−0.01	0.95	0.49	0.10	0.12(0.01)	−0.01	0.97
	$CPQL$	0.50	0.12	0.12(0.01)	0.00	0.96	0.50	0.10	0.12(0.01)	0.00	0.98
	AML	0.54	0.13	0.13(0.01)	0.04	0.97	0.53	0.13	0.13(0.01)	0.02	0.97
	$TPQL$	0.54	0.12	0.13(0.01)	0.04	0.96	0.53	0.11	0.13(0.01)	0.03	0.98
200	FL	0.51	0.09	0.09(0.00)	0.01	0.91	0.51	0.10	0.09(0.00)	0.01	0.94
	PQL	0.47	0.07	0.08(0.00)	−0.03	0.90	0.47	0.08	0.08(0.00)	−0.03	0.93
	$CPQL$	0.47	0.07	0.08(0.00)	−0.03	0.89	0.47	0.08	0.08(0.00)	−0.03	0.93
	AML	0.51	0.09	0.09(0.00)	0.01	0.92	0.51	0.10	0.09(0.00)	0.01	0.94
	$TPQL$	0.51	0.09	0.09(0.00)	0.01	0.91	0.51	0.10	0.09(0.00)	0.01	0.95

Table B.3: Inference on β_5 at Scenario I: $(\beta_2, \beta_5, \beta_6, \beta_7) = (0.9, 0.9, 0.9, 0.9)$ by BIC. SD, the Monte Carlo median absolute deviation divided by 0.6745; SE, the mean of 100 estimated standard error by sandwich formula; SD_e, the empirical standard error estimate of the sandwich standard error estimate; CP, Monte Carlo coverage probability of 95% confidence interval; FL, full likelihood approach; PQL, penalized quasi-likelihood approach; $CPQL$, PQL after bias correction; AML, approximate marginal likelihood approach; $TPQL$, two-stage penalized quasi-likelihood approach

m	*method*	$\theta_1 = 1.96$					$\theta_2 = 1.00$				
		$\hat{\beta}_5$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$	$\hat{\beta}_5$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$
100	FL	0.92	0.21	0.19(0.02)	0.02	0.94	0.94	0.18	0.18(0.02)	0.04	0.97
	PQL	0.78	0.17	0.17(0.01)	−0.12	0.88	0.85	0.16	0.16(0.01)	−0.05	0.97
	$CPQL$	0.81	0.19	0.17(0.01)	−0.09	0.90	0.87	0.16	0.16(0.01)	−0.03	0.97
	AML	0.92	0.19	0.19(0.02)	0.02	0.92	0.94	0.17	0.18(0.02)	0.04	0.96
	$TPQL$	0.92	0.21	0.19(0.02)	0.02	0.93	0.95	0.18	0.18(0.02)	0.05	0.95
200	FL	0.91	0.13	0.13(0.01)	0.01	0.94	0.93	0.13	0.13(0.01)	0.03	0.92
	PQL	0.77	0.10	0.11(0.01)	−0.13	0.77	0.83	0.10	0.11(0.01)	−0.07	0.92
	$CPQL$	0.79	0.11	0.12(0.01)	−0.11	0.83	0.84	0.11	0.11(0.01)	−0.06	0.93
	AML	0.90	0.14	0.13(0.01)	0.00	0.93	0.92	0.13	0.13(0.01)	0.02	0.93
	$TPQL$	0.90	0.14	0.13(0.01)	0.00	0.93	0.92	0.13	0.13(0.01)	0.02	0.93
		$\theta_3 = 0.81$					$\theta_4 = 0.64$				
		$\hat{\beta}_5$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$	$\hat{\beta}_5$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$
100	FL	0.94	0.18	0.18(0.01)	0.04	0.96	0.94	0.19	0.18(0.02)	0.04	0.95
	PQL	0.86	0.15	0.16(0.01)	−0.04	0.95	0.87	0.18	0.16(0.01)	−0.03	0.96
	$CPQL$	0.87	0.16	0.16(0.01)	−0.03	0.95	0.88	0.18	0.16(0.01)	−0.02	0.98
	AML	0.93	0.16	0.18(0.01)	0.03	0.96	0.94	0.19	0.18(0.02)	0.04	0.96
	$TPQL$	0.94	0.17	0.18(0.01)	0.04	0.96	0.94	0.19	0.18(0.02)	0.04	0.96
200	FL	0.92	0.11	0.12(0.01)	0.02	0.97	0.92	0.12	0.12(0.01)	0.02	0.98
	PQL	0.84	0.10	0.11(0.01)	−0.06	0.94	0.85	0.12	0.11(0.01)	−0.05	0.94
	$CPQL$	0.85	0.10	0.11(0.01)	−0.05	0.94	0.86	0.12	0.11(0.01)	−0.04	0.94
	AML	0.92	0.11	0.12(0.01)	0.02	0.97	0.92	0.12	0.12(0.01)	0.02	0.98
	$TPQL$	0.92	0.11	0.12(0.01)	0.02	0.97	0.92	0.12	0.12(0.01)	0.02	0.98

Table B.4: Inference on β_5 at Scenario II: $(\beta_2, \beta_5, \beta_6, \beta_7) = (0.5, 0.5, 0.5, 0.5)$ by BIC. SD, the Monte Carlo median absolute deviation divided by 0.6745; SE, the mean of 100 estimated standard error by sandwich formula; SD_e, the empirical standard error estimate of the sandwich standard error estimate; CP, Monte Carlo coverage probability of 95% confidence interval; FL, full likelihood approach; PQL, penalized quasi-likelihood approach; $CPQL$, PQL after bias correction; AML, approximate marginal likelihood approach; $TPQL$, two-stage penalized quasi-likelihood approach

m	*method*	$\theta_1 = 1.96$					$\theta_2 = 1.00$				
		$\hat{\beta}_5$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$	$\hat{\beta}_5$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$
100	FL	0.60	0.16	0.15(0.01)	0.10	0.85	0.56	0.16	0.15(0.01)	0.06	0.94
	PQL	0.50	0.14	0.14(0.01)	0.00	0.99	0.50	0.13	0.14(0.01)	0.00	0.98
	$CPQL$	0.51	0.14	0.14(0.01)	0.01	0.97	0.51	0.13	0.14(0.01)	0.01	0.95
	AML	0.58	0.15	0.15(0.01)	0.08	0.90	0.55	0.16	0.15(0.01)	0.05	0.94
	$TPQL$	0.58	0.16	0.15(0.01)	0.08	0.91	0.55	0.14	0.15(0.01)	0.05	0.94
200	FL	0.53	0.11	0.11(0.01)	0.03	0.91	0.52	0.10	0.10(0.01)	0.02	0.96
	PQL	0.45	0.08	0.10(0.00)	−0.05	0.96	0.46	0.08	0.09(0.00)	−0.04	0.94
	$CPQL$	0.46	0.09	0.10(0.00)	−0.04	0.96	0.47	0.09	0.10(0.00)	−0.03	0.94
	AML	0.52	0.09	0.11(0.01)	0.02	0.90	0.52	0.10	0.10(0.01)	0.02	0.96
	$TPQL$	0.53	0.10	0.11(0.01)	0.03	0.93	0.52	0.09	0.10(0.01)	0.02	0.96
		$\theta_3 = 0.81$					$\theta_4 = 0.64$				
		$\hat{\beta}_5$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$	$\hat{\beta}_5$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$
100	FL	0.55	0.14	0.14(0.01)	0.05	0.94	0.55	0.13	0.14(0.01)	0.05	0.95
	PQL	0.51	0.13	0.13(0.01)	0.01	0.99	0.50	0.12	0.13(0.01)	0.00	0.98
	$CPQL$	0.52	0.12	0.14(0.01)	0.02	0.97	0.50	0.13	0.13(0.01)	0.00	0.98
	AML	0.56	0.15	0.14(0.01)	0.06	0.93	0.54	0.14	0.14(0.01)	0.04	0.98
	$TPQL$	0.56	0.14	0.14(0.01)	0.06	0.95	0.54	0.14	0.14(0.01)	0.04	0.97
200	FL	0.51	0.10	0.10(0.00)	0.01	0.98	0.50	0.09	0.10(0.00)	0.00	0.99
	PQL	0.47	0.09	0.09(0.00)	−0.03	0.97	0.46	0.07	0.09(0.00)	−0.04	0.98
	$CPQL$	0.47	0.09	0.09(0.00)	−0.03	0.97	0.47	0.08	0.09(0.00)	−0.03	0.98
	AML	0.51	0.10	0.10(0.00)	0.01	0.98	0.50	0.08	0.10(0.00)	0.00	0.99
	$TPQL$	0.51	0.10	0.10(0.00)	0.01	0.98	0.50	0.08	0.10(0.00)	0.00	0.99

Table B.5: Inference on β_6 at Scenario I: $(\beta_2, \beta_5, \beta_6, \beta_7) = (0.9, 0.9, 0.9, 0.9)$ by BIC. SD, the Monte Carlo median absolute deviation divided by 0.6745; SE, the mean of 100 estimated standard error by sandwich formula; SD_e, the empirical standard error estimate of the sandwich standard error estimate; CP, Monte Carlo coverage probability of 95% confidence interval; FL, full likelihood approach; PQL, penalized quasi-likelihood approach; $CPQL$, PQL after bias correction; AML, approximate marginal likelihood approach; $TPQL$, two-stage penalized quasi-likelihood approach

m	*method*	$\theta_1 = 1.96$					$\theta_2 = 1.00$				
		$\hat{\beta}_6$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$	$\hat{\beta}_6$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$
100	FL	0.92	0.18	0.21(0.02)	0.02	0.94	0.90	0.18	0.20(0.02)	0.00	0.97
	PQL	0.79	0.15	0.18(0.01)	−0.11	0.93	0.82	0.16	0.18(0.01)	−0.08	0.94
	$CPQL$	0.82	0.16	0.19(0.01)	−0.08	0.93	0.83	0.17	0.18(0.01)	−0.07	0.95
	AML	0.93	0.19	0.21(0.02)	0.03	0.95	0.90	0.19	0.20(0.02)	0.00	0.97
	$TPQL$	0.93	0.18	0.21(0.02)	0.03	0.96	0.91	0.18	0.20(0.02)	0.01	0.98
200	FL	0.91	0.15	0.14(0.01)	0.01	0.97	0.89	0.13	0.14(0.01)	−0.01	0.98
	PQL	0.77	0.12	0.13(0.01)	−0.13	0.84	0.80	0.11	0.12(0.00)	−0.10	0.90
	$CPQL$	0.79	0.12	0.13(0.01)	−0.11	0.86	0.81	0.11	0.12(0.00)	−0.09	0.92
	AML	0.90	0.15	0.14(0.01)	0.00	0.97	0.89	0.13	0.14(0.01)	−0.01	0.98
	$TPQL$	0.90	0.15	0.14(0.01)	0.00	0.97	0.89	0.13	0.14(0.01)	−0.01	0.98
		$\theta_3 = 0.81$					$\theta_4 = 0.64$				
		$\hat{\beta}_6$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$	$\hat{\beta}_6$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$
100	FL	0.91	0.18	0.19(0.01)	0.01	0.94	0.90	0.16	0.19(0.01)	0.00	0.99
	PQL	0.83	0.17	0.18(0.01)	−0.07	0.92	0.83	0.14	0.18(0.00)	−0.07	0.92
	$CPQL$	0.85	0.18	0.18(0.01)	−0.03	0.94	0.84	0.15	0.18(0.00)	−0.06	0.93
	AML	0.90	0.17	0.19(0.01)	0.00	0.95	0.90	0.17	0.19(0.01)	0.00	0.99
	$TPQL$	0.91	0.18	0.19(0.01)	0.01	0.95	0.90	0.16	0.19(0.01)	0.00	0.99
200	FL	0.90	0.13	0.13(0.01)	0.00	0.97	0.89	0.13	0.13(0.01)	−0.01	0.99
	PQL	0.82	0.12	0.12(0.00)	−0.08	0.95	0.82	0.11	0.12(0.00)	−0.08	0.95
	$CPQL$	0.83	0.12	0.12(0.01)	−0.07	0.96	0.83	0.12	0.12(0.01)	−0.07	0.95
	AML	0.90	0.13	0.13(0.01)	0.00	0.97	0.89	0.13	0.13(0.01)	−0.01	0.99
	$TPQL$	0.90	0.13	0.13(0.01)	0.00	0.97	0.89	0.13	0.13(0.01)	−0.01	0.99

Table B.6: Inference on β_6 at Scenario II: $(\beta_2, \beta_5, \beta_6, \beta_7) = (0.5, 0.5, 0.5, 0.5)$ by BIC. SD, the Monte Carlo median absolute deviation divided by 0.6745; SE, the mean of 100 estimated standard error by sandwich formula; SD_e, the empirical standard error estimate of the sandwich standard error estimate; CP, Monte Carlo coverage probability of 95% confidence interval; FL, full likelihood approach; PQL, penalized quasi-likelihood approach; $CPQL$, PQL after bias correction; AML, approximate marginal likelihood approach; $TPQL$, two-stage penalized quasi-likelihood approach

m	*method*	$\theta_1 = 1.96$					$\theta_2 = 1.00$				
		$\hat{\beta}_6$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$	$\hat{\beta}_6$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$
100	FL	0.62	0.22	0.17(0.02)	0.12	0.84	0.57	0.15	0.16(0.01)	0.07	0.94
	PQL	0.50	0.11	0.15(0.01)	0.00	0.98	0.50	0.13	0.15(0.01)	0.00	0.97
	$CPQL$	0.51	0.12	0.16(0.01)	0.01	0.98	0.50	0.13	0.15(0.01)	0.00	0.97
	AML	0.59	0.17	0.17(0.01)	0.09	0.93	0.56	0.16	0.16(0.01)	0.06	0.93
	$TPQL$	0.58	0.16	0.17(0.01)	0.08	0.90	0.55	0.15	0.16(0.01)	0.05	0.95
200	FL	0.50	0.10	0.12(0.01)	0.00	0.97	0.50	0.09	0.11(0.01)	0.00	0.98
	PQL	0.43	0.09	0.11(0.01)	−0.07	0.97	0.45	0.08	0.10(0.00)	−0.05	0.93
	$CPQL$	0.45	0.10	0.11(0.01)	−0.05	0.96	0.46	0.08	0.11(0.00)	−0.04	0.94
	AML	0.51	0.11	0.12(0.01)	0.01	0.98	0.50	0.08	0.11(0.01)	0.00	0.98
	$TPQL$	0.51	0.12	0.12(0.01)	0.01	0.97	0.50	0.09	0.11(0.01)	0.00	0.97
		$\theta_3 = 0.81$					$\theta_4 = 0.64$				
		$\hat{\beta}_6$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$	$\hat{\beta}_6$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$
100	FL	0.55	0.14	0.16(0.01)	0.05	0.95	0.58	0.18	0.15(0.01)	0.08	0.88
	PQL	0.50	0.12	0.15(0.01)	0.00	0.98	0.53	0.16	0.15(0.01)	0.03	0.98
	$CPQL$	0.51	0.12	0.15(0.01)	0.01	0.98	0.53	0.16	0.15(0.01)	0.03	0.98
	AML	0.58	0.15	0.16(0.01)	0.08	0.89	0.55	0.18	0.16(0.01)	0.05	0.90
	$TPQL$	0.55	0.14	0.16(0.01)	0.05	0.94	0.57	0.17	0.16(0.01)	0.07	0.92
200	FL	0.50	0.08	0.11(0.01)	0.00	0.97	0.50	0.10	0.11(0.01)	0.00	1.00
	PQL	0.45	0.08	0.10(0.00)	−0.05	0.96	0.46	0.09	0.10(0.00)	−0.04	0.97
	$CPQL$	0.46	0.09	0.10(0.00)	−0.04	0.96	0.47	0.09	0.10(0.00)	−0.03	0.97
	AML	0.50	0.08	0.11(0.01)	0.00	0.97	0.51	0.10	0.11(0.01)	0.01	0.98
	$TPQL$	0.50	0.08	0.11(0.01)	0.00	0.97	0.50	0.10	0.11(0.01)	0.00	0.99

Table B.7: Inference on β_7 at Scenario I: $(\beta_2, \beta_5, \beta_6, \beta_7) = (0.9, 0.9, 0.9, 0.9)$ by BIC. SD, the Monte Carlo median absolute deviation divided by 0.6745; SE, the mean of 100 estimated standard error by sandwich formula; SD_e, the empirical standard error estimate of the sandwich standard error estimate; CP, Monte Carlo coverage probability of 95% confidence interval; FL, full likelihood approach; PQL, penalized quasi-likelihood approach; $CPQL$, PQL after bias correction; AML, approximate marginal likelihood approach; $TPQL$, two-stage penalized quasi-likelihood approach

m	*method*	$\theta_1 = 1.96$					$\theta_2 = 1.00$				
		$\hat{\beta}_7$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$	$\hat{\beta}_7$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$
100	FL	0.95	0.20	0.19(0.02)	0.05	0.94	0.94	0.17	0.18(0.02)	0.04	0.97
	PQL	0.81	0.16	0.16(0.01)	−0.09	0.91	0.86	0.15	0.16(0.01)	−0.04	0.95
	$CPQL$	0.83	0.18	0.17(0.01)	−0.07	0.93	0.87	0.15	0.16(0.01)	−0.03	0.95
	AML	0.94	0.19	0.19(0.02)	0.04	0.95	0.94	0.16	0.18(0.02)	0.04	0.96
	$TPQL$	0.95	0.20	0.19(0.02)	0.05	0.94	0.95	0.16	0.18(0.02)	0.05	0.98
200	FL	0.93	0.11	0.13(0.01)	0.03	0.93	0.92	0.11	0.13(0.01)	0.02	0.98
	PQL	0.78	0.08	0.11(0.01)	−0.12	0.83	0.83	0.09	0.11(0.00)	−0.07	0.88
	$CPQL$	0.81	0.09	0.12(0.01)	−0.09	0.87	0.84	0.09	0.11(0.01)	−0.06	0.91
	AML	0.92	0.10	0.13(0.01)	0.02	0.93	0.92	0.10	0.12(0.01)	0.02	0.98
	$TPQL$	0.92	0.10	0.13(0.01)	0.02	0.94	0.92	0.10	0.12(0.01)	0.02	0.98
		$\theta_3 = 0.81$					$\theta_4 = 0.64$				
		$\hat{\beta}_7$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$	$\hat{\beta}_7$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$
100	FL	0.94	0.17	0.18(0.01)	0.04	0.97	0.95	0.17	0.18(0.01)	0.05	0.98
	PQL	0.87	0.16	0.16(0.01)	−0.03	0.97	0.88	0.15	0.16(0.01)	−0.02	0.96
	$CPQL$	0.88	0.16	0.16(0.01)	−0.02	0.96	0.89	0.15	0.16(0.01)	−0.01	0.97
	AML	0.94	0.17	0.18(0.01)	0.04	0.96	0.95	0.17	0.18(0.02)	0.05	0.98
	$TPQL$	0.95	0.17	0.18(0.01)	−0.05	0.96	0.95	0.17	0.18(0.02)	0.05	0.98
200	FL	0.93	0.10	0.12(0.01)	0.03	0.97	0.93	0.12	0.12(0.01)	0.03	0.96
	PQL	0.84	0.09	0.11(0.00)	−0.06	0.91	0.86	0.11	0.11(0.00)	−0.04	0.96
	$CPQL$	0.86	0.09	0.11(0.01)	−0.04	0.92	0.87	0.12	0.11(0.01)	−0.03	0.96
	AML	0.93	0.10	0.12(0.01)	0.03	0.97	0.93	0.12	0.12(0.01)	0.03	0.96
	$TPQL$	0.93	0.10	0.12(0.01)	0.03	0.97	0.93	0.12	0.12(0.01)	0.03	0.96

Table B.8: Inference on β_7 at Scenario II: $(\beta_2, \beta_5, \beta_6, \beta_7) = (0.5, 0.5, 0.5, 0.5)$ by BIC. SD, the Monte Carlo median absolute deviation divided by 0.6745; SE, the mean of 100 estimated standard error by sandwich formula; SD_e, the empirical standard error estimate of the sandwich standard error estimate; CP, Monte Carlo coverage probability of 95% confidence interval; FL, full likelihood approach; PQL, penalized quasi-likelihood approach; $CPQL$, PQL after bias correction; AML, approximate marginal likelihood approach; $TPQL$, two-stage penalized quasi-likelihood approach

m	*method*	$\theta_1 = 1.96$					$\theta_2 = 1.00$				
		$\hat{\beta}_7$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$	$\hat{\beta}_7$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$
100	FL	0.59	0.14	0.15(0.01)	0.09	0.89	0.55	0.14	0.15(0.01)	0.05	0.94
	PQL	0.48	0.12	0.14(0.01)	−0.02	0.98	0.48	0.12	0.14(0.01)	−0.02	0.98
	$CPQL$	0.50	0.12	0.14(0.01)	0.00	0.99	0.49	0.12	0.14(0.01)	−0.01	0.98
	AML	0.58	0.14	0.15(0.01)	0.08	0.90	0.54	0.13	0.15(0.01)	0.04	0.98
	$TPQL$	0.56	0.14	0.15(0.01)	0.06	0.94	0.53	0.13	0.15(0.01)	0.03	0.98
200	FL	0.50	0.12	0.11(0.01)	0.00	0.96	0.49	0.11	0.10(0.01)	−0.01	0.98
	PQL	0.43	0.10	0.10(0.00)	−0.07	0.89	0.44	0.09	0.09(0.00)	−0.06	0.95
	$CPQL$	0.44	0.11	0.10(0.00)	−0.06	0.91	0.45	0.09	0.10(0.00)	−0.05	0.96
	AML	0.50	0.12	0.11(0.01)	0.00	0.96	0.49	0.10	0.10(0.01)	−0.01	0.96
	$TPQL$	0.50	0.13	0.11(0.01)	0.00	0.97	0.50	0.10	0.10(0.01)	0.00	0.99
		$\theta_3 = 0.81$					$\theta_4 = 0.64$				
		$\hat{\beta}_7$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$	$\hat{\beta}_7$	SD	$SE(SD_e)$	*bias*	$CP(95\%)$
100	FL	0.53	0.15	0.14(0.01)	0.03	0.95	0.54	0.12	0.14(0.01)	0.04	0.96
	PQL	0.49	0.12	0.13(0.01)	−0.01	0.95	0.49	0.11	0.13(0.01)	−0.01	0.98
	$CPQL$	0.49	0.12	0.14(0.01)	−0.01	0.95	0.49	0.12	0.13(0.01)	−0.01	0.98
	AML	0.54	0.14	0.14(0.01)	0.04	0.94	0.52	0.13	0.14(0.01)	0.02	0.99
	$TPQL$	0.54	0.14	0.14(0.01)	0.04	0.96	0.53	0.12	0.14(0.01)	0.03	0.98
200	FL	0.50	0.10	0.10(0.01)	0.00	0.96	0.50	0.10	0.10(0.01)	0.00	0.99
	PQL	0.45	0.09	0.09(0.00)	−0.05	0.94	0.46	0.09	0.09(0.00)	−0.04	0.98
	$CPQL$	0.46	0.09	0.09(0.00)	−0.04	0.95	0.46	0.09	0.09(0.00)	−0.04	0.98
	AML	0.49	0.11	0.10(0.01)	−0.01	0.96	0.50	0.10	0.10(0.01)	0.00	0.99
	$TPQL$	0.50	0.11	0.10(0.01)	0.00	0.97	0.50	0.10	0.10(0.01)	0.00	0.99