Lecture Notes: Mixed Effects Models and Longitudinal Data Analysis

Jiming Jiang
University of California, Davis, U. S. A.
March 2005

1 Linear Mixed Models

1.1 Introduction

The term “mixed model” is sometimes confused with “mixture model”, although the two are not unrelated. The main difference is that, while a mixture model is often defined through conditional distributions, a mixed model almost always involves random effects.

There is no general consensus among mixed-model users on the roles that the random effects play in a mixed model. Some believe that the random effects represent unobserved variables of practical interest, which for good reasons should be considered random. This is what we call the first-type usage. Others use the random effects as ways of modelling the correlations among the observations, but are not interested in the random effects themselves. Such a usage is called the second-type. Robinson (1991) gives a wide-ranging account of the first-type usage of random effects. As for the second-type, one of the fast-developing areas is the analysis of longitudinal data (e.g., Laird and Ware 1982, Diggle et al. 1996). Several books have been published on mixed models in general. See, Rao and Kleffe (1988), Searle et al. (1992), Khuri et al. (1998), McCulloch and Searle (2000), among others.

\[\text{The notes are based on a tutorial lecture series given at the workshops on Semi-parametric Methods for Survival and Longitudinal Data held at the Institute for Mathematical Sciences, National University of Singapore, March 29-30, 2005.}\]
1.2 Some real-life examples

1.2.1 Effect of air pollution episodes on children

Laird and Ware (1982) discussed an example of analysis of the effect of air pollution episodes on pulmonary function. About two hundred school children were examined under normal conditions, then during an air pollution alert and on three successive weeks following the alert. One of the objectives was to determine whether the volume of air exhaled in the first second of a forced exhalation, denoted by FEV$_1$, was depressed during the alert.

Note that in this case the data is longitudinally collected with five observational times common for all the children. Laird and Ware proposed the following simple linear mixed model for analysis of the longitudinal data: $y_{ij} = \beta_j + \alpha_i + \epsilon_{ij}$, $i = 1, \ldots, m$, $j = 1, \ldots, 5$, where $\beta_j$ is the mean FEV$_1$ for the $j$th observational time, $\alpha_i$ is a random effect associated with the $i$th child, $\epsilon_{ij}$ is a error term and $m$ is the total number of children. It is assumed that the random effects are independent and distributed as $N(0, \tau^2)$, the errors are independent and distributed as $N(0, \sigma^2)$, and the random effects and errors are independent. It should be mentioned that some measurements were missing in this study. However, the above model can be modified to take this into account. In particular, the number of observations for the $i$th individual may be denoted by $n_i$, where $1 \leq n_i \leq 5$.

Based on the model, Laird and Ware were able to analyze the data using methods that will be described in the sequel, with the following findings: (i) a decline in mean FEV$_1$ was observed on and after the alert day; (ii) the variances and covariances for the last four measurements were larger than those involving the baseline day. The two authors further studied the problem of identification of sensitive subgroups or individuals most severely affected by the pollution episode using a more complicated linear mixed model. The detail will be given in Chapter 5.

1.2.2 Prediction of maize single-cross performance

Bernardo (1996) reported results of best linear unbiased prediction, or BLUP, of single-cross performance using a linear mixed model and genetic relationship among parental inbreds. Grain yield, moisture, stalk lodging and root lodging data were obtained for 2043 maize single crosses evaluated in the multilocation testing program of Limagrain Genetics, from 1990 to 1994. The objective of the study was to investigate the robustness of BLUP
for identifying superior single crosses when estimates of genetic relationship among inbreds are erroneous.

In analysis of the single-cross data, Bernardo proposed the following linear mixed model: \( y = X\beta + Z_0c + Z_1g_1 + Z_2g_2 + Zd + e \), where \( y \) is a vector of observed performance for a given trait (i.e., hybrid by multilocation trial means); \( \beta \) is a vector of fixed effects of multilocation yield trials; \( c \) is a vector of check effects; \( g_j \) is a vector of general combining ability effects of Group \( j \), \( j = 1, 2 \); \( d \) is a vector of specific combining ability effects; \( e \) is a vector of residual effects; and \( X, Z_j, j = 0, 1, 2 \) and \( Z \) are (known) incidence matrices of 0’s and 1’s. Here \( c, g_1, g_2, \) and \( d \) are treated as random effects. It is assumed that \( c, g_1, g_2, \) and \( e \) are uncorrelated with their covariance matrices modeled according to the genetic relationship.

Based on the linear mixed model, Bernardo was able to predict the performance of untested single crosses given the average performance of the tested ones, using the BLUP method (see §2.3.1 for detail). The results indicated that BLUP is robust when inbred relationships are erroneously specified. More discussion will come in Chapter 5.

1.2.3 Small area estimation of income

Large scale sample surveys are usually designed to produce reliable estimates of various characteristics of interest for large geographic areas. However, for effective planning of health, social and other services, and for apportioning government funds, there is a growing demand to produce similar estimates for smaller geographic areas and subpopulations for which adequate samples are not available. The usual design-based small-area estimators [e.g., Cochran (1977)] are unreliable since they are based on a very few observations that are available from the area. This makes it necessary to "borrow strength" from related small areas to find indirect estimators that increase the effective sample size and thus precision. Such indirect estimators are typically based on linear mixed models or generalized linear mixed models that provide a link to related small area through the use of supplementary data such as recent census data and current administrative records. See Ghosh and Rao (1994) for a review.

Among many examples of applications, Fay and Herriot (1979) used a linear mixed model for estimating per-capita income (PCI) for small places from the 1970 Census of Population and Housing. In the 1970 Census, income was collected on the basis of a 20 percent sample. However, of the estimates
required, more than one-third, or approximately 15,000, are for places with population of fewer than 500 persons. With such small populations, the sampling error of the direct estimates is quite significant. For example, Fay and Herriot estimated that for a place of 500 persons, the coefficient of variation of the direct estimate of PCI was about 13 percent; for a place of 100 persons, that coefficient increased to 30 percent. In order to “borrow strength” from related places and other sources, Fay and Herriot proposed the following linear mixed model: 

\[ y_i = x_i' \beta + v_i + e_i, \]

where \( y_i \) is the natural logarithm of the sample estimate of PCI for the \( i \)th place (the logarithm transformation stabilized the variance); \( x_i \) is a vector of known covariates related to the place; \( \beta \) is a vector of unknown regression coefficients; \( v_i \) is a random effect associated with the place; and \( e_i \) represents the sampling error. It is assumed that \( v_i \) and \( e_i \) are distributed independently such that \( v_i \sim N(0, A) \), \( e_i \sim N(0, D_i) \), where \( A \) is unknown but \( D_i \)'s are known.

1.3 Linear mixed model

The best way to understand a linear mixed model, or mixed linear model in some earlier literature, is to first recall a linear regression model. The latter can be expressed as \( y = X\beta + \epsilon \), where \( y \) is a vector of observations, \( X \) is a matrix of known covariates, \( \beta \) is a vector of unknown regression coefficients and \( \epsilon \) is a vector of (unobservable random) errors. In this model, the regression coefficients are considered fixed. However, there are cases in which it makes sense to assume that some of these coefficients are random. These cases typically occur when the observations are correlated. For example, in medical studies observations are often collected from the same individuals over times. It may be reasonable to assume that correlations exist among the observations from the same individual, especially if the times at which the observations are collected are relatively close. In animal breeding, lactation yields of dairy cows associated with the same sire may be correlated. In educational research, test scores of the same student may be related.

Now, let us see how a linear mixed model may be useful for modelling the correlations among the observations. Consider, for example, the example above regarding medical studies. Assume that each individual is associated with a random effect whose value is unobservable. Let \( y_{ij} \) denote the observation from the \( i \) individual collected at time \( t_j \), and \( \alpha_i \) the random effect associated with the \( i \)th individual. Assume that there are \( m \) individuals. For simplicity, let us assume that the observations from all individuals are col-
lected at a common set of times, say, \( t_1, \ldots, t_k \). Then, a linear mixed model may be expressed as \( y_{ij} = x'_{ij} \beta + \alpha_i + \epsilon_{ij}, \ i = 1, \ldots, m, \ j = 1, \ldots, k, \) where \( x_{ij} \) is a vector of known covariates; \( \beta \) is a vector of unknown regression coefficients; the random effects \( \alpha_1, \ldots, \alpha_m \) are assumed to be i.i.d. with mean zero and variance \( \sigma^2 \); the \( \epsilon_{ij} \)'s are errors which are i.i.d. with mean zero and variance \( \tau^2 \); and the random effects and errors are independent. It follows that the correlation between any two obervations from the same individual is \( \sigma^2/(\sigma^2 + \tau^2) \), while observations from different individuals are uncorrelated. This model is a special case of the so-called longitudinal linear mixed model, which will be discussed much in detail in the sequel. Of course, this is only a simple model and it may not capture all the correlations among the observations. Therefore, we would like to have a richer class of models that allow further complications.

A general linear mixed model may be expressed as

\[
y = X\beta + Z\alpha + \epsilon,
\]

where \( y \) is a vector of observations, \( X \) is a matrix of known covariates, \( \beta \) is a vector of unknown regression coefficients, which are often called the fixed effects, \( Z \) is known matrix, \( \alpha \) is a vector of random effects and \( \epsilon \) is a vector errors. Both \( \alpha \) and \( \epsilon \) are unobservable. Compared with the linear regression model, it is clear that the difference is \( Z\alpha \), which may take many different forms, thus creates a rich class of models, as we shall see. A statistical model must come with assumptions. The basic assumptions for (1.1) are that the random effects and errors have mean zero and finite variances. Typically, the covariance matrices \( G = \text{Var}(\alpha) \) and \( R = \text{Var}(\epsilon) \) involve some unknown dispersion parameters, or variance components. It is also assumed that \( \alpha \) and \( \epsilon \) are uncorrelated.

**Example 1.1 (One-way random effects model).** A model is called a random effects model if the only fixed effect is an unknown mean. Suppose that the observations \( y_{ij}, \ i = 1, \ldots, m, \ j = 1, \ldots, k_i \) satisfy \( y_{ij} = \mu + \alpha_i + \epsilon_{ij} \) for all \( i \) and \( j \), where \( \mu \) is an unknown mean; \( \alpha_i, \ i = 1, \ldots, m \) are random effects which are distributed independently as \( N(0, \sigma^2) \); \( \epsilon_{ij} \)'s are errors which are distributed independently as \( N(0, \tau^2) \); and the random effects are independent with the errors. Typically, the variances \( \sigma^2 \) and \( \tau^2 \) are unknown. To express the model in terms of (1.1), let \( y_i = (y_{ij})_{1 \leq j \leq k_i} \) be the (column) vector of observations from the \( i \)th group or cluster, and similarly \( \epsilon_i = (\epsilon_{ij})_{1 \leq j \leq k_i} \). Then, let \( y = (y_1', \ldots, y_m')', \ \alpha = (\alpha_i)_{1 \leq i \leq m} \) and \( \epsilon = (\epsilon_1', \ldots, \epsilon_m')' \). It is easy
to show that the model can be expressed as (1.1) with $\beta = \mu$ and suitable $X$ and $Z$, in which $\alpha \sim N(0, \sigma^2 I_m)$ and $\epsilon \sim N(0, \tau^2 I_n)$ with $n = \sum_{i=1}^{m} k_i$.

One special case is when $k_i = k$ for all $i$. This is called the balanced case. It can be shown that, in the balanced case, the model can be expressed as (1.1) with $X = 1_m \otimes 1_k = 1_{mk}$, $Z = I_m \otimes 1_k$ and everything else as before (with $k_i = k$). Note that in this case $n = mk$.

### 1.4 Types of linear mixed models

There are different types of linear mixed models, and different ways of classifying them. One way of classification is according to whether or not normality assumption is made. As will be seen, normality provides more flexibility in modelling, while models without normality are more robust to violation of distributional assumptions.

#### 1.4.1 Gaussian mixed models

Under a Gaussian linear mixed model, or simply Gaussian mixed model, both the random effects $\alpha$ and errors $\epsilon$ are assumed normal, and $\alpha$ and $\epsilon$ are uncorrelated. Alternatively, a Gaussian mixed model may be expressed by its marginal distribution. To see this, note that under (1.1) and normality, we have $y \sim N(X\beta, V)$ with $V = R + ZGZ'$, which gives an alternative definition of a Gaussian mixed model.

A disadvantage of the marginal model is that the random effects are not explicitly defined. In many cases, these random effects have practical meanings and the inference about them may be of interest. For example, in small area estimation (e.g., Ghosh and Rao 1994), the random effects are associated with the small area means, which are often of main interest.

#### 1.4.2 Non-Gaussian linear mixed models

Under non-Gaussian linear mixed models, the random effects and errors are assumed to be independent, or simply uncorrelated, but their distributions are not assumed to be normal. As a result, the (joint) distribution of the data may not be fully specified up to a set of parameters. The following are two specific cases.
1. **Mixed ANOVA model.** Following Jiang (1996), a non-Gaussian (linear) mixed ANOVA model is defined by (1.1) with

\[ Z \alpha = Z_1 \alpha_1 + \cdots + Z_s \alpha_s, \]

where \( Z_1, \ldots, Z_s \) are known matrices; the components of \( \alpha_i \) are i.i.d. with mean 0 and variance \( \sigma^2_i \), \( 1 \leq i \leq s \); the components of \( \epsilon \) are i.i.d. with mean 0 and variance \( \sigma^2_0 \); and \( \alpha_1, \ldots, \alpha_s, \epsilon \) are independent. Denote the common distribution of the components of \( \alpha_i \) by \( F_i \) (\( 1 \leq i \leq s \)) and that of the components of \( \epsilon \) by \( G \). If the parametric forms of \( F_1, \ldots, F_s, G \) are not assumed, the distribution of \( y \) is not specified up to a set of parameters. In fact, even if the parametric forms of the \( F_i \)'s and \( G \) are known, as long as they are not normal, the (joint) distribution of \( y \) may not have an analytic expression. The vector \( \theta \) is often defined as \( \theta = (\sigma^2_0, \sigma^2_1, \ldots, \sigma^2_s)' \).

2. **Longitudinal model.** A general non-Gaussian longitudinal mixed model is defined as follows. Let \( y_i \) be the vector of observations corresponding to the \( i \)th subject. Suppose that \( y_1, \ldots, y_m \) are independent such that

\[ y_i = X_i \beta + Z_i \alpha_i + \epsilon_i, \quad i = 1, \ldots, m, \]

where \( X_i \) and \( Z_i \) are known matrices; \( \beta \) is an unknown vector of regression coefficients; \( \alpha_i \) is a vector of random effects; and \( \epsilon_i \) is a vector of errors. It is assumed that \( \alpha_i, \epsilon_i, i = 1, \ldots, m \) are independent with \( \alpha_i \sim N(0, G_i), \epsilon_i \sim N(0, R_i) \), where the covariance matrices \( G_i \) and \( R_i \) are known up to a vector \( \theta \) of variance components.

**Example 1.2 (Growth curve model).** For simplicity, suppose that for each of the \( m \) individuals, the observations are collected over a common set of times \( t_1, \ldots, t_k \). Suppose that \( y_{ij} \), the observation collected at time \( t_j \) from the \( i \)th individual, satisfies \( y_{ij} = \xi_i + \eta_i x_{ij} + \zeta_{ij} + \epsilon_{ij} \), where \( \xi_i \) and \( \eta_i \) represent, respectively, a random intercept and a random slope; \( x_{ij} \) is a known covariate; \( \zeta_{ij} \) corresponds to a serial correlation; and \( \epsilon_{ij} \) is an error. For each \( i \), it is assume that \( \xi_i \) and \( \eta_i \) are jointly normally distributed with means \( \mu_1, \mu_2 \), variances \( \sigma^2_1, \sigma^2_2 \), respectively, and correlation coefficient \( \rho \); and \( \epsilon_{ij} \)'s are independent and distributed as \( N(0, \tau^2) \). As for the \( \zeta_{ij} \)'s, it is assumed that they satisfy the following relation of the first order autoregressive process, or AR(1): \( \zeta_{ij} = \phi \zeta_{ij-1} + \omega_{ij} \), where \( \phi \) is a constant such that \( 0 < \phi < 1 \), and \( \omega_{ij} \)'s are independent and distributed as \( N\{0, \sigma^2_3(1 - \phi^2)\} \). Furthermore, the three random components \( (\xi, \eta, \zeta) \) and \( \epsilon \) are independent, and observations from different individuals are independent. There is a slight departure of this
model from the standard linear mixed model in that the random intercept and slope may have nonzero means. However, by subtracting the means and thus defining new random effects, this model can be expressed in the standard form. In particular, the fixed effects are $\mu_1$ and $\mu_2$, and the (unknown) variance components are $\sigma_j^2$, $j = 1, 2, 3$, $\tau^2$, $\rho$ and $\phi$.

If normality is not assumed, but the assumptions about the means and covariance matrices of $\alpha_i$ and $\epsilon_i$ are maintained, one has a non-Gaussian longitudinal model.

1.5 Estimation of variance components

A main problem in the analysis of linear mixed models is the estimation of the variance components. In many cases (e.g., quantitative genetics), the variance components are of practical interest. In some other cases (e.g., longitudinal data analysis), the variance components themselves are not of main interest, but their existence affects quality of inference about other parameters, such as the fixed effects, which are of main interest. Some of the earlier methods in mixed model analysis did not require the normality assumption. These include the analysis of variance (ANOVA) method, or Henderson’s methods (Henderson 1953), and minimum norm quadratic unbiased estimation (MINQUE) method, proposed by C. R. Rao (e.g., Rao 1972). However, the ANOVA method is known to produce inefficient estimators of the variance components when the data is unbalanced. It also has the unpleasant feature of estimators sometimes falling outside the parameter space, such as negative estimators of variances. The MINQUE method, on the other hand, depends on some initial values of the variance components. If it is run iteratively, each time using the current values to update the initial values, the procedure, called I-MINQUE, is known to produce estimators that are identical to the REML estimators, if they belong to the parameter space (e.g., Searle et al. 1992, section 11.3). However, the latter property is not guaranteed.

If normality is assumed, the efficient estimators of the variance components are the maximum likelihood estimators (MLE). However, the latter had not been in serious use in linear mixed models until Hartley and Rao (1967). The main reason was that, unlike the ANOVA estimator, the MLE under a linear mixed model was not easy to handle computationally in the earlier days. There was also an issue of asymptotic behavior of the MLE, because, unlike the traditional i. i. d. case, the observations are correlated under a
linear mixed model. Both the computational and asymptotic issues were addressed by the Hartley-Rao paper. Asymptotic properties of the MLE were further studied by Miller (1977) for a wider class of linear mixed models.

The MLE of the variance components are, in general, biased. Such a bias will not vanish as the sample size increases, if the number of the fixed effects is proportional to the sample size. In fact, in the latter case the MLE will be inconsistent. The following is a well-known example.

Example 1.3 (The Neyman-Scott problem). Nayman and Scott (1948) gave the following example which shows that, when the number of parameters increases with the sample size, the MLE may not be consistent. Suppose that two observations are collected from \( m \) individuals. Each individual has its own (unknown) mean, say, \( \mu_i \) for the \( i \)th individual. Suppose that the observations are independent and normally distributed with variance \( \sigma^2 \). The problem of interest is to estimate \( \sigma^2 \). The model may be expressed as the following: \( y_{ij} = \mu_i + \epsilon_{ij} \), where \( \epsilon_{ij} \)'s are independent and distributed as \( N(0, \sigma^2) \). Note that this may be viewed as a special case of the linear mixed model (1.1), in which \( Z = 0 \). However, it can be shown that the MLE of \( \sigma^2 \) is inconsistent.

Furthermore, in some cases such as animal genetics the parameters of main interest are the variance components, while the fixed effects are considered as nuisance parameters. It would be nice to have a method that can focus on the variance components without having to simultaneously estimate the nuisance parameters. Thompson (1962) proposed a method, which was later put on a broader basis by Patterson and Thompson (1971), known as restricted or residual maximum likelihood, or REML in either case. The idea is to consider a transformation of the data that is orthogonal to the design matrix of the fixed effects, that is, \( X \). In other words, each component of such a transformation is an error contrast. To formulate the REML procedure, let the dimensions of \( y \) and \( \beta \) be \( n \) and \( p \), respectively. Without loss of generality, assume that \( \text{rank}(X) = p \). Let \( A \) be a \( n \times (n - p) \) matrix of full rank such that \( A'X = 0 \). The REML estimators of the variance components are simply the MLE based on \( z = A'y \). It is easy to show that the REML estimators do not depend on the choice of \( A \). Furthermore, several authors have argued that there is no loss of information by doing REML in estimating the variance components (e.g., Patterson and Thompson 1971, Harville 1977, Jiang 1996). Different derivations of REML have been given by Harville (1974), Barndorff-Nielsen (1983), Verbyla (1990), Heyde (1994), and Jiang (1996). In addition, several authors have written review articles
on REML, see Harville (1977), Khuri and Sahai (1985), Robinson (1987) and Speed (1997).

1.5.1 Inference about non-Gaussian linear mixed models

The ML and REML methods are developed under the normality assumption. However, the normality assumption is likely to be violated in real life. For example, Lange and Ryan (1989) gave several examples showing that non-normality of the random effects is, indeed, encountered in practice. Due to such concerns as well as that about efficiency, which the ANOVA estimator lacks, some researchers have taken a quasi-likelihood approach. The idea is to use the Gaussian ML or REML estimators in nonnormal situations. See Richardson and Welsh (1994), Jiang (1996, 1997a), Heyde (1994, 1997), among others. Throughout this review, these estimators will also be called ML and REML estimators, even if normality does not hold. Jiang (1996, 1997a) established consistency and asymptotic normality of REML estimators in nonnormal situations. In exactly the same situation, the author also gave necessary and sufficient conditions for the ML estimators to be consistent and asymptotically normal. These conditions are in terms of the rate at which \( p \) goes to infinity with \( n \). In particular, Jiang (1996) derived the asymptotic covariance matrix (ACM) of the REML estimator of the variance components. See Jiang (1998b) for the ML analogue. The ACM is important to various inferences about the model, including interval estimation and hypothesis testing. Unfortunately, the ACM under nonnormality involves parameters other than the variance components, for example, the third and fourth moments of the random effects. Note that standard procedures such as ML and REML do not produce estimators of these additional parameters.

To see where exactly the problem occurs, consider the mixed ANOVA model (see §1.4.2.1). Without loss of generality, consider the Hartley-Rao form of variance components (Hartley and Rao 1967): \( \lambda = \sigma^2_j \), \( \gamma_j = \sigma^2_j / \sigma^2_0 \), \( 1 \leq j \leq s \). Let \( \theta = (\lambda, \gamma_1, \ldots, \gamma_s)' \). According to Jiang (1996), the ACM of the REML estimator, \( \hat{\theta} \), is given by

\[
\Sigma_R = \left\{ \text{E} \left( \frac{\partial^2 l_R}{\partial \theta \partial \theta'} \right) \right\}^{-1} \text{Var} \left( \frac{\partial l_R}{\partial \theta} \right) \left\{ \text{E} \left( \frac{\partial^2 l_R}{\partial \theta \partial \theta'} \right) \right\}^{-1},
\]

where \( l_R \) is the Gaussian restricted log-likelihood function. The matrix \( \mathcal{I}_2 = \text{E}(\partial^2 l_R / \partial \theta \partial \theta') \) only depends on \( \theta \), whose estimator is already available.
However, unlike $\mathcal{I}_2$, the matrix $\mathcal{I}_1 = \text{Var}(\partial l_r / \partial \theta)$ depends on, in addition to $\theta$, the kurtoses of the random effects and errors.

It is clear that the key issue is how to estimate $\mathcal{I}_1$, which we call quasi information matrix (QUIM) for an obvious reason. If $l_R$ were the true restricted log-likelihood, then we would have $\mathcal{I}_1 = -\mathcal{I}_2$, which is the Fisher information matrix. Traditionally, there are two ways to estimate the Fisher information. They are (i) the estimated information and (ii) the observed information. See, for example, Efron and Hinkley (1978) for a discussion and comparison of the two methods in the i.i.d. case. However, as explained by Jiang (2005), neither method applies to the above non-Gaussian linear mixed models. The author further proposed a method called partially observed information to solve the problem. To explain the approach, let us first consider a simple example.

**Example 1.4.** Consider the following model with crossed random effects:

$$ y_{ij} = \mu + v_i + w_j + e_{ij}, \quad i = 1, \ldots, m, \quad j = 1, \ldots, n, $$

where $\mu$ is an unknown mean, $v_i$, $w_j$ are random effects, and $e_{ij}$ is an error. It is assumed that the $v_i$'s are i.i.d. with mean 0 and variance $\sigma_1^2$, $w_j$'s are i.i.d. with mean 0 and variance $\sigma_2^2$, $e_{ij}$'s are i.i.d. with mean 0 and variance $\sigma_0^2$, and $v$, $w$, $e$ are independent. Consider an element of the QUIM $\text{Var}(\partial l_R / \partial \theta)$ for REML estimation, say, $\text{var}(\partial l_R / \partial \lambda)$, where $l_R$ is the restricted Gaussian log-likelihood and $\theta = (\lambda, \gamma_1, \gamma_2)'$ ($\lambda$ and $\gamma$'s are defined earlier). It can be shown that $\text{var}(\partial l_R / \partial \lambda) = S_1 + S_2$, where

$$ S_1 = E \left\{ (a_0 + a_1 + a_2) \sum_{i,j} u_{ij}^4 \sum_i u_{ij} \left( \sum_j u_{ij} \right)^4 \right\}, $$

while $S_2$ and $a_r, \ r = 0, 1, 2$ depend only on $\theta$. It is clear that $S_2$ can be estimated by replacing the variance components by their REML estimators, which are already available. As for $S_1$, it cannot be estimated in the same way. However, the form of $S_1$ suggests an “observed” estimator by taking out the expectation sign and replacing the parameters involved by their REML estimators. In fact, as $m, n \to \infty$, this observed $S_1$, say, $\hat{S}_1$, is consistent in the sense that $\hat{S}_1 / S_1 \to 1$ in probability. It is interesting to note that $S_2$ cannot be consistently estimated by an observed form. In conclusion, $\hat{S}_1$ cannot be estimated by an estimated form but can be estimated by an
observed form; $S_2$ can be estimated by an estimated form but not by an observed form.

In general, write $I_{1,jk} = \text{cov} (\partial l_R / \partial \theta_j, \partial l_R / \partial \theta_k)$, which is the $j,k$ element of the QUIM $I_1 = \text{Var} (\partial l_R / \partial \theta)$. Let $V$ denote the covariance matrix of $y$. The kurtoses of the random effects and errors are defined as $\kappa_t = E (\alpha_t^4) - 3 \sigma_t^4$, $0 \leq t \leq s$, where $\alpha_0 = \epsilon$. Let $x_i'$ be the $i$th row of $X$, and $z_i'$ and $z_{i}'$ be the $i$th row and $l$th column of $Z_l$, respectively. Let $u_i = y_i - x_i' \beta$, where $y_i$ is the $i$th component of $y$. Let $f_1, \ldots, f_L$ be the different nonzero functional values of the following, considered as a function of $\kappa$:

$$f(i_1, \ldots, i_4) = \sum_{t=0}^{s} \kappa_t z_{i_1} \cdots z_{i_4}.$$  

**Theorem 1.1.** For any non-Gaussian mixed ANOVA model, we have

$$I_{1,jk} = E \left\{ \sum_{f(i_1, \ldots, i_4) \neq 0} c_{j,k}(i_1, \ldots, i_4) u_{i_1} \cdots u_{i_4} \right\}$$

$$+ \left\{ 2 \text{tr}(B_j V B_k V) - 3 \lambda^2 \sum_{f(i_1, \ldots, i_4) \neq 0} c_{j,k}(i_1, \ldots, i_4) \Gamma(i_1, i_3) \Gamma(i_2, i_4) \right\}$$

$$= I_{1,1,jk} + I_{1,2,jk},$$

$0 \leq j, k \leq s$, where $c_{j,k}(i_1, \ldots, i_4) = c_{j,k,l}$, if $f(i_1, \ldots, i_4) = f_l, 1 \leq l \leq L$ with

$$c_{j,k,l} = \frac{1}{|\{f(i_1, \ldots, i_4) = f_l\}|} \sum_{f(i_1, \ldots, i_4) = f_l} B_{j,i_1,i_2} B_{k,i_3,i_4}.$$ 

$B_j, 0 \leq j \leq s$ are matrix-valued functions of $\theta$, $B_{j,i_1,i_2}$ is the $(i_1, i_2)$ element of $B_j$, and $\Gamma(i_1, i_3)$, etc. are functions of $\theta$.

Note that $2\text{tr}(B_j V B_k V)$ is the Gaussian covariance between $\partial l_R / \partial \theta_j$ and $\partial l_R / \partial \theta_k$. This means that, under normality, $I_{1,1,jk}$ is identical to the second term in $I_{1,2,jk}$ without the negative sign. However, without normality, $I_{1,1,jk}$ may involve higher moments of the random effects and errors, and this is why the expectation is not taken inside the summation. Instead, we propose to estimate $I_{1,1,jk}$ by taking out the expectation sign, and replacing any parameter involved by its REML estimator. More precisely, we define

$$\hat{I}_{1,1,jk} = \sum_{f(i_1, \ldots, i_4) \neq 0} \hat{c}_{j,k}(i_1, \ldots, i_4) \hat{u}_{i_1} \cdots \hat{u}_{i_4},$$

12
where \( \hat{c}_{j,k}(i_1, \ldots, i_4) \) is defined in the same way as \( c_{j,k}(i_1, \ldots, i_4) \) except with \( \theta \) replaced by \( \hat{\theta} \), and \( \hat{u}_i = y_i - x_i' \hat{\beta} \). Here \( \hat{\beta} = (X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}y \), and \( \hat{V} \) is \( V \) with \( \theta \) replaced by \( \hat{\theta} \). This is the observed part. On the other hand, \( I_{1,2,jk} \) only depends on \( \theta \), and therefore can be estimated by replacing \( \theta \) by \( \hat{\theta} \). The result, denoted by \( \hat{I}_{1,2,jk} \), is the estimated part.

An estimator of \( I_{1,jk} \) is then \( \hat{I}_{1,1,jk} + \hat{I}_{1,2,jk} \), hence an estimator of \( I_1 \) is given by \( \hat{I}_1 = \hat{I}_{1,1} + \hat{I}_{1,2} \), where \( \hat{I}_{1,r} = (\hat{I}_{1,r,jk})_{0 \leq j,k \leq s} \), \( r = 1,2 \). Because the estimator consists partially of an observed form and partially of an estimated one, it is called partially observed quasi information matrix, or POQUIM. For more detail, see Jiang (2005).

### 1.6 Prediction of random effects

The prediction of random effects, or mixed effects in a more general context, has a long history dating back to C. R. Henderson in his early work in the field of animal breeding (e.g., Henderson 1948). The best-known method for the prediction of mixed effects is best linear unbiased prediction, or BLUP. Robinson (1991) gives a wide-ranging account of BLUP with examples and applications. Also see Ghosh and Rao (1994) for applications in small area estimation.

A mixed effect may be expressed as \( \eta = b'\beta + a'\alpha \), where \( a, b \) are known vectors. If the fixed effects and variance components are both known, the best predictor for \( \eta \), under the normality assumption, is given by \( \hat{\eta} \), where \( \hat{\eta} \) is \( \eta \) with \( \alpha \) replace by \( E(\alpha|y) \), i.e.,

\[
\hat{\alpha} = GZ'V^{-1}(y - X\beta).
\]

Without assuming normality, \( \hat{\eta} \) is the best linear predictor of \( \eta \) in the sense that it minimizes the mean squared error (MSE) of a predictor that is linear in \( y \) (e.g., Searle et al. 1992, §7.3). Of course, \( \beta \) is unknown in practice. It is then customary to replace \( \beta \) by

\[
\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y,
\]

which is the MLE of \( \beta \) under normality, provided that \( \theta \) is known. The result is BLUP, in other words, the BLUP of \( \eta \) is given by \( \hat{\eta} \) with \( \beta \) replaced by \( \hat{\beta} \). The original derivation of BLUP was given by Henderson (1950), in which he proposed to find the “maximum likelihood estimates” of the random effects.
Of course, these are not the MLE in the usual sense, because the random effects are different from fixed parameters. Later, in Henderson (1973), he showed that BLUP is, indeed, what is meant by the name, that is, (i) it is linear in \( y \); (ii) its expected value is equal to that of \( \eta \); and (iii) it minimizes the MSE among all linear unbiased predictors of \( \eta \). Different derivations of BLUP were also given by Harvill (1990) and Jiang (1997b), among others.

Robinson (1991) used the following example to illustrate the calculation of BLUP.

Example 1.5. Consider a linear mixed model for the first lactation yields of dairy cows with sire additive genetic merits being treated as random effects and herd effects being treated as fixed effects. The herd effects are represented by \( \beta_j \), \( j = 1, 2, 3 \) and sire effects by \( \alpha_i \), \( i = 1, 2, 3, 4 \), which correspond to sires A, B, C, D. The matrix \( R \) is taken to be the identity matrix, while the matrix \( G \) is assumed to be 0.1 times the identity matrix. This would be a reasonable assumption, provided that the sires were unrelated and that the variance ratio \( \sigma^2/\tau^2 \) had been estimated previously, where \( \sigma^2 = \text{var}(\alpha_i) \) and \( \tau^2 = \text{var}(\epsilon_{ij}) \). Suppose that the data is given below.

<table>
<thead>
<tr>
<th>Herd</th>
<th>1</th>
<th>1</th>
<th>2</th>
<th>2</th>
<th>2</th>
<th>3</th>
<th>3</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sire</td>
<td>A</td>
<td>D</td>
<td>B</td>
<td>D</td>
<td>D</td>
<td>C</td>
<td>C</td>
<td>D</td>
</tr>
<tr>
<td>Yield</td>
<td>110</td>
<td>100</td>
<td>110</td>
<td>100</td>
<td>100</td>
<td>110</td>
<td>110</td>
<td>100</td>
</tr>
</tbody>
</table>

The BLUP for a mixed effect \( \eta = b'\beta + a'\alpha \) is given by \( \hat{\eta} = b'\hat{\beta} + a'\hat{\alpha} \), where

\[
\hat{\beta} = (105.64, 104.28, 105.46)', \\
\hat{\alpha} = (0.40, 0.52, 0.76, -1.67)'.
\]

The expression of BLUP involves \( \theta \), the vector of variance components, which is typically unknown in practice. It is customary to replace \( \theta \) by a consistent estimator, \( \hat{\theta} \). The resulting predictor is often called empirical BLUP, or EBLUP, denoted by \( \hat{\eta} \). Kackar and Harville (1981) showed that, if \( \hat{\theta} \) is an even and translation invariant estimator and the data is normal, EBLUP remains unbiased. Some of the well-known estimators of \( \theta \), including ANOVA, ML and REML estimators, are even and translation invariant. In their arguments, however, Kackar and Harville had assumed existence of the expected value of EBLUP, which is not obvious. The existence of the expected value of EBLUP was later proved by Jiang (2000a). Asymptotic behavior of EBLUP was studied by Jiang (1998b).
Harville (1991) considered a one-way random effects model and showed
that, in this case, the EBLUP is identical to an empirical Bayes (EB) es-
timator. He further noted that while much of the work on EB had been
carried out by professional statisticians in relatively simple cases, such as the
one-way random effects model, EBLUP has been used by practitioners under
relatively complex models. One of the areas in which EBLUP has been exten-
sively used, and studied, is small area estimation. See, for example, Ghosh
and Rao (1994) and Rao (2003). While EBLUP is fairly easy to calculate, its
MSE is complicated. On the other hand, estimation of the MSE of EBLUP is
of significant practical interest. For example, over $7 billion dollars of funds
are allocated annually based on EBLUP estimators of school-age children in
poverty at the county and school district levels (National Research Council
2000). Kackar and Harville (1984) gave an approximation to the MSE of
EBLUP under the linear mixed model (1), taking account of the variability
in $\hat{\theta}$, and proposed an estimator of $\text{MSE}(\hat{\eta})$ based on this approxima-
tion. But the approximation is somewhat heuristic, and the accuracy of the
approximation and the associated MSE estimator was not studied. Prasad and
Rao (1990) studied the accuracy of a second order approximation to $\text{MSE}(\hat{\eta})$
for two important special cases of linear mixed models that are frequently
used in small area estimation, that is, the Fay-Herriot model (Fay and Her-
riot 1979) and nested error regression model (e.g., Battese et al. 1988). The
results were extended by Das et al. (2004) to general linear mixed models.
For example, for Gaussian mixed ANOVA models with REML estimation of
$\theta$, the latter authors showed that $\text{MSE}(\hat{\eta}) = g_1(\theta) + g_2(\theta) + g_3(\theta) + o(d_*^{-2})$, where

$$
g_1(\theta) = a'(G - GZV^{-1}ZG)a,
g_2(\theta) = (b - X'V^{-1}ZGa)'(X'V^{-1}X)^{-1}(b - X'V^{-1}ZGa),
g_3(\theta) = tr \left[ \left\{ (\partial/\partial \theta')V^{-1}ZGa \right\}'V \left\{ (\partial/\partial \theta')V^{-1}ZGa \right\} H^{-1} \right],
$$

where $H = E(\partial^2 l_\theta / \partial \theta \partial \theta')$, $d_* = \min_{1 \leq i \leq n} d_i$ with $d_i = ||Z_i'PZ_i||_2$ and

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

Based on the approximation, the authors obtained an estimator of $\text{MSE}(\hat{\eta})$
whose bias is corrected to second order. More specifically, an estimator
$\widehat{\text{MSE}}(\hat{\eta})$ was obtained such that $E\{\widehat{\text{MSE}}(\hat{\eta})\} = \text{MSE}(\hat{\eta}) + o(d_*^{-2})$. See Das et
al. (2004) for details.
Alternatively, Jiang, Lahiri and Wan (2002) proposed a jackknife method which led to second order approximation and estimation of the MSE of EBLUP in the case of longitudinal linear mixed models. Denote MSE(\hat{\eta}) by \hat{b}(\theta). The jackknife estimator of the MSE of \hat{\eta} is given by \hat{\text{MSE}}(\hat{\eta}) = \text{MSE}^{\text{MSE}}(\hat{\eta}) + \text{MSE}(\hat{\eta}), where

\[
\text{MSE}^{\text{MSE}}(\hat{\theta}) = \frac{m - 1}{m} \sum_{i=1}^{m} (\hat{\eta}_{-i} - \hat{\eta})^2 ,
\]

\[
\hat{\text{MSE}}(\hat{\eta}) = b(\hat{\theta}) - \frac{m - 1}{m} \sum_{i=1}^{m} \{ b(\hat{\theta}_{-i}) - b(\hat{\theta}) \} .
\]

Here \( m \) represents the number of clusters (e.g., number of small areas), \( \hat{\theta}_{-i} \) denotes an M-estimator of \( \theta \) using the data without the \( i \)th cluster (e.g., the \( i \)th small area), and \( \hat{\eta}_{-i} \) the EBLUP of \( \eta \) in which the fixed parameters are estimated using the data without the \( i \)th cluster. It was shown that \( \text{E}\{\hat{\text{MSE}}(\hat{\eta})\} = \text{MSE}(\hat{\eta}) + o(m^{-1}) \), that is, the bias of the MSE estimator is corrected to the second order. The result holds, in particular, when \( \hat{\theta} \) is either the REML or ML estimator, and also for non-Gaussian (longitudinal) linear mixed models. In fact, the jackknife method also applies to longitudinal generalized linear mixed models, in which EBLUP is replaced by empirical best predictor (EBP). See Jiang, Lahiri and Wan (2002) for detail.

### 1.7 Other types of inference

Most of the literature on analysis of linear mixed models has been focusing on estimation and prediction problems. However, virtually every other types of analysis were also considered. These include:

(i) **Bayesian inference.** A linear mixed model can be naturally formulated as a hierarchical model under the Bayesian framework. Such a model usually consists of three levels, or stages of hierarchies. At the first stage, a linear model is set up given the fixed and random effects; at the second stage, the distribution of the fixed and random effects is specified given the variance component parameters; finally, at the last stage, a prior distribution is assumed for the variance components. The inference includes that about the fixed and random effects and that about the variance components. See, for example, Hill (1965), Tiao and Tan (1965, 1966), Gianola and Fernando (1986), and Gelman et al. (2003).
(ii) Tests in linear mixed models. Various statistical tests under Gaussian mixed models were discussed, e.g., in Khuri et al. (1998). For non-Gaussian linear mixed models, Jiang (2003a) developed an empirical method of moments and use it to tests of hypotheses regarding the dispersion parameters; Jiang (2005) proposed robust dispersion tests using the POQUIM method.

(iii) Confidence intervals. Burdick and Graybill (1992) discussed confidence intervals in Gaussian mixed models. For non-Gaussian linear mixed models, large sample confidence intervals may be obtained using the POQUIM method developed in Jiang (2005).

(iv) Prediction intervals. Jeske and Harville (1988) considered prediction intervals for a mixed effect, assuming that the joint distribution of α and y − E(y) is known up to a vector of unknown parameters. Jiang and Zhang (2002) proposed a distribution-free method for constructing prediction intervals for a future observation under a non-Gaussian linear mixed model.

(v) Mixed model diagnostics. Some authors have proposed to use EBLUP or EB as tools for informal checking of the distributional assumptions regarding the random effects. See, for example, Dempster and Ryan (1985), Lange and Ryan (1989), and Calvin and Sedransk (1991). As for formal model checking, some methods were recently developed. Jiang, Lahiri and Wu (2001) developed an asymptotic theory of Pearson's χ²-test with estimated cell frequencies for formally checking the distributional assumptions. The procedure requires splitting the data into two parts, one used for estimation and the other for testing. Jiang (2001a) developed a χ² goodness-of-fit test that applies to a general mixed ANOVA model (2), which does not require data splitting.

(vi) Mixed model selection. In a way, model selection and estimation are viewed as two components of a process called model identification. While there is extensive literature on parameter estimation in linear mixed models, the other component, that is, mixed model selection, has received little attention. Only recently have some results emerge. See Jiang and Rao (2003). Clearly, further research is needed in this area.
2 Generalized linear mixed models

2.1 Introduction

For the most part, linear mixed models have been used in situations where the observations are continuous. However, there are cases in practice where the observations are discrete, or categorical. McCullagh and Nelder (1989) proposed an extension of linear models, call generalized linear models, or GLM. They noted that the key elements of a classical linear model, that is, a linear regression model, are that (i) the observations are independent; (ii) the mean of the observation is a linear function of some covariates; and (iii) the variance of the observation is a constant. The extension to GLM consists of modification of (ii) and (iii) above, by that (ii)' the mean of the observation is associated with a linear function of some covariates through a link function; and (iii)' the variance of the observation is a function of the mean. GLM include a variety of models that include normal, binomial, Poisson and multinomial as special cases. Therefore, these models are applicable to cases where the observations may not be continuous.

One element that GLM have in common with linear models is that the observations are assumed to be independent. In many cases, however, the observations, or responses, are correlated, as well as discrete or categorical. For example, if $y_{i1}, \ldots, y_{i10}$ indicate whether or not the $i$th individual (person) visited a doctor during each of the past 10 years, that is, $y_{ij} = 1$ if the $i$th individual visited a doctor within the $j$th year in the past, and $y_{ij} = 0$ otherwise, then the responses from the same individual are likely to be correlated. On the other hand, the responses are binary instead of continuous. As mentioned earlier, the linear mixed models discussed in the previous section do not apply to such cases. It is clear now that what one needs is an extension of the linear mixed model to cases where the responses are both correlated and, at the same time, discrete or categorical.

2.2 Examples of applications

2.2.1 Foetal mortality in mouse litters

Brooks et. al. (1997) presented six data sets recording foetal mortality in mouse litters. Here we consider the HS2 dataset from Table 4 of their paper, which reports the number of dead implants in 1328 litters of mice from untreated experimental animals.
The data may be considered as being summaries of the individual responses $y_{ij}$, $i = 1, \ldots, 1328$, $j = 1, \ldots, n_i$, where $n_i$ is the size of the $i$th litter; $y_{ij} = 1$ if the $j$th implant in the $i$th litter is dead, and $y_{ij} = 0$ otherwise. The total number of responses is $N = \sum_{i=1}^{1328} n_i = 10533$. For simplicity, the $n_i$'s are considered nonrandom.

Brooks et. al. (1997) use a beta-binomial model to model the correlation among responses from the same litter. Alternatively, one may use a mixed logistic model, which is a special case of generalized linear mixed model, to model the correlation. See Jiang and Zhang (2001) for detail.

2.2.2 The salamander mating data

McCullagh & Nelder (1989, §14.5) have published a dataset involving samples from mating experiments of two salamander populations, Rough Butt and Whiteside. Three experiments were conducted during 1986, one in the summer and two in the autumn. In each experiment there were 10 males and 10 females from each population. They were paired according to the design given by Table 14.3 in McCullagh & Nelder (1989). The same 40 salamanders were used for the summer and first autumn experiments. A new set of 40 animals was used in the second autumn experiment. For each pair, it was recorded whether a mating occurred, 1, or not, 0.

The responses are binary and clearly correlated, so that neither linear mixed models nor generalised linear models apply. McCullagh & Nelder (1989) proposed the following mixed logistic model with crossed random effects. For each experiment, let $u_i$ and $v_j$ be the random effects corresponding to the $i$th female and $j$th male involved in the experiment. Then, on the logistic scale, the probability of successful mating is modelled in term of fixed effects $+u_i+v_j$. It was further assumed that the random effects are independent and normally distributed with means 0 and variances $\sigma^2$ for the females and $\tau^2$ for the males. The problem then is to estimate the fixed effects and the variances $\sigma^2$ and $\tau^2$.

For more applications, see, for example, Breslow and Clayton (1993), Lee and Nelder (1996), Malec et al. (1997), Ghosh et al. (1998) and McCulloch and Searle (2000).
2.3 Definition of GLMM

Suppose that given a vector of random effects, \( \alpha \), the responses, \( y_1, \ldots, y_n \) are conditionally independent such that the conditional distribution of \( y_i \) given \( \alpha \) is a member of the exponential family with pdf

\[
f_i(y_i|\alpha) = \exp \left\{ \frac{y_i \xi_i - b(\xi_i)}{a_i(\phi)} + c_i(y_i, \phi) \right\},
\]

where \( b(\cdot), a_i(\cdot), c_i(\cdot, \cdot) \) are known functions, and \( \phi \) is a dispersion parameter which may or may not be known. According to the properties of the exponential family, the quantity \( \xi_i \) is associated with the conditional mean \( \mu_i = \mathbb{E}(y_i|\alpha) \), which, in turn, is associated with a linear predictor

\[
\eta_i = x_i' \beta + z_i' \alpha,
\]

where \( x_i \) and \( z_i \) are known vectors and \( \beta \) a vector of unknown parameters (the fixed effects), through a known link function \( g(\cdot) \) such that

\[
g(\mu_i) = \eta_i.
\]

Furthermore, it is assumed that \( \alpha \sim N(0, G) \), where the covariance matrix \( G \) may depend on a vector \( \theta \) of unknown variance components. Note that, according to the properties of exponential family, one has \( b'(\xi_i) = \mu_i \). In particular, under the so-called canonical link, one has \( \xi_i = \eta_i \), that is, \( g = h^{-1} \), where \( h(\cdot) = b'(\cdot) \).

2.4 Likelihood function under GLMM

Despite the usefulness of GLMMs, inference about these models has encountered some difficulties. This is because, unlike linear mixed models, the likelihood function under a GLMM typically does not have a closed-form expression (with, of course, the exception of the normal case). In fact, such a likelihood may involve high dimensional integrals which cannot be evaluated analytically. To understand the computational difficulties, consider the following simple example.

Example 2.1. Suppose that, given the random effects \( u_1, \ldots, u_{m_1} \) and \( v_1, \ldots, v_{m_2} \), binary responses \( y_{ij}, i = 1, \ldots, m_1, j = 1, \ldots, m_2 \) are conditionally independent such that, with \( p_{ij} = P(y_{ij} = 1|u, v) \), \( \text{logit}(p_{ij}) = \mu + u_i + v_j \),

20
where $\mu$ is an unknown parameter, $u = (u_i)_{1 \leq i \leq m_1}$ and $v = (v_j)_{1 \leq j \leq m_2}$. Furthermore, the random effects $u_1, \ldots, u_{m_1}$ and $v_1, \ldots, v_{m_2}$ are independent such that $u_i \sim N(0, \sigma_1^2)$, $v_j \sim N(0, \sigma_2^2)$, where the variances $\sigma_1^2$ and $\sigma_2^2$ are unknown. Thus, the unknown parameters involved in this model are $\psi = (\mu, \sigma_1^2, \sigma_2^2)'$. It can be shown that the likelihood function under this model for estimating $\psi$ can be expressed as

$$
c - \frac{m_1}{2} \log(\sigma_1^2) - \frac{m_2}{2} \log(\sigma_2^2) + \mu y_\cdot + \log \int \cdots \int \prod_{i=1}^{m_1} \prod_{j=1}^{m_2} \{1 + \exp(\mu + u_i + v_j)\}^{-1} \times \exp \left( \sum_{i=1}^{m_1} u_i y_{i}\cdot + \sum_{j=1}^{m_2} v_j y_{j}\cdot - \frac{1}{2\sigma_1^2} \sum_{i=1}^{m_1} u_i^2 - \frac{1}{2\sigma_2^2} \sum_{j=1}^{m_2} v_j^2 \right) \, du_1 \cdots du_{m_1} dv_1 \cdots dv_{m_2},$$

where $c$ is a constant, $y_\cdot = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} y_{ij}$, $y_{i}\cdot = \sum_{j=1}^{m_2} y_{ij}$ and $y_{j}\cdot = \sum_{i=1}^{m_1} y_{ij}$. The multi-dimensional integral involved has no closed-form expression, and it cannot be further simplified. Furthermore, such an integral is difficult to evaluate even numerically. For example, if $m_1 = m_2 = 40$, the dimension of the integral will be 80. To make it even worse, the integrand involves a product of 1600 terms with each term less than one. Such a product is numerically zero, making it difficult to evaluated with Monte-Carlo method. Because of the computational difficulties, there have been two main approaches in inference about GLMM. The first focuses on developing computational method for the maximum likelihood. The second tries to avoid the computational difficulties of the likelihood-based inference by considering approximate or other alternative methods.

### 2.5 Monte-Carlo EM algorithm

For relatively simple models, the likelihood function may be evaluated by numerical integration techniques. However, numerical integration is generally intractable in GLMM if the dimension of integrals involved is greater than two. Alternatively, the integrals may be evaluated by Monte-Carlo methods. It should be pointed out that, for problems involving irreducibly high-dimensional integrals, naive Monte-Carlo usually does not work. For example, the high-dimensional integral in Example 2.1 cannot be evaluated
by a naive Monte-Carlo method. This is because a product of 1,600 terms
with each term less than one is numerically zero. Thus, an i. i. d. sum of
such terms will not yield anything but zero without a huge simulation size!

McCulloch (1994) proposed a Monte-Carlo EM algorithm for evaluating
the likelihood function under a threshold model, which is a special case of
GLMM with binary responses. A key element in the EM algorithm is the so-
called “complete data”. Usually, this consists of the observed data, denoted
by y, and some unobserved random variables, denoted by ξ. For example,
ξ may be a vector of missing observations; or a vector of random effects.
The idea is to choose ξ appropriately so that maximum likelihood becomes
trivial for the complete data. Let w = (y, ξ) denote the complete data,
which is assumed to have a probability density f(w|θ) depending on a vector
θ of unknown parameters. In the E-step of the algorithm, one computes the
conditional expectation

$$Q(θ|θ^{(k)}) = \text{E} \left\{ \log f(w|θ)|y, θ^{(k)} \right\},$$

where θ^{(k)} is the estimated θ at step k (the current step). Note that Q
is a function of θ. Then, in the M-step, one maximizes Q(θ|θ^{(k)}) with re-
spect to θ to obtain the next step estimator θ^{(k+1)}. The process is iterated
until convergence. In GLMM, it is the E-step that causes difficulties. Mc-
Culloch (1994) used Gibbs sampler to approximate the E-step. Later, in
McCulloch (1997), the author improved his earlier algorithm by replacing
the Gibbs sampler with a Metropolis-Hastings algorithm (e. g., Gelman
et al. 2003) to fit more general models. In addition, the author proposed
a Monte-Carlo analogue of the Newton-Raphson algorithm (MCNR), again
based on a Metropolis-Hastings algorithm, and a method called simulated
maximum likelihood (SML) based on importance sampling (see below). Sim-
ulation results showed that SML worked poorly compared to MCEM and
MCNR.

Subsequently, Booth and Hobert (1999) proposed two Monte-Carlo EM
algorithms. Unlike McCulloch (1994, 1997), Booth and Hobert used i. i.
d. sampling to construct Monte-Carlo approximations at the E-step. More
specifically, the authors used two methods to generate the Monte-Carlo sam-

1. Importance sampling. The idea of importance sampling is the follow-
ing. Suppose that one needs to evaluate an integral of the form

\[ I(f) = \int f(x)dx \]

for some function \( f(x) \geq 0 \). Note that the integral may be expressed as

\[ I(f) = \int \frac{f(x)}{h(x)}h(x)dx = E \left\{ \frac{f(\xi)}{h(\xi)} \right\}, \]

where \( h \) is a pdf such that \( h(x) > 0 \) if \( f(x) > 0 \), and \( \xi \) is a random variable with pdf \( h \). Thus, if one can generate a sequence of i. i. d. samples \( \xi_1, \ldots, \xi_K \) with the pdf \( h \), one can approximate the integral by

\[ E_p(f) \approx \frac{1}{K} \sum_{k=1}^{K} \frac{f(\xi_k)}{h(\xi_k)}. \]

See Gelman et al. (2003) for more detail. Note that the E-step is all about the calculation of \( Q\{\psi|\psi^{(l)}\} = E[\log\{f(y, \alpha|\psi)\}|y; \psi^{(l)}] \), where \( \psi = (\beta', \phi, \theta')' \) and \( l \) represents the current step. The expected value is computed under conditional distribution of \( \alpha \) given \( y \), which has density

\[ f(\alpha|y; \psi) \propto f(y|\alpha; \beta, \phi)f(\alpha|\theta). \]

There is a normalizing constant involved in the above expression, which is the (marginal) density function \( f\{y|\psi^{(l)}\} \). (This is why \( \alpha \) is used instead of \( = \) ) However, as the authors pointed out, the constant does not play a role in the next M-step, because it depends only on \( \psi^{(l)} \), while the next-step maximization is over \( \psi \). Let \( \alpha_1^*, \ldots, \alpha_K^* \) be an i. i. d. sample generated from \( g \), the importance sampling distribution. Then, we have the approximation

\[ Q\{\psi|\psi^{(l)}\} \approx \frac{1}{K} \sum_{k=1}^{K} w_{kl} \log\{f(y, \alpha_k^*|\psi)\}, \tag{2.1} \]

where \( w_{kl} = f\{\alpha_k^*|y; \psi^{(l)}\}/g(\alpha_k^*) \), known as the importance weights. The right side of (2.1) is then maximized with respect to \( \phi \) in the M-step to obtain \( \psi^{(l+1)} \). Note that the right side of (2.1) is not a completely known function (of \( \phi \)), but subject to an unknown constant which is \( f\{y|\psi^{(l)}\} \).
However, as noted earlier, this constant makes no difference in the M-step, therefore we simply ignore it. In other words, the function that is actually maximized is the right side of (2.1) with \( f\{ y | \psi^{(l)} \} \) replaced by 1. As for the importance sampling distribution \( g \), Booth and Hobert proposed to use a multivariate t-distribution whose mean and covariance matrix match the Laplace approximations of the mean and covariance matrix of \( f(\alpha|y; \psi) \).

2. Rejection sampling. Alternatively, i. i. d. samples may be generated from \( f(\alpha|y; \psi) \) by multivariate rejection sampling as follows (Geweke 1996, §3.2). Write the conditional density as \( f = cf_1f_2 \), where \( c \) is the normalizing constant, and \( f_1, f_2 \) are the two factors on the right side of (4.11). (i) First draw \( \alpha \) from \( f_2 \) and, independently, \( u \) from the Uniform[0, 1] distribution. (ii) If \( u \leq f_1(\alpha)/\tau \), where \( \tau = \sup_\alpha f_1(\alpha) \), accept \( \alpha \). Otherwise, return to (i). Note that \( f_1 \) corresponds to a likelihood function under a GLM. Therefore, \( \tau \) can be found using the iterative WLS procedure for fitting the GLMs (McCullagh and Nelder 1989, pp. 206). Furthermore, it can be shown that \( \tau \) needs not to change at each step of the MCEM algorithm (Booth and Hobert 1999, pp. 271-272).

Furthermore, the latter authors proposed a rule that automatically increases the Monte-Carlo sample size as the algorithm proceeds, whenever necessary. They showed that the new algorithms have the following advantages over the Markov chain methods. First, the assessment of the Monte-Carlo errors is straightforward when i. i. d. samples are used. Note that such an assessment is critical for the automated method. A related theoretical advantage is that conditions for the central limit theorem in the i. i. d. case is much easier to verify than under a Markov chain. In terms of computational speed, they showed that for the same example that was considered by McCulloch (1997), the rejection sampling and importance sampling methods are about 2.5 times and 30 times faster, respectively, than the Metropolis-Hastings sampling method of McCulloch (1997).

2.6 Maximization by parts

Although MCEM methods have been fairly well-developed in likelihood-based inference about GLMM, alternative procedures have also been proposed. Here we introduce a method proposed by Song et al. (2005), which they called maximization by parts (MBP). Again, the objective was to overcome some of the computational difficulties in maximum likelihood estimation. One of such difficulties is the computation of the second derivatives

24
of the log-likelihood function. For example, the Newton-Raphson procedure requires calculations of both the first and second derivatives. If the likelihood function is complicated, the derivation and calculation of its derivatives, especially the second derivatives, can be both analytically and computationally challenging.

The idea of MBP is very easy to illustrate. Write the log-likelihood function as

$$ l(\theta) = l_w(\theta) + l_\epsilon(\theta). $$

Let \( \hat{i} \) denotes the vector of first (partial) derivatives. Then, the likelihood equation

$$ \hat{i}(\theta) = 0 $$

can be written as

$$ \hat{i}_w(\theta) = -\hat{i}_\epsilon(\theta). $$

Here the \( \theta \)'s on both sides of (2.4) are supposed to be the same, but they do not have to be so in an iterative equation, and this is the idea of MBP. The initial estimator, \( \hat{\theta}^{(1)} \), is a solution to \( \hat{i}_w(\theta) = 0 \). Then, use the equation

$$ \hat{i}_w\{\hat{\theta}^{(2)}\} = -\hat{i}_\epsilon\{\hat{\theta}^{(1)}\} $$

to update to get the next step estimator \( \hat{\theta}_2 \), and so on. It is easy to see that if BMP converges, the limit, say, \( \hat{\theta} \), satisfies (2.3).

From a practical standpoint, the most important issue regarding MBP seems to be the decomposition (2.2), that is, the choice of \( l_w(\theta) \). A condition for a good choice of \( l_w \) is the so-called information dominance. In other words, \( \hat{i}_w \) needs to be larger than \( \hat{i}_\epsilon \) in certain sense (Song et al. 2005, Theorem 2). Another condition for choosing \( l_w \) is that \( \hat{i}_w(\theta) = 0 \) is an unbiased estimating equation, or, alternatively, that \( \hat{\theta}_1 \) is a consistent estimator.

The MBP method is potentially applicable to at least some class of GLMMs. It is suggested that the hierarchical log-likelihood of Lee and Nelder (1996) may be used as \( l_w \). However, if the random effects in the GLMM are normally distributed, this will lead to a biased estimating equation. In fact, the solution to such an equation may not be consistent (Clayton 1996, Jiang 1999). The choice of \( l_w \) or the performance of MBP with the proposed hierarchical log-likelihood \( l_w \) remain unclear to date.
Assuming that MBP is applicable to GLMM, the next question is how much does MBP help. As noted earlier, the procedure has computational advantage in situations where $\hat{l}$ is much more difficult to deal with (numerically or analytically) than $\tilde{l}$. An example of such cases is the Gaussian copula model (e. g., Song 2000). Now let us consider a GLMM example.

**Example 2.2.** Suppose that, given the random effects $u_i, 1 \leq i \leq a$ and $v_j, 1 \leq j \leq b$, $y_{ij}$ are conditionally independent such that

$$
\logit\{P(y_{ij} = 1 | u, v)\} = \beta_0 + \beta_1 x_{ij} + u_i + v_j,$$

where $u = (u_i)_{1 \leq i \leq a}, v = (v_j)_{1 \leq j \leq b}, x_{ij}$ is a known covariate, and $\beta_0$ and $\beta_1$ are unknown coefficients. Furthermore, suppose that the random effects are independent with $u_i \sim N(0, \sigma^2), v_j \sim N(0, \tau^2)$. It is more convenient to use the following expressions: $u_i = \sigma \xi_i, v_j = \tau \eta_j$, where $\xi_1, \ldots, \xi_a$ and $\eta_1, \ldots, \eta_b$ are i. i. d. $N(0, 1)$ random variables. Then, the log-likelihood function under this GLMM has the following expression:

$$
l = c + \log \left[ \int \cdots \int \exp \left\{ \sum_{i=1}^{a} \sum_{j=1}^{b} \phi_{ij}(y_{ij}, \beta, \sigma \xi_i, \tau \eta_j) - \frac{1}{2} \sum_{i=1}^{a} \xi_i^2 - \frac{1}{2} \sum_{j=1}^{b} \eta_j^2 \right\} d\xi_1 \cdots d\xi_a d\eta_1 \cdots d\eta_b \right],$$

where $c$ is a constant, and

$$
\phi_{ij}(y_{ij}, \beta, u_i, v_j) = y_{ij}(\beta_0 + \beta_1 x_{ij} + u_i + v_j) - \log\{1 + \exp(\beta_0 + \beta_1 x_{ij} + u_i + v_j)\}.
$$

For simplicity, let us assume that the variance components $\sigma$ and $\tau$ are known, so that $\beta_0$ and $\beta_1$ are the only unknown parameters. It can be shown that the first and second derivatives of $l$ have the following forms:

$$
\frac{\partial l}{\partial \beta_s} = \frac{\int \cdots \int \exp\{\cdots\} \psi_s d\xi \eta}{\int \cdots \int \exp\{\cdots\} d\xi \eta},
$$

$$
\frac{\partial^2 l}{\partial \beta_s \partial \beta_t} = \frac{\int \cdots \int \exp\{\cdots\} \psi_s \psi_t d\xi \eta}{\int \cdots \int \exp\{\cdots\} d\xi \eta} - \left[ \frac{\int \cdots \int \exp\{\cdots\} \psi_s d\xi \eta}{\int \cdots \int \exp\{\cdots\} d\xi \eta} \right] \times \left[ \frac{\int \cdots \int \exp\{\cdots\} \psi_t d\xi \eta}{\int \cdots \int \exp\{\cdots\} d\xi \eta} \right],
$$

26
\( s, t = 0, 1, \) where \( d\xi = d\xi_1 \cdots d\xi_a, \) \( d\eta = d\eta_1 \cdots d\eta_b, \)

\[
\psi_s = \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{\partial \phi_{ij}}{\partial \beta_s},
\]

\[
\psi_{s,t} = \left( \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{\partial \phi_{ij}}{\partial \beta_s} \right) \left( \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{\partial \phi_{ij}}{\partial \beta_t} \right) + \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{\partial^2 \phi_{ij}}{\partial \beta_s \partial \beta_t}.
\]

Now the fact is, the integrals involved in \( \hat{I} \) are equally difficult to evaluate as those involved in \( \hat{I}. \) (Note that these are \( a + b \) dimensional integrals.) Nevertheless, new integrals do emerge in \( \hat{I}, \) that is, there are three different integrals in \( \hat{I}, \) and six different ones in \( \hat{I}. \) In general, if there are \( p \) unknown parameters, there may be as many as \( p + 1 \) different integrals in \( \hat{I}, \) and as many as \( p + 1 + p(p + 1)/2 = (1/2)(p + 1)(p + 2) \) different integrals in \( \hat{I}. \) If \( p \) is large, it is quite a saving in computation, provided that any single one of the integrals (involved in \( \hat{I} \)) can be evaluated.

### 2.7 Bayesian inference

GLMM can be naturally formulated in a Bayesian framework, and thus analyzed using the Bayesian methods. The main difference in the model is that a (joint) prior is assumed for \( \beta \) and \( G, \) the covariance matrix of \( \alpha. \) Let \( \pi(\beta, G) \) be the (joint) prior density for \((\beta, G). \) For example, a flat prior is sometime used, i.e., \( \pi(\beta, G) \propto \text{constant}. \) The main objective of the Bayesian inference is to obtain the posteriors for \( \beta, G \) and \( \alpha, \) which are given by

\[
f(\beta, G | y) = \frac{f(y | \beta, \alpha) f(\alpha | G) \pi(\beta, G) d\alpha}{\int f(y | \beta, \alpha) f(\alpha | G) \pi(\beta, G) d\alpha d\beta dG},
\]

\[
f(\alpha | y) = \frac{\int f(y | \beta, \alpha) f(\alpha | G) \pi(\beta, G) d\beta dG}{\int \int \int f(y | \beta, \alpha) f(\alpha | G) \pi(\beta, G) d\alpha d\beta dG}.
\]

The posteriors are typically numerically intractable, especially when the dimension of the integrals involved is greater than one. Therefore, Monte-Carlo methods were proposed to handle the computation. For example, Zeger and Karim (1991) used Gibbs sampler to evaluate the posteriors. A main advantage of the Bayesian method is that one obtains a posterior (distribution)
rather than just a point estimator. A disadvantage is the computational intensiveness. There is also an issue of improper posteriors (e. g., Hobert and Casella 1996).

2.8 Approximate inference

When the exact likelihood function is difficult to compute, approximation becomes one of the natural alternatives. A well-known method of approximate integrals is named after Laplace. Suppose that one wishes to approximate an integral of the form

\[ \int \exp\{-q(x)\} dx, \]

where \( q(\cdot) \) is a “well-behaved” function in the sense that it achieves its minimum value at \( x = \bar{x} \) with \( q' (\bar{x}) = 0 \) and \( q''(\bar{x}) > 0 \). Then, we have, by Taylor expansion,

\[ q(x) = q(\bar{x}) + \frac{1}{2} q''(\bar{x}) (x - \bar{x})^2 + \cdots, \]

which yields an approximation to (2.5)

\[ \int \exp\{-q(x)\} dx \approx \sqrt{\frac{2\pi}{q''(\bar{x})}} \exp\{-q(\bar{x})\}. \]  

There is a multivariate extension of (2.6), which is more useful in our case. Let \( q(\alpha) \) be a “well-behaved” function which attains its minimum value at \( \alpha = \bar{\alpha} \) with \( q'(\bar{\alpha}) = 0 \) and \( q''(\bar{\alpha}) > 0 \), where \( q' \) and \( q'' \) denote the gradient (i.e., the vector of first derivatives) and Hessian (i.e., the matrix of second derivatives) of \( q \), respectively, and the notation \( A > 0 \) means that the matrix \( A \) is positive definite. Then, we have

\[ \int \exp\{-q(\alpha)\} d\alpha \approx c |q''(\bar{\alpha})|^{-1/2} \exp\{-q(\bar{\alpha})\}, \]

where \( c \) is a constant depending only on the dimension of the integral, and \( |A| \) denotes the determinant of matrix \( A \). Several authors have used this method to approximate the likelihood function, and then treat the approximate likelihood as the true likelihood for inference about GLMM. The method may also be viewed as estimation of both fixed and random effects via maximization of the joint density functions of the observations and random effects,
penalized quasi-likelihood (PQL), or maximum hierarchical likelihood. See, Schall (1991), Breslow and Clayton (1993), Wolfinger and O’Connell (1993), McGilchrist (1994), and Lin and Breslow (1996), among others. Lee and Nelder’s (1996) hierarchical likelihood method is similar to PQL in spirit but allows non-Gaussian distributions for the random effects. These methods are computationally attractive. They also provide estimates of the random effects (see below), which in some cases are of practical interest. Unfortunately, the Laplace-approximation based methods are known to have some unsatisfactory properties. In particular, the resulting estimators are inconsistent under standard asymptotic assumptions (e.g., Jiang 1998a). Furthermore, Lin and Breslow (1996) showed that their PQL method works well when the variances of the random effects are close to zero; otherwise, the bias can be substantial. Also see Kuk (1995).

2.9 Estimating equations

On the other hand, generalized estimating equations (GEE) have been used in the analysis of longitudinal data (Liang and Zeger 1986, Prentice 1988). Such a method applies to a special class of GLMM, in which the observations are independently clustered. In other words, the covariance matrix of the observations is block-diagonal. Jiang (1998a) proposed estimating equations that apply to GLMMs not necessarily having a block-diagonal covariance structure, such as the one in the salamander mating problem. He showed that the estimators, which are solution to the estimating equations, are consistent. However, the estimators are inefficient. Later, Jiang and Zhang (2001) proposed a two-step procedure to obtain more efficient estimators. They considered a broader class of models, which does not require full specification of the conditional distribution of the responses given the random effects. Therefore, the method applies to a broader class of models than the GLMM. To illustrate their method, let $S$ be a vector of base statistics. The first-step estimator of $\theta$, $\hat{\theta}$, is a solution to the estimating equation $B\{S - u(\theta)\} = 0$, where $\theta$ is the vector of parameters under the GLMM which include fixed effects and variance components, $B$ is a known matrix, and $u(\theta) = E_{\theta}(S)$. As it turns out, if the base statistics are chosen as those in Jiang (1998a), then, to obtain the first-step estimator of $\theta$, one only needs to assume that the form of the conditional mean of the responses given the random effects is specified. For example, if the responses $y_i$ are binary, one only needs to assume that $\logit(p_i) = x_i^T\beta + z_i^T\alpha$, where $p_i$ is the conditional

29
probability of $y_i = 1$ given a vector of random effects $\alpha, \beta$ is a vector of fixed effects, and $x_i, z_i$ are known vector of covariates. Jiang and Zhang showed that, under mild conditions, the first-step estimator is consistent. In fact, the estimator of Jiang (1998a) is a special case of the first-step estimator. On the other hand, the efficiency of the first-step estimator may be improved by choosing a "good" $B$. It can be shown that the optimal $B$, in the sense of minimum asymptotic variance, is $B^* = U'V^{-1}$, where $U = \partial u/\partial \theta'$ and $V = \text{Var}_\theta(S)$. It is clear that $B^*$ depends on $\theta$, i.e., $B^* = B^*(\theta)$, hence is not available. However, one may replace $\theta$ by $\bar{\theta}$, the first-step estimator. This results in the second-step estimator, which is a solution to $B^* \{S - u(\bar{\theta})\} = 0$, where $\tilde{B}^* = B^*(\bar{\theta})$. To obtain the second-step estimator, one needs to specify the forms of the first two conditional moments of the responses given the random effects. It was shown that, under regularity conditions, the second-step estimator is asymptotically as efficient as the solution to the optimal estimating equation with $B^*$, which is not computable. Furthermore, Jiang and Zhang reported results from two simulated examples; in each case the second-step estimator had about 40% reduction of the MSE compared to the first-step estimator.

### 2.10 Prediction of random effects

The method of joint estimation of the fixed and random effects discussed earlier, which was originally due to Henderson (1950), provides estimators, or predictors, of the random effects. Jiang et al. (2001) took another look at the method as maximization of a posterior of the random effects (under a non-informative prior). Let $y$ be a vector of responses, and $\theta$ a vector of dispersion parameters. Write $L_J(\alpha, \beta) = f(y, \alpha|\beta, \theta)$, the joint density function of $y$ and $\alpha$ given $\beta$ and $\theta$, where $\alpha$ is a vector of random effects, and $\beta$ a vector of fixed effects. Since

$$\max_{\alpha, \beta} L_J(\alpha, \beta) = \max_\beta \max_\alpha L_J(\alpha, \beta),$$

the maximization can be done by two steps. In the first step one finds $\alpha = \bar{\alpha}(\beta)$ that maximizes $L_J(\alpha, \beta)$ for fixed $\beta$. In the second step, one finds $\beta$ that maximizes $L_J(\bar{\alpha}, \beta)$, and lets $\hat{\alpha} = \bar{\alpha}(\bar{\beta})$. Now consider the first step. Observe that

$$f(y, \alpha|\beta, \theta) = f(y|\beta, \theta)f(\alpha|y, \beta, \theta),$$

(2.7)
The first factor on the right side of (2.7) corresponds to the likelihood function for estimating $\beta$ and $\theta$, while the second factor to the posterior density of $\alpha$ given $y$ (if one would like, assuming that a noninformative prior has been assigned to $\alpha$). The idea of Henderson (1950) was to find $\alpha$ and $\beta$ that jointly maximize $f(y, \alpha|\beta, \theta)$. Since the first factor does not depend on $\alpha$, maximizing $f(y, \alpha|\beta, \theta)$ is equivalent to maximizing the posterior, and then the profile joint density $f(\alpha|\beta, \theta)$. Note that, although in linear mixed models the maximizer $\hat{\alpha}$ and $\hat{\beta}$ correspond to the BLUP and BLUE, they are no longer such predictor and estimator in nonlinear cases, such as GLMM. Still, the method is intuitive in the sense that $\hat{\alpha}$ maximizes the posterior. For such a reason, Jiang et al. called $\hat{\alpha}$ and $\hat{\beta}$ maximum a posterior estimators, or MPE.

The MPE are typically obtained by solving the following system of nonlinear equations:

$$\frac{\partial l_J}{\partial \beta} = 0, \quad (2.8)$$

$$\frac{\partial l_J}{\partial \alpha} = 0, \quad (2.9)$$

where $l_J = \log(L_J)$. In practice, there are often a large number of random effects involved in a GLMM. For example, in the salamander mating problem that was discussed earlier (McCullagh and Nelder 1989, §14.5), the number of random effects associated with the female and male salamanders is 80. In an NHIS problem considered by Malec et al. (1997), the number of random effects corresponding to the small areas is about 600. This means that the first step of MPE, i.e., the maximization of $L_J(\alpha, \beta)$ for fixed $\beta$ is over a high dimensional space. In other words, one has to simultaneously solve a large number of nonlinear equations (2.8) and (2.9). It is well known that standard methods of solving nonlinear systems such as Newton-Raphson (N-R) may be inefficient and extremely slow when the dimension of the solution is high. In fact, even in the linear case directly solving the BLUP equations may involve inversion of a large matrix, which can be computationally burdensome. There are other disadvantages of N-R. First, convergence of the N-R is sensitive to the initial values. When the dimension of the solution is high, it can be very difficult to find the initial values that will result in convergence. Second, N-R requires computation of partial derivatives, the analytic derivation of which can be tedious, and errors are often made in the process as well as in programming.
Jiang (2000b) proposed a nonlinear Gauss-Seidel algorithm for computing the MPE, which is an extension of the Gauss-Seidel algorithm in numerical analysis for solving large linear systems. We use an example to illustrate the algorithm.

**Example 2.1 (continued).** Consider, once again, Example 2.1. To compute the MPE, one needs to solve the following system of nonlinear equations:

\[
\frac{u_i}{\sigma_1} + \sum_{j=1}^{m_2} \frac{\exp(\mu + u_i + v_j)}{1 + \exp(\mu + u_i + v_j)} = y_i, \quad 1 \leq i \leq m_1, \quad (2.10)
\]

\[
\frac{v_j}{\sigma_2} + \sum_{i=1}^{m_1} \frac{\exp(\mu + u_i + v_j)}{1 + \exp(\mu + u_i + v_j)} = y_j, \quad 1 \leq j \leq m_2, \quad (2.11)
\]

where \( y_i = \sum_{j=1}^{m_2} y_{ij} \) and \( y_j = \sum_{i=1}^{m_1} y_{ij} \). Note that given the \( v_j \)'s, each equation in (2.10) is univariate, which can be easily solved (e.g., by bisection or one-dimensional N-R). A similar fact is observed in (2.11). This motivates the following algorithm: Starting with initial values \( v_j^{(0)}, 1 \leq j \leq m_2 \), solve (2.10) with \( v_j^{(0)} \) in place of \( v_j, 1 \leq j \leq m_2 \) to get \( u_i^{(1)}, 1 \leq i \leq m_1 \); then (2.11) with \( u_i^{(1)} \) in place of \( u_i, 1 \leq i \leq m_1 \) to get \( v_j^{(1)}, 1 \leq j \leq m_2 \), and so on. It is clear that the algorithm does not require the calculation of derivatives. Each step of the algorithm is easy to operate and, in fact, has a unique solution. Finally, it can be shown that the convergence of the algorithm is not affected by initial values. In other words, one has **global convergence.** See Jiang (2000b) for detail.

One of the main areas of applications of prediction of random effects, or mixed effects, is small area estimation (e. g., Rao 2003). In this context, a method known as empirical best prediction (EBP) has been developed. It is a two-step procedure. In the first-step, one derives an expression for the best predictor which is the conditional expectation of the random (or mixed) effect given the data. The expression is likely dependent on a vector of unknown parameters. Therefore, in the second-step, one replace the unknown parameters by consistent estimators. See Jiang and Lahiri (2001), Jiang (2003b). A feature of EBP is that it is model-based. If the assumed model fails, the predictor may perform poorly. Jiang and Lahiri (2005) developed a model-assisted EBP method that has the property of design-consistency, that is, even under model failure, the predictor of a domain mean is approximately equal to a design-based estimator as long as the domain sample size is large. The design-based estimator (e. g., Hajek 1971), on the other hand, is known
to approach the true domain mean when the domain sample size is large regardless of the assumed model. In both the EBP and model-assisted EBP, the authors have obtained estimators of the MSE of the EBP whose bias is corrected to the second-order.

2.11 Future research and open problems

There is a lack of studies on theoretical properties, such as asymptotic behavior, of the MLE under a GLMM, despite considerable effort made in developing computing algorithms for these estimators. The problem is relatively straightforward for longitudinal GLMMs, in which the responses may be divided into independent clusters (e.g., Jiang 2001b). What is really challenging is the case with crossed random effects, such that there is no independent clusters or groups. Note that, unlike linear mixed models, the likelihood function can only be expressed as (multi-dimensional) integrals under a GLMM. General asymptotic results such as those in the linear mixed model case (e.g., Miller 1977, Jiang 1996) do not exist for GLMM. For example, the salamander mating data has been analyzed by many authors; some others use the same model and data structure for simulations (e.g., McCullagh and Nelder 1989, Karim and Zeger 1992, Drum and McCullagh 1993, Lin and Breslow 1996, Lin 1997, Jiang 1998a, and Jiang and Zhang 2001). Furthermore, Monte-Carlo EM algorithms have been developed to obtain the MLE with the salamander data (Booth and Hobert 1999). However, one fundamental question has yet to be answered: Is the MLE consistent? The answer is not obvious at all.

Unlike estimation problems, the literature on testing in GLMM is not extensive. For example, Lin (1997) used PQL method for testing the hypothesis that all the variance components are zero. While PQL is known to yield inconsistent point estimators, its usage in testing such a hypothesis is entirely appropriate. This is because the random effects become zero under the null hypothesis, so a GLMM becomes a GLM, for which the likelihood-based methods are well justified. Also see Lin and Carroll (1999). However, it is difficult to extend the method to testing complex hypotheses, under which the random effects may not vanish. Note that the lack of asymptotic theory on MLE also contributed to the difficulties in testing, because large sample tests rely on the asymptotic theory. On the other hand, once again, testing for longitudinal GLMMs is relatively easier (e.g., Song and Jiang 2000).

Other underdeveloped areas include model diagnostics and model section
for GLMM as well as some prediction problems.

3 Application to longitudinal data I

3.1 Preliminary

The defining feature of longitudinal study is that individuals are measured repeatedly over time. We first consider an example.

Example 3.1. Kaslow et al. (1987) reported a data set regarding CD4+ cell numbers. It is known that human immune deficiency virus (HIV) attacks an immune cell called CD4+. A normal individual has about 1,100 CD4+ cells per millilitre of blood. The number decreases over time as the HIV disease progresses. The data set contains measures of CD4+ cell numbers from 369 infected men over time since seroconversion (time when HIV becomes detectable). A total of 2,376 counts were recorded. For more detail, see, for example, Diggle et al. (1996).

Unlike time series, longitudinal data is both cross-sectional (i.e., involving multiple individuals or subjects) and longitudinal (i.e., over time for each individual). Both patterns should be explored in the data analysis.

Initial analysis of a longitudinal data set includes graphical tools such as scatter plots against time, residual plots (against time and individuals) and various smooth curves such as smoothing splines and loess (local weighted LS lines; Cleveland 1979). In addition, some time series analysis tools may be used. For example, the autocorrelation function is defined as

\[ \hat{\rho}(s) = \text{cor}\{y_{it}, y_{it+s}\}, \]

where \( y_{it} \) represents the observation of the \( i \)th individual at time \( t \).

3.2 A linear mixed model for longitudinal data

Because of the nature of longitudinal data, observations collected from the same individual (subject) over time may be correlated. These correlations, in general, have significant impact on the analysis of longitudinal data. For example, by ignoring the correlations within the subjects, the standard errors of the estimates of the model parameters may be incorrectly estimated. In the case of continuous observations, one way of modelling the mean response
as well as the correlations within the subjects is to use a linear mixed model. We first consider an example.

**Example 3.2.** Let \( y_{ij} \) be the observation from subject \( i \) at time \( t_j, i = 1, \ldots, m, j \in J_i \subset J = \{1, \ldots, b\} \). We assume that \( y_{ij} \) satisfies the following

\[
y_{ij} = x_{ij}^\prime \beta + u_i + e_{ij},
\]

where \( x_{ij} \) is a vector of subject/time dependent covariates, \( \beta \) is a vector of regression coefficients, \( u_i \) is a subject-specific random effect, and \( e_{ij} \) is an (unexplained) error. Furthermore, it is assumed that \( u_i \)'s and \( e_{ij} \)'s are independent with \( u_i \sim N(0, \sigma_u^2) \) and \( e_{ij} \sim N(0, \sigma_e^2) \). Then, we have \( \text{cor}(y_{ij}, y_{i'j'}) = 0 \) if \( i \neq i' \), and

\[
\text{cor}(y_{ij}, y_{ik}) = \frac{\sigma^2_u}{\sigma_u^2 + \sigma_e^2}, \quad j \neq k.
\]

The model in Example 3.2 is, of course, very simple and it may not capture all the correlations among the observations. For example, there may be serial correlations among the observations from the same subject given the subject-specific random effects. Such a serial correlation is taken into account in the next example.

**Example 3.3.** Suppose that the observation times are equally spaced. In such a case we may assume, without loss of generality, that \( t_j = j \). Suppose that the observations \( y_{ij} \) satisfy

\[
y_{ij} = x_{ij}^\prime \beta + u_i + w_{ij} + e_{ij},
\]

\( i = 1, \ldots, m, j \in J_i \subset T = J \) (\( T \) is defined in the first paragraph of this section), where \( u_i \) is an individual-specific random effect, \( w_{ij} \) corresponds to a serial correlation, and \( e_{ij} \) represents a measurement error. It is assumed that the \( u_i \)'s are independent \( N(0, \sigma_u^2) \), and the \( e_{ij} \)'s are independent \( N(0, \sigma_e^2) \). As for the \( w_{ij} \)'s, it is assumed that they satisfy the following relation of the first order autoregressive process, or AR(1):

\[
w_{ij} = \phi w_{i,j-1} + z_{ij},
\]

where \( \phi \) is a constant and \( \phi \in (0, 1) \), and \( z_{ij} \)'s are independent \( N\{0, \sigma_w^2(1 - \phi^2)\} \). Finally, \( u, w \) and \( e \) are independent. We have (e.g., Anderson 1971, pp. 174)

\[
\text{cov}(y_{ij}, y_{ir}) = \sigma^2_u + \sigma^2_w \phi^{|j-r|} + \sigma^2_e \delta_{ij},
\]

35
where $\delta_{q,r} = 1$ if $q = r$ and 0 otherwise. It is seen that the correlation within the subject decreases as the times of observations depart.

Following Datta and Lahiri (2000), a general linear mixed model for longitudinal data may be expressed as follows:

$$y_i = X_i \beta + Z_i \alpha_i + \epsilon_i, \quad i = 1, \ldots, m, \quad (3.1)$$

where $y_i$ represents the vector of observations from the $i$th individual; $X_i$ and $Z_i$ are known matrices; $\beta$ is an unknown vector of regression coefficients; $\alpha_i$ is a vector of random effects; and $\epsilon_i$ is a vector of errors. It is assumed that $\alpha_i$, $\epsilon_i$, $i = 1, \ldots, m$ are independent with $E(\alpha_i) = 0$, $\text{Var}(\alpha_i) = G_i$; $E(\epsilon_i) = 0$, $\text{Var}(\epsilon_i) = R_i)$, where the covariance matrices $G_i$ and $R_i$ are known up to a vector $\theta$ of dispersion parameters. It is easy to see that Example 3.2 and Example 3.3 are special cases of this general longitudinal model. Also note that (3.1) is a special case of (1.1) with $y = (y_i)_{1 \leq i \leq m}$, $X = (X_i)_{1 \leq i \leq m}$, $Z = \text{diag}(Z_1, \ldots, Z_m)$, $\alpha = (\alpha_i)_{1 \leq i \leq m}$ and $\epsilon = (\epsilon_i)_{1 \leq i \leq m}$.

In the above model, the correlations, or more generally, covariances, are modelled parametrically, that is, $G_i = G_i(\theta)$, $R_i = R_i(\theta)$, where $G_i(\cdot)$ and $R_i(\cdot)$ are known functions. Alternatively, the covariances may be modelled nonparametrically. In such a case, one simply assumes that $V_i = \text{Var}(y_i)$ is unspecified, $1 \leq i \leq m$.

### 3.3 Estimation of the regression coefficients under linear model

A traditional method of estimating the regression coefficients $\beta$ in the case of correlated observations is called weighted least squares, or WLS. Suppose that the observations are collected from individuals over time. Let $y$ denote the vector of observations, which may be correlated, and $X$ a matrix of known covariates. Suppose that $E(y) = X\beta$, where $\beta$ is a vector of unknown regression coefficients. The WLS estimator of $\beta$ is obtained by minimizing

$$(y - X\beta)'W(y - X\beta), \quad (3.2)$$

where $W$ is a known symmetric weighting matrix. As before, suppose, without loss of generality, that $X$ is of full rank. Then, for any nonsingular $W$, the minimizer of (3.2) is given by

$$\hat{\beta}_W = (X'WX)^{-1}X'Wy.$$
This estimator is unbiased with
\[
\text{Var}(\hat{\beta}_W) = (X'WX)^{-1}X'WVWX(X'WX)^{-1}.
\]

As a special case, the ordinary least squares (OLS) estimator is obtained by choosing \( W = I \), the identity matrix. This gives
\[
\hat{\beta}_I = (X'X)^{-1}X'y.
\]

On the other hand, the optimal choice of \( W \) in the sense of minimum variance is known to be \( W = V^{-1} \), where \( V = \text{Var}(y) \). This is known as the best linear unbiased estimator, or BLUE, given by
\[
\hat{\beta}_{\text{BLUE}} = (X'V^{-1}X)^{-1}X'V^{-1}y. \tag{3.3}
\]

However, since \( V \) is typically unknown, the BLUE is not computable.

Under a parametric model, that is \( V = V(\theta) \), where \( \theta \) is a vector of dispersion parameters and \( V \) is specified up to \( \theta \), and the normality assumption, the unknown covariances may be estimated by the ML or REML methods (see §1). We consider some examples.

Example 3.4 (Growth of trees). Diggle et al. (1996) reported two data sets collected from 79 Sitka spruce trees over two growth periods. The objective was to assess the effect of ozone pollution on tree growth. 54 of the trees were grown with ozone exposure at 70 ppb. The other 25 trees were grown without ozone exposure (the control group). The two growth periods were 1988 with five observational times, and 1989 with eight observational times. Diggle et al. (1996) assumed the following models for the mean responses. For the 1988 data, the model is
\[
\begin{align*}
\mu_1(t_j) &= \beta_j, \quad i = 1, \ldots, 5; \\
\mu_2(t_j) &= \beta_j + \tau + \gamma t_j, \quad i = 1, \ldots, 5.
\end{align*}
\]

For the 1989 data, the model is
\[
\begin{align*}
\mu_1(t_j) &= \beta_j, \quad i = 6, \ldots, 13; \\
\mu_2(t_j) &= \beta_j + \tau, \quad i = 6, \ldots, 13.
\end{align*}
\]

Furthermore, the authors assumed an unspecified covariance matrix \( V^{(1)} \) (same for all subjects) for the 1988 data, and another unspecified covariance matrix \( V^{(2)} \) (again, same for all subjects) for the 1989 data. REML method was used in estimating both covariance matrices.
Example 3.5 (Effect of air pollution episodes on children). Laird and Ware (1982) consider data from an observational study, in which approximately 200 school children were examined under normal conditions, then during an air pollution alert and three successive weeks following the alert. The objective of the study was to determine whether FEV$_1$, defined as the volume of air exhaled in the first second of a forced exhalation, was depressed during the alert. Two simple linear mixed models were considered for this longitudinal data. In the first model, it is assumed that

$$y_{ij} = \beta_j + u_i + \epsilon_{ij},$$

$i = 1, \ldots, m$, $j = 1, \ldots, 5$, where $u_i$ is a subject-specific random effect and $\epsilon_{ij}$ is an error. It is assumed that $u_i$'s and $\epsilon_{ij}$'s are independent with $u_i \sim N(0, \sigma^2)$, $\epsilon_{ij} \sim N(0, \tau^2)$. Note that this is a fully parametric model with $\theta = (\sigma^2, \tau^2)'$. The latter was estimated by REML. In the second model, another random effect was introduced to quantify the mean decline in FEV$_1$ for each child. The model can be expressed as

$$y_i = X_i \beta + Z_i v_i + e_i,$$

where $y_i$ represents the vector of responses for the $i$th child, $X_i = I_5$, $v_i = (v_{i1}, v_{i2})' \sim N(0, D)$, where $D$ is a $2 \times 2$ unknown covariance matrix, and

$$Z' = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 \end{pmatrix}.$$

Again, the model is fully parametric with $\theta = (\sigma_1^2, \sigma_2^2, \rho, \tau)'$, where $\sigma_j^2 = \text{var}(v_{ij})$, $j = 1, 2$, $\rho = \text{cor}(v_{i1}, v_{i2})$, and $\tau^2 = \text{var}(\epsilon_{ij})$. Thus, we have $G_i(\theta) = Z_i D Z_i'$ and $R_i(\theta) = \tau^2 I_5$. Again, REML was used to estimate $\theta$.

4 Application to longitudinal data II

4.1 Iterative weighted least squares

Estimation of the within-subject covariances under a parametric covariance model may be sensitive to model misspecifications. In this subsection, we develop an efficient estimation procedure under unspecified (nonparametric) covariances. Recall that, if the $V_i$'s were known, one could estimate $\beta$ by
the BLUE given by (3.3). Since the observations from different subjects are assumed independent, the BLUE can be further expressed as

$$\hat{\beta}_{\text{BLUE}} = \left( \sum_{i=1}^{m} X_i'V_i^{-1}X_i \right)^{-1} \sum_{i=1}^{m} X_i'V_i^{-1}y_i. \quad (4.1)$$

On the other hand, it would be easier to estimate $V$ if $\beta$ were known. For example, an unbiased estimator of $V$ is given by $\hat{V} = (y - X\hat{\beta})(y - X\beta)'$. However, this is not a consistent estimator. In fact, if $V$ is completely unknown, there are $n(n+1)/2$ unknown parameters in $V$, which is (far) more than the sample size $n$. Therefore, in such a case, one is not expected to have a consistent estimator of $V$ no matter what. It is clear that some information about $V$ must be available. For simplicity, let us first consider a special case.

**Example 4.1 (Balanced data).** Suppose that the observations are collected over a common set of times. Let $y_{ij}$, $j = 1, \ldots, k$ be the measures collected from the $i$th individual over times $t_1, \ldots, t_k$, respectively, and $y_i = (y_{ij})_{1 \leq j \leq k}$, $i = 1, \ldots, m$. Suppose that the vectors $y_1, \ldots, y_m$ are independent with

$$E(y_i) = X_i\beta \quad \text{and} \quad \text{Var}(y_i) = V_0,$$

where $X_i$ is a matrix of known covariates, and $V_0 = (v_{qr})_{1 \leq q, r \leq k}$ is an unknown covariance matrix. It follows that $V = \text{diag}(V_0, \ldots, V_0)$. Now the good thing is that $V$ may be estimated consistently, if $k$ is fixed. In fact, if $\beta$ were known, a method of moments estimator of $V$ would be the following:

$$\hat{V} = \text{diag}(\hat{V}_0, \ldots, \hat{V}_0), \quad \text{where}$$

$$\hat{V}_0 = \frac{1}{m} \sum_{i=1}^{m} (y_i - X_i\beta)(y_i - X_i\beta)'.$$

In general, if the data is unbalanced, i.e., if the observations are not necessarily collected over a common set of times, method of moments estimators of the $V_i$’s may still be obtained, given $\beta$. See Jiang et al. (2005) for detail.

To summarize the main idea, if $V$ were known, one could use (4.1) to compute the BLUE of $\beta$; if $\beta$ were known, one could use the method of moments to estimate $V$. It is clear that there is a cycle, which motivates the following algorithm when neither $V$ nor $\beta$ is assumed known: Start with the OLS estimator $\hat{\beta}_I$ (see §3.3), and estimate $\hat{V}$ by the method of moments, in which $\beta$ is replaced by $\hat{\beta}_I$; then replace the $V$ on the right side of (4.1) by $\hat{V}$ just obtained to get the next step estimator of $\beta$; and repeat the process. We
call such a procedure iterative weighted least squares, or I-WLS. Jiang et al. (2005) showed that, under very mild conditions, the probability that I-WLS converges at an exponential rate tends to one as the sample size increases. Furthermore, the limiting I-WLS estimator is consistent and asymptotically as efficient as the BLUE.

### 4.2 Binary responses and counts

In many cases, the responses collected over time are binary (i.e., 0 or 1, or YES/NO) or counts. The following is an example.

**Example 4.2 (Seizure counts).** Thall and Vail (1990) presented data from a clinical trial involving 59 epileptics who were randomized to a new drug or a placebo. A multivariate response variable consisted of the logarithm of a baseline seizures counts as well as seizure counts during the two-weeks before each of four clinic visits.

For longitudinal data that are binary or counts, such as in Example 4.2, generalized linear mixed models (GLMMs) discussed earlier may be used in the analysis. More specifically, one may consider a longitudinal GLMM defined as follows.

Suppose that there are $m$ independent clusters such that, within the $i$th cluster, the responses $y_{ij}$, $1 \leq j \leq n_i$ are conditionally independent given a $d$-dimensional vector $\alpha_i$ of random effects with conditional density

$$f(y_{ij} | \alpha_i) = \exp \left\{ \frac{y_{ij} \theta_{ij} - b(\theta_{ij})}{\phi} + c(y_{ij}, \phi) \right\},$$

where $\phi$ is a dispersion parameter, and the functions $b(\cdot)$ and $c(\cdot, \cdot)$ are the same as before. Furthermore, let $\mu_{ij} = E(y_{ij}|\alpha_i)$ and assume that

$$g(\mu_{ij}) = x'_{ij} \beta + z'_{ij} \alpha_i,$$

where $g$ is the link function, and $x_{ij}$ and $z_{ij}$ are known vectors. The random effects $\alpha_i$ is assumed to be distributed as $N(0, G)$.

Two issues deserve some attention. First, as discussed earlier, maximum likelihood estimator (MLE) under a longitudinal GLMM may be difficult to compute if $d$ is greater than one. Second, in many cases of longitudinal data, there may be serial correlations among the responses from the same subject. Such a serial correlation may not be taken into account by a GLMM. Note that, under the GLMM assumption, the repeated measures are conditionally
independent given the random effects, which means that no (additional) serial
correlation exists once the values of the random effects are specified. We
consider an example.

Example 4.3 (The salamander-mating data). Consider the salamander-
mating experiments discussed in subsection 2.2.2. McCullagh and Nelder
(1989) proposed a GLMM for analyzing the data, in which random effects
corresponding to the female/male animals were introduced. The dataset
and model have been extensively studied. However, in most cases it was assumed
that a different set of animals (20 for each sex) are used in each mating
experiment, although, in reality, the same set of animals were repeatedly used
in two of the experiments (McCullagh and Nelder 1989, §14.5). Furthermore,
most of the GLMMs used in this context (with the exception of, perhaps,
Jiang and Zhang 2001) assumed that no further correlation among the data
exists given the random effects. However, the responses in this case should
be considered longitudinal, because repeated measures were collected from
the same subjects (once in the summer, and once in the autumn of 1986).
Therefore, serial correlation may exist among the repeated responses even
given the random effects (i.e., the animals). In other words, the true
correlations among the data may not have been adequately addressed by
the GLMMs.

In the next subsection, we consider a method that may take into account
such serial correlations.

4.3 Generalized estimating equations

4.3.1 General

The general framework of estimating functions was set up by V. P. Go-
dambe some thirty years before that of generalized linear mixed models (Go-
dambe 1960). In Godambe (1991), the author viewed the approach as an
extension of the Gauss-Markov theorem. An estimating function is a func-
tion, possibly vector valued, that depends both on $y = (y_i)_{1 \leq i \leq n}$, a vector of
observations, and $\theta$, a vector of parameters. Denoted by $g(y, \theta)$, the estimat-
ing function is required to satisfy

$$E_\theta \{g(y, \theta)\} = 0 \quad (4.2)$$

for every $\theta$. For simplicity, let us first consider the case that $y_1, \ldots, y_n$ are
independent with $E(y_i) = \theta$, a scalar. Let $\mathcal{G}$ denote the class of estimating
functions of the form

\[ g(y, \theta) = \sum_{i=1}^{n} a_i(\theta)(y_i - \theta), \]

where \( a_i(\theta) \) are differentiable functions with \( \sum_i a_i(\theta) \neq 0 \). Then, an extension of the Gauss-Markov theorem states the following (Godambe 1991).

**Theorem 4.1.** If \( \text{var}(y_i) = \sigma^2, 1 \leq i \leq n, g^* = \sum_{i=1}^{n} (y_i - \theta) \) is an optimal estimating function within \( \mathcal{G} \) and the equation \( g^* = 0 \) provides \( \bar{y} \), the sample mean, as an estimator of \( \theta \).

The equation

\[ g(y, \theta) = 0 \quad (4.3) \]

to be solved for \( \theta \) is called an estimating equation. In Theorem 4.1, the optimality is in the following sense, which was also introduced by Godambe. Note that for (4.3) to be used as an estimating equation, the corresponding estimating function should be as close to zero as possible, if \( \theta \) is the true parameter. In view of (4.2), this means that one needs to minimize \( \text{var}(g) \).

On the other hand, in order to distinguish the true \( \theta \) from a false one, it makes sense to maximize \( \partial g / \partial \theta \), or the absolute value of its expected value. When both are put on the same scale, the two criteria for optimality can be combined by considering

\[ \frac{\text{var}(g)}{[E(\partial g / \partial \theta)]^2} = \text{var}(g_n), \quad (4.4) \]

where \( g_n = g / E(\partial g / \partial \theta) \) is a standardized version of \( g \). Thus, the optimality in Theorem 4.2 is in the sense that

\[ \text{var}(g_n^*) \leq \text{var}(g_n) \quad \text{for any } g \in \mathcal{G}. \]

Now consider a multivariate version of the estimating function. Let \( y \) be a vector of responses that is associated with a vector \( x \) of explanatory variables. Here we allow \( x \) to be random as well. Suppose that the (conditional) mean of \( y \) given \( x \) is associated with \( \theta \), a vector of unknown parameters. For notation simplicity, write \( \mu = E_\theta(y|x) = \mu(x, \theta) \), and \( V = \text{Var}(y|x) \). Here \( \text{Var} \) represents the covariance matrix, and \( \text{Var} \) or \( E \) without subscript \( \theta \) mean to be taken at the true \( \theta \). Let \( \mu \) denote the matrix of partial derivatives, i.e., \( \dot{\mu} = \partial \mu / \partial \theta \). Consider the following class of vector-valued estimating
functions \( \mathcal{H} = \{ G = A(y - \mu) \} \), where \( A = A(x, \theta) \), such that \( \text{E}(\hat{G}) \) is nonsingular. The following theorem can be established.

**Theorem 4.2.** Suppose that \( V \) is known, and that \( \text{E}(\mu'V^{-1}\mu) \) is nonsingular. Then, the optimal estimating function within \( \mathcal{H} \) is given by \( G^* = \mu'V^{-1}(y - \mu) \), that is, with \( A = A^* = \mu'V^{-1} \).

Here the optimality is in the similar sense as the univariate case. Define the partial order of nonnegative definite matrices as \( A \geq B \) if \( A - B \) is nonnegative definite. Then, the optimality in Theorem 4.2 is in the sense that the estimating function \( G^* \) maximizes, in the partial order of nonnegative definite matrices, the generalized information criterion

\[
\mathcal{I}(G) = \{ \text{E}(\hat{G}) \} \{ \text{E}(GG^*) \}^{-1} \{ \text{E}(\hat{G}) \}, \tag{4.5}
\]

where \( \hat{G} = \partial G/\partial \theta' \). It is easy to see that (4.5) is, indeed, the Fisher information matrix when \( G \) is the score function corresponding to a likelihood, that is, \( G = \partial \log(L)/\partial \theta \), where \( L \) is the likelihood function - and this provides another view at Godambe’s criterion of optimality. Also, (4.5) is equal to the reciprocal of (4.4) in the univariate case, so that maximizing (4.5) is equivalent to minimizing (4.4). See Heyde (1997) for more detail.

### 4.3.2 Generalized estimating equations (GEE)

In the case of longitudinal GLMM, the optimal estimating function according to Theorem 4.2 can be expressed as

\[
G^* = \sum_{i=1}^{m} \mu_i'V_i^{-1}(y_i - \mu_i),
\]

where \( y_i = (y_{ij})_{1 \leq j \leq n_i}, \mu_i = \text{E}(y_i) \) and \( V_i = \text{Var}(y_i) \). Here, as in the earlier sections, the covariates \( x_i \) are considered fixed rather than random. The corresponding estimating equation is known as generalized estimating equation, or GEE (Liang and Zeger 1986), given by

\[
\sum_{i=1}^{m} \mu_i'V_i^{-1}(y_i - \mu_i) = 0. \tag{4.6}
\]

In (4.6), it is assumed that \( V_i, 1 \leq i \leq m \) are known because, otherwise, the equation cannot be solved. However, the true \( V_i \)’s are unknown in practice. Note that, under a GLMM, the \( V_i \)’s may depend on a vector of
variance components, $\theta$, in addition to $\beta$, that is, $V_i = V_i(\beta, \theta), 1 \leq i \leq m$. If a GLMM is not assumed and neither is any other parametric model for the covariances, the $V_i$'s may be completely unknown. Liang and Zeger proposed to use “working” covariance matrices instead of the true $V_i$'s to obtained the GEE estimator. For example, one may use the identity matrices which correspond to a model assuming independent errors with equal variance. The method is justified in the following sense. As is shown by Liang and Zeger, under some regularity conditions, the resulting GEE estimator is consistent despite that the working covariances misspecify the true $V_i$'s. However, the estimator based on working $V_i$'s may be inefficient as compared to that based on the true $V_i$’s.

Alternatively, one may replaced the $V_i$’s in (4.6) by their consistent estimators, say, $\hat{V}_i$. For example, under a GLMM, if $\theta$ is replaced by a $\sqrt{m}$-consistent estimator, say, $\hat{\theta}$ (i. e., $\sqrt{m}(\hat{\theta} - \theta)$ is bounded in probability), the resulting GEE estimator is asymptotically as efficient as the GEE estimator based on the true $V_i$’s. (Of course, the latter is not an estimator unless the $V_i$’s are known.) This means that $\sqrt{m}(\hat{\beta} - \beta)$ has the same asymptotic covariance matrix as $\sqrt{m}(\hat{\beta} - \beta)$, where $\beta$ is the solution to (4.6) with $\theta$ replaced by $\hat{\theta}$, and $\hat{\beta}$ is that with the true $V_i$’s (e. g., Liang and Zeger 1986). However, to find a $\sqrt{m}$-consistent estimator one typically needs to assume a parametric model for the $V_i$’s, which increases the risk of model misspecifications. Even under a parametric covariance model, the $\sqrt{m}$-consistent estimator may not be easy to compute, especially if the model is complicated. Recently, Jiang et al. (2005) proposed an iterative alternative which offers a more robust and computationally attractive solution. The approach may be regarded as an extension of the I-WLS method discussed in subsection 4.1. We conclude this subsection with an example

**Example 4.4 (Mothers’ stress and children’s morbidity).** Zeger and Liang (1986) applied the GEE method to a dataset from a study on the association between mothers’ stress and children’s morbidity. The study involved 167 mothers with infants between ages of 18 months and 5 years. Each mother was asked to keep diary on whether her child was ill as well as her own stressness. The first 9 days of the diaries were used. The authors analyzed the data using the GEE with working covariance matrices. More specifically, four different working covariances were considered. There were (i) independence; (ii) 1-dependence; (iii) stationary; and (iv) exchangeable. See Zeger and Liang (1986) for more detail.
4.4 Informative missing data

In this section, we consider extension of the GEE method to the situation where information about the process of the missing data is taken into account. Following Robins et al. (1995), we consider a follow-up study, in which the responses are denoted by $y_{it}$, $0 \leq t \leq T$ with $y_{i0}$ being the measurement just prior to the start of the follow-up. Again, let $X_i = (X'_{it})_{0 \leq t \leq T}$ denote a matrix of explanatory variables associated with the $i$th subject, where $X_{it} = (X_{it})_{i \leq t \leq p}$. We assume that $X_i$ is completely observed, $1 \leq i \leq n$.

It is known that a fully parametric likelihood method still provides valid inference about $\beta$, if the model is correctly specified and the prob. of missing at time $t$ does not depend on the current and future data (i. e., at times $s \geq t$; Rubin 1976). However, with incomplete data, the likelihood-based methods can be sensitive to model misspecification and, even with complete data, computational challenging.

Furthermore, the GEE method if valid only under the stronger assumption that the data is missing completely at random, i. e., the nonresponse process is independent of the data process.

We assume a semiparametric regression model:

$$E(y_{it}|X_i) = g_t(X_i, \beta). \tag{4.7}$$

Write the left side of (4.7) as $\mu_{it}$, and let $\mu_i = (\mu_{it})_{0 \leq t \leq T}$. Furthermore, we assume that, in addition to $y_{it}$ and $X_i$, measures on a vector of time-dependent covariates $V_{it}, 0 \leq t \leq T$ are available. Let $W_{i0} = (X'_{i0}, \ldots, X'_{iT}, y_{i0}, V'_{i0})'$, and $W_{it} = (y_{it}, V'_{it})', 1 \leq t \leq T$. The notation $W_{it}$ denotes $\{W'_{i0}, \ldots, W'_{it-1}\}'$, that is, the vector of all the data up to time $t - 1$.

Define $R_{it} = 1$ if subject $i$ is observed at time $t$, that is, if both $y_{it}$ and $V_{it}$ are observed, and $R_{i0} = 0$ otherwise. We assume that $y_{it}$ and $V_{it}$ are both observed or both missing, and $R_{i0} = 1$. We also assume that, once a subject leaves the study, the subject will not return. This means that $R_{it} = 1$ implies $R_{i(t-1)} = 1, \ldots, R_{i1} = 1$. The following assumptions are made regarding the missing process $R_{it}$:

$$P\{R_{it} = 1|R_{i(t-1)} = 1, W_{it}, y_{it}\} = P\{R_{it} = 1|R_{i(t-1)} = 1, W_{it}\}, \tag{4.8}$$

and, denoting the right side of (4.8) by $\lambda_{it}$,

$$\lambda_{it} \geq \delta, \quad \lambda_{it} = \lambda_{it}(W_{it}, \psi), \tag{4.9}$$

$$\lambda_{it} = \lambda_{it}(W_{it}, \psi), \tag{4.10}$$

45
where $\delta$ is some positive constant, and $\psi$ an unknown vector of parameters. Conditions (4.8) - (4.10) are discussed in Robins et al. (1995). According to the latter, $\psi$ can be estimated by the maximum partial likelihood method. The resulting estimator, denoted by $\hat{\psi}$, does not depend on the estimation of the other parameters. Write $\pi_{it} = \prod_{s=1}^{t} \lambda_{is}$, and $\Delta_{it} = \pi_{it}^{-1}R_{it}$. Define $\tilde{y}_{it} = \Delta_{it}y_{it}$. According to Lemma A.1 of Robins et al. (1995), we have $E(\tilde{y}_{it}|X_i) = E(y_{it}|X_i) = \mu_{it}$. Also, let $\tilde{V}_i = \text{Var}(\tilde{y}_i|X_i)$, where $\tilde{y}_i = (\tilde{y}_{it})_{0 \leq t \leq T}$. Then, according to the previous subsection, when $\psi$ and $\tilde{V}_i$'s are known, the following estimating equation is optimal:

$$
\sum_{i=1}^{n} \mu_i^t \tilde{V}_i^{-1}(\tilde{y}_i - \mu_i) = 0.
$$

If $\psi$ is unknown, since $\tilde{y}_i$ depends on $\psi$, i.e., $\tilde{y}_i = \tilde{y}_i(\psi)$, we replace $\psi$ by $\hat{\psi}$, the maximum partial likelihood estimator, to get $\hat{y}_i = \hat{y}_i(\hat{\psi})$. Thus, the estimating equation becomes

$$
\sum_{i=1}^{n} \mu_i^t \tilde{V}_i^{-1}(\hat{y}_i - \mu_i) = 0. \tag{4.11}
$$

Note that the only real difference between equations (4.6) and (4.11) is that $y_i$ is replaced by $\hat{y}_i$.

Robins et al. (1995) showed that, after incorporating information about the missing data as above, the resulting GEE estimator for $\beta$ is consistent, under some regularity conditions. However, the asymptotic covariance matrix of their GEE estimator with estimated missing probabilities (by $\hat{\psi}$) is “smaller” than that of the GEE estimator with the true missing probabilities. In other words, the GEE estimator with the estimated missing probabilities is asymptotically at least as efficient as that with the true missing probabilities (see the discussion on page 110 of the above reference). Also see Ying (2003, section 2).

References


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


**Department of Statistics**  
**University of California**  
**One Shields Avenue**  
**Davis, CA 95616**  
**U. S. A.**