The performance of estimation methods for generalized linear mixed models

Damian Collins
University of Wollongong

The performance of estimation methods for generalized linear mixed models

A thesis submitted in fulfillment of the

requirements for the award of the degree

Doctor of Philosophy

from

University of Wollongong

by

Damian Collins BSc (Hons) UNSW

School of Mathematics and Applied Statistics

June 2008

THIS PAGE IS BLANK

I, Damian Paul Collins, declare that this thesis, submitted in fulfillment of the requirements for the award of Doctor of Philosophy in the School of Mathematics and Applied Statistics, University of Wollongong, is wholly my own work unless otherwise referenced or acknowledged. This document has not been submitted for qualifications at any other academic institution.

Damian Collins

June, 2008

Acknowledgments

I would first like to thank my supervisors, Ken Russell, Robin Thompson and Brian Cullis, for giving me this wonderful opportunity to improve myself, and for supporting me throughout this process. I am sorry that it was such a saga for you all, and that I didn't listen to your advice as well as I should have.

I would also like to thank Idris Barchia, Paul Nicholls and other biometricians in NSW DPI for taking care of the consulting workload during my absence.

I would also like to thank Gwenda Thompson for her friendship and camaradarie during the latter stages.

I would also like to thank the staff at SMAS at UOW for providing a supportive environment for research.

Finally, I would like to thank my parents, for always being there when I needed someone to talk to and providing a home away from home.

The NSW DPI provided me with a generous financial assistance package during my years of research.

Abstract

Generalised linear models (GLMs) are a flexible class of non-linear models for non-normally distributed response data. GLMs encompass models for discrete response data which takes one of several values rather than being measured on a continuous scale. Discrete response data is abundant in agricultural and biological research, for instance, in the mortality of animals and plants (binary/binomial data) and the scoring of disease (ordinal data).

Generalised linear mixed models (GLMMs) are an extension of GLMs which include additional random effects in the (conditional) linear predictor. Some examples of where GLMMs may be useful include the analysis of designed experiments, surveys, spatial data and longitudinal or repeated measures data.

The fundamental difficulty in using GLMMs is that no closed analytical expression for the likelihood is available. A variety of approaches have been proposed to circumvent this difficulty, including approximate likelihood approaches, such as penalized quasi-likelihood (PQL), numerical approaches, such as Gauss-Hermite quadrature (GHQ), and approaches based on the use of Monte Carlo methods, such as modern Bayesian approaches implementing Markov Chain Monte Carlo (MCMC) techniques.

Although in recent years more attention in the literature has been given to Bayesian approaches and other approaches based on Monte Carlo techniques for GLMMs, there is still widespread interest amongst practitioners in the use of approximate likelihood approaches, especially with the work of Lee & Nelder (2001, 2006). The objective of this PhD is primarily to explore the approximate likelihood approaches, as well as comparing and contrasting them with numerical and Monte Carlo approaches.

The most widely known approximate likelihood approach, PQL, is well-known to give biased estimators of the GLMM parameters for binary grouped data when the group size is small. However, the other two groups of approaches for GLMMs are not without problems. Numerical approaches such as GHQ are only suitable for GLMMs with nested random effects only, and often require very good starting values to achieve convergence. Approaches based on Monte Carlo techniques can be very computational intensive and also have convergence problems, as well as being sensitive to the choice of priors, when used within the Bayesian paradigm. The approximate likelihood approach of Lee and Nelder is claimed, by its proponents, to enjoy the computational efficiency of PQL whilst not suffering from the estimation bias issues that PQL experiences.

A background to the GLMM and inferential issues is provided in Chapter 1, with theoretical material and alternative approaches for modelling correlation in non-normal data, such as the generalized estimating equation (GEE) approach. It is argued that the GLMM is the most generally applicable model for modelling correlation and clustering in non-normal data available at present. The second chapter reviews the main estimation approaches for GLMMs, discussing in more detail the issues associated with each of the approaches already highlighted above.

Chapters 3 and 4 focus on the two most popular approximate likelihood approaches, PQL and the hierarchical GLM (HGLM) approach of Lee & Nelder (2001, 2006) respectively. Simulation studies are presented in Chapter 3 for binary and sparse Poisson data from a range of designs. These studies show that the two main factors associated with estimation biases are the group sizes and the relative magnitude of the variance components (as well as the sparcity of the Poisson data). These studies also suggest that hypothesis testing for fixed effects, against the usual null hypothesis of zero effect, can be reliably conducted using Wald tests using the estimated variance-covariance matrix of the fixed effects from PQL. Finally, they also indicate that the first order Laplace approximation may be useful for calculating approximate likelihood ratio tests for testing variance components. Chapter 4 contains discussion

of the HGLM approach of Lee and Nelder, which relies on either a first or second order approximation of the likelihood. Computational issues associated with the use of the HGLM approach are discussed in the context of a Fortran 90 implementation. Further simulation studies show that estimation biases for HGLM approaches are generally much smaller in magnitude than PQL, but the HGLM estimators can also be unstable for binary models with conditional expectations near 0 or 1. Some heuristic arguments for the relative performance of the HGLM approaches versus PQL are also presented.

Estimation biases for the PQL and the HGLM approaches are compared with Bayesian and GHQ approaches in Chapter 5 using a series of case studies. The approximate likelihood approaches performed reasonably well against Bayesian and GHQ approaches for all case studies presented, with the exception of the Rodriguez & Goldman (2001) datasets, with no finite maximum for the likelihood found using the (second order) HGLM approaches. The second order HGLM approach gave similar estimates to the Bayesian and GHQ approaches in a paired binary simulation study. Despite greater estimation biases, the PQL estimators had lower MSE than the GHQ estimators in a second paired binary (and Poisson) simulation study, in which the Bayesian estimator, with default priors, suffered estimation bias as well. PQL also performed relatively well against other approaches in a simulation study involving a randomised complete block design (RCBD) and in a simulation study involving a spatial GLMM, where PQL was compared with a much more computationally intensive Bayesian approach. These simulations also showed that the "REML-like" correction to the likelihood used by the HGLM and Bayesian approaches can give some positive estimation bias.

Whilst both approximate likelihood approaches had difficulties either in terms of estimation bias or instability, in general they perform relatively well against the other approaches and provide a useful and efficient way of fitting a wide variety of GLMMs. The use of a first or second order HGLM approach is generally preferable to PQL to achieve lower estimation biases. If PQL is employed, it is suggested that the

first order Laplace approximation be calculated for approximate testing of variance components.

Contents

List of Figures

List of Tables

THIS PAGE IS BLANK

Chapter 1

Review of basic elements of theory

This chapter provides a review of background theory necessary for a discussion of generalized linear mixed models (GLMMs) and approaches in following chapters. GLMMs are a fusion of generalized linear models (GLMs) and linear mixed models (LMMs). This chapter summarises inferential techniques for these two parent model classes, since these inferential techniques are also the basis of some of the approaches devised for GLMMs.

1.1 Linear and generalized linear models and classical inferential approaches

1.1.1 Linear models

Linear models with independent normally distributed errors and common variance, which will be referred to as "normal linear models", have been central to applied statistical work for several generations of applied statisticians. They encompass a wide range of statistical models, including, for instance, Analysis of Variance (ANOVA) models, which are particularly useful for analysing designed experiments. Recent

advances in computational speed, as well as widespread availability of software, have made it computationally trivial to fit a normal linear model, even for relatively large datasets.

The normal linear model is illustrated in the analysis of data from a completely randomised design (CRD). It is assumed that t treatments were randomised to n experimental units. In addition, a covariate is available to allow for differences between the units prior to treatment application. The linear model for the ith experimental unit is

$$y_i = \tau_0 + \tau_{j(i)} + x_i\tau_{t+1} + e_i, \qquad i = 1, \ldots, n, \tag{1.1}$$

where y_i is the response, τ_0 is the overall mean, $\tau_{j(i)}$ is the effect of the treatment assigned to the ith unit, x_i is the covariate value with associated regression coefficient τ_{t+1}, and $e_i \sim N(0, \sigma^2)$ are random errors.

A general formulation for a normal linear model is

$$y_i = \boldsymbol{x}_i^T\boldsymbol{\tau} + e_i, \qquad i = 1, \ldots, n,$$

where $\boldsymbol{x}_i$ is a vector of values of the regressor variables for the ith observation with associated vector of coefficients $\boldsymbol{\tau}$ of length p, and $e_i \sim N(0, \sigma^2)$ is the normally distributed error for the ith observation. This general formulation can also be expressed in matrix/vector notation as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{e},$$

where $\boldsymbol{y} = (y_1 \ldots y_n)^T$ is a vector of responses, $\boldsymbol{X}$ is the $n \times p$ design matrix whose ith row is $\boldsymbol{x}_i^T$, and $\boldsymbol{e} = (e_1, \ldots, e_n)^T \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ is a vector of independent and identically distributed (i.i.d.) errors. To illustrate this general formulation, the above example (1.1) will be used. Assuming that there are $r = n/t$ replications of each treatment and that the units are ordered by treatment, the matrix $\boldsymbol{X}$ and vector $\boldsymbol{\tau}$

would be as follows:

$$\boldsymbol{X} = \begin{pmatrix} \mathbf{1}_r & \mathbf{1}_r & \mathbf{0}_r & \dots & \mathbf{0}_r & \boldsymbol{x}_{(1)} \\ \mathbf{1}_r & \mathbf{0}_r & \mathbf{1}_r & \vdots & \vdots & \boldsymbol{x}_{(2)} \\ \vdots & \vdots & \mathbf{0}_r & \ddots & \mathbf{0}_r & \vdots \\ \mathbf{1}_r & \vdots & \vdots & \mathbf{0}_r & \mathbf{1}_r & \boldsymbol{x}_{(t)} \end{pmatrix}, \qquad \boldsymbol{\tau} = \begin{pmatrix} \tau_0 \\ \tau_1 \\ \tau_2 \\ \vdots \\ \tau_t \\ \tau_{t+1} \end{pmatrix},$$

where $\mathbf{1}_r$ and $\mathbf{0}_r$ are vectors of length r consisting of 1 or 0 respectively and $\boldsymbol{x}_{(i)}$ is the vector of covariate values for units assigned the ith treatment. Note that the matrix $\boldsymbol{X}$ is not full rank, since the first column is the sum of the next t columns. However, $\boldsymbol{X}$ can easily be made full rank by dropping the first column and, correspondingly, dropping τ_0 from the vector of coefficients, for instance. In further discussion, it is assumed that $\boldsymbol{X}$ is always full rank.

Amongst the numerous texts available on normal linear models, Draper & Smith (1998) is considered one of the standard references.

1.1.2 Generalized linear models

Generalized linear models (GLMs) are, as the name suggests, a generalization or extension of normal linear models. GLMs incorporate normal linear models as a special case, but also cater for other error distributions, in particular, error distributions catering for discrete data such as the Poisson and binomial distributions. GLMs originated from a variety of different analysis problems, including dilution assay to determine infective organism concentration, probit analysis in toxicology experiments and log-linear models for cross-tabulations of counts (McCullagh & Nelder, 1989, chapter 1). Nelder & Wedderburn (1972) were the first to propose the generalized linear model to encompass these different models under one unified mathematical framework.

To illustrate the formulation of a GLM, the example (1.1) for the normal linear model

above will be modified. It is now assumed that the responses y_i follow a Poisson distribution, and that the effects of treatments and covariates are multiplicative rather than additive. Hence, we assume that

$$\mathrm{E}(y_i) = \mu_i = \exp(\tau_0 + \tau_{j(i)} + x_i\tau_{t+1}), \tag{1.2}$$

where μ_i, τ_0, $\tau_{t(i)}$, x_i and τ_{t+1} are as before.

The specification of a generalized linear model is:

1. the probability distribution of the data is a member of the exponential family of distributions, where the probability density function (PDF) can be written in the following general form for one observation y_i, $i = 1, \ldots, n$:

$$f(y_i; \theta_i) = \exp\left(\frac{(y_i\theta_i - b(\theta_i))}{a_i(\phi)} + c(y_i, \phi_i)\right), \tag{1.3}$$

   where θ_i and ϕ are often called the canonical and dispersion parameters respectively, and a_i, b and c are arbitrary functions.

2. the mean of the response $\mu_i = \mathrm{E}(y_i)$ is related to a linear function of regressor variables in the vector $\boldsymbol{x}_i$ of length p via a link function $g(\mu_i) = E(y_i) = \boldsymbol{x}_i^T\boldsymbol{\tau}$, where the elements of $\boldsymbol{\tau}$ are the associated regression coefficients. The linear function $\eta_i = \boldsymbol{x}_i^T\boldsymbol{\tau}$ is generally referred to as the *linear predictor*. In the above example (1.2), the link function is the logarithmic function.

The exponential family of distributions contains many of the well-known statistical distributions, including the normal, binomial, Poisson and gamma distributions. Distributions in this family have a number of desirable mathematical and statistical features, as discussed in Barndorff-Nielsen (1978); for instance, if $y_1, \ldots, y_n$ have PDF $f(y; \theta)$, then the mean $\overline{y}$ is a sufficient statistic for θ.

The first two moments of the general distributional form in (1.3) are $\mathrm{E}(y_i) = \mu_i = b'(\theta_i)$ and $\mathrm{Var}(y_i) = b''(\theta_i)a_i(\phi)$. The canonical parameter θ can be expressed as a function of the linear predictor $\theta = f(\eta)$ where $f(\cdot) = b_d^{-1}\{g^{-1}(\cdot)\}$ and $b_d = b'$. A

canonical link function g satisfies $g^{-1} = b'$, and so the canonical parameter equals the linear predictor $\theta = \eta$. Other properties of GLMs can be found in a standard reference on GLMs, for instance, McCullagh & Nelder (1989), pp 29-32.

Let $v(\cdot) = b''(b_d^{-1}(\cdot))$ be the "variance function", where $b_d = b'$ as before. The relation between the variance and the mean, encapsulated in the variance function $v = v(\mu)$, uniquely identifies the exponential family distribution concerned. For instance, $v(\mu) = 1$, μ and μ^2 are the variance functions for the normal, Poisson and gamma distributions respectively.

The functions $a_i(\cdot)$, $i = 1, \ldots, n$, are most commonly of the form $a_i(\phi) = a_i\phi$, where a_i are known constants, and this will be assumed from now on. The dispersion parameter ϕ can be either known or estimated from the data. For instance, it is equal to 1 for the Poisson and binomial distributions, but equivalent to the residual variation σ^2 for Normal data. The constants a_i are also usually equal to 1, but for binomial (grouped binary) or other grouped data, $a_i = 1/m_i$, where m_i is the binomial denominator or group size respectively.

One of the uses of the inverse link g^{-1} is to map the linear predictor $\eta = \boldsymbol{x}^T\boldsymbol{\tau}$ to a valid range for the response variable. For instance, for binary data, the probit link function $\Phi(\mu) = \eta$, where Φ is the cumulative density function (CDF) of a $N(0, 1)$ distribution, transforms the linear predictor to take values between 0 and 1, since $\mu = \Phi^{-1}(\eta)$ represents the probability of a successful response. The GLM model can also be interpreted as a model for data generated from a "latent" continuous variable. For instance, binary data arising from a model with probit link function can be generated by dichotomising a normally distributed variable with mean $\boldsymbol{x}^T\boldsymbol{\tau}$ and variance 1, according to whether it is above or below 0.

1.1.3 Maximum likelihood estimation

A statistical model can be defined by a PDF $f(\boldsymbol{y}; \boldsymbol{\theta})$ for data $\boldsymbol{y}$ with fixed, but unknown, parameters $\boldsymbol{\theta}$ which delineate the model. The likelihood function is equivalent to the PDF for fixed observed data $\boldsymbol{y}$, that is, $L(\boldsymbol{\theta}; \boldsymbol{y}) = f(\boldsymbol{y}; \boldsymbol{\theta})$, and $\boldsymbol{\theta}$ is allowed

to vary. A log-likelihood function $\ell = \log L(\boldsymbol{\theta}; \boldsymbol{y})$ is more useful in practice than the unlogged version because of its statistical properties, and this will be referred to subsequently as the "log-likelihood".

Maximum likelihood (ML) estimates of the unknown parameters $\boldsymbol{\theta}$ are defined as estimates that maximize $\ell(\boldsymbol{\theta}; \boldsymbol{y})$ (or $L(\boldsymbol{\theta}; \boldsymbol{y})$). The score statistic and Fisher information matrix are the vector of first derivatives of the likelihood, $\partial\ell/\partial\boldsymbol{\theta}$, and the negative expectation of the matrix of second derivatives, $\boldsymbol{I} = -E(\partial^2\ell/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T)$, respectively. Two key results for the score statistic and information matrix respectively are:

1. the expectation of the score statistic is zero (where the expectation is taken over the distribution of $\boldsymbol{y}$ at $\boldsymbol{\theta}$):

$$\mathrm{E}_{\boldsymbol{\theta}}\left(\frac{\partial\ell}{\partial\boldsymbol{\theta}}\right) = \mathbf{0};$$

2. the information is the variance of the score statistic:

$$-\mathrm{E}\left(\frac{\partial^2\ell}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right) = \mathrm{E}\left(\frac{\partial\ell}{\partial\boldsymbol{\theta}}\frac{\partial\ell}{\partial\boldsymbol{\theta}^T}\right)^2 = \mathrm{Var}\left(\frac{\partial\ell}{\partial\boldsymbol{\theta}}\right).$$

By the Cramer-Rao Lower Bound (CRLB) theorem, the diagonal elements of the inverse of the Fisher information matrix are the minimum variance of each element of $\boldsymbol{\theta}$ that any estimator of $\boldsymbol{\theta}$ can attain. Maximum likelihood may not necessarily provide the best unbiased estimator, and the estimator may even be biased for small sample sizes. However, asymptotically, maximum likelihood estimators of $\boldsymbol{\theta}$ are unbiased and have variance-covariance matrix equal to the CRLB of $\boldsymbol{\theta}$.

1.1.3.1 Methods of computing ML estimates

To obtain ML estimates of the unknown parameters $\boldsymbol{\theta}$, one can solve $\partial\ell/\partial\boldsymbol{\theta} = \mathbf{0}$ for $\boldsymbol{\theta}$ (assuming that $\partial^2\ell/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T \leq 0$ for all $\boldsymbol{\theta}$). Often, however, no closed form expression for the solution to the score function $\partial\ell/\partial\boldsymbol{\theta} = \mathbf{0}$ exists. Therefore, an iterative method of solution is required, such as Fisher scoring or the EM algorithm.

Fisher scoring and quasi-Newton techniques: Fisher scoring, also known as the Gauss-Newton algorithm, is a variation on the Newton-Raphson iterative technique for finding the maximum or minimum of a non-linear function $f(\boldsymbol{\theta})$. Here it is assumed that $f(\boldsymbol{\theta})$ is being maximized. Given an estimate $\hat{\boldsymbol{\theta}}^{(k-1)}$ from the $(k-1)$th iteration, the Newton-Raphson technique calculates the estimate $\hat{\boldsymbol{\theta}}^{(k)}$ for the kth iteration as

$$\hat{\boldsymbol{\theta}}^{(k)} = \hat{\boldsymbol{\theta}}^{(k-1)} - \left\{ \left(\frac{\partial^2 f}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right)^{-1} \left(\frac{\partial f}{\partial \boldsymbol{\theta}} \right) \right\}_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(k-1)}}.$$

In Fisher scoring, the negative inverse of the Hessian, $-(\partial^2 f / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T)^{-1}$, is replaced by the information matrix, $\boldsymbol{I}(\boldsymbol{\theta}) = -E(\partial^2 f / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T)$, since it is often easier to compute. Fisher scoring is an example of a *quasi-Newton* method, where the Hessian is replaced by an alternate formulation or an approximation $\boldsymbol{A}$, therefore the step size is $-\boldsymbol{A}^{-1}(\partial f / \partial \boldsymbol{\theta})$. Quasi-Newton techniques can be shown to have quadratic convergence near the true solution, but can also easily "overshoot" the maximum. One approach to correcting the latter is to reduce the step size until the function increases. Smyth (1997) reviews quasi-Newton and other optimisation techniques with emphasis on statistical applications. One important implementation of Fisher scoring is the iteratively reweighted least squares (IRLS) technique to solve GLMs (section 1.1.3.2).

The EM algorithm: The EM (expectation-maximization) algorithm (Dempster, Laird & Rubin, 1977) obtains maximum likelihood estimates by creating auxiliary or "missing" data. The data vector $\boldsymbol{y}$ is augmented with missing data, denoted $\boldsymbol{z}$, so that the "complete" likelihood is based on the joint PDF, $f(\boldsymbol{y}, \boldsymbol{z}; \boldsymbol{\theta})$.

Heuristically, the EM algorithm repeatedly "guesses", and then maximizes, the complete log-likelihood with respect to $\boldsymbol{\theta}$. Formally, the kth iteration of the EM algorithm involves two steps:

1. The E Step: determine $E_{z|y}\{\log f(\boldsymbol{y}, \boldsymbol{z}; \boldsymbol{\theta})\}$ over the conditional distribution of $\boldsymbol{z}$ given $\boldsymbol{y}$ and current estimates $\boldsymbol{\theta}^{(k-1)}$, i.e.

$$\ell_E^{(k)} = \int \log f(\boldsymbol{y}, \boldsymbol{z}; \boldsymbol{\theta}) f(\boldsymbol{z}|\boldsymbol{y}; \boldsymbol{\theta}^{(k-1)}) d\boldsymbol{y}.$$

2. The M Step: maximize $\ell_E^{(k)}$ to give $\boldsymbol{\theta}^{(k)}$.

The EM algorithm is particularly useful for mixed models, for instance linear mixed models (section 1.2), where the random effects constitute the missing data. It can be shown that the EM algorithm will always converge to a maximum likelihood solution (Dempster *et al.*, 1977), but may do so very slowly (e.g. Lindstrom & Bates, 1988). Recent attempts to improve the EM algorithm include the "working parameter" method (Meng & van Dyk, 1997) or the PX-EM algorithm (Liu, Rubin & Wu, 1998; Liu & Wu, 1999), which essentially apply a suitable transformation to $\boldsymbol{z}$ to improve the convergence rate. Other variations include using Monte Carlo to evaluate the E step (Wei & Tanner, 1990) and using ECM (expectation-conditional maximisation) for the M step (Meng & Rubin, 1993).

1.1.3.2 Maximum likelihood estimation in linear and generalized linear models

Maximum likelihood techniques are well-defined for normal linear and generalized linear models. For normal linear models where $y_i \sim N(\boldsymbol{x}_i^T\boldsymbol{\tau}, \sigma^2)$, $i = 1, \ldots, n$, maximizing the likelihood is equivalent to minimizing the sum of squares of the residuals $\sum_{i=1}^n (y_i - \boldsymbol{x}_i^T\boldsymbol{\tau})^2$, with the ML estimator being $\hat{\boldsymbol{\tau}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ if $\boldsymbol{X}$ is full rank. It may be necessary to use unequal weights, w_i, for the observations, for instance, if there is heterogeneity in the variability of the residuals, that is, $e_i \sim N(0, \sigma^2\xi_i)$, in which case the ML estimator is $\hat{\boldsymbol{\tau}} = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{y}$, where $\boldsymbol{W} = \mathrm{diag}(w_i)$ and $w_i = \xi_i^{-1}$.

For generalized linear models, $\partial\ell/\partial\boldsymbol{\theta} = \mathbf{0}$ generally yields no closed form solution for $\boldsymbol{\theta}$, so an iterative method of solution is required. The standard technique for solving a GLM is called iteratively reweighted least squares (IRLS) and is derived using Fisher scoring (section 1.1.3.1). It can be shown that the estimate of $\boldsymbol{\tau}$ at the kth iteration is

$$\hat{\boldsymbol{\tau}}^{(k)} = (\boldsymbol{X}^T\boldsymbol{W}^{(k-1)}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}^{(k-1)}\boldsymbol{\psi}^{(k)},$$

where $\boldsymbol{W}^{(k-1)} = \text{diag}\{w_i^{(k-1)}\}$ is a diagonal matrix of GLM weights,

$$w_i^{(k-1)} = \left\{ a_i g'(\mu_i^{(k-1)}) v(\mu_i^{(k-1)}) \right\}^{-1},$$

and $\boldsymbol{\psi}^{(k-1)} = (\psi_1^{(k-1)} \dots \psi_n^{(k-1)})^T$ is a so-called "working variable" with elements

$$\psi_i^{(k)} = \eta_i^{(k-1)} + g'(\mu_i^{(k-1)})(y_i - \mu_i^{(k-1)}).$$

The quantities $\eta_i^{(k-1)} = \boldsymbol{x}_i^T \hat{\boldsymbol{\tau}}^{(k-1)}$ and $\mu_i^{(k-1)} = g^{-1}(\eta_i^{(k-1)})$ are calculated using the current estimates $\hat{\boldsymbol{\tau}}^{(k-1)}$. It is also easy to extend a GLM to allow for under- or over-dispersion of the data y_i, for distributions where ϕ is assumed to be 1, such as the Poisson and binomial distributions. For these distributions, the dispersion component ϕ can be estimated from the data, using either the deviance or Pearson residuals – see, for instance, McCullagh & Nelder (1989) pp 124-127 for details. Note that, by estimating ϕ, we no longer have a closed form for the distribution of the response, and so estimation for this extended model should be considered an instance of maximum quasi-likelihood (see section 1.1.3.4 below) rather than maximum likelihood.

1.1.3.3 Marginal, conditional, integrated and profile likelihood

Where $\boldsymbol{\theta} = (\theta_1 \dots \theta_p)^T$ is a vector (i.e. $p > 1$), the use of ML estimation may result in biased estimators for an individual parameter θ_i. For instance, where $Y_i \sim N(\mu, \sigma^2)$, $i = 1, \dots, n$, the parameter vector is $\boldsymbol{\theta} = (\mu, \sigma^2)^T$, and the ML estimator, $\hat{\sigma}^2 = \sum_{i=1}^n (y_i - \bar{y})^2/n$, is negatively biased.

A number of general strategies to provide an improved "likelihood" to overcome this bias are described briefly below. It is assumed that $\boldsymbol{\theta}$ can be split into two components, firstly, the parameters of interest $\boldsymbol{\psi}$, and secondly, the "nuisance" parameters $\boldsymbol{\lambda}$. In the example, it is assumed that σ is the component of interest, and μ is the nuisance parameter.

1. Marginal likelihood: A marginal likelihood is formed from the PDF of a statistic $\boldsymbol{T}(\boldsymbol{y})$ which depends only on the parameters of interest. For instance, if $Y_i \sim$

$N(\mu, \sigma^2)$, $i = 1 \ldots n$, the statistic

$$\boldsymbol{T}(\boldsymbol{y}) = (Y_1 - \overline{Y}, Y_2 - \overline{Y}, \ldots, Y_n - \overline{Y})^T \sim N\left(\boldsymbol{0}, \sigma^2 \left[\boldsymbol{I}_n - \boldsymbol{J}_{nn}/n\right]\right),$$

has a distribution which does not depend on the nuisance parameter μ. The $n \times n$ matrices $\boldsymbol{I}_n$ and $\boldsymbol{J}_{nn}$ here are the identity matrix and a matrix of 1s respectively.

2. Conditional likelihood: A conditional likelihood is formed using the conditional PDF of the data given $\boldsymbol{S}(\boldsymbol{y})$, a sufficient statistic for the nuisance parameters $\boldsymbol{\lambda}$. If $y_i \sim N(\mu, \sigma^2)$, $i = 1 \ldots n$, the conditional PDF of $\boldsymbol{y}$ given a sufficient statistic for μ, $\hat{\mu} = \overline{y}$, is

$$-\frac{n-1}{2} \log \sigma^2 - \frac{\sum_i (y_i - \overline{y})^2}{2\sigma^2},$$

which is a function of σ^2 alone.

Both the formation of marginal and conditional likelihoods for inference can be considered as specific cases of Cox's partial likelihood approach (Cox, 1975).

3. Integrated likelihood: Motivated partly by Bayesian theory, an integrated likelihood can be formed by simply integrating out the nuisance parameters from the likelihood function. A weighting function, or non-informative prior, may need to be applied in the integrand (Berger, Liseo & Wolpert, 1999).

4. Profile likelihood: A profile likelihood for $\boldsymbol{\psi}$ is formed by substituting the maximum likelihood estimate of $\boldsymbol{\lambda}$ given $\boldsymbol{\psi}$, $\hat{\boldsymbol{\lambda}}_\psi$, into the joint log-likelihood $\ell_P(\boldsymbol{\psi}) = \ell(\boldsymbol{\psi}, \hat{\boldsymbol{\lambda}}_\psi)$. A profile likelihood is thus always trivial to obtain, provided that there is an analytical expression for $\hat{\boldsymbol{\lambda}}_\psi$. However it will be subject to the same biases as applying maximum likelihood on the full parameter vector $\boldsymbol{\theta}$. Corrections to the profile likelihood which attempt to ameliorate this bias include modified profile likelihood (Barndorff-Neilsen, 1983), conditional profile likelihood (Cox & Reid, 1987) and adjusted profile likelihood (McCullagh &

Tibshirani, 1990).

1.1.3.4 Quasi-likelihood

The use of quasi-likelihood requires only second order distributional assumptions, rather than a full distributional assumption for the response. Like the exponential family of distributions (section 1.1.2), it is assumed that the variance is a function of the mean, that is $\mathrm{E}(y_i) = \mu_i$, and $\mathrm{var}(y_i) = a_i \phi v(\mu_i)$, $i = 1 \ldots n$, where, as before, $v(\mu_i)$ is the variance function, ϕ is a known dispersion parameter and a_i are known constants.

As noted at the beginning of section 1.1.3, a log-likelihood ℓ is characterised by score equations which have expectation zero and variance equal to the information. A "quasi-score" function which possesses these two attributes can be formed using only second order assumptions. Its integral is the quasi-likelihood function,

$$Q(\boldsymbol{\mu}; \boldsymbol{y}, \phi) = \sum_{i=1}^{n} \int_{y_i}^{\mu_i} \frac{(y_i - t)}{a_i \phi v(t)} dt,$$

with associated quasi-score equations

$$U(\tau_k) = \frac{\partial}{\partial \tau_k} Q(\boldsymbol{\mu}; \boldsymbol{y}, \phi) = \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \tau_k} \frac{(y_i - \mu_i)}{a_i \phi v(\mu_i)}, \qquad k = 1 \ldots p. \tag{1.4}$$

As the quasi-score equations are generally non-linear in $\boldsymbol{\tau}$, an iterative solution, such as Fisher scoring (section 1.1.3.1), is required.

To illustrate the formation of a quasi-likelihood function, the second order assumptions for a Poisson distribution are used, where $v(\mu) = \mu$ and $a_i = \phi = 1$. The quasi-likelihood function with these second order assumptions is

$$Q(\boldsymbol{\mu}; \boldsymbol{y}, \phi) = \sum_{i} \int_{y_i}^{\mu_i} \frac{(y_i - t)}{t} dt = \sum_{i} \left[y_i \left\{ \log(\mu_i) - \log(y_i) \right\} - (\mu_i - y_i) \right],$$

with corresponding PDF proportional to that arising from a Poisson distribution,

$$e^{Q(\boldsymbol{\mu})} = \prod_i \frac{\mu_i^{y_i} e^{-\mu_i}}{e^{-y_i} y_i^{y_i}} \propto \prod_i \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}.$$

Of all linear estimating equations of the form $\boldsymbol{H}\left[\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\tau})\right] = \boldsymbol{0}$, where $\boldsymbol{H}$ is a $p \times n$ matrix, it can be shown that the quasi-score equation (1.4), where $\boldsymbol{H} = \boldsymbol{D}^T \boldsymbol{V}^{-1}$ and $\boldsymbol{D} = \partial\boldsymbol{\mu}/\partial\boldsymbol{\tau}^T$ and $\boldsymbol{V} = \text{diag}\left\{v\left(\mu_i\right)\right\}$, is the optimal estimating equation, in the sense of achieving estimators with minimum variance (Godambe & Heyde, 1987).

Extended quasi-likelihood: Where ϕ is unknown, Nelder & Pregibon (1987) proposed an extended quasi-likelihood function for estimating both $\boldsymbol{\tau}$ and ϕ,

$$Q^+(\boldsymbol{\mu}, \phi; \boldsymbol{y}) = Q(\boldsymbol{\mu}; \boldsymbol{y}, \phi) - \sum_i \log\left\{2\pi a_i \phi v(y_i)\right\}/2.$$

The use of the extended quasi-likelihood Q^+ gives the same score equations for $\boldsymbol{\tau}$ as ordinary quasi-likelihood, since $\partial Q^+/\partial\boldsymbol{\mu} = \partial Q/\partial\boldsymbol{\mu}$. For the Poisson example, the extended quasi-likelihood function is

$$Q^+(\boldsymbol{\mu}, \phi, \boldsymbol{y}) = \sum_i \left[y_i\left\{\log(\mu_i) - \log(y_i)\right\} - (\mu_i - y_i)\right] - \sum \log(2\pi y_i)/2.$$

The corresponding PDF is again proportional to that arising from the Poisson distribution,

$$e^{Q^+} = \prod_i \frac{\mu_i^{y_i} e^{-\mu_i}}{e^{-y_i} y_i^{y_i} \sqrt{2\pi y_i}} \propto \prod_i \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}.$$

As shown in this example, the extended quasi-likelihood is equal to the likelihood of the corresponding exponential family, but with any factorial terms replaced by a Stirling approximation,

$$k! \doteq (2\pi k)^{1/2} k^k e^{-k}.$$

The extended quasi-likelihood for dispersion modelling: Nelder & Pregibon (1987) also used the extended quasi-likelihood for dispersion modelling, in two ways. Firstly, the variance function $v(\mu)$ may involve unknown parameters $\boldsymbol{\xi}$, such as a

member of the Tweedie family (Tweedie, 1984), where $v(\mu) = \mu^{\xi}$. Secondly, the dispersion parameter ϕ may vary across observations so that, if $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)^T$, then $g_\phi(\phi_i) = \boldsymbol{v}_i^T \boldsymbol{\xi}$, where $\boldsymbol{v}_i$ is a vector of regression variables for the ith observation and g_ϕ is a link function, usually either the identity or logarithmic link function. In both cases, extended quasi-likelihood can be used to estimate the parameters $\boldsymbol{\xi}$. In each case, estimation would proceed by alternatively solving for $\boldsymbol{\tau}$, the regression coefficients parametrizing the mean $\boldsymbol{\mu}$, and $\boldsymbol{\xi}$, the regression coefficients for the dispersion parameter $\boldsymbol{\phi}$, as in, for instance, Smyth (1989).

1.2 Linear mixed models

The linear model (section 1.1.1) has one source of variation, the residual error $\boldsymbol{e}$. However, many data analysis problems required models with more than one source of variation. In the analysis of designed experiments, factors can be randomised at different levels of aggregation or strata, owing to physical or logistical constraints, such as in a split plot design. In survey data, there are often clustering effects in the data where individual observations can not be assumed to be independent, for instance, in a household survey where people within the same household are more likely to provide similar responses. For some types of data, standard ANOVA techniques may be applied for each stratum in turn. However, for many types of data, such as where there are missing values, more advanced approaches are required. The use of a linear mixed model (LMM) is one such approach, and this is introduced along with classical likelihood techniques.

1.2.1 Specification

To illustrate a simple LMM, the example used for the normal linear model of section 1.1.1 is extended so that the experimental units are assumed to be in r blocks, with each block comprising $t = n/r$ experimental units. It is assumed that the treatments have been randomised to experimental units according to a randomised complete

block design (RCBD), so that each treatment appears once in each block. The linear mixed model is

$$y_i = \tau_0 + \tau_{j(i)} + u_{k(i)} + x_i\tau_{(t+1)} + e_i,$$

where $u_{k(i)}$ is the effect of the block in which the ith experimental unit resides. The block effects are assumed to be normally distributed, $u_k \sim N(0, \sigma_b^2)$, $k = 1 \ldots r$, to allow for the randomisation of the design, as outlined in Nelder (1965a,b). The parameters τ_0 and $\tau_{j(i)}$, and random errors $e_i \sim N(0, \sigma^2)$, are as for the normal linear model.

The general formulation of the LMM for data $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e},$$

where $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the design matrices for the fixed and random components and are of dimension $n \times p$ and $n \times b$ respectively, with associated vectors of coefficients $\boldsymbol{\tau}$ and $\boldsymbol{u}$ respectively, where $\boldsymbol{u} \sim N\left(\boldsymbol{0}, \sigma^2\boldsymbol{G}(\boldsymbol{\gamma})\right)$ and $\boldsymbol{e} \sim N\left(\boldsymbol{0}, \sigma^2\boldsymbol{R}(\boldsymbol{\phi})\right)$, and $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$ are vectors of variance parameters of length q and s respectively. Let $\boldsymbol{\kappa} = (\sigma^2, \boldsymbol{\gamma}^T, \boldsymbol{\phi}^T)^T$.

The LMM can be defined in two stages:

- conditional on the random effects $\boldsymbol{u}$, the data $\boldsymbol{y}$ is normally distributed with mean $\boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}\boldsymbol{u}$ and variance $\sigma^2\boldsymbol{R}(\boldsymbol{\phi})$;
- it is further assumed that $\boldsymbol{u}$ is normally distributed with mean $\boldsymbol{0}$ and covariance matrix $\sigma^2\boldsymbol{G}$, where $\boldsymbol{G}$ is parametrized by variance parameters $\boldsymbol{\gamma}$.

For the example above, assume that the units are ordered by block within treatment. Then $\boldsymbol{X}$ and $\boldsymbol{\tau}$ are as given in section 1.1.1. The matrix $\boldsymbol{Z}$ and vector $\boldsymbol{u}$ are

$$\boldsymbol{Z} = \begin{pmatrix} \boldsymbol{I}_t \\ \boldsymbol{I}_t \\ \vdots \\ \boldsymbol{I}_t \end{pmatrix} \text{ and } \boldsymbol{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_r \end{pmatrix},$$

where $\boldsymbol{I}_t$ is the $t \times t$ identity matrix.

1.2.2 Estimation and Prediction

1.2.2.1 ML and Residual ML (REML) for LMMs

In order to make formal inference for the linear mixed model of section 1.2, the likelihood needs to be evaluated. The LMM specification involves two distributions, the distribution of the data given the random effects, $\boldsymbol{y}|\boldsymbol{u} \sim N\left(\boldsymbol{X\tau} + \boldsymbol{Zu}, \sigma^2\boldsymbol{R}(\boldsymbol{\phi})\right)$, and the distribution of the random effects, $\boldsymbol{u} \sim N\left(\boldsymbol{0}, \sigma^2\boldsymbol{G}(\boldsymbol{\gamma})\right)$. Denote their corresponding PDFs $f_{Y|U}$ and f_U respectively. The joint PDF of the data and random effects is simply $f_{Y,U} = f_{Y|U}f_U$. However, $f_{Y,U}$ is not a likelihood, since it involves the unknown random effects $\boldsymbol{u}$. The random effects need to be integrated out, to give the marginal, or unconditional, PDF f_Y of the data $\boldsymbol{y}$:

$$\begin{aligned} f_Y &= \int f_{Y|U} f_U d\boldsymbol{u} \\ &= \int \left\{ (2\pi)^{-n/2} \left|\sigma^2\boldsymbol{R}\right|^{-1/2} \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X\tau}-\boldsymbol{Zu})^T\boldsymbol{R}^{-1}(\boldsymbol{y}-\boldsymbol{X\tau}-\boldsymbol{Zu})\right] \right. \\ &\quad \left. \left|\sigma^2\boldsymbol{G}\right|^{-1/2} \exp\left[-\frac{1}{2\sigma^2}\boldsymbol{u}^T\boldsymbol{G}^{-1}\boldsymbol{u}\right] \right\} d\boldsymbol{u} \\ &= (2\pi)^{-n/2}|\boldsymbol{V}|^{-1/2} \exp\left[-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X\tau})^T\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{X\tau})\right], \end{aligned} \tag{1.5}$$

where $\text{var}(\boldsymbol{y}) = \boldsymbol{V} = \sigma^2\left(\boldsymbol{R} + \boldsymbol{ZGZ}^T\right)$ is the marginal variance of $\boldsymbol{y}$. Alternatively, since the model is linear, the marginal expectation and variance can be easily determined without integration, viz.

$$\begin{aligned} \text{E}(\boldsymbol{y}) &= \text{E}\left\{\text{E}(\boldsymbol{y}|\boldsymbol{u})\right\} = \text{E}\left\{\boldsymbol{X\tau} + \boldsymbol{Zu}\right\} = \boldsymbol{X\tau} \\ \text{var}(\boldsymbol{y}) &= \text{var}\left\{E(\boldsymbol{y}|\boldsymbol{u})\right\} + \text{E}\left\{\text{Var}(\boldsymbol{y}|\boldsymbol{u})\right\} = \sigma^2\boldsymbol{ZGZ}^T + \sigma^2\boldsymbol{R}. \end{aligned}$$

Using standard statistical theory concerning mixtures of normal distributions, it is also clear that the marginal distribution should be a normal distribution. Note also that a linear mixed model can be specified using a marginal formulation as $\boldsymbol{y} = \boldsymbol{X\tau}+\boldsymbol{e}$ where $\boldsymbol{e} \sim N\left(\boldsymbol{0}, \boldsymbol{V}(\boldsymbol{\kappa})\right)$. This marginal formulation is widely used in many references

on LMMs.

The log-likelihood is simply $\ell(\boldsymbol{\tau}, \boldsymbol{\kappa}; \boldsymbol{y}) = \log f_Y$. Assuming that $\boldsymbol{X}$ is full rank, the ML estimator for $\boldsymbol{\tau}$, given $\boldsymbol{\kappa}$, is

$$\hat{\boldsymbol{\tau}}_{ML} = (\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{y},$$

with estimated variance $(\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X})^{-1}$. The ML estimator for a variance parameter κ_i, $\hat{\kappa}_{i,ML}$, is well-known to be downwardly biased, that is $\mathrm{E}(\hat{\kappa}_{i,ML}) < \kappa_i$, since it ignores the loss of degrees of freedom in estimating $\boldsymbol{\tau}$. This is an example of the failure of ML estimation in the presence of nuisance effects (section 1.1.3.3). REML (restricted maximum likelihood), first proposed by Patterson & Thompson (1971), is a likelihood for estimating the variance components $\boldsymbol{\kappa}$, formed from the probability distribution of the tranformed data, $\boldsymbol{T}(\boldsymbol{y}) = \boldsymbol{L}^T \boldsymbol{y}$, where $\boldsymbol{L}$ is a matrix such that $\boldsymbol{L}^T \boldsymbol{X} = \boldsymbol{0}$. REML can be considered as an example of a marginal likelihood, in the sense of section 1.1.3.3, since the distribution of $\boldsymbol{T}(\boldsymbol{y})$ is a function of $\boldsymbol{\kappa}$ alone, and does not involve $\boldsymbol{\tau}$. The REML score equations for the variance components are

$$\begin{aligned} \partial\ell/\partial\sigma^2 = \partial\ell/\partial\kappa_1 &= -\frac{1}{2}((n-p)/\sigma^2 - \boldsymbol{y}^T \boldsymbol{P} \boldsymbol{y}/\sigma^4) = 0, \\ \partial\ell/\partial\kappa_i &= -\frac{1}{2}[\mathrm{tr}(\boldsymbol{P}\boldsymbol{V}_i) - \boldsymbol{y}^T \boldsymbol{P} \boldsymbol{V}_i \boldsymbol{P} \boldsymbol{y}/\sigma^2] = 0, \quad i > 1 \end{aligned}$$

where $\boldsymbol{P} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}$, and $\boldsymbol{V}_i = \partial\boldsymbol{V}/\partial\kappa_i$. The ML score equations are similar, but with $\boldsymbol{P}$ replaced by $\boldsymbol{V}^{-1}$. Note that, except for $\partial\ell/\partial\sigma^2 = 0$, these equations are non-linear for κ_i. In general, no closed form solution for $\boldsymbol{\kappa}$ exists, and therefore an iterative solution is required.

Note that, in parallel to the development of LMM theory, the iterative generalized least squares (IGLS) (Goldstein, 1986) has been developed for so-called "multilevel models". The multilevel model formulation can be shown to be equivalent to the LMM formulation above. The restricted IGLS algorithm (RIGLS) of Goldstein (1989) has been shown to be equivalent to solving a REML likelihood for the variance parameters.

1.2.2.2 Best Linear Unbiased Predictors (BLUPs)

Estimation of the random effects $\boldsymbol{u}$ is referred to as "prediction", since $\boldsymbol{u}$ is a vector of random variables. It is assumed that REML estimation is used for the variance parameters, and that $\boldsymbol{y}_2 = \boldsymbol{L}^T\boldsymbol{y}$ where $\boldsymbol{L}$ is defined so that $\boldsymbol{L}^T\boldsymbol{X} = \boldsymbol{0}$. The best linear unbiased predictor (BLUP) of $\boldsymbol{u}$ is the mean of the posterior distribution of $\boldsymbol{u}$ given $\boldsymbol{y}_2$, $\tilde{\boldsymbol{u}} = E(\boldsymbol{u}|\boldsymbol{y}_2)$, where "best" is defined in the sense of being the predictor with lowest mean squared error. Note that this results holds for a more general mixed model than the normal linear mixed model. (In the case of ML estimation, $\tilde{\boldsymbol{u}} = E(\boldsymbol{u}|\boldsymbol{y})$ instead.) The BLUP is unbiased in the sense that $E(\tilde{\boldsymbol{u}}) = E(\boldsymbol{u}) = \boldsymbol{0}$, but not in the usual sense, since $E(\tilde{\boldsymbol{u}}) \neq \boldsymbol{u}$. Note that the expression for the BLUP implies a regression of the random effects $\boldsymbol{u}$ upon $\boldsymbol{y}_2$ (or $\boldsymbol{y}$ in the case of ML estimation).

For the normal linear mixed model, and for given $\boldsymbol{\kappa}$, the BLUP is

$$\tilde{\boldsymbol{u}} = E(\boldsymbol{u}|\boldsymbol{y}) = \left(\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\tau}}_{ML}),$$

where $\hat{\boldsymbol{\tau}}_{ML}$ was defined in the previous section. Henderson, Kempthorne, Searle & Von Krisig (1959) showed that differentiating the joint "log-likelihood" $\log f_{Y,U}$ with respect to $\boldsymbol{\tau}$ and $\boldsymbol{u}$ leads to the *mixed model equations,*

$$\begin{bmatrix} \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1} \end{bmatrix} \begin{pmatrix} \boldsymbol{\tau} \\ \boldsymbol{u} \end{pmatrix} = \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{y} \end{bmatrix}, \tag{1.6}$$

whose solution is $(\hat{\boldsymbol{\tau}}_{ML}^T, \tilde{\boldsymbol{u}}^T)^T$, thus avoiding the need to invert the large matrix $\boldsymbol{V}$. Harville (1977) showed that the solutions to the mixed model equations can also be used to provide ML or REML estimates of $\boldsymbol{\kappa}$, and suggested an iterative solution for the mixed model which alternates between solving for $\boldsymbol{\kappa}$ and solving for $\boldsymbol{\tau}$ and $\boldsymbol{u}$. Fisher scoring is preferable to Harville's method for maximising the (REML) log-likelihood with respect to the variance parameters $\boldsymbol{\kappa}$, since Harville's method is essentially an application of the EM algorithm, which is known to have slow convergence properties (section 1.1.3.1). More recently, Gilmour, Thompson & Cullis

(1995) proposed the "average information" REML algorithm (AI-Reml), which further improves the calculation of the information matrices involved in Fisher scoring by avoiding the evaluation of traces of large matrices.

1.2.3 Usefulness of the linear mixed model

As well as being useful for models where there are multiple sources of variation, the LMM has proved useful in other ways. It is sometimes desirable to fit some effects as random rather than fixed, such as in early generational plant variety trials with little replication, where variety effects are often fitted as random. The reason for doing so is to maximise the correlation between the estimated or predicted variety means and their true values, since the objective of early generational variety trials is varietal selection (Gilmour *et al.*, 1997). Random effect modelling is also useful for small area estimation to improve robustness (Ghosh & Rao, 1994). Linear mixed models specified with general covariance structures also allow the fitting of spatial and/or temporal correlation patterns. Finally, the linear mixed model also provides a way of fitting cubic smoothing splines, where the smoothing parameter is "automatically" determined by the variance component estimates (Verbyla, Cullis, Kenward & Welham, 1999).

1.3 Further issues

1.3.1 Bayesian estimation

In this section, Bayesian inference is discussed with reference to the linear mixed model of section 1.2.

In order to implement a Bayesian approach, prior distributions are required for all parameters in the model, since, under the Bayesian paradigm, all parameters are treated as random variables rather than fixed unknowns. For the random effects $\boldsymbol{u}$, the distributional assumption $\boldsymbol{u} \sim \boldsymbol{N}\left(\boldsymbol{0}, \sigma^2 \boldsymbol{G}(\boldsymbol{\gamma})\right)$ provides an informative prior distribution for $\boldsymbol{u}$. For the fixed effects $\boldsymbol{\tau}$, an uninformative prior is normally used,

such as $\boldsymbol{\tau} \sim N\left(\boldsymbol{\tau}_0, \sigma_\tau^2 \boldsymbol{I}\right)$, where $\boldsymbol{\tau}_0 = 0$ and σ_τ^2 is suitably large. It should be noted here that there is less distinction between fixed and random effects under a Bayesian framework,. The *hyper-parameters* of the model are those that delineate these two prior distributions, and comprise the variance parameters $\boldsymbol{\kappa}$, as well as τ_0 and σ_τ^2, the prior mean and variance of $\boldsymbol{\tau}$. Hyper-parameters may either take pre-specified values, or alternatively, be assigned a prior distribution which is delineated by further hyper-parameters.

Inference for one or more parameters in the model, say $\boldsymbol{\psi}$, is based on calculating its posterior PDF given the data $\boldsymbol{y}$,

$$f(\boldsymbol{\psi}|\boldsymbol{y}) = \frac{f(\boldsymbol{\psi}, \boldsymbol{y})}{f(\boldsymbol{y})} = \frac{\int f(\boldsymbol{y}|\boldsymbol{\psi}, \boldsymbol{\lambda}) f(\boldsymbol{\psi}, \boldsymbol{\lambda}) d\boldsymbol{\lambda}}{\int f(\boldsymbol{y}|\boldsymbol{\psi}, \boldsymbol{\lambda}) f(\boldsymbol{\psi}, \boldsymbol{\lambda}) d\boldsymbol{\lambda} d\boldsymbol{\psi}}, \tag{1.7}$$

where $\boldsymbol{\lambda}$ represents the remaining parameters in the model.

To compare classical likelihood and Bayesian inference in linear mixed models, it is useful to consider estimation of the variance components, $\boldsymbol{\kappa}$. It can be shown that ML estimation of the variance components corresponds to using a Bayesian framework with a point-mass prior for $\boldsymbol{\tau}$ at $\boldsymbol{\tau}_0$ where $\sigma_\tau^2 = 0$. Restricted maximum likelihood corresponds to using a Bayesian framework with a uniform prior for $\boldsymbol{\tau}$ over $(-\infty, \infty)$. See Searle, Casella & McCulloch (1992, Chapter 9) for more discussion.

1.3.1.1 Empirical Bayes estimation

Empirical Bayes estimation, in the context of a mixed linear model, refers to Bayesian estimation of $\boldsymbol{\tau}$ and $\boldsymbol{u}$ after calculation of a ML (or REML) estimate of the hyper-parameters $\boldsymbol{\kappa}$, and is very similar to BLUP estimation (section 1.2.2.2). See Searle *et al.* (1992, Chapter 9) for more detail.

1.3.1.2 Need for Monte Carlo approaches

The integrals required to evaluate the posterior PDF in (1.7) may not be analytically tractable. Markov Chain Monte Carlo (MCMC) techniques are Monte Carlo sam-

pling techniques which provide an indirect means of obtaining this posterior PDF. The Gibbs sampler is the most widely used MCMC technique. At each iteration of the Gibbs sampler, a simulated value of each parameter is drawn from its conditional PDF given the data and all the other parameters in the model. In the case of a linear mixed model, one may simulate, for the kth sample, from $f(\boldsymbol{u}|\boldsymbol{y}, \boldsymbol{\tau}^{(k-1)}, \boldsymbol{\kappa}^{(k-1)})$, $f(\boldsymbol{\tau}|\boldsymbol{y}, \boldsymbol{u}^{(k)}, \boldsymbol{\kappa}^{(k-1)})$ and $f(\boldsymbol{\kappa}|\boldsymbol{y}, \boldsymbol{u}^{(k)}, \boldsymbol{\tau}^{(k)})$ in turn. It can be shown that the conditional PDFs converge to the required posterior PDFs (e.g. Gelfand & Smith, 1990).

The conditional PDFs required for implementing the Gibbs sampler can be determined readily from the full joint PDF. In the case of just two parameter vectors $\boldsymbol{\psi}$ and $\boldsymbol{\lambda}$, the required conditional PDF for sampling $\boldsymbol{\psi}$ is $f(\boldsymbol{\psi}|\boldsymbol{\lambda}, \boldsymbol{y}) = f(\boldsymbol{\psi}, \boldsymbol{\lambda}, \boldsymbol{y})/f(\boldsymbol{\lambda}, \boldsymbol{y})$. Therefore, the conditional distribution for sampling $\boldsymbol{\psi}$ can usually be inferred solely from the numerator of this expression $f(\boldsymbol{\psi}, \boldsymbol{\lambda}, \boldsymbol{y})$, since the denominator is independent of $\boldsymbol{\psi}$, and thus can be regarded as a nuisance scaling factor.

Both the number of samples drawn during the initial "burn-in" period and drawn after burn-in, the latter which is also often called the length of the chain, need to be sufficiently large, to ensure that the conditional PDF has converged to a stable distribution. Both graphical and numerical approaches have been proposed to test for convergence of the chain, as reviewed in Mengersen, Robert & Guihenneuc-Jouyaux (1998), for instance.

One of the widely-touted advantages of using a Bayesian approach for linear mixed models, and mixed models in general, is that it allows for the fact that the variance parameters $\boldsymbol{\kappa}$ have also been estimated from the same data when making inference for the fixed and random effects $\boldsymbol{\tau}$ and $\boldsymbol{u}$. For instance, in the classical approach, $\boldsymbol{\tau}$ is estimated with the assumption that $\boldsymbol{\kappa}$ is known, $\hat{\tau} = \left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}$, and so $\text{var}(\hat{\boldsymbol{\tau}}) = \left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}$. In practice, however, $\boldsymbol{\kappa}$ is also estimated from the data, and $\boldsymbol{\tau}$ is estimated by substituting $\hat{\boldsymbol{\kappa}}$ in place of $\boldsymbol{\kappa}$. If $\hat{\boldsymbol{V}} = \boldsymbol{V}(\hat{\boldsymbol{\kappa}})$ then $\hat{\tau} = \left(\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{y}$ with estimated variance $\left(\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}\right)^{-1}$. This estimated

variance is too small, since the true variance of $\hat{\boldsymbol{\tau}}$ is

$$\text{var}(\hat{\boldsymbol{\tau}}) = \text{E}\left[\text{var}(\hat{\boldsymbol{\tau}}|\hat{\boldsymbol{\kappa}})\right] + \text{var}\left[\text{E}(\hat{\boldsymbol{\tau}}|\hat{\boldsymbol{\kappa}})\right] \approx \left(\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}\right)^{-1} + \text{var}\left[\text{E}(\hat{\boldsymbol{\tau}}|\hat{\boldsymbol{\kappa}})\right] \geq \left(\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}\right)^{-1}.$$

Therefore, standard errors and confidence intervals for $\boldsymbol{\tau}$ under the classical approach (and, using a similar argument, for $\boldsymbol{u}$ as well) can be too narrow. (Kenward & Roger (1997), for instance, describe adjustments in the classical approach to allow satisfactory inference concerning the fixed effects.)

1.3.2 Integral approximations

Integral approximations are useful for statistical inference when the integrals are analytically intractable. Two widely used integral approximations are outlined in this section, the Laplace approximation and Gauss-Hermite quadrature. Both are particularly useful for GLMMs to evaluate the likelihood (section 1.4.1). The objective is to obtain an analytical approximation for the integral, rather than just a number, which can then be differentiated in order to apply classical likelihood techniques.

1.3.2.1 The Laplace approximation

This subsection outlines the "first order" Laplace approximation, which is often synonymous with "Laplace approximation". A univariate integral

$$\int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{\infty} e^{g(x)}dx$$

can be approximated using a second order Taylor series expansion of the log-integrand $g(x) = \log f(x)$ around its mode $\hat{x}$,

$$\begin{aligned}\int_{-\infty}^{\infty} e^{g(x)}dx &\approx \int_{-\infty}^{\infty} \exp\left\{g(\hat{x}) + g'(\hat{x})(x-\hat{x}) + g''(\hat{x})(x-\hat{x})^2/2\right\}dx \\ &= e^{g(\hat{x})}\int_{-\infty}^{\infty} e^{g''(\hat{x})(x-\hat{x})^2/2}dx = e^{g(\hat{x})}\sqrt{\frac{2\pi}{-g''(\hat{x})}},\end{aligned}$$

noting that $g'(\hat{x}) = 0$ since $\hat{x}$ is the mode. The resultant integrand is a Gaussian distribution with mean $\hat{x}$ and variance $-1/g''(\hat{x})$. The extension to multivariate integrals is straightforward, giving

$$\int_{-\infty}^{\infty} e^{g(\boldsymbol{x})} d(\boldsymbol{x}) = e^{g(\hat{\boldsymbol{x}})} (2\pi)^{p/2} \left| -g''(\hat{\boldsymbol{x}}) \right|^{-1/2},$$

where p is the dimension of the vector $\boldsymbol{x}$, and $g''(\hat{\boldsymbol{x}}) = \partial^2 g / \left(\partial \boldsymbol{x} \partial \boldsymbol{x}^T \right)$. Heuristically, the first order Laplace approximation involves replacing the integrand with a scaled normal distribution of the same mode and curvature at the mode, as shown in Figure 1.1.

Note that this approximation is generally referred to as a first order Laplace approximation, despite involving a second order Taylor series expansion to approximate $g(x)$. The reason why it is referred to as a "first order" approximation is that the error of the approximation is $O(n^{-1})$ when $g(\boldsymbol{x})$ is the sum of n identical and independent components, such as in many simple statistical applications. (Some researchers, such as Raudenbush, Yang & Yosef (2000), still refer to the approximation according to the order of the Taylor series terms used.)

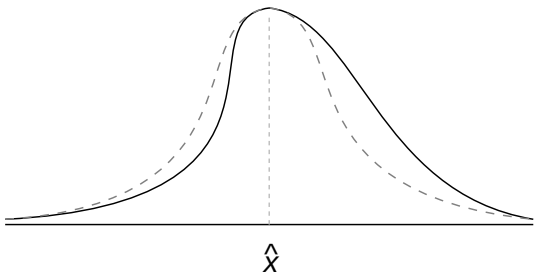


Figure 1.1: A heuristic explanation of the first order Laplace approximation. The integrand $f(x)$, displayed as a solid line, is approximated by a scaled normal distribution with the same mode, $\hat{x}$, and the same curvature at $\hat{x}$, shown as a dashed line.

1.3.2.2 Gauss-Hermite quadrature

Gauss-Hermite quadrature (GHQ) is a numerical integration technique where an integral is approximated as a weighted sum of integrand calculations, similar to the

more familiar Simpson and trapezoidal rules of integration. An m-point GHQ approximation for a univariate integral is

$$\int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{\infty} g(x)\exp(-x^2)dx \simeq \sum_{i=1}^{m} w_i g(x_i),$$

where $g(x) = f(x)\exp(x^2)$. The nodes x_i are the roots of the mth order Hermite polynomial, and the w_i are corresponding weights, which can be found in standard references such as Abramowitz & Stegun (1972) (p 890), or can be calculated in statistical or mathematical software packages. For instance, in the `R` statistical package (R Development Core Team, 2008), the weights and nodes can be obtained using the `gausquad` function from the `statmod` package.

One major disadvantage of GHQ is that the nodes x_i are distributed around 0, and not around the mode of $f(x)$, where most of the integral is concentrated. Adaptive GHQ (section 2.2) reparametrizes the integral so that the quadrature points are distributed around the mode of the integral. Adaptive GHQ is also a generalisation of the first order Laplace approximation, since one point adaptive GHQ corresponds to a first order Laplace approximation.

1.4 The generalized linear mixed model

1.4.1 Specification

As noted in section 1.2, linear mixed models present a useful extension to normal linear models for data where observations are non-independent, due to grouping or other sources of correlation. However, there is also a need to cater for additional sources of variation when the data are not normally distributed. A natural extension which combines the generalized linear model (GLM) and the linear mixed model (LMM) is the generalized linear mixed model (GLMM).

To continue the example used in section 1.2, assume that t treatments have been randomised to $r = n/t$ blocks according to an RCBD and that x_i is a covariate.

As in the example in section 1.1.2, assume that y_i represents count data to which a Poisson distribution applies, and that the treatment, covariate and block effects are multiplicative. A generalized linear mixed model for this data can be written

$$E(y_i) = \mu_i = \exp(\tau_0 + \tau_{j(i)} + u_{k(i)} + x_i\tau_{t+1}),$$

where τ_0 is the grand mean on the log scale, $\tau_{j(i)}$ and $u_{k(i)}$ are the multiplicative effects of the treatment and block for the ith experimental unit, and τ_{t+1} is the coefficient corresponding to the covariate x_i. The block effects are assumed to be normally distributed, $u_k \sim N(0, \sigma^2)$, $k = 1 \dots r$.

A GLMM can be specified as follows. As before, let y_i represent the ith observation, $i = 1 \dots n$, and vectors $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$, of length p and b respectively, represent covariates in the model corresponding to fixed and random coefficients $\boldsymbol{\tau}$ and $\boldsymbol{u}$ respectively. As with a linear mixed model, a GLMM can be specified in two stages:

- Conditional on the random effects $\boldsymbol{u}$, the model for $\boldsymbol{y} = (y_1 \dots y_n)^T$ corresponds to a generalized linear model.

  The distribution of y_i given $\boldsymbol{u}$ can be from either an exponential family or indirectly specified via a quasi-likelihood function. The first and second order moments of y_i, conditional on the random effects, are $\mu_i^u = \mathrm{E}(y_i|\boldsymbol{u})$ and $\mathrm{var}(y_i|\boldsymbol{u})$, where the "$u$" superscript indicates "conditional on $\boldsymbol{u}$" and

  $$\begin{aligned} g(\mu_i^u) &= \boldsymbol{x}_i^T\boldsymbol{\tau} + \boldsymbol{z}_i^T\boldsymbol{u}, \\ \mathrm{var}(y_i|\boldsymbol{u}) &= a_i\phi v(\mu_i^u), \end{aligned} \tag{1.8}$$

  where $v(\cdot)$ is a known function and ϕ and a_i represent known constants. The function $g(\cdot)$ is the link function, as for a GLM. The corresponding PDF is denoted $f_{Y|U}$.

- The random effects $\boldsymbol{u} = (u_1, \dots, u_b)^T$ have a multivariate normal distribution with mean $\boldsymbol{0}$ and covariance matrix $\boldsymbol{G} = \boldsymbol{G}(\boldsymbol{\gamma})$, where the vector $\boldsymbol{\gamma}$ contains

the variance components and is of length q. The corresponding PDF is denoted f_U.

1.4.2 The problem of likelihood inference for GLMMs

As for linear mixed models in section 1.2.2.1, a likelihood for the data is formed by integrating out the random effects $\boldsymbol{u}$ from the joint distribution of $\boldsymbol{y}$ and $\boldsymbol{u}$. The log-likelihood is

$$\begin{aligned} \ell(\boldsymbol{\tau}, \boldsymbol{\gamma}; \boldsymbol{y}) &= \log\left(\int f_{Y|U} f_U d\boldsymbol{u}\right) \\ &\propto \log\left(|\boldsymbol{G}|^{-1/2} \int \exp\left[-\frac{1}{2\phi}\sum_{i=1}^{n} d_i(y_i; \mu_i^u) - \frac{1}{2}\boldsymbol{u}^T \boldsymbol{G}^{-1}\boldsymbol{u}\right] d\boldsymbol{u}\right), \quad (1.9) \end{aligned}$$

where $d_i(y_i; \mu_i^u)$ is the contribution to the deviance measure of fit from the ith observation. For a distribution from an exponential family, the function $d_i(y; \mu)$ is

$$d_i(y; \mu) = \frac{(y\theta - b(\theta))}{a_i} + c(y, \phi),$$

whereas for the more general case of quasi-likelihood,

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

The problem with the expression for the likelihood in (1.9) is that it is generally not analytically tractable, and so the likelihood cannot be expressed in closed form. This creates a severe impediment for classical likelihood inference. This has motivated the application of approximate approaches based on the use of the Laplace approximation (section 1.3.2.1), such as penalized quasi-likelihood and Hierarchical GLM approaches, as well as other approaches, such as the use of Gauss-Hermite quadrature (section 1.3.2.2) and Bayesian methods implemented using MCMC approaches (section 1.3.1). These methods will be explored in the following chapters.

1.4.3 Alternatives to GLMMs

GLMMs are by no means the only approach to modelling correlation for non-normal data arising from a generalized linear model. The problem with obtaining an expression for the GLMM likelihood (section 1.4.2) motivates consideration of other alternative models. Two prominent alternatives will be discussed here, marginal models and the use of non-normal random effects.

1.4.3.1 Marginal models

A marginal model models the (unconditional) mean, $\mu_i = E(y_i)$, of the data as a function of the covariates $\boldsymbol{x}_i$ corresponding to fixed effects alone, that is,

$$g(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\tau}^*, \tag{1.10}$$

where $\boldsymbol{\tau}^*$ are the corresponding marginal fixed coefficients and $g(\cdot)$ is a link function, as before. This "marginal" model (1.10) for the unconditional mean μ_i can be contrasted with the GLMM, where the "conditional" model (1.8) for the conditional mean μ_i^u is a function of both the covariates corresponding to fixed and random effects, $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$, respectively. To account for correlation or clustering in the data, the marginal model assumes that the covariance of the data is unrestricted in form, that is $\text{cov}(\boldsymbol{y}) = \boldsymbol{V}$ where $\boldsymbol{V}$ is an arbitrary $n \times n$ matrix, whereas for a GLM the covariance matrix is restricted to be diagonal, corresponding to independent responses. A structured form of $\boldsymbol{V}$ is also possible, for instance, one that is based on a corresponding GLMM.

For an LMM, which, as noted above, is a special case of a GLMM, the marginal specification is equivalent to the conditional specification. That is, a conditional specification of an LMM is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}, \quad \boldsymbol{u} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{G}),\ \boldsymbol{e} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{R}),$$

with an equivalent marginal specification given by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau}^* + \boldsymbol{e}, \qquad \boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{V}),$$

where $\boldsymbol{V} = \sigma^2\boldsymbol{R} + \sigma^2\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T$. Here $\boldsymbol{\tau}^* = \boldsymbol{\tau}$, but this is not true for other GLMMs. For GLMMs in general, the marginal coefficients $\boldsymbol{\tau}^*$ reflect the observable attributes of the data, such as the "raw" means or averages. For this reason, proponents of the marginal approach argue that the marginal coefficients $\boldsymbol{\tau}^*$ are more readily interpretable than the conditional coefficients $\boldsymbol{\tau}$ (e.g. Carlin, Wolfe, Brown & Gelman, 2001).

Liang & Zeger (1986) outline the Generalized Estimating Equations (GEE) method for solving the marginal model, which Zeger, Liang & Albert (1988) also call the "population averaged" model. The main focus of the GEE approach is to estimate the regression parameters $\boldsymbol{\tau}^*$ — the covariances between observations are considered nuisance parameters. Proponents of the marginal approach also argue that the GEE approach for estimating the marginal model is more robust to misspecification of the covariance structure than using a GLMM (e.g. Neuhaus, Hauck & Kalbfleish, 1992*)*.

The primary problem with the marginal model is that it does not have a direct probabilistic interpretation like the conditional GLMM model. Lindsey & Lambert (1998) argue that marginal models can give highly misleading answers when the aim of the analysis is to establish causal relations, due to problems akin to Simpson's paradox or the ecological fallacy, and this argument is further reinforced in Lee & Nelder (2004). If the sample is not representative or random, the marginal quantities may not be generalizable, as pointed out in McCulloch & Searle (2001). Nevertheless, others such as Carlin *et al.* (2001) still prefer the marginal specification, arguing that the inherent distributional assumptions of the conditional GLMM may not be credible. For the purposes of agricultural and biological research, however, the conditional GLMM model is arguably more preferable, since it is easier to choose structured covariance matrices to match the design of the data. Note that these structured covariance matrices are obviously on the link scale, rather than on the original scale of the data.

Some GLMM proponents have extended the GLMM to incorporate a marginal component. That is, the model for the conditional mean is as specified in section 1.4.1, but, in addition, the conditional variance is no longer assumed to be diagonal, so that $\text{var}(\boldsymbol{y}|\boldsymbol{u}) = \boldsymbol{V} = \boldsymbol{V}_\mu^{1/2}\boldsymbol{R}_\phi\boldsymbol{V}_\mu^{1/2}$, say, where $\boldsymbol{V}_\mu = \text{diag}\{a_1 v(\mu_1), \ldots, a_n v(\mu_n)\}$ and $\boldsymbol{R}_\phi$ is a function of variance parameters $\boldsymbol{\phi}$ alone (e.g. Wolfinger & O'Connell, 1993). Along similar lines, Candy (1997) proposes the additive GLMM (AGLMM), where the random effects $\boldsymbol{u}$ are fitted on the scale of the response, rather than within the linear predictor, so that the conditional mean is defined as

$$\mu_i^u = h(\boldsymbol{x}_i^T\boldsymbol{\tau}) + \boldsymbol{z}_i^T\boldsymbol{u},$$

where $h = g^{-1}$ is the inverse link function. In this model, the marginal mean is just $\mu_i = h(\boldsymbol{x}_i^T\boldsymbol{\tau})$.

1.4.3.2 Non-normal random effects

Some researchers have questioned the assumption of normality for the random effects in the GLMM specification.

One alternative to the normality assumption is to allow the random effects to arise from any distribution in the exponential or quasi-likelihood families. Hierarchical GLMs (Lee & Nelder, 1996, 2001, 2006) are an extension of GLMMs which do this. In particular, a conjugate distribution could be employed. For instance, in the case of binomial and Poisson distributed data, the conjugate distributions for the random effects are the beta and gamma distributions respectively. The advantage of a conjugate distribution is that the integral required to evaluate the likelihood (equation 1.9) is analytically tractable. Another alternative is to make no distributional assumptions whatever, as in the non-parametric approach of Aitkin (1999).

The advantage of normally distributed random effects is that a fuller range of multivariate random effects models can be examined, such as those involving temporal or spatial correlation. The merits of using non-normally distributed over normally

distributed random effects have still not been fully assessed, and are outside the scope of this thesis. For many real-life datasets, where many random terms may be required but the main inferential focus concerns the fixed effects, the assumption of normality appears to be a reasonable one.

1.5 Objectives of this research

Classical statistical analysis of the GLMM has been hindered by the intractability of the likelihood expression. However, GLMMs present a general modelling approach for correlated non-normal data which cannot be accommodated by other approaches, such as marginal models or by using non-normal random effects. In particular, as indicated above, GLMMs may be more suitable for agricultural and biological data than marginal models, because of the explicit fitting of random terms to account for design strata and other known sources of variation. Therefore, methods to fit GLMMs which work around the intractability of the GLMM likelihood are still worthy of attention.

In this thesis, GLMM approaches are divided into two groups:

- Approximate likelihood approaches: These include penalized quasi-likelihood (PQL) and the methodology proposed by Lee and Nelder for Hierarchical GLMs (HGLMs, which are a broader class of models than GLMMs). Many approximate methods, including these two, are based on the Laplace approximation.
- Other approaches: This group includes numerical approaches, such as the use of Gauss-Hermite quadrature (GHQ), and stochastic approaches, such as Bayesian approaches implementing Markov Chain Monte Carlo (MCMC) techniques.

This research primarily focuses on the first group, the approximate likelihood approaches. As will be seen in Chapters 3 and 4, approximate likelihood approaches can suffer estimation bias problems with some GLMMs. However, the other approaches can suffer from either excessive computational requirements and/or lack of generalizability.

The rest of this thesis is arranged as follows. The second chapter reviews the major approaches for GLMMs. The third and fourth chapters examines two approximate likelihood approaches, penalized quasi-likelihood (PQL) and the HGLM methodology of Lee & Nelder (2001) respectively. The fifth chapter compares the performance of these two approximate likelihood approaches with the most prominent alternative approaches, GHQ and Bayesian approaches using MCMC techniques.

Chapter 2

Review of approaches to estimation for GLMMs

This chapter reviews the range of approaches to estimation available for generalized linear mixed models (GLMMs), including the approximate likelihood approaches and the other approaches as noted at the end of chapter 1. It also discusses some of the relative merits and disadvantages of each approach raised in the literature.

2.1 Approximate approaches (Laplace based)

This section outlines two approximate approaches for maximum likelihood estimation based on the Laplace approximation, penalized quasi-likelihood (PQL) and the hierarchical GLM approach of Lee & Nelder (2001). More detail on the latter is reserved for chapters 3 and 4. The GLMM specification is as in section 1.4.1, and is repeated here for convenience.

Let the data be denoted y_i, $i = 1 \dots n$ and let vectors $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ represent covariates in the model corresponding to fixed and random coefficients $\boldsymbol{\tau}$ and $\boldsymbol{u}$ of length p and b respectively. A GLMM is defined by two properties:

- The probability density function (PDF) $f_{Y|U}$ of the data y_i, $i = 1, \dots, n$, conditional on the random effects, corresponds to that of a distribution from an

exponential family or a quasi-likelihood distribution. The first two conditional moments are defined as $\mu_i^u = \mathrm{E}(y_i|\boldsymbol{u})$ and $\mathrm{var}(y_i|\boldsymbol{u})$, where

$$g(\mu_i^u) \quad = \quad \boldsymbol{x}_i^T\boldsymbol{\tau} + \boldsymbol{z}_i^T\boldsymbol{u}$$

and

$$\mathrm{var}(y_i|\boldsymbol{u}) = a_i\phi v(\mu_i^u).$$

The function $v(\cdot)$ is assumed known, and ϕ and a_i represent known constants. The function $g(\cdot)$ is the link function as for a GLM.

- The random effects $\boldsymbol{u} = (u_1, \ldots, u_b)^T$ have a multivariate normal PDF f_U with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{G} = \boldsymbol{G}(\boldsymbol{\gamma})$, where the elements of $\boldsymbol{\gamma}$ are called the variance components of length q.

In addition to this, let $\boldsymbol{X}$ denote the $n \times p$ fixed design matrix with ith row $\boldsymbol{x}_i^T$, and similarly let $\boldsymbol{Z}$ denote the $n \times b$ random design matrix with ith row $\boldsymbol{z}_i^T$.

2.1.1 Penalized quasi-likelihood

Penalized quasi-likelihood (PQL) is an iterative approach for solving a GLMM which is similar to the IRLS approach for solving a GLM (section 1.1.3.2). PQL can easily be implemented with repeated calls to a linear mixed model (LMM) package. At the kth iteration, the working variate $\boldsymbol{\psi}^{(k)} = (\psi_1^{(k)}, \ldots, \psi_n^{(k)})$ is generated, with elements

$$\psi_i^{(k)} = g\left(\mu_i^{u,(k-1)}\right) + \left(y_i - \mu_i^{u,(k-1)}\right) g'\left(\mu_i^{u,(k-1)}\right),$$

where $\mu_i^{u,(k-1)}$ is the conditional mean of y_i given estimates $\hat{\boldsymbol{\tau}}^{(k-1)}$ and $\tilde{\boldsymbol{u}}^{(k-1)}$ from the $(k-1)$th iteration. A weighted LMM is fitted to this working variable,

$$\psi_i^{(k)} = \boldsymbol{x}_i^T\boldsymbol{\tau} + \boldsymbol{z}_i^T\boldsymbol{u} + e_i^*, \tag{2.1}$$

with weights $w_i^{(k-1)} = \left\{\phi a_i v(\mu_i^{u,(k-1)})[g'(\mu_i^{u,(k-1)})]^2\right\}^{-1}$, implying that the errors e_i^* are independent with known variance, $\text{var}(e_i^*) = \left(w_i^{(k)}\right)^{-1}$. The fitting of this LMM generates updated estimates $\hat{\boldsymbol{\tau}}^{(k)}$ and $\tilde{\boldsymbol{u}}^{(k)}$, which are used to calculate the working variate $\boldsymbol{\psi}^{(k+1)} = \left(\psi_1^{(k+1)}, \ldots \psi_n^{(k+1)}\right)^T$ for the $(k+1)$th iteration. These two steps, the formation of the working variable and the fitting of a LMM, are repeated until convergence of the parameter estimates, or the deviance, is obtained. An analogous technique has been developed for normal non-linear mixed models (Lindstrom & Bates, 1990). As for a LMM, initial values for $\boldsymbol{\tau}$ and $\boldsymbol{u}$ can both be set to $\mathbf{0}$, and the initial values of the variance parameters $\boldsymbol{\gamma}$ can either be set to a small positive value (e.g. 0.1) or user-specified. It should be noted that the initial values of the conditional means μ_i are also required, but can be set in a similar fashion as for a GLM, using the data y_i or some minor modification of it – for instance, see McCullagh & Nelder (1989, p. 41).

The simplest derivation for the PQL technique is based on a first order Taylor series approximation (Goldstein, 1991, 1995; Goldstein & Rasbash, 1996; Wolfinger & O'Connell, 1993). If the GLMM is approximated as

$$y_i \approx \mu_i^u + e_i = h\left(\eta_i\right) + e_i, \qquad e_i \sim N\left(0, \phi a_i v(\mu_i)\right)$$

where $h = g^{-1}$ is the inverse link function, then an expansion around the current value of the linear predictor, $\eta_i^{(k-1)} = \boldsymbol{x}_i^T \boldsymbol{\tau}^{(k-1)} + \boldsymbol{z}_i^T \boldsymbol{u}^{(k-1)}$, gives

$$y_i \approx h\left(\eta_i^{(k-1)}\right) + h'\left(\eta_i^{(k-1)}\right)\left(\boldsymbol{x}_i^T \boldsymbol{\tau} + \boldsymbol{z}_i^T \boldsymbol{u} - \eta_i^{(k-1)}\right) + e_i. \tag{2.2}$$

Re-arranging this expression gives (2.1).

The standard implementation of PQL outlined above assumes that the dispersion parameter ϕ is equal to 1. However, ϕ can also be estimated from the data, as for the analysis of a GLM, to allow for under- or over-dispersion of the data not explained by the random effects u_i. Some implementations of PQL, such as that available in the `lmer` function of the `lme4` R package (Bates & Sarkar, 2006) and the SAS

`glimmix` macro (Wolfinger & O'Connell, 1993), estimate the dispersion component ϕ by default.

2.1.1.1 Theoretical development of PQL

Joint maximisation: Early proponents for PQL, previously referred to as "joint maximisation" (Harville & Mee, 1984), argued that estimates of $\boldsymbol{\tau}$ and $\boldsymbol{u}$, given variance parameters $\boldsymbol{\gamma}$, can be determined by maximising the joint likelihood $\ell_J = \log f_{Y,U}$. A Bayesian justification of this assertion is as follows. The best unbiased predictor of $\boldsymbol{u}$ is the posterior mean

$$E(\boldsymbol{u}|\boldsymbol{y}) = \int \boldsymbol{u} f_{U|Y}(\boldsymbol{u}|\boldsymbol{y}) d\boldsymbol{u} \propto \int \boldsymbol{u} f_{Y,U} d\boldsymbol{u}.$$

If $f_{U|Y}$ is approximately normal, $E(\boldsymbol{u}|\boldsymbol{y})$ can be approximated by the mode of $f_{Y,U}$, or, equivalently, of ℓ_J. Similarly, $E(\boldsymbol{\tau}|\boldsymbol{y})$ can be approximated with the mode of $f_{Y,U}$, or of ℓ_J, if a flat prior for $\boldsymbol{\tau}$ is used. These estimates of $\boldsymbol{\tau}$ and $\boldsymbol{u}$ were termed *maximum a posteriori* (MAP) estimates (Laird, 1978; Stiratelli, Laird & Ware, 1984; Harville & Mee, 1984; Schall, 1991).

It can readily be shown that maximising ℓ_J with respect to $\boldsymbol{\tau}$ and $\boldsymbol{u}$ yields the updating equations for $\boldsymbol{\tau}$ and $\boldsymbol{u}$ in the linear mixed model (2.1) above. Now, ignoring terms not involving $\boldsymbol{\tau}$ or $\boldsymbol{u}$,

$$\ell_J \propto -\frac{1}{2}\sum_{i=1}^{n} d_i(y_i; \mu_i^u) - \frac{1}{2}\boldsymbol{u}^T \boldsymbol{G}^{-1}\boldsymbol{u},$$

where

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

assuming that $a_i = 1$ and $\phi = 1$ for simplicity. Differentiation with respect to $\boldsymbol{\tau}$ and

$\boldsymbol{u}$ yields

$$\begin{pmatrix} \partial \ell_J/\partial \boldsymbol{\tau} \\ \partial \ell_J/\partial \boldsymbol{u} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{D}(\boldsymbol{y}-\boldsymbol{\mu}^u) \\ \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{D}(\boldsymbol{y}-\boldsymbol{\mu}^u) - \boldsymbol{G}^{-1}\boldsymbol{u} \end{pmatrix},$$

$$E\begin{pmatrix} \partial^2\ell_J/\left(\partial\boldsymbol{\tau}\partial\boldsymbol{\tau}^T\right) & \partial^2 l_J/\left(\partial\boldsymbol{\tau}\partial\boldsymbol{u}^T\right) \\ \partial^2\ell_J/\left(\partial\boldsymbol{u}\partial\boldsymbol{\tau}^T\right) & \partial^2\ell_J/\left(\partial\boldsymbol{u}\partial\boldsymbol{u}^T\right) \end{pmatrix} = \begin{pmatrix} -\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} & -\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{Z} \\ -\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{X} & -\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z}-\boldsymbol{G}^{-1} \end{pmatrix},$$

where $\boldsymbol{W} = \text{diag}\{v(\mu_i^u)[g'(\mu_i^u)]^2\}^{-1}$ contains the GLM weights, and $\boldsymbol{D} = \text{diag}\{\partial\eta_i^u/\partial\mu_i^u\}$. If the current estimates are denoted $\hat{\boldsymbol{\tau}}^{(k-1)}$ and $\tilde{\boldsymbol{u}}^{(k-1)}$, then re-arrangement of the Fisher scoring equations gives

$$\begin{pmatrix} \boldsymbol{X}^T\boldsymbol{W}^{(k-1)}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{W}^{(k-1)}\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{W}^{(k-1)}\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{W}^{(k-1)}\boldsymbol{Z}+\boldsymbol{G}^{-1} \end{pmatrix}\begin{pmatrix} \boldsymbol{\tau}-\hat{\boldsymbol{\tau}}^{(k-1)} \\ \boldsymbol{u}-\tilde{\boldsymbol{u}}^{(k-1)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}^T\boldsymbol{W}^{(k-1)}\boldsymbol{D}^{(k-1)}(\boldsymbol{y}-\boldsymbol{\mu}^{(k-1)}) \\ \boldsymbol{Z}^T\boldsymbol{W}^{(k-1)}\boldsymbol{D}^{(k-1)}(\boldsymbol{y}-\boldsymbol{\mu}^{(k-1)}) - \boldsymbol{G}^{-1}\tilde{\boldsymbol{u}}^{(k-1)} \end{pmatrix},$$

where the $(k-1)$ superscript indicates evaluation at $\hat{\boldsymbol{\tau}}^{(k-1)}$ and $\tilde{\boldsymbol{u}}^{(k-1)}$. Substituting

$$\boldsymbol{\psi}^{(k)} = \boldsymbol{X}\hat{\boldsymbol{\tau}}^{(k-1)} + \boldsymbol{Z}\tilde{\boldsymbol{u}}^{(k-1)} + \boldsymbol{D}^{(k-1)}(\boldsymbol{y}-\hat{\boldsymbol{\mu}}^{(k-1)})$$

gives

$$\begin{pmatrix} \boldsymbol{X}^T\boldsymbol{W}^{(k-1)}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{W}^{(k-1)}\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{W}^{(k-1)}\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{W}^{(k-1)}\boldsymbol{Z}+\boldsymbol{G}^{-1} \end{pmatrix}\begin{pmatrix} \hat{\boldsymbol{\tau}} \\ \tilde{\boldsymbol{u}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}^T\boldsymbol{W}^{(k-1)}\boldsymbol{\psi}^{(k)} \\ \boldsymbol{Z}^T\boldsymbol{W}^{(k-1)}\boldsymbol{\psi}^{(k)} \end{pmatrix}, \quad (2.3)$$

which are the mixed model equations (1.6) for the linear mixed model in (2.1).

Breslow and Clayton (1993) As noted previously, Breslow & Clayton (1993) first coined the term "penalized quasi-likelihood" (PQL), by which the approach is best-known today. The term "penalized quasi-likelihood" reflects the use of a quasi-likelihood distributional assumption for the "conditional" likelihood $\ell_c = \log f_{Y|U}$, but with "penalized" estimates of $\boldsymbol{u}$ through the addition of $\log f_U$ in the joint likelihood

ℓ_J.

Breslow & Clayton (1993) applied the first order Laplace approximation (section 1.3.2.1) to the integral expression for the log-likelihood (equation 1.9), written as

$$\ell(\boldsymbol{\tau}, \boldsymbol{\gamma}; \boldsymbol{y}) = \log \int e^{h(\boldsymbol{u})} d\boldsymbol{u},$$

where $h(\boldsymbol{u}) = \log f_{Y,U}$. This gives

$$\begin{aligned} \ell(\boldsymbol{\tau}, \boldsymbol{\gamma}; \boldsymbol{y}) &\approx h(\tilde{\boldsymbol{u}}_{\tau,\gamma}) - \frac{1}{2} \log \mid -h''(\tilde{\boldsymbol{u}}_{\tau,\gamma}) \mid \\ &\approx -\frac{1}{2\phi} \sum_{i=1}^{n} d_i(y_i, \tilde{\mu}_i^u) - \frac{1}{2} \log |\boldsymbol{G}| - \frac{1}{2} \tilde{\boldsymbol{u}}_{\tau,\gamma}^T \boldsymbol{G}^{-1} \tilde{\boldsymbol{u}}_{\tau,\gamma} \\ &\quad -\frac{1}{2} \log \left| \boldsymbol{Z}^T \tilde{\boldsymbol{W}} \boldsymbol{Z} + \boldsymbol{G}^{-1} \right|, \end{aligned} \tag{2.4}$$

where $h''(\boldsymbol{u})$ is the second derivative with respect to $\boldsymbol{u}$, $\tilde{\boldsymbol{u}}_{\tau,\gamma}$ is the mode of $h(\boldsymbol{u})$ and $\boldsymbol{W} = \text{diag}(w_i, \ldots, w_n)$, where $w_i = \left\{ \phi a_i v(\mu_i^u) [g'(\mu_i^u)]^2 \right\}^{-1}$. The use of the tilde notation in $\tilde{\mu}_i^u$ and $\tilde{\boldsymbol{W}}$ indicates evaluation at $\tilde{\boldsymbol{u}}_{\tau,\gamma}$. A term in $h''(\boldsymbol{u})$,

$$-\sum_{i=1}^{n} (y_i - \mu_i^u) \boldsymbol{z}_i \frac{\partial}{\partial \boldsymbol{u}} \left[\frac{1}{\phi a_i v(\mu_i^u) g'(\mu_i^u)} \right],$$

is ignored here as it is equal to zero for canonical links and has zero expectation for non-canonical links.

As noted in the previous section, the PQL estimating equations for $\boldsymbol{\tau}$ and $\boldsymbol{u}$ are based on maximising $\ell_J = \log f_{Y,U}$. To justify the use of ℓ_J for estimating $\boldsymbol{\tau}$ and $\boldsymbol{u}$, Breslow & Clayton (1993) argued that the last term in (2.4) can be ignored, since the GLM weights w_i should vary little with $\boldsymbol{\tau}$ and $\boldsymbol{u}$, and so $\ell(\boldsymbol{\tau}, \boldsymbol{\gamma}; \boldsymbol{y}) \approx h(\tilde{\boldsymbol{u}}_{\tau,\gamma})$.

To determine estimating equations for the variance parameters in $\boldsymbol{\gamma}$, they replaced the deviance $\sum_i d_i(y_i, \mu_i^u)$ with the Pearson chi-squared statistic, $\sum_i (y_i - \mu_i^u)^2 / a_i v(\mu_i^u)$, in (2.4). This yields, after simplication, a profile likelihood of $\boldsymbol{\gamma}$ which corresponds to the profile log-likelihood for $\boldsymbol{\gamma}$ in the linear mixed model (2.1),

$$\ell\left(\hat{\boldsymbol{\tau}}_\gamma, \boldsymbol{\gamma}; \boldsymbol{y}\right) = -\frac{1}{2} \log \left| \tilde{\boldsymbol{V}} \right| - \frac{1}{2} (\tilde{\boldsymbol{\psi}} - \boldsymbol{X} \hat{\boldsymbol{\tau}}_\gamma)^T \tilde{\boldsymbol{V}}^{-1} (\tilde{\boldsymbol{\psi}} - \boldsymbol{X} \hat{\boldsymbol{\tau}}_\gamma), \tag{2.5}$$

where $\boldsymbol{V} = \boldsymbol{W}^{-1} + \boldsymbol{ZGZ}^T$, $\boldsymbol{\tau}_\gamma$ solves $\partial h(\tilde{\boldsymbol{u}}_{\tau,\gamma})/\partial \boldsymbol{\tau} = \boldsymbol{0}$, and the tilde notation indicates evaluation at $\hat{\boldsymbol{\tau}}_\gamma$ and $\tilde{\boldsymbol{u}}_{\tau,\gamma}$. The working variate $\tilde{\boldsymbol{\psi}}$ consists of elements $\tilde{\psi}_i^{(k)} = g\left(\tilde{\mu}_i^u\right) + \left(y_i - \tilde{\mu}_i^u\right) g'\left(\tilde{\mu}_i^u\right)$, as before. A "REML" type correction, $-\frac{1}{2}\log\left|\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right|$, can then be added to this log-likelihood. Most implementations of PQL use this approximate REML correction.

Other justifications and comments There are numerous alternative arguments to Breslow & Clayton (1993) for PQL. McGilchrist (1994) advocated PQL by arguing that the conditional likelihood, $\ell_C = \log f_{Y|U}$, should have a well defined maximum, and hence can be approximated as a normal likelihood. Schall (1991) also endorsed the PQL approach, and so the approach is also sometimes called "Schall's approach" instead of PQL. Both Schall (1991) and McGilchrist (1994) argued that PQL is robust to the mis-specification of the distribution of the random effects, and can be applied even when the normality assumption is invalid. Engel & Keen (1994) called the approach "IRREML" (iterative reweighted REML), which corresponded to a Genstat macro of the same name. Wolfinger & O'Connell (1993) preferred "pseudo-likelihood" (PL), or "restricted pseudo-likelihood" (REPL) when the "REML" type adjustment is applied. "Pseudo-likelihood", in the sense of Carroll & Ruppert (1988), may be more appropriate than "penalized quasi-likelihood", since it reflects the implicit normality assumption on the working variate $\boldsymbol{\psi}$. The Wolfinger & O'Connell (1993) paper is associated with a SAS macro `glimmix`, which is one of the most widely used PQL implementations.

Others have been less forthcoming about the benefits of PQL. McCulloch & Searle (2001) (page 233) argued that, for the working variate

$$\psi_i = \boldsymbol{x}_i^T\boldsymbol{\tau} + \boldsymbol{z}_i^T\boldsymbol{u} + g'(\mu_i^u)(y_i - \mu_i^u),$$

the PQL approach assumes that

$$Var(\psi_i) = \boldsymbol{z}_i^T\boldsymbol{G}\boldsymbol{z}_i + g'(\mu_i^u)^2 V(\mu_i^u),$$

but this assumption ignores the fact that μ_i^u is a function of the random effects $\boldsymbol{u}$, not a constant. Engel & Keen (1994) suggested a modification to PQL to allow for the dependence of the GLM weights, $w_i = \{a_i v(\mu_i^u) g'(\mu_i^u)^2\}^{-1}$, on the estimates of the random effects $\tilde{\boldsymbol{u}}_{\tau,\gamma}$. They suggested using the expected weights $\mathrm{E}(w_i^{-1})$, instead of w_i^{-1}, where the expectation was calculated over the distribution of $\boldsymbol{u}$ using a bootstrap technique. However, simulation studies in Engel & Buist (1998) showed that this technique did not perform as well as anticipated in practice.

2.1.1.2 Estimating (and correcting) the bias

The penalized quasi-likelihood technique can suffer from large estimation biases for certain GLMMs, in particular for binary data with small cluster sizes, or more generally, where the number of observations per random effect is small. This has been documented in numerous studies, notably Rodriguez & Goldman (2001). This has motivated the exploration techniques to correct this bias, such as CPQL, PQL2 and the iterated bootstrap correction.

Corrected PQL – CPQL: Breslow & Lin (1995) and Lin & Breslow (1996) derived approximate expressions for the PQL biases by using a Taylor series approximation of the likelihood around $\boldsymbol{\gamma} = \mathbf{0}$. They consequently provided corrections to the estimates of $\boldsymbol{\tau}$ and $\boldsymbol{\gamma}$ in turn.. For simplicity, these correction are shown here in the case of a GLMM with one variance component, as in Breslow & Lin (1995). Let the data be y_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$ where $\sum_i n_i = n$. The model for the conditional mean $\mu_{ij}^u = E(y_{ij}|u_i)$ is assumed to be $g(\mu_{ij}^u) = \boldsymbol{x}_{ij}^T\boldsymbol{\tau} + u_i$, where $u_i \sim N(0, \gamma_1)$ and $\mathrm{var}(y_{ij}|u_i) = a_{ij}\phi v(\mu_{ij}^u)$. The corrections for $\boldsymbol{\tau}$ and $\boldsymbol{\gamma}$ are as follows.

1. A correction to $\boldsymbol{\tau}$: They provided a correction for the estimation of $\boldsymbol{\tau}$ using ℓ_J rather than the true likelihood ℓ,

$$\hat{\boldsymbol{\tau}}_{CP} = \hat{\boldsymbol{\tau}}_P - \frac{\gamma_1}{2}(\boldsymbol{X}^T\boldsymbol{W}^0\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{t}^0,$$

where $\hat{\boldsymbol{\tau}}_P$ is the PQL estimate, $\mu_{ij}^0 = g^{-1}(\boldsymbol{x}_{ij}^T \hat{\boldsymbol{\tau}}_P)$, $\boldsymbol{W}^0 = \text{diag}\left\{a_{ij} v\left(\mu_{ij}^0\right)/\phi\right\}$, and $\boldsymbol{t}^{(0)}$ is an $n \times 1$ vector with elements $a_{ij} v'(\mu_{ij}^o) v(\mu_{ij}^o)/\phi$. (The "0" superscript indicates evaluation at $\boldsymbol{\tau} = \hat{\boldsymbol{\tau}}_P$ and $u_i = 0$, $i = 1, \ldots, m$.) It can be seen that the magnitude of the correction term is proportional to the size of the variance component γ_1. This correction is based on a "first order" Taylor series around $\gamma_1 = 0$, but Lin & Breslow (1996) also provided a correction based on a second order Taylor series.

2. A correction to $\boldsymbol{\gamma}$: Breslow & Lin provided a correction for estimating $\boldsymbol{\gamma}$, to account for the usage of the approximate profile likelihood (2.5) rather than the true likelihood ℓ,

$$\hat{\gamma}_{1,CP} = \frac{D}{B - C} \hat{\gamma}_{1,P},$$

where $\hat{\gamma}_{1,P}$ is the PQL estimate and the quantities B, C and D are given in Breslow & Lin (1995), page 88.

Lin and Breslow made some suggestions on how to implement these two adjustments in practice. They suggested first applying the correction for $\boldsymbol{\gamma}$, followed by re-estimation of $\boldsymbol{\tau}$, and then finally applying a correction for $\boldsymbol{\tau}$. For data where the estimated variance components using PQL are greater than 1, they recommended ignoring the correction for $\boldsymbol{\tau}$.

Given that their derivation used an approximation around $\boldsymbol{\gamma} = 0$, their bias correction factors are limited in applicability to cases where the (estimated) variance components are relatively small. For binary data, for instance, Lin & Breslow (1996) suggested that their correction factors work satisfactorily when the estimated variance components are less than 1. In addition, Lin and Breslow noted (section 5.2 of Lin & Breslow, 1996) that their derivation holds only when the random effects design matrix $\boldsymbol{Z}$ is highly sparse. This requirement implies that, for grouped data, for instance, the number of groups should increase with the sample size or, equivalently, that the group size remains small and constant. As noted earlier, PQL is expected to incur large estimation biases for such cases.

We are not aware of any statistical software available which applies this correction technique – it is left to the practitioner to write the necessary code to apply these correction factors for themselves.

PQL2: Goldstein & Rasbash (1996) proposed a "second order" approximation, PQL2. This is an extension to the Taylor series derivation of PQL in (2.2), but using a second order Taylor series expansion. As for the previous PQL derivation, let $\eta_i^{(k-1)} = \boldsymbol{x}_i^T \hat{\boldsymbol{\tau}}^{(k-1)} + \boldsymbol{z}_i^T \tilde{\boldsymbol{u}}^{(k-1)}$, where $\hat{\boldsymbol{\tau}}^{(k-1)}$ and $\tilde{\boldsymbol{u}}^{(k-1)}$ are the current estimates and $h = g^{-1}$ is the inverse link function. A second order expansion around $\eta_i = \eta_i^{(k-1)}$ gives

$$y_i \approx h\left(\eta_i^{(k-1)}\right) + h'\left(\eta_i^{(k-1)}\right)\left(\eta_i - \eta_i^{(k-1)}\right) + \frac{1}{2}h''\left(\eta_i^{(k-1)}\right)\left(\eta_i - \eta_i^{(k-1)}\right)^2 + e_i,$$

which, after ignoring second order terms involving $\boldsymbol{\tau}$, becomes

$$\begin{aligned} y_i \approx\ & h\left(\eta_i^{(k-1)}\right) + h'\left(\eta_i^{(k-1)}\right)\boldsymbol{x}_i^T\left(\boldsymbol{\tau} - \boldsymbol{\tau}^{(k-1)}\right) + h'\left(\eta_i^{(k-1)}\right)\boldsymbol{z}_i^T\left(\boldsymbol{u} - \boldsymbol{u}^{(k-1)}\right) \\ & + \frac{1}{2}h''\left(\eta_i^{(k-1)}\right)\boldsymbol{z}_i^T\left(\boldsymbol{u} - \boldsymbol{u}^{(k-1)}\right)\left(\boldsymbol{u} - \boldsymbol{u}^{(k-1)}\right)^T\boldsymbol{z}_i + e_i. \end{aligned}$$

This expression is further simplified by replacing the last term by its expectation, and so a modified "working" variable can be formed:

$$\begin{aligned} \psi_i^{(k)} =\ & \left(\boldsymbol{x}_i^T\boldsymbol{\tau}^{(k-1)} + \boldsymbol{z}_i^T\boldsymbol{u}^{(k-1)}\right) + g'\left(\mu_i^{u,(k-1)}\right)\left(y_i - \mu_i^{u,(k-1)}\right) \\ & - \frac{1}{2}g'\left(\mu_i^{u,(k-1)}\right)h''\left(\eta_i^{(k-1)}\right)\boldsymbol{z}_i^T\mathrm{Var}\left(\boldsymbol{u} - \boldsymbol{u}^{(k-1)}\right)\boldsymbol{z}_i. \end{aligned}$$

As with PQL, a linear mixed model (2.1) can then be fitted to this modified "working" variable, $\boldsymbol{\psi}^{(k)} = (\psi_1^{(k)}, \ldots, \psi_n^{(k)})^T$, to generate new estimates $\hat{\boldsymbol{\tau}}^{(k)}$ and $\tilde{\boldsymbol{u}}^{(k)}$. The same iterative procedure is used as in PQL – the only difference to PQL is in the formation of the working variable, as above.

PQL2 has been implemented in the MLwiN statistical package (Goldstein *et al.*, 1998).

The iterated bootstrap method (Kuk, 1995): Kuk (1995) proposed a computationally intensive simulation approach to correct the bias. This approach is a general statistical approach for correcting estimation bias, not just for PQL.

Let $\boldsymbol{\varphi} = (\boldsymbol{\tau}^T, \boldsymbol{\gamma}^T)^T$. Using the PQL estimate denoted $\hat{\boldsymbol{\varphi}}^{(0)}$, a set of simulated datasets are generated from the model $f_Y\left(\boldsymbol{y}; \hat{\boldsymbol{\varphi}}^{(0)}\right)$. PQL is then applied to each simulated dataset, with the average PQL estimate denoted $\tilde{\boldsymbol{\varphi}}^{(0)}$ and estimated bias $\boldsymbol{b}^{(0)} = \tilde{\boldsymbol{\varphi}}^{(0)} - \hat{\boldsymbol{\varphi}}^{(0)}$, giving a revised estimate $\hat{\boldsymbol{\varphi}}^{(1)} = \hat{\boldsymbol{\varphi}}^{(0)} - \boldsymbol{b}^{(0)}$. A further set of datasets are generated from $f_Y(\boldsymbol{y}; \hat{\boldsymbol{\psi}}^{(1)})$, and PQL is applied to each of these, with average PQL estimate $\tilde{\boldsymbol{\varphi}}^{(1)}$ and revised bias $\boldsymbol{b}^{(1)} = \tilde{\boldsymbol{\varphi}}^{(1)} - \hat{\boldsymbol{\varphi}}^{(1)}$, giving a new revised estimate $\hat{\boldsymbol{\varphi}}^{(2)} = \hat{\boldsymbol{\varphi}}^{(0)} - \boldsymbol{b}^{(1)}$. This process is repeated until the estimated bias converges, that is, $\boldsymbol{b}^{(k)} = \boldsymbol{b}^{(k-1)}$ at a given iteration k (within a suitable tolerance). Equivalently, the process is repeated until the average PQL estimate at a given iteration k, $\tilde{\boldsymbol{\varphi}}^{(k)}$, is equal to the original PQL estimate, $\hat{\boldsymbol{\varphi}}^{(0)}$ (again, within a suitable tolerance).

Kuk (1995) showed that, provided convergence is achieved, this method will yield consistent estimates. However, the computational burden of this approach can obviously be enormous, as noted by Rodriguez & Goldman (2001). Goldstein (1996) discussed the calculation of standard errors and hypotheses tests using this technique, which has been implemented in the MLwiN package (Goldstein *et al.*, 1998).

2.1.2 Hierarchical GLM approach of Lee and Nelder

"Hierarchical GLMs" (HGLMs) of Lee & Nelder (1996, 2001) encompass GLMMs as a special case. As already noted in section 1.4.2, HGLMs generalize GLMMs to include models with non-normal random effects. In addition, HGLMs encompass flexible modelling of the dispersion ϕ as well as the mean μ_i^u. This allows for systematic changes in the variability of the response not accounted for by the quasi-likelihood distribution or by the inclusion of random effects into the model. An even broader model class called "Double hierarchical GLMs" (DHGLMs) has been developed and is discussed in Lee & Nelder (2006), as well as in their recent book, Lee, Nelder & Pawitan (2006). Hierarchical GLMs are also illustrated in several papers dealing with

simulated or real datasets (e.g. Yun & Lee, 2004; Noh, Yip, Lee & Pawitan, 2006; Lee & Nelder, 2003, 2005; Lee, Yun & Lee, 2003; Noh & Lee, 2007) and have been implemented in the software package GenStat, as described in Payne *et al.* (2006).

It should be noted here that the use of the qualifier "hierarchical" by Lee & Nelder has caused some confusion. In Bayesian inference, the term "hierarchical models" defines an even broader class of models, for instance, as used in the paper by Hobert (2000). HGLMs, in the more restrictive Lee & Nelder sense, have some appeal to statisticians who are not inclined towards Bayesian modelling, and in particular amongst statisticians in agricultural or biological applications.

Lee & Nelder (2001) also outlined a systematic approach to the analysis of HGLMs. In this section, the discussion of this approach is restricted to its application for GLMMs.

2.1.2.1 The h-likelihood and profile likelihoods for $\boldsymbol{\tau}$ and $\boldsymbol{\gamma}$

In the HGLM approach, the joint likelihood ℓ_J is denoted as the "hierarchical likelihood" or h-likelihood, h. The use of the term "likelihood" here has resulted in confusion and criticism from some statisticians, such as Lindsey (2004) and Kuk & Cheng (1999) (with the latter paper rebutted in Lee & Nelder, 2005), who are quick to point out that $h = \ell_J$ is not a likelihood in the standard sense.

Similar to the Breslow & Clayton (1993) derivation of PQL, the Lee & Nelder approach is based on the use of a Laplace approximation. To approximate the likelihood ℓ, they suggested a (first order) Laplace approximation (section 1.3.2.1)

$$\begin{aligned} \ell \approx p_u(h) &= \left(h - \frac{1}{2} \log \left| -\frac{1}{2\pi} \frac{\partial^2 h}{\partial \boldsymbol{u} \partial \boldsymbol{u}^T} \right| \right)_{\tilde{\boldsymbol{u}}_{\tau,\gamma}} \\ &\approx \left(h - \frac{1}{2} \log \left| \boldsymbol{Z}^T \boldsymbol{W} \boldsymbol{Z} + \boldsymbol{G}^{-1} \right| + \frac{b}{2} \log 2\pi \right)_{\tilde{\boldsymbol{u}}_{\tau,\gamma}}, \end{aligned} \tag{2.6}$$

where $\boldsymbol{W}$ is, as before, a diagonal matrix of GLM weights, and $\tilde{\boldsymbol{u}}_{\tau,\gamma}$ is the mode of h for given $\boldsymbol{\tau}$ and $\boldsymbol{\gamma}$, and b is the length of $\boldsymbol{u}$. Note that an approximation sign is used in the second line of (2.6), since $\partial^2 h / \partial \boldsymbol{u} \partial \boldsymbol{u}^T = \boldsymbol{Z}^T \boldsymbol{W} \boldsymbol{Z} + \boldsymbol{G}^{-1} + \boldsymbol{R}$, where $\boldsymbol{R}$

is a remainder term which is $\mathbf{0}$ for canonical links and has expectation $\mathbf{0}$ for non-canonical links. In keeping with Lee & Nelder's notation, the use of $p_u(h)$ reflects a transformation $p_u(\cdot)$ of h where the random effects $\boldsymbol{u}$ are effectively "integrated" out of the h-likelihood.

Lee & Nelder suggested the use of $p_u(h)$ for inference concerning $\boldsymbol{\tau}$. However, for inference concerning $\boldsymbol{\gamma}$, they suggested a further approximation which "conditions" out $\boldsymbol{\tau}$ from the estimated likelihood $p_u(h)$, in the sense of an adjusted Cox-Reid profile likelihood (Cox & Reid, 1987). The form of this approximation is the same as that used to "integrate out" $\boldsymbol{u}$ to form $p_u(h)$. Let $\boldsymbol{\beta} = (\boldsymbol{\tau}^T, \boldsymbol{u}^T)^T$. The likelihood for inference concerning γ is thus

$$p_\tau\left(p_u(h)\right) \quad = \quad \left(p_u(h) - \frac{1}{2}\log\left|-\frac{1}{2\pi}\frac{\partial^2 p_u(h)}{\partial\boldsymbol{\tau}\partial\boldsymbol{\tau}^T}\right|\right)_{\hat{\boldsymbol{\tau}}_\gamma}.$$

where $\hat{\boldsymbol{\tau}}_\gamma$ satisfies $\partial p_u(h)/\partial\boldsymbol{\tau} = \mathbf{0}$. However, this expression could be tedious to compute. Instead, Lee & Nelder (2001) argued for a further approximation on $p_\tau\left(p_u(h)\right)$:

$$\begin{aligned} p_\tau\left(p_u(h)\right) &\approx p_\beta(h) \\ &= \left(h - \frac{1}{2}\log\left|-\frac{1}{2\pi}\frac{\partial^2 h}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right|\right)_{\hat{\boldsymbol{\beta}}_\gamma} \qquad (2.7) \\ &\approx \left(h - \frac{1}{2}\log\left|\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z} + \boldsymbol{G}^{-1}\right| - \frac{1}{2}\log\left|\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right| + \frac{p+b}{2}\log 2\pi\right)_{\hat{\boldsymbol{\beta}}_\gamma} \end{aligned}$$

where $\boldsymbol{V} = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T + \boldsymbol{W}^{-1}$ and $\hat{\boldsymbol{\beta}}_\gamma = (\hat{\boldsymbol{\tau}}_\gamma^T, \tilde{\boldsymbol{u}}_{\tau,\gamma}^T)^T$. For a normal linear mixed model, both these expressions correspond to the REML likelihood (section 1.2.2.1). For this reason, Lee & Nelder (2003) described their approach as an extension of REML to non-normal models.

For models where there are few observations per random effect, such as grouped binary data with small group sizes, Lee & Nelder suggested the use of a second order Laplace approximation to obtain a better approximation to the true likelihood ℓ. Therefore, they calculated

$$p_u^s(h) \quad = \quad p_u(h) - F/24,$$

where $-F/24$ represents the difference between a first and second order Laplace approximation. This term is mathematically complex, involving third order and fourth order derivatives of h with respect to the vector $\boldsymbol{u}$, and so its definition will be deferred until Chapter 4. This term is similarly included to form a "second order" likelihood for $\boldsymbol{\gamma}$,

$$p^s_\beta(h) = p_\beta(h) - F/24.$$

Raudenbush *et al.* (2000) and Shun (1997) also examined the application of second order Laplace approximations for GLMMs, but only for a two level (one-way classification) and simple crossed model respectively.

2.1.2.2 Levels of approximation

Lee & Nelder's approach allows for inferences with several different levels of approximation.

Here the approach will be denoted as $\texttt{HG}(m,d)$, where m and d each take values 0, 1 or 2 to indicate the level of approximation used for making inference about $\boldsymbol{\tau}$ and $\boldsymbol{\gamma}$ respectively. (As in Lee & Nelder's papers, "m" and "d" represent "mean" and "dispersion" parameters respectively.). The likelihood used for inference about $\boldsymbol{\tau}$ and $\boldsymbol{\gamma}$ for each of the methods are as in the following table. Here, only the cases where $m \leq d$ are shown, since it is deemed more important to use higher order approximations for $\boldsymbol{\gamma}$ than it is for $\boldsymbol{\tau}$ to correct the biases.

Approximation	*Likelihood for* $\boldsymbol{\tau}$	*Likelihood for* $\boldsymbol{\gamma}$
HG(0,0)	h	PQL likelihood
HG(0,1)	h	$p_\beta(h)$
HG(0,2)	h	$p^s_\beta(h)$
HG(1,1)	$p_u(h)$	$p_\beta(h)$
HG(1,2)	$p_u(h)$	$p^s_\beta(h)$
HG(2,2)	$p^s_u(h)$	$p^s_\beta(h)$

Noh & Lee (2007) showed the details of the implementation of each of these levels of approximation, including score and iterative updating equations. Their derivations showed that these approximations can be implemented in a similar way to PQL, with alternate estimation of $\boldsymbol{\tau}$ and $\boldsymbol{u}$ and $\boldsymbol{\gamma}$.

As yet, the methodology outlined above by Lee & Nelder has received little critical assessment in the statistical community, partly perhaps because of confusion of their methodology with PQL, the emphasis on Bayesian and Monte Carlo approaches in the current literature, and because of the authors' perceived anti-Bayesian biases.

2.2 Numerical methods – Gauss-Hermite quadrature

As in section 1.3.2.2, an m-point Gauss-Hermite quadrature (GHQ) approximation to an integral of $f(x) = g(x)\exp(-x^2)$ is

$$\int_{-\infty}^{\infty} g(x)\exp(-x^2)dx \simeq \sum_{i=1}^{m} w_i g(\xi_i),$$

where the "nodes" ξ_i are the roots of the mth order Hermite polynomial with corresponding weights w_i, both of which are available from standard references (e.g. Abramowitz & Stegun, 1972, p 890). GHQ can be applied to multivariate integrals by applying it to each univariate integral in turn (Pan & Thompson, 2003, p 62),

$$\begin{aligned}
\int f(\boldsymbol{x})d\boldsymbol{x} &= \int \dots \int f(x_1, \dots, x_q)dx_1 \dots dx_q \\
&= \int \dots \int g(x_1, \dots, x_q)e^{-\boldsymbol{x}^T\boldsymbol{x}}dx_1 \dots dx_q \\
&\simeq \sum_{i_1=1}^{m_1} w_{i_1}^{(1)} \sum_{i_2=1}^{m_2} w_{i_2}^{(2)} \dots \sum_{i_q=1}^{m_q} w_{i_q}^{(q)} g(\xi_{i_1}^{(1)}, \xi_{i_2}^{(2)}, \dots, \xi_{i_q}^{(q)}),
\end{aligned}$$

where $\xi_{i_j}^{(j)}$ and $w_{i_j}^{(j)}$ $(i = 1 \dots m_j)$ are the nodes and weights for m_j-point GHQ at the jth coordinate of $\boldsymbol{x}$, $j = 1, \dots, q$. However, this expression requires evaluation of $g(x_1, x_2, \dots, x_q)$ at $m_1 m_2 \dots m_q$ nodes, and so can be very computationally intensive.

2.2.1 Quadrature for nested random effects models

For a general GLMM, the expression for the log-likelihood,

$$\ell = \log \int f_{Y,U}(\boldsymbol{y}, \boldsymbol{u}) d\boldsymbol{u},$$

is a multivariate integral. If $\boldsymbol{u}$ is of length q, application of an m-point GHQ approximation for each univariate integral would require m^q evaluations of $f_{Y,U}(\boldsymbol{y}, \boldsymbol{u})$. For GLMMs involving only nested random effects, the hierarchical structure can be utilised to dramatically reduce the number of function evaluations required.

To illustrate the use of quadrature for nested random effects models, a two-way nested classification model will be used. The conditional mean $\mu^u_{ijk} = E(y_{ijk}|u_{1i}, u_{2ij})$ for the kth sub-unit within the jth unit within the ith cluster satisfies

$$g(\mu^u_{ijk}) = \tau_0 + u_{1i} + u_{2ij}, \qquad i = 1 \dots b_g,\ j = 1 \dots m_g, k = 1 \dots m_s,$$

where $u_{1i} \sim N(0, \gamma_1)$ and $u_{2ij} \sim N(0, \gamma_2)$.

Let the random effect vectors be represented as $\boldsymbol{u}_1 = (u_{11}, \dots, u_{1b_g})^T$ and $\boldsymbol{u}_2 = (u_{211}, \dots, u_{21m_g}, \dots, u_{2b_g1}, \dots, u_{2b_gm_g})^T$, with corresponding PDFs $f_{U_1}(\boldsymbol{u}_1) = \prod_i f_{u_1}(u_{1i})$ and $f_{U_2}(\boldsymbol{u}_2) = \prod_i \prod_j f_{u_2}(u_{2ij})$, where f_{u_1} and f_{u_2} are normal PDFs. Let the conditional PDF of the data $\boldsymbol{y}$ given random effect vectors $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ be given by

$$f_{Y|U_1,U_2}(\boldsymbol{y}|\boldsymbol{u}_1, \boldsymbol{u}_2) = \prod_i \prod_j \prod_k f_{y|u_1,u_2}(y_{ijk}|u_{1i}, u_{2ij}).$$

The expression for the likelihood can be represented as a product of univariate integrals (over u_{2ij}) nested within another univariate integral (over u_{1i}), that is,

$$\begin{aligned}
&\int \dots \int f_{Y|U_1,U_2}(\boldsymbol{y}|\boldsymbol{u}_1, \boldsymbol{u}_2) f_{U_1}(\boldsymbol{u}_1) f_{U_2}(\boldsymbol{u}_2) d\boldsymbol{u}_1 d\boldsymbol{u}_2 \\
&= \prod_i \int_{u_{1i}} \left(\prod_j \int_{u_{2ij}} \left\{ \prod_k f(y_{ijk}|u_{1i}, u_{2ij}) \right\} f_{u_2}(u_{2ij}) du_{2ij} \right) f_{u_1}(u_{1i}) du_{1i}.
\end{aligned}$$

If GHQ is applied to each univariate integral, this becomes, after multiplying and

dividing each integral by $\exp(-u_{1i}^2)$ or $\exp(-u_{2ij}^2)$,

$$\begin{aligned}
&\simeq \prod_i \sum_{l_1=1}^{m} w_{l_1,m} \exp(\xi_{l_1,m}^2) f_{u_1}(\xi_{l_1,m}) \\
&\quad \left(\prod_j \sum_{l_2=1}^{m} w_{l_2,m} \exp(\xi_{l_2,m}^2) \left\{ \prod_k f_{y|u_1,u_2}(y_{ijk}|\xi_{l_1,m}, \xi_{l_2,m}) \right\} f_{u_2}(\xi_{l_2,m}) \right),
\end{aligned}$$

where m point GHQ is used throughout, and $w_{l,m}$ and $\xi_{l,m}$ are the lth weight and lth node for m-point quadrature. The utilisation of the nested structure has reduced the number of function evaluations from $O\left(m^{U+UV}\right)$ to just $O\left(m + m^2\right)$.

Note that an alternative way to apply GHQ above is to perform a change of variable in each integral. For instance, for the "inner" integral with respect to u_{2ij},

$$\begin{aligned}
&\int_{v_{ij}} \left\{ \prod_k f(y_{ijk}|u_{1i}, u_{2ij}) \right\} f(u_{2ij}) du_{2ij} \\
= &\int \left\{ \prod_k f(y_{ijk}|u_{1i}, u_{2ij}) \right\} \frac{1}{\sqrt{2\pi\gamma_2}} \exp\left(-u_{2ij}^2/2\gamma_2\right) du_{2ij} \\
= &\int \left\{ \prod_k f(y_{ijk}|u_{1i}, \sqrt{2\gamma_2} v_{2ij}) \right\} \frac{1}{\sqrt{\pi}} \exp\left(-v_{2ij}^2\right) dv_{2ij} \\
= &\sum_{l_2=1}^{m} \frac{w_{l_2,m}}{\sqrt{\pi}} \left\{ \prod_k f(y_{ijk}|u_{1i}, \sqrt{2\gamma_2}\xi_{l_2,m}) \right\},
\end{aligned}$$

where the change of integration variable, $v_{2ij} = u_{2ij}/\sqrt{2\gamma_2}$, has been applied. A similar change of variable can be applied to the outer integral with respect to u_{1i}, i.e. $v_{1i} = u_{1i}/\sqrt{2\gamma_1}$. The change of variable may be preferable in general, since the points where $f_{y|u_1,u_2}$ are evaluated, with respect to u_{2ij}, are chosen according to the magnitude of the variance γ_1. That is, these points are at $\sqrt{2\gamma_2}\xi_{l_2,m}$, rather than at $\xi_{l_2,m}$.

When the random effects are not independent, such as when spatial or temporal correlation is modelled, a re-parametrization of the random effects can be applied. Let $\boldsymbol{G} = \boldsymbol{L}\boldsymbol{L}^T$ be the Cholesky decomposition of $\mathbf{var}(\boldsymbol{u}) = \boldsymbol{G}$, where $\boldsymbol{L}$ is lower triangular. The use of a re-parametrization, $\boldsymbol{u} = \boldsymbol{L}^T\boldsymbol{v}$, gives a new linear predictor $g(\mu_i^u) = \boldsymbol{x}_i^T\boldsymbol{\tau} + \boldsymbol{z}_i^{*T}\boldsymbol{v}$, where $\boldsymbol{z}_i^* = \boldsymbol{L}\boldsymbol{z}_i$. The re-parametrized random effects, $\boldsymbol{v} \sim$

$N(\mathbf{0}, \boldsymbol{I})$, are uncorrelated, and so standard GHQ can be applied for each univariate integral in turn.

2.2.1.1 Quadrature for crossed effects

For crossed random effects, GHQ is less useful, since it is not possible to simplify the multivariate integral required to evaluate the likelihood. Therefore, GHQ is not feasible for most models with crossed random effects. Consider the two-way crossed model for data y_{ij}, $i = 1, \ldots, b_1$, $j = 1, \ldots, b_2$, where the model for the conditional mean $\mu^u_{ij} = E(y_{ij}|u_{1i}, u_{2j})$ is

$$g(\mu^u_{ij}) = \tau_0 + u_{1i} + u_{2j},$$

where $u_{1i} \sim N(0, \gamma_1)$ and $u_{2j} \sim N(0, \gamma_2)$. Let $\boldsymbol{u}_1 = (u_{11}, \ldots, u_{1b_1})$ and $\boldsymbol{u}_2 = (u_{21}, \ldots, u_{2b_2})$ be the vectors of random effects with densities $f_{U_1}(\boldsymbol{u}_1) = \prod_i f_{u_1}(u_{1i})$ and $f_{U_2}(\boldsymbol{u}_2) = \prod_j f_{u_2}(u_{2j})$. Let $f_{Y|U_1,U_2}(\boldsymbol{y}|\boldsymbol{u}_1, \boldsymbol{u}_2) = \prod_i \prod_j f_{y|u_1,u_2}(y_{ij}|u_{1i}, u_{2j})$ be the conditional PDF of the data $\boldsymbol{y}$ given the random effects vectors $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$.. In contrast to the nested model, the expression for the likelihood,

$$\begin{aligned}
&\int \ldots \int f_{Y|U_1,U_2}(\boldsymbol{y}|\boldsymbol{u}_1, \boldsymbol{u}_2) f_{U_1}(\boldsymbol{u}_1) f_{U_2}(\boldsymbol{u}_2) d\boldsymbol{u}_1 d\boldsymbol{u}_2 \\
&= \int \ldots \int \left(\prod_i \prod_j f_{y|u_1,u_2}(y_{ij}|u_{1i}, u_{2j}) \right) \left(\prod_i f_{u_1}(u_{1i}) \right) \left(\prod_j f_{u_2}(u_{2j}) \right) d\boldsymbol{u}_1 d\boldsymbol{u}_2,
\end{aligned}$$

cannot be simplified any further.

2.2.2 Adaptive Gauss-Hermite quadrature

Naylor & Smith (1982) and Liu & Pierce (1994) suggested an improvement to standard GHQ. The idea is simply to re-parametrize the integral before applying GHQ, in such a way that the nodes are centred around the mean or mode of the integrand, rather than around zero.

For simplicity, the technique is described for a univariate integral $\int_{-\infty}^{\infty} f(x)dx$. Let

$\hat{\mu}$ be the mode of $f(x)$ or the mean of a variable with probability density function (PDF) proportional to $f(x)$, and let $\hat{\sigma}^2$ represent either the estimated curvature of $f(x)$ at $\hat{\mu}$,

$$\hat{\sigma}^2 = \left(-\partial^2 \log f(x)/\partial x^2\right)^{-1}\Big|_{x=\hat{\mu}},$$

or the variance of a variable with PDF proportional to $f(x)$. Let the integral be

$$\int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{\infty} \frac{f(x)}{\phi(x;\hat{\mu},\hat{\sigma}^2)}\phi(x;\hat{\mu},\hat{\sigma}^2)dx = \int_{-\infty}^{\infty} h(x)\frac{1}{\sqrt{2\pi}\hat{\sigma}}e^{-(x-\hat{\mu})^2/2\hat{\sigma}^2}dx,$$

where $h(x) = f(x)/\phi(x;\hat{\mu},\hat{\sigma})$ and $\phi(x;\mu,\sigma^2)$ is a normal PDF with parameters μ and σ^2. If a re-parametrization of the integration variable $t = (x-\hat{\mu})/\left(\sqrt{2}\hat{\sigma}\right)$ is applied, then this expression becomes

$$\begin{aligned} \ldots &= \int_{-\infty}^{\infty} h(\sqrt{2}\hat{\sigma}t + \hat{\mu})\frac{1}{\sqrt{2\pi}\hat{\sigma}}e^{-t^2}\sqrt{2}\hat{\sigma}dt \\ &= \int_{-\infty}^{\infty} h\left(\sqrt{2}\hat{\sigma}t + \hat{\mu}\right)\frac{1}{\sqrt{\pi}}e^{-t^2}dt, \end{aligned}$$

and, with the application of GHQ, this becomes

$$\ldots \simeq \sum_{l=1}^{m}\frac{w_{l,m}}{\sqrt{\pi}}h\left(\hat{\mu} + \sqrt{2}\hat{\sigma}\xi_{l,m}\right) = \sqrt{2}\hat{\sigma}\sum_{l=1}^{m} w^*_{l,m} f\left(\hat{\mu} + \sqrt{2}\hat{\sigma}\xi_{l,m}\right),$$

where $w^*_{i,m} = w_{i,m}\exp(\xi^2_{i,m})$, and $w_{l,m}$ and $\xi_{l,m}$ are the lth weights and lth node respectively for m point GHQ. Lesaffre & Spiessens (2001) demonstrated the deficiency of standard GHQ, and the superiority of adaptive GHQ, for one particular dataset with large variance components. They showed that standard GHQ was numerically unstable in this case, unless a very large number of quadrature points was used.

2.2.2.1 Adaptive GHQ for nested random effects models

To illustrate the application of adaptive GHQ for nested GLMMs, a one way classification model will be used. Let data y_{ij}, $i = 1,\ldots,b_g$, $j = 1,\ldots,m_g$, have conditional

mean $\mu_{ij}^u = E(y_{ij}|u_i)$, where

$$g(\mu_{ij}^u) = \tau_0 + u_i,$$

and $u_i \sim N(0, \sigma_u^2)$. Let $f_{Y|U}(\boldsymbol{y}|\boldsymbol{u}) = \prod_i \prod_j f_{y|u}(y_{ij}|u_i)$ be the conditional PDF of the data given the random effects, and $f_U(\boldsymbol{u}) = \prod_i f_u(u_i)$ be the PDF of the random effects. The likelihood is

$$\begin{aligned}
&\int \dots \int f_{Y|U}(\boldsymbol{y}|\boldsymbol{u}) f_U(\boldsymbol{u}) d\boldsymbol{u} \\
&= \int \dots \int \left(\prod_i \prod_j f_{y|u}(y_{ij}|u_i) \right) \left(\prod_i f_u(u_i) \right) d\boldsymbol{u} \\
&= \prod_i \int \frac{\left(\prod_j f_{y|u}(y_{ij}|u_i) \right) f_u(u_i)}{\phi(u_i; \tilde{u}_i, \hat{\sigma}_i^2)} \frac{1}{\sqrt{2\pi}\hat{\sigma}} e^{-(u_i - \tilde{u}_i)^2 / 2\hat{\sigma}_i^2} du_i,
\end{aligned}$$

where $\hat{\mu}_i$ and $\hat{\sigma}_i^2$ are the estimated mode and dispersion of $\left\{ \prod f_{y|u}(y_{ij}|u_i) \right\} f_u(u_i)$, and $\phi(x; \mu, \sigma^2)$ is the normal PDF with mean μ and variance σ^2. Using a transformation $t_i = (u_i - \hat{\mu}_i)/\sqrt{2}\hat{\sigma}_i$, and applying GHQ to the resultant integral, gives

$$\begin{aligned}
\dots &= \prod_i \int e^{t_i^2} \sqrt{2}\hat{\sigma}_i \left(\prod_j f_{y|u}(y_{ij}|\sqrt{2}\hat{\sigma}_i t_i + \hat{\mu}_i) \right) f_u(\sqrt{2}\hat{\sigma}_i t_i + \hat{u}_i) e^{-t_i^2/2} dt_i \\
&\simeq \prod_i \sum_{s=1}^{S} w_{s,m} e^{\xi_s^2} \sqrt{2}\hat{\sigma}_i \left(\prod_j f_{y|u}(y_{ij}|\sqrt{2}\hat{\sigma}_i \xi_{s,m} + \hat{\mu}_i) \right) f_u(\sqrt{2}\hat{\sigma}_i \xi_{s,m} + \hat{u}_i),
\end{aligned}$$

where, as before, $w_{s,m}$ and $\xi_{s,m}$ represent the weights and nodes for m-point GHQ.

2.2.3 Implementation of Gauss-Hermite quadrature for GLMMs

Most GHQ software currently implements standard GHQ. These include the older packages `Mixor` (Hedeker & Gibbons, 1996) and `Egret` (Statistics and Epidemiology Research Corporation, 1993), as well as newer packages such as `AML` (Lillard & Panis, 2003). Some more recent GHQ software, however, implements adaptive GHQ, such as the SAS `NLMIXED` procedure (Wolfinger, 1999), the R function `lmer` from the `lme4` package (Bates & Sarkar, 2006), and the Stata `GLLAMM` procedure (Rabe-Hesketh *et al.*, 2001). However, the `NLMIXED` and `lmer` implementations are limited to GLMMs

with one random classification only, although the `GLLAMM` package can fit nested two way classifications as well.

Maximisation of the GHQ-derived likelihood with respect to the parameters $\boldsymbol{\tau}$ and $\boldsymbol{\kappa}$ can be achieved using derivative-free optimisation, as in `lmer`, or by using Newton-Raphson or quasi-Newton approaches, which requires analytical expressions for the score equations and information matrices. Such analytical expressions can also be derived using GHQ in the same way as the analytical approximation for the likelihood is derived. However, Lesaffre & Spiessens (2001) indicated that using standard GHQ to evaluate the score equations, which involves the approximation of one integral over the approximation of another integral, can lead to severe estimation biases.

Adaptive GHQ also incurs an additional complication over standard GHQ in the estimation of the mode and curvature of the integrand with respect to each random effect u_i. In GLLAMM, as described in Rabe-Hesketh *et al.* (2002), the mean and standard deviation are re-calculated after each iteration of the Newton-Raphson technique, and these calculations in turn requires an iterative technique. Pan & Thompson (2000) described a modification to adaptive GHQ where a further approximation is applied, resulting in score equations of a similar form to the score equations used in PQL, and so denote their technique "GH-PQL". Finally, Clarkson & Zhan (2002) described the application of "spherical-radial quadrature" (Monahan & Genz, 1997) to GLMMs. This is a variant on adaptive GHQ, where the multivariate integral for the likelihood is re-expressed as

$$\ell = \log \int f_{Y,U}(\boldsymbol{y}, \boldsymbol{u}) d\boldsymbol{u} = \log \int_{r=0}^{\infty} \left(\int_{A(\boldsymbol{v})} f_{Y,V}(\boldsymbol{y}, r\boldsymbol{v}) d\boldsymbol{v} \right) dr,$$

and where $r\boldsymbol{v} = \boldsymbol{B}^{-1}(\boldsymbol{u} - \tilde{\boldsymbol{u}})$ is a centered version of $\boldsymbol{u}$, with $\tilde{\boldsymbol{u}}$ being the mode of $\log f_{Y,U}$ and $\boldsymbol{B} = \partial^2 \log f_{Y,U} / \left\{ \partial \boldsymbol{u} \partial \boldsymbol{u}^T \right\}$ is the Hessian. The domain of integration for $\boldsymbol{v}$, $A(\boldsymbol{v})$, is the unit sphere in q dimensions, and is approximated by taking symmetrically spaced points over the unit sphere. GHQ is applied to the outer integral with respect to the "radius" r.

Usefulness of standard GHQ in calculating (marginal) probabilities As above, GHQ is useful for numerically evaluating the GLMM likelihood, and so facilitating the calculation of (approximate) maximum likelihood estimates. Standard GHQ is also useful for GLMMs for calculating marginal probabilities of events for fixed, or known, values of $\boldsymbol{\gamma}$ and $\boldsymbol{\tau}$. For instance, consider the simple one-way classification model for discrete data y_{ij}, $i = 1, \ldots, b_g$, $j = 1, \ldots, m_g$, where the model for $\mu_{ij}^u = E(y_{ij}|u_i)$ is given by

$$g(\mu_{ij}^u) = \tau_0 + \tau_1 x_{ij} + u_i,$$

where $u_i \sim N(0, \gamma_1)$. Let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{im_g})^T$. The marginal probability $P(\boldsymbol{y}_i = \boldsymbol{c})$ for some $\boldsymbol{c}$ is given by

$$P(\boldsymbol{y}_i = \boldsymbol{c}) = \int P(\boldsymbol{y}_i = \boldsymbol{c}|u_i) f_u(u_i) du_i$$

where $f_u(u_i)$ is a normal PDF with mean 0 and variance γ_1. Like the expression for the GLMM likelihood (1.9), this is an intractable integral. However, it can be calculated using GHQ as

$$\begin{aligned} P(\boldsymbol{y}_i = \boldsymbol{c}) &= \int \exp(u_i^2) P(\boldsymbol{y}_i = \boldsymbol{c}|u_i) f_u(u_i) \exp(-u_i^2) du_i \\ &= \sum_{t=1}^{m} w_t \exp(\xi_t) P(\boldsymbol{y}_i = \boldsymbol{c}|\xi_t) f_u(\xi_t) \end{aligned}$$

or, using a change of variables approach, where $v_i = u_i/\sqrt{2\gamma_i}$,

$$\begin{aligned} P(\boldsymbol{y}_i = \boldsymbol{c}) &= \int P(\boldsymbol{y}_i = \boldsymbol{c}|u_i) \frac{1}{\sqrt{2\pi\gamma_1}} \exp(-u_i^2/2\gamma_i) du_i \\ &= \int P(\boldsymbol{y}_i = \boldsymbol{c}|\sqrt{2\gamma_1} v_i) \frac{1}{\sqrt{\pi}} \exp(-v_i^2) dv_i \\ &= \sum_{t=1}^{m} \frac{w_t}{\sqrt{\pi}} P(\boldsymbol{y}_i = \boldsymbol{c}|\sqrt{2\gamma_1}\xi_t). \end{aligned}$$

2.3 Stochastic methods (including full Bayesian MCMC)

2.3.1 Monte Carlo methods

In this subsection, the use of Monte Carlo methods to evaluate the integral expression for the GLMM likelihood are reviewed.

2.3.1.1 Sampling techniques

The use of Monte Carlo methods requires the ability to draw random samples from any distribution with a given PDF $f(x)$. Specific techniques are available for some standard PDFs, such as the normal PDF. Alternatively, if the inverse CDF $F^{-1}(c)$ is available, where $F(y) = \int_{-\infty}^{y} f(x)dx$, the "inversion technique" can be used, where the sampled values are $F^{-1}(y)$ for $y \sim \text{Uniform}(0, 1)$.

In other cases, it may not be possible to sample from $f(x)$ directly, but from another similar PDF $g(x)$, denoted the "proposal density function". For instance, if $f(x)$ is a unimodal pdf, a possible choice for $g(x)$ may be a normal PDF with the same mode as $f(x)$, and the same curvature (second derivative) at the mode as $f(x)$. A number of sampling methods are available when a similar PDF $g(x)$ is available. *Importance sampling* samples from $g(x)$ and weights the sampled values $x^{(i)}$ by $f(x^{(i)})/g(x^{(i)})$. For instance, the mean of $f(x)$ can be estimated as $\left(\sum_i x^{(i)} f(x^{(i)})/g(x^{(i)})\right)/N$ if $x^{(i)} \sim g(x^{(i)})$. The adequacy of importance sampling depends on how close $g(x)$ is to $f(x)$. *Rejection sampling* involves application of a scaling factor c so that $cg(x) > f(x)$ for all x. For generating the ith sample $x^{(i)}$, a test sample x^* is drawn from $g(x)$ and a second sample y^* is drawn from Uniform $[0, cg\,(x^*)]$. If $y^* \leq f\,(x^*)$, then x^* is "accepted" and $x^{(i)} = x^*$, otherwise it is rejected and a new x^* is sampled until acceptance. A modified version of rejection sampling called derivative-free *adaptive rejection sampling* or ARS (Gilks & Wild, 1992; Gilks, 1992) is used in the popular Gibbs sampling package `BUGS` (Spiegelhalter, Thomas, Best & Gilks, 1995). ARS caters for the lack of an obvious choice for $g(x)$ by building $g(x)$ using the tangents of $f(x)$ at the sampled points.

More information on importance and rejection sampling can be found in many standard texts, for instance, in MacKay (1998).

2.3.1.2 Markov Chain Monte Carlo (MCMC)

Markov Chain Monte Carlo (MCMC) is a branch of Monte Carlo methods where sampling is performed from a transition PDF $T(x^*|x)$ whose stationary PDF is $f(x)$. At the ith iteration, the Metropolis-Hastings (MH) algorithm (Hastings, 1970) generates a new candidate sample x^* from $T(x^*|x^{(i-1)})$, where $x^{(i-1)}$ is the result of the previous iteration, and accepts it with probability

$$P_A = \min\left(1, \frac{f(x^*)T(x^{(i-1)}|x^*)}{f(x^{(i-1)})T(x^*|x^{(i-1)})}\right).$$

If the candidate sample is accepted, then $x^{(i)} = x^*$, else $x^{(i)} = x^{(i-1)}$. A special case of the MH algorithm is the Metropolis algorithm where $T(x|x^*) = T(x^*|x)$ so that $P_A = \min\left(1, f(x^*)/f(x^{(i-1)})\right)$. Another special case is the Gibbs sampler (section 1.3.1.2), where $\boldsymbol{x} = (x_1, \ldots, x_n)^T$ is multivariate and the MH algorithm is performed on each component x_j in turn. That is, if $\boldsymbol{x}_{-j} = (x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n)^T$, then $T(x_j|\boldsymbol{x}_{-j}) = f(x_j|\boldsymbol{x}_{-j})$, and so the probability of acceptance P_A is always equal to 1. In contrast to importance and rejection sampling where statistically independent samples are generated at each iteration, MCMC approaches generate serially dependent samples.

2.3.1.3 Monte Carlo approaches for GLMMs

Monte Carlo EM (MCEM) Monte Carlo implementations of the EM algorithm for GLMMs were described in Wei & Tanner (1990), McCulloch (1997), Booth & Hobert (1999) and Chen (2006), amongst others. In all these papers, Monte Carlo methods are used to approximate the intractable E-step of the EM algorithm as follows

$$\mathrm{E}\left\{\log f_{Y,U}(\boldsymbol{y}, \boldsymbol{u})|\boldsymbol{y}\right\} = \frac{1}{m}\sum_{i=1}^{m} \log f_{Y,U}\left(\boldsymbol{y}, \boldsymbol{u}^{(i)}\right), \tag{2.8}$$

where the $\boldsymbol{u}^{(i)}$ are sampled from the posterior PDF $f_{U|Y}(\boldsymbol{u}|\boldsymbol{y};\boldsymbol{\tau}_0,\boldsymbol{\kappa}_0)$, and $\boldsymbol{\tau}_0$ and $\boldsymbol{\kappa}_0$ are the current estimates. McCulloch (1997) used a dependent Metropolis-Hasting algorithm for sampling $\boldsymbol{u}^{(i)}$, with proposal density f_U, whereas Booth & Hobert (1999) proposed independent sampling, such as importance or rejection sampling. Chen (2006) suggested sampling from the distribution of the standardised random effects, $\boldsymbol{u}^* = \boldsymbol{G}^{-1/2}\boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{I})$. This standardised distribution is independent of the variance parameters $\boldsymbol{\kappa}$, and so removes the need to generate new samples $\boldsymbol{u}^{(i)}$ at each E-step. By removing the need for re-sampling, it also makes it easier to assess convergence, a stumbling block in the implementation of Monte Carlo algorithms. The simulated likelihood in (2.8) can be maximised readily with respect to the fixed effects $\boldsymbol{\tau}$, using an IRLS approach; with respect to the variance parameters $\boldsymbol{\kappa}$, it can be maximised directly without iteration.

Other Monte Carlo approaches for GLMMs McCulloch (1997) also suggested a Monte Carlo version of the Newton-Raphson equations for estimating the fixed effects $\boldsymbol{\tau}$. McCulloch (1997) and Ng, Carpenter, Goldstein & Rasbash (2006) explored the use of simulated maximum likelihood (SML), where the marginal likelihood is approximated using importance sampling,

$$\log \int f_{Y,U} d\boldsymbol{u} \simeq \log \left(\frac{1}{N} \sum_{i=1}^{N} \frac{f_{\boldsymbol{y}|\boldsymbol{u}}(\boldsymbol{y}|\boldsymbol{u}^{(i)}) f_U(\boldsymbol{u}^{(i)})}{h_u(\boldsymbol{u}^{(i)})} \right), \tag{2.9}$$

and $\boldsymbol{u}^{(i)}$ is sampled from a proposal PDF h_u. McCulloch (1997) demonstrated, through simulation, that SML works poorly unless the initial values are close to the optimal values, and suggested the preliminary use of MCEM (or MCNR) to achieve good starting values. Ng *et al.* (2006) alternatively recommended the preliminary use of approximate likelihood approaches such as PQL2 (section 2.1.1.2) to obtain reasonable starting values.

Delyon, Lavielle & Moulines (1999) outlined "Stochastic Approximate EM" (SAEM) where, at each E-step, a new function to be maximised is formed by accumulating the results from previous E-steps. If $Q(\boldsymbol{\tau}, \boldsymbol{\kappa}; \boldsymbol{\tau}^{(k-1)}, \boldsymbol{\kappa}^{(k-1)}) = \sum_{i=1}^{m} \log f_{Y,U}(\boldsymbol{y}, \boldsymbol{u}^{(i)})/m$ is

the expectation produced at the kth step where $\boldsymbol{u}^{(i)} \sim f_{U|Y}(\boldsymbol{u}|\boldsymbol{y}; \boldsymbol{\tau}^{(k-1)}, \boldsymbol{\kappa}^{(k-1)})$, then the new function to be maximised is

$$P^{(k)}(\boldsymbol{\tau}, \boldsymbol{\kappa}) = (1 - \alpha_k) P^{(k-1)}(\boldsymbol{\tau}, \boldsymbol{\kappa}) + \alpha_k Q(\boldsymbol{\tau}, \boldsymbol{\kappa}; \boldsymbol{\tau}^{(k-1)}, \boldsymbol{\kappa}^{(k-1)}), \qquad k > 1,$$

where $0 \leq \alpha_k \leq 1$, $\sum_k \alpha_k = \infty$ and $\sum_k \alpha_k^2 < \infty$ and $P^{(1)}(\boldsymbol{\tau}, \boldsymbol{\kappa}) = Q(\boldsymbol{\tau}, \boldsymbol{\kappa}; \boldsymbol{\tau}^{(0)}, \boldsymbol{\kappa}^{(0)})$ where $\boldsymbol{\tau}^{(0)}$ and $\boldsymbol{\kappa}^{(0)}$ are the initial values of $\boldsymbol{\tau}$ and $\boldsymbol{\kappa}$.

A variant on simulated maximum likelihood above, often called Quasi-Monte Carlo (QMC), is to select the points $\boldsymbol{u}^{(i)}$ deterministically rather than stochastically (e.g. Pan & Thompson, 2000; Kuk, 1999), which can both reduce the computational load and improve the convergence rate.

2.3.2 Full Bayesian approaches

In this sub-section, the implementation of full Bayesian approaches with the aid of Markov Chain Monte Carlo sampling is described.

2.3.2.1 Further sampling methods and issues

Following on from section 1.3.1, this section discusses sampling methods and issues for applying Bayesian approaches with MCMC, but not limited specifically to GLMMs. Research into improving MCMC sampling approaches is currently very active, and a few important developments of particular relevance to Bayesian inference are listed here. The use of *hierarchical centering* has been advocated to reduce correlations between successive samples and improve mixing (Gelfand, Sahu & Carlin, 1995), but may not always lead to improvements (Papaspiliopoulos, Roberts & Skold, 2003). The basic idea of hierarchical centering, when applied to GLMMs, is to sample from the posterior PDFs of re-parametrized random effects, $\boldsymbol{u}^* = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{u}$, instead of those from $\boldsymbol{u}$. Damien, Wakefield & Walker (1999), and Albert & Chib (1993) for probit GLMs, suggest the use of *auxiliary variables* to simplify the conditional densities required in a Gibbs sampler. This is the Monte Carlo analogue of the EM

algorithm (section 1.1.3.1). *Blocking* is a technique to improve the convergence rate of the Gibbs sampler by sampling from the conditional density of a group of parameters, rather than sampling from the conditional densities of each individual parameter in turn. Chib & Carlin (1999) indicated that, compared with standard Gibbs sampling, Gibbs sampling with blocking is more difficult to code, but should converge faster. *Slice sampling* (Neal, 2003) is a relatively new technique similar to rejection sampling which also appears particularly promising, as reviewed in Zhao, Staudenmayer, Coull & Wand (2003).

The choice of prior density can be a major stumbling block in implementing a Bayesian approach. Hobert & Casella (1996) cautioned against indiscriminate use of "flat" (supposed uninformative) priors in mixed models, which can result in an improper posterior density and hence give meaningless results. Kass & Wasserman (1996) discussed the choice of "uninformative" priors which lead to proper posterior densities. More recently, Gelman (2005) discussed the difficulties in choosing an uninformative prior for the variance parameters.

2.3.2.2 Bayesian techniques specifically for GLMMs

The use of a full Bayesian approach for GLMMs was first put forward by Zeger & Karim (1991), soon after the seminal paper by Gelfand & Smith (1990) which first advocated the Gibbs sampling technique. Clayton (1996) also advocated a full Bayesian approach for GLMMs using Gibbs sampling, arguing that the conditional distributions are log-concave and quite amenable to adaptive rejection sampling (section 2.3.2.1).

Zhao *et al.* (2003) reviewed the use of Bayesian MCMC approaches for a broad range of GLMMs, including those where cubic smoothing splines were fitted as random terms in the model (e.g. Verbyla *et al.*, 1999), as well as GLMMs which modelled spatial or temporal correlation patterns, which involve non-diagonal covariance structures for the random effects. They investigated "off the shelf" products such as `BUGS` (Spiegelhalter *et al.*, 1995) as well as other MCMC approaches, including slice sam-

pling, and concluded that BUGS performed relatively favourably. Browne & Draper (2000, 2006) compared Bayesian approaches empirically to two approximate likelihood techniques, PQL and PQL2, and found that the Bayesian approaches perform better with respect to both reduced estimation bias and improved 95% confidence interval coverage.

2.4 Marginal approaches and other approaches

2.4.1 Marginal approaches

As discussed previously in section 1.4.3.1, a marginal model is an alternative to a "conditional" GLMM for modelling correlation or clustering in non-normal data. This section examines marginal approaches for solving a (conditional) GLMM, not marginal models. Two marginal approaches are investigated, the Maximisation-Expectation (ME) approach and marginal quasi-likelihood (MQL).

2.4.1.1 The Maximisation-Expectation approach

The Maximisation-Expectation (ME) approach (Gilmour, 1983; Gilmour, Anderson & Rae, 1985) is also often referred to as the GAR approach after its authors. The basic idea behind the approach is to form the marginal means and variances (and covariances) and then to use iteratively reweighted least squares (IRLS) as for a standard GLM.

To illustrate the formation of the marginal moments, consider a binary model with probit link where the conditional mean, or probability of success, μ_{ij}^u for data y_{ij} is

$$\mu_{ij}^u = \Phi(\boldsymbol{x}_{ij}^T\boldsymbol{\tau} + u_i), \quad i = 1 \ldots n,\ j = 1 \ldots m,\ u_i \sim N\left(0, \sigma^2\right).$$

The marginal, or observed, mean, μ_{ij}, is analytically tractable with a probit link (but

not with a logit link):

$$\begin{aligned}\mu_{ij} &= \mathrm{E}_u(\mu_{ij}^u) = \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\boldsymbol{x}_{ij}^T\boldsymbol{\tau}+u_i}\frac{1}{\sqrt{2\pi}}e^{-t^2/2}dt\right)\frac{1}{\sqrt{2\pi\sigma_u^2}}e^{-u^2/2\sigma_u^2}du \\ &= \Phi\left(\frac{\boldsymbol{x}_{ij}^T\boldsymbol{\tau}}{\sqrt{1+\sigma_u^2}}\right).\end{aligned}$$

This expression indicates that the marginal mean μ_{ij} is attenuated towards 0.5 in comparison with the conditional mean μ_{ij}^u. The marginal covariance, $\boldsymbol{V}$, is more complicated, but can derived using

$$\begin{aligned}\boldsymbol{V} &= \mathrm{var}(\boldsymbol{y}) = \mathrm{E}_u\left\{\mathrm{var}(\boldsymbol{y}|\boldsymbol{u})\right\} + \mathrm{var}_u\left\{E(\boldsymbol{y}|\boldsymbol{u})\right\} \\ &= \mathrm{E}_u\left\{\mathrm{diag}\left[\mu_{ij}^u(1-\mu_{ij}^u)\right]\right\} + \mathrm{var}_u\left\{\mathrm{vec}\left(\mu_{ij}^u\right)\right\},\end{aligned}$$

as discussed in Trottier (1998). As indicated in Engel *et al.* (1995), this expression is analytically tractable for the probit link – in this case, $\boldsymbol{V}$ will be block-diagonal with non-zero elements being the marginal covariances of observations in the same group, viz.

$$\mathrm{cov}(y_{ij}, y_{ik}) = \boldsymbol{\Phi}_2\left(\lambda\boldsymbol{x}_{ij}^T\boldsymbol{\tau}, \lambda\boldsymbol{x}_{ik}^T\boldsymbol{\tau}; \rho_u\right) - p_i p_j,$$

where $\lambda = \sqrt{1+\sigma_u^2}$, $\rho = \sigma_u^2/\left(\sigma_u^2+1\right)$ and $\boldsymbol{\Phi}_2(a,b,\rho)$ is the CDF of the standard bivariate normal distribution with correlation ρ. Engel *et al.* (1995) also provides a Taylor expression (top of p.20) for the double integral in $\boldsymbol{\Phi}_2\left(\lambda\boldsymbol{x}_{ij}^T\boldsymbol{\tau}, \lambda\boldsymbol{x}_{ik}^T\boldsymbol{\tau}; \rho_u\right)$ when ρ_u^2 is small. Trottier (1998) also suggested a general approximation to the marginal variance

$$\boldsymbol{V} \approx \mathrm{E}\left(\boldsymbol{V}^u\right) + \boldsymbol{D}^{-1}\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{D}^{-1},$$

where $\boldsymbol{V}^u = \mathrm{diag}\left[\mu_{ij}^u(1-\mu_{ij}^u)\right]$ and $\boldsymbol{D} = \partial\boldsymbol{\eta}/\partial\boldsymbol{\mu}\big|_{\boldsymbol{\mu}=\hat{\boldsymbol{\mu}}}$. For other types of GLMMs, the expressions for the marginal moments are generally intractable. Trottier (1998) suggested first approximating the conditional mean and variance by conditional moments from which marginal moments can easily be derived, such as for the probit link.

Once the marginal mean and variance are determined, either exactly or approxi-

mately, the IRLS technique can proceed by forming the working variable at the kth iteration,

$$\boldsymbol{\psi}_i^{(k)} = \hat{\boldsymbol{\eta}}^{(k-1)} + \boldsymbol{D}\left(\boldsymbol{y} - \hat{\boldsymbol{\mu}}^{(k-1)}\right),$$

where $\hat{\boldsymbol{\mu}}^{(k-1)}$ is the estimate of the marginal mean, $\boldsymbol{D} = \partial\boldsymbol{\eta}/\partial\boldsymbol{\mu}|_{\boldsymbol{\mu}=\hat{\boldsymbol{\mu}}^{(k-1)}}$ and $\boldsymbol{\eta}^{(k-1)} = \boldsymbol{X}\boldsymbol{\tau}^{(k-1)}$. Note that $\boldsymbol{D}$ is not, in general, a diagonal matrix. Fitting a weighted linear model

$$\boldsymbol{\psi}^{(k)} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{E}, \tag{2.10}$$

where $\boldsymbol{E} \sim (\boldsymbol{0}, \boldsymbol{D}\boldsymbol{V}\boldsymbol{D}^T)$, results in a weighted generalized least squares solution for $\boldsymbol{\tau}$ at the kth iteration, $\boldsymbol{X}^T\boldsymbol{T}^{-1}\boldsymbol{X}\hat{\boldsymbol{\tau}}^{(k)} = \boldsymbol{X}^T\boldsymbol{T}^{-1}\boldsymbol{\psi}^{(k)}$ where $\boldsymbol{T} = \boldsymbol{D}\boldsymbol{V}\boldsymbol{D}^T$. Gilmour (1983), pp 45-46, showed that this solution for $\boldsymbol{\tau}$ is equivalent to forming Henderson-like mixed model equations for $\boldsymbol{\tau}$ and $\boldsymbol{u}$:

$$\begin{bmatrix} \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z} + \boldsymbol{G}^{-1} \end{bmatrix} \begin{pmatrix} \hat{\boldsymbol{\tau}}^{(i+1)} \\ \tilde{\boldsymbol{u}}^{(i+1)} \end{pmatrix} = \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{\psi}^{(i)} \\ \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{\psi}^{(i)} \end{bmatrix}, \tag{2.11}$$

where $\boldsymbol{W} = \boldsymbol{D}\boldsymbol{E}^{-1}\boldsymbol{D}^T$, $\boldsymbol{E} = \mathrm{E}(\boldsymbol{V}^u)$ and $\boldsymbol{V}^u = \mathrm{diag}\{a_i v(\mu_i^u)\}$. Trottier (1998) also confirmed this solution, arguing that, since $\boldsymbol{V} \approx \boldsymbol{E} + \boldsymbol{D}^{-1}\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{D}^{-1}$ as noted above, it follows that $\boldsymbol{T} = \boldsymbol{D}\boldsymbol{E}\boldsymbol{D}^T + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T$, and so the solution for $\boldsymbol{\tau}$ from (2.10) is also the solution to (2.11). Estimation of the variance parameters $\boldsymbol{\kappa}$ can be performed at each iteration using the REML score equations corresponding to the implicit normal linear model for the working variate, $\boldsymbol{\psi} \sim N(\boldsymbol{X}\boldsymbol{\tau}, \boldsymbol{D}\boldsymbol{V}\boldsymbol{D}^T)$, for instance, as shown in section 3 of Gilmour *et al.* (1985).

Therefore, the updating equations for the ME approach take a similar form to those for PQL. The major difference between the ME and PQL approaches is that the ME approach requires expressions for the marginal moments, whereas PQL does not. Gilmour (1983) argued that the ME method should give more stable estimates than PQL (joint maximisation) when the number of observations per random effect is small, but will perform less well as the number of observations per random effect increases. Gilmour *et al.* (1985) argued that PQL effectively calculates a marginal variance for

the data on the basis that the random effects are known, resulting in underestimation of the variance components with small cluster sizes. Simulation results comparing ME and PQL, demonstrating these effects, may be found in Gilmour *et al.* (1985) and Engel *et al.* (1995).

2.4.1.2 Marginal Quasi-likelihood (MQL)

Marginal quasi-likelihood (MQL) was proposed by Longford (1988) and Goldstein (1991), but also reviewed by Breslow & Clayton (1993), where the name "marginal quasi-likelihood" was first coined. An approximate Taylor series expansion of the data is obtained:

$$\boldsymbol{y} \approx \boldsymbol{\mu}^u + \boldsymbol{e}^* \approx h(\boldsymbol{X}\boldsymbol{\tau}) + h'(\boldsymbol{X}\boldsymbol{\tau})\boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}^* = h(\boldsymbol{X}\boldsymbol{\tau}) + \boldsymbol{e}^{**},$$

where $h = g^{-1}$ is the inverse link function, implying the following marginal moments:

$$\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{y}) \approx h(\boldsymbol{X}\boldsymbol{\tau})$$

and

$$\boldsymbol{V} = \mathrm{var}(\boldsymbol{y}) \approx \boldsymbol{V}^u + \boldsymbol{\Delta}^{-1}\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{\Delta}^{-1},$$

where $\boldsymbol{\Delta} = \mathrm{diag}\{g'(\mu_i)\}$. As for PQL, an iterative process is required. At each iteration, a working variate is formed,

$$\psi_i^{(k)} = \eta_i^{(k-1)} + g'\left(\mu_i^{(k-1)}\right)\left(y_i - \mu_i^{(k-1)}\right),$$

where $\eta_i^{(k-1)} = \boldsymbol{x}_i^T\boldsymbol{\tau}^{(k-1)}$ and $\mu_i^{(k-1)} = g^{-1}(\eta_i^{(k-1)})$, to which a weighted linear mixed model is fitted

$$\psi_i^{(k)} = \boldsymbol{x}_i^T\boldsymbol{\tau} + \boldsymbol{z}_i^T\boldsymbol{u} + e_i,$$

with weights $w_i^{(k-1)} = \left\{\phi a_i v\left(\mu_i^{(k-1)}\right)[g'(\mu_i^{(k-1)})]^2\right\}^{-1}$, or $e_i \sim N\left(0, \{w_i^{(k-1)}\}^{-1}\right)$. This yields updated estimates $\boldsymbol{\tau}^{(k)}$ and $\boldsymbol{u}^{(k)}$, and consequently an updated working variate for the next iteration, as for PQL. This cycle is repeated until convergence.

The use of marginal quasi-likelihood will result in marginal estimates of the regression coefficients $\boldsymbol{\tau}$, as for a GEE (section 1.4.3.1). MQL has been shown to give even more biased estimates of the variance components than PQL for grouped binary data with small group sizes (e.g. Rodriguez & Goldman, 1995, 2001).

Differences between MQL, ME and PQL are discussed in Engel & Keen (1996). All three approaches share a commonality in iteratively fitting a weighted linear mixed model to a working variate. The working variate used for ME and MQL is similar, as both are marginal approaches. However, there are differences in the weights between the two methods, as shown in Engel & Keen (1996) in the case of a probit link.

2.4.2 Non-parametric GLMM – Aitkin (1999)

Aitkin (1999) considered a non-parametric approach, which utilizes the GHQ approximation to the likelihood but makes no distributional assumption about the random effects. For simplicity of exposition, a simple GLMM for data y_{ij} , $i = 1 \dots m$, $j = 1 \dots n_i$ is assumed, where $g(\mu_{ij}) = \boldsymbol{x}_{ij}^T \boldsymbol{\tau} + u_i$, $\boldsymbol{\tau}$ is a vector of fixed effects of length p and $u_i \sim N(0, \gamma_1)$. Letting $\upsilon_i = u_i/\gamma_1 \sim N(0,1)$, the likelihood can be approximated using standard m-point GHQ ,

$$\begin{aligned} \ell(\boldsymbol{\tau}, \gamma_1) &= \log\left(\prod_i \int \left[\prod_i f_{y|u}(y_{ij}; \boldsymbol{\tau}, \gamma_1, \upsilon_i)\right] f_\upsilon(\upsilon_i) d\upsilon_i\right) \\ &\approx \log\left(\prod_i \sum_{q=1}^{m} w_q \prod_j f_{y|u}(y_{ij}; \boldsymbol{\tau}, \gamma_1, \xi_q)\right) \\ &= \sum_i \log\left(\sum_q w_q f_{iq}\right), \end{aligned} \tag{2.12}$$

with weights and nodes w_q and ξ_q respectively and $f_{iq} = \prod_j f_{y|u}(y_{ij}; \boldsymbol{\tau}, \gamma_1, \xi_q)$. The score equations for $\boldsymbol{\tau}$ and γ_1 are weighted sums of standard GLM-like score equations. For instance, for τ_l, $l = 1, \dots, p$, the score is

$$\frac{\partial \ell}{\partial \tau_l} = \sum_i \sum_q w_{iq} s_{iq}(\tau_l),$$

where each GLM-like score equation is

$$s_{iq}(\tau_l) = \partial \log f_{iq}/\partial \tau_l = \sum_j (y_{ij} - \mu_{ijq})\, x_{ijl}/v(\mu_{ijq})g'(\mu_{ijq}),$$

and where $g(\mu_{ijq}) = \boldsymbol{x}_{ij}^T\boldsymbol{\tau} + \gamma\xi_q$ and $v(\cdot)$ is the variance function. The corresponding weights are $w_{iq} = w_q f_{iq}/\sum_r w_r f_{ir}$. An EM algorithm can be used to estimate $\boldsymbol{\tau}$ and γ, alternating between the estimation of the weights w_{iq} and the estimation of $\boldsymbol{\tau}$ and γ.

Aitkin's innovation is to utilize the GHQ-approximated expression for the likelihood in (2.12), but to let the scaled GHQ nodes $\alpha_q = \gamma\xi_q$, and their corresponding weights w_q, be unknown parameters rather than fixed values. The likelihood now corresponds to that of a discrete mixture model, and can be readily maximised for $\boldsymbol{\tau}$, α_q and w_q for any given number of quadrature points m. To choose the appropriate number of quadrature points m, Aitkin suggests running separate analyses with increasing values of m, $m = 3, 4, 5, \ldots$, and stopping when there is a suitably small change in the deviance between successive values of m.

Aitkin's approach is appropriate where the focus is the estimation of $\boldsymbol{\tau}$ – the variance component γ_1 is no longer estimated, and the ξ_q and w_q are essentially nuisance parameters. His approach is readily applicable to GLMMs with either independent nested random effects or specific types of correlated nested random effects such as random coefficient models. However, Aitkin acknowledges that the approach would require more methodological development to apply it to other types of GLMMs, such as those with spatially or temporally correlated random effects.

2.4.3 Modified EM approach – Steele (1996)

Steele (1996) outlined a modified EM algorithm where the E-step, $\int \ell_J f_{U|Y} d\boldsymbol{u}$, is approximated using a version of the Laplace approximation developed by Tierney *et al.* (1989). He noted that his first-order approximation yields the same score equations for $\boldsymbol{\tau}$ and $\boldsymbol{u}$ when a canonical link is used.

2.5 Discussion

A number of alternative approaches for estimation in GLMMs have been discussed in this chapter. The two approximate likelihood approaches described above, PQL and HGLM approaches, appear to be the most useful and flexible approaches for fitting a broad range of GLMMs at present. The other approaches discussed in this chapter are either limited in the possible GLMMs they can fit, or appear to be potentially computationally intensive. For instance, numerical approaches, such as GHQ, are effectively limited to the subset of GLMMs involving nested random terms. Also, approaches based on Monte Carlo methods, including Bayesian approaches, are widely considered to be very computationally intensive.

As indicated above, approximate likelihood approaches are well-known to suffer from estimation bias problems in some cases, especially PQL. In the next chapter, the estimation biases for PQL will be explored, in order to try to determine under what conditions the estimation biases will be a significant problem for inference. In addition, other inferential issues using PQL, such as hypothesis testing, will also be examined. In Chapter 4, the HGLM approach of Lee & Nelder (2001) will be explored, and compared against PQL, especially with regard to estimation bias.

Chapter 3

The use of PQL for GLMMs

This chapter investigates the use of penalized quasi-likelihood (PQL) for GLMMs. PQL is the most well-known approximate likelihood approach for GLMMs. It is considered to be a computationally efficient way of fitting a wide variety of GLMMs. The biggest known problem with the PQL approach, highlighted in previous literature, is the potential for large estimation biases for some GLMMs, such as for binary grouped data with small group sizes. However, in GLMMs with general designs, it is still not entirely clear under what conditions PQL can reliably be used in practice. To help determine some guidelines in this area, the studies in this chapter explore the magnitude of the estimation biases for a variety of designs, and with changing design parameters, for binary and Poisson data. The chapter also looks at the issue of hypothesis testing of variance components and fixed effects when using PQL.

3.1 Factors affecting estimation bias

3.1.1 Background

Large estimation biases for PQL when fitting some GLMMs have been reported in previous studies in the literature. These previous studies mainly considered binary GLMMs for grouped data where the group size is relatively small. Two of the most

important papers investigating PQL estimation bias are Breslow & Lin (1995) (and sequel, Lin & Breslow (1996)), who devised an approximate correction (section 2.1.1.2), and Rodriguez & Goldman (2001), who demonstrated how large the PQL estimation biases can be for nested two-way (or "three level") binary data with small group sizes at each level of classification. These and other studies showed that estimation biases can be especially large (over 50%) for the variance parameter estimators. However, under other conditions, the PQL approach has been reported to provide adequate estimators, giving similar results to GHQ and Bayesian approaches, such as in the examples in Breslow & Clayton (1993).

Despite the knowledge that PQL estimation biases will be severe for binary grouped data with small group sizes, it is not immediately clear under what conditions PQL estimation biases will be severe for GLMMs in general. PQL is still attractive to practitioners due to its general availability, ease of use and computational efficiency. It would be desirable to be able to offer some guidelines on when PQL can be reliably used, that is, for what types of data, and designs, will the estimation biases be small enough to be of no consequence. Breslow (2003) tried to offer a general rule of thumb. He suggested that PQL will fail to provide reliable estimators when the conditional PDF of the data given the random effects, $f_{Y|U}$, is far from normality. His rule of thumb is therefore similar to the rule of thumb used to judge the adequacy of the χ^2 distributional approximation of the goodness of fit statistic for contingency table data. For binomial data, he suggested that, for PQL to provide adequate estimators, both the (conditional) expected numbers of successes and failures should be generally greater than 5, and for Poisson data he likewise suggested that the conditional means should be generally greater than 5.

Lee & Nelder (1996) argued that PQL-like estimators are consistent when the diagonal elements of the inverse of the second derivative of the joint likelihood with respect to $\boldsymbol{u}$, $\left(\partial l_J/\partial \boldsymbol{u}\partial \boldsymbol{u}^T\right)^{-1}$, are $O(n^{-1})$, where n is the number of observations in the dataset. This condition corresponds to the situation when the sample size n goes to infinity, but the number of random effects b remains constant. However, Jiang (1998),

and discussants to Lee & Nelder (1996), indicated that PQL-like estimators were not consistent in cases where b was also going to infinity.

3.1.2 Aims

The assertions of Breslow (2003) and Lee & Nelder (1996) offer an explanation of the PQL estimation bias problems for binary grouped data with small group sizes. But neither assertion has been explored for GLMMs in general.

Critical examination of these assertions, or development of a new rule of thumb, could be done either analytically or empirically, or using a combination of these. The use of an analytical approach appears to be unpromising. As indicated in section 2.1, the PQL approach does not maximise a fixed criterion such as a log-likelihood, or even an approximate log-likelihood (such as the Laplace approximation to the likelihood), and so it is difficult to find an analytic expression for the bias. Attempting to use the iterative formulae quickly becomes unmanageable, as seen in Engel (1998) for a simple one-way classification. Breslow & Lin (1995) and Lin & Breslow (1996) attempted to determine general analytic expressions for the bias, but these were based on approximate expansions around $\boldsymbol{\gamma} = \mathbf{0}$ and for designs where the group sizes were small. Lin & Breslow (1996) demonstrated that their correction formulae, based on these analytic expressions, worked poorly in some cases. In any case, if a general basis or rule of thumb could be devised by analytic means, it would be desirable to validate such a rule using Monte Carlo simulation studies.

It was therefore decided to proceed on an empirical basis, using Monte Carlo simulation to examine the estimation biases for a range of simple designs for binary and Poisson data. For each design, the design parameters were varied to determine which design parameters had the most influence on the estimation bias. If patterns emerged in the dependence of the estimation biases on the design parameters across designs, this would allow the extrapolation of these patterns to more complex GLMMs, and provide a basis for establishing general guidelines for when PQL is adequate or not. To our best knowledge, such an empirical approach for exploring PQL estimation

biases has not been performed before: previous simulation studies in the literature have dealt with specific designs and specific parameter values, rather than exploring a variety of designs with varying parameters as done here.

The designs considered range from the very simplest, the one-way classification, to designs involving nested, crossed or correlated random effects. In each design, the parameters of the model are varied in a factorial arrangement in order to allow for the detection of possible interaction effects on the estimation bias. For instance, in the one-way classification model (equation 3.1.4.1), the design parameters included the number of groups (denoted b_g), the number of observations per group (denoted m_g) and the variance component relating to group effects (denoted γ_1). The choice of design parameter values was, in part, guided by previous literature, in that some parameter values were chosen to be those which were known to result in large PQL estimation biases. For instance, large numbers of groups with relatively small group sizes were chosen, given the previous literature findings on high estimation bias under these conditions. Given the use of a factorial structure, the results of each of these simulation studies could have been analysed using a formal analysis of variance (ANOVA): however, the majority of interactions were found to be statistically significant, up to five-way interactions. Attempting to describe all the statistically significant interactions would have obscured the salient factors affecting the estimation bias. Therefore, for each design, an exploratory analysis of the simulation results was conducted to highlight the main contributing factors to the magnitude of the estimation biases.

It is important to note here that we are considering estimation bias in a relative sense. Therefore, the bias for each parameter estimator was calculated relative to the magnitude of the parameter, that is, the bias for parameter θ was estimated as

$$\widehat{\text{Bias for } \hat{\theta}_T} = \frac{\left(\overline{\hat{\theta}}_T - \theta_T\right)}{\theta_T},$$

where θ_T is the true value for parameter θ and $\overline{\hat{\theta}}_T$ is the average estimate for θ, for a given subset of simulation datasets from the total set of datasets where $\theta = \theta_T$.

The use of relative estimation bias appears to be intuitively sensible, since it reflects the importance of the estimation bias in a practical sense: a relative estimation bias of less than 5% in absolute value might be considered ignorable, say, but a relative estimation bias of over 50% in absolute value would definitely be considered severe, no matter how small the parameter's true value was.

3.1.3 Methodology

Some of the methodological details in common to all the Monte Carlo simulations are as follows.

In each design, the data either represents binary or Poisson data, with the corresponding link function g being either the logit and logarithmic function respectively. The simulated datasets were analysed according to the same model that generated the data. For all simulations, the linear mixed model package ASReml version 2.0 (Gilmour *et al.*, 2006) was used. Results for simulated datasets where ASReml reported non-convergence or singularities were omitted in the calculation of estimation biases. These represented a very small fraction (<0.1%) of the total number of datasets in most cases. For each design, 200 simulations were conducted for each combination of simulation parameters in each model. The starting values for variance components were either set to be 0.1 or 0.00001 for the binary and Poisson models respectively in each design, with the latter chosen to reduce divergence problems for the Poisson models. However, apart from this divergence problem, preliminary simulations suggested that the estimates were not affected by the choice of starting values for any of these models.

It should be noted that estimates of variance parameters, apart from the correlation parameter γ_ρ in the correlated and random coefficient models, were constrained to be positive, since this was the default setting of ASReml. Even though many current linear mixed model and GLMM implementations like ASReml do restrict variance components to take positive values by default, it may have been more correct, in hindsight, to have allowed at least some of the estimated variance components to

take any negative or positive value. For instance, Nelder (1954) discussed the interpretation of negative variance components in the context of the analysis of designed experiments, making the case that estimated variance components should always be allowed to take negative values. (However, as Jiang pointed out in the discussion to Lee & Nelder (2004), the Genstat HGLM implementation, which Nelder himself was involved in developing, only allows positive variance components, at least at present.) Therefore negative biases of the variance components reported in the simulation studies below may be considered under-estimates of the potential negative biases, although the magnitude of this underestimation is probably not large. It is important to note that when allowing for negative variance component estimates, the BLUPs for the associated random effects do not exist. Another issue with the design-based approach in Nelder (1954, 1965b,a) is that it implies an analysis of the response as a normal random variable. It is unclear how to extend Nelder's design-based approach to allow for the analysis of the response as a non-normal random variable, such as in a GLMM.

Also, it should be noted that the estimation biases of the variance estimates (i.e. $\hat{\gamma}$) are investigated, rather than the corresponding standard deviations $\sqrt{\hat{\gamma}}$, despite some authors who prefer the latter, such as Yun & Lee (2004), who argued that the distribution of the standard deviation estimate $\sqrt{\hat{\gamma}}$ is more symmetrical.

To conduct the simulations, Fortran 90 code was written for each design, which, for each combination of simulation parameter values, generated a simulated dataset, called ASReml to analyse it, and then parsed the ASReml output to obtain the PQL estimates, which it then wrote to an output file. All random variables were generated by calling routines in the Fortran 90 version of the `randlib` library, freely available on the internet[1]. The `barossa` machine on the NSW academic ac3 system[2] was used remotely to perform most of the simulations.

[1] `http://biostatistics.mdanderson.org/SoftwareDownload/`
[2] `http://www.ac3.edu.au/`

3.1.4 Designs with independent random effects

The designs considered in this section all have independent random effects, that is, the variance matrix of the random effects, $\boldsymbol{G}(\boldsymbol{\gamma})$, is diagonal.

3.1.4.1 One-way classification

We begin with the simplest GLMM, the one-way classification, and consider scenarios where the number of observations per group is small. Let y_{ij}, $i = 1, \ldots, b_g$, $j = 1, \ldots, m_g$, (the sub-script g here indicating "group") represent the jth observation from the ith group, generated from a model for the conditional mean $\mu_{ij}^u = E(y_{ij}|u_i)$ given by

$$g(\mu_{ij}^u) = \tau_0 + \tau_1 x_{1ij} + \tau_2 x_{2ij} + u_i, \tag{3.1}$$

where $u_i \sim N(0, \gamma_1)$. The covariates $x_{1ij} = 2(j-1)/(m_g - 1) - 1$ and $x_{2ij} = 2(i-1)/(b_g - 1) - 1$ take values between -1 and 1, and their values vary within and between groups respectively. The values of the simulation parameters used in the study are given in Table 3.1.

Parameter	*Binary model*	*Poisson model*
b_g	50, 100, 200, 500, 1000	...
m_g	2, 4, 8, 16, 32, 64	...
γ_1	0.25, 1, 4	...
τ_0	0, 2	0.1, 1
τ_1, τ_2	0, 2	0, 1

Table 3.1: Values of the simulation parameters used for binary and Poisson models for the one-way classification study (3.1). The use of ... in the 3rd (Poisson model) column indicates the same values were used as for the binary model.

The rationale for the parameter settings given in Table 3.1 was as follows. The values for b_g and m_g were chosen, on the basis of previous literature, as those in which high levels of PQL estimation bias would be expected, as indicated earlier. The values of γ_1 chosen here represent data with relatively small, medium and large variation between groups. For the binary model, the two values of τ_i, $i = 0, 1, 2$, were selected sufficiently widely apart to reasonably determine whether changing that parameter

had any effect on the bias. For the Poisson model, the values of τ_0 were selected to represent sparse Poisson data with low average rates ($\tau_0 = 0.1$ rather than 0, was selected so that relative bias could be calculated). Without loss of generality, only positive values of τ_i, $i = 0, 1, 2$, were examined. That is, there would be no reason to expect, because of the symmetry around 0.5 on the probability scale, that the bias for $\tau_0 = -2$ for the binary model should not be exactly the same as that for $\tau_0 = 2$. For Poisson data, selecting $\tau_0 < 0$ would have resulted in Poisson data which was unrealistically sparse. For $i = 1$ and 2, the choice of $\tau_i = -2$ for either binary or Poisson models would have been resulted in data from the same distribution as for $\tau_i = 2$, because the covariates x_{1ij} and x_{2ij} were centered around 0.

For the binary model, the average estimation biases for all parameter estimators were, almost always, negative. These estimation biases varied considerably with the group size m_g and the variance component γ_1, with the negative bias increasing with increasing γ_1 but decreasing with increasing m_g (Figure 3.1). For the biases of $\hat{\tau}_i$, $i = 0, 1, 2$, a multiplicative interaction is also apparent between m_g and γ_1 – that is, there was little effect of m_g on the biases when $\gamma_1 = 0.25$, but a large decrease in the negative biases with m_g when $\gamma_1 = 4$. Note that the bias for each $\hat{\tau}_i$ is shown only where $\tau_i = 2$, since there was no apparent bias of $\hat{\tau}_i$ when $\tau_i = 0$. The bias for $\hat{\gamma}_1$ was much greater in magnitude than for each $\hat{\tau}_i$ across the range of parameter values. This was to be expected, since the $\hat{\tau}_i$ are naturally constrained to be at least as great as the GLM or "marginal" estimates (where $\hat{\gamma}_1 = 0$), whereas values of $\hat{\gamma}_1$ were only constrained to be greater than 0. The absolute biases for $\hat{\tau}_i$ were less than 5% when $\gamma_1 = 0.25$, even for small m_g. There were similar biases for each $\hat{\tau}_i$, $i = 0, 1, 2$, when $m_g = 2$, but the bias for $\hat{\tau}_1$ decreased more rapidly with increasing m_g than did the biases for $\hat{\tau}_0$ or $\hat{\tau}_2$. The effects of other simulation parameters on the biases were either less pronounced, or only occurred under certain combinations of the other simulation parameter values. For instance, the biases for $\hat{\gamma}_1$ tended to increase with the number of groups b_g, but this was only really noticeable where both m_g and γ_1 were small, and was relatively modest in other cases (Figure 3.2). Two other influences on the biases noted here are an increase in the bias with τ_0 (for instance, by -4% for $\hat{\gamma}_1$ from

$\tau_0 = 0$ to $\tau_0 = 2$) and with τ_1 where $m_g = 2$ and $\tau_0 = 2$ (Figure 3.3). Note that the biases for γ_1 are positive for $\gamma_1 = 0.25$ and $m_g = 2$ in Figure 3.2 and 3.3.

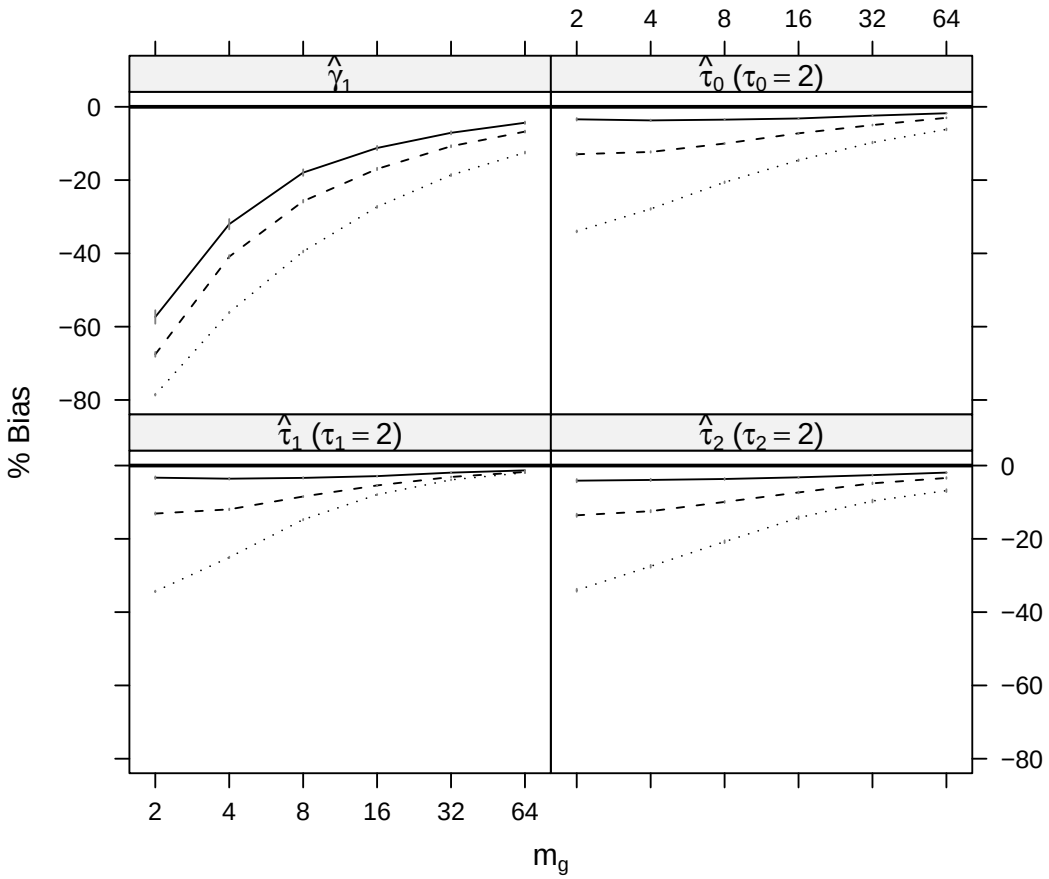
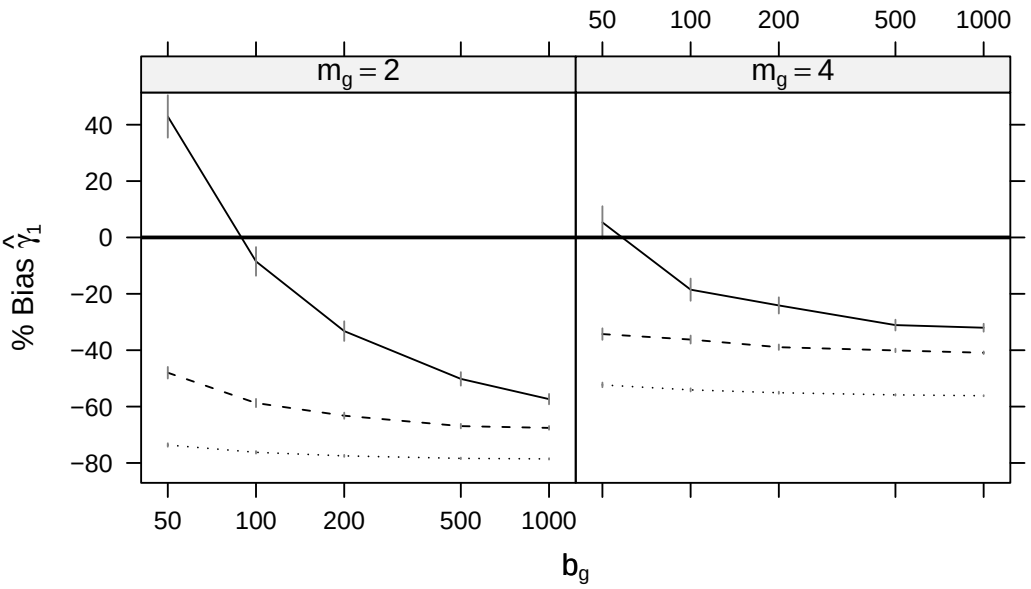


Figure 3.1: Biases in the binary one-way classification model (3.1): the interactions between the effects of m_g and γ_1 on the biases for $\hat{\gamma}_1$ and $\hat{\tau}_i$ (when τ_i=2), $i = 0, 1, 2$. (γ_1=0.25: solid, 1: dashed, 4: dotted). Note that these biases are calculated where b_g=1000. (Error bars are ±2SE.)

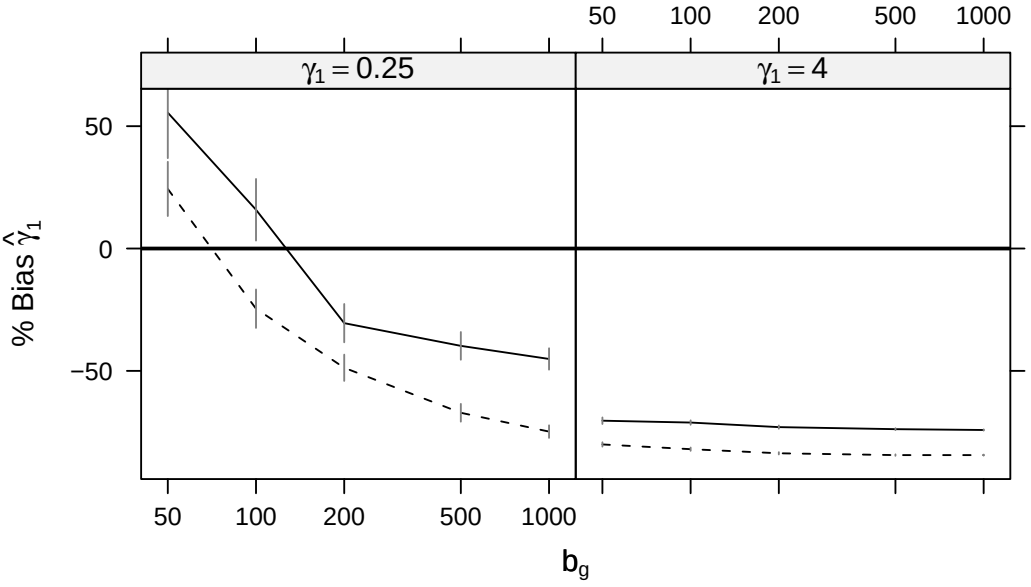
For the Poisson model, there were sizeable estimation biases for $\hat{\gamma}_1$, $\hat{\tau}_0$ and $\hat{\tau}_2$, but little, if any, evidence of bias for $\hat{\tau}_1$ (Figure 3.4). The biases for $\hat{\gamma}_1$ and $\hat{\tau}_2$ were always negative, similar to the biases for these parameters in the binary model, but the bias for $\hat{\tau}_0$ was always positive. As for the binary model, the biases for $\hat{\gamma}_1$, $\hat{\tau}_0$ and $\hat{\tau}_2$ varied considerably with m_g and γ_1, increasing with increasing γ_1 and decreasing with increasing m_g. However, the bias also varied considerably with the intercept τ_0, decreasing with increasing τ_0.

In summary, this simulation study confirmed that the estimation biases for GLMMs



(a) Interaction of n and σ on the biases for σ^2 when m=2 and 4.

Figure 3.2: Biases in the binary one-way classification model (3.1): the interaction between the effects of b_g and γ_1 on the biases for $\hat{\gamma}_1$ for $m_g = 2$ and $m_g = 4$ respectively. (γ_1=0.25: solid, 1: dashed, 4: dotted). (Error bars are ±2SE.)



(a) Effect of τ_1 on the biases for σ^2 when $m = 2$ and $\tau_0 = 2$ for $\sigma = 0.5$ and 2 respectively (

Figure 3.3: Biases in the binary one-way classification model (3.1): the interaction between the effects of τ_1, b_g and γ_1 on the bias of $\hat{\gamma}_1$ at $m_g = 2$ and $\tau_0 = 2$. ($\tau_1 = 0$: solid, 2: dashed). (Error bars are ±2SE.)

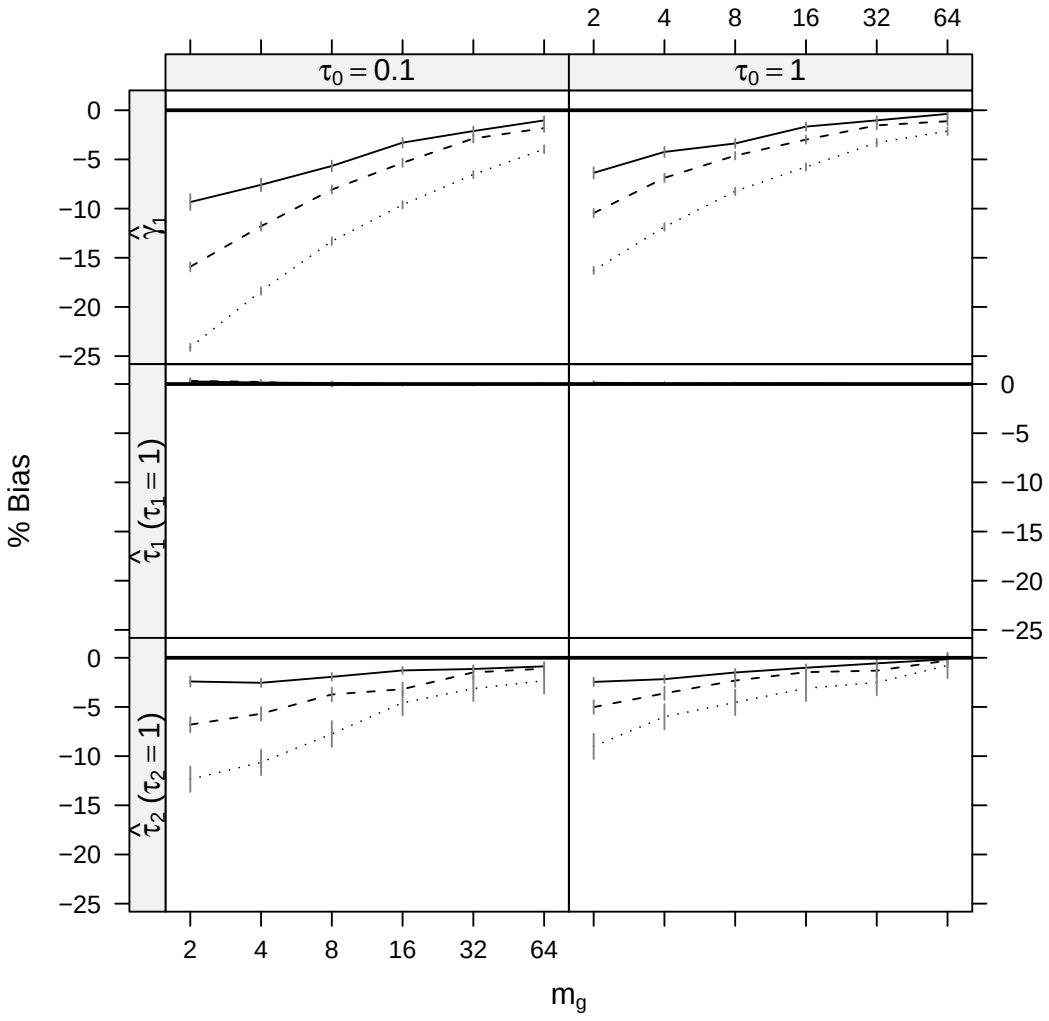


Figure 3.4: Biases in the Poisson one-way classification model (3.1): the interaction between the effects of m_g, γ_1 and τ_0 on the biases for $\hat{\gamma}_1$, $\hat{\tau}_1$ ($\tau_1 = 1$) and $\hat{\tau}_2$ ($\tau_2 = 1$). (γ_1=0.25: solid, 1: dashed, 4: dotted). (Error bars are $\pm$2SE.)

in simple grouped data increase markedly with decreasing group size m_g for both binary and sparse Poisson models. As already discussed, this result has been shown empirically in the literature for binary data, but not for sparse Poisson data, although the Breslow (2003) rule of thumb anticipated higher levels of estimation bias for sparse Poisson data..

The study also demonstrated that the estimation biases also increase markedly with the between group heterogeneity, that is, with the magnitude of the true variance component γ_1, for both binary and Poisson models. This effect of γ_1 on the biases

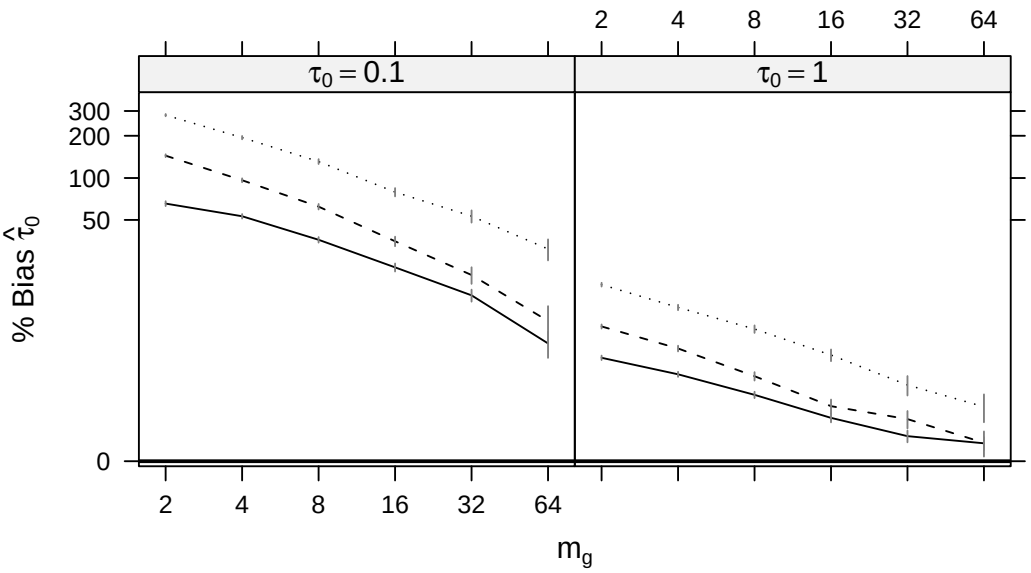


Figure 3.5: Biases in the Poisson one-way classification model (3.1): the interaction between the effects of m_g, γ_1 and τ_0 on the biases for $\hat{\tau}_0$. (γ_1=0.25: solid, 1: dashed, 4: dotted). (Error bars are ±2SE.) (Note: log scale used for y-axis.)

has been less well identified in the literature than the effect of group size m_g. This effect also suggests that Breslow (2003)'s hypothesis, concerning the adequacy of PQL, may be deficient. Breslow's hypothesis is based solely upon the conditional PDF of the data given the random effects, $f_{Y|U}$, that is, he asserts that PQL will fail when $f_{Y|U}$ is far from normality. However, the dependence of the estimation biases on γ_1 in this study shows that the magnitude of the estimation biases, and the determination of when PQL will "fail", cannot be predicted on the conditional PDF $f_{Y|U}$ alone, since $f_{Y|U}$ is not a function of $\boldsymbol{\gamma}$. This issue is discussed further below. Similarly, both the hypothesis of Lee & Nelder (1996) and the corrected PQL technique of Breslow & Lin (1995) and Lin & Breslow (1996) also ignore the potential influence of $\boldsymbol{\gamma}$ on the biases. For the corrected PQL technique, the correction factors given for the variance parameters are independent of $\hat{\boldsymbol{\gamma}}$, and therefore do not take into account increasing estimation bias with $\boldsymbol{\gamma}$ – see, for instance, the formula for a single variance component γ_1, reproduced in section 2.1.1.2.

For the Poisson model, the size of the intercept τ_0 also contributed strongly to the magnitude of the estimation bias, which was to be expected since it reflects the spar-

sity of the Poisson data and is consistent with Breslow's hypothesis. It is interesting that, despite the sparcity of the Poisson data examined here, there is no indication of bias for the estimate of the coefficient $\hat{\beta}_1$, which corresponds to the covariate x_{1ij} which changed within groups. Lower levels of bias were also observed for $\hat{\beta}_1$ than either $\hat{\beta}_0$ or $\hat{\beta}_2$ in the binary model.

Testing Breslow's hypothesis We consider Breslow's hypothesis in the light of the above results, restricting discussion to the binary case for simplicity.

As noted above, the hypothesis of Breslow (2003) concerning the adequacy of PQL, is that PQL will fail when the conditional distribution $f_{Y|U}$ is far from normality. Since $f_{Y|U}$is not a function of γ_1, this hypothesis does not account for the effects of γ_1 on the magnitude of the estimation bias demonstrated in the one-way classification study above. However, changes to the variance component γ_1 will affect the marginal, or observed, distribution of the data. Increasing γ_1 should increase the marginal probability of having a more "extreme" observation, that is, the probability of observing a low number of successes or failures in a group. One could therefore argue that increasing γ_1 increases the estimation bias because it increases the probability of observing a low number of successes and failures in a group. Breslow's rule of thumb, that PQL is adequate when the expected number of successes or failures is generally greater than 5, may consequently have some justification.

The aim of this section is to test whether Breslow's rule of thumb is adequate, by comparing the probabilities of observing a low number of successes or failures in a group with the corresponding estimation biases for some selected simulation parameter values. Consider the binary one-way classification model (3.1), but with no covariates, where

$$\text{logit}(\mu_{ij}^u) = \tau_0 + u_i. \tag{3.2}$$

Since the number of successes $Y_{i.} = \sum_j Y_{ij}$ in a group is conditionally binomial,

$$Y_{i.}|u_i \sim \text{Binomial}\left(m_g, \mu_i^u\right),$$

where $\mu_i^u = \text{logit}^{-1}(\tau_0 + u_i)$, the marginal probability of having less than c successes in a group is defined as

$$P\left(Y_{i.} \leq c\right) \quad = \quad \int P\left(Y_{i.} \leq c|u_i\right) f(u_i) du$$

where $P(Y_{i.} \leq c|u_i)$ is the binomial CDF with parameters m_g and μ_i^u, and $f(u_i)$ is a normal PDF where the corresponding distribution has variance γ_1. The probability of less than c failures, $P\left(\{m_g - Y_{i.}\} \leq c\right)$, is defined similarly. Like the GLMM likelihood, both of these probabilities have no closed analytical form. However, they can be approximated using standard GHQ, that is,

$$\int P\left(Y_{i.} \leq c|u_i\right) P(u_i) du = \sum_{j=1}^{k} w_{j,k} P\left(Y_{i.} \leq c|\xi_{j,k}\right)$$

where k is the number of quadrature points (e.g. 20), and the $w_{j,k}$ and $\xi_{j,k}$ are the scaled weights and nodes for $k-$point GHQ (section 2.2.3). Let $P_c = P(Y_{i.} \leq c) + P\left(\{m_g - Y_{i.}\} \leq c\right)$ denote the probability of having a low number of successes or failures.

Table 3.2 shows the probability P_c where $c = 5$ and the corresponding estimation bias for four combinations of γ_1 and τ_0, $\gamma_1 \times \tau_0 = (1, 4) \times (1, 2)$, where the other parameters are set at $m_g = 32$ and $b_g = 500$. Note that, although the probability of having a low number of successes or failures for $(\gamma_1, \tau_0) = (1, 2)$, $P_c = 0.65$, is higher than for $(\gamma_1, \tau_0) = (4, 1)$, where $P_c = 0.48$, the magnitude of the PQL biases are larger in the latter. Therefore a higher probability of observing a low numbers of successes or failures does not appear to imply a larger PQL bias, and so Breslow's rule of thumb appears to be deficient.

The greater bias experienced when $(\gamma_1, \tau_0) = (4, 1)$, than when $(\gamma_1, \tau_0) = (1, 2)$, must therefore be related to some difference in the marginal distributions of the data, $P(Y_{i.} = c)$, that is not encapsulated by Breslow's rule of thumb. The two marginal distributions are plotted in Figure 3.6. It can be seen that the latter, $(\gamma_1, \tau_0) = (1, 2)$, has the more "normal-like" distribution of the two distributions, in that it has a

distinct "peak" near $c = 30$, whereas the other distribution does not have such a peak.

				PQL Bias (%)	
γ_1	τ_0	$P(Y_{i.} < 5)$	$P(\{m_g - Y_{i.}\} < 5)$	$\hat{\gamma}_1$	$\hat{\tau}_0$
1	1	0.0057	0.30	-9.5±0.2	-4.1±0.1
1	2	0.00028	0.65	-14.8±0.2	-5.2±0.1
4	1	0.088	0.39	-17.8±0.1	-8.9±0.2
4	2	0.032	0.58	-23.6±0.1	-10.2±0.1

Table 3.2: Illustrating the deficiency of Breslow (2003)'s hypothesis using a binary one-way classification model (equation 3.2) for four combinations $\gamma_1 \times \tau_0 = (1, 4) \times (1, 2)$ where $m_g = 32$ and $b_g = 500$.

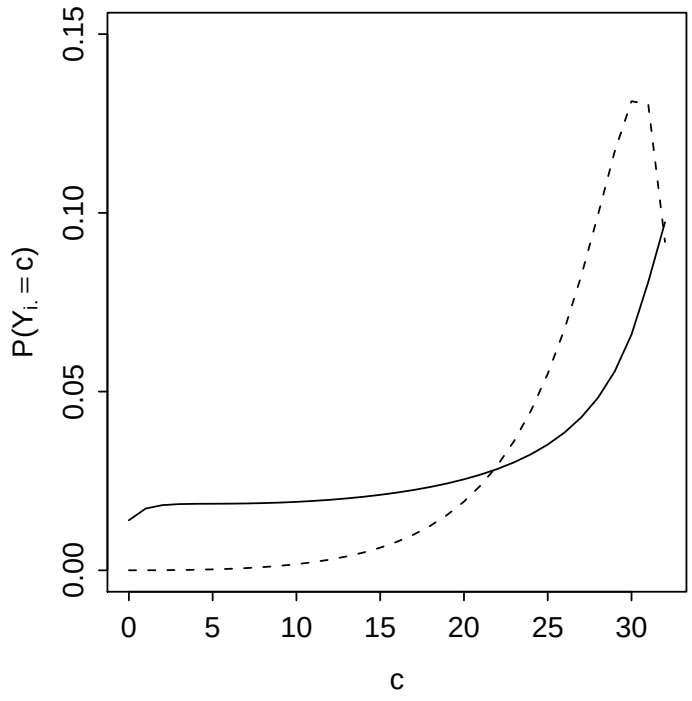


Figure 3.6: Marginal probability distributions $P(Y_{i.} = c)$ where $Y_{i.} = \sum Y_{ij}$ in model 3.2 for $(\gamma_1, \tau_0) = (4, 1)$ (solid line) and $(\gamma_1, \tau_0) = (1, 2)$ (dotted line).

3.1.4.2 Nested two-way classification

A nested two-way classification is now considered. Data y_{ijk}, $i = 1, \ldots, b_g$, $j = 1, \ldots, m_g$, $k = 1, \ldots, m_s$, were generated, and analysed, according to the following

model for the conditional mean $\mu^u_{ijk} = E(y_{ijk}|u_{1i}, u_{2ij})$,

$$g(\mu^u_{ijk}) \quad = \quad \tau_0 + u_{1i} + u_{2ij}, \qquad u_{1i} \sim N(0, \gamma_1), \qquad u_{2ij} \sim N(0, \gamma_2), \qquad (3.3)$$

with parameter values given in Table 3.3.

Parameter	*Binary model*	*Poisson model*
b_g	50, 100, 200, 500	...
m_g, m_s	2, 4, 8, 16	...
γ_1, γ_2	0, 4	...
τ_0	0, 2	0.1, 1

Table 3.3: Values of the simulation parameters used for the nested two-way classification study (3.3). The use of ... in the 3rd (Poisson model) column indicates the same values were used as for the binary model.

For the binary model, the average estimation biases for $\hat{\gamma}_1$,$\hat{\gamma}_2$ and $\hat{\tau}_0$ are only shown where $\gamma_1 = 2$, $\gamma_2 = 2$ and $\tau_0 = 2$ respectively. As for the one-way classification, negative estimation biases were observed for all parameters. The magnitude of the negative bias for $\hat{\gamma}_1$ increased with increasing γ_1 but decreased with increasing m_g and m_s (Figure 3.7), which was consistent with the one-way classification results above. The bias for $\hat{\gamma}_2$, when $\gamma_1 = 0$, was of a similar magnitude to the bias for γ_1 in the one-way classification for each value of m_s (corresponding to m_g for the one-way classification). However, the bias for $\hat{\gamma}_2$ was larger when $\gamma_1 = 2$ than when $\gamma_1 = 0$. The magnitude of the negative bias for $\hat{\tau}_0$ (τ_0=2) increased more rapidly with increasing γ_2 than with increasing γ_1.

For the Poisson model, increasing γ_2 increased the magnitude of the estimation bias for $\hat{\gamma}_1$, however, in contrast, increasing γ_1 tended to reduce the magnitude of the bias for $\hat{\gamma}_2$ (Figure 3.8). Both of these effects were similar for both $\tau_0 = 0.1$ and $\tau_0 = 1$. The magnitude of the bias for $\hat{\tau}_0$ increased more with increasing γ_2 than increasing γ_1 (Figure 3.9).

Even greater negative biases were observed in this study than in the one-way classification models, which is perhaps to be expected since there are two sources of heterogeneity, not one. The results of this simulation study were consistent with the

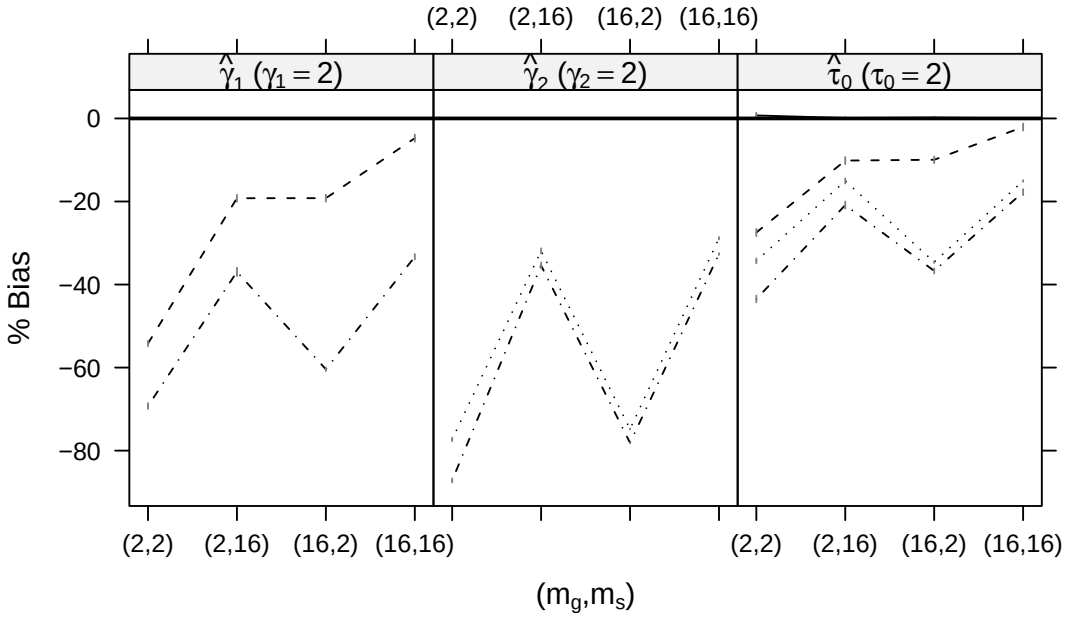


Figure 3.7: Biases in the binary nested two-way model (3.3) : interactions of the effects of m_g, m_s, γ_1, γ_2 on the biases for $\hat{\gamma}_1$, $\hat{\gamma}_2$ and $\hat{\tau}_0$. ($(\gamma_1, \gamma_2) = (4, 4)$: dot-dashed, $(4, 0)$: dashed; $(0, 4)$: dotted). (Error bars are $\pm$2SE.)

results of the previous one-way classification – the group sizes m_s (for γ_2) and $m_s m_g$ (for γ_1) and the variance components γ_1 and γ_2 were the main factors determining the magnitude of the estimation bias. The magnitude of the variance component at the lower level, γ_2, had more influence on the magnitude of the biases than that at the higher level of aggregation, γ_1. This also is to be expected since the effective group size m_s for γ_2 is lower than the group size $m_g m_s$ for γ_1. Similarly, increasing m_s reduced the magnitude of the biases more than increasing m_g.

3.1.4.3 Crossed two-way classification

A crossed two-way classification is now considered. Data y_{ijk} was generated, and analysed, according to the following model for the conditional mean $\mu^u_{ijk} = E(y_{ijk}|u_{1i}, u_{2j})$:

$$g(\mu^u_{ijk}) = \tau_0 + u_{1i} + u_{2j}, \qquad i = 1 \ldots b_1,\ j = 1 \ldots b_2,\ k = 1 \ldots m_s \qquad (3.4)$$
$$u_{1i} \sim N(0, \gamma_1),\ u_{2j} \sim N(0, \gamma_2).$$

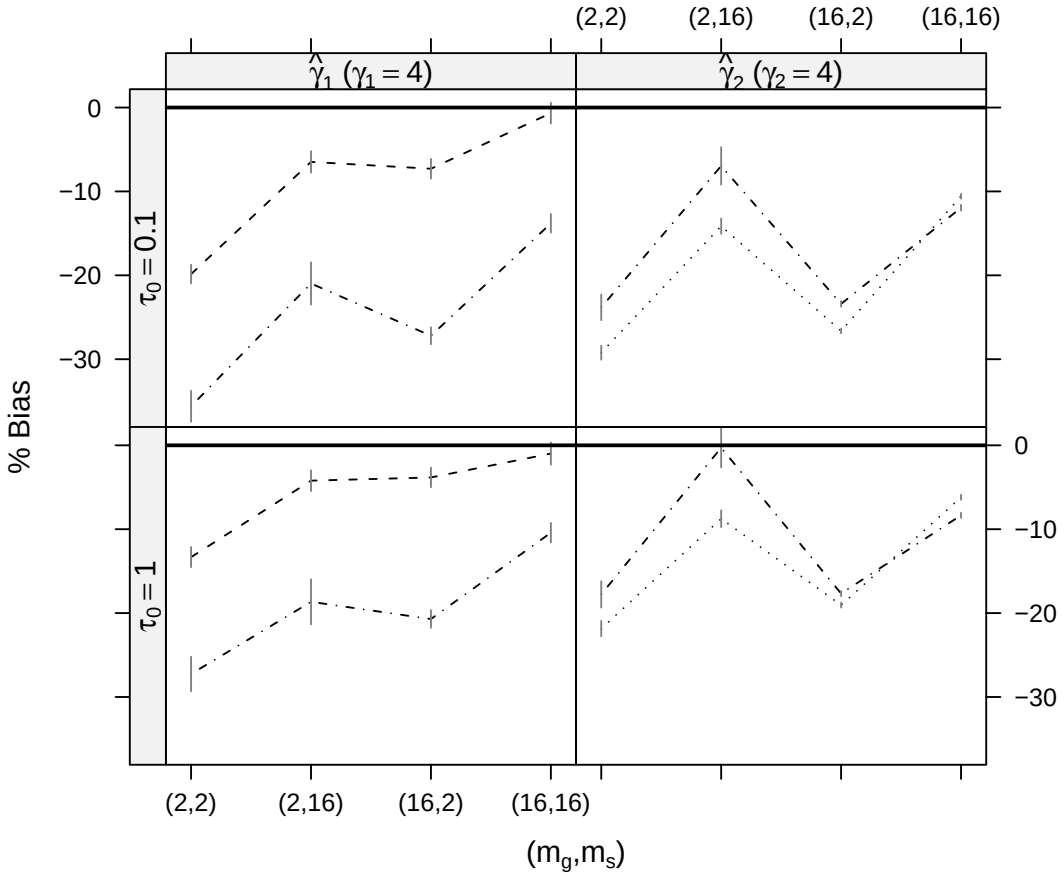


Figure 3.8: Biases for the Poisson nested two-way model (3.3): interactions of the effects of m_g, m_s, γ_1, γ_2 and τ_0 on the biases for $\hat{\gamma}_1$ and $\hat{\gamma}_2$. $((\gamma_1, \gamma_2) = (4, 4)$: dot-dashed, $(4, 0)$: dashed; $(0, 4)$: dotted). (Error bars are $\pm$2SE.)

The simulation parameter values used are given in Table 3.4. The values of the parameters b_1 and b_2 were chosen so that the variance component γ_1 represents variation between groups which are large in number but small in size and, conversely, γ_2 represents variation between groups which are small in number but large in size.

Parameter	*Binary model*	*Poisson model*
b_1	50, 100, 200	...
b_2	3, 10, 25	...
m_s	1, 2, 4	...
γ_1, γ_2	0, 4	...
τ_0	0, 2	0.1, 1

Table 3.4: Values of the simulation parameters used for the crossed two-way classification study (3.3). The use of ... in the 3rd (Poisson model) column indicates the same values were used as for the binary model.

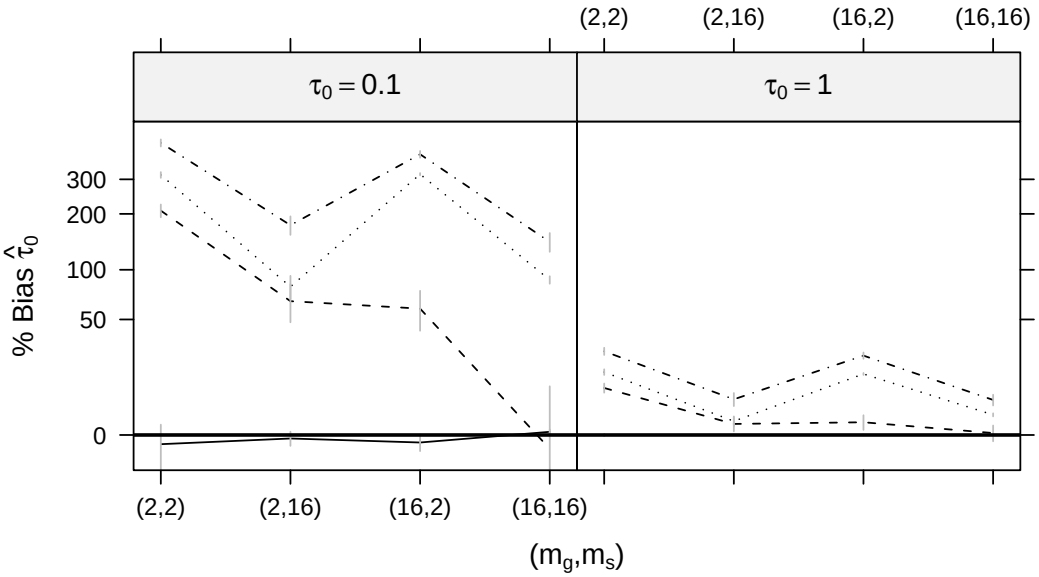


Figure 3.9: Biases for the Poisson nested two-way model (3.3): Interactions of the effects of m_g, m_s, γ_1, γ_2 and τ_0 on the biases for $\hat{\tau}_0$. ($(\gamma_1, \gamma_2) = (4, 4)$: dot-dashed, $(4, 0)$: dashed; $(0, 4)$: dotted). (Error bars are $\pm$2SE.)

For the binary model, the biases for $\hat{\gamma}_1$, where $\gamma_1 = 4$, were consistent with the one-way classification results (Figure 3.10). For both $\hat{\gamma}_2$ and $\hat{\tau}_0$, where $\gamma_2 = 4$ and $\tau_0 = 2$ respectively, there was considerable negative bias when $\gamma_1 = 4$ but minimal bias when $\gamma_1 = 0$. For the Poisson model, the biases for $\hat{\gamma}_1$ were consistent with the Poisson one-way classification model (Figure 3.11). There was no evidence of bias for $\hat{\gamma}_2$ (Figure 3.11), but there was positive bias for $\hat{\tau}_0$ when $\gamma_1 = 4$ (Figure 3.12).

As for the nested two-way classification, the biases in this study are consistent with those from the one-way classification. The effects on the estimation biases of the group size and variance component corresponding to the smaller groups, that is, b_2 and γ_1, are stronger than the ones corresponding to the larger groups (b_1 and γ_2).

3.1.4.4 Designs with many fixed effects for binary data

For all the previous designs above with binary data, negative biases are consistently seen for each of the variance parameters γ_i. A common feature of all these designs is a limited number of fixed effects. However, Engel & Buist (1998) reported positive bias for the variance parameter in their simulation study, which used a design with

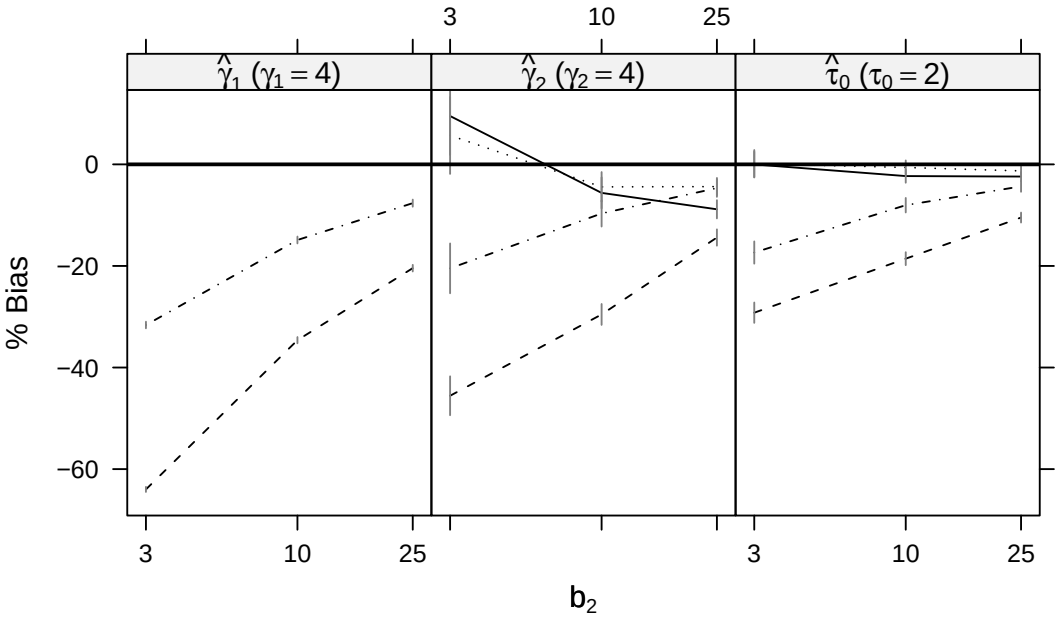


Figure 3.10: Biases for the binary crossed two-way model (3.4): interactions between the effects of b_2, m_s, γ_1, and γ_2 on the biases for $\hat{\gamma}_1$ ($\gamma_1 = 4$), $\hat{\gamma}_2$ ($\gamma_2 = 4$) and $\hat{\tau}_0$ ($\tau_0 = 2$). ($(\gamma_1, m_s) = (4,4)$: dot-dashed, $(4,1)$: dashed, $(1,4)$: dotted, $(1,1)$: solid). (Error bars are $\pm$2SE.)

many fixed effects. They claimed that such a design is common, for instance, in the analysis of breeding trials. An auxiliary simulation study was performed to validate their finding.

The following two designs were examined:

- a crossed two-way design where data y_{ijk}, $i = 1 \dots p,\ j = 1 \dots b,\ k = 1 \dots m_s$, is generated, and analysed, from the following model for $\mu^u_{ijk} = E(y_{ijk}|u_j)$,

$$\text{logit}(\mu^u_{ijk}) = \tau_0 + \tau_i + u_j, \tag{3.5}$$

  where $\tau_i = \sigma_\tau \Phi^{-1}((i - 0.5)/p)$ are the quantiles from a $N(0, \sigma^2_\tau)$ distribution and $u_j \sim N(0, \gamma_1)$.

- nested two-way design where data y_{ijk}, $i = 1 \dots p,\ j = 1 \dots b_s,\ k = 1 \dots m_s$, is generated, and analysed, from the following model for $\mu^u_{ijk} = E(y_{ijk}|u_{ij})$,

$$\text{logit}(\mu^u_{ijk}) = \tau_0 + \tau_i + u_{ij}, \tag{3.6}$$

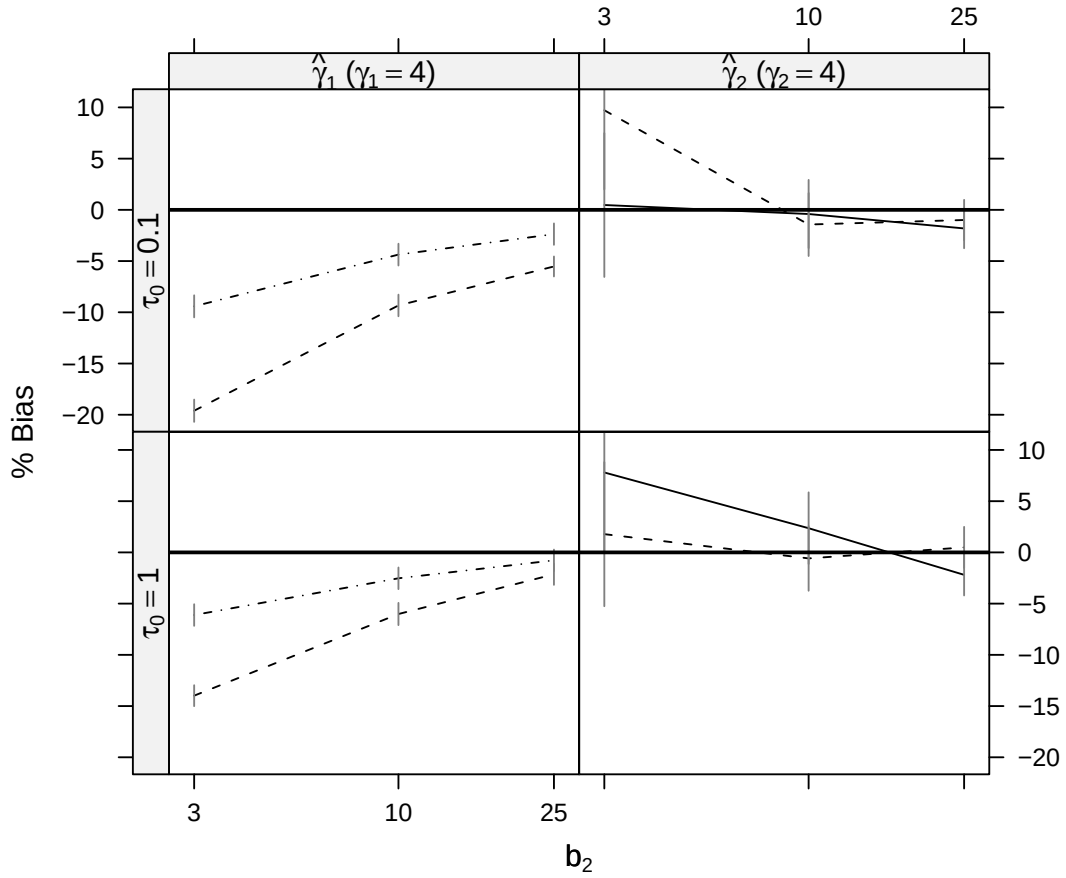


Figure 3.11: Biases for the Poisson crossed two-way model (3.4): interactions of the effects of b_2, m_s, γ_1, and γ_2 on the biases for $\hat{\gamma}_1$ and $\hat{\gamma}_2$. ($(\gamma_1, m_s) = (4,4)$: dot-dashed, $(4,1)$: dashed, $(1,4)$: dotted, $(1,1)$: solid). (Error bars are $\pm$2SE.)

with $\tau_i = \sigma_\tau \Phi^{-1}((i-0.5)/p)$ and $u_{ij} \sim N(0, \gamma_1)$.

In the analysis of each simulated dataset from each design, the τ_i were fitted as fixed effects, and the u_i, or u_{ij}, were treated as random effects. The values of the simulation parameters used for each design are given in Table 3.5. The values of p, b (or b_s), σ_τ^2 and γ_1 were arbitrarily chosen to get a good spread of different conditions.

For the crossed model, where there were relatively few random effects ($b = 10$), a positive bias for $\hat{\gamma}_1$ was observed (Figure 3.13). Increasing the number of fixed effects p from 10 to 200 increased this positive bias, and also reduced the negative bias for $\hat{\gamma}_1$ where $b = 200$. Increasing the variability of fixed effects, σ_τ, also increased the positive bias. Conversely, when the number of fixed effects was small ($p = 10$) and

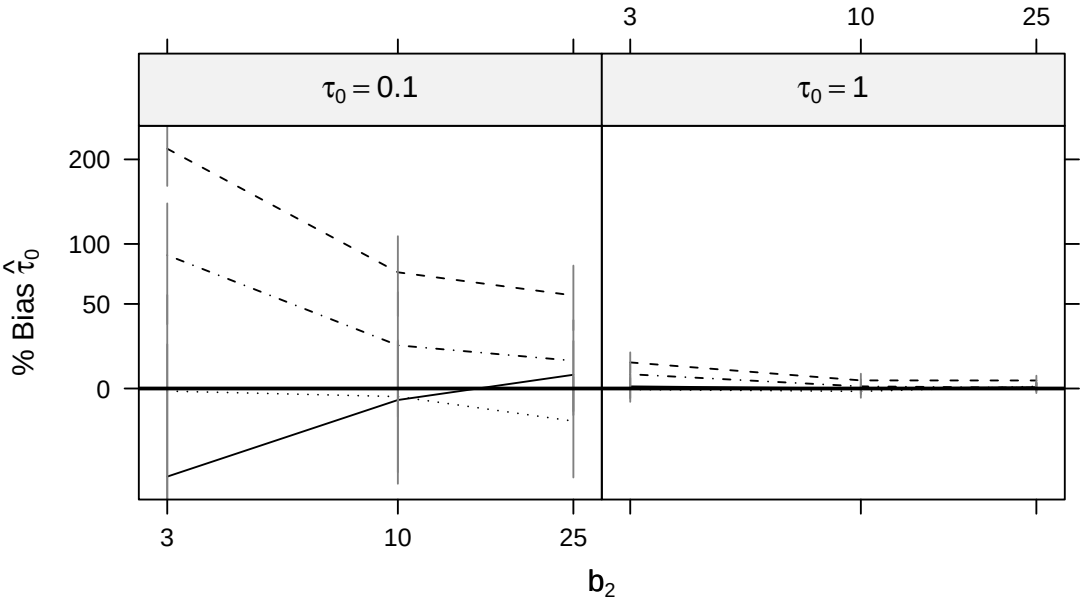


Figure 3.12: Biases for the Poisson crossed two-way model (3.4): interactions of the effects of b_2, m_s, γ_1, γ_2 on the biases for $\hat{\gamma}_1$ and $\hat{\gamma}_2$. ($(\gamma_1, m_s) = (4, 4)$: dot-dashed, $(4, 1)$: dashed, $(1, 4)$: dotted, $(1, 1)$: solid) (Error bars are $\pm$2SE.)

Parameter	*Crossed model*	*Nested model*
p	10, 50, 200	20, 50, 100, 200
b or b_s	10, 50, 200	2, 4, 8, 16
m_s	1, 4	2, 4, 8, 16
σ^2_τ, γ_1	0, 4	...
τ_0	0	...

Table 3.5: Values of the simulation parameters used for the crossed (3.5) and nested (3.6) binary models with many fixed effects. The use of ... in the 3rd (nested model) column indicates the same values were used as for the binary model.

the number of random effects was large ($b = 200$), the bias was negative, as for the previous simulation studies in this chapter. Increasing the number of replicates, m_s, reduced the magnitude of the bias in all cases.

For the nested model, negative biases increased with b_s, which here represented the ratio of the number of random effects to the number of fixed effects (Figure 3.13, where $m_s = 1$). Where $\gamma_1 = 0$, large positive biases for γ_1 were also observed in the nested model; these biases decreased with increasing b_s (Figure 3.14).

Therefore, this study confirms the results of Engel & Buist (1998), showing that increasing the number of fixed effects in the model may serve to reduce the magnitude

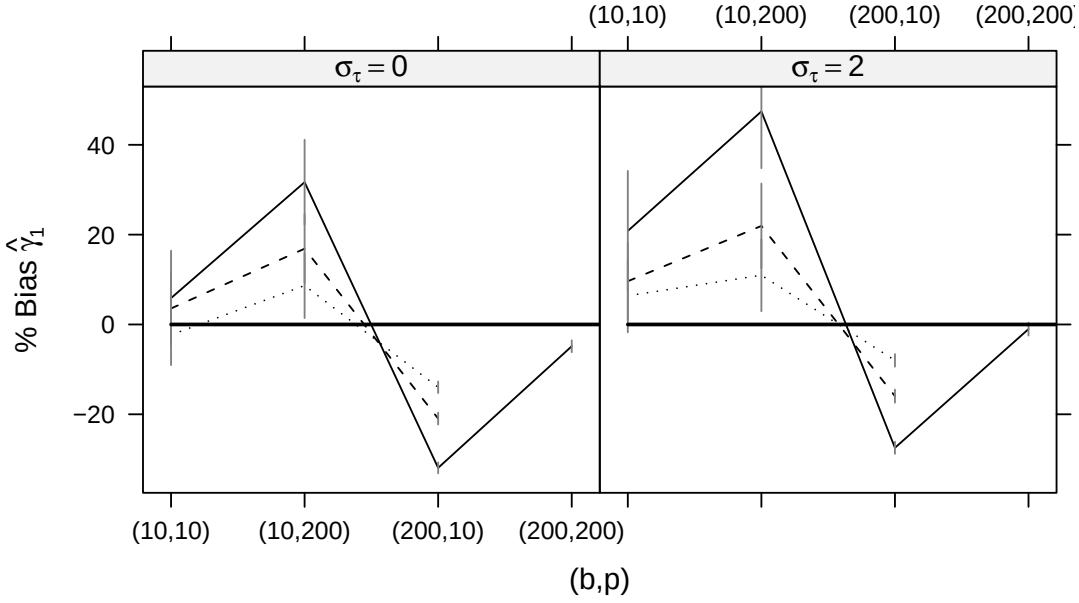


Figure 3.13: Biases for a crossed binary model with many fixed effects (3.5): interactions between the effects of b, p, m_s and σ_τ on the bias for $\hat{\gamma}_1$ ($m_s = 1$: solid, 2: dashed; 4: dotted). (Error bars are ±2SE.)

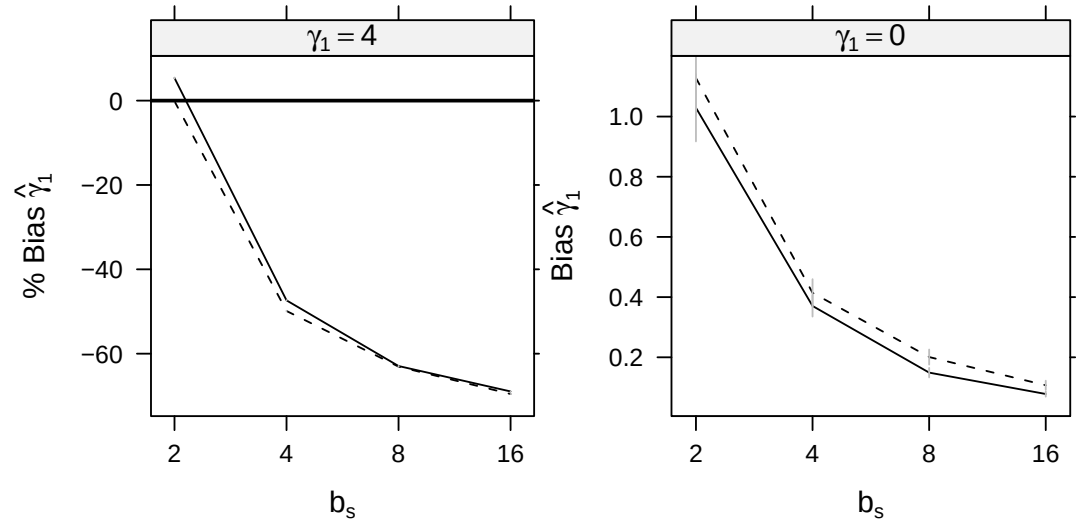


Figure 3.14: Biases for a nested binary model with many fixed effects (3.6): interactions between the effects of b and σ_u on the bias for $\hat{\gamma}_1$ where $\gamma_1 = 4$ and $\gamma_1 = 0$ respectively ($\sigma_u = 0$: solid, 2: dashed). (Error bars are ±2SE.)

of the negative biases for binary GLMMs, and even induce positive biases. The reduction in the negative bias with an increasing number of fixed effects may be related to the inconsistency of maximum likelihood estimation with a large number of "nuisance" effects (section 1.1.3.3). For instance, consider the binary matched pairs

data Y_{ij}, $i = 1, \ldots, b_g$, $j = 1, 2$, where $\mu_{ij} = E(y_{ij})$, and the following GLM for μ_{ij},

$$\text{logit}(\mu_{ij}) = \tau_0 + \tau_i + \tau_{b_g+1} x_{ij},$$

where x_{ij} is 1 if $j = 2$ or 0 if $j = 1$. It is well-known that the expectation of the ML estimator of τ_{b_g+1} is equal to $2\tau_{b_g+1}$ (Andersen, 1973). Similarly, the inclusion of many fixed effects into the GLMM, as in the studies in this section, may induce positive bias which offsets the negative bias induced by using PQL.

3.1.5 Designs with correlated random effects

The following designs involve correlated random effects, that is, where the variance covariance matrix of the random effects, $\boldsymbol{G}$, is non-diagonal. In the context of PQL estimation biases, these designs have been less explored in the literature than models with independent random effects.

3.1.5.1 Random coefficients design

A simple random coefficients design is examined here. Data y_{ij}, $i = 1, \ldots, b_g$, $j = 1, \ldots, m_g$, were generated, and analysed, according to the following model for the conditional mean $\mu^u_{ij} = E(y_{ij}|u_{1i}, u_{2i})$:

$$g(\mu^u_{ij}) = \tau_0 + \tau_1 x_{ij} + u_{1i} + x_{ij} u_{2i}, \tag{3.7}$$

where $x_{ij} = (j-1)/(m_g - 1) - 1$ is a covariate which varies within groups and $(u_{1i}, u_{2i})^T \sim N(\mathbf{0}, \boldsymbol{D})$ where

$$\boldsymbol{D} = \begin{pmatrix} \gamma_1 & \gamma_\rho \sqrt{\gamma_1 \gamma_2} \\ \gamma_\rho \sqrt{\gamma_1 \gamma_2} & \gamma_2 \end{pmatrix}.$$

The simulation parameter values chosen are given in Table 3.6.

For the binary model, simulations where either ASReml reported non-convergence or where the parameter estimates diverged ($\hat{\tau}_i > 1000$, or $\hat{\gamma}_i > 10^5$, $i = 1, 2$) were

Parameter	*Binary model*	*Poisson model*
b_g	50, 100, 200, 500	...
m_g	2, 4, 8, 16	...
(γ_1, γ_2)	(0,4), (4,0), (4,4)	...
γ_ρ	-0.7, 0, 0.7	...
τ_0	0, 2	0.1, 1
τ_1	0, 2	0, 1

Table 3.6: Values of the simulation parameters used for the random coefficients model (3.7) . The use of ... in the 3rd (Poisson model) column indicates the same values were used as for the binary model.

discarded. Across all simulations, only 0.6% of simulations were discarded. However, almost all the discarded simulations had simulation parameter values of $m_g = 2$, $\tau_0 = 2$ and $\tau_1 = 2$, and the proportion of discarded simulations was higher at low b_g. Instead of reporting the bias for the correlation estimator $\hat{\gamma}_\rho$, which is well-known to be too unstable, even for normal linear mixed models (Brian Cullis, personal communication), the bias is reported for the covariance estimator $\hat{\gamma}_{12}$, where $\gamma_{12} = \gamma_\rho\sqrt{\gamma_1\gamma_2}$.

For the binary model, the biases for $\hat{\gamma}_1$ ($\gamma_1 = 4$) and $\hat{\gamma}_2$ ($\gamma_2 = 4$) were strongly negative, and increased in magnitude with increasing γ_2 and γ_1 respectively (Figure 3.15). The bias for the covariance estimator $\hat{\gamma}_{12}$ was also strongly negative. The biases for $\hat{\tau}_0$ ($\tau_0 = 2$) and $\hat{\tau}_1(\tau_1 = 2)$ were also negative, and increased in magnitude with γ_1 and γ_2, but more so with increasing γ_1 (Figure 3.16). When $\tau_i = 2$, $i = 1, 2$, $\gamma_1 = 4$ and $\gamma_2 = 4$ and $\gamma_\rho \neq 0$, the biases for $\hat{\tau}_i$ were smaller in magnitude when $\gamma_\rho = -0.7$ but larger when $\gamma_\rho = 0.7$. There were also non-negligible biases for $\hat{\tau}_i$ when $\tau_i = 0$ and $\gamma_\rho \neq 0$, with positive bias observed when $\gamma_\rho = -0.7$ and negative bias when $\gamma_\rho = 0.7$ (Figure 3.17).

For the Poisson model, an unanticipated trend in the bias was observed – the magnitude of the negative bias for $\hat{\gamma}_1$ ($\gamma_1 = 4$) decreased with increasing γ_2 from 0 to 4 (Figure 3.18, left plot). Closer inspection of the results for the case when $\gamma_1 = 4$ and $\gamma_2 = 0$ revealed that, in over 80% of cases, the estimate $\hat{\gamma}_1$ was less than 0.1 (in contrast,when $\gamma_1 = \gamma_2 = 4$, less than 0.01% of $\hat{\gamma}_1$ values were less than 1). Therefore, when $\gamma_1 = 4$ and $\gamma_2 = 0$, PQL very often failed to provide meaningful estimates of

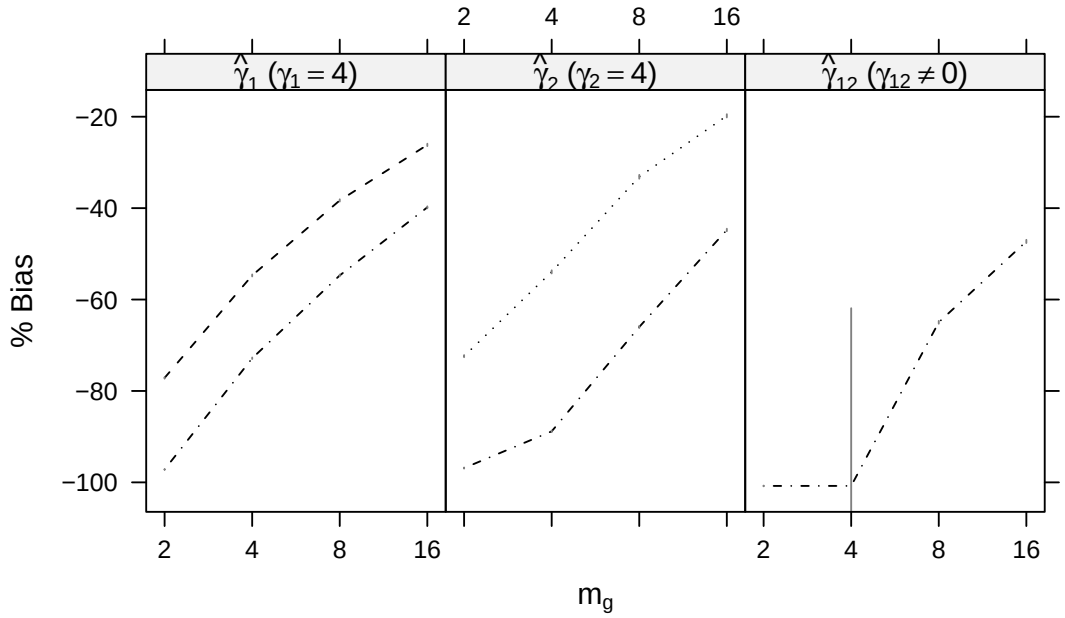


Figure 3.15: Biases for the binary random coefficients model (3.7): interactions between the effects of m_g and γ_2 or γ_1 on biases for $\hat{\gamma}_1$, $\hat{\gamma}_2$ and $\hat{\gamma}_{12} = \hat{\gamma}_\rho\sqrt{\hat{\gamma}_1\hat{\gamma}_2}$. ($(\gamma_1, \gamma_2) = (4, 4)$: dot-dashed; $(4, 0)$: dashed; $(0, 4)$: dotted). (Error bars are $\pm$2SE.)

γ_1; even at higher values of m_g, the estimates did not improve. In contrast, the bias for $\hat{\gamma}_2$ ($\gamma_2 = 4$), as expected, increased with increasing γ_1 (middle plot). For both $\hat{\gamma}_1$ and $\hat{\gamma}_2$, the magnitude of the negative bias decreased slightly with increasing τ_0. The magnitude of the negative bias for $\hat{\gamma}_{12}$ ($\gamma_{12} > 0$) decreased with increasing τ_0 (right plot), also as expected.

Referring now to the biases for $\hat{\tau}_i$, $i = 1, 2$, the negative bias for $\hat{\tau}_0$ was largest where $\gamma_1 = 4$ and $\gamma_2 = 0$, most likely associated with the corresponding high negative bias for $\hat{\gamma}_1$ at these values of γ_1 and γ_2 (Figure 3.19). There was little or no bias for $\hat{\tau}_1$ ($\tau_1 = 1$) when $\gamma_1 = 0$, and only negative bias when $\gamma_1 = 4$ (Figure 3.20). Bias for $\hat{\tau}_1$ ($\tau_1 = 1$) also varied with γ_ρ, with higher and lower levels of negative bias when $\gamma_\rho = -0.7$ and 0.7 respectively.

As for the previous studies in this chapter, this study showed that there were strong effects of the group size m_g on the biases. This study also shows that increasing either of the variance parameters, γ_1 or γ_2, leads to an increase in the biases, consistent with previous studies. For the Poisson model, the effect of the intercept τ_0 on the biases

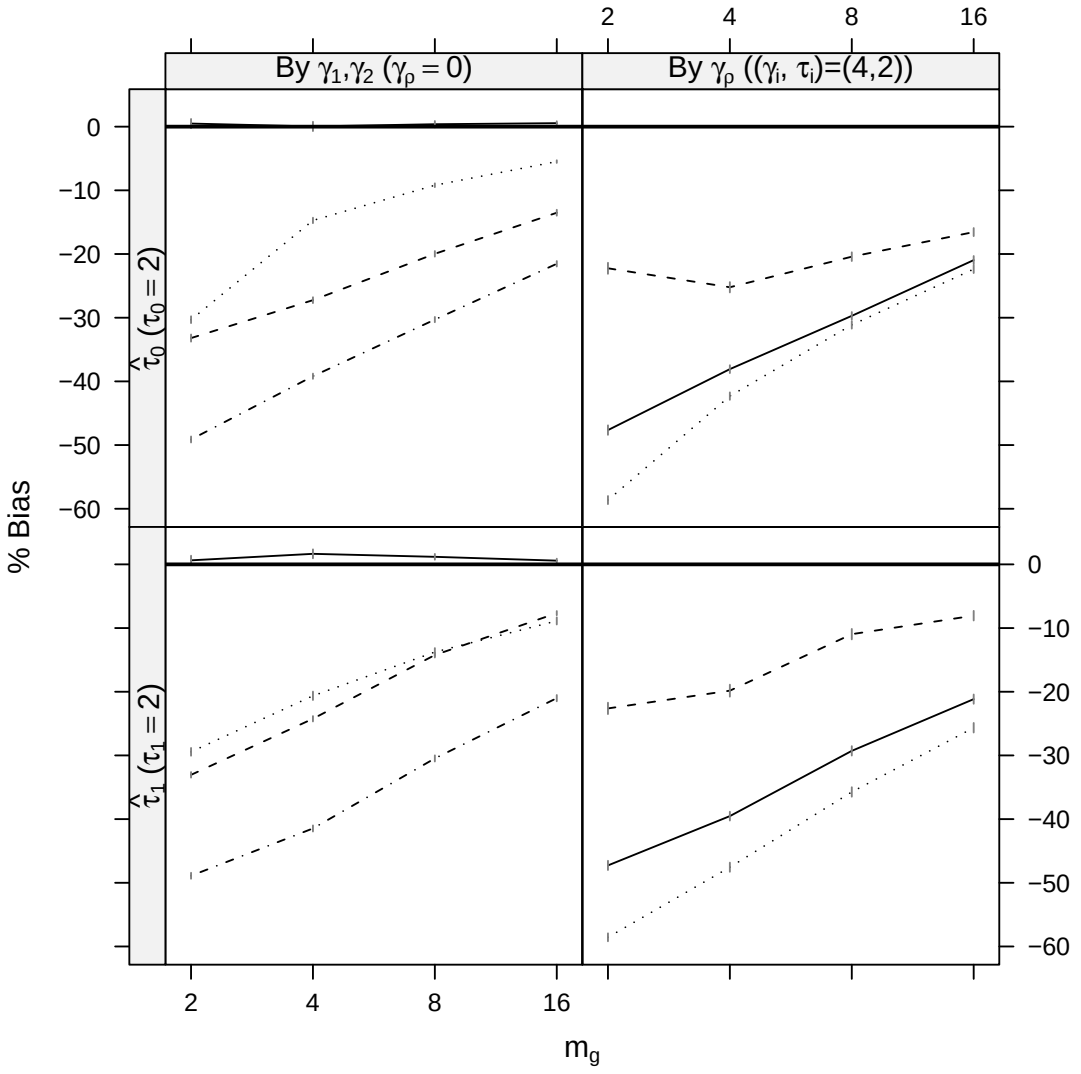


Figure 3.16: Biases for the binary random coefficients model (3.7): interactions between the effects of

(a) m_g and γ_1 and γ_2 where $\gamma_\rho = 0$ (Left-column)
($(\gamma_1, \gamma_2) = (4,4)$: dot-dashed, $(4,0)$: dashed, $(0,4)$: dotted, $(0,0)$: solid.)

(b) m_g and γ_ρ where $\tau_i = 2$ $(i = 0, 1)$ and $\gamma_j = 2$ $(j = 0, 1)$ (Right-column)
($\gamma_\rho = 0$: solid; 0.7: dotted; -0.7: dashed).

on the biases for $\hat{\tau}_0$ ($\tau_0 = 2$) and $\hat{\tau}_1$ ($\tau_1 = 2$). (Error bars are $\pm$2SE.)

was also present, as in previous studies. The effects of the correlation parameter γ_ρ on the biases for $\hat{\tau}_i$ cannot be related to the results of the previous studies, but a probable reason for this effect is as follows. Since the estimates of the variance parameters γ_ρ, γ_1 and γ_2 are negatively biased, the estimates of the fixed parameter estimates τ_i are attenuated towards the estimates of τ_i from fitting a corresponding

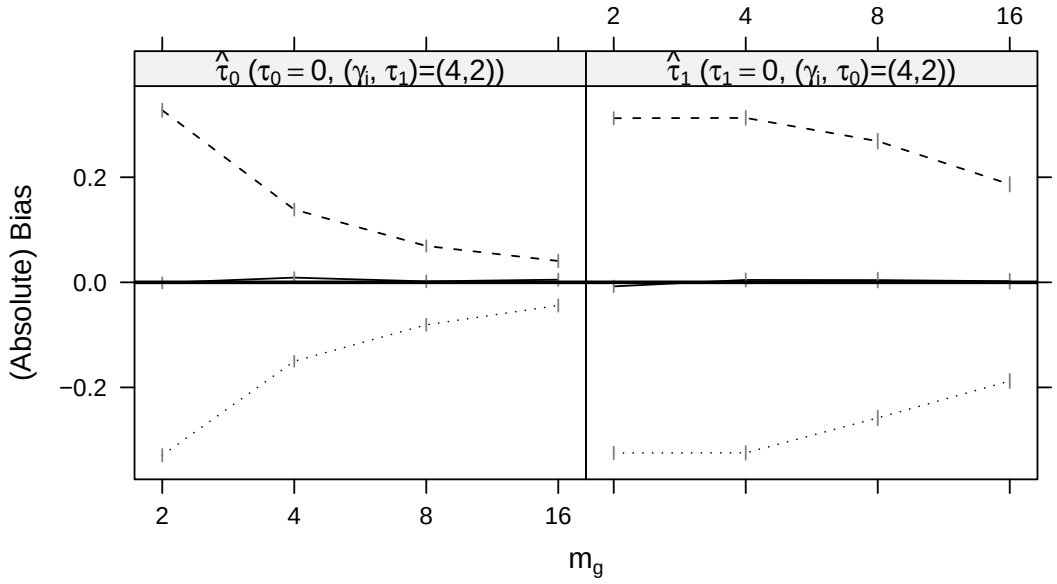


Figure 3.17: Biases for the binary random coefficients model (3.7): interactions between the effects of γ_ρ and m_g on the absolute biases for $\hat{\tau}_0$ ($\tau_0 = 0$, $(\gamma_i, \tau_1) = (4, 2)$) and $\hat{\tau}_1$ ($\tau_1 = 0$, $(\gamma_i, \tau_1) = (4, 2)$). ($\gamma_\rho = 0$: solid; $\gamma_\rho = 0.7$: dotted; $\gamma_\rho = -0.7$: dashed). (Error bars are ±2SE.)

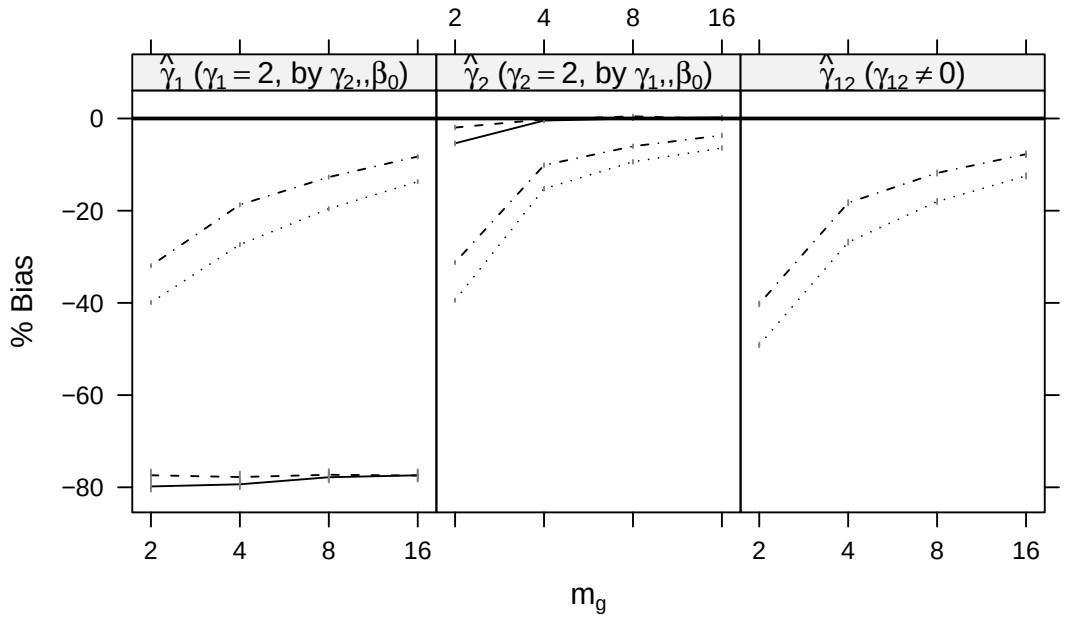


Figure 3.18: Biases for the Poisson random coefficient model (3.7): interactions between the effects of m_g, γ_1, γ_2 and τ_0 on the biases for $\hat{\gamma}_1$, $\hat{\gamma}_2$ and $\hat{\gamma}_{12}$. ($\gamma_i, \tau_0 = 0, 0.1$: solid; $\gamma_i, \tau_0 = 0, 1$: dashed; $\gamma_i, \tau_0 = 4, 0.1$: dotted; $\gamma_i, \tau_0 = 4, 1$: dot-dashed). (Error bars are ±2SE.)

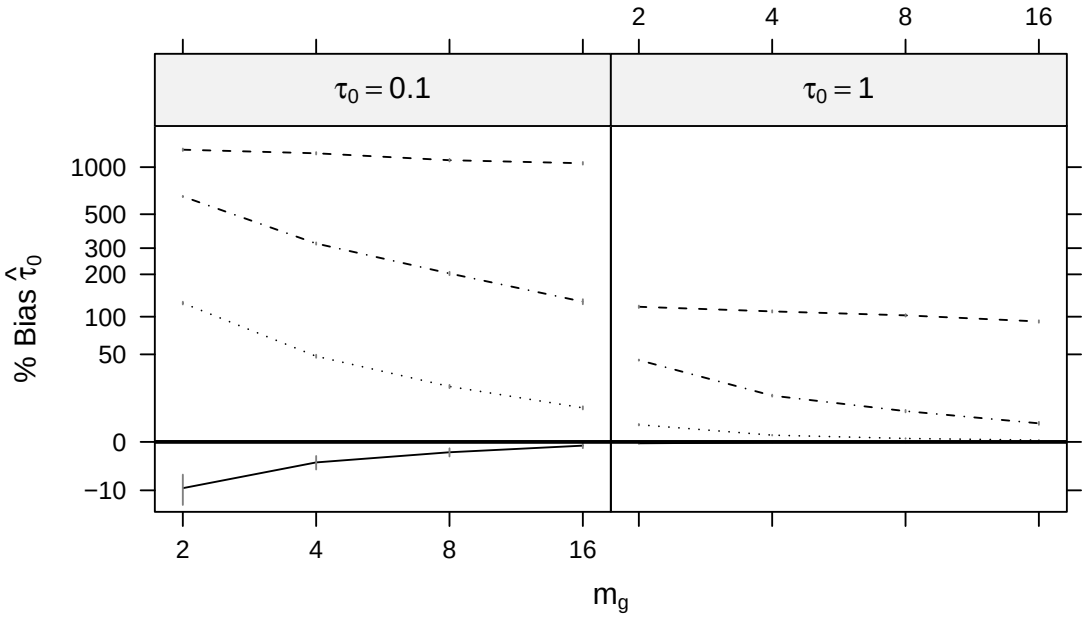


Figure 3.19: Biases for the Poisson random coefficient model (3.7): interactions of the effects of m_g, γ_1, γ_2 and τ_0 on biases for $\hat{\tau}_0$ ($\gamma_1, \gamma_2 = 0, 0$: solid; $\gamma_1, \gamma_2 = 0, 4$: dashed; $\gamma_1, \gamma_2 = 0, 4$: dotted; $\gamma_1, \gamma_2 = 4, 4$: dot-dashed). (Error bars are $\pm$2SE.)

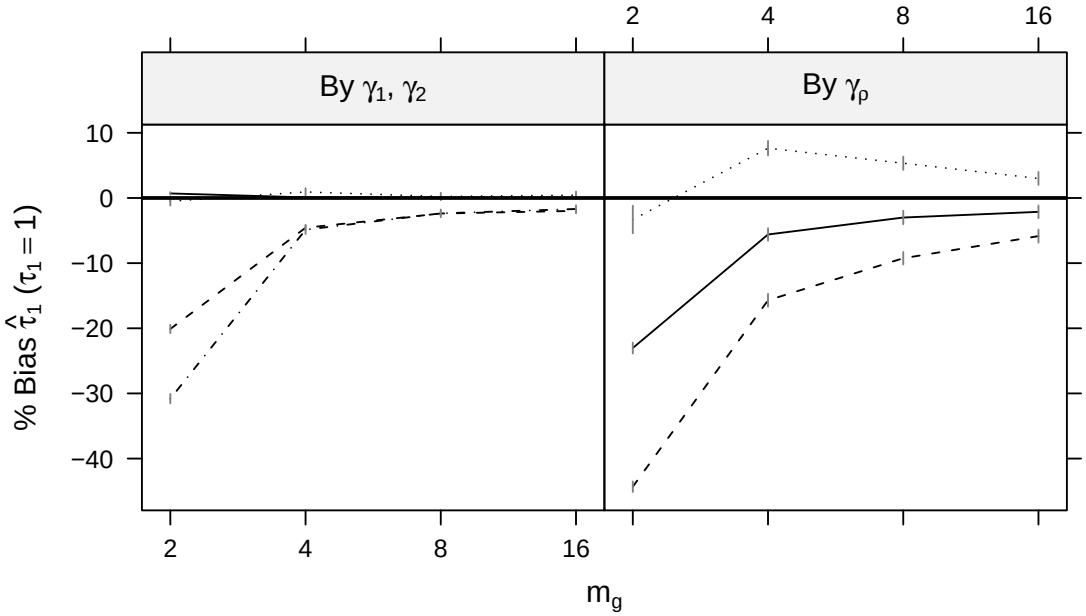


Figure 3.20: Biases for the Poisson random coefficient model (3.7): interactions between the effects of m_g, γ_1 and γ_2 or γ_ρ on the biases for $\hat{\tau}_1$ (Left plot: $(\gamma_1, \gamma_2) = (0, 0)$: solid; $(0, 4)$: dashed; $(0, 4)$: dotted; $(4, 4)$: dot-dashed. Right plot: $\gamma_\rho = 0$: solid; -0.7: dashed; 0.7: dotted). (Error bars are $\pm$2SE.)

GLM, that is, model (3.7) without the random effects u_{1i} and u_{2ij}. A bias is observed when $\tau_i = 0$ and $\gamma_\rho \neq 0$, since the expected value of the GLM estimates of τ_i in this case are not 0.

The complete failure of PQL for some of the Poisson simulations, where $\gamma_1 = 4$ and $\gamma_2 = 0$, is perhaps not simply a problem with PQL, but rather with the nature of the data generated and the model being fitted. One of the simulated datasets using these parameters was captured; the other parameter value settings for this simulated dataset were $(b_g, m_g, \gamma_\rho, \tau_1, \tau_2) = (50, 16, -0.7, 0.1, 0)$. The variance parameter estimates for this dataset were $(\hat{\gamma}_1, \hat{\gamma}_\rho, \hat{\gamma}_2) = (0.012, 0, 0)$. We also tried to analyse this dataset, using the same model (3.7), with adaptive GHQ, as implemented in SAS's `NLMIXED` procedure (section 2.2.3). However, all attempts at fitting this dataset using `NLMIXED` resulted in non-convergence (error message: "`optimisation cannot improve the function value`"), even when overly generous starting values of $(\gamma_1, \gamma_\rho, \gamma_2) = (4, 0.1, 0.1)$ were used. However, refitting this dataset in `ASReml`, but using a simpler model which ignored γ_ρ, that is, model (3.7) with no correlation between u_{1i} and u_{2i}, where

$$\boldsymbol{D} = \begin{pmatrix} \gamma_1 & 0 \\ 0 & \gamma_2 \end{pmatrix},$$

gave much more credible answers of $(\hat{\gamma}_1, \hat{\gamma}_2) = (3.49, 0)$.

3.1.5.2 Design with auto-regressive correlated errors

A simple model with autoregressive correlated errors within groups is examined, corresponding to repeated measures data, or data with spatial dependence in only one direction. Let data Y_{ij}, $i = 1 \ldots b_g$, $j = 1 \ldots m_g$, be generated, and analysed, from the following model for the conditional mean $\mu^u_{ij} = E(y_{ij}|u_{ij})$,

$$g(\mu^u_{ij}) = \tau_0 + u_{ij}, \tag{3.8}$$

where $u_{ij} \sim N(0, \gamma_1)$ is a Gaussian process with $\text{cov}\,(u_{ij}, u_{ik}) = \gamma_1 \gamma_\rho^{|k-j|}$. The simulation parameter values are shown in Table 3.7.

Parameter	*Binary model*	*Poisson model*
b_g	50, 100, 200, 500	...
m_g	2, 4, 8, 16, 32, 64	...
γ_1	1, 4	...
γ_ρ	0.5, 0.8	...
τ_0	0, 2	0.1, 1

Table 3.7: Values of the simulation parameters used for the correlated AR model (3.8) . The use of ... in the 3rd (Poisson model) column indicates the same values were used as for the binary model.

For the binary model, there were negative biases for $\hat{\gamma}_1$ and $\hat{\tau}_0$ (Figure 3.21) and positive biases for $\hat{\gamma}_\rho$ (Figure 3.22). The biases for all three parameters were excessively large at all values of m_g, although less so at $\gamma_\rho = 0.8$ than at $\gamma_\rho = 0.5$ for $\hat{\gamma}_1$ and $\hat{\gamma}_\rho$. As for the one-way classification, there was no evidence of bias for $\hat{\tau}_0$ when $\tau_0 = 0$, so only biases for $\hat{\tau}_0$ at $\tau_0 = 2$ are shown. For the Poisson model, there were large negative and positive biases for $\hat{\gamma}_1$ and $\hat{\tau}_0$ respectively, and the magnitude of these biases increased with γ_1 but decreased with γ_ρ and τ_0 (Figures 3.23 and 3.24). There was little or no evidence of bias for $\hat{\gamma}_\rho$.

In this study, the effects of the variance parameter γ_1, and the effects of τ_0 for Poisson data, were consistent with the results for the one-way classification (section 3.1.4.1). The one-way classification model (3.1) can be considered a special case of the correlated AR model (3.8) where $\gamma_\rho = 1$, that is, perfect correlation between units in a group, and so $u_{ij} = u_i$. The main difference between the results for this study, compared to the one-way classification results, is that the magnitude of the bias does not reduce much with increasing group size m_g. To understand why the group size m_g makes so little difference to the bias in this study, one might consider an "effective" group size, that is, the number of units per independent random effect. For instance, a correlation of $\gamma_\rho = 1$ corresponds to the one-way classification, where the "effective" group size is m_g. At the other extreme, a correlation of $\gamma_\rho = 0$ corresponds to the situation where all the u_{ij} are independent and the "effective" group size is 1.

Values of γ_ρ between these two extremes, such as $\gamma_\rho = 0.5$ and 0.8, should therefore result in effective group sizes lying between 1 and m_g. One approach to determining

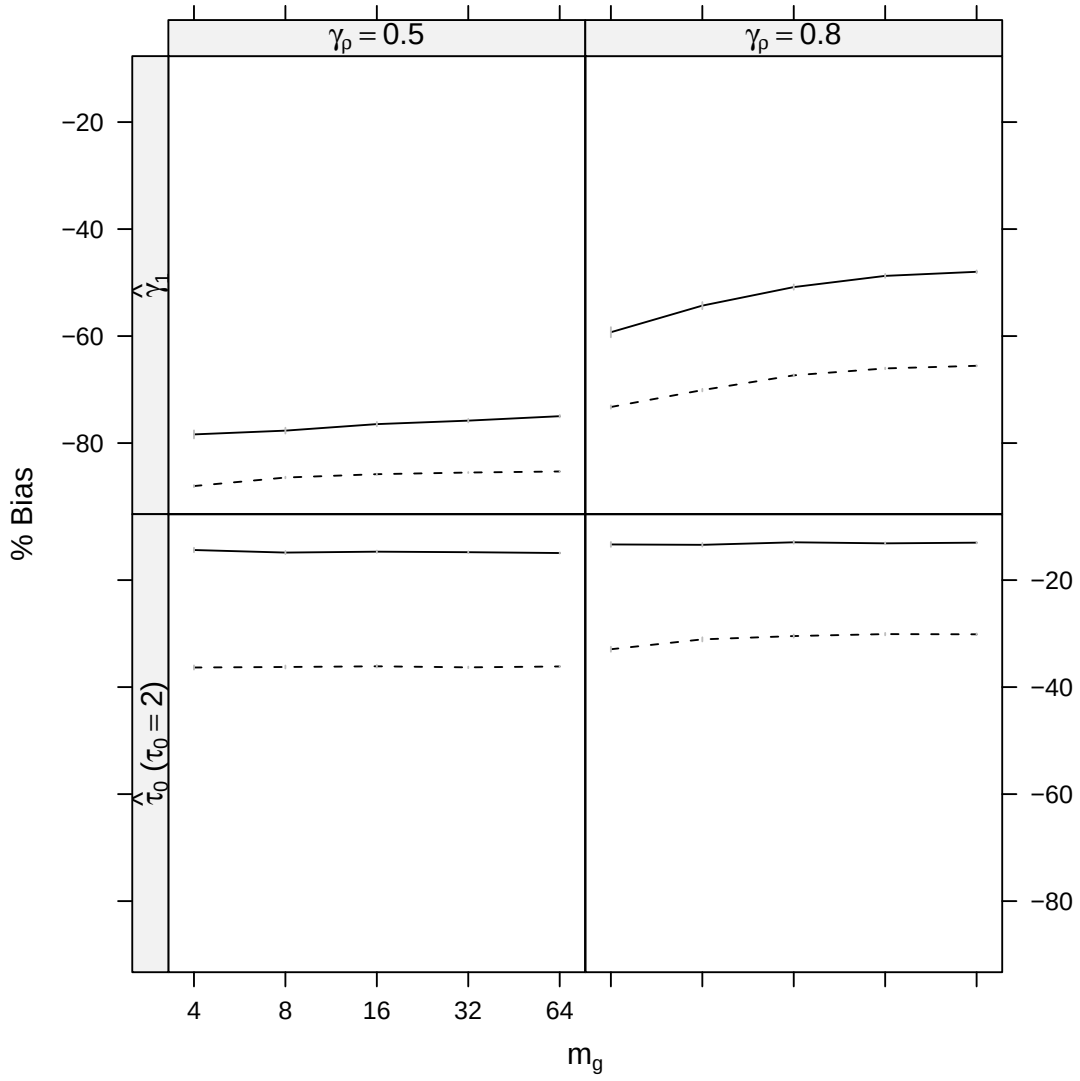


Figure 3.21: Biases for the binary AR correlated model (3.8): interactions between the effects of m_g, γ_1 and γ_ρ on the biases for $\hat{\gamma}_1$ and $\hat{\tau}_0$ ($\gamma_1 = 1$: solid; 4: dotted). (Error bars are $\pm$2SE.)

an effective group size is to calculate the number of independent random effects per group using elementary time series theory. The random effects in group i can be generated using the following model,

$$u_{ij} = \gamma_\rho u_{i,j-1} + e_{ij},$$

where $e_{ij} \sim N\left(0, \gamma_1^2(1-\gamma_\rho^2)\right)$. Therefore, the first random effect, u_{i1}, contributes 1 "independent" random effect, and each subsequent random effect u_{ij}, $j > 1$, con-

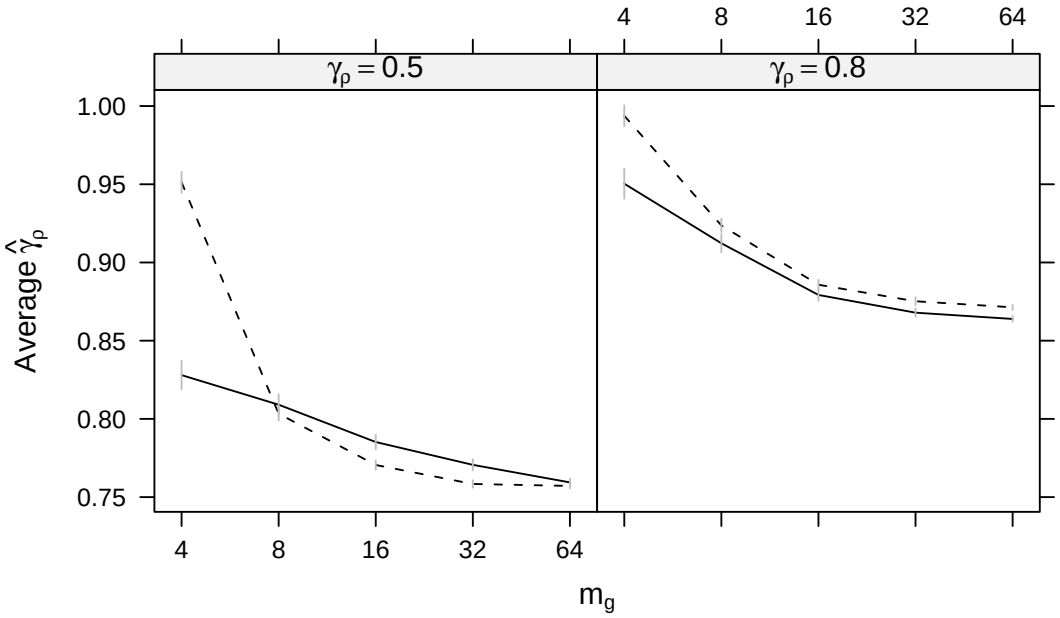


Figure 3.22: Biases for the binary AR correlated model (3.8): interactions between the effects of m_g, γ_1 and γ_ρ on the bias for $\hat{\gamma}_\rho$ for the binary AR correlated model ($\gamma_1 = 1$: solid; $\gamma_1 = 4$: dotted). (Error bars are $\pm$2SE.)

tributes an additional "independent" $(1-\gamma_\rho^2)$ of a random effect. So the total number of independent random effects in a group is $1 + (m_g - 1)(1 - \gamma_\rho^2)$, and therefore the effective group size is $m_e = m_g / \left[1 + (m_g - 1)(1 - \gamma_\rho^2)\right]$. For $m_g = 64$, this gives effective group sizes of 1.3 and 2.7 for $\gamma_\rho = 0.5$ and 0.8 respectively. Comparing the biases for γ_1 in Figure 3.21 at $m_g = 64$ against the biases for γ_1 in the one-way classification (Figure 3.1) for $m_g = 2$ and 4 respectively shows that this rule appears to work reasonably well.

3.1.6 Discussion

The simulation studies in this section explored the effects of design parameters on the PQL estimation biases for relatively simple GLMMs. As already indicated, some of the parameter settings used in these simulations, such as the group size m_g, were chosen based on previous literature to induce large PQL estimation biases. Therefore, it is expected that the large magnitudes of the estimation biases shown here will generally not be indicative or representative of the estimation biases arising in the

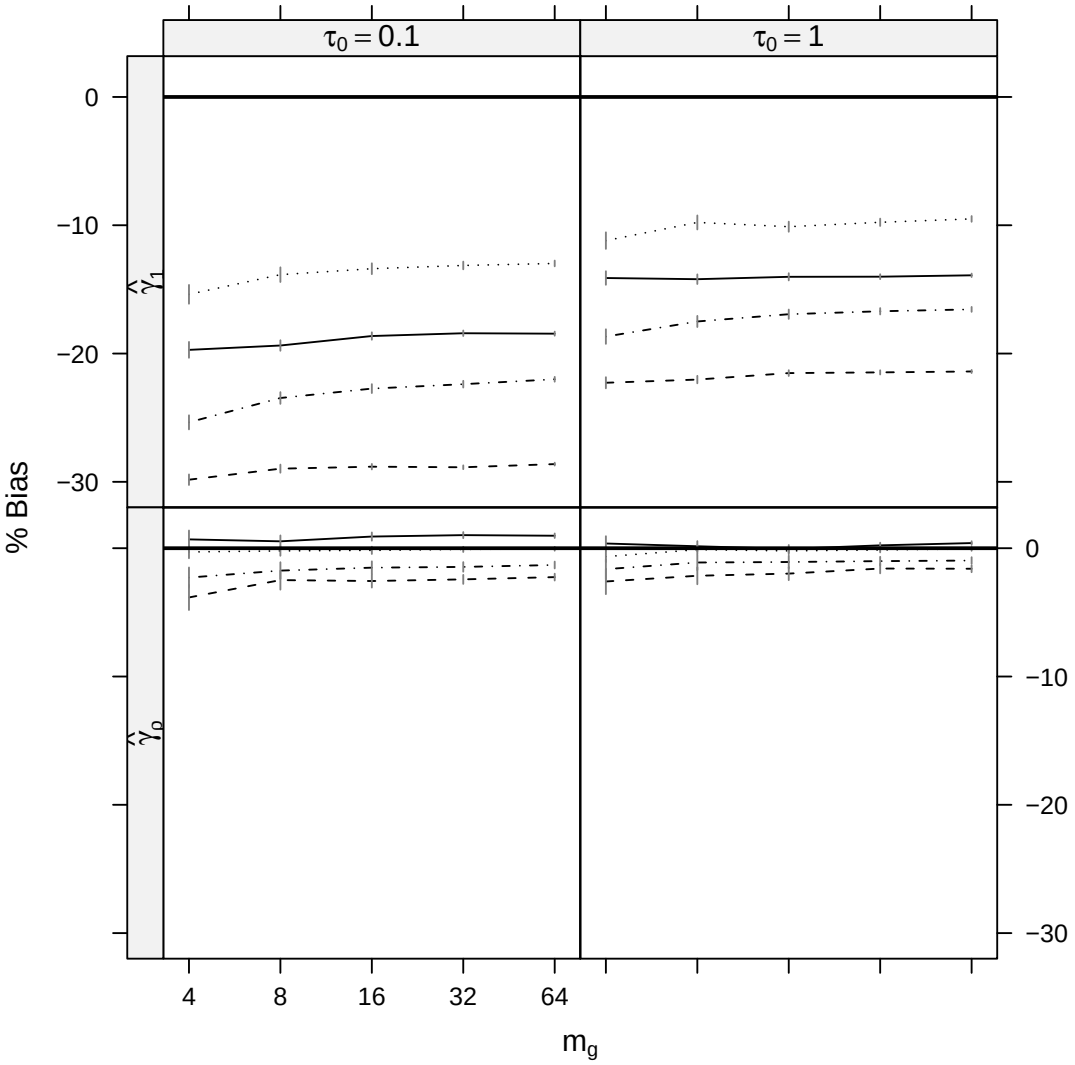


Figure 3.23: Biases for the Poisson AR correlated model (3.8): interactions between the effects of m_g, τ_0, γ_1 and γ_ρ on the biases for $\hat{\gamma}_1$ and $\hat{\gamma}_\rho$ ((γ_1, γ_ρ)=(1,0.5): solid; (2,0.5): dashed; (1,0.8): dotted; (2,0.8): dot-dashed). (Error bars are $\pm$2SE.)

analysis of "real data" .

The two major influences on the estimation biases, across all simulation studies reported here, are the magnitude of the variance parameters and the group sizes. For designs with independent random effects, negative estimation biases were recorded for the variance parameter estimators for both binary and Poisson models, with generally negative estimation biases for the fixed effect estimators as well (the main exception being the intercept parameter estimator in the Poisson models, which invariably had positive estimation biases). These negative estimation biases increased in magnitude

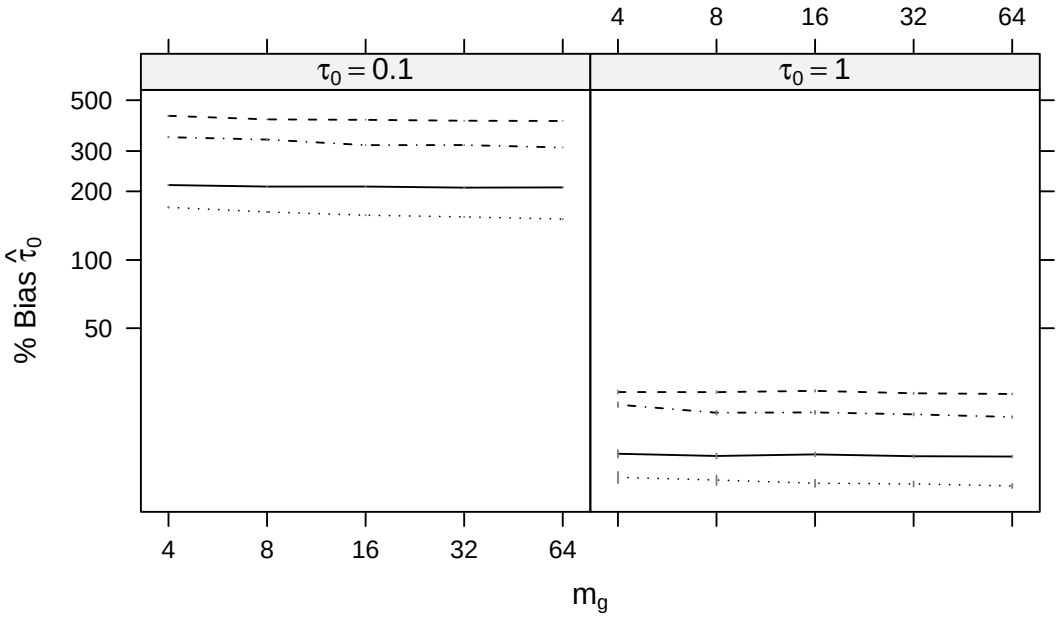


Figure 3.24: Biases for the Poisson AR correlated model (3.8): interactions between the effects of m_g, τ_0, γ_1 and γ_ρ on the biases for $\hat{\gamma}_1$ and $\hat{\gamma}_\rho$ ((γ_1, γ_ρ)=(1,0.5): solid; (2,0.5): dashed; (1,0.8): dotted; (2,0.8): dot-dashed). (Error bars are $\pm$2SE.)

with the magnitude of each of the variance parameters and as each of the group sizes decreased. This was demonstrated for the one-way classification study (section 3.1.4.1), and confirmed for the nested and crossed two-way classification studies. In both the two-way studies, there are multiple levels of grouping, and it was the group size at the lowest level of aggregation, and the magnitude of the corresponding variance parameter, which has the most impact on the the biases. The effect of having a large number of fixed effects in the model (section 3.1.4.4), on reducing and over-compensating the biases for the binary model, may be related to the inconsistency of the ML estimator for some GLMs where there are many fixed effects in the models.

As shown at the end of the one-way classification results (section 3.1.4.1), the rule of thumb proposed by Breslow (2003), regarding when PQL can be considered "adequate", has scope for improvement. That is, whether the data generally has counts of successes or failures above 5 for binomial data, or has rates generally above 5 for Poisson data, may not necessarily indicate that estimation biases will be small and ignorable, and vice versa. Breslow's hypothesis, that PQL will "fail" when the con-

ditional PDF of the data given the random effects $f_{Y|U}$ is far from normal, ignores the influence of the variance parameters on the bias. The assertion of Lee & Nelder (1996), presented at the start of this chapter, similarly ignores the effects of variance parameters on the biases, as do the CPQL correction factors presented in section 2.1.1.2.

For designs with correlated random effects, similar trends were observed as for designs with independent random effects. However, the results of the correlated AR study (section 3.1.5.2) were somewhat different, with little reduction in the bias with increasing group size. In this case, it is more appropriate to consider an "effective" group size, representing the "information" available for estimating a single random effect. A rule for calculating an "effective" group size, in the case of correlated AR1 errors, has been given at the end of section 3.1.5.2. This result suggests that relatively larger biases may be expected in GLMMs where correlated random effects were used.

3.1.7 Case study : Beitler-Landis dataset

A case study in Breslow (2003) is used to demonstrate effects of the design parameters on the estimation bias. The Beitler & Landis (1985) multi-centre clinical trial on topical cream effectiveness was used in Breslow (2003) to demonstrate how well PQL can perform in a binomial example. The data are reproduced in Table 3.8 for convenience.

	Treated			*Control*		
Clinic	*Success*	*Total*	*%*	*Success*	*Total*	*%*
1	11	36	31	10	37	27
2	16	20	80	22	32	69
3	14	19	74	7	19	37
4	2	16	13	1	17	6
5	6	17	35	0	12	0
6	1	11	9	0	10	0
7	1	5	20	1	9	11
8	4	6	67	6	7	86

Table 3.8: The Beitler & Landis (1985) dataset used in Breslow (2003).

The first model proposed by Breslow (2003) will be explored, ignoring the need to

allow for possible treatment by clinic interactions. Let $\mu_{ij} = \mathrm{E}(y_{ij}|u_i)$ be the conditional mean for Y_{ij}, the jth observation within the ith clinic, with model

$$\mathrm{logit}(\mu_{ij}) = \tau_0 + x_{ij}\tau_1 + u_i, \qquad i = 1, \ldots, 8,\ j = 1, 2 \tag{3.9}$$

where $u_i \sim N(0, \gamma_1)$ are the random clinic effects, τ_0 is the intercept and τ_1 is the effect of treatment corresponding to a centred covariate $x_{ij} = -0.5$ where $j = 1$ (control) and $x_{ij} = 0.5$ where $j = 2$ (treatment). Estimates from PQL and adaptive GHQ (AGHQ), using the SAS `NLMIXED` procedure (Wolfinger, 1999) with default options, are reproduced in Table 3.9. In addition, estimates (posterior means) from using a Bayesian approach are also shown. (The Bayesian approach was implemented using the classic `BUGS` (Spiegelhalter *et al.*, 1995) program using 20,000 samples from the Gibbs sampler after a 2,000 sample burn-in, with an Inverse Gamma (IG) prior for γ_1, and IG parameter settings $\mu = 0.001$ and $r = 0.001$.) As Breslow (2003) noted, there is remarkable concordance between the PQL and GHQ estimates, showing no evidence of PQL bias, which he also confirms using simulations based on this design. Note that the Bayesian estimate (posterior mean) of γ_1 is noticeably different from the others, with a relatively high SE associated with it. However, further examination of Bayesian estimation, including the exploration of alternative priors, will be deferred until chapter 5.

	$\hat{\tau}_0$	$\hat{\tau}_1$	$\hat{\gamma}_1$	**LRT** $\hat{\gamma}_1$
PQL	-0.784 ± 0.537	0.724 ± 0.296	2.033 ± 1.250	50.4
GHQ	-0.828 ± 0.533	0.739 ± 0.300	1.960 ± 1.190	55.4
Bayes	-0.828 ± 0.636	0.753 ± 0.303	3.248 ± 2.751	

Table 3.9: Estimates of the parameters in (3.9) for the analysis of the Beitler & Landis (1985) dataset using PQL, GHQ and Bayesian approaches.

To verify that the concordance between GHQ and PQL estimates was not a "fluke", 1000 simulated datasets were generated from the design. The true value of each parameter was set equal to the PQL estimates in Table 3.9. The results (line 1, Table 3.10) show that PQL estimation biases are minor. The results of the one-way classification simulation study would suggest that these low biases can be attributed

to this design having both a small number of groups (clinics) and a relatively large average group size (patients per clinic).

However, if the simulation study is repeated with 32 clinics instead of 8, negative biases are now observed (scenario 2, Table 3.10). Similarly, repeating the study with each of the binomial denominators reduced to one fourth of their magnitude (rounded up to the nearest integer) results in slight negative biases (scenario 3). Finally, repeating the simulation study with both these changes results in the worst estimation biases (scenario 4). Therefore, increasing the number of groups (clinics) or decreasing the group size (patients per clinic) both increase the magnitude of the estimation biases, in line with the results of the earlier simulation studies.

Scenario	$\hat{\tau}_0$ **(-0.784)**	$\hat{\tau}_1$ **(0.724)**	$\hat{\gamma}_1$ **(2.033)**
(1): Original design	-0.731 ± 0.016	0.724 ± 0.010	2.046 ± 0.039
(2): (1) with 32 clusters	-0.750 ± 0.008	0.706 ± 0.005	1.840 ± 0.017
(3): (1) with denoms/4	-0.709 ± 0.018	0.722 ± 0.019	1.957 ± 0.047
(4): (2) and (3)	-0.703 ± 0.009	0.664 ± 0.008	1.574 ± 0.018

Table 3.10: Average parameter estimates from simulations conducted using the Beitler/Landis dataset (3.9) as the design.

3.2 Other statistical inference using PQL

This section will consider other statistical inference using PQL. In some cases, the primary aim may be to test whether an effect is non-zero, with the precise estimate of the effect being of secondary importance.

3.2.1 Inference concerning variance components

In normal linear mixed models, testing of variance components is generally performed either with a likelihood ratio test (LRT) or a Wald test, the latter using the estimated variance-covariance matrix of the variance parameter estimates. However, the use of a Wald test for testing variance components is generally not recommended, since the distributions of variance component estimators are not symmetric.

This study therefore only examines the use of a LRT for testing variance components in GLMMs. Only tests of single variance components, that is, hypothesis tests with null hypothesis

$$H_0 : \gamma_i = 0, \quad i \in 1, \ldots, q,$$

are considered here. In order to apply a LRT, a likelihood calculation is required, with the analytical intractability of the likelihood an obvious hindrance. Since PQL does not maximise a likelihood or a fixed criterion, a first order Laplace approximation to the likelihood is proposed. The first order Laplace approximation can be readily calculated at the PQL estimates of $\boldsymbol{\gamma}$, $\boldsymbol{\tau}$ and $\boldsymbol{u}$.

Binary or Poisson data were generated from a null model for data Y_{ij}, $i = 1, \ldots, b_g$, $j = 1, \ldots, m_g$, where the model for the conditional mean $\mu^u_{ij} = E(y_{ij}|u_i)$ was

$$g(\mu^u_{ij}) = \tau_0. \tag{3.10}$$

The link function g was the logit or log function for binary and Poisson data respectively. The parameter settings are given in Table 3.11.

Parameter	*Binary Model*	*Poisson model*
b_g	50, 100, 200, 500, 1000	...
m_g	2, 4, 8, 16	...
τ_0	0, 0.5, 1, 1.5	...

Table 3.11: Values of the simulation parameters used for data generation in model (3.10) for testing a single variance component in the one-way classification model (3.11) using PQL. The use of ... in the 3rd (Poisson model) column indicates the same values were used as for the binary model.

To calculate an LRT, both the data generation model (3.10) and an alternative model

$$g(\mu^u_{ij}) = \tau_0 + u_i, \qquad u_i \sim N(0, \gamma_1), \tag{3.11}$$

were fitted using PQL. The LRT statistic was calculated as $2(\ell_A - \ell_N)$, where $\ell_N = \log f_Y$ is the GLM likelihood for the null model and ℓ_A is the Laplace approximation

to the likelihood for the alternative model. If $h = \log f_{Y,U}$, then

$$\ell_A = \left\{ h - \frac{1}{2}\log\left| -\frac{\partial^2 h}{\partial \boldsymbol{u}\partial \boldsymbol{u}^T} \right| \right\}_{\hat{\tau}_0, \tilde{\boldsymbol{u}}} = \left\{ \log f_{Y,U} - \frac{1}{2}\log \sum_{i=1}^{n} (m\hat{w}_i + \hat{\sigma}^{-2}) \right\}_{\hat{\tau}_0, \tilde{\boldsymbol{u}}},$$

where $\hat{w}_i = \hat{\mu}_i(1 - \hat{\mu}_i)$ is the estimated GLM weight for the ith group, $i = 1, \ldots, n$, given PQL estimates $\hat{\gamma}_1$ and $\hat{\tau}_0$ and predictions $\tilde{\boldsymbol{u}}$. Note that ℓ_A omits a REML-like correction, but since there is only one fixed effect in the model, this correction should be negligible. The estimated variance component was restricted to be positive in the alternative model. For linear mixed models where the variance component is restricted to be positive, Stram & Lee (1994) suggested a critical value for the LRT is $\chi^2_{1;0.9}$, the 95% quantile of a 50:50 mixture of the χ^2_0 and χ^2_1 distributions, and this will be applied here.

The use of the LRT based on the Laplace approximation for the binary model resulted in conservative tests, particularly as m decreased and as τ_0 increased (Table 3.12). For the Poisson model, the effects of τ_0 and m were less apparent. The conservativeness of these tests most likely resulted from the PQL estimation biases, which is supported by the fact that the LRT was less conservative for the Poisson model.

m	Mean	Type I (%)	τ_0	Mean	Type I (%)
2	0.24 (0.41)	1.1 (3.9)	0	0.38 (0.41)	3.6 (4.2)
4	0.34 (0.41)	2.5 (3.7)	0.5	0.35 (0.43)	2.6 (3.8)
8	0.38 (0.44)	3.2 (4.3)	1	0.34 (0.41)	2.5 (3.9)
16	0.41 (0.44)	4.3 (4.4)	1.5	0.31 (0.44)	2.4 (4.4)

Table 3.12: Means of the LRT statistic and the estimated Type I error rates for testing $H_0 : \gamma_1 = 0$ in (3.11) from simulations for the binary model (Poisson in brackets) (3.10). The LRT is calculated using a Laplace approximation and estimates from PQL.

Note that the efficient score approach of Lin (1997) is an alternative, and possibly better, approach for testing variance parameters which has not been examined here. This approach also suffers from conservative tests in the case of multiple variance components because it relies on the application of a first order Laplace approximation. Engel & Buist (1996) and Engel & Keen (1996) also suggest an approach based on the working variate likelihood, that is, the REML likelihood based on the working variate

as shown in (2.5), but with a REML correction added. They suggest the addition of a further "REML" step after convergence of the PQL algorithm, where the working variate and weights are held constant. This technique has not been investigated.

3.2.2 Inference concerning the fixed effects

For inference concerning fixed effects, a by-product of the PQL estimation algorithm is an estimate of the variance-covariance matrix of the fixed effect estimate $\hat{\boldsymbol{\tau}}$,

$$\text{var}(\hat{\boldsymbol{\tau}}) = \left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1},$$

where $\boldsymbol{V} = \boldsymbol{W}^{-1} + \boldsymbol{ZGZ}^T$ and $\boldsymbol{W}$ is the diagonal matrix of GLM weights. In this study, the application of a Wald test to test a single fixed effect, using the test statistic $\hat{\tau}_i/\text{SE}(\hat{\tau}_i)$, was examined. An approximate critical value of 2 was used for the application of a 95% two-sided Wald test. (The choice of 2, corresponding to the 97.5% cutoff point for a t_{100} distribution, is slightly more conservative than the standard normal 1.96 critical value.) This study estimates the Type I error rate, $\Pr\left(\hat{\tau}_i/\hat{\text{SE}}(\hat{\tau}_i) > 2|\tau_i = 0\right)$, from the results of each of the simulation studies above. This, of course, only uses the simulations where the fixed parameter concerned was zero, that is, $\tau_i = 0$.

In addition, the estimated $\text{SE}(\hat{\tau}_i)$ were compared with the true Monte Carlo estimates of $\text{SE}(\hat{\tau}_i)$ from the results of each of the simulation studies above. This was performed by plotting/regressing the Monte Carlo SE estimates, calculated for each combination of simulation parameters, against their corresponding average estimated SE.

3.2.2.1 Type I error rates

The simulation parameters generally had little effect on the average Type I error rates for each study, apart from the random coefficient models. Therefore, Type I error rates calculated across all values of the simulation parameters in each study are shown in Table 3.13. For the binary one-way classification, the Wald test was slightly

conservative for all parameters. For τ_1, however, the Type I error rates were even more conservative at low m_g and large γ_1, as low as 1.37% for $m_g = 2$ and $\gamma_1 = 4$. For the nested and crossed two-way classifications (the latter where $m_g = 25$), and the correlated AR model, the Type I error rates were also conservative (although less so for $\gamma_1 = 4$ for the nested two-way classification or where $\gamma_1 = 0$ and $\gamma_2 = 4$ in the crossed classification). Note that setting the critical value to $t_{n-1;0.975}$ for τ_0, instead of 2, would have resulted in anti-conservative tests where $b_g > 200$ (and $t_{n-1;0.975} < 2$). For the Poisson one-way classification, the test statistics for τ_1 and τ_2 were also slightly conservative (with a minor increase in Type I error rate with b_g for τ_1).

Binary						*Poisson*	
One-way model			Other models (τ_0)			One-way	
τ_0	τ_1	τ_2	Nest	Cross ($m_g = 25$)	Corr	τ_1	τ_2
4.65	3.74	4.56	4.24	4.75	4.37	4.55	4.80

Table 3.13: Average type I error rates, expressed as percentages, for testing $H_0 : \tau_i = 0$ using $|\hat{\tau}_i/\text{SE}(\hat{\tau}_i)| > 2$ as rejection region. (One-way=one-way classification (3.1), Nest=two-way nested model (3.3), Cross=two-way crossed model (3.4), Corr=AR Correlated model (3.8), RC=random coefficient model (3.7).)

For the random coefficient models (3.7), there were some noticeable effects of the simulation parameters on the average type I error rates (Table 3.14). For the binary random coefficient model, the type I error rates for τ_0 were conservative where either γ_1 or γ_2 were equal to 0, but when $\gamma_1 = \gamma_2 = 4$ the type I error rates increased over their nominal rates as m_g decreased. For τ_1, the type I error rates were also conservative when either γ_1 or γ_2 equaled 0, but were well over the nominal rate when $\gamma_1 = \gamma_2 = 4$ and $\tau_1 = 1$. For the Poisson random coefficient model, the type I error rates for τ_1 were conservative, except when $\gamma_1 = 4$ and $\gamma_2 = 0$.

3.2.2.2 Estimated versus Monte Carlo SEs

For the one-way classification, the average estimated SEs predicted the true Monte Carlo SEs extremely well: there was little or no evidence that the relation between the average estimated SEs and the true Monte Carlo SEs for τ_i varied from a one-one

(a) τ_0 (binary)

$\gamma_1 = 0$ or	$\gamma_1 = 4$ and $\gamma_2 = 4$, by m_g			
$\gamma_2 = 0$	m_g =2	$m_g = 4$	$m_g = 8$	$m_g = 16$
4.06	42.2	17.3	8.42	5.67

(b) τ_1 (binary) by (γ_1, γ_2)

$(0,0)$	$(0,4)$	$(4,0)$	$(4,4)$ $(\tau_0 = 0)$	$(4,4)$ $(\tau_0 = 1)$
4.04	5.60	2.60	4.22	33.4

(c) τ_1 (Poisson) by (γ_1, γ_2)

(0,0)	$(4,0)$	$(0,4)$ $(m_g = 2)$	$(0,4)$ $(m_g > 2)$	$(4,4)$
4.21	10.9	7.77	4.93	4.42

Table 3.14: Average type I error rates for random coefficients model (3.7) using $|\hat{\tau}_i/\mathrm{SE}(\hat{\tau}_i)| > 2$ as rejection region:
(a) the effect of m on the Type I error rates for τ_0 for the binary model
(b) effects of γ_1 and γ_2 on the Type I error rates for τ_1 for the binary model
(c) effects of γ_1 and γ_2 on the Type I error rates for τ_1 for the Poisson model.

relationship for either the binary (Figure 3.25) or Poisson (Figure 3.26) models.

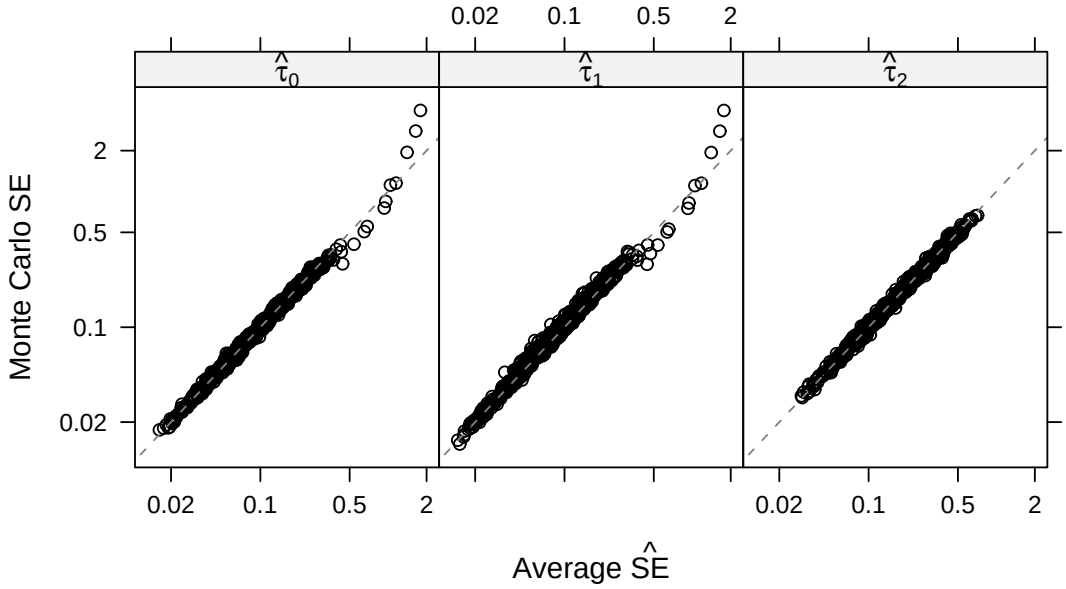


Figure 3.25: Average estimated SEs versus Monte Carlo SEs for each of the fixed parameters $\hat{\tau}_i$, $i = 0, 1, 2$, in the binary one-way classification model. (Note: log scale used for both axes.)

For the nested two-way, correlated AR and crossed two-way models, the average estimated SEs also predicted the true Monte Carlo SEs for τ_0 very well for both the binary (Figure 3.27) and Poisson models (Figure 3.28). There are a large number of points for the correlated AR Poisson model (Figure 3.28) where the Monte Carlo SEs

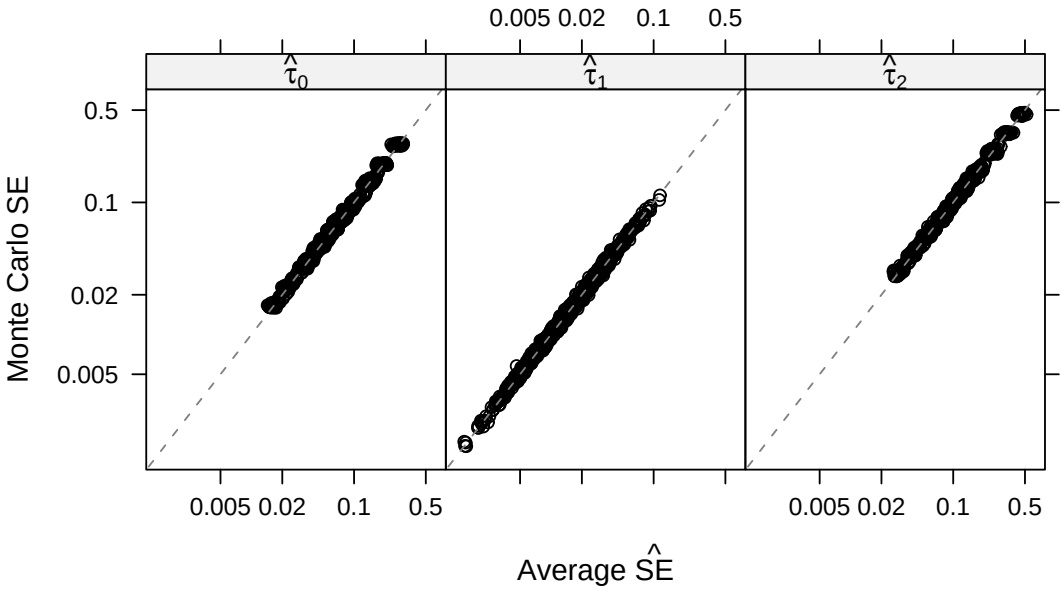


Figure 3.26: Average estimated SEs versus Monte Carlo SEs for all combinations of simulation parameters for the Poisson one-way classification model (3.1). (Note: log scale used for both axes.)

exceeded the estimated SEs. These correspond to $m_g = 2$, more so when $\gamma_1 = 4$ than $\gamma_1 = 0$.

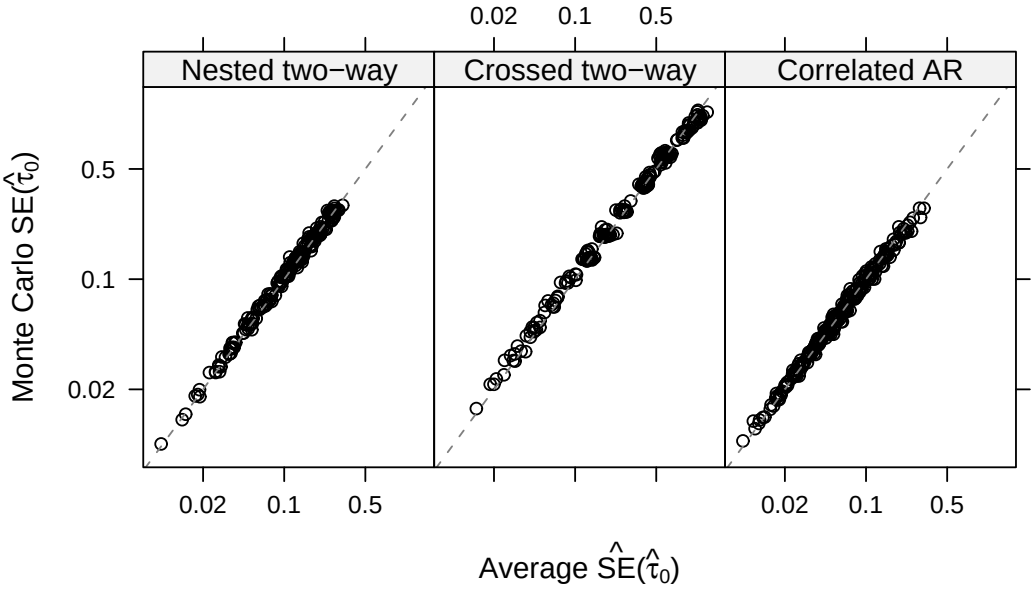


Figure 3.27: Average estimated SEs versus Monte Carlo SEs for all combinations of simulation parameters for the binary nested two way (3.3), crossed two-way (3.4), and AR correlated (3.8) models. (Note: log scale used for both axes.)

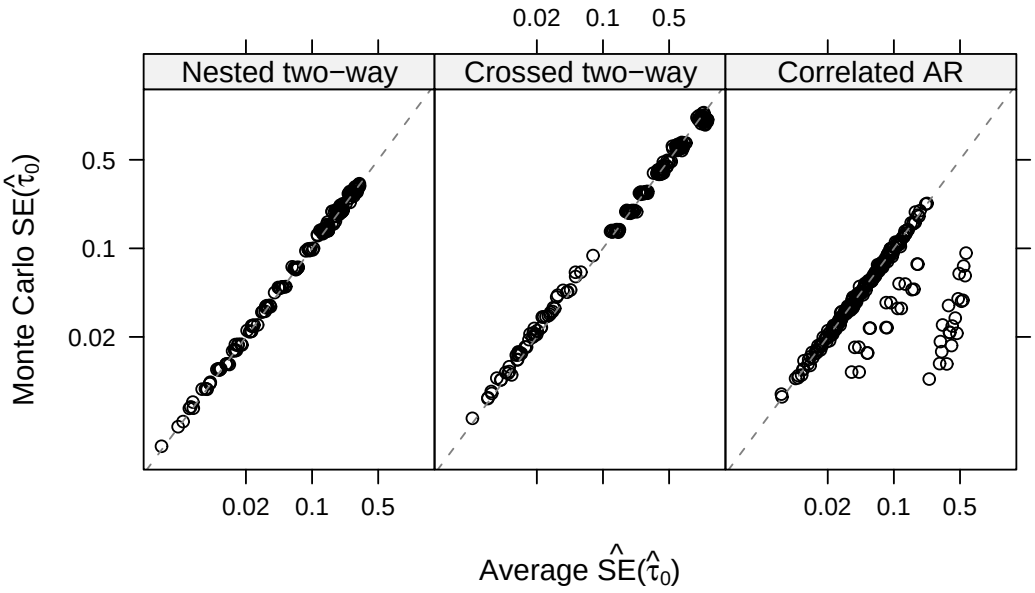


Figure 3.28: Average estimated SEs versus Monte Carlo SEs for all combinations of simulation parameters for the Poisson nested two way (3.3), crossed two-way (3.4), and AR correlated (3.8) models. (Note: log scale used for both axes.)

For the binary random coefficient model, the estimated SEs appeared generally adequate (Figure 3.29). For the Poisson random coefficients model, the estimated SEs were generally adequate, except when $\gamma_1 = 4$ and $\gamma_2 = 0$, where they severely underestimated the Monte Carlo SEs (Figure 3.30). This corresponded to the case where PQL estimation failed, as already discussed in section 3.1.5.1.

Hypothesis testing of fixed effects could also be performed using a likelihood ratio test, where the Laplace approximation is used to form a likelihood, as was done for testing the variance parameters. Lee & Nelder (1996) suggest the use of the h-likelihood, $h = \log f_{Y,U}$, for testing fixed effects, which ignores a correction term in the Laplace approximation. This has not been explored here.

3.3 Discussion

As already indicated, the large estimation biases in these simulation studies are not representative of estimation biases to be expected in the analysis of any real-life datasets. The two main factors affecting the estimation biases throughout these

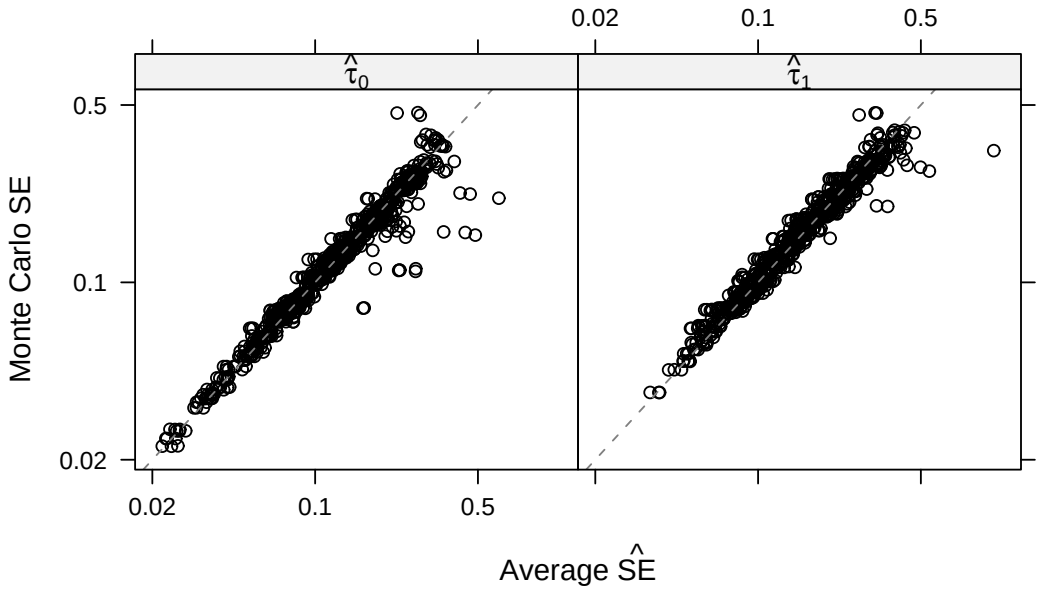


Figure 3.29: Average estimated SEs versus Monte Carlo SEs for τ_i for all combinations of simulation parameters for the binary random coefficients model (3.7). (Note: log scale used for both axes.)

studies are the group size(s) and the magnitude of the variance parameters. The rule of thumb proposed by Breslow (2003) to indicate whether PQL provides "reliable" estimation appears to be inadequate, as discussed earlier. For auto-regressively correlated data, increasing the group size has much less effect on the bias since the "effective" group size does not increase at the same rate. With respect to hypothesis testing and estimation of SEs for fixed effects, PQL appears to do reasonably well, even when there are strong estimation biases. PQL may be adequate for hypothesis testing of fixed effects against a null hypothesis of $H_0 : \tau_i = 0$. In the analysis of experimental data, the detection of treatment differences is often the main focus, and so PQL could be useful in this regard. One caveat of the simulation studies performed in this chapter is that, for all designs, equal group sizes were used: clearly, in real-life datasets, especially observational studies, group sizes would often not be equal. However, it is anticipated that the estimation biases seen here, for a given design and group size, would generally reflect the estimation biases one would expect from an equivalent design with unequal group sizes, but with an average group size equal to the given group size.

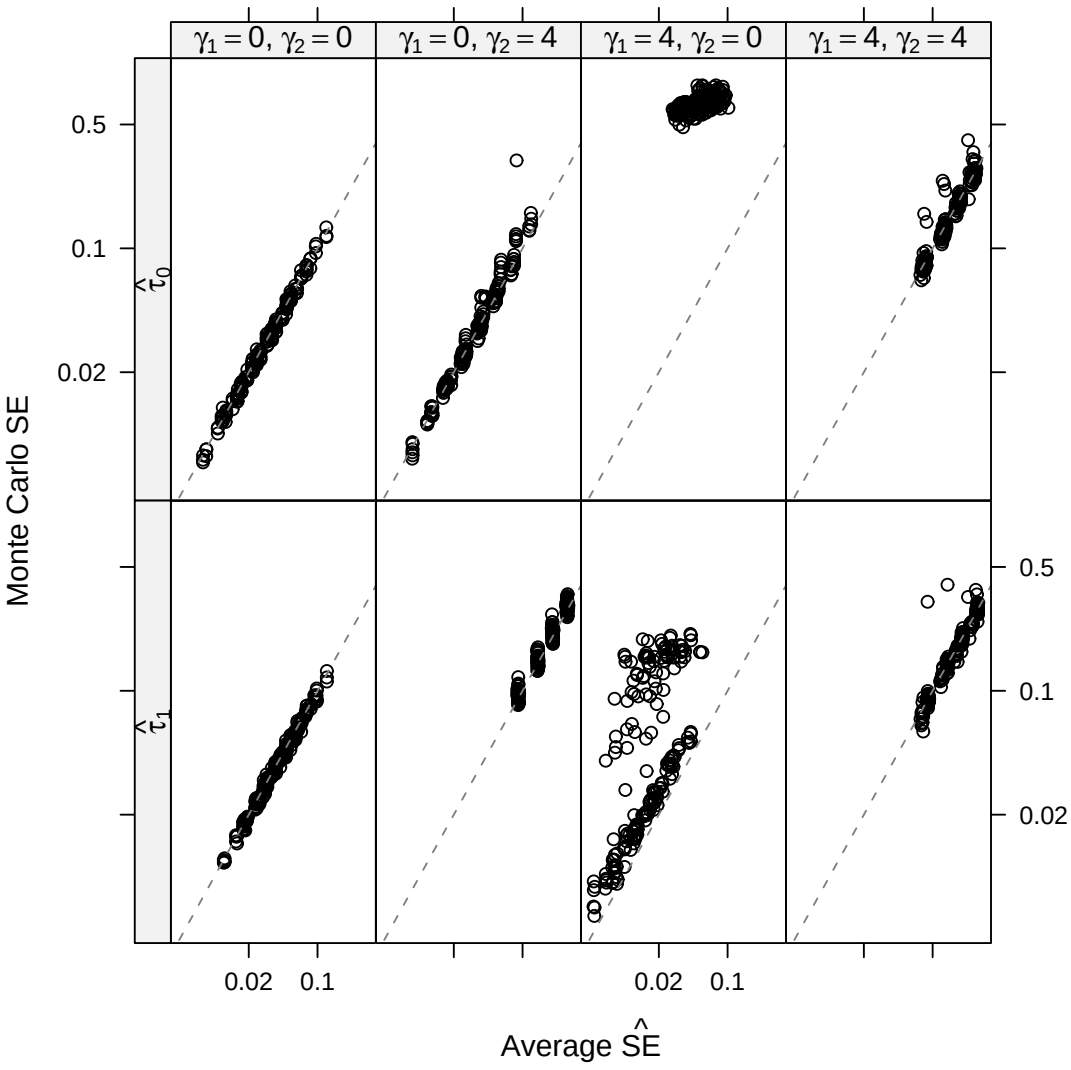


Figure 3.30: Average estimated SEs versus Monte Carlo SEs for τ_i for all combinations of simulation parameters for the Poisson random coefficients model (3.7), by γ_1 and γ_2. (Note: log scale used for both axes.)

It might be useful to speculate on where PQL may be adequate, with respect to estimation bias, in an agricultural and/or biological applications. Firstly, consider the analysis of non-normal data from an RCBD design in an agricultural setting, where there are less than 20 blocks and greater than 5 treatments, say. Owing to the small number of groups (blocks) and not too small group size (block size or number of treatments), it is expected that PQL estimation biases will not be large, even for sparse Poisson data or binomial data with small denominators. This is actually demonstrated in a later chapter (section 5.2.5) using a simulation study with

10 blocks and 5 treatments. For a complete block design, treatment contrasts are "within-group" comparisons. In the one-way classification model (3.1), the bias for the within-group coefficient β_1 was less than the between-group coefficient β_2 in the binary model, and was negligible in the Poisson model. Therefore, it may be expected that estimation of treatment contrasts will be less affected by bias than estimation of the overall mean. In the case of incomplete block or other unbalanced designs, treatment contrasts do involve between-group comparisons, and so may incur greater estimation bias than treatment contrasts for complete block designs. This interpretation of the results, however, needs to be qualified by the fact that a GLMM is a non-linear model, and therefore orthogonality of parameters with respect to design, such as between treatments and blocks in a randomised complete block design, does not equate to orthogonality between the respective parameter estimates. For experimental designs with multiple strata, and where the average number of units per strata is not too small (e.g. >10), PQL may still provide adequate estimators, provided the variance components are not too large (e.g. $\gamma_i < 1$).

For the analysis of simple experimental designs, the estimation of variance parameters associated with design strata is often not of direct interest. However, analysis of repeated measures data and allowing for spatial trends requires the fitting of correlated random effects. Given the results of the correlated AR study in section 3.1.5.2, there may be more significant estimation biases incurred by PQL for such analyses, especially where the correlations between random effects are low. As discussed at the end of section 3.1.5.2, as the correlation between random effects decreases, the "effective" group size tends to an extreme of 1. The fitting of cubic smoothing splines to model trends using a linear mixed model (Verbyla *et al.*, 1999), not investigated in this chapter, may also be expected to incur estimation bias problems, since such models often involve fitting a large number of random effects, and hence few observations per random effect.

Given the PQL bias problems in under-estimating the variance parameters, it is expected that PQL will be often inadequate where the main objective is estimating

variance components or functions of the variance components. Such a case is in the analysis of breeding data, with the dual problems of a small average group size (offspring per sire), and a large variance component relating to groups (sires). The biases would be especially severe for binomial data with small denominators or sparse Poisson data. The heritabilities will be under-estimated in general, although, as in Engel & Buist (1998) and the simulation study of section 3.1.4.4, the PQL negative bias may be offset by a positive bias induced by having a large number of fixed effects in the model. Where the prediction of random effects was of interest, such as in the prediction of spatial or temporal trends, the underestimation of variance parameters would also lead to under-estimation of the random effects. Although in section 3.1.5.2, the spatial correlation parameter γ_ρ was less affected by bias than the spatial variance γ_1, so under- or over-smoothing of spatial or temporal trends may not be a large issue.

For a final assessment of the benefits and risks in using PQL for fitting GLMMs, it is critical to compare the performance of PQL against other GLMM approaches. This is deferred to Chapter 5, where it is considered using a series of case studies.

Chapter 4

The HGLM approach of Lee and Nelder

This chapter discusses the use of the approximate likelihood methodology of Lee & Nelder (2001, 2006) for HGLMs. HGLMs include GLMMs as a special case where the random effects are normally distributed. Lee & Nelder argue that their HGLM approach performs better for GLMMs than PQL, with respect to lower estimation biases. Like PQL, their HGLM approach is an approximate likelihood approach which is based on the Laplace approximation. This chapter examines their HGLM approach, including an examination of computational aspects as well as simulation studies to compare the estimation biases of the HGLM approach against PQL.

4.1 Review of the HGLM methodology, and comparison with PQL

The HGLM approach of Lee & Nelder, in the context of GLMMs, has already been outlined in section 2.1.2. This section reviews the key formulae and contrasts the HGLM approach to PQL.

The so-called "h-likelihood", $h = \log f_{Y,U}$, is the cornerstone of inference using the HGLM methodology. Two likelihood expressions, derived from the h-likelihood, are

proposed for making inference concerning the fixed effects $\boldsymbol{\tau}$ and the variance parameters $\boldsymbol{\gamma}$.

For inference concerning $\boldsymbol{\tau}$ given $\boldsymbol{\gamma}$, the HGLM approach uses the following (first order) Laplace approximation (section 1.3.2.1) to approximate the true likelihood ℓ:

$$\begin{aligned} p_u(h) &= \left(h - \frac{1}{2} \log \left| -\frac{1}{2\pi} \frac{\partial^2 h}{\partial \boldsymbol{u} \partial \boldsymbol{u}^T} \right| \right)_{\tilde{\boldsymbol{u}}_{\tau,\gamma}} \\ &\approx \left(h - \frac{1}{2} \log \left| \boldsymbol{Z}^T \boldsymbol{W} \boldsymbol{Z} + \boldsymbol{G}^{-1} \right| + \frac{b}{2} \log 2\pi \right)_{\tilde{\boldsymbol{u}}_{\tau,\gamma}}, \end{aligned} \tag{4.1}$$

where $\boldsymbol{W}$ is a diagonal matrix of GLM weights, $\tilde{\boldsymbol{u}}_{\tau,\gamma}$ is the mode of h given $\boldsymbol{\tau}$ and $\boldsymbol{\gamma}$, and $b = \dim(\boldsymbol{u})$. Note that there is no closed expression for $\tilde{\boldsymbol{u}}_{\tau,\gamma}$, but it can be found iteratively using Fisher scoring, for instance. In contrast, PQL uses h alone for inference concerning $\boldsymbol{\tau}$, which assumes that $\boldsymbol{W}$ varies little with $\boldsymbol{\tau}$ and so the $\log \left| \boldsymbol{Z}^T \boldsymbol{W} \boldsymbol{Z} + \boldsymbol{G}^{-1} \right|$ term can be ignored. Note that an approximation sign $\approx$ has been used on the second line of (4.1) since, in general, $\partial^2 h / \partial \boldsymbol{u} \partial \boldsymbol{u}^T = \boldsymbol{Z}^T \boldsymbol{W} \boldsymbol{Z} + \boldsymbol{G}^{-1} + \boldsymbol{R}$, where

$$\boldsymbol{R} = \sum (y_i - \mu_i) \, \boldsymbol{z}_i \frac{\partial}{\partial \boldsymbol{u}} \left(\frac{1}{\phi a_i v(\mu_i) g'(\mu_i)} \right)$$

is a remainder term, and is only equal to $\mathbf{0}$ for canonical links, but otherwise has expectation $\mathbf{0}$.

For inference concerning $\boldsymbol{\gamma}$, the HGLM approach uses a similar Laplace-type approximation as the one for $\boldsymbol{\tau}$:

$$\begin{aligned} p_\beta(h) &= \left(h - \frac{1}{2} \log \left| -\frac{1}{2\pi} \frac{\partial^2 h}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right| \right)_{\tilde{\boldsymbol{\beta}}_\gamma} \\ &\approx \left(h - \frac{1}{2} \log \left| \boldsymbol{Z}^T \boldsymbol{W} \boldsymbol{Z} + \boldsymbol{G}^{-1} \right| - \frac{1}{2} \log \left| \boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X} \right| + \frac{p+b}{2} \log 2\pi \right)_{\tilde{\boldsymbol{\beta}}_\gamma}, \end{aligned} \tag{4.2}$$

where $\boldsymbol{\beta} = (\boldsymbol{\tau}^T, \boldsymbol{u}^T)^T$, $\boldsymbol{V} = \boldsymbol{Z} \boldsymbol{G} \boldsymbol{Z}^T + \boldsymbol{W}^{-1}$ and $p = \dim(\boldsymbol{\tau})$. The vector $\tilde{\boldsymbol{\beta}}_\gamma$ equals $(\hat{\boldsymbol{\tau}}_\gamma^T, \tilde{\boldsymbol{u}}_{\tau,\gamma}^T)^T$, where $\hat{\boldsymbol{\tau}}_\gamma^T$ satisfies $\partial p_u(h) / \partial \boldsymbol{\tau} = \mathbf{0}$. In contrast, PQL uses some further approximations and assumptions to form a likelihood criterion for $\boldsymbol{\gamma}$:

1. Firstly, the conditional likelihood $2 \log f_{Y|U}$ can be replaced by the Pearson

chi-squared statistic, $\sum_i (y_i - \mu_i)^2 / V(\mu_i)$.

2. If $-1/2 \log |\boldsymbol{W}|$ is then added , an expression corresponding to the likelihood based on the working variate $\boldsymbol{\psi} = (\psi_1 \dots \psi_n)^T$, where $\psi_i = g(\mu_i) + g'(\mu_i)(y_i - \mu_i)$, is obtained,

$$-\frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}(\boldsymbol{\psi} - \boldsymbol{X}\boldsymbol{\tau})^T \boldsymbol{V}^{-1}(\boldsymbol{\psi} - \boldsymbol{X}\boldsymbol{\tau}) - \frac{1}{2}\log\left|\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right|, \qquad (4.3)$$

where $\boldsymbol{V} = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T + \boldsymbol{W}^{-1}$.

3. The expression in (4.3) has the form of a REML likelihood for a normally distributed working variate $\boldsymbol{\psi}$ with mean $\mathbf{0}$ and variance $\boldsymbol{V} = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T + \boldsymbol{W}^{-1}$. To derive updating equations, it is further assumed that $\boldsymbol{W}$ is not a function of $\boldsymbol{\gamma}$, which enables the use of standard updating equations for a linear mixed model.

The two HGLM likelihoods above, (4.1) and (4.2), correspond to first order approximations. In some cases, a second order approximation is required for ℓ:

$$p_u^s(h) \quad = \quad p_u(h) - F/24,$$

with the same correction for $p_\beta(h)$,

$$p_\beta^s(h) = p_\beta(h) - F/24,$$

where $-F/24$ represents the additional second order correction factor, described in section 4.3.1 below.

As part of their HGLM approach, and implemented in the HG-system of the `GenStat` statistical package (Payne *et al.*, 2006), Lee & Nelder (2001, 2006) define different levels of approximation for inference concerning $\boldsymbol{\tau}$ and $\boldsymbol{\gamma}$. These are denoted `HG`(m, d)`,` where $m, d = 0, 1, 2$, $m \leq d$, where m and d indicate the level of approximation for inference concerning $\boldsymbol{\tau}$ and $\boldsymbol{\gamma}$ respectively. These are shown in Table 4.1. For instance, HG(0,0) corresponds to a PQL approach, HG(0,1) uses h for inference concerning $\boldsymbol{\tau}$,

like PQL, but uses $p_\beta(h)$ for inference concerning $\boldsymbol{\gamma}$, and so on. Noh & Lee (2007) provide details of the updating equations required to apply an iterative solution for each level of approximation. The iterative solution they propose requires alternate estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, similar to an EM approach for PQL.

Approximation	*Likelihood for* $\boldsymbol{\tau}$	*Likelihood for* $\boldsymbol{\gamma}$
HG(0,0)	h	PQL*
HG(0,1)	h	$p_\beta(h)$
HG(0,2)	h	$p^s_\beta(h)$
HG(1,1)	$p_u(h)$	$p_\beta(h)$
HG(1,2)	$p_u(h)$	$p^s_\beta(h)$
HG(2,2)	$p^s_u(h)$	$p^s_\beta(h)$

Table 4.1: The different levels of approximate inference in the HGLM approach proposed by Lee & Nelder (2001, 2006), with the corresponding (approximate) likelihood expressions for fixed effects $\boldsymbol{\tau}$ and variance components $\boldsymbol{\gamma}$. (*: Note that there is no explicit expression for the likelihood associated with PQL; only the score equations at each iteration are defined.)

Note also that the functions $p_\beta(h)$ and $p_u(h)$ can be derived as approximations of the posterior distributions for $\boldsymbol{\tau}$ and $\boldsymbol{\gamma}$, but only if the prior densities of the fixed effects $\boldsymbol{\tau}$ and variance parameters $\boldsymbol{\gamma}$ are flat. For instance, consider the posterior distribution of $\boldsymbol{\gamma}$. Assuming that the prior $f_\gamma(\boldsymbol{\gamma}) \propto 1$, the posterior distribution of $\boldsymbol{\gamma}$ simplifies to

$$f_{\gamma|Y}(\boldsymbol{\gamma}|\boldsymbol{y}) \propto f_{Y|\gamma}(\boldsymbol{y}|\boldsymbol{\gamma}) f_\gamma(\boldsymbol{\gamma}) = f_{Y|\gamma}(\boldsymbol{y}|\boldsymbol{\gamma}).$$

If the conditional prior $f_{\tau|\gamma}(\boldsymbol{\tau}|\boldsymbol{\gamma}) \propto 1$, then this further simplifies to the expression which $p_\beta(h)$ is approximating, viz.

$$\begin{aligned} f_{Y|\gamma}(\boldsymbol{y}|\boldsymbol{\gamma}) &= \int f_{Y|\beta,\gamma}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\gamma}) f_{\beta|\gamma}(\boldsymbol{\beta}|\boldsymbol{\gamma}) d\boldsymbol{\beta} \\ &= \int f_{Y|\beta,\gamma}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\gamma}) f_{U|\gamma}(\boldsymbol{u}|\boldsymbol{\gamma}) f_\tau(\boldsymbol{\tau}) d\boldsymbol{\beta} \\ &= \int f_{Y|\beta,\gamma}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\gamma}) f_{U|\gamma}(\boldsymbol{u}|\boldsymbol{\gamma}) d\boldsymbol{\beta}. \end{aligned}$$

4.2 First order HGLM approaches

This section discusses the implementation and performance of the first order HGLM approaches, that is, the approximations HG(i,1) where $i \leq 1$. Firstly, an implementation in Fortran 90 using quasi-Newton optimisation with numerical derivatives is described. Secondly, results of selected simulation studies to compare the magnitude of the estimation biases against PQL are shown. Finally, the calculation of the analytical derivatives are presented, along with the corresponding updating equations.

4.2.1 A Fortran 90 implementation with numerical derivatives

An implementation of the HGLM approach for GLMMs was written in Fortran 90, since the current GenStat implementation was found to be too slow for performing extensive simulations. Initially, this Fortran code only implemented the first-order HGLM approaches, however, additional code was added later which implemented the second-order approaches described in the next section (section 4.3.1).

As noted above, Noh & Lee (2007) presented updating equations for implementing the HGLM approach with GLMMs for the approximations listed in table 4.1 . The expressions for derivatives and updating equations presented in this paper were difficult to follow (using a draft version of this paper kindly supplied by Dr. Noh), as discussed below (section 4.2.3). Alternative techniques for optimising the likelihood criteria, which did not require analytical derivatives, were considered. Initially the Nelder & Mead (1964) simplex optimisation technique was used, as implemented in Alan Miller's Fortran 90 code[1]. However, this optimisation technique was found to be unduly slow, and so was replaced by a quasi-Newton approach called L-BFGS-B (Zhu *et al.*, 1997)[2]. The L-BFGS-B approach required expressions for both the likelihood and its first order derivatives. The latter were calculated using finite differences. That is, if a criterion $f(x)$ was to be optimised over a parameter x, then the first derivative was calculated as $f'(x) = \{f(x+c) - f(x)\}/c$, with an increment c chosen

[1] Available from `http://users.bigpond.net.au/amiller/minim.f90`
[2] Available from `http://www.ece.northwestern.edu/~nocedal/lbfgsb.html`

empirically as 10^{-5} (the often recommended value for finite difference derivatives, the squareroot of machine precision, was found to be too small here).

To implement the HG(0,1) approximation, the likelihood criterion $p_\beta(h)$ was maximised over $\boldsymbol{\gamma}$ using the L-BFGS-B optimisation approach (with numerical derivatives, as above). To calculate $p_\beta(h)$ at a given value of $\boldsymbol{\gamma}$, the estimates $\hat{\boldsymbol{\tau}}_\gamma$ and $\tilde{\boldsymbol{u}}_{\tau,\gamma}$ were required, and were determined by repeatedly forming and solving the mixed model equations,

$$\begin{bmatrix} \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z}+\boldsymbol{G}^{-1} \end{bmatrix} \begin{pmatrix} \hat{\boldsymbol{\tau}} \\ \tilde{\boldsymbol{u}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{\psi} \\ \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{\psi} \end{pmatrix}, \tag{4.4}$$

until convergence of $\hat{\boldsymbol{\tau}}$ and $\tilde{\boldsymbol{u}}$, with $\hat{\boldsymbol{\tau}}_\gamma$ and $\tilde{\boldsymbol{u}}_{\tau,\gamma}$ being the final values of $\hat{\boldsymbol{\tau}}$ and $\tilde{\boldsymbol{u}}$ at convergence. Here, $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_n)^T$ is the "working variate" with elements $\psi_i = \eta_i + g'(\mu_i)(y_i - \mu_i)$. The formation of these mixed model equations, at each iteration, required the recalculation of the quantities $\boldsymbol{\psi}$ and $\boldsymbol{W}$, as well as $\eta_i = g^{-1}(\mu_i)$ and μ_i, $i = 1 \ldots n$, using the current estimates $\hat{\boldsymbol{\tau}}$ and $\tilde{\boldsymbol{u}}$. Initial values of $\hat{\boldsymbol{\tau}} = \boldsymbol{0}$, $\tilde{\boldsymbol{u}} = \boldsymbol{0}$ and $\gamma_i = 0.1$, $i = 1, \ldots, q$, were used. Good initial values of μ_i, $i = 1, \ldots, n$, were also required to achieve convergence, and were derived based on the recommendations given in McCullagh & Nelder (1989).

For the HG(1,1) approximation, a two stage approach was required. As for HG(0,1), the likelihood criterion $p_\beta(h)$ was maximised over $\boldsymbol{\gamma}$ using L-BFGS-B with numerical derivatives. As for HG(0,1), the estimate $\hat{\boldsymbol{\tau}}_\gamma$ was required to calculate $p_\beta(h)$ at a given $\boldsymbol{\gamma}$. However, for this approximation, $\hat{\tau}_\gamma$ was the maximum of the criterion $p_u(h)$ with respect for $\boldsymbol{\tau}$, for given $\boldsymbol{\gamma}$. Hence, for a given value of $\boldsymbol{\gamma}$, the criterion $p_u(h)$ was maximised with respect to $\boldsymbol{\tau}$, also using L-BFGS-B with numerical derivatives. The calculation of $p_u(h)$ and $p_\beta(h)$ required, at a given $\boldsymbol{\tau}$ and $\boldsymbol{\gamma}$, an estimate $\tilde{\boldsymbol{u}}_{\tau,\gamma}$, which was calculated from repeated evaluation of

$$\tilde{\boldsymbol{u}} \;=\; \left(\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z}+\boldsymbol{G}^{-1}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{W}(\boldsymbol{\psi}-\boldsymbol{X}\boldsymbol{\tau})$$

until convergence, where $\tilde{\boldsymbol{u}}_{\tau,\gamma} = \tilde{\boldsymbol{u}}$ at convergence. The quantities $\boldsymbol{\psi}$, η_i, μ_i , $i =$

$1 \ldots n$, and $\boldsymbol{W}$ are as given above, and were re-calculated at each re-evaluation of $\tilde{\boldsymbol{u}}$. Initial values were as for the HG(0,1) approximation.

The exercise of writing computationally efficient code in a lower level language, such as Fortran, made apparent the need to decompose mathematical expressions, particularly those involving matrices, into simple floating point scalar calculations. Initial code was written using "dense" matrix operations: that is, matrices such as $\boldsymbol{Z}$ were stored as written, and computations were performed using Fortran 90's matrix capabilities. However, it was soon realized that this wasted a great deal of computational resources, so much so that the program was much slower to run than notoriously computational intensive MCMC approaches. The code was subsequently revised to implement sparse matrix calculations: only the non-zero elements of large sparse matrices such as $\boldsymbol{Z}$ were stored, and their sparcity was exploited in the required matrix multiplications, such as in the calculation of $\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z}$. Inversion and calculation of log-determinants of the square matrix

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z} + \boldsymbol{G}^{-1} \end{bmatrix}$$

and its RHS lower block component, $\boldsymbol{C}_{22} = \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z} + \boldsymbol{G}^{-1}$, were performed using the public domain sparse matrix library of Misztal (1999)[3].

Cross-checking of the estimates obtained from the above Fortran 90 implementation for some simulated datasets, with those obtained using the current GenStat implementation, showed some stark differences, particularly for simulated datasets generated for nested two-way classification designs (section 4.2.2.2). Fortunately, other GLMM approaches were available, such as the `lmer` function in the `lme4` R package (Bates, 2007), which also maximised a first order Laplace approximation to the likelihood, similar to the HGLM approach described above, but with no REML-like correction. The use of these other GLMM approaches on the simulated datasets allowed us to verify that our estimates were correct. Further, by constructing a profile likelihood for

[3]This library is available from `http://nce.ads.uga.edu/~ignacy/newprograms.html`

$\boldsymbol{\gamma}$ using the GenStat implementation, it was shown that GenStat was not optimising the approximate likelihood $p_\beta(h)$ that it was reporting. At the time of writing, the reason for the GenStat estimation problem has not been resolved, but it presumably indicates a problem with the computations of the likelihood derivatives and updating equations. A derivation of these updating equations, not yet implemented in code, is presented below in section 4.2.3. The problem has also been reported to the GenStat developers.

4.2.2 Performance in simulation studies compared to PQL

Two of the designs used for the simulation studies in Chapter 3 are re-examined, to compare the estimation biases of the first order HGLM approaches, HG(0,1) and HG(1,1), with PQL.

As in Chapter 3, 200 simulated datasets were generated and analysed according to the respective model for each combination of simulated parameter values. The link function g represented the logit and logarithmic link for binary and Poisson data respectively. PQL was implemented using `ASReml` version 2 (Gilmour *et al.*, 2006). In keeping with the simulation studies conducted in Chapter 3, the estimates of the variance parameters were constrained to be positive.

4.2.2.1 One-way classification design

The one-way classification design (3.1) for binary or Poisson data y_{ij} was re-examined, minus the covariates, where $\mu_{ij} = E(y_{ij}|u_i)$, $i = 1 \ldots b_g$, $j = 1 \ldots m_g$ and

$$g(\mu_{ij}) = \tau_0 + u_i, \qquad u_i \sim N(0, \gamma_1). \tag{4.5}$$

The simulated parameter values used are shown in Table 4.2.

For the binary model, the negative biases for $\hat{\gamma}_1$ for the HG(0,1) and HG(1,1) approximations were smaller in magnitude than the corresponding biases for PQL for all values of m_g, γ_1 and τ_0 (Figure 4.1). As with PQL, the negative bias for $\hat{\gamma}_1$ decreased

Parameter	*Binary model*	*Poisson model*
b_g	50, 100, 200, 500	...
m_g	2, 4, 8, 16	...
γ_1	0.25, 1, 4, 9	...
τ_0	0, 1	...

Table 4.2: Values of the simulation parameters used for binary and Poisson models for the one-way classification study (4.5) comparing first order HGLM approaches and PQL. The use of ... in the 3rd (Poisson model) column indicates the same values were used as for the binary model.

in magnitude with increasing m_g. However, in contrast to PQL, the negative bias for $\hat{\gamma}_1$ did *not* increase with γ_1 for either of the HGLM approximations. The absolute magnitude of the bias of $\hat{\gamma}_1$ for either HGLM approximation was never more than 50%, in contrast to PQL, where the absolute bias exceeded 80% when $\gamma_1 = 9$. When $\tau_0 = 0$, there was negligible difference in the bias between the HG(0,1) and HG(1,1) estimators, and only a slight difference when $\tau_0 = 1$ and $m \leq 4$. There was also a noticeable increase in the magnitude of the negative bias with b_g (not shown) for the HG(i,1) approximations, although this had generally flattened out by $b_g = 500$. The increase in the negative bias with b_g reflected positive skewness of the distribution of $\hat{\gamma}_1$ at lower b_g, hence $b_g = 500$ was chosen in Figure 4.1 as an approximation of the asymptotic biases ($b_g \to \infty$).

Figure 4.2 shows the biases for $\hat{\tau}_0$ in the binary model where $\tau_0 = 1$ (there was no apparent bias for any HGLM approach when $\tau_0 = 0$). Note that there is a distinct difference between HG(0,1) and HG(1,1) approximations: the biases for HG(1,1) are negligible, whereas the negative bias for HG(0,1) increases with increasing γ_1, as for PQL.

For the Poisson model, like the binary model, the negative bias for $\hat{\gamma}_1$ for either HGLM approximation did not increase in magnitude with γ_1 either, and was less than 5% at $b_g = 500$ and $\tau_0 = 0$ (Figure 4.3). The positive bias for $\hat{\tau}_0$ ($\tau_0 = 0$) for HG(0,1) was slightly smaller than that of PQL, but the bias of $\hat{\tau}_0$ for HG(1,1) was almost negligible (Figure 4.4). (The biases at $\tau_0 = 1$ were similar to those shown for $\tau_0 = 0$, but less pronounced.)

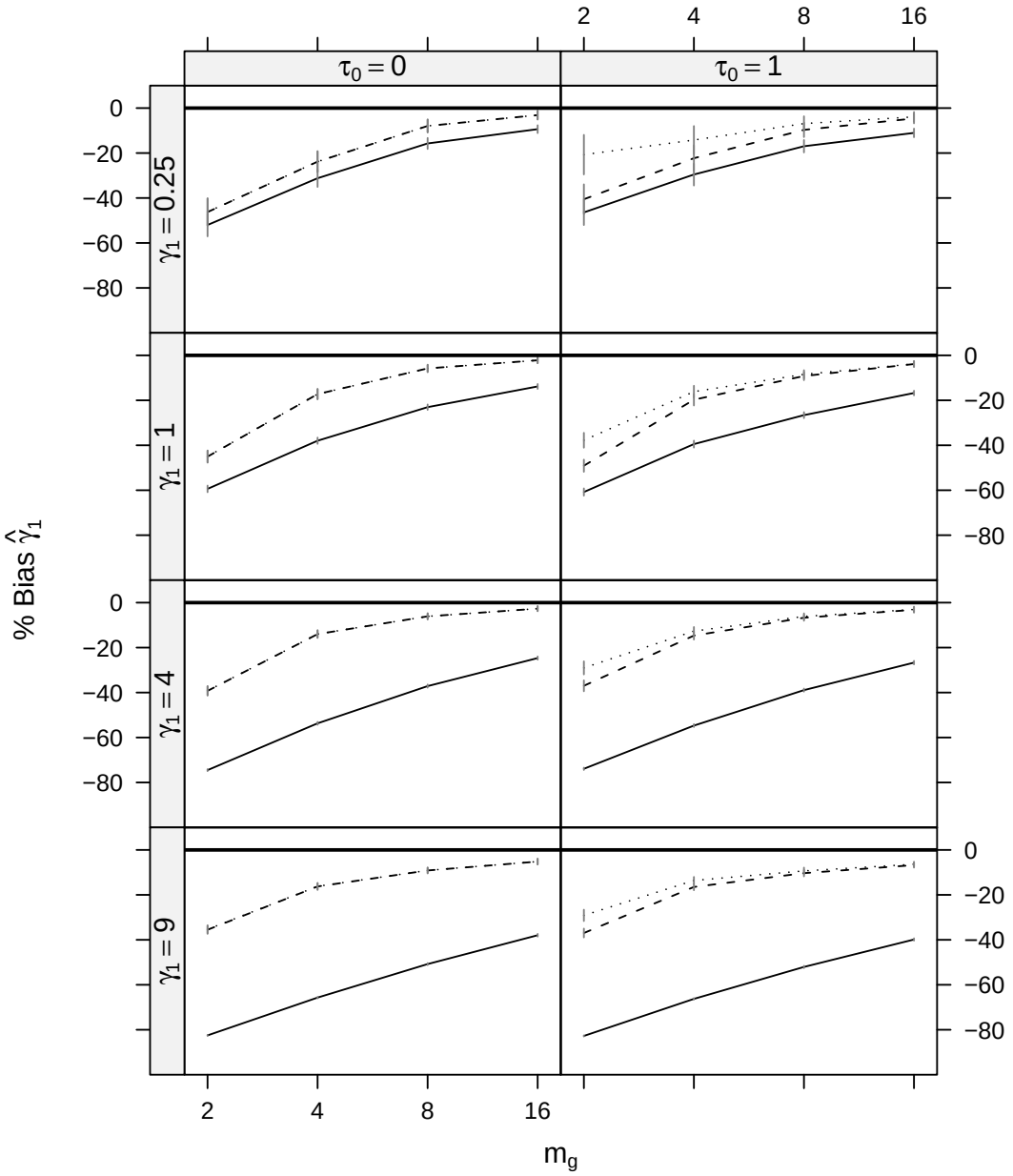


Figure 4.1: Interactions of the effects of m_g, γ_1 and τ_0 on the biases for $\hat{\gamma}_1$ for the binary one-way classification model (4.5) for $b_g = 500$ for PQL and first-order HGLM approximations. (PQL: solid; HG(0,1): dashed; HG(1,1): dotted) Error bars are $\pm$2SE.

In their papers, Lee & Nelder assert that the use of a first order approximation is adequate for estimating fixed effects, and second order approximations are only required for estimating variance parameters. The results of this study support their assertion, with low or negligible biases for the estimation of τ_0 for both binary and

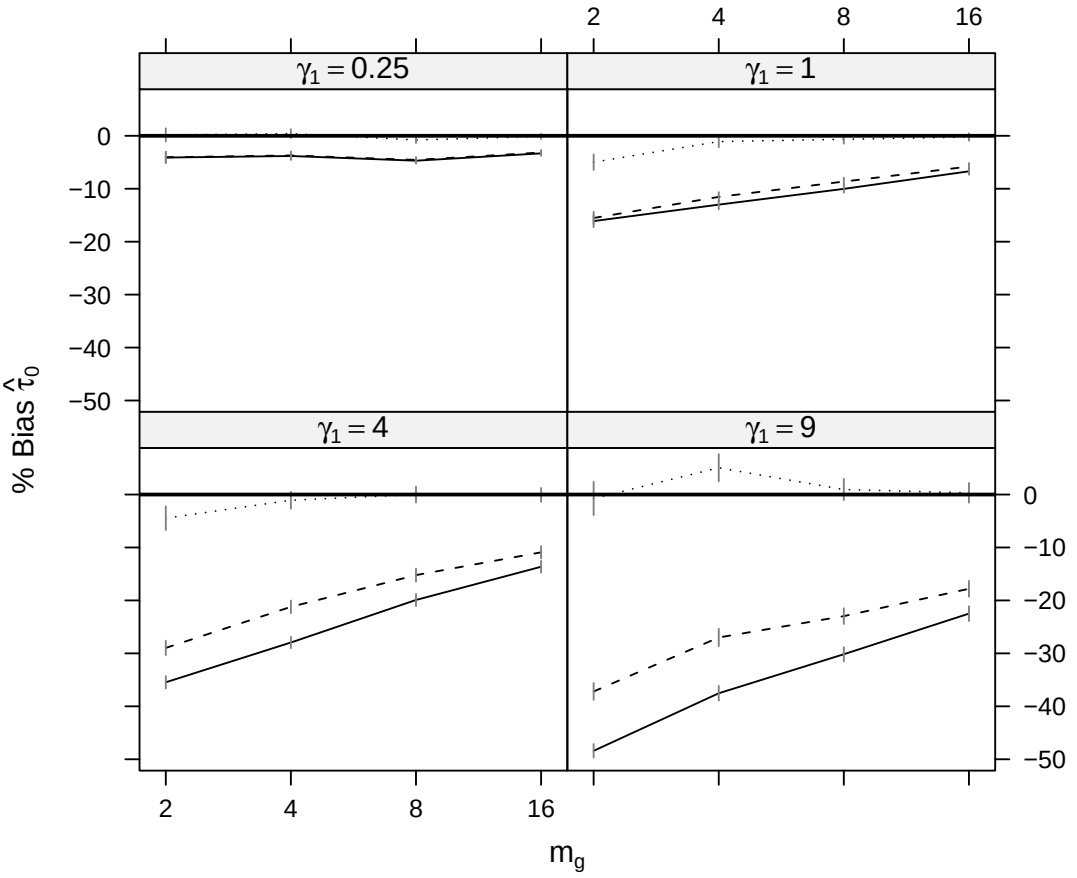


Figure 4.2: Interactions of the effects of m_g and γ_1 on the biases for $\hat{\tau}_0$ for the binary one-way classification model (4.5) for $b_g = 500$ for PQL and the first order HGLM approximations. (PQL: solid; HG(0,1): dashed; HG(1,1): dotted). Error bars are $\pm$2SE.

Poisson data. For the Poisson model, the use of HG(1,1) reduces the magnitude of the bias for the variance component $\hat{\gamma}_1$ to an adequately low level as well. It is important to note that the magnitude of the negative bias for the variance parameter estimator $\hat{\gamma}_1$ does not increase with γ_1 for either HGLM approach, although the negative (or positive for Poisson) bias for the intercept $\hat{\tau}_0$ does increase with γ_1 using HG(0,1) for either the binary or Poisson models.

It might be useful to consider the differences in the biases between PQL and the HGLM approximations in the light of their different likelihood criteria shown in Table 4.1. Firstly, consider the differences between the biases of PQL versus the biases of HG(0,1). As Table 4.1 shows, the only difference between PQL and the HG(0,1) is in

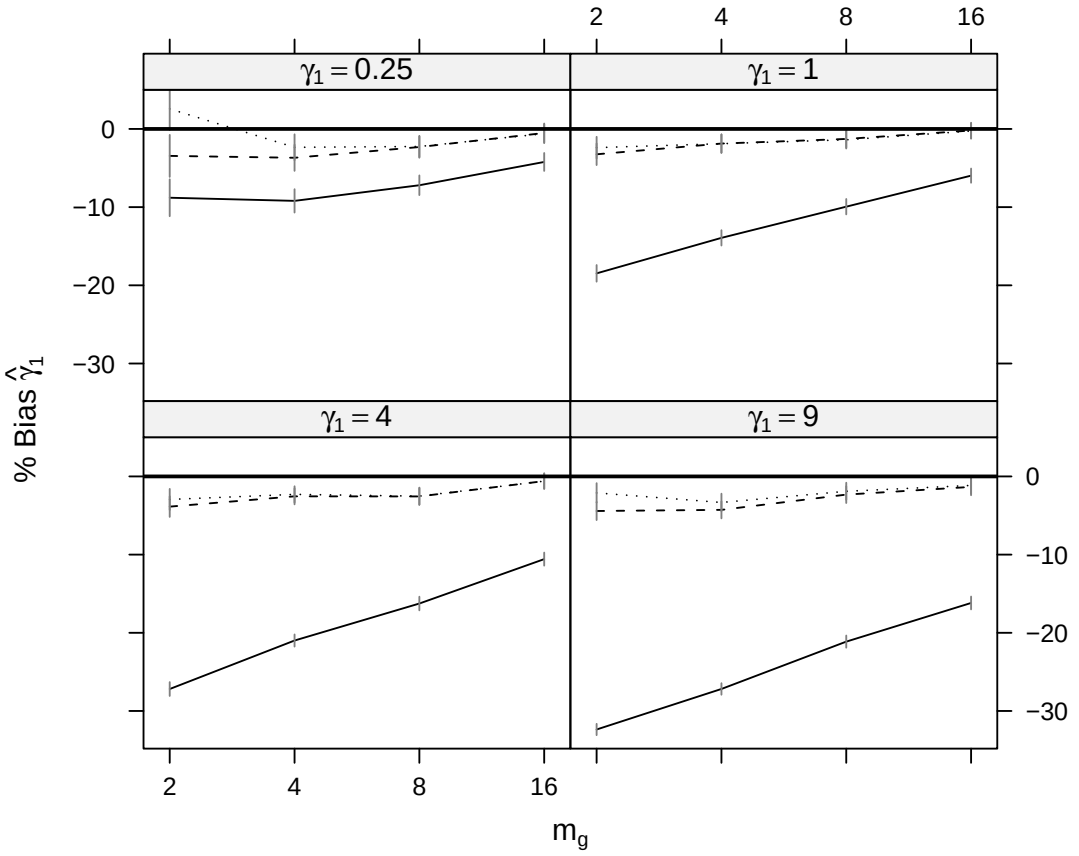


Figure 4.3: Interactions of the effects of m_g and γ_1 on the biases for $\hat{\gamma}_1$ for the Poisson one-way classification model (4.5) for $b_g = 500$ and $\tau_0 = 0$ for PQL and the first order HGLM approximations. (PQL: solid; HG(0,1): dashed; HG(1,1): dotted). Error bars are ±2SE.

the estimation of $\boldsymbol{\gamma}$. For HG(0,1), $p_\beta(h)$ is maximised with respect to $\boldsymbol{\gamma}$, whereas for PQL an approximation to $p_\beta(h)$, at each iteration, is maximised instead, as described in points 1-3 on page 115. This approximation incurs additional negative bias, as seen by comparing the biases for $\hat{\gamma}_1$ for PQL against HG(0,1) in Figures 4.1 (binary) and 4.3 (Poisson). In addition, these figures show that this additional negative bias increases with the magnitude of the variance parameter γ_1. A reason for this increase in the bias with γ_1, or with $\boldsymbol{\gamma}$ for a general GLMM design, is given below in section 4.2.3.4.

Secondly, consider the differences in the biases between HG(0,1) and HG(1,1) approaches. As in Table 4.1, the only difference between HG(0,1) and HG(1,1) is in the

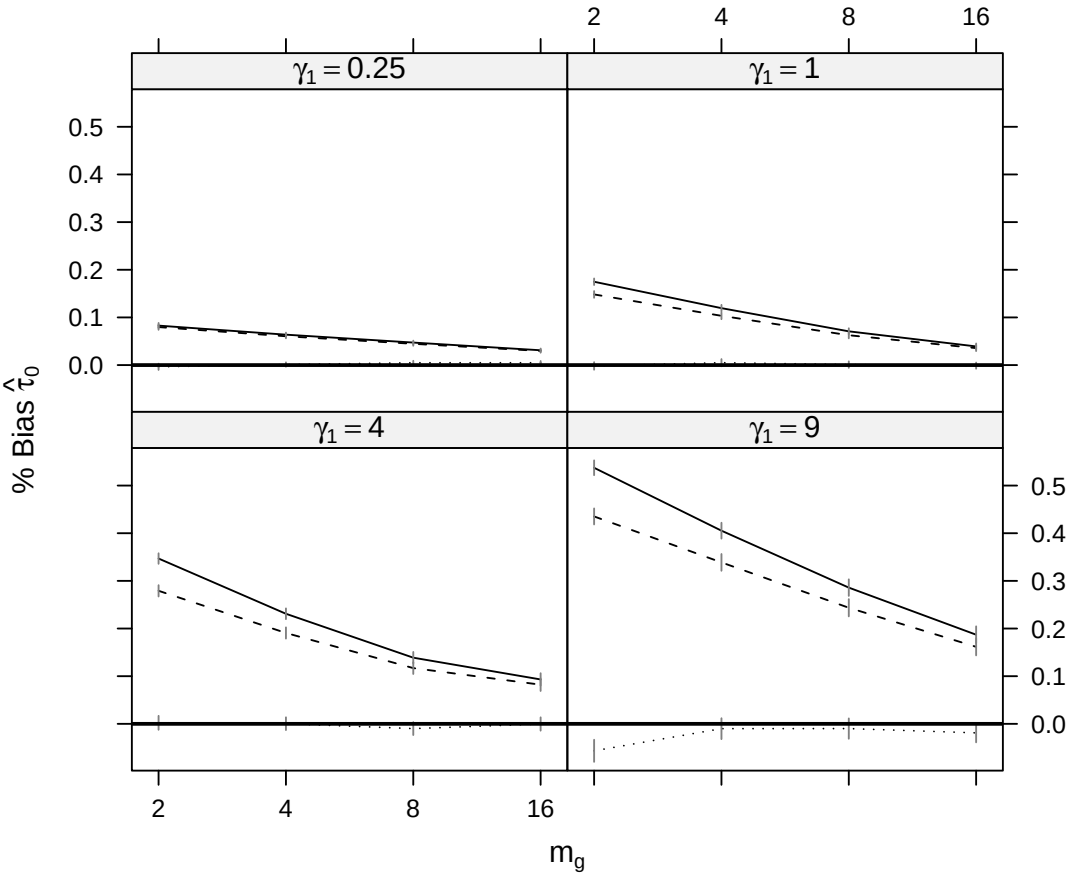


Figure 4.4: Interactions of the effects of m_g and γ_1 on the biases for $\hat{\tau}_0$ for the Poisson one-way classification model (4.5) for $b_g = 500$ and $\tau_0 = 0$ for PQL and the first order HGLM approximations. (PQL: solid; HG(0,1): dashed; HG(1,1): dotted). Error bars are ±2SE.

estimation of $\boldsymbol{\tau}$ (given $\boldsymbol{\gamma}$). For HG(1,1), the criterion $p_u(h)$ is maximised with respect to $\boldsymbol{\tau}$ whereas, for HG(0,1), h is maximised instead, ignoring the $-1/2 \log \left| \boldsymbol{Z}^T \boldsymbol{W} \boldsymbol{Z} + \boldsymbol{G}^{-1} \right|$ term. By comparing the biases for $\hat{\tau}_0$ for HG(0,1) against HG(1,1) in Figures 4.2 (binary) and 4.4 (Poisson), it is seen that the difference between the biases for both approximations increases with γ_1. This suggests that the importance of the ignored term, $-1/2 \log \left| \boldsymbol{Z}^T \boldsymbol{W} \boldsymbol{Z} + \boldsymbol{G}^{-1} \right|$, increases with γ_1. For the one-way classification model with no covariates, (4.5), the conditional means are equal for all units in a group, that is, $\mu_{ij} = \mu_i$. Therefore, this term simplifies to

$$-1/2 \log \left| \boldsymbol{Z}^T \boldsymbol{W} \boldsymbol{Z} + \boldsymbol{G}^{-1} \right| = -1/2 \sum_i \log \left(m_g w_i + \gamma_1^{-1} \right),$$

where $w_i = g'(\mu_i)^{-2}V(\mu_i)^{-1}$ are the GLM weights. Note that

$$\lim_{\gamma_1 \to 0} \log\left(m_g w_i + \gamma_1^{-1}\right) = \log\left(\gamma_1^{-1}\right),$$

which is independent of τ_0. So this term will change relatively little with τ_0 when γ_1 is small, and so ignoring this term will make little difference to the estimate of τ_0 when γ_1 is small. Conversely,

$$\lim_{\gamma_1 \to \infty} \log\left(m_g w_i + \gamma_1^{-1}\right) = \log\left(m_g w_i\right),$$

which obviously does depend on τ_0 via the GLM weights w_i. Therefore, as γ_1 increases, the change of $-1/2\sum_i \log\left(m_g w_i + \gamma_1^{-1}\right)$ with τ_0 will be larger in magnitude. For the binary logit model, where $w_i = \mu_i(1-\mu_i)$, smaller values of τ_0 in absolute value will be favoured by HG(0,1) compared to HG(1,1), since $1/2\sum_i \log\left(m_g w_i + \gamma_1^{-1}\right)$ is a decreasing function of $|\tau_0|$. For the Poisson log model, where $w_i = \mu_i$, higher values of τ_0 will be favoured by HG(0,1) compared to HG(1,1), since $1/2\sum_i \log\left(m_g w_i + \gamma_1^{-1}\right)$ is an increasing function of τ_0.

This intuition should apply generally for GLMMs in the estimation of $\boldsymbol{\tau}$. As $\boldsymbol{\gamma}$ decreases, $\boldsymbol{G}^{-1} \to \boldsymbol{\infty}$, and so $-1/2\log\left|\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z} + \boldsymbol{G}^{-1}\right| \to -1/2\log|\boldsymbol{G}^{-1}|$. As $\boldsymbol{\gamma}$ increases, $\boldsymbol{G}^{-1} \to \boldsymbol{0}$ and so $-1/2\log\left|\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z} + \boldsymbol{G}^{-1}\right| \to -1/2\log\left|\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z}\right|$, and therefore this term becomes increasingly important.

One-way classification for binary-logit data with $\tau_0 = 2$: A further set of simulations was conducted with the one-way classification (4.5) using $\tau_0 = 2$, with the same values of all the other simulation parameters as given in Table 4.2.

The estimation biases for the HG(0,1) approximation were similar to those observed when $\tau_0 = 0$ or 1 (Figure 4.5). However, the estimation biases for HG(1,1) when $m_g = 2$ were unexpectedly large and positive. Inspection of the individual estimates showed that the HG(1,1) approximation had high rates of divergence in these cases, that is, the HG(1,1) estimates of γ_1 and τ_0 took unexpectedly high values, well above

the HG(0,1) estimates and the true value of γ_1. For instance, when $\gamma_1 = 9$ and $m_g = 2$, divergence of the HG(1,1) estimator was observed in about 50% of cases, where the HG(1,1) estimates of γ_1 were between 50 and 100, well above the true value of 9. A sample of the datasets where the HG(1,1) estimates diverged were examined more closely. For each dataset, the profile likelihood $p_\beta(h)$ was plotted with respect to γ_1 for a range of values around the HG(1,1) estimate, and it was found that the HG(1,1) estimate corresponded with the observed mode of $p_\beta(h)$ with respect to γ_1. This confirmed that there was a finite maximum of $p_\beta(h)$ with respect to γ_1 for these datasets, and so presumably one could infer that this was true for all datasets where the HG(1,1) estimates diverged. The high rate of divergence observed here where $\tau_0 = 2$, however, is in line with the expected instability of the first order Laplace approximation in the extreme regions of the binary parameter space, noted by Breslow & Lin (1995) (see figure 2 on page 89 of Breslow & Lin (1995), and the associated description on page 88 of their paper).

Note that the HGLM approximations incorporate a REML-like correction in the likelihood criterion for γ_1, $p_\beta(h)$. Using $p_u(h)$ as the criterion for γ_1 instead provides a "non-REML" version. In order to test whether the divergence problem resulted from the inclusion of the REML-like correction, a non-REML HG(1,1) approximation was also implemented (which required a simple modification of the Fortran 90 code). A sample of simulated datasets where divergence occurred were examined, and it was found that the removal of the REML-like correction made little difference to the estimates, and so the problem with divergence remained. Using the non-REML version of HG(1,1) also allowed us to verify our estimates against estimates from other GLMM implementations. As noted earlier, other GLMM applications also use a Laplace approximation to the likelihood, but with no REML-like correction, and include the `lmer` and `glmmadmb` functions, in the packages `lme4` and `glmmadmb` respectively, for the `R` statistical package (R Development Core Team, 2008). The non-REML HG(1,1) estimates obtained using the modified Fortran implementation agreed with these other Laplace-based GLMM applications, `lmer` and `glmmadmb`, for both datasets where the HG(1,1) diverged and where it did not diverge.

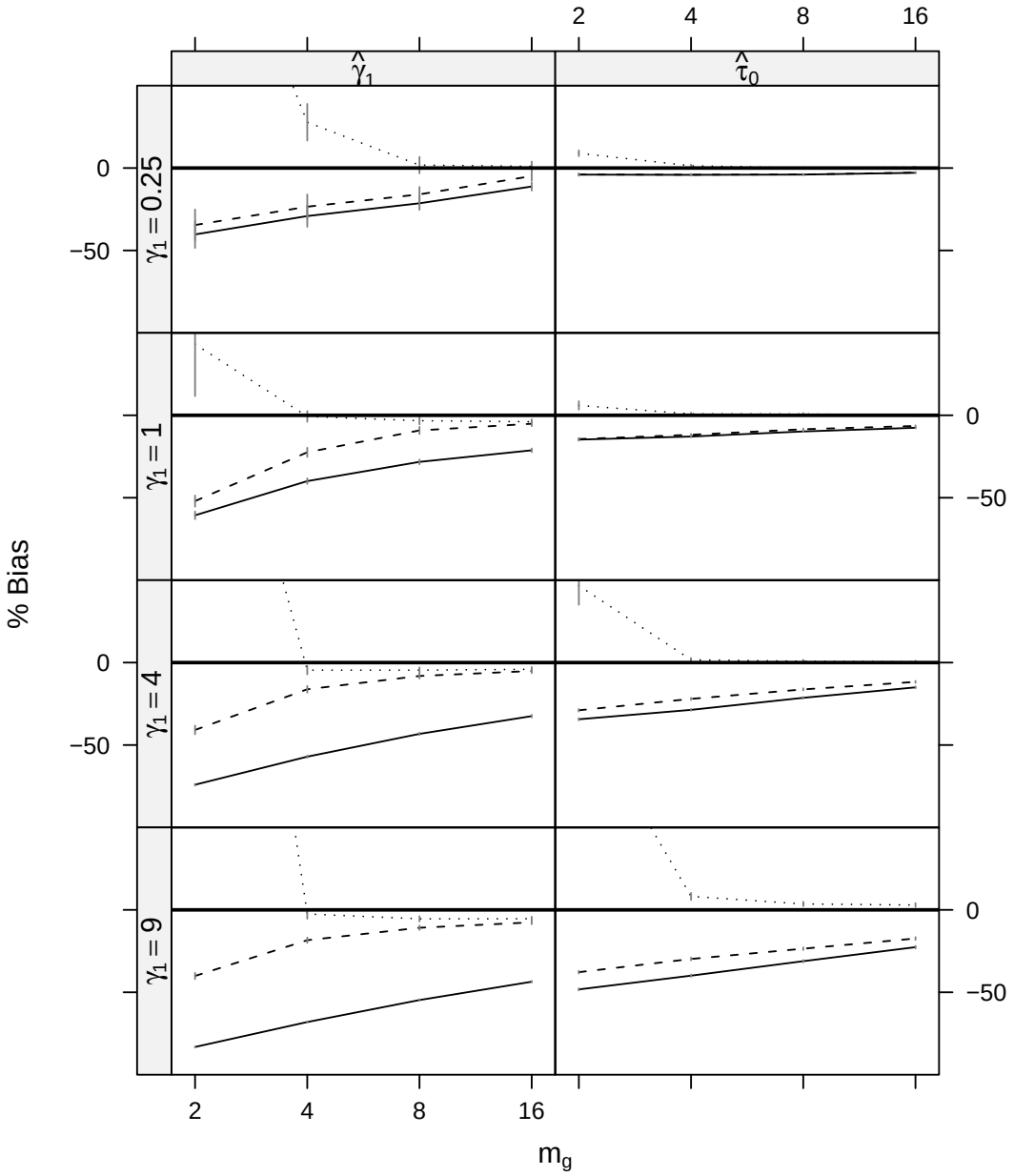


Figure 4.5: Interactions of the effects of m_g and γ_1 on average biases for $\hat{\gamma}_1$ and $\hat{\tau}_0$ for the binary one-way classification model (4.5) where $\tau_0 = 2$ and $b_g = 500$. (PQL: solid; HG(0,1): dashed; HG(1,1): dotted) (Error bars are $\pm$2SE.)

The divergence of the HG(1,1) estimates observed here contradict some extraordinarily good simulation results reported in Noh *et al.* (2005) for a similar binary logit one-way classification model. Their "full sample" model (ignoring the ascertainment issue) for binary data y_{ij} corresponds to equation (4.5) above, where $\tau_0 = -5$, $\gamma_1 = 4.5$, $b_g = 100,000$ and $m_g = 5$. It appears that a first order HG(1,1) approxima-

tion was used in this paper. We replicated their simulation study with $b_g = 10,000$ (to reduce computational time), and similar divergence problems were found, as for $\tau_0 = 2$ simulations above.

4.2.2.2 Nested two-way classification model (binary data only)

The nested two-way classification design from the previous chapter (section 3.1.4.2) is examined for binary data only, since the magnitude of the biases for Poisson data were much smaller.

Binary data y_{ijk}, $i = 1 \ldots b_g$, $j = 1 \ldots m_g$, $k = 1 \ldots m_s$ were generated, and analysed, according to the model for the conditional mean $\mu^u_{ijk} = E(y_{ijk}|u_{1i}, u_{2ij})$:

$$\text{logit}(\mu^u_{ijk}) \quad = \quad \tau_0 + u_{1i} + u_{2ij}, \qquad u_{1i} \sim N(0, \gamma_1), \qquad u_{2ij} \sim N(0, \gamma_2). \quad (4.6)$$

The values of the simulation parameters are given in Table 4.3.

Parameter	*Binary model*	*Poisson model*
b_g	50, 100, 200	...
m_g, m_s	2, 4, 8	...
γ_1, γ_2	1, 4	...
τ_0	0, 1	...

Table 4.3: Values of the simulation parameters used for the nested two-way classification study (4.6) to compare first order HGLM approaches against PQL. The use of ... in the 3rd (Poisson model) column indicates the same values were used as for the binary model.

Figure 4.6 shows that the negative bias for $\hat{\gamma}_1$ was substantially lower in magnitude for both HGLM approximations compared with that for PQL, with negligible difference between HG(0,1) and HG(1,1) approximations . It is also noticeable that the negative bias for $\hat{\gamma}_1$ is larger when $(\gamma_1, \gamma_2) = (1, 4)$ than when $(\gamma_1, \gamma_2) = (4, 1)$ for both PQL and the HGLM approximations. However, for $\hat{\gamma}_2$, the bias was still substantial when $m_s = 2$ for both HGLM approximations. Surprisingly, the negative bias of $\hat{\gamma}_2$ for the HGLM approximations was larger in magnitude when $\gamma_2 = 1$, where the bias was over 60% when $m_s = 2$, than when $\gamma_2 = 4$. The negative bias of $\hat{\tau}_0$ for HG(0,1) was similar to that of PQL, but the bias of $\hat{\tau}_0$ for HG(1,1) appeared to be negligible.

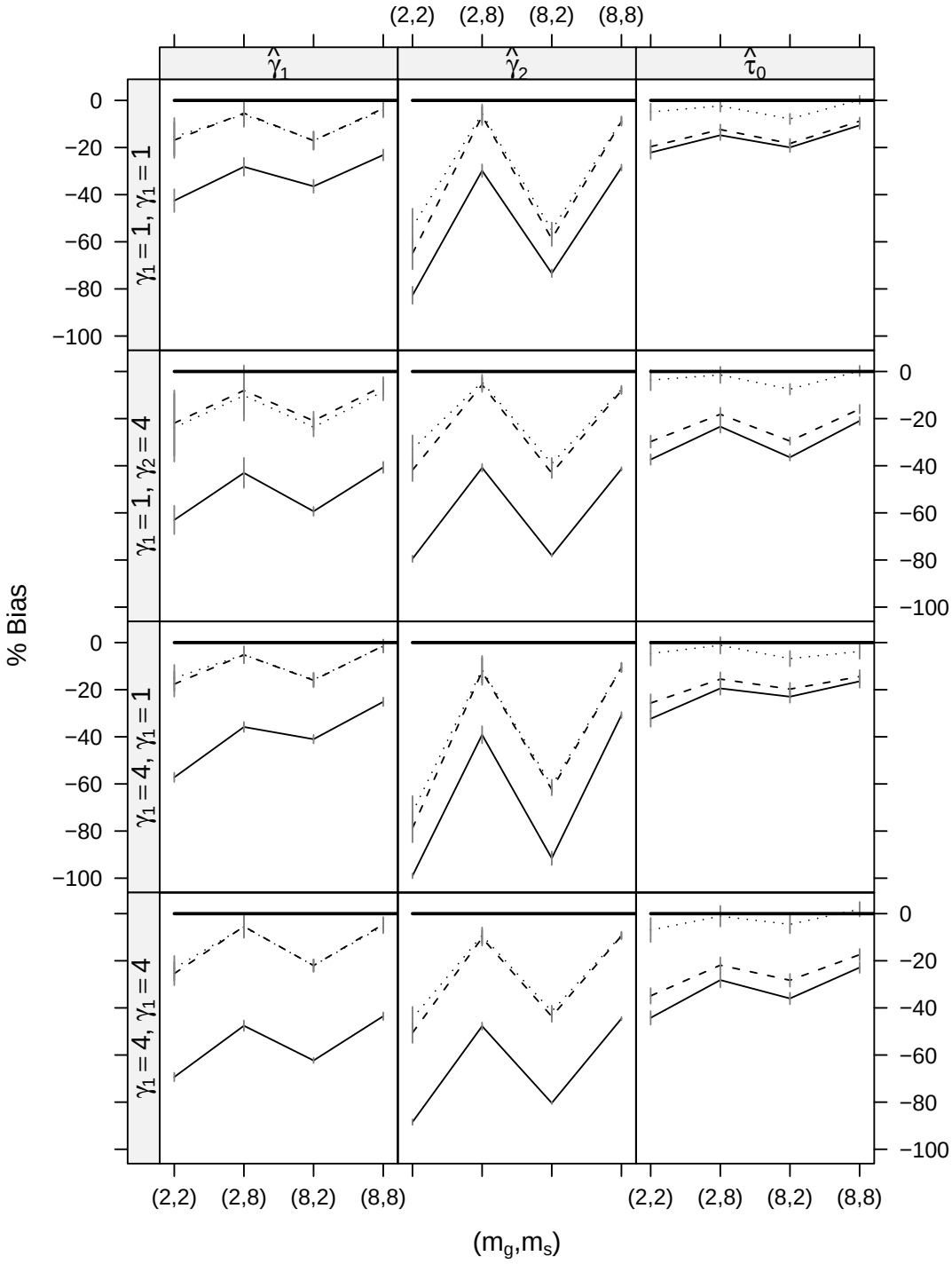


Figure 4.6: Interactions of the effects of m_g, m_s, γ_1 and γ_2 on average biases for $\hat{\gamma}_1$, $\hat{\gamma}_2$ and $\hat{\tau}_0$ for the binary nested two-way classification model (4.6) where $b_g = 200$ and $\tau_0 = 1$ for PQL and first order HGLM approximations. (PQL: solid; HG(0,1): dashed; HG(1,1): dotted) Error bars are ±2SE.

This simulation study, like the previous one-way classification study, demonstrated that the estimation biases of the fixed parameter estimator $\hat{\tau}_0$ using the HG(1,1) approximation were small and, for practical purposes, virtually ignorable. However, there are still significant biases for the variance parameters $\hat{\gamma}_1$ and $\hat{\gamma}_2$ using HG(1,1). It will be of interest to see whether the use of the second order HGLM approximations remove this bias, which will be explored later in the chapter.

4.2.3 Analytical expressions for the score equations

In this section the score equations for first order HG(i,1) approaches, $i = 0, 1$, are derived. As noted previously, these score equations were not used in the Fortran 90 implementation described in section 4.2.1, which used numerical derivatives based on finite differencing. It is assumed here that a canonical link is being used, so that

$$-\frac{\partial^2 h}{\partial \boldsymbol{u} \partial \boldsymbol{u}^T} = \left| \boldsymbol{Z}^T \boldsymbol{W} \boldsymbol{Z} + \boldsymbol{G}^{-1} \right|.$$

For HG(1,1), derivatives of

$$\ell \approx p_u(h) = \left(h - \frac{1}{2} \log \left| \boldsymbol{Z}^T \boldsymbol{W} \boldsymbol{Z} + \boldsymbol{G}^{-1} \right| + \frac{b}{2} \log 2\pi \right)_{\tilde{\boldsymbol{u}}_{\tau,\gamma}}$$

with respect to $\boldsymbol{\tau}$ are presented, as well as derivatives of

$$p_\beta(h) \approx \left(h - \frac{1}{2} \log \left| \boldsymbol{Z}^T \boldsymbol{W} \boldsymbol{Z} + \boldsymbol{G}^{-1} \right| - \frac{1}{2} \log \left| \boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X} \right| + \frac{p+b}{2} \log 2\pi \right)_{\hat{\boldsymbol{\beta}}_\gamma},$$

with respect to $\boldsymbol{\gamma}$, where $\boldsymbol{\beta} = \left(\boldsymbol{\tau}^T, \boldsymbol{u}^T \right)^T$ as before. Let the h-likelihood be written as

$$h \propto -\frac{1}{2} \sum_{i=1}^{n} d_i(y_i; \mu_i^u) - \frac{1}{2} \boldsymbol{u}^T \boldsymbol{G}^{-1} \boldsymbol{u} - \frac{1}{2} \log |\boldsymbol{G}|,$$

where

$$d_i(y, \mu) = -2 \int_y^\mu \frac{y - \nu}{a_i v(\nu)} d\nu,$$

and $v(\cdot)$ is the variance function corresponding to the GLM or quasi-likelihood dis-

tribution of $f_{Y|U}$, with the a_i being known constants.

4.2.3.1 Score equations for the fixed effects using HG(1,1)

We require $\partial p_u(h)/\partial \boldsymbol{\tau}$. Let $\tilde{h} = h|_{\boldsymbol{u}=\tilde{\boldsymbol{u}}_{\boldsymbol{\tau},\boldsymbol{\gamma}}}$. Firstly, note that

$$\frac{\partial \tilde{h}}{\partial \boldsymbol{\tau}} = \frac{\partial h}{\partial \boldsymbol{\tau}}\bigg|_{\tilde{\boldsymbol{u}}_{\tau,\gamma}} + \frac{\partial h}{\partial \boldsymbol{u}}\bigg|_{\tilde{\boldsymbol{u}}_{\tau,\gamma}} \frac{\partial \tilde{\boldsymbol{u}}_{\tau,\gamma}}{\partial \boldsymbol{\tau}}\bigg|_{\tilde{\boldsymbol{u}}_{\tau,\gamma}} = \frac{\partial h}{\partial \boldsymbol{\tau}}\bigg|_{\tilde{\boldsymbol{u}}_{\tau,\gamma}},$$

since $\partial h/\partial \boldsymbol{u}|_{\tilde{\boldsymbol{u}}_{\tau,\gamma}} = \mathbf{0}$, by the definition of $\tilde{\boldsymbol{u}}_{\tau,\gamma}$. So

$$\frac{\partial p_u(h)}{\partial \boldsymbol{\tau}} = \frac{\partial h}{\partial \boldsymbol{\tau}}\bigg|_{\tilde{\boldsymbol{u}}_{\tau,\gamma}} - \frac{1}{2}\frac{\partial}{\partial \boldsymbol{\tau}}\left(\log\left|\boldsymbol{Z}^T\tilde{\boldsymbol{W}}\boldsymbol{Z} + \boldsymbol{G}^{-1}\right|\right),$$

where $\tilde{\boldsymbol{W}} = \boldsymbol{W}|_{\boldsymbol{u}=\tilde{\boldsymbol{u}}_{\tau.\gamma}}$ and

$$\frac{\partial h}{\partial \boldsymbol{\tau}} = \sum \frac{(y_i - \mu_i)}{a_i v(\mu_i)} \frac{\boldsymbol{x}_i}{g'(\mu_i)}.$$

Now, for $j \in \{1, \ldots, p\}$,

$$\frac{\partial}{\partial \tau_j}\left(\log\left|\boldsymbol{Z}^T\tilde{\boldsymbol{W}}\boldsymbol{Z} + \boldsymbol{G}^{-1}\right|\right) = \text{trace}\left\{\left(\boldsymbol{Z}^T\tilde{\boldsymbol{W}}\boldsymbol{Z} + \boldsymbol{G}^{-1}\right)^{-1}\boldsymbol{Z}^T\frac{\partial \tilde{\boldsymbol{W}}}{\partial \tau_j}\boldsymbol{Z}\right\}.$$

Let $\tilde{w}_i$ be the ith diagonal of $\tilde{\boldsymbol{W}}$. Then

$$\frac{\partial \tilde{w}_i}{\partial \tau_j} = \frac{\partial w_i}{\partial \tau_j}\bigg|_{\tilde{\boldsymbol{u}}_{\tau,\gamma}} + \frac{\partial w_i}{\partial \boldsymbol{u}}\bigg|_{\tilde{\boldsymbol{u}}_{\tau,\gamma}} \frac{\partial \tilde{\boldsymbol{u}}_{\tau,\gamma}}{\partial \tau_j}.$$

Using implicit differentiation, with the knowledge that $\partial h/\partial \boldsymbol{u}|_{\tilde{\boldsymbol{u}}_{\tau,\gamma}} = \mathbf{0}$, it can be shown (Appendix A.1) that

$$\frac{\partial \tilde{\boldsymbol{u}}_{\tau,\gamma}}{\partial \tau_j} = -\left(\boldsymbol{Z}^T\tilde{\boldsymbol{W}}\boldsymbol{Z} + \boldsymbol{G}^{-1}\right)^{-1}\boldsymbol{Z}^T\tilde{\boldsymbol{W}}\boldsymbol{x}_{.,j},$$

where $\boldsymbol{x}_{.,j}$ is the jth column of $\boldsymbol{X}$. Therefore

$$\begin{aligned}\frac{\partial \tilde{w}_i}{\partial \tau_j} &= \frac{\partial w_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \tau_j}\bigg|_{\tilde{\boldsymbol{u}}_{\tau,\gamma}} + \frac{\partial w_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \boldsymbol{u}}\bigg|_{\tilde{\boldsymbol{u}}_{\tau,\gamma}} \frac{\partial \tilde{\boldsymbol{u}}_{\tau,\gamma}}{\partial \boldsymbol{\tau}_j} \\ &= \tilde{w}_i' \left(x_{ij} - \boldsymbol{z}_i^T \left(\boldsymbol{Z}^T \tilde{\boldsymbol{W}} \boldsymbol{Z} + \boldsymbol{G}^{-1} \right)^{-1} \boldsymbol{Z}^T \tilde{\boldsymbol{W}} \boldsymbol{x}_{.,j} \right),\end{aligned}$$

where $\tilde{w}_i' = \partial w_i / \partial \eta_i |_{\tilde{\boldsymbol{u}}_{\tau,\gamma}}$ and $\boldsymbol{z}_i^T$ is the ith row of $\boldsymbol{Z}$. Therefore

$$\frac{\partial \tilde{\boldsymbol{W}}}{\partial \tau_j} = \tilde{\boldsymbol{W}}' \left(\boldsymbol{X}_j - \boldsymbol{Z} \left(\boldsymbol{Z}^T \tilde{\boldsymbol{W}} \boldsymbol{Z} + \boldsymbol{G}^{-1} \right)^{-1} \boldsymbol{Z}^T \tilde{\boldsymbol{W}} \boldsymbol{X}_j \right),$$

where $\tilde{\boldsymbol{W}}' = \text{diag}\{\tilde{w}_i'\}$ and $\boldsymbol{X}_j = \text{diag}\{x_{ij}\}$. So

$$\begin{aligned}\frac{\partial}{\partial \tau_j}\left(\log \left| \boldsymbol{Z}^T \tilde{\boldsymbol{W}} \boldsymbol{Z} + \boldsymbol{G}^{-1} \right| \right) &= \text{trace} \left\{ \boldsymbol{Z} \left(\boldsymbol{Z}^T \tilde{\boldsymbol{W}} \boldsymbol{Z} + \boldsymbol{G}^{-1} \right)^{-1} \boldsymbol{Z}^T \frac{\partial \tilde{\boldsymbol{W}}}{\partial \tau_j} \right\} \\ &= \sum_{i,k,l} z_{ik} \tilde{d}_{kl} z_{il} \tilde{w}_i' \left(x_{ij} - \sum_{m,n} z_{im} \tilde{d}_{mn} z_{in} \tilde{w}_i x_{ij} \right) \\ &= \sum_i \tilde{w}_i' x_{ij} \tilde{\xi}_i \left(1 - \tilde{w}_i \tilde{\xi}_i \right),\end{aligned}$$

where $\tilde{\xi}_i = \sum_{k,l} z_{ik} \tilde{d}_{kl} z_{il}$ and $\tilde{d}_{kl}$ is the (k,l)th element of the matrix $\left(\boldsymbol{Z}^T \tilde{\boldsymbol{W}} \boldsymbol{Z} + \boldsymbol{G}^{-1} \right)^{-1}$.

Correction to the mixed model equations for HG(1,1) As noted in section 4.2.1, the joint solutions $\boldsymbol{\beta}_\gamma = \left(\hat{\boldsymbol{\tau}}_\gamma^T, \tilde{\boldsymbol{u}}_{\tau,\gamma}^T \right)^T$ of $\partial h / \partial \boldsymbol{\beta} = 0$, for fixed $\boldsymbol{\gamma}$, can be found by repeatedly solving the linear set of equations

$$\begin{bmatrix} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} & \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{Z} \\ \boldsymbol{Z}^T \boldsymbol{W} \boldsymbol{X} & \boldsymbol{Z}^T \boldsymbol{W} \boldsymbol{Z} + \boldsymbol{G}^{-1} \end{bmatrix} \begin{pmatrix} \hat{\boldsymbol{\tau}} \\ \tilde{\boldsymbol{u}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{\psi} \\ \boldsymbol{Z}^T \boldsymbol{W} \boldsymbol{\psi} \end{pmatrix} \tag{4.7}$$

for $\hat{\boldsymbol{\tau}}$ and $\tilde{\boldsymbol{u}}$ until convergence, and $\hat{\boldsymbol{\tau}}_\gamma = \hat{\boldsymbol{\tau}}$ and $\tilde{\boldsymbol{u}}_{\tau,\gamma} = \tilde{\boldsymbol{u}}$ at convergence. The vector $\boldsymbol{\psi} = (\psi_1 \dots \psi_n)^T$ has elements $\psi_i = \eta_i + g'(\mu_i)(y_i - \mu_i)$, $\eta_i = g^{-1}(\mu_i)$ and μ_i , $i = 1 \dots n$, and $\boldsymbol{W}$ are evaluated at the current estimates $\hat{\boldsymbol{\tau}}$ and $\tilde{\boldsymbol{u}}$, as before.

To solve the score equations $\partial p_u(h) / \partial \boldsymbol{\tau}$ for $\boldsymbol{\tau}$, required for the HG(1,1) approximation,

the RHS of (4.7) can be replaced by

$$\begin{pmatrix} \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{\psi} - \boldsymbol{\zeta} \\ \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{\psi} \end{pmatrix},$$

where $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_p)^T$ and

$$\zeta_j = \frac{\partial}{\partial \tau_j}\left(\log\left|\boldsymbol{Z}^T\tilde{\boldsymbol{W}}\boldsymbol{Z} + \boldsymbol{G}^{-1}\right|\right) = \sum_i w_i' x_{ij}\xi_i\left(1 - w_i\xi_i\right). \tag{4.8}$$

Here, $w_i' = \partial w_i/\partial\eta_i$ and $\xi_i = \sum_{k,l} z_{ik}d_{kl}z_{il}$, where d_{kl} is the (k,l)th element of $\boldsymbol{D} = \left(\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z} + \boldsymbol{G}^{-1}\right)^{-1}$. (Note that, for simplicity of notation, the tilde $\tilde{}$ notation used previously has been dropped – all quantities are calculated at the current estimates of $\boldsymbol{\tau}$ and $\boldsymbol{u}$.)

Note that Noh & Lee (2007) assume that no change to the LHS of (4.7) is required to implement HG(1,1). Therefore the updating equations use the same information matrix for Fisher scoring as for HG(0,1), that is,

$$-\mathrm{E}\left\{\frac{\partial^2 h}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right\} = \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z} + \boldsymbol{G}^{-1} \end{bmatrix},$$

where $\boldsymbol{\beta} = (\boldsymbol{\tau}^T, \boldsymbol{u}^T)^T$. This ignores the complication arising from the fact that there are two separate likelihood criteria for maximising $\boldsymbol{\tau}$ and $\boldsymbol{u}$, $p_u(h)$ and h respectively.

Fortran-style pseudo-code to compute $\boldsymbol{\zeta}$, is shown in Table 4.4. This pseudocode assumes that $\boldsymbol{Z}$ is stored sparsely with non-zero elements $\boldsymbol{z} = (z_1 \ldots z_{nsp})^T$ and corresponding row and column positions $\boldsymbol{r} = (r_1 \ldots r_{nsp})^T$ and $\boldsymbol{c} = (c_1 \ldots c_{nsp})^T$. It also assumes that $\boldsymbol{z}$, $\boldsymbol{r}$ and $\boldsymbol{c}$ are stored in row-order, that is, the non-zero elements of the first row from left to right, then the non-zero elements of the second row etc. Note that the variable `xi` represents ξ_i, $i = 1, \ldots, n$, and the variable `zt(j)` represent ζ_j, $j = 1, \ldots, p$.

```
xi=0
starte=1
i=r(1)
xi = z(1)**2*D(c(1),c(1))
do j=2,nsp
  newrow=r(j)
  ! accumulate contributions from row j of Z to zeta_i's
  if(newrow/=i.or.j==nsp) then
     temp= e3(i)*xi*(1-wt(i)*xi)
     do l=1,p
        zt(l) = zt(l) + temp*X(i,l)
     enddo
     starte=j  ! only search from elements of z(starte:...on)....
     i=newrow
     xi = z(j)**2*D(c(j),c(j))
  endif
  if(j>starte) then
    zj=z(j)  ! store zj, cj in cache
    cj=c(j)
    do k=starte,j-1
      xi = xi + 2*zj*z(k)*D(cj,c(k))
    enddo
    xi = xi + zj**2*D(cj,cj)
  endif
enddo
```

Table 4.4: Fortran-style pseudo-code required to compute the adjustment $\boldsymbol{\zeta}$ (4.8) to the mixed model equations (4.7) for using the HG(1,j) approaches ($j \geq 1$).

4.2.3.2 Score equations for the variance components

To derive score equations for $\boldsymbol{\gamma}$, let $\boldsymbol{\beta} = (\boldsymbol{\tau}^T, \boldsymbol{u}^T)^T$ as before and

$$\begin{aligned} p_\beta(h) &= \left(h - \frac{1}{2} \log \left| -\frac{1}{2\pi} \frac{\partial^2 h}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right| \right)_{\tilde{\boldsymbol{\beta}}_\gamma} \\ &= \left(h - \frac{1}{2} \log |\boldsymbol{C}| + \frac{p+b}{2} \log 2\pi \right)_{\tilde{\boldsymbol{\beta}}_\gamma}, \end{aligned} \tag{4.9}$$

where

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} & \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{Z} \\ \boldsymbol{Z}^T \boldsymbol{W} \boldsymbol{X} & \boldsymbol{Z}^T \boldsymbol{W} \boldsymbol{Z} + \boldsymbol{G}^{-1} \end{bmatrix}. \tag{4.10}$$

Let

$$\begin{aligned} p_\beta &= h - \frac{1}{2}\log|\boldsymbol{C}| + \frac{p+b}{2}\log 2\pi, \\ p_u &= h - \frac{1}{2}\log\left|\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z} + \boldsymbol{G}^{-1}\right| + \frac{b}{2}\log 2\pi. \end{aligned}$$

Let the vector $\tilde{\boldsymbol{\beta}}_\gamma$ equal $(\hat{\boldsymbol{\tau}}_\gamma^T, \tilde{\boldsymbol{u}}_{\tau,\gamma}^T)^T$, where $\hat{\boldsymbol{\tau}}_\gamma^T$ either satisfies $\partial h/\partial\boldsymbol{\tau} = 0$ for HG(0,1) or $\partial p_u(h)/\partial\boldsymbol{\tau} = 0$ for HG(1,1). Then, for HG(1,1),

$$\begin{aligned} \frac{\partial p_\beta(h)}{\partial\boldsymbol{\gamma}} &= \left.\frac{\partial p_\beta}{\partial\boldsymbol{\gamma}}\right|_{\hat{\boldsymbol{\beta}}_\gamma} + \left.\frac{\partial p_\beta}{\partial\boldsymbol{\tau}}\right|_{\hat{\boldsymbol{\beta}}_\gamma}\frac{\partial\hat{\boldsymbol{\tau}}_\gamma}{\partial\boldsymbol{\gamma}} + \left.\frac{\partial p_\beta}{\partial\boldsymbol{u}}\right|_{\hat{\boldsymbol{\beta}}_\gamma}\frac{\partial\hat{\boldsymbol{u}}_{\tau,\gamma}}{\partial\boldsymbol{\gamma}} \\ &= \left.\frac{\partial p_\beta}{\partial\boldsymbol{\gamma}}\right|_{\hat{\boldsymbol{\beta}}_\gamma} + \left(\left.\frac{\partial p_u}{\partial\boldsymbol{\tau}}\right|_{\hat{\boldsymbol{\beta}}_\gamma} + \left.\frac{\partial(p_\beta - p_u)}{\partial\boldsymbol{\tau}}\right|_{\hat{\boldsymbol{\beta}}_\gamma}\right)\frac{\partial\hat{\boldsymbol{\tau}}_\gamma}{\partial\boldsymbol{\gamma}} \\ &\quad + \left(\left.\frac{\partial h}{\partial\boldsymbol{u}}\right|_{\hat{\boldsymbol{\beta}}_\gamma} + \left.\frac{\partial(p_\beta - h)}{\partial\boldsymbol{u}}\right|_{\hat{\boldsymbol{\beta}}_\gamma}\right)\frac{\partial\hat{\boldsymbol{u}}_{\tau,\gamma}}{\partial\boldsymbol{\gamma}} \\ &= \left.\frac{\partial p_\beta}{\partial\boldsymbol{\gamma}}\right|_{\hat{\boldsymbol{\beta}}_\gamma} + \left.\frac{\partial(p_\beta - p_u)}{\partial\boldsymbol{\tau}}\right|_{\hat{\boldsymbol{\beta}}_\gamma}\frac{\partial\hat{\boldsymbol{\tau}}_\gamma}{\partial\boldsymbol{\gamma}} + \left.\frac{\partial(p_\beta - h)}{\partial\boldsymbol{u}}\right|_{\hat{\boldsymbol{\beta}}_\gamma}\frac{\partial\hat{\boldsymbol{u}}_{\tau,\gamma}}{\partial\boldsymbol{\gamma}}, \end{aligned}$$

since $\partial p_u/\partial\boldsymbol{\tau}|_{\hat{\boldsymbol{\beta}}_\gamma} = \mathbf{0}$ and $\partial h/\partial\boldsymbol{u}|_{\hat{\boldsymbol{\beta}}_\gamma} = \mathbf{0}$ by definition. Similarly, for HG(0,1),

$$\frac{\partial p_\beta(h)}{\partial\boldsymbol{\gamma}} = \left.\frac{\partial p_\beta}{\partial\boldsymbol{\gamma}}\right|_{\hat{\boldsymbol{\beta}}_\gamma} + \left.\frac{\partial(p_\beta - h)}{\partial\boldsymbol{\tau}}\right|_{\hat{\boldsymbol{\beta}}_\gamma}\frac{\partial\hat{\boldsymbol{\tau}}_\gamma}{\partial\boldsymbol{\gamma}} + \left.\frac{\partial(p_\beta - h)}{\partial\boldsymbol{u}}\right|_{\hat{\boldsymbol{\beta}}_\gamma}\frac{\partial\hat{\boldsymbol{u}}_{\tau,\gamma}}{\partial\boldsymbol{\gamma}},$$

since $\partial h/\partial\boldsymbol{\tau}|_{\hat{\boldsymbol{\beta}}_\gamma} = \mathbf{0}$ by definition. Note that $p_\beta - h \propto -1/2\log|\boldsymbol{C}|$ and $p_\beta - p_u \propto -1/2\log\left|\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right|$, where $\boldsymbol{V} = \boldsymbol{W}^{-1} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T$. As in Noh & Lee (2007) and Lee & Nelder (2001), the dependence of $\hat{\boldsymbol{\tau}}_\gamma$ on $\boldsymbol{\gamma}$ is ignored, so that

$$\begin{aligned} \frac{\partial p_\beta(h)}{\partial\boldsymbol{\gamma}} &\simeq \left.\frac{\partial p_\beta}{\partial\boldsymbol{\gamma}}\right|_{\hat{\boldsymbol{\beta}}_\gamma} + \left.\frac{\partial(p_\beta - h)}{\partial\boldsymbol{u}}\right|_{\hat{\boldsymbol{\beta}}_\gamma}\frac{\partial\hat{\boldsymbol{u}}_{\tau,\gamma}}{\partial\boldsymbol{\gamma}} \\ &= \frac{\partial h}{\partial\boldsymbol{\gamma}} - \frac{1}{2}\left.\frac{\partial\log|\boldsymbol{C}|}{\partial\boldsymbol{\gamma}}\right|_{\hat{\boldsymbol{\beta}}_\gamma} - \frac{1}{2}\left.\frac{\partial\log|\boldsymbol{C}|}{\partial\boldsymbol{u}}\right|_{\hat{\boldsymbol{\beta}}_\gamma}\frac{\partial\hat{\boldsymbol{u}}_{\tau,\gamma}}{\partial\boldsymbol{\gamma}} \end{aligned}$$

for both HG(0,1) and HG(1,1). Ignoring the dependence of $\hat{\boldsymbol{\tau}}_\gamma$ on $\boldsymbol{\gamma}$ tacitly assumes that $\boldsymbol{\tau}$ and $\boldsymbol{\gamma}$ are independent, which, Lee & Nelder (2001) argued, is a reasonable assumption. It may be useful to explore whether this assumption is valid, but this has not been done here.

We will now consider the score equation for single variance component, $\partial p_\beta(h)/\partial\gamma_i$,

$$\frac{\partial p_\beta(h)}{\partial\gamma_j} \simeq \frac{\partial h}{\partial\gamma_j} - \frac{1}{2}\frac{\partial \log|\boldsymbol{C}|}{\partial\gamma_j}\bigg|_{\hat{\boldsymbol{\beta}}_\gamma} - \frac{1}{2}\frac{\partial \log|\boldsymbol{C}|}{\partial \boldsymbol{u}}\bigg|_{\hat{\boldsymbol{\beta}}_\gamma}\frac{\partial \hat{\boldsymbol{u}}_{\tau,\gamma}}{\partial\gamma_j}. \tag{4.11}$$

Note that

$$\frac{\partial h}{\partial\gamma_j} = -\frac{1}{2}\left\{\boldsymbol{u}^T\boldsymbol{G}^{-1}\frac{\partial \boldsymbol{G}}{\partial\gamma_j}\boldsymbol{G}^{-1}\boldsymbol{u} + \text{trace}\left(\boldsymbol{G}^{-1}\frac{\partial \boldsymbol{G}}{\partial\gamma_j}\right)\right\}. \tag{4.12}$$

Let

$$\boldsymbol{T} = \begin{pmatrix} \boldsymbol{X} & \boldsymbol{Z} \\ \boldsymbol{0} & \boldsymbol{I} \end{pmatrix}, \quad \boldsymbol{W}_a = \begin{pmatrix} \boldsymbol{W} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{G}^{-1} \end{pmatrix},$$

as in Lee & Nelder (2001) and Noh & Lee (2007), and so $\boldsymbol{C} = \boldsymbol{T}^T\boldsymbol{W}_a\boldsymbol{T}$. Let

$$\boldsymbol{C}^{-1} = \begin{pmatrix} \boldsymbol{C}^{11} & \boldsymbol{C}^{12} \\ \boldsymbol{C}^{21} & \boldsymbol{C}^{22} \end{pmatrix}.$$

Then

$$\begin{aligned} \frac{\partial}{\partial\gamma_j}\left(\log|\boldsymbol{C}|\right) &= \text{trace}\left\{\boldsymbol{C}^{-1}\left(\boldsymbol{T}^T\frac{\partial \boldsymbol{W}_a}{\partial\gamma_j}\boldsymbol{T}\right)\right\} \\ &= -\text{trace}\left\{\boldsymbol{C}^{22}\left(\boldsymbol{G}^{-1}\frac{\partial \boldsymbol{G}}{\partial\gamma_j}\boldsymbol{G}^{-1}\right)\right\}. \end{aligned} \tag{4.13}$$

Now consider the kth element of $\partial\left(\log|\boldsymbol{C}|\right)/\partial\boldsymbol{u}$,

$$\begin{aligned} \frac{\partial \log|\boldsymbol{C}|}{\partial u_k} &= \text{trace}\left\{\boldsymbol{C}^{-1}\left(\boldsymbol{T}^T\frac{\partial \boldsymbol{W}_a}{\partial u_k}\boldsymbol{T}\right)\right\} \\ &= \text{trace}\left\{\boldsymbol{C}^{-1}\left(\begin{bmatrix} \boldsymbol{X}^T \\ \boldsymbol{Z}^T \end{bmatrix}\frac{\partial \boldsymbol{W}}{\partial u_k}\begin{bmatrix} \boldsymbol{X} & \boldsymbol{Z} \end{bmatrix}\right)\right\}, \end{aligned} \tag{4.14}$$

where the ith diagonal element of $\boldsymbol{W}_k' = \partial\boldsymbol{W}/\partial u_k$ is

$$\frac{\partial w_i}{\partial u_k} = \frac{\partial w_i}{\partial\eta_i}\frac{\partial\eta_i}{\partial u_k} = \frac{\partial w_i}{\partial\eta_i}z_{ij}.$$

Using implicit differentiation, it can be shown that (Appendix A.1)

$$\frac{\partial \tilde{\boldsymbol{u}}_{\tau,\gamma}}{\partial \gamma_j} = \left(\boldsymbol{Z}^T \tilde{\boldsymbol{W}} \boldsymbol{Z} + \boldsymbol{G}^{-1}\right)^{-1} \boldsymbol{G}^{-1} \frac{\partial \boldsymbol{G}}{\partial \gamma_j} \boldsymbol{G}^{-1} \tilde{\boldsymbol{u}}_{\tau,\gamma}. \tag{4.15}$$

So, combining (4.11) with its components (4.12), (4.13), (4.14) and (4.15), we obtain

$$\begin{aligned}
\frac{\partial p_\beta(h)}{\partial \gamma_j} &\simeq \frac{\partial h}{\partial \gamma_j} - \frac{1}{2} \frac{\partial \log |\boldsymbol{C}|}{\partial \gamma_j}\bigg|_{\hat{\boldsymbol{\beta}}_\gamma} - \frac{1}{2} \frac{\partial \log |\boldsymbol{C}|}{\partial \boldsymbol{u}}\bigg|_{\hat{\boldsymbol{\beta}}_\gamma} \frac{\partial \hat{\boldsymbol{u}}_{\tau,\gamma}}{\partial \gamma_j} \\
&= -\frac{1}{2} \left\{ \boldsymbol{u}^T \boldsymbol{G}^{-1} \frac{\partial \boldsymbol{G}}{\partial \gamma_j} \boldsymbol{G}^{-1} \boldsymbol{u} + \operatorname{trace}\left(\boldsymbol{G}^{-1} \frac{\partial \boldsymbol{G}}{\partial \gamma_j} \right) \right\}\bigg|_{\hat{\boldsymbol{\beta}}_\gamma} \\
&\quad + \frac{1}{2} \operatorname{trace}\left\{ \boldsymbol{C}^{22} \left(\boldsymbol{G}^{-1} \frac{\partial \boldsymbol{G}}{\partial \gamma_j} \boldsymbol{G}^{-1} \right) \right\}\bigg|_{\hat{\boldsymbol{\beta}}_\gamma} \\
&\quad - \frac{1}{2} \sum_k \operatorname{trace}\left\{ \boldsymbol{C}^{-1} \left(\begin{bmatrix} \boldsymbol{X}^T \\ \boldsymbol{Z}^T \end{bmatrix} \frac{\partial \boldsymbol{W}}{\partial u_k} \begin{bmatrix} \boldsymbol{X} & \boldsymbol{Z} \end{bmatrix} \right) \right\}\bigg|_{\hat{\boldsymbol{\beta}}_\gamma} \\
&\quad \left\{ \left(\boldsymbol{Z}^T \tilde{\boldsymbol{W}} \boldsymbol{Z} + \boldsymbol{G}^{-1}\right)^{-1} \boldsymbol{G}^{-1} \frac{\partial \boldsymbol{G}}{\partial \gamma_j} \boldsymbol{G}^{-1} \tilde{\boldsymbol{u}}_{\tau,\gamma} \right\}.
\end{aligned}$$

4.2.3.3 Use of the score equations

Note that the analytic expressions for the derivatives derived above, $\partial p_u(h)/\partial \boldsymbol{\tau}$ and $\partial p_\beta(h)/\partial \boldsymbol{\gamma}$, could be used to replace the finite difference derivative calculations in the current Fortran 90 implementation (section 4.2.1). Alternatively, the derivatives $\partial p_u(h)/\partial \boldsymbol{\tau}$ could be incorporated into the mixed model equations, as outlined at the end of section 4.2.3.1. For computational speed and efficiency, the implementation of the analytic derivatives $\partial p_u(h)/\partial \boldsymbol{\tau}$ would take higher priority over the implementation of $\partial p_\beta(h)/\partial \boldsymbol{\gamma}$, since the updated estimates of $\boldsymbol{\tau}$, given $\boldsymbol{\gamma}$, are generated more frequently than the estimates of $\boldsymbol{\gamma}$ themselves.

Further analytical work could be undertaken to determine expressions for the expectations of the second derivatives, $-\mathrm{E}\left\{\partial^2 p_u(h)/\partial \boldsymbol{\tau} \partial \boldsymbol{\tau}^T\right\}$ and $-\mathrm{E}\left\{\partial^2 p_\beta(h)/\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T\right\}$. The derivation of the latter information matrix would enable the direct application of Fisher scoring for finding the maximum of $p_\beta(h)$ with respect to $\boldsymbol{\gamma}$, and would obviate the current need for using a quasi-Newton approach like L-BFGS-B. This option has

not been explored.

4.2.3.4 Differences between the HGLM and PQL score equations

As noted in section 4.2.2, an intuitive explanation for why the estimation biases for $\hat{\boldsymbol{\gamma}}$ increased with the magnitude of $\boldsymbol{\gamma}$ for PQL, but did not do so for either HGLM approximation, is developed in this sub-section.

The difference between HG(0,1) and HG(1,1), with respect to the magnitude of the biases for $\hat{\boldsymbol{\tau}}$, was already discussed in section 4.2.2.1. An intuitive explanation was provided for why the magnitude of the bias for HG(0,1) increased in magnitude with $\boldsymbol{\gamma}$, whereas the bias for HG(1,1) did not increase with $\boldsymbol{\gamma}$. A summary of this argument is as follows. The HG(0,1) approximation uses h for inference concerning $\boldsymbol{\tau}$, whereas HG(1,1) uses $p_u(h)$. The difference between the two criteria, $p_u(h) - h = -1/2 \log \left| \boldsymbol{Z}^T \boldsymbol{W} \boldsymbol{Z} + \boldsymbol{G}^{-1} \right|$, increases with $\boldsymbol{\gamma}$ since $\lim_{\gamma \to 0} \log \left| \boldsymbol{Z}^T \boldsymbol{W} \boldsymbol{Z} + \boldsymbol{G}^{-1} \right| = \log \left| \boldsymbol{G}^{-1} \right|$ and $\lim_{\gamma \to \infty} \log \left| \boldsymbol{Z}^T \boldsymbol{W} \boldsymbol{Z} + \boldsymbol{G}^{-1} \right| = \log \left| \boldsymbol{Z}^T \boldsymbol{W} \boldsymbol{Z} \right|$. Therefore, the change in $\log \left| \boldsymbol{Z}^T \boldsymbol{W} \boldsymbol{Z} + \boldsymbol{G}^{-1} \right|$ with $\boldsymbol{\tau}$ is ignorable when $\boldsymbol{\gamma}$ is small, but becomes more important as $\boldsymbol{\gamma}$ increases.

Now the difference between PQL and HG(0,1) will be explored. The difference between these two approaches is in the estimation of $\boldsymbol{\gamma}$, with HG(0,1) using $p_\beta(h)$ and PQL using an approximation to $p_\beta(h)$. The score equations for $\boldsymbol{\gamma}$ when HG(0,1) is used, as in section 4.2.3.2, are

$$\begin{aligned} \frac{\partial p_\beta(h)}{\partial \boldsymbol{\gamma}} &= \left. \frac{\partial p_\beta}{\partial \boldsymbol{\gamma}} \right|_{\hat{\boldsymbol{\beta}}_\gamma} + \left. \frac{\partial (p_\beta - h)}{\partial \boldsymbol{\beta}} \right|_{\hat{\boldsymbol{\beta}}_\gamma} \frac{\partial \hat{\boldsymbol{\beta}}_\gamma}{\partial \boldsymbol{\gamma}} \\ &= \frac{\partial h}{\partial \boldsymbol{\gamma}} + \left. \frac{\partial \left(p_\beta - h \right)}{\partial \boldsymbol{\gamma}} \right|_{\hat{\boldsymbol{\beta}}_\gamma} + \left. \frac{\partial (p_\beta - h)}{\partial \boldsymbol{\beta}} \right|_{\hat{\boldsymbol{\beta}}_\gamma} \frac{\partial \hat{\boldsymbol{\beta}}_\gamma}{\partial \boldsymbol{\gamma}}. \end{aligned} \tag{4.16}$$

As before, $\boldsymbol{\beta} = \left(\boldsymbol{\tau}^T, \boldsymbol{u}^T \right)^T$, $p_\beta \propto h - \log |\boldsymbol{C}|/2$, $\boldsymbol{C}$ is the matrix defined in (4.10) and h, the h-likelihood, can be written as

$$h \propto -\frac{1}{2} \sum_{i=1}^{n} d_i(y_i; \mu_i^u) - \frac{1}{2} \boldsymbol{u}^T \boldsymbol{G}^{-1} \boldsymbol{u} - \frac{1}{2} \log |\boldsymbol{G}| .$$

Firstly, note that (4.16) shows that the dependence of terms in h on $\boldsymbol{\gamma}$, through $\hat{\boldsymbol{\beta}}_\gamma$, can be ignored. We briefly review the approximations to $p_\beta(h)$ used for implementing PQL.

Firstly, twice the conditional likelihood, $2f_{Y|U} = -2\sum_{i=1}^{n} d_i(y_i;\mu_i^u)$, is replaced by the Pearson χ^2 statistic, $\sum (y_i-\mu_i)^2/V(\mu_i)$. Since the conditional likelihood $f_{Y|U}$ is only dependent of $\boldsymbol{\gamma}$ through $\hat{\boldsymbol{\beta}}_\gamma$, this change makes no difference to inference for $\boldsymbol{\gamma}$, as in (4.16). Secondly, an additional $-1/2\log|\boldsymbol{W}|$ is added. These two changes create a modified PQL "likelihood", which resembles a normal LMM likelihood for a "working" variate $\boldsymbol{\psi}$. However, in exploiting this resemblance, the corresponding PQL score equations ignore the dependence of $\boldsymbol{W}$ on $\boldsymbol{\gamma}$, and so ignore terms in the last component of (4.16). Specifically, it ignores the difference

$$\frac{\partial(p_\beta - h)}{\partial\boldsymbol{\beta}} = -\frac{1}{2}\left\{\frac{\partial\log\left|\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z}+\boldsymbol{G}^{-1}\right|}{\partial\boldsymbol{\beta}} + \frac{\partial\log\left|\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right|}{\partial\boldsymbol{\beta}}\right\},$$

where $\boldsymbol{V} = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T + \boldsymbol{W}^{-1}$. We assume now that $\log\left|\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right|$ is "small". (For most of the simulation studies in chapter 3, this is valid since there were few fixed effects.)

The reason for the increase in the magnitude of the bias of $\hat{\gamma}$ with $\boldsymbol{\gamma}$ when using PQL, but not HG(0,1), can now be seen to be due to the fact that $\lim_{\gamma\to\infty}\log\left|\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z}+\boldsymbol{G}^{-1}\right| = \log\left|\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z}\right|$. Therefore, the size of the component omitted by PQL, $\partial(p_\beta - h)/\partial\boldsymbol{\beta}$, increases with $\boldsymbol{\gamma}$.

For the binary logit model, increasing $\boldsymbol{\gamma}$ results in increases in the absolute values of both the fixed and random effect estimates, $|\hat{\tau}_i|$, $i = 1,\ldots,p$, and $|\hat{u}_j|$, $j = 1,\ldots,b$ respectively, and so a decrease in the weights $w_i = \mu_i(1-\mu_i)$, and, subsequently, a decrease in $\log\left|\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z}\right|$. So the omission of $\log\left|\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z}+\boldsymbol{G}^{-1}\right|$ in PQL would cause PQL to favour lower estimates of $\boldsymbol{\gamma}$ relative to the HG(0,1) approach. For Poisson log models, increasing $\boldsymbol{\gamma}$ results in increasing $|u_j|$ but decreasing τ_i, and the net result is, as for the binary model, also a decrease in $\log\left|\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z}\right|$. Therefore, the negative biases for $\hat{\gamma}$ will increase with increasing $\boldsymbol{\gamma}$ when using PQL for either binary and

Poisson models, but not using either of the HGLM approximations.

4.2.4 Adequacy of the (first order) Laplace approximation

The first order HGLM approaches are based on the (first order) Laplace approximation of the likelihood. Some intuition for the performance of the first order HGLM approaches in the simulation studies above may be gained by consideration of the one-way classification model (4.5).The contribution to the likelihood from the ith group is

$$L_i = \int \prod_{j=1}^{m_g} f_{y|u}(y_{ij}|u_i) f_u(u_i) du_i, \tag{4.17}$$

where $f_{y_||u}(y_{ij}|u_i)$ is the conditional PDF of the data y_{ij} given u_i and $f_u(u_i) = (2\pi\gamma_1)^{-1/2} \exp\left(-u_i^2/2\gamma_1\right)$ is the normal PDF for u_i. For binary data with a logit link,

$$\prod_{j=1}^{m_g} f_{y|u}(y_{ij}|u_i) = \prod_{j=1}^{m_g} \exp\left[\sum y_{ij}\left(\tau_0 + u_i\right)\right] \left[1 + \exp(\tau_0 + u_i)\right]^{-m_g},$$

and for Poisson data with a logarithmic link,

$$\prod_{j=1}^{m_g} f_{y|u}(y_{ij}|u_i) \propto \prod_{j=1}^{m_g} \exp\left[-m_g \exp\left(\tau_0 + u_i\right)\right] \exp\left[\sum y_{ij}\left(\tau_0 + u_i\right)\right].$$

The first order Laplace approximation involves approximating the logarithm of the integrand, the ith component of the h-likelihood,

$$h_i = \log\left\{\prod_{j=1}^{m_g} f_{y_j|u}(y_{ij}|u_i) f_u(u_i)\right\},$$

in (4.17) with a quadratic approximation around the mode, that is,

$$h_i \approx h_i|_{u_i=\tilde{u}_i} + \frac{1}{2} \left.\frac{\partial^2 h_i}{du_i^2}\right|_{u_i=\tilde{u}_i} \left(u_i - \tilde{u}_i\right)^2.$$

We can compare the true log-integrand, h_i, against its quadratic approximation at the mode, for given values of τ_0 and γ_1, and for given data y_{ij}, $j = 1, \ldots, m_g$. For

instance, for paired binary data ($m_g = 2$), ignoring terms not involving u_i,

$$h_i = \left(\sum_{j=1}^{2} y_{ij} \right) (\tau_0 + u_i) - 2 \log \left(1 + \exp \left[\tau_0 + u_i \right] \right) - \frac{1}{2\gamma_1} u_i^2,$$

with second derivative

$$\frac{\partial^2 h_i}{\partial u_i^2} = -2 \exp \left(\tau_0 + u_i \right) / \left(1 + \exp \left[\tau_0 + u_i \right] \right)^2 - \frac{1}{\gamma_1}.$$

When $y_{ij} = 1$, $j = 1, 2$, $\gamma_1 = 4$ and $\tau_0 = 0$, the quadratic approximation at the mode tends to under-estimate the true h_i away from the mode $\tilde{u}_i$ (Figure 4.7a), since $\tilde{u}_i$ is close to the point of maximum curvature of h_i (Figure 4.7b), and so the curvature of h_i is over-estimated. Therefore, the Laplace approximation will under-estimate the contribution to the likelihood for $\gamma_1 = 4$ and $\tau_0 = 0$ where $y_{ij} = 1$, $j = 1, 2$. By contrast, where $\tau_0 = 4$, and y_{ij} and γ_1 are as before, the quadratic approximation will over-estimate the true h_i (Figure 4.8a), since $\tilde{u}_i$ is far from the point of maximum curvature (Figure 4.8b).

It is the relative under- or over-estimation of the Laplace approximation for different values of τ_0 and γ_1 which determines its adequacy. For instance, if the Laplace approximation under-estimated the true likelihood contribution L_i by a constant proportion across all possible values of τ_0 and γ_1 and data y_{ij}, then the estimates derived using the Laplace approximated likelihood would be equal to the true maximum likelihood estimates. This is clearly not the case, as shown in Figures 4.7 and 4.8 for two different values of τ_0. The latter plot also suggests that the Laplace approximation may over-estimate the true likelihood where γ_1 and τ_0 are both increasing, and so might explain the divergence of the Laplace approximation seen in the simulations of section 4.2.2.1 where $\tau_0 = 2$.

Figures 4.7 and 4.8 also suggest that the reason why the Laplace approximation has difficulty with paired binary data is that the second derivative of h_i has a pronounced peak, in this case at $u_i = -\tau_0$. So the difference between the true h_i and its Laplace approximation will depend on the proximity of $\tilde{u}_i$ to $-\tau_0$. Other GLMMs may not

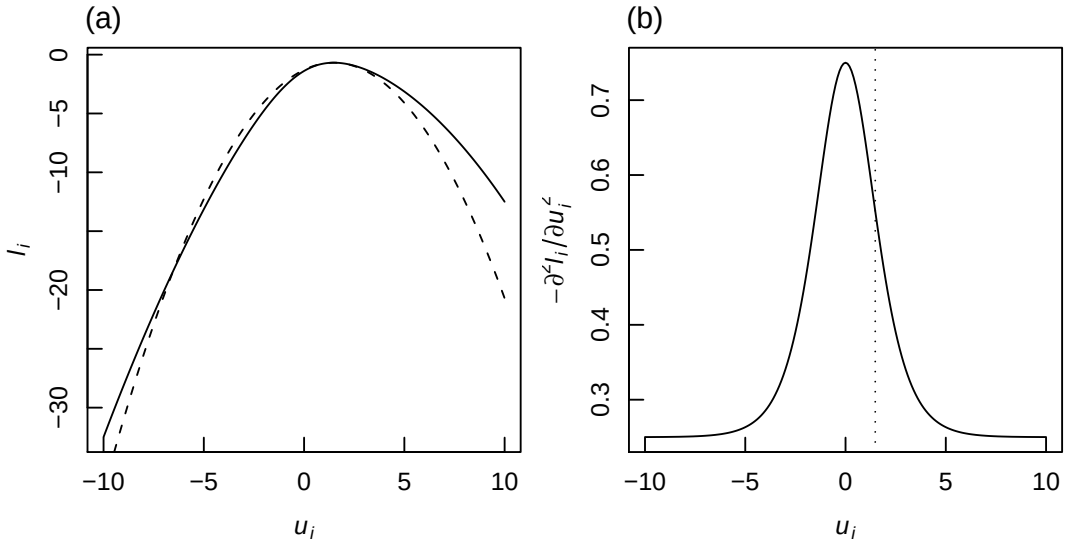


Figure 4.7: The Laplace approximation for a single group i in binary logit one-way classification (4.5) where $m_g = 2$, $\tau_0 = 0$, $\gamma_1 = 4$ and $y_{ij} = 1$, $j = 1, 2$
(a) The true log-integrand h_i of the contribution to the likelihood (4.17) versus u_i (solid). The quadratic approximation of h_i around the mode is superimposed (dotted).
(b) The curvature of h_i versus u_i. The position of the mode of h_i, $\tilde{u}_i = 1.48$, is shown as a dotted line.

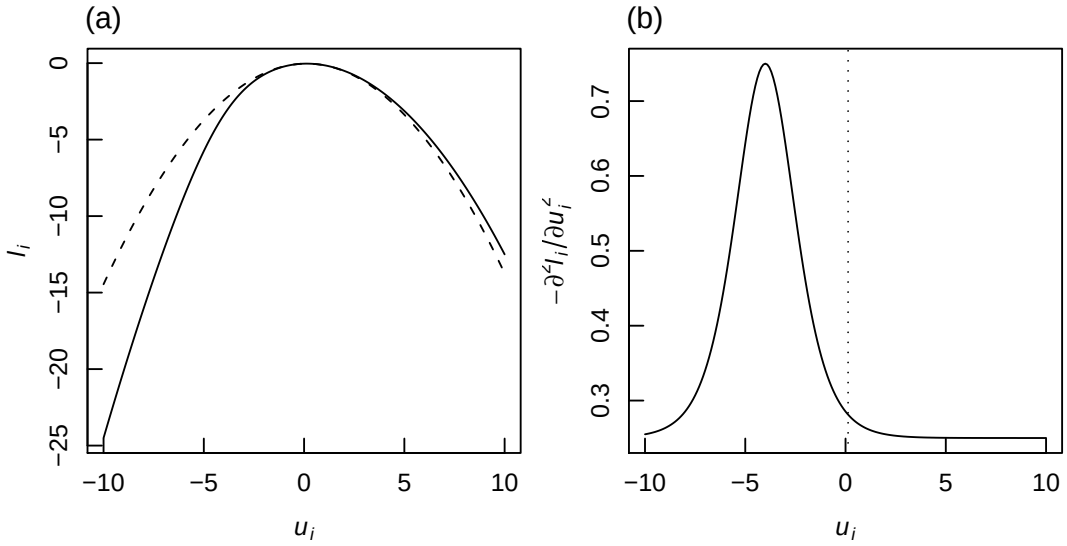


Figure 4.8: As for Figure 4.7, but where $\tau_0 = 4$.

have such a pronounced peak. For instance, for paired Poisson data with log link,

and ignoring terms not involving u_i,

$$h_i = \left(\sum_{j=1}^{2} y_{ij}\right)(\tau_0 + u_i) - 2\exp(\tau_0 + u_i) - \frac{1}{2\gamma_1}u_i^2,$$

with a monotically increasing second derivative with respect to u_i,

$$\frac{\partial^2 h_i}{\partial u_i^2} = -2\exp(\tau_0 + u_i) - \frac{1}{\gamma_1}.$$

The absence of a pronounced peak in the second derivative may explain the better performance of the Laplace approximation for Poisson data (compared to binary). For instance, Figure 4.9 shows the corresponding plots for $y_{ij} = 0$, $\tau_0 = 0$ and $\gamma_1 = 4$.

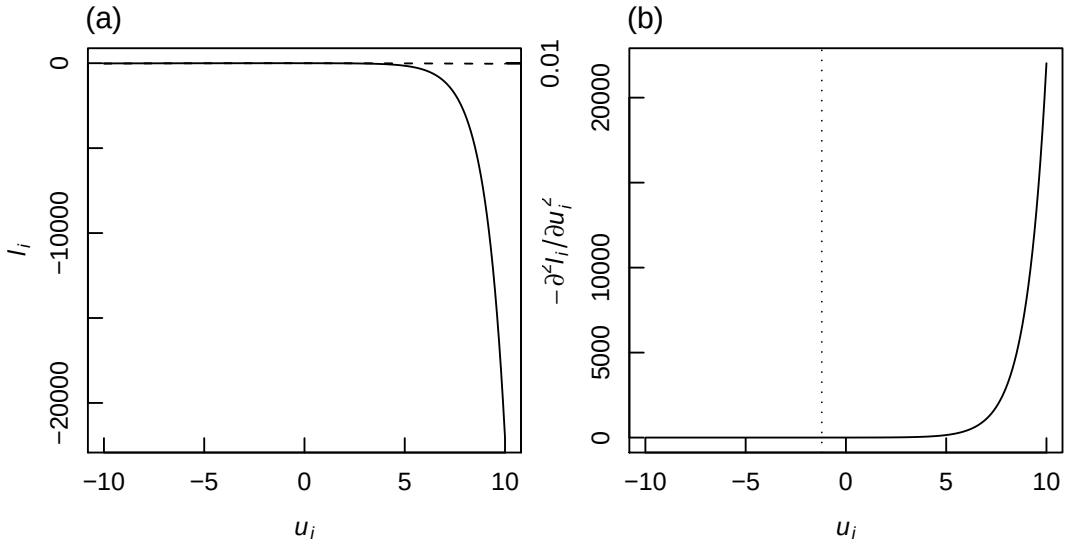


Figure 4.9: As for Figure 4.7, but for the Poisson log one-way classification where $m_g = 2$, $\tau_0 = 0$, $\gamma_1 = 4$ and $y_{ij} = 0$, $j = 1, 2$.

4.3 Second order HGLM approaches

This section discusses the implementation of the second order HGLM approaches HG(i,2), $i = 0, 1, 2$. Firstly, the calculations required to form the correction term in the second order Laplace approximation, denoted as $-F/24$ in Lee & Nelder (2001), will be shown. Secondly, the computation of this term is shown, with some Fortran style pseudo-code and suggestions on how it can be more efficiently computed. Fi-

nally, simulation studies, using the binary one-way classification and nested two-way classification designs, are presented.

4.3.1 An expression for the second order Laplace correction term

A second order Laplace approximation for the true likelihood ℓ is required to implement methods HG(i,2), $i = 0, 1, 2$ in Table 4.1:

$$p_u^s(h) \quad = \quad p_u(h) - F/24, \tag{4.18}$$

where $-F/24$ represents the difference between a first and second order Laplace approximation. Lee and Nelder's $-F/24$ notation is retained, however, it remains unknown what "F" refers to and why Lee & Nelder (2001) use "$-F/24$" instead of simply "F". The article by Reid (1991), that they cite with regard to this formula, does not allude to this notation. To add further confusion, Noh & Lee (2007) use trace(F)/24 instead of $F/24$, turning F into a matrix instead of a scalar.

The expression for F given in Lee & Nelder (2001, p. 996) and Noh & Lee (2007, p.898) is, using their own notation,

$$F = \text{trace}\left[-\left\{3\frac{\partial^4 h}{\partial \upsilon^4} + 5\frac{\partial^3 h}{\partial \upsilon^3} D(h,\upsilon)^{-1}\frac{\partial^3 h}{\partial \upsilon^3}\right\} D(h,\upsilon)^{-2}\right]_{\upsilon=\hat{\upsilon}}$$

where (again in their notation) $D(h,\upsilon) = -\partial^2 h/\partial \upsilon^2$, and υ represents the random effects on the scale of the linear predictor. This expression is a heuristic univariate style expression, and it is difficult to see exactly what calculations are required. For instance, since the higher order derivatives $\partial^3 h/\partial \upsilon^3$ and $\partial^4 h/\partial \upsilon^4$ are not intrinsically matrices, it is unclear what elements of these derivatives are being multiplied. However, result 1 in Noh & Lee (2007, p. 898) is more helpful, but it is overly complicated as described below.

A second order Laplace approximation of an integral can be created by extending the first order Laplace approximation, as presented in section 1.3.2.1, and taking higher order terms in the Taylor series expansion of the log of the integrand (see

Appendix A.2 for the derivation in the case of a univariate integral). This results in an expression for the correction factor $-F/24$:

$$\begin{aligned} -F/24 &= \frac{1}{8}\sum_{j,k,l,m} h^{(4)}_{jklm} g_{jk} g_{lm} \\ &\quad + \frac{1}{2}\left(\frac{1}{4}\sum_{j,k,l,r,s,t} h^{(3)}_{jkl} h^{(3)}_{rst} g_{jk} g_{lr} g_{st} + \frac{1}{6}\sum_{j,k,l,r,s,t} h^{(3)}_{jkl} h^{(3)}_{rst} g_{jr} g_{ks} g_{lt}\right), \end{aligned} \tag{4.19}$$

where

$$\begin{aligned} g_{jk} &= \left\{-\frac{\partial^2 h}{\partial \boldsymbol{u}\partial \boldsymbol{u}^T}\right\}^{-1}_{jk} \approx \left\{\boldsymbol{Z}^T \boldsymbol{W} \boldsymbol{Z} + \boldsymbol{G}^{-1}\right\}^{-1}_{jk}, \\ h^{(3)}_{jkl} &= \frac{\partial^3 h}{\partial u_j \partial u_k \partial u_l} = \sum_i \frac{\partial^3 h}{\partial \eta_i^3} z_{ij} z_{ik} z_{il}, \\ h^{(4)}_{jklm} &= \frac{\partial^4 h}{\partial u_j \partial u_k \partial u_l \partial u_m} = \sum_i \frac{\partial^4 h}{\partial \eta_i^4} z_{ij} z_{ik} z_{il} z_{im}. \end{aligned} \tag{4.20}$$

The second order Laplace approximation can also be expressed in matrix form as in Raudenbush *et al.* (2000), or using tensor notation as in Shun & McCullagh (1995), but the above univariate expression is the most convenient for computational purposes.

Noh & Lee (2007, p. 899) suggest a formulation of the term $-F/24$ which eliminates some of the unnecessary calculations, such as those involving zero cells. However, their formulation is based on their vector/matrix formulation, given in result 1 of their paper (p. 898 of Noh & Lee, 2007). This formulation suggests the need to calculate intermediary results, denoted C_{1i} and $C_{2i,i'}$ using Kronecker multiplications involving vectors $R_{(i,j)}$. (Note that there is an error in their formula for $C_{2i,i'}$. The term $R_{(i,i)} \otimes R_{(i,i')} \otimes R_{(i,i')}/8$ should be $R_{(i,i)} \otimes R_{(i,i')} \otimes R_{(i',i')}/8$.) However, their formulation is overly complex, since it apparently ignores a simple result concerning Kronecker multiplication of vectors: if $\boldsymbol{a} = (a_1 \dots a_n)^T$ and $\boldsymbol{b} = (b_1 \dots b_m)^T$ are both vectors and $\mathbf{1}_{nm}$ is the vector of length nm consisting of 1s, then

$$(\boldsymbol{a} \otimes \boldsymbol{b})^T \mathbf{1}_{nm} = \left(\sum_{i=1}^{n} a_i\right)\left(\sum_{j=1}^{m} b_j\right).$$

The expression for the second order Laplace approximation in (4.19) above can be further simplified. Let $R_{i,i'} = \sum_{j,k} z_{ij} z_{i'k} g_{jk}$, $e_{3i} = \partial^4 h / \partial \eta_i^3$ and $e_{4i} = \partial^4 h / \partial \eta_i^4$. Combining (4.19) with (4.20) gives

$$\begin{aligned}
\sum_{j,k,l,m} h^{(4)}_{jklm} g_{jk} g_{lm} &= \sum_{j,k,l,m} \left(\sum_i \frac{\partial^4 h}{\partial \eta_i^4} z_{ij} z_{ik} z_{il} z_{im} \right) g_{jk} g_{lm} \\
&= \sum_i \frac{\partial^4 h}{\partial \eta_i^4} \left(\sum_{j,k} z_{ij} z_{ik} g_{jk} \right)^2 = \sum_i e_{4i} R_{i,i}^2.
\end{aligned}$$

In addition,

$$\begin{aligned}
&\sum_{j,k,l,r,s,t} h^{(3)}_{jkl} h^{(3)}_{rst} g_{jk} g_{lr} g_{st} \\
= &\sum_{j,k,l,r,s,t} \left(\sum_i \frac{\partial^3 h}{\partial \eta_i^3} z_{ij} z_{ik} z_{il} \right) \left(\sum_{i'} \frac{\partial^3 h}{\partial \eta_{i'}^3} z_{i'r} z_{i's} z_{i't} \right) g_{jk} g_{lr} g_{st} \\
= &\sum_{i,i'} \left(\frac{\partial^3 h}{\partial \eta_i^3} \right) \left(\frac{\partial^3 h}{\partial \eta_{i'}^3} \right) \left(\sum_{j,k} z_{ij} z_{ik} g_{jk} \right) \left(\sum_{l,r} z_{il} z_{i'r} g_{lr} \right) \left(\sum_{s.t} z_{i's} z_{i't} g_{st} \right) \\
= &\sum_{i,i'} e_{3i} e_{3i'} R_{i,i} R_{i,i'} R_{i'i'} = \sum_i e_{3i}^2 R_{i,i}^3 + 2 \sum_{i'<i} e_{3i} e_{3i'} R_{i,i} R_{i,i'} R_{i',i'},
\end{aligned}$$

and similarly

$$\begin{aligned}
&\sum_{j,k,l,r,s,t} h^{(3)}_{jkl} h^{(3)}_{rst} g_{jr} g_{ks} g_{lt} \\
= &\sum_{i,i'} \left(\frac{\partial^3 h}{\partial \eta_i^3} \right) \left(\frac{\partial^3 h}{\partial \eta_{i'}^3} \right) \left(\sum_{j,r} z_{ij} z_{i'r} g_{jr} \right) \left(\sum_{k,s} z_{ik} z_{i's} g_{ks} \right) \left(\sum_{l.t} z_{il} z_{i't} g_{lt} \right) \\
= &\sum_{i,i'} e_{3i} e_{3i'} R_{i,i'} R_{i,i'} R_{i,i'} = \sum_i e_{3i}^2 R_{i,i}^3 + 2 \sum_{i'<i} e_{3i} e_{3i'} R_{i,i'} R_{i,i'} R_{i,i'}.
\end{aligned}$$

Therefore the resultant expression is

$$\begin{aligned}
-F/24 &= \sum_i \Bigg\{ R_{i,i}^2 \left(\frac{1}{8} e_{4i} + \frac{5}{24} e_{3i}^2 R_{i,i} \right) \\
&\quad + \sum_{i':i'<i} e_{3i} e_{3i'} \left(\frac{1}{4} R_{i,i} R_{i,i'} R_{i',i'} + \frac{1}{6} R_{i,i'} R_{i,i'} R_{i,i'} \right) \Bigg\}.
\end{aligned} \tag{4.21}$$

4.3.2 Computation of the second order Laplace correction term

Fortran 90 style pseudo-code to compute the second order Laplace correction factor in (4.21) is given in Table 4.5, assuming that $\boldsymbol{Z}$ is stored sparsely with non-zero elements $\boldsymbol{z} = (z_1 \dots z_{nsp})^T$ and corresponding row and column positions $\boldsymbol{r} = (r_1 \dots r_{nsp})^T$ and $\boldsymbol{c} = (c_1 \dots c_{nsp})^T$. The code assumes that the elements in $\boldsymbol{z}$, $\boldsymbol{r}$ and $\boldsymbol{c}$ are ordered by column within row. Define the matrix $\boldsymbol{D} = \left(\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z} + \boldsymbol{G}^{-1}\right)^{-1}$ so that d_{jk} is the (j,k)th element of $\boldsymbol{D}$. Let $\boldsymbol{Rd} = (R_{1,1}, R_{2,2}, \dots R_{n,n})^T$ be the "diagonal" elements where $i = i'$. The pseudo-code calculates $-F/24$ by proceeding row by row through $\boldsymbol{Z}$, with `i` storing the current row of $\boldsymbol{Z}$ and $\boldsymbol{Rn} = (R_{i,1} \dots R_{i,i-1})^T$. Note that the $R_{i,i'}$, $i' < i$ only need to be stored until contributions to $F/24$ of the current row are accumulated (that is, when we move to the next row of $\boldsymbol{Z}$, where `newrow!=i` in the pseudo-code).

Some modifications were made to the first order Fortran 90 code discussed in section 4.2.1 to implement the second order approximations HG(i,2), $i = 0, 1, 2$. Despite the use of sparse matrix techniques, the resulting implementation of the second order HGLM approaches was still rather slow, when combined with the two-stage numerical derivative approach required to implement HG(i, j) approximations where $i \geq 1$. There may, however, be further efficiencies to be made in the calculation of $R_{(i,i')}$, by creating a matrix/table of indices of $\boldsymbol{z}$ or $\boldsymbol{D}$ that need to be multiplied in the first iteration. This would avoid the need in subsequent iterations to doubly traverse $\boldsymbol{z}$, as is done in the above pseudocode. This has not been implemented.

4.3.3 Performance in simulation studies

The simulation studies conducted for the first order HGLM approximations (section 4.2.2) were repeated using the second order HGLM approximations, HG(i,2), $i = 0, 1, 2$, but for binary data only, since the first order HGLM approaches appeared to be adequate for the Poisson models.

```
Rd=0; Rn=0;
i=r(1)
Rd(i) = z(1)**2*D(c(1),c(1))
do j=2,nsp
  newrow=r(j)
  ! accumulate contributions from row j of Z to F24
  if(newrow!=i) then
     F24 = F24 + Rd(i)*Rd(i)* (e4(i)/8+5*(e3(i)**2)*Rd(i)/24)
     if(i>1) then
       do l=1,i-1
         temp1 = Rd(i)*Rn(l)*Rd(l)/4
         temp2 = Rn(l)*Rn(l)*Rn(l)/6
         F24 = F24 + e3(i)*e3(l)*(temp1+temp2)
       enddo
     endif
     Rn=0     !reset Rn for the next row of Z
     i=newrow
  endif
  zj=z(j)     ! save as scalars (in cache) instead of
  cj=c(j)    ! looking up z(j),c(j) for each k=1,j-1
  do k=1,j-1
    ip=r(k)
    if(ip==i) Rd(i)  = Rd(i)  + 2*zj*z(k)*D(cj,c(k))
    if(ip!=i) Rn(ip) = Rn(ip) + 2*zj*z(k)*D(cj,c(k))
  enddo
  Rd(i) = Rd(i) + zj**2*D(cj,cj)
enddo
```

Table 4.5: Fortran style pseudo-code required to compute the correction term (4.21) for the second order Laplace approximation.

4.3.3.1 One way classification (binary data only)

The one way classification model (4.5) is again considered, using the simulation parameter values given in Table 4.6.

Parameter	*Binary model*
b_g	50, 100, 200, 500
m_g	2, 4, 8, 16
γ_1	0.25, 1, 4, 9
τ_0	0, 2

Table 4.6: Values of the simulation parameters for the one-way classification study (4.5) comparing second order HGLM approaches and PQL.

Biases for $\hat{\gamma}_1$ are shown in Figure 4.10 for $b_g = 500$. Surprisingly, all second order

HGLM approximations performed better when $\tau_0 = 2$ than when $\tau_0 = 0$, in contrast to the performance of the first order approximations (section 4.2.2). For $\tau_0 = 0$, all the second order HGLM approximations gave similar average biases, and large positive biases for $\hat{\gamma}_1$ were observed when $m_g = 2$ or $\gamma_1 = 9$, probably due to diverging, or unusually high, estimates for some datasets (similar to the HG(1,1) approximation when $\tau_0 = 2$). When $\tau_0 = 2$, the HG(0,2) approximation was little better than PQL for small m_g. However, HG(1,2) and HG(2,2) estimators both had little or no bias, except where $m_g = 2$ or $\gamma_1 = 9$. For $\hat{\tau}_0$, the use of HG(1,2) and HG(2,2) also generally resulted in little bias, except when $\gamma_1 = 9$, where some positive bias of the HG(2,2) (and HG(0,2)) estimators is observed (Figure 4.11).

4.3.3.2 Nested two-way classification model (binary data only)

Simulations using the nested two-way classification model (4.6) were also performed. Owing to the current computational slowness of the second order HGLM implementation, only the combinations $(m_g, m_s) = (2, 2)$, (2,4) and (4,2) were explored, with the remaining parameter value settings as given in Table 4.3.

Figure 4.12 shows the estimation biases for the second order HGLM approximations against PQL. The second order HGLM approximations generally had low estimation biases, except when $\gamma_1 = 1$ and $\gamma_2 = 4$, where there was strong negative and positive bias for $\hat{\gamma}_1$ and $\hat{\gamma}_2$ respectively. Further examination of these simulated datasets in this case showed that there were many simulations where $\hat{\gamma}_1$ went to 0 and the estimate of $\hat{\gamma}_2$ diverged.

4.4 Discussion

The simulation studies in this chapter show that first order HGLM approaches generally do better than PQL with respect to having estimation biases of smaller magnitude. One major difference in the estimation biases between PQL and first order HGLM approaches is that the magnitude of the biases for PQL increase markedly

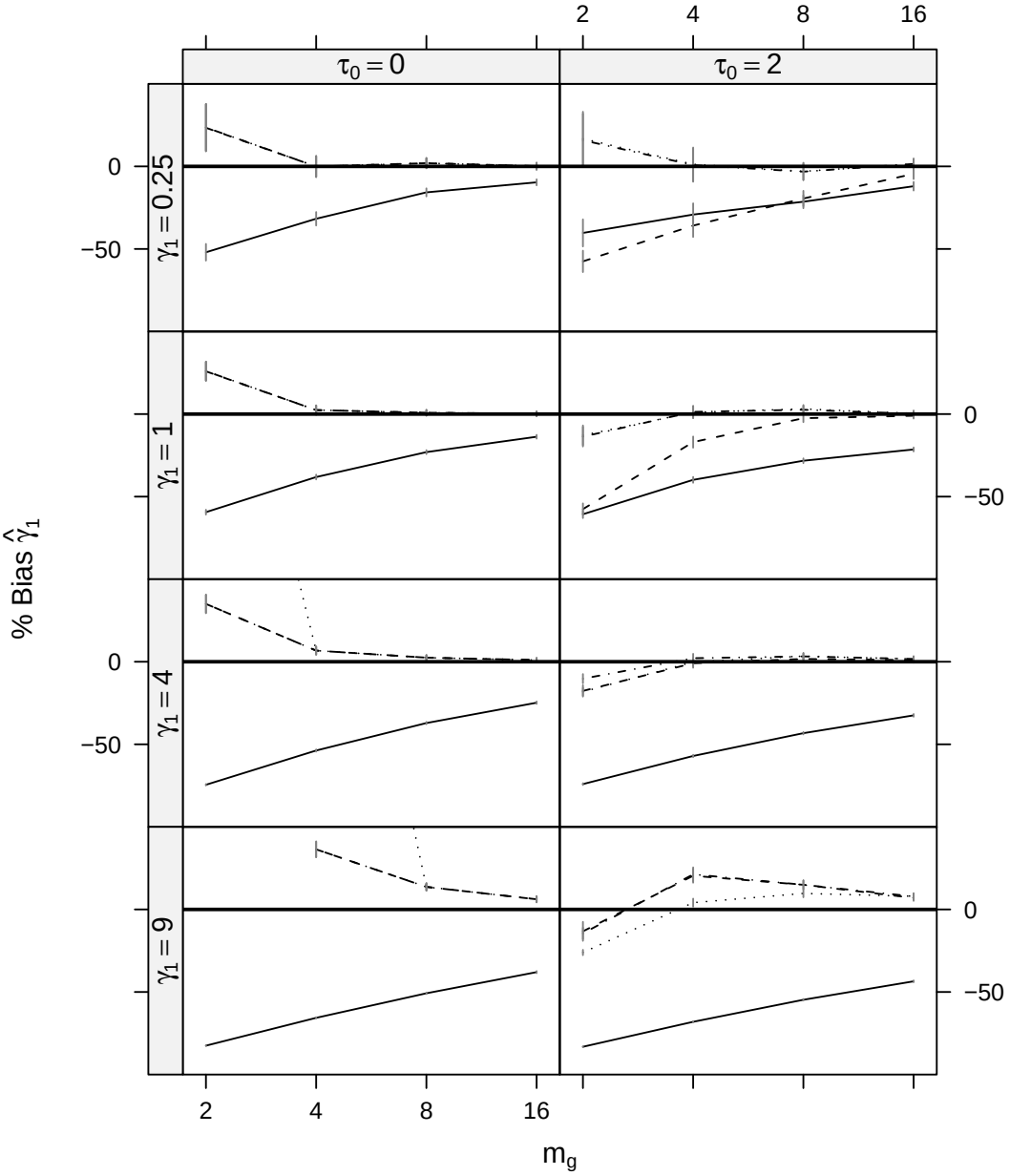


Figure 4.10: Interactions of the effects of m_g and γ_1 on average biases for $\hat{\gamma}_1$ for the binary one way classification model (4.5) where $b_g = 500$. (PQL: solid; HG(0,2): dashed; HG(1,2): dotted; HG(2,2): dot-dashed). Error bars are ±2SE.

with the magnitude of the variance component, $\boldsymbol{\gamma}$, whereas the biases for HG(1,1), and HG(0,1) for $\hat{\boldsymbol{\gamma}}$, are stable across values of $\boldsymbol{\gamma}$. An explanation for the increase in the biases with $\boldsymbol{\gamma}$ for PQL has been provided in section 4.2.3.4. It would be appear that using the first order HGLM approaches would be much more preferable to PQL when the variance components are expected to be large. However, some caution is

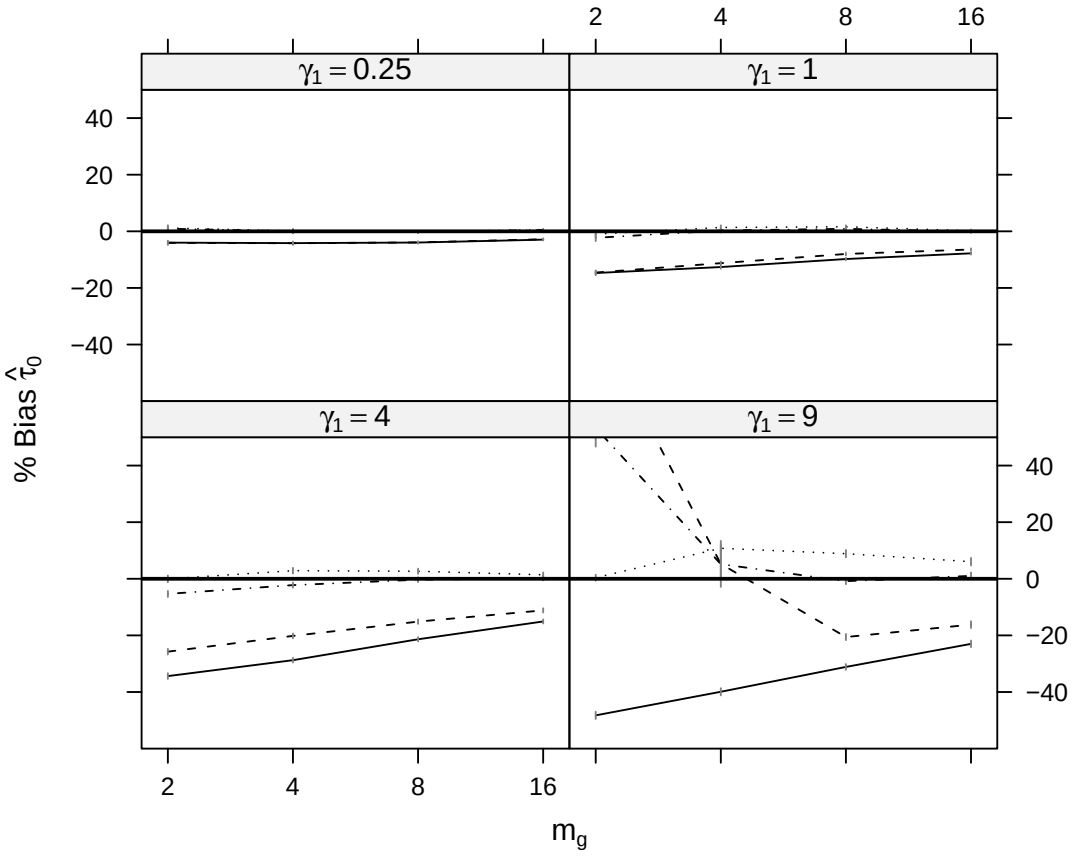


Figure 4.11: Interactions of the effects of m_g and γ_1 on average biases for $\hat{\gamma}_1$ for the binary one way classification model (4.5) where $b_g = 500$. (PQL: solid; HG(0,2): dashed; HG(1,2): dotted; HG(2,2): dot-dashed) Error bars are ±2SE.

advised when dealing with binary data, or binomial data with low denominators, given the instability of the HG(1,1) estimators for the binary one-way classification when $\tau_0 = 2$ and $m_g \leq 4$ (Figure 4.5 and accompanying text). Binary models in general with low group sizes and average probabilities well away from 0.5 may also be susceptible to this instability. The intuitive arguments given in section 4.2.4 suggest that this instability may be restricted to binary models (or binomial with low denominators), where the curvature or second derivative of the likelihood integrand with respect to a random effect, i.e. $-\partial^2 h/\partial u_i^2$ **,** has a sharp peak.

For HG(1,1), the biases for $\hat{\tau}_0$, for both binary and Poisson one way classification and the nested two way binary model, are relatively small in magnitude. Therefore, it appears that the first order Laplace approximation may be adequate for estimation of

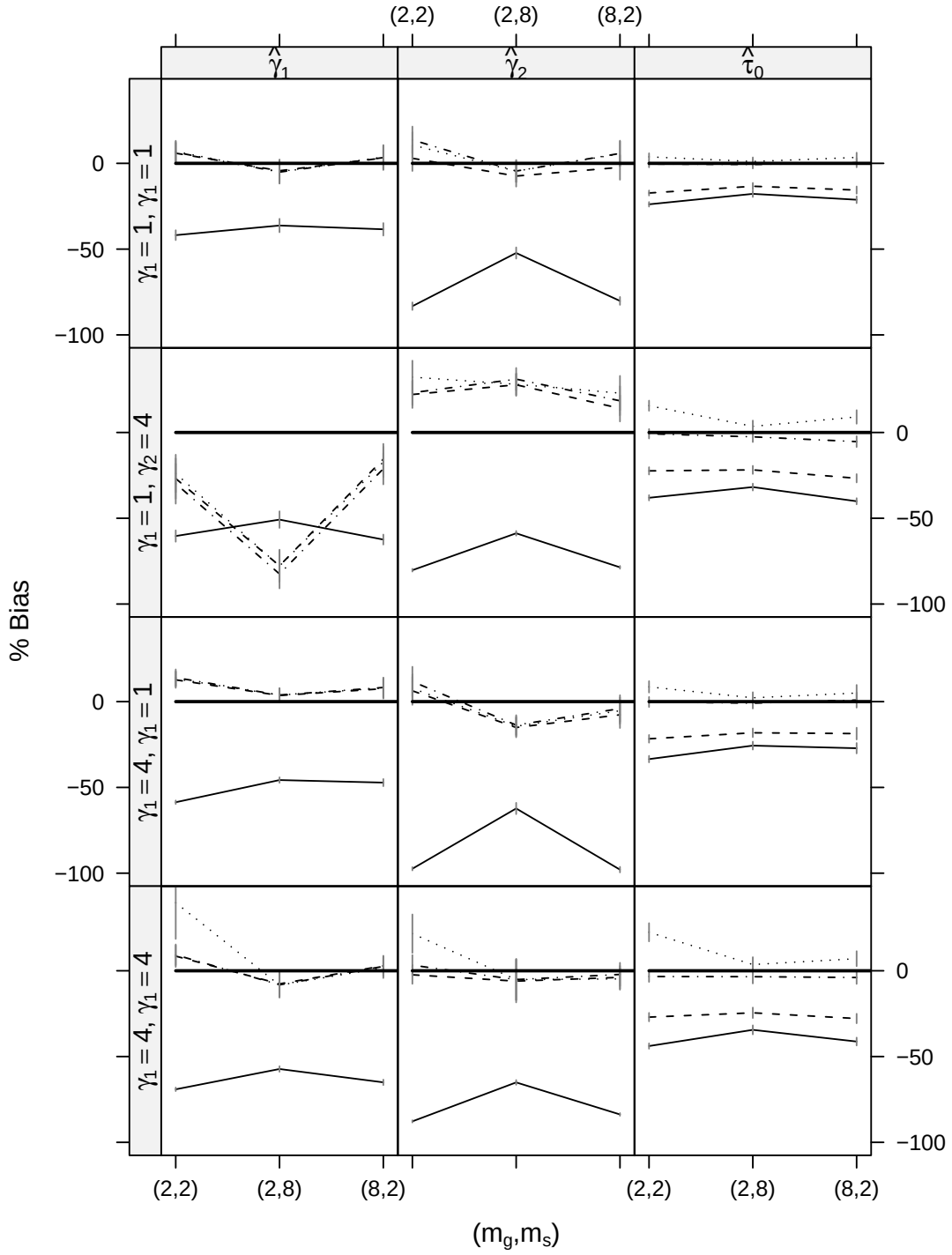


Figure 4.12: Interactions of the effects of m_g, m_s, γ_1 and γ_2 on the biases for $\hat{\gamma}_1$, $\hat{\gamma}_2$ and $\hat{\tau}_0$ for the binary nested two-way classification model (4.6) where $b_g = 200$ and $\tau_0 = 1$. (PQL: solid; HG(0,2): dashed; HG(1,2): dotted; HG(2,2): dot-dashed) Error bars are ±2SE.

fixed coefficients in GLMMs. However, the accuracy of the estimated variances and SEs of the fixed coefficients, and the use of approximate Wald tests for hypothesis testing of fixed coefficients, have not been examined here.

Although first order HGLM approaches might be adequate for estimation of the fixed coefficients, and for estimating the variance components in Poisson models, it appears second order approaches are still required for estimating the variance parameters in binary models. The two simulation studies in section 4.3.3 show that second order approaches can substantially reduce the biases for the variance parameter estimators, but at the cost of instability and divergence (where unusually large estimates of the parameters were found) in some cases. More work is required to establish when second order approaches are preferable over first order approaches, and whether this added instability and divergence negates the beneficial decreases in the magnitude of the estimation biases.

Comparison to other approaches for GLMMs, such as GHQ and Bayesian MCMC approaches, is necessary to determine the relative merits of approximate likelihood approaches such as PQL and the HGLM approach. This will be explored in the next chapter, using a series of case studies.

Chapter 5

Case studies

Chapters 3 and 4 explored two approximate likelihood approaches, PQL and HGLM approaches, especially with respect to problems with estimation biases. However, a fuller assessment of the merits of using these approximate likelihood approaches requires comparison to other alternative GLMM approaches. In this chapter, such comparisons will be endeavoured against two prominent alternatives, Gauss-Hermite quadrature (GHQ) and Bayesian approaches, using a select group of case studies.

5.1 Preliminaries

5.1.1 Review of alternative approaches

A review of some key issues concerning the use of each of the alternatives to approximation likelihood approaches for GLMMs, GHQ and Bayesian approaches, is presented in this section.

5.1.1.1 Gauss-Hermite Quadrature

As outlined in the background theory of section 2.2, Gauss-Hermite quadrature (GHQ) is a numerical integration approach for evaluating the GLMM likelihood expression. The major disadvantage of GHQ, as noted in section 2.2, is that it is only

feasible for GLMMs involving nested random effects. In addition, most current implementations of GHQ only cater for, at most, two-way nested random effects models, such as the nested two-way classification (3.3). This severely restricts the utility of GHQ for agricultural and biological data, where multiple, and often non-nested, sources of variation are present.

Of the two forms of GHQ, standard and adaptive, the adaptive version is now strongly preferred in the literature. Standard GHQ can suffer appreciably from numerical instability (Rabe-Hesketh *et al.*, 2005). This has been demonstrated in the case of grouped data when the group sizes are large – in order for GHQ to converge in these cases, good starting values close to the optimum values are often required. Rabe-Hesketh *et al.* (2005) argued that, in general, standard GHQ will have difficulty in situations where there is ample "information" concerning each random effect, giving a very "peaked" integrand in the expression for the GLMM likelihood (1.9). Lesaffre & Spiessens (2001) illustrated the numerical problems associated with GHQ for a simple grouped GLMM, and demonstrated the superiority of adaptive GHQ, which had less numerical instability using fewer quadrature points. It is interesting to note here that, in contrast to standard GHQ, approximate likelihood techniques perform better when there is ample information per random effect (as confirmed in the simulation studies of Chapter 3). Despite the benefits of using adaptive GHQ over standard GHQ, however, there are relatively few implementations of adaptive GHQ at present. Two implementations of adaptive GHQ are the `NLMIXED` subroutine (Wolfinger, 1999) in the SAS statistical package (SAS Institute Inc., 2000) and the `gllamm` suite of functions (Rabe-Hesketh *et al.*, 2001) in the Stata statistical package (StataCorp, 2007). Of the two, only the `gllamm` implementation caters for GLMMs with more than one level of nested random classification. The current lack of implementations catering for GLMMs with multiple random classifications may be a consequence of the difficulties in implementation for such GLMMs, as discussed in Rabe-Hesketh *et al.* (2002, 2005).

5.1.1.2 Bayesian/MCMC approaches

Background theory in the use of full Bayesian approaches for GLMMs has been provided in section 2.3.2. The application of full Bayesian, and other MCMC, approaches to inference for GLMMs has been popular in recent years in the research literature. For instance, Rodriguez & Goldman (2001) and Browne & Draper (2006) demonstrated the superiority of full Bayesian approaches over PQL for their simulation studies, in terms of both reduced estimation bias and providing 95% coverage intervals which are closer to the nominal coverage rates. However, despite their apparent popularity in the research literature, the use of full Bayesian approaches for GLMMs in applied statistical work still appears to be limited, possibly due to inertia, perceptions of difficulty and/or perceptions regarding the influence of the priors on inference. Some of this perception is well-founded, since the use of Bayesian approaches requires some additional user expertise not required when using likelihood-based approaches, for instance, in choosing appropriate priors and monitoring convergence. Another impediment to the adoption of Bayesian approaches is the current lack of available "off the shelf" software, apart from `WinBUGS` (Spiegelhalter *et al.*, 1995).

There are important methodological issues in using Bayesian approaches which are still being resolved. For instance, the influence of the Inverse Gaussian prior for variance parameters, which is the standard prior used for variance parameters, has been recently questioned (Gelman, 2005). However, Browne & Draper (2006) showed that the impact of the prior choice appeared to be minor in their simulations, especially when compared to the large estimation biases associated with PQL.

5.1.2 Software used in these case studies

The software used in the case studies which follow, and some implementational issues associated with their use, are discussed in this section.

5.1.2.1 Approximate likelihood approaches

As in previous chapters, PQL was implemented using ASReml version 2.0 (Gilmour *et al.*, 2006).

The HGLM approaches were implemented using the Fortran 90 code outlined in the previous chapter. In addition, non-REML versions of the HGLM approximations are also examined, that is, with the REML-like correction removed, in order to examine the importance of this correction to the estimation of $\boldsymbol{\gamma}$. For instance, the first order approximation HG(1,1) uses $p_\beta(h)$ (4.2) as the likelihood criterion for $\boldsymbol{\gamma}$. A non-REML version of HG(1,1) uses $p_u(h)$ (4.1) as the likelihood criterion for $\boldsymbol{\gamma}$ instead, which is also the likelihood criterion for $\boldsymbol{\tau}$. The non-REML versions also shared more similarity with the GHQ approach, in that the GHQ approach does not apply a REML-like correction either. It should be noted that other implementations of the first order Laplace approximation for GLMMs also use no REML-like correction, such as the `glmmadmb` and `lmer` functions in the R statistical package. Similarly, non-REML versions of the second order approximations are examined where the likelihood criterion for $\boldsymbol{\gamma}$ is $p_u^s(h)$ instead of $p_\beta^s(h)$. To implement in the non-REML versions of the HGLM approximations in the Fortran 90 implementation (section 4.2.1), a simple modification of the code was required.

In addition, a simple implementation of PQL was implemented in Fortran 90, in addition to ASReml. This implementation used an estimation scheme which alternated between estimation of the variance parameters $\boldsymbol{\gamma}$ and estimation of the fixed and random effects $\boldsymbol{\tau}$ and $\boldsymbol{u}$. It used the same matrix libraries as for the Fortran 90 implementation of the HGLM approach, as described in section 4.2.1. The development of this PQL implementation also allowed us to create a non-REML version of PQL, similar to the non-REML versions of HGLM described above. This non-REML version of PQL is used in the RCBD case study (section 5.2.5).

5.1.2.2 Bayesian/MCMC

For the full Bayesian approaches, the classic `BUGS` program[1] was used in preference to `WinBUGS`, since it could be called non-interactively using a script on a Unix system. (At the time of performing simulations, the newer `OpenBUGS` program[2] appeared to be quite unstable and prone to give erroneous results.)

Standard priors were used for the parameters in the model, as used in the library of examples included with the `BUGS` software. For fixed coefficients τ_i, normal priors with large variance (10,000) were used. For the variance parameters, the standard Inverse Gamma prior (IG) was used, where the prior for the reciprocal of the variance parameter was a gamma distribution,

$$f(x; r, \mu) = \frac{\mu^r x^{r-1} e^{-ux}}{\Gamma(r)}.$$

The Inverse Gamma prior for the variance parameter is subsequently denoted IG(r, μ). The parameter settings $r = 0.001$, $\mu = 0.001$ were used by default, but in some studies, alternative settings $r = 0.1$, $\mu = 0.1$ were also examined, to test the robustness of the estimates to the choice of r and μ. Initial values for the parameters were 1 for variance parameters and 0 for fixed coefficients and random effects. Experimentation with different initial values for some sample datasets generated in these studies showed that there was no dependence of the generated posterior distributions on the initial values chosen. The choice of the number of samples using the Gibbs sampler, including the number of "burn-in" samples, was made informally via examination of trace plots for a few of the simulated datasets in each study. However, for some studies, as noted below, the number of samples was increased by 50 or 100% to determine whether there was any change in the average (posterior) estimates. Unless otherwise stated, the estimators from the Bayesian approach are the means of the respective posterior distributions, however, in most cases the posterior medians are presented as well, for reasons to be discussed.

[1] `http://www.mrc-bsu.cam.ac.uk/bugs/classic/contents.shtml`
[2] `http://www.mathstat.helsinki.fi/openbugs/`

In recent years, the routine use of the Inverse Gamma prior as an "uninformative" prior for a variance parameter has been questioned, and other diffuse priors have been suggested. Gelman (2005) suggests the use of either a Uniform(0,A) prior on the squareroot of the variance parameter (i.e. the standard deviation), or using a prior in the "half-t" family. Preliminary investigations of these priors showed that the Uniform prior fared more poorly with respect to bias than the Inverse Gamma prior, with the results being highly dependent on the choice of the upper limit A. (A report on one investigation is available from the author.) However, an alternative "half-Cauchy" prior, from the half-t family, has been investigated for some of the simulation studies here, where the prior distribution is

$$f(x; A) \propto \left(1 + \frac{x}{A^2}\right)^{-1}.$$

`BUGS` code to implement this prior is given in the appendix of Gelman (2005). Two arbitrary values of A, 3 and 30, were chosen, to examine the sensitivity of the results to the setting of this parameter. These two priors are denoted HC(3) and HC(30) respectively.

5.1.2.3 Quadrature

For GHQ approaches, SAS's `NLMIXED` procedure (Wolfinger, 1999) was used. This procedure implemented adaptive GHQ (AGHQ). By default, the `NLMIXED` procedure automatically determines an "appropriate" number of quadrature points based on the data. However, the user can also specify the number of quadrature points in the call to `NLMIXED`. `NLMIXED` only allows for non-nested grouped data, and so could not be used for a nested two-way classification model, such as used for the Rodriguez-Goldman datasets examined in section 5.2.4. For these models, the software `AML` (Lillard & Panis, 2003) was used, which implements standard (non-adaptive) quadrature.

5.1.2.4 Other notes

As for simulation studies of previous chapters, the estimates of the variance parameters were constrained to be positive, unless the variance parameter was a correlation coefficient. This was required here anyway, since most of the software implementations above did not allow negative variance parameter estimates.

5.2 Simple comparisons

5.2.1 The Beitler-Landis dataset

We return to the Beitler & Landis (1985) dataset (3.9) given in section 3.1.7. As already noted in section 3.1.7, there is a good level of agreement between the PQL and AGHQ estimates (see Table 5.1). The test statistics for testing $H_0 : \gamma_1 = 0$ (last column of Table 5.1) are also remarkably similar between PQL and AGHQ. However the Bayesian approach using `BUGS` gives very different estimates, and much wider SEs. (Here, 20,000 iterations of the Gibbs sampler were used to calculate the posterior distributions, after a 2,000 iteration burn-in. A IG(0.1,0.1) prior was also tried, but resulted in negligible change to the estimates using the IG(.001,.001) prior. Finally, an alternative HC(3) prior was also tried, but this resulted in a somewhat higher posterior mean for γ_1 of 3.565.

	$\hat{\tau}_0$	$\hat{\tau}_1$	$\hat{\gamma}_1$	**LRT** $H_0 : \gamma_1 = 0$
PQL (`ASReml`)	-0.784 ± 0.537	0.724 ± 0.296	2.033 ± 1.250	50.4
AGHQ (`NLMixed`)	-0.828 ± 0.533	0.739 ± 0.300	1.960 ± 1.190	55.4
Bayes (`BUGS`)	-0.828 ± 0.636	0.753 ± 0.303	3.248 ± 2.751	

Table 5.1: Estimates from the analysis of Beitler/Landis data (table 3.8, model 3.9) using PQL, adaptive GHQ and Bayesian approaches.

An explanation for the difference between the Bayesian and the other estimates may be found in the (estimated) profile likelihoods for γ_1. The estimated profile likelihoods for γ_1 (Figure 5.1) using AGHQ, HG(0,1) ($p_\beta(h)$) and PQL likelihoods are reasonably similar, with similar modes (AGHQ$\approx$ 2.0, HG(0,1)$\approx$2.3, PQL$\approx$1.5). Each profile

likelihood is relatively flat with respect to γ_1 for values greater than the mode. Thus, there is little change in the likelihood between the Bayesian estimate and the two other estimates. The Bayesian estimates presented in Table 5.1 were the means of the respective posterior distributions. The median of the posterior distribution for γ_1 (2.44) was much more similar to the other non-Bayesian estimates. This median was also consistent with the mode of the HG(1,2) likelihood (2.48) (for simplicity, this profile likelihood is omitted from Figure 5.1).

In addition, the mode for the HG(0,1) likelihood, or HG(0,1) estimate, is greater (2.3) than that for AGHQ (2.0). The difference between the two modes could be due to the fact that HG(0,1) has an (approximate) REML-like correction in the likelihood for $\boldsymbol{\gamma}$, whereas AGHQ does not. If a non-REML HG(0,1) is used, with the REML-like correction is omitted, as discussed in section 5.1.2.1, the mode of the non-REML HG(0,1) likelihood is similar to the AGHQ mode (1.9). Also, note that the PQL estimate of 2.03 is not the mode of the PQL profile likelihood (1.5), since the PQL approach does not maximise its own likelihood. However, the PQL profile likelihood has a similar shape to AGHQ likelihood, and so the LRT statistic calculated from the PQL profile likelihood for testing $H_0 : \gamma_1 = 0$ in Table 5.1 is also similar to the LRT statistic calculated from the AGHQ likelihood.

5.2.2 A paired binary simulation study

A paired binary example was used to compare the different GLMM approaches. As discussed in chapter 3, the paired binary case is a well-known example where PQL exhibits significant estimation biases, and so this was seen as a suitably challenging comparison of the different approaches. A total of 200 simulated datasets was generated from the following model for data y_{ij}, $i = 1 \dots 100$, $j = 1, 2$, with conditional mean $\mu_{ij} = \mathrm{E}(y_{ij}|u_i)$:

$$\mathrm{logit}(\mu_{ij}) = \tau_0 + \tau_1 x_{1ij} + \tau_2 x_{2ij} + u_i, \tag{5.1}$$

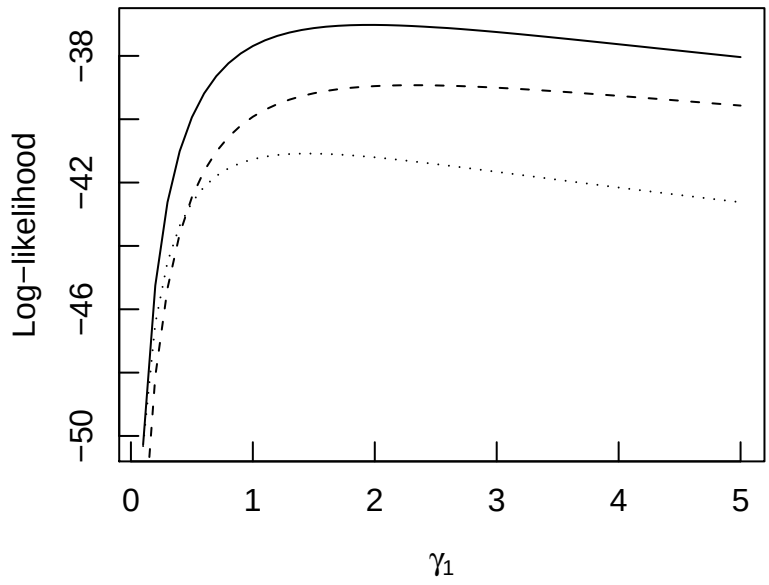


Figure 5.1: Profile likelihoods of γ_1 in the Landis-Beitler model (3.9) for AGHQ, PQL and HG(0,1) approaches (AGHQ: solid; HG(0,1): dashed; PQL: dotted) (A constant of 29 has been added to the PQL profile likelihood.)

and $u_i \sim N(0, \gamma_1)$, $(\tau_0, \tau_1, \tau_2)^T = (1.5, 1.0, 0.01)^T$, and $\gamma_1 = 2.25$. Each of the covariates, $x_{1ij} = (i - 50.5)$ and $x_{2ij} = 2(j - 1.5)$, were centred, that is, $\sum_{i,j} x_{kij} = 0$, for $k = 1, 2$, and their values varied between and within pairs respectively. Each of these 200 simulated datasets was analysed according to (5.1) using PQL, HGLM approaches, Bayesian (`BUGS`) and AGHQ (`NLMIXED`) approaches, with the average estimates from each approach presented in Table 5.2. For the Bayesian (`BUGS`) approach, 20,000 samples were generated to calculate the posterior distributions, after a 2,000 iteration burn-in.

Of the 200 simulations, only 196 converged for Bayesian and AGHQ approaches (the latter using the default `NLMIXED` calculation of quadrature points). For the other four simulations, `BUGS` reported "`sing error`" and gave estimates of γ_1 over 10,000. Results are reported for the remaining 196 simulations. Of the HGLM approaches, the HG(0,2), HG(2,2) and HG(1,1) approaches also had divergence problems with a large number of simulated datasets (>10), where the estimates of γ_1 greatly exceeded the true value of 2.25, and so for brevity are omitted.

As expected from previous simulation studies, the PQL estimator for each of the

parameters exhibited very large negative biases, especially for $\hat{\gamma}_1$. The biases for the HG(0,1) estimators were also negative, but considerably smaller than PQL for $\hat{\gamma}_1$. The Bayesian (posterior mean) and HG(1,2) estimators both had positive biases for each of the parameters. For the Bayesian approaches, the median of the posterior distribution had less positive bias than the posterior mean. Using an IG(0.1, 0.1) prior, instead of the default IG(.001, .001) prior, only exacerbated the positive biases, and so are not presented. The use of either a HC(3) or HC(30) prior also exacerbated the positive biases – for instance, for the HC(3) prior resulted in average estimates for γ_1 of 4.924±0.285 for the posterior mean and 3.997±0.222 for the posterior median, with the HC(30) average estimates even higher still. Using a longer Gibbs sampler chain was also investigated, but made little difference to the average estimates, and so also are not presented. The default AGHQ estimator appeared to be slightly negatively biased for γ_1. Further investigation showed that `NLMIXED` had used only 3 quadrature points for each dataset (as noted above in section 5.1.2, `NLMIXED` automatically determines a suitable number of quadrature points to use). Repeating the AGHQ analyses with 9 quadrature points for each dataset resulted in non-trivial changes to the average estimates, indicating that the default choice of 3 quadrature points was too small. The use of non-REML versions of the HG(0,1) and HG(1,2) approximations reduced the average estimate of γ_1 for both HG(0,1) and HG(1,2). These reductions in the average estimates lead to greater negative bias for γ_1 for the non-REML version of HG(0,1) than for the standard HG(0,1) approximation, but lower positive bias in the non-REML HG(1,2) approximation than for the standard HG(1,2) approximation. The average HG(1,2) estimates were more similar to the average Bayesian (posterior median) estimates than with the average AGHQ estimates, whereas the average non-REML HG(1,2) estimates were more similar to the average AGHQ estimates.

To illustrate some of the differences between the AGHQ estimates and the other estimates for γ_1, Figure 5.2 shows scatterplots of the 200 estimates of γ_1 for PQL, Bayesian, HG(1,2) and non-REML HG(1,2) approaches plotted against the corresponding 200 AGHQ estimates. The scatterplot of the PQL vs AGHQ estimates (top

	γ_1 (2.25)	τ_0 (1.5)	τ_1 (1)	τ_2 (0.01)
PQL	0.696 ± 0.021	1.157 ± 0.015	0.793 ± 0.012	0.008 ± 0.000
HG(0,1)	1.476 ± 0.075	1.213 ± 0.016	0.840 ± 0.013	0.008 ± 0.001
HG(0,1) (noREML)	1.062 ± 0.051	1.185 ± 0.015	0.815 ± 0.012	0.008 ± 0.000
Bayesian (mean)	4.375 ± 0.308	1.730 ± 0.040	1.157 ± 0.028	0.012 ± 0.001
Bayesian (median)	3.358 ± 0.238	1.659 ± 0.037	1.124 ± 0.026	0.011 ± 0.001
AGHQ (default)	2.171 ± 0.094	1.510 ± 0.024	1.031 ± 0.019	0.010 ± 0.001
AGHQ (points=9)	2.668 ± 0.133	1.577 ± 0.028	1.062 ± 0.020	0.011 ± 0.001
HG(1,2)	3.824 ± 0.174	1.798 ± 0.033	1.167 ± 0.024	0.012 ± 0.001
HG(1,2) (noREML)	2.774 ± 0.126	1.646 ± 0.028	1.083 ± 0.020	0.011 ± 0.001

Table 5.2: Average estimates (±SE) for 196 simulations from (5.1) using the PQL, Bugs, AGHQ and HGLM approaches.

left in Figure 5.2) suggests that a non-linear correction to the PQL estimates would be required to correct the PQL bias (assuming the AGHQ estimates represent the "gold standard"), that is, the negative PQL bias appears to increase in magnitude with the AGHQ estimate of γ_1. The plots of the Bayesian (posterior mean) and HG(1,2) estimates against AGHQ show that both the Bayesian and HG(1,2) estimates increasingly diverge from the AGHQ estimates as the AGHQ estimate of γ_1 increases. The non-REML HG(1,2) estimates appear to have good consistency with the AGHQ estimates across the entire range of AGHQ estimates for γ_1.

5.2.3 Further paired binary (and Poisson) simulation studies

A further paired data simulation study was performed, using a simpler design with no covariates. Both Poisson and binary data were generated and analysed in this study. A total of 200 datasets were simulated from the following model for the conditional mean $\mu_{ij} = E(y_{ij}|u_i)$ of data y_{ij}, $i = 1, \ldots, 100$, $j = 1, 2$,

$$g(\mu_{ij}) = \tau_0 + u_i, \tag{5.2}$$

where $u_i \sim N(0, \gamma_1)$. The link function $g(\cdot)$ was the logit and logarithmic link for binary and Poisson data respectively. The settings $\tau_0 = 0$ and $\gamma_1 = 0.80$ were chosen. Each dataset was analysed using PQL (ASReml), Bayesian (`BUGS`), AGHQ (`NLMIXED`), HG(1,1) and HG(1,2) approaches, along with non-REML versions of the

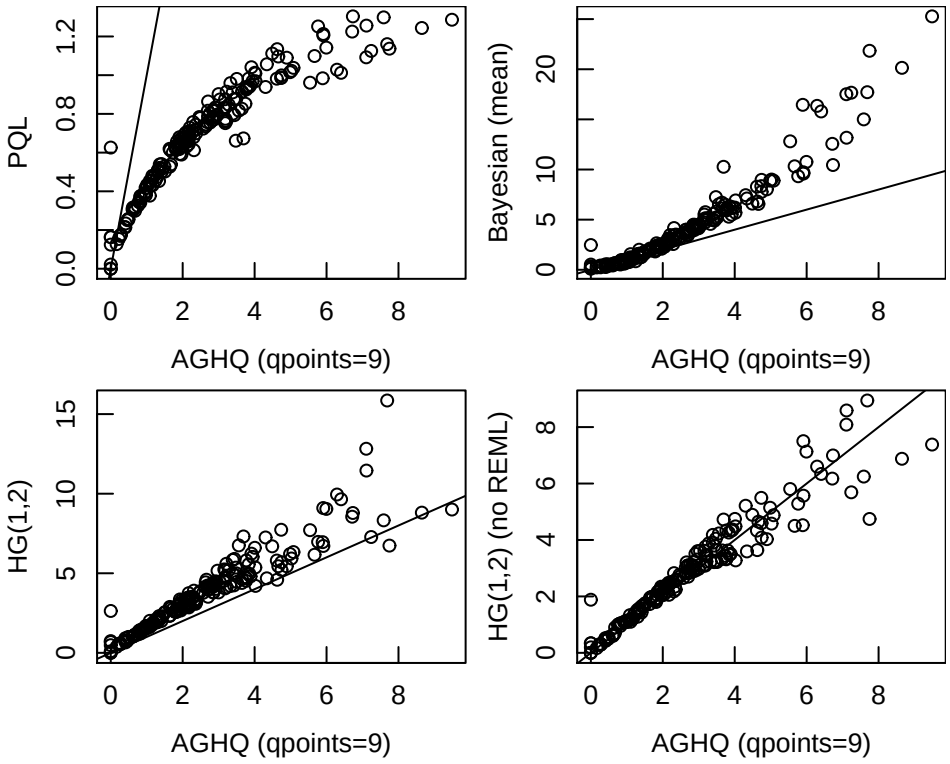


Figure 5.2: AGHQ estimates of γ_1 from 196 simulations of (5.1) versus those from PQL, Bayesian (mean), HG(1,2) and HG(1,2) (no REML correction) approaches. The one to one line is shown on each plot.

HG(1,1) and HG(1,2) approaches. For the Bayesian (`BUGS`) approach, 10,000 samples were generated to calculate the posterior distributions, after a 1,000 iteration burn-in.

Average estimates for each approach are shown in Table 5.3. As for the previous simulation study, the PQL estimator for γ_1 showed strong negative bias in the binary case, and also negative bias in the Poisson case, but to a much lesser extent than the binary case. For the binary case, both the standard and non-REML HG(1,1) estimators were also negatively biased for γ_1, but these estimators showed little apparent bias in the Poisson case. Both standard and non-REML HG(1,2) estimators, by contrast, exhibited positive bias for γ_1 in the binary case, but, like the HG(1,1) estimators, had no apparent bias in the Poisson case. The AGHQ estimator was positively biased for γ_1 in the binary case, though less so than the HG(1,2) estimators, but had no apparent bias in the Poisson case.

For the binary case, the Bayesian estimator for γ_1 (using an IG(0.001,0.001) prior)

was also negatively biased. Increasing the number of samples of the Gibbs sampler to 20,000 (with 2,000 sample burnin) had little effect on the average estimates. Using an IG(0.1,0..1) prior instead resulted in some positive bias, although the median of this posterior distribution did not appear to be biased. These biases were somewhat surprising, and may suggest that the influence of the choice of prior is greater when the variance parameter is relatively small. The use of a HC(3) or HC(30) prior also resulted in positive biases – for instance, the average estimates for γ_1 using the HC(3) prior were 1.178±0.058 for the posterior mean and 0.990±0.055 for the posterior median, with even higher average estimates using a HC(30) prior. An interesting aspect of the results in this study is that, despite the large negative bias of the PQL estimator of γ_1 for the binary case study, it nevertheless had a lower MSE than the AGHQ estimator of γ_1, which can be readily seen by comparing the boxplots of the 200 estimates for each estimator presented in Figure 5.3. Other estimators for γ_1 in this study, such as the HG(1,2) estimators, also have a higher MSE than PQL. The boxplots in Figure 5.3 also shows some skewness in the distribution of the AGHQ estimates, perhaps explaining why the AGHQ estimator is positively biased. Callens & Croux (2005) also showed that PQL performs better than AGHQ, with respect to the MSE, across a range of parameter values in their simulation studies.

	Binary		Poisson	
	$\hat{\gamma}_1$ (0.80)	$\hat{\tau}_0$ (0)	$\hat{\gamma}_1$ (0.80)	$\hat{\tau}_0$ (0)
PQL	0.354 ± 0.016	0.025 ± 0.011	0.671 ± 0.010	0.165 ± 0.008
HG(1,1)	0.526 ± 0.031	0.028 ± 0.012	0.796 ± 0.013	0.005 ± 0.009
HG(1,1) (no REML)	0.469 ± 0.028	0.027 ± 0.012	0.780 ± 0.013	0.009 ± 0.009
Bayes IG(.001)	0.659 ± 0.028	0.027 ± 0.012	0.843 ± 0.014	0.000 ± 0.009
Bayes IG(0.1) (mean)	1.032 ± 0.050	0.028 ± 0.013	0.844 ± 0.014	0.000 ± 0.009
Bayes IG(0.1) (median)	0.844 ± 0.046	0.028 ± 0.012	0.818 ± 0.014	0.004 ± 0.009
HG(1,2)	1.184 ± 0.067	0.030 ± 0.013	0.822 ± 0.014	-0.002 ± 0.009
HG(1,2) (no REML)	1.045 ± 0.061	0.030 ± 0.013	0.805 ± 0.014	0.003 ± 0.009
AGHQ	0.865 ± 0.046	0.028 ± 0.013	0.799 ± 0.013	0.008 ± 0.009

Table 5.3: Average estimates (±SE) for 200 simulated datasets from (5.2) using PQL, Bayesian, AGHQ and HGLM approaches.

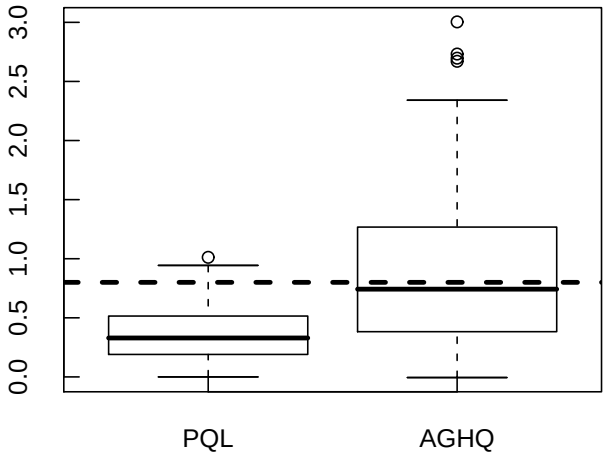


Figure 5.3: Box plots of the estimates of γ_1 for PQL and AGHQ for the binary logit model in (5.2). The true value ($\gamma_1 = 0.80$) is shown as a dotted line.

5.2.4 The Rodriguez-Goldman datasets

Rodriguez & Goldman (2001) analyse two datasets, from the 1987 National Survey of Maternal Health in Guatemala, to illustrate the use of a variety of GLMM approaches. Both datasets have a binary response with a multi-level structure; that is, data was collected on children within families within communities, with a very low average number of children per family (<2) in each. The binary response in the first dataset recorded whether or not complete immunisation had been performed for the child, and for the second dataset recorded whether or not modern prenatal care had been used for the child. The first dataset consisted of records for each of 2159 children within 1595 families within 161 communities, with the second dataset similarly having 2449 children within 1558 families within 160 communities. There were 15 and 21 covariates for the first and second datasets respectively. The list of covariates and the estimates for different GLMM approaches are shown in Tables 2 and 3 of Rodriguez & Goldman (2001).

For each dataset, the model for the conditional mean $\mu_{ijk} = E(y_{ijk}|u_i, v_{ij})$ for the kth child within the jth family within the ith community was

$$\text{logit}(\mu_{ijk}) = \boldsymbol{x}_{ijk}^T\boldsymbol{\beta} + u_{1i} + u_{2ij}, \tag{5.3}$$

where the vector $\boldsymbol{x}_{ijk}$ represents the set of covariates for the child, with associated coefficients $\boldsymbol{\beta}$, and $u_{1i} \sim N(0, \gamma_C)$ and $u_{2ij} \sim N(0, \gamma_F)$ were family and community effects respectively. An analysis of these datasets was conducted using the HGLM approaches, as well as verifying the PQL, GHQ (maximum likelihood) and Bayesian (Gibbs) estimates that Rodriguez & Goldman (2001) present, using `ASReml`, `AML` and `BUGS` respectively.

Results for the estimation of the variance components γ_F and γ_C for the two datasets are shown in Table 5.4. Firstly, note that Rodriguez & Goldman (2001) presented estimates of the standard deviations, $\sqrt{\gamma_F}$ and $\sqrt{\gamma_C}$, whereas here estimates of the variance components, γ_C and γ_F, are shown.

The PQL estimates generated from `ASReml` are equivalent to those in Rodriguez & Goldman (2001), and took less than 1 second to produce for each model. The GHQ results for the first dataset were reproduced exactly using `AML` with only 10 quadrature points, whereas 20 points were required according to Rodriguez & Goldman (2001). However, 20 quadrature points were required for the second dataset, as Rodriguez & Goldman (2001) noted. The settings used for the `BUGS` analysis were as described in Rodriguez & Goldman (2001), with an IG(.1,.1) prior used for both γ_C and γ_F, and a total of 5,000 iterations of the Gibbs sampler after a 200 iteration burnin. Note that the differences between the GHQ and Bayesian estimators reflect the lack of any implicit REML-type correction in the former.

Since the datasets are moderately large, it is interesting to make some comparisons in the time taken for each approach, also shown in Table 5.4. All timings shown here were for Pentium II computer. The PQL approach in `ASReml` took less than 1 second for each model. The GHQ approach were relatively fast as well, taking 13 seconds for the first dataset. For the Bayesian approach, the time taken to fit the first dataset was relatively modest (7 minutes), compared to the time quoted in Rodriguez & Goldman (2001) (5 hours), suggesting that their criticisms regarding the slowness of the Bayesian approach are somewhat outdated with advances in computational speed. (For the second dataset, the time taken (12.5 minutes) is also for 5200 iterations,

since increasing the number of iterations to 11000 resulted in negligible change to the estimates.)

The HGLM estimators performed relatively poorly for these datasets compared to Bayesian or GHQ approaches. For the first dataset, both the HG(0,1) and HG(1,1) estimators of γ_F and γ_C appear to have significant negative bias (>50%) compared with the Bayesian or GHQ approaches. This is probably to be expected, given the large negative biases of the first order HGLM approaches for the variance parameters in the binary nested two-way classification model (section 4.2.2.2) when the lower-level group size (children per family) was small . For the second dataset, the HG(1,1) estimator for γ_F was strongly positively biased, even after removing the REML correction. This is another case where the HG(1,1) estimator "diverges" to a very large estimate, similar to the divergence seen for the one-way classification study (4.5) where $\tau_0 = 2$. In addition, the HG(1,1) approach, as implemented in Fortran, was relatively slow, possibly due to the two-stage numerical optimisation required, in addition to a relatively large number of fixed effects being estimated. The second order HGLM approaches fared no better, with none of these approaches achieving a finite maximum to the likelihood for either dataset – that is, the second order approximated likelihood $p^s_\beta(h)$ increased monotonically with both γ_F and γ_C. One possible explanation for this lack of a solution may be as follows. Based on the magnitudes of the Bayesian and GHQ estimates, the "true" value of γ_F appears to be much larger than γ_C for both datasets. In the nested two-way nested classification study of the previous chapter (section 4.3.3.2), divergence of the second order HGLM estimators was also apparent when the lower level variance component, γ_2, was larger than the higher level variance component, γ_1. Therefore, the second order approaches may be more prone to instability when the variance component at the lower level, γ_F in this case, is relatively large compared to the higher level variance component, γ_c, as appears to be the case for the both these datasets here. The profile likelihoods for $\gamma_F \times \gamma_C$ for the first data for the GHQ and the non-REML HG(1,2) approaches are provided in Figure 5.4. It shows that the profile likelihood for non-REML HG(1,2) bears some resemblance to the "true" GHQ profile likelihood, with an apparent local

minimum at the AML estimate (marked "X").

In summary, the results are quite disappointing for both the first and second order HGLM approaches, and the GHQ or Bayesian approaches both appeared to be adequate with respect to computational speed.

	Immunisation			Pre-natal care		
	γ_F	γ_C	Run-time	γ_F	γ_C	Run-time
PQL (ASReml)	0.584	0.335	<1 sec	1.640	0.787	1 sec
PQL2	3.063	0.706	-	7.563	2.924	-
PQLB (Kuk's)	7.236	1.124	-	44.35	12.11	-
GHQ/ML (AML)	5.427	1.056	13 secs	53.85	13.76	77 secs
Bayesian (mean)	6.926	1.316	7 mins	104.7	28.68	12.5 mins
Bayesian (median)	6.587	1.256	...	101.8	27.45	...
HG(0,1)	1.27	0.57	17 secs	14.81	5.620	34 secs
HG(1,1)	1.62	0.61	3 mins	331.3	5.892	2 hr 20m
HG(1,1) (noREML)				214.0	5.109	...
HG(1,2)	NA – no maximum			NA – no maximum		

Table 5.4: Estimates of variance parameters for the Rodriguez & Goldman (2001) datasets (5.3) using PQL, Bayesian, GHQ and HGLM approaches.

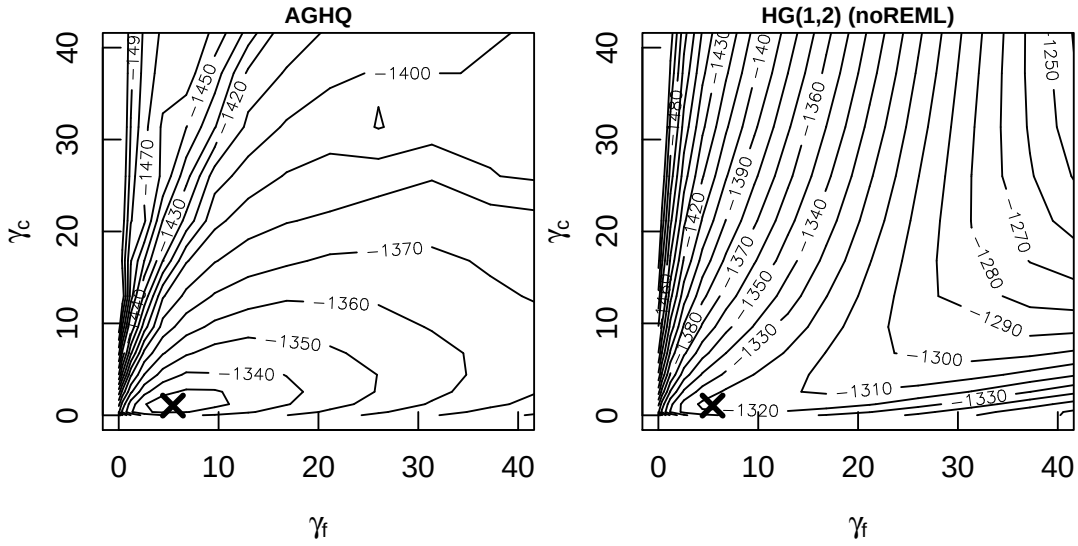


Figure 5.4: Profile likelihoods for $\gamma_F \times \gamma_G$ from the GHQ and non-REML HG(1,2) approaches for the first Rodriguez-Goldman dataset (5.3). The GHQ estimate is shown as a cross in both. (Profile likelihoods were generated on a grid $\gamma_F \times \gamma_C = (0.1, 0.6, \ldots, 6.6)^2$ using the R `contour` function with default settings.)

5.2.5 Simulation study using a "typical" RCBD

This simulation study used a randomised complete block design (RCBD), with settings that may be similar to those in some agricultural experiments.

Data Y_{ij}, $i = 1, \ldots, b$, $j = 1, \ldots, p$ were simulated with conditional mean $\mu_{ij} = E(y_{ij}|u_i)$ where

$$g(\mu_{ijk}) = \tau_{t(i,j)} + u_i, \tag{5.4}$$

The parameters $\tau_t = \sigma_t \Phi^{-1}\left((t - 0.5)/p\right)$, $t = 1, \ldots, p$, represented the "treatment" effects, and the $u_i \sim N(0, \gamma_1)$ represented "block" effects. The data y_{ij} were generated as either binary, Poisson or Binomial data (with a denominator of 4), with the link functions $g(\cdot)$ being logit, logarithmic and logit link respectively. It was decided to use $b = 10$ and $p = 5$, the number of "blocks" and "treatments" respectively, since these might be considered somewhat typical of a small agricultural experiment. The parameters σ_t^2 and γ_1 were both arbitrarily set to 1. A total of 200 simulated datasets were generated and analysed according to (5.4). In the analysis of each simulated dataset, the "treatment" parameters τ_i were estimated as fixed parameters, whereas the block parameters u_i were estimated as random effects.

Note that, initially, a simulation study involving a split-plot design was planned, to give a more challenging comparison between the different approaches. The split-plot design is the equivalent of a nested two-way classification design (sections 3.1.4.2 and 4.2.2.2), but with additional fixed (treatment) effects. Such a design would have required the use of a standard GHQ package such as `AML`, as `NLMIXED` cannot analyse GLMMs with more than one level of random classification. Before proceeding with such a simulation study, a few preliminary datasets were generated using a sample split-plot design and analysed using the same model in `AML`. However, regardless of the settings and the starting values chosen, it was not possible to get AML to converge to a solution for any of these preliminary datasets. Note also that, even for simulated datasets using the simpler RCBD design (5.4) above, convergence for GHQ was only achievable if the ASReml estimates were used as starting values, and even

then convergence was not always guaranteed. However, the AGHQ program `NLMIXED` could be used for the RCBD design instead, and it was found that it always converged for these datasets without the need of specifying good starting values. This exercise highlighted the instability of the standard GHQ approach for some GLMMs, and the advantage of using adaptive GHQ, as was already discussed in section 5.1.1.1.

Returning to the RCBD design (5.4), mean estimates for τ_1, τ_3, τ_5 and γ_1 for a range of approaches are shown in Table 5.5. For the Poisson case, estimates for 3 of the 200 simulations did not converge for most approaches, and so were excluded. For the binary case, both PQL and AGHQ were positively biased, with, surprisingly, larger positive bias for AGHQ. For the Poisson and binomial cases, the estimation biases for PQL and AGHQ were much smaller in magnitude, with some negative bias apparent for both estimators in the Poisson case and no apparent bias for either in the binomial case. For both Poisson and binomial cases, the average PQL and AGHQ estimates were similar. Plotting the individual estimates from PQL against the AGHQ estimates for the 200 simulated datasets (not shown) confirmed a very high level of concordance between the two estimators. The smaller biases observed here for PQL are in stark contrast to the large estimation biases observed in the simulation studies of Chapter 3.

For the binomial and Poisson cases, the average estimates from Bayesian and HGLM approaches are also presented. For the Bayesian approach, 5,000 samples of the Gibbs sampler were taken, with a 1,000 sample burn-in (and an IG(0.001,0.001) prior for γ_1, as in previous studies). It was decided to only present the median of the posterior distributions in this case, since the posterior means displayed a greater degree of positive bias. For the binomial case, both the Bayesian (median) and the HGLM approaches gave positively biased estimators, similar to the paired binary example of section 5.2.2. The use of other priors only exacerbated the positive bias for the Bayesian estimator: for instance, using a HC(3) prior gave an average posterior median of 1.548 ± 0.082 for γ_1. However, the non-REML versions of the HGLM approaches provided estimators with no apparent bias for the binomial case, and similar to PQL and AGHQ. For the

Poisson case, however, both the Bayesian (median) and standard HGLM approaches had no apparent negative bias, unlike the PQL and AGHQ estimators.

The relatively good performance of PQL here can probably be attributed to the relatively small number of groups (blocks), and perhaps also to having a greater number of fixed effects compared to the simulation studes in chapter 3, where large negative estimation biases were observed. As seen in section 3.1.4.4, the negative biases of PQL for binary GLMMs can be offset in designs with many fixed effects. The reason why the inclusion of many fixed effects offsets the negative bias, and sometimes leads to positive bias, may be related to the inconsistency of ML estimation in the binary matched-paired problems where pair effects are fitted as fixed effects, as shown in Andersen, 1973. Like the HGLM approaches, an implicit REML-like correction is also incorporated in the PQL estimate. A non-REML version of the PQL approach can also be used, as discussed in section 5.1.2.1, and also shown in Table 5.5 for the Poisson and binomial cases. Comparison of the average PQL and non-REML PQL estimates in Table 5.5 shows the large impact of the REML-like correction for PQL in this case.

5.2.6 The Salamander dataset

The "salamander dataset" was originally presented in McCullagh & Nelder (1989) and has become a standard dataset to test GLMM approaches. It has been presented and analysed in, amongst many others, Drum & McCullagh (1993), Schall (1991), Breslow & Clayton (1993), Karim & Zeger (1992), Shun (1997) and Noh & Lee (2007). It has attracted interest due to the crossed design which renders numerical integration approaches such as GHQ impractical. In summary, the dataset consists of repeated matings of female and male salamanders from two populations, Roughbutt (R) and Whiteside (W). The dataset consists of three experiments, with one conducted in the summer and the other two in following autumn. In each experiment, 10 salamanders from each population and sex were used. The experimental design was such that each salamander was mated with six salamanders of the opposite sex, being three

	$\hat{\tau}_1$(-1.28)	$\hat{\tau}_3$(0)	$\hat{\tau}_5$ (1.28)	$\hat{\gamma}_1$ (1)
		Binary		
PQL	-1.571 ± 0.132	-0.116 ± 0.061	2.205 ± 0.203	1.486 ± 0.112
AGHQ	-1.646 ± 0.193	1.453 ± 0.202	4.341 ± 0.367	2.358 ± 0.294
		Poisson		
PQL	-1.370 ± 0.042	0.055 ± 0.026	1.352 ± 0.022	0.946 ± 0.041
AGHQ	-1.433 ± 0.042	-0.008 ± 0.027	1.289 ± 0.023	0.931 ± 0.043
Bayesian (median)	-1.575 ± 0.048	-0.042 ± 0.028	1.281 ± 0.024	1.162 ± 0.056
HG(1,1)	-1.450 ± 0.042	-0.025 ± 0.027	1.272 ± 0.024	1.068 ± 0.051
HG(1,2)	-1.452 ± 0.042	-0.027 ± 0.027	1.270 ± 0.024	1.086 ± 0.052
HG(1,1) (nonREML)	-1.433 ± 0.042	-0.008 ± 0.027	1.289 ± 0.023	0.921 ± 0.043
HG(1,2) (nonREML)	-1.435 ± 0.042	-0.009 ± 0.027	1.288 ± 0.023	0.934 ± 0.044
PQL (nonREML)	-1.358 ± 0.042	0.068 ± 0.026	1.365 ± 0.022	0.823 ± 0.036
		Binomial (denom=4)		
PQL	-1.259 ± 0.038	-0.011 ± 0.031	1.354 ± 0.034	1.084 ± 0.054
AGHQ	-1.304 ± 0.039	-0.011 ± 0.032	1.405 ± 0.035	1.047 ± 0.060
Bayesian (median)	-1.334 ± 0.041	-0.004 ± 0.033	1.451 ± 0.037	1.350 ± 0.081
HG(1,1)	-1.321 ± 0.040	-0.010 ± 0.033	1.423 ± 0.036	1.303 ± 0.072
HG(1,2)	-1.325 ± 0.040	-0.010 ± 0.033	1.428 ± 0.036	1.376 ± 0.076
HG(1,1) (nonREML)	-1.303 ± 0.039	-0.011 ± 0.032	1.403 ± 0.035	1.048 ± 0.057
HG(1,2) (nonREML)	-1.308 ± 0.040	-0.011 ± 0.032	1.408 ± 0.035	1.107 ± 0.060
PQL (nonREML)	-1.248 ± 0.037	-0.011 ± 0.030	1.342 ± 0.034	0.924 ± 0.047

Table 5.5: Average estimates of $\hat{\tau}_1$, $\hat{\tau}_3$, $\hat{\tau}_5$ and γ_1 for 200 simulated datasets from an RCBD design (5.4) using PQL, Bayesian, GHQ and HGLM approaches.

salamanders from each population. Each experiment comprised 120 observations, and so there are a total of 360 observations in total. Further details on the design can be obtained in McCullagh & Nelder (1989). The primary objective of the study was to determine whether there were differences in mating success between the four population by sex combinations, allowing for differences between individual salamanders in their mating success. Here, we use the dataset simply to demonstrate similarities and differences between the estimates from Bayesian and HGLM approaches.

We follow Noh & Lee (2007) in presenting the results of both the summer (120 observations, 40 salamanders) and the pooled salamander dataset (360 observations, 120 salamanders). The model for the (conditional) probability of mating success μ_{ijk} between the ith female and jth male in experiment k is

$$\text{logit}\,(\mu_{ijk}) = \tau_0 + \tau_1 x_{Fik} + \tau_2 x_{Mjk} + \tau_3 x_{Cijk} + u_{Fik} + u_{Mjk}, \tag{5.5}$$

where $x_{Fik} = I(i\text{th female of the } k\text{th experiment is Whitehead})$, $x_{Fik} = I(j\text{th male of the } k\text{th experiment is Whitehead})$, $x_{Cijk} = x_{Fik} \times x_{Mjk}$ and $I(\cdot)$ is the indicator function. The $u_{Fik} \sim N(0, \gamma_F)$ and $u_{Mjk} \sim N(0, \gamma_M)$ are random effects pertaining to the ith female and jth male in the kth experiment respectively.

Estimates from PQL, Bayesian and HGLM approaches are presented in Table 5.6. For comparison, the MCEM estimates and estimates from Drum & McCullagh (1993) (D&M) are taken from Noh & Lee (2007) for comparison. (Note that Noh & Lee (2007) present the variance parameters as standard deviations, that is, $\sqrt{\gamma_F}$ and $\sqrt{\gamma_M}$.) The MCEM estimates can be considered maximum likelihood (ML) estimates, in place of GHQ estimates which cannot be obtained here due to the crossed design. To demonstrate the influence of the prior distribution, four prior distributions have been used, the Inverse Gamma priors IG(.001,0.001) and IG(0.1,0.1), and the Half Cauchy priors HC(3) and HC(30). A total of 10,000 samples of the Gibbs sampler were taken for each, with a 1,000 sample burn-in. Posterior medians are presented in all cases. For both datasets, the PQL estimates of all parameters are consistently lower than the rest. There is some variation in the Bayesian estimates between the different priors, in particular there are higher estimates of all parameters for the Half-Cauchy priors than Inverse Gamma priors. This is more evident for the summer dataset, which is to be expected, since the choice of prior would have more influence when there are fewer datapoints. For both datasets, the HG(1,2) estimates are similar to the Bayesian estimates, and the HG(1,2) "non-REML" estimates are similar to the MCEM estimates (the latter which are ML, not REML estimates, as already indicated). The concordance of the HG(1,2) estimates with Bayesian and ML estimates is similar to the concordance noted in the paired binary study in section 5.2.2.

Finally, it should also be noted that the HGLM estimates presented in Table 5.6 are different from the ones presented in Table 2 of Noh & Lee (2007), which can be obtained from the (current) Genstat HGLM implementation. The latter estimates are also shown in Table 5.6, marked (Noh/Lee). The difference between the estimates

from the Fortran 90 implementation, which is used here, to the current Genstat implementation has already been discussed in section 4.2.1.

	$\hat{\tau}_0$	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}_3$	$\hat{\gamma}_F$	$\hat{\gamma}_M$
			Summer			
PQL	1.16	-2.57	-0.38	2.81	1.42	0.09
Bayes IG(.001)	1.40	-3.15	-0.44	3.51	2.40	0.052
Bayes IG(.1)	1.56	-3.44	-0.51	3.74	2.50	0.049
Bayes HC(3)	1.61	-3.63	-0.56	3.92	2.93	0.058
Bayes HC(30)	1.62	-3.67	-0.57	4.07	3.23	0.076
HG(1,1)	1.48	-3.26	-0.48	3.53	2.31	0.30
HG(1,2)	1.57	-3.45	-0.53	3.74	2.71	0.52
HG(1,1) (nonREML)	1.34	-2.94	-0.42	3.18	1.57	0.07
HG(1,2) (nonREML)	1.39	-3.07	-0.45	3.32	1.78	0.19
MCEM *	1.38	-3.04	-0.45	3.29	1.74	0.23
D&M *	1.42	-3.08	-0.47	3.30	1.69	0.34
HG(1,1) (Noh/Lee)*	1.45	-3.19	-0.48	3.48	2.22	0.25
HG(1,2) (Noh/Lee)*	1.39	-3.05	-0.42	3.25	1.77	0.22
			Pooled			
PQL	0.79	-2.29	-0.54	2.82	0.72	0.63
Bayes IG(.001)	1.05	-3.05	-0.073	3.75	1.55	1.38
Bayes IG(.1)	1.04	-3.03	-0.071	3.75	1.53	1.37
Bayes HC(3)	1.06	-3.10	-0.072	3.80	1.69	1.51
Bayes HC(30)	1.04	-3.08	-0.072	3.81	1.73	1.53
HG(1,1)	1.05	-3.01	-0.73	3.72	1.39	1.22
HG(1,2)	1.09	-3.15	-0.77	3.89	1.66	1.49
HG(1,1) (nonREML)	1.01	-2.90	-0.70	3.59	1.17	1.04
HG(1,2) (nonREML)	1.05	-3.03	-0.74	3.75	1.41	1.27
MCEM *	1.02	-2.96	-0.70	3.63	1.39	1.23
D&M *	1.06	-3.05	-0.72	3.77	1.66	1.49
HG(1,1) (Noh/Lee)*	1.04	-2.98	-0.74	3.71	1.37	1.21
HG(1,2) (Noh/Lee)*	1.02	-2.97	-0.72	3.66	1.39	1.21

Table 5.6: Estimates for the summer and pooled salamander datasets from model (5.5) using PQL, Bayesian and HGLM approaches. The MCEM, D&M, HG(1,1) (Noh/Lee) and HG(1,2) (Noh/Lee) estimates (starred) are taken from Noh & Lee (2007).

5.3 A simulation study using spatially correlated errors

The modelling of spatially correlated data is an important application of GLMMs. Several papers endorse the use of PQL for fitting GLMMs to spatially correlated

data (e.g. Kneib & Fahrmeir (2004); Paciorek (2007); Ainsworth & Dean (2006)), especially in comparison to Bayesian approaches. The simulation study presented here compared a Bayesian implementation in the R function `geoRglm` (Christensen & Ribeiro Jr., 2002) against PQL, as implemented in `ASReml`, for data generated using a Matérn correlation function, as suggested by Stein (1999).

5.3.1 Methods

Data $Y_i \sim \text{Poisson}(\mu_i)$, $i = 1 \ldots 400$, was generated on a 20×20 grid where

$$\log(\mu_i) = \eta_i = \tau_0 + S(\boldsymbol{\ell}_i), \tag{5.6}$$

and $\boldsymbol{\ell}_i$ denoted a two-dimensional location . The $S(\boldsymbol{\ell}_i)$ represented a Gaussian process with mean 0 and covariance function

$$\text{cov}\left(S(\boldsymbol{\ell}_i), S(\boldsymbol{\ell}_j)\right) = \gamma_1 \rho(d_{ij}) = \gamma_1 \rho(d_{ij}),$$

where $d_{ij} = \parallel \boldsymbol{\ell}_i - \boldsymbol{\ell}_j \parallel$ is the Euclidean distance between locations $\boldsymbol{\ell}_i$ and $\boldsymbol{\ell}_j$. The function $\rho(\cdot)$ was the Matérn correlation function with range γ_ϕ and smoothness γ_ν (also referred to as κ). When $\gamma_\nu = 1.5$, the Matérn correlation function is

$$\rho(d) = \left(1 + \frac{d}{\gamma_\phi}\right) e^{-d/\gamma_\phi},$$

which is shown in Figure 5.5b where $\gamma_\phi = 2$.

The parameter settings chosen in this study used were $\tau_0 = 0$, $\gamma_1 = 1$, γ_ϕ=2 and $\gamma_\nu = 1.5$. A total of 200 simulated datasets were generated from model (5.6). Each dataset was analysed according to the same model which generated the data, using PQL as implemented in `ASReml` (version 1.63) and `geoRglm` (version 0.8-11). Since the current version of `geoRglm` could not estimate the smoothness parameter γ_ν, it was fixed at $\gamma_\nu = 1.5$ in both the `geoRglm` and `ASReml` analyses. The fixing of the smoothness parameter also removed the need to have some of the design points $\boldsymbol{\ell}_i$ at

closer proximities, which would have been required if the smoothness parameter was to be estimated (Stein, 1999).

As well as comparing the two methods in the estimation of the parameters γ_ϕ, γ_1 and τ_0, it was also deemed important to measure how well each method predicted the spatial trend and the coverage of 95% confidence intervals for the spatial predictions. An additional 25 points $\ell_{401}, \ldots, \ell_{425}$ were generated at positions $(0.5, 4.5, 8.5, 12.5, 16.5) \times (0.5, 4.5, 8.5, 12.5, 16.5)$ as shown in Figure 5.5a. Let $\boldsymbol{S} = (S(\boldsymbol{\ell}_1), \ldots, S(\boldsymbol{\ell}_{400}))^T$, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_{400})^T$, $\boldsymbol{S}^* = (S(\boldsymbol{\ell}_{401}), \ldots, S(\boldsymbol{\ell}_{425}))^T$ and $\boldsymbol{\eta}^* = (\eta_{401}, \ldots, \eta_{425})^T$. The "delta" method of Ainsworth & Dean (2006) was used for calculating the standard errors and confidence intervals of the predicted trend $\boldsymbol{\eta}^*$, and is described in Appendix A.3.

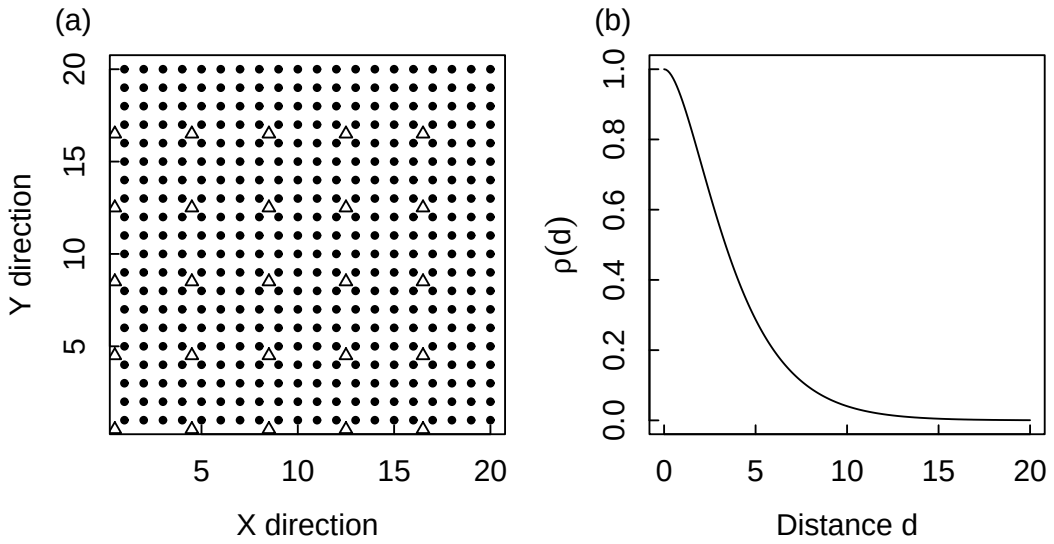


Figure 5.5: (a) The 20×20 grid of sampled locations (filled dots) and the 25 locations to predict at (triangles) for the spatial case study (5.6) (b) the Matérn correlation function with $\gamma_\phi = 2$ and $\gamma_\nu = 1.5$.

5.3.1.1 Implementation of Bayesian approach using geoRGLM

The settings used for the R package `geoRglm` were chosen according to the recommended settings specified in the `geoRGLM` documentation. Letting $\pi(\cdot)$ denote the prior PDF for the relevant parameter, these settings were:

- a uniform prior for τ_0. (Note that this is different from the standard recommendation for location parameters, which is a normal prior with zero mean and very large variance.)

- a squared reciprocal prior for γ_ϕ, $\pi(\gamma_\phi) = 1/\gamma_\phi^2$. This was also recommended in Diggle *et al.* (2002).

- a scaled inverse chisquared prior for γ_1,

$$\pi(\gamma_1; \nu, \sigma^2) = \frac{\left(0.5\sigma^2\nu\right)^{v/2}}{\Gamma(\nu/2)} \frac{\exp\left(-\nu\sigma^2/2\gamma_1\right)}{\gamma_1^{1+\nu/2}},$$

  where ν and σ^2 are the "degrees of freedom" and "scale" parameters respectively. Settings of $\nu = \sigma^2 = 1$ were selected to obtain an uninformative prior.

- A burn-in of 10,000 samples was used. After the burn-in, another 100,000 samples were generated, but only every 1,000th sample was retained to estimate the posterior distributions.

The `geoRGLM` package implemented the alternate use of random walk sampling of the conditional distribution of γ_ϕ and Metropolis-Hastings sampling of the conditional distribution of γ_1. Sampling of $\boldsymbol{S}$ and τ_0 was performed in a separate later stage. The `geoRglm` documentation recommended acceptance rates are $\sim$23% for γ_ϕ and $\sim$60% for γ_1. Preliminary analysis on a few simulated datasets was used to select initial proposal variances for γ_ϕ and γ_1 to achieve these recommended acceptance rates, which were then used for all simulated datasets. However, it was found that the acceptance rates still varied considerably between simulations, and so the proposal variances were set separately for each simulated dataset, based on the acceptance rate for the dataset obtained in the first run. For most simulated datasets, the final acceptance rates were not too far away from the recommended rates (γ_ϕ: 0.253, SE 0.035; γ_1: 0.656, SE 0.076). Mixing was relatively poor for some simulations, but this appeared to be unrelated to the achieved acceptance rate.

Estimates and 95% confidence intervals for $\hat{\mu}_i$, or Y_i, were obtained from the posterior distributions produced by `geoRglm`. The median, rather than the mean, of the

posterior distribution was used as a point estimate, $\hat{\mu}_i$.

It should be noted that `geoRglm` was very computationally intensive. The analysis for each simulation took over 30 minutes to run, and required over 500MB RAM during processing on a Pentium class computer. In contrast, PQL using ASReml took only about 5 minutes per simulation, with 11MB RAM required during processing.

5.3.1.2 Estimation and prediction error

For both PQL and Bayesian approaches, two measures of error were determined, denoted here as "estimation error" and "prediction error". For each simulation, the "estimation error" measured the average log-proportional error of predicting the underlying trend μ_i^*, $i = 401 \ldots 425$, for each of the 25 additional points,

$$\text{Estimation error} = \sqrt{\frac{1}{25}\sum_{i=401}^{425}\left\{\log\left(\frac{\hat{\mu}_i^*}{\mu_i^*}\right)\right\}^2} = \sqrt{\frac{1}{25}\sum_{i=401}^{425}\left\{\log\left(\hat{\mu}_i^*\right) - \log\left(\mu_i^*\right)\right\}^2}.$$

A visual demonstration of this calculation for one simulation is shown in Figure 5.6.

The prediction error was defined similarly, but measured the error in predicting the generated data Y_i, $i = 401, \ldots, 425$, and only for non-zero Y_i,

$$\begin{aligned}\text{Prediction error} &= \sqrt{\frac{1}{\left(\sum_{i=401}^{425} N_i\right)}\sum_{i=401}^{425} N_i\left\{\log\left(\frac{Y_i}{\hat{\mu}_i^*}\right)\right\}^2} \\ &= \sqrt{\frac{1}{\left(\sum_{i=401}^{425} N_i\right)}\sum_{i=401}^{425} N_i\left\{\log\left(\hat{\mu}_i^*\right) - \log\left(Y_i\right)\right\}^2},\end{aligned}$$

where $N_i = I(Y_i > 0)$ and $I(\cdot)$ is the indicator function.

5.3.2 Results

Both methods gave similar average estimates of the parameters (Table 5.7). The use of PQL resulted in some underestimation of the spatial variance γ_1, whilst the use of `geoRglm` resulted in some underestimation of the Matérn range, γ_ϕ. The two methods

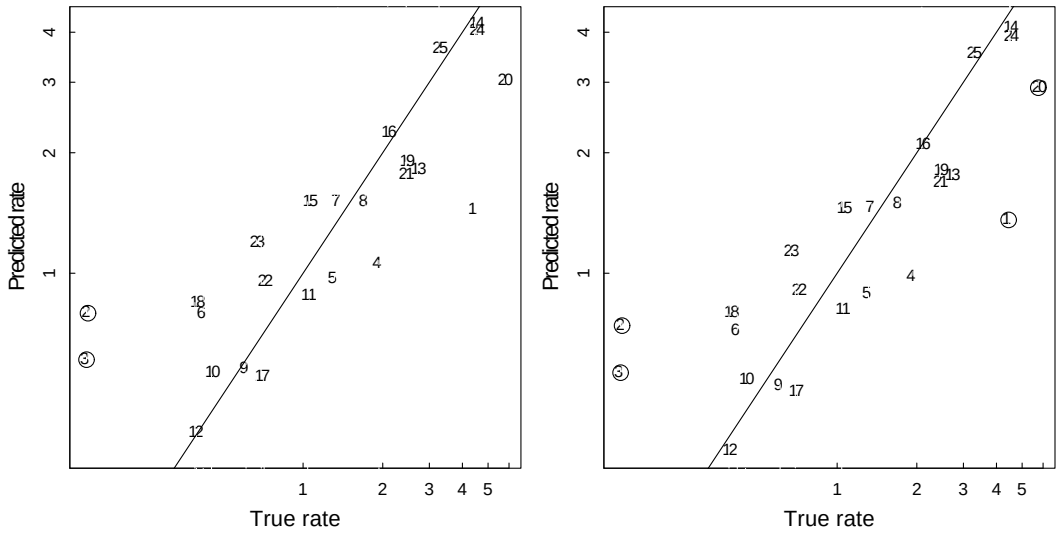


Figure 5.6: A scatterplot of the true rates, μ_i, against the predicted rates, $\hat{\mu}_{400+i}$, $i = 1, \ldots, 25$, for PQL (left plot) and Bayesian (right plot) approaches, for the 25 additional points in one simulated dataset. The log-log scale is used in both plots. A one-one line is also shown (solid). Observations where the true rate fell outside of its respective 95% confidence interval are circled.
(The PQL estimates for this dataset were $(\hat{\gamma}_\phi, \hat{\gamma}_1, \hat{\tau}_0)^T = (1.81, 0.59, 0.23)^T$, with 58.4% estimation error, and the Bayesian estimates were $(\hat{\gamma}_\phi, \hat{\gamma}_1, \hat{\tau}_0)^T = (1.82, 0.72, 0.15)^T$, with 58.7% estimation error. The achieved acceptance rates were 0.281 and 0.601 for γ_ϕ and γ_1 respectively.)

were very similar in both the average estimation and prediction errors. The similarity between the PQL and Bayesian predictions of $\hat{\mu}_i$ can be seen in Figure 5.6 for one simulated dataset. The coverage of the prediction intervals for PQL, however, tended to be conservative.

	γ_1 (1)	γ_ϕ (2)	Error(%)		Coverage	
	Est (RMSE)	Est (RMSE)	Est	Pred	Est	Pred
PQL (ASReml)	0.94 (0.064)	2.01 (0.037)	42.8	71	99	97
Bayesian (geoRGLM)	1.02 (0.036)	1.91 (0.095)	42.5	75	95	68

Table 5.7: Average estimates of parameters (RMSE in brackets), average estimation and prediction errors, and true coverage rates of the 95% confidence intervals of the rates μ_i and realized values y_i for PQL and Bayesian approaches for the spatial case study (5.6).

5.4 A "real-life" dataset with an ordinal response

A supplementary study is presented here, examining the performance of the PQL approach in the analysis of a "real-life" ordinal disease dataset with spatial correlation and the use of an XFA factor analytic variance structure (Thompson *et al.*, 2003).

One aim of this study was to examine PQL estimation biases associated with ordinal data for a "real-life" design, and compare them to the estimation biases observed for PQL in earlier studies. A supplementary objective was to examine the use of the (first order) Laplace approximation for hypothesis testing of variance parameters, in combination with PQL estimation, as a follow-up to the study in chapter 3 (section 3.2.1).

5.4.1 Description of the dataset

The dataset was kindly provided by Steven Harden, biometrician at NSW DPI, Tamworth. The data arise from a chickpea variety trial conducted using three standard varieties (Howzat, Jimbour and Tyson) and 179 new varieties. During the course of the trial phytophera root damage was visually assessed on four occasions for each plot. The objective is to rate the chickpea varieties in their susceptibility to phytophera disease.

The trial area consisted of a 24×40 grid of plots divided into 4 replicates of 6×40 plots as shown in Figure . The three standard varieties (Howzat, Jimbour and Tyson), were replicated 77, 80 and 141 times respectively, and the remaining 179 varieties were replicated between 2 and 4 times. Phytophera damage was visually assessed on a 9 point scale where 1=no damage, 3=0-10% dead, 5=20-40% dead, 7=60-80% dead, 9= all plants dead (2,4,6,8 were intermediary values). Some missing values were also present due to misadventure.

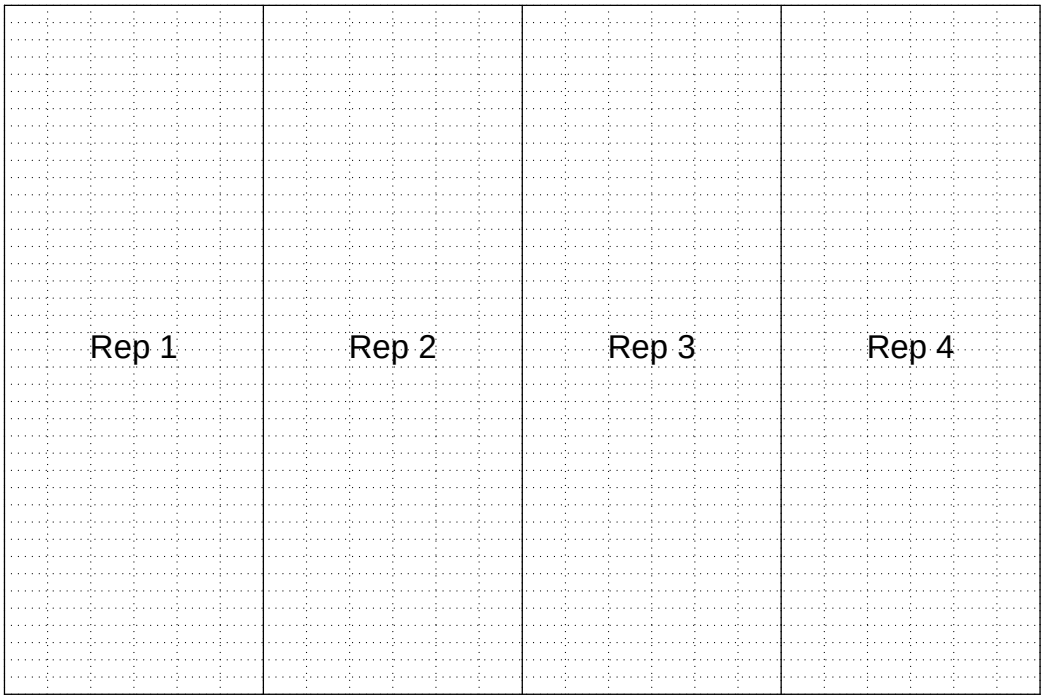


Figure 5.7: Design of the phytophera trial. A 24×40 grid of plots was divided into 4 replicates each of 6×40 plots as shown.

5.4.2 Analysis of the "real-life" dataset

5.4.2.1 Methods

For simplicity, an analysis of the first measurement in time is conducted.

Using the framework for accounting for natural spatial variation in field trials set out in Gilmour *et al.* (1997), a suitable model for a normally distributed response ψ_{ij} for the ith, jth plot would be

$$\psi_{ij} = \tau_0 + u_{Vm(i,j)} + u_{Rn(i,j)} + S\left(\boldsymbol{\ell}_{ij}\right) + u_{Oij} \tag{5.7}$$

where variety, replicate and unit effects are assumed to be random, that is, $u_{Vm} \sim N(0, \gamma_v)$, $m = 1, \ldots, 182$, $u_{Rn} \sim N(0, \gamma_r)$, $n = 1, \ldots, 4$ and $u_{Oij} \sim N(0, \gamma_u)$. The term $S\left(\boldsymbol{\ell}_{ij}\right)$ represents a Gaussian process at location $\boldsymbol{\ell}_{ij}$ having a separable auto-

regressive correlation structure in each direction, that is,

$$\text{cov}\left\{S\left(\boldsymbol{\ell}_{ij}\right), S\left(\boldsymbol{\ell}_{kl}\right)\right\} = \gamma_S \gamma_{\rho R}^{|j-l|} \gamma_{\rho C}^{|k-i|},$$

where γ_S is the spatial variance, and $\gamma_{\rho R}$ and $\gamma_{\rho C}$ are the spatial correlations along rows and columns respectively. The addition of u_{oij} allows for non-spatial variation, such as measurement error. A natural extension of this model to the ordinal case is to assume the ordinal response represents grouped normal data. That is, there is a latent normally distributed variable ψ_{ij} with expectation, without loss of generality,

$$E\left(\psi_{ij}|\tau_0, u_{Vm(i,j)}, u_{Rn(i,j)}, S\left(\boldsymbol{\ell}_{ij}\right), u_{Oij}\right) = u_{Vm(i,j)} + u_{Rn(i,j)} + S\left(\boldsymbol{\ell}_{ij}\right) + u_{Oij}$$

and a residual variance of 1. An ordinal response y_{ij} can been generated by categorising ψ_{ij} using cut-points $\tau_1, \tau_2, \ldots, \tau_8$, that is,

$$y_{ij} = 1 + \sum_k I(\psi_{ij} > \tau_k),$$

and so the ordinal response falls in one of nine ordered classes, $y_{ij} \in \{1, \ldots, 9\}$. If this generation model is assumed, an ordinal-probit model for y_{ij} is appropriate, where the model for the cumulative probabilities $\mu_{ijk} = P(y_{ij} \leq k|u_{vm(i,j)}, u_{rn(i,j)}, S(\ell_{ij}), u_{oij})$, $k = 1, \ldots, 8$, is

$$\Phi^{-1}\left(\mu_{ijk}\right) = \tau_k - u_{Vm(i,j)} - u_{Rn(i,j)} - S\left(\boldsymbol{\ell}_{ij}\right). \tag{5.8}$$

Note that the u_{oij} term has now been removed from this model, since it is completely confounded with the residual variation and cannot be estimated. In the absense of software to fit the ordinal mixed model (5.8), the equivalent normal linear mixed model (5.7) may be adequate, especially as the number of levels in the ordinal response increases. The fitting a normal linear mixed model to an ordinal response treats the ordinal classes {1,...,9} as numerical scores, or otherwise equidistant. As an interesting aside, it was decided to fit the normal linear mixed model (5.7), as well as the ordinal model (5.8), to the ordinal response to determine if there is any difference

in this particular case.

The ordinal model (5.8) above assumes the cutpoints τ_k are invariant with respect to the other factors in the model. This assumption could be relaxed, although the interpretation of the ordinal data as grouped normal data would no longer hold. One possible extension is to assume that the cutpoints vary between varieties in some systematic way, so extending 5.8 as

$$\Phi^{-1}(\mu_{ijk}) = \tau_k - u_{Rn(i,j)} - u_{Vkm(i,j)} - S(\boldsymbol{\ell}_{ij}), \tag{5.9}$$

where $u_{Vkm} \sim N(0, \gamma_{vt})$ are random cutpoints for each variety. To impose a systematic constraint on the u_{Vkm} effects, a first order factor-analytic structure was used (Thompson *et al.*, 2003), that is,

$$u_{Vkm} = \lambda_k u^*_{Vm},$$

where λ_k are the factor loadings, and represent variance components to be estimated, and $u^*_{vm} \sim N(0, 1)$. In order to determine whether there was a significant improvement in the likelihood between models (5.8) and (5.9), a likelihood ratio test (LRT) was conducted, using the first-order Laplace approximation of the likelihood calculated at the PQL estimates of each model. The formulae to determine the first order Laplace approximation for each model is given in Appendix A.4. The use of a first order Laplace approximation, in conjunction with PQL estimation for testing variance components, has already been shown to give conservative tests for a one-way classification model in section 3.2.1. The null distribution of the Laplace approximation will be determined in the simulation study below.

All models were fitted using the beta version of `ASReml` version 3, which can fit ordinal models. More information about the use of factor analytic structures to parsimoniously parametrize variance structures is given in Thompson *et al.* (2003).

5.4.2.2 Results

The results for fitting the normal (5.7) and ordinal (5.8) models are shown in table 5.8. The estimates of the variance components are not directly comparable, being on different scales. However, there is a reasonable level of consistency in the estimates of the spatial correlations $\gamma_{\rho R}$ and $\gamma_{\rho C}$ between the two models. In addition, the ratio of the variance component over its estimated standard error (shown in brackets, also produced in ASReml) is also reasonably consistent between both models for all parameters. The additional measurement error in the ordinal scoring process is effectively absorbed as the measurement error variance γ_u in the normal model. So it appears that the treatment of the ordinal data as equidistant scores, as assumed by fitting a normal linear mixed model (5.7) to the data, may not to be an unreasonable assumption in this example. However, in general, the use of a normal linear mixed model for fitting ordinal score data is a dangerous practise and should be not condoned, especially when the scores regularly take high or low values, since the model does not allow for non-linearity or variance heterogeneity.

The deviance for the ordinal model (5.8), $-2\sum_{i,j}\log(\hat{\mu}_{ijy_{ij}})$, is 2197.7 with 950 degrees of freedom, suggesting the possibility of some lack of fit. The estimates for the alternative ordinal model (5.9) are also shown in Table 5.8. The alternative ordinal model (5.9) gives similar estimates of the replicate and spatial variance parameters as the original ordinal model (5.8). The estimates of the factor loadings λ_k, $k = 1, \ldots, 8$, (Table 5.9) shown an almost linear decline with cutoff k. The deviance for (5.9) is 2110.07 , a reduction of 77.7 from (5.8). The calculation of the first order Laplace likelihood for each model, using the calculations given in Appendix A.4, also suggested the inclusion of the cutoff by variety interaction was highly significant, with a likelihood ratio test statistic of $2 \times (-1413.6 - (-1459.4)) = 91.6$.

5.4.3 Simulation study

A simulation study was conducted using the ordinal model (5.8) as the design, with true parameter settings being the PQL estimates in Table 5.8. The objectives of

	γ_R	γ_V	γ_S	$\gamma_{\rho R}$	$\gamma_{\rho C}$
Normal (5.7)	0.54 (1.08)	1.32 (7.51)	0.44 (3.90)	0.70 (6.51)	0.88 (19.5)
Ordinal (5.8)	0.23 (0.80)	2.19 (7.53)	0.58 (2.69)	0.84 (11.2)	0.94 (33.7)
Ordinal FA (5.9)	0.41 (0.96)	–	0.50 (2.94)	0.77 (8.13)	0.92 (29.6)

	γ_u
Normal (5.7)	1.01 (14.6)
Ordinal (5.8)	–
Ordinal FA (5.9)	–

Table 5.8: Estimates of variance components from fitting models (5.7), (5.8) and (5.9) to the phytophera dataset using PQL. The ratio of the estimate divided by its SE, also reported in `ASReml`, is shown in brackets.

λ_1	λ_2	λ_3	λ_4
2.38 (14.17)	1.58 (13.45)	1.15 (12.50)	0.86 (13.02)
λ_5	λ_6	λ_7	λ_8
0.57 (7.19)	0.45 (4.59)	0.14 (1.18)	-0.16 (-0.90)

Table 5.9: Estimates of the variance components λ_i, $i = 1, \ldots, 8$, associated with the factor-analytic ordinal model (5.9) to the phytophera data using PQL. The ratio of the estimate divided by its SE, also reported in `ASReml`, is shown in brackets.

the simulation study were two-fold. The first objective was to examine estimation biases associated with this ordinal model and design, and compare them to the PQL estimation biases in previous studies. The second objective was to determine the null distribution of the likelihood ratio test (LRT), generated using the first order Laplace approximation, for testing the alternative ordinal model (5.9) against the ordinal model (5.8) from which the data was generated. Model (5.8) can be viewed as a nested version of (5.9) where all the factor loadings are equal, $\lambda_k = \sqrt{\gamma_V}$, and, since there are no constraints on the λ_k in (5.9), the LRT here should approximate a χ^2_7 distribution.

A total of 500 datasets were generated from model (5.8) and analysed using PQL with the u_{oij} removed, since the estimate of γ_o was 0. For comparison of the estimation biases, simulated data $Y_{ij} \sim \text{Binomial}(m, \mu_{ij})$ was generated using an equivalent model for μ_{ij},

$$\text{logit}(\mu_{ij}) = v_{v(i,j)} + r_{r(i,j)} + S(\boldsymbol{\ell}_{ij}) \tag{5.10}$$

and where the denominator m was either 4 or 32. In contrast to previous simulation

studies reported in this thesis, PQL (as implemented currently in `ASReml`) did exhibit some dependence on the starting values for some simulated datasets, especially for the spatial correlation parameters, $\gamma_{\rho R}$ and $\gamma_{\rho C}$, which tended to hit the boundary at 1 whenever poor starting values were chosen, with the corresponding spatial variance γ_S going to infinity. To maximise the convergence rate, it was decided to use starting values of $\gamma_{\rho R} = \gamma_{\rho C} = 0.8$ for the spatial correlation parameters, $\lambda_i = 1.0$, $i = 1, \ldots, 8$ for the XFA parameters, and $\gamma_V = \gamma_R = 0.1$ for the other two variance parameters. In addition, to also improve convergence, a "`!step 0.01`" qualifier was used in ASReml to reduce the update step sizes for the first few iterations. For the ordinal simulations, convergence was determined on the basis of change of the parameter values from the previous iteration, as reported by ASReml – only 432 of the 500 simulations where the combined change of parameter values was <10% were retained.

5.4.3.1 Estimation biases

Table 5.10 shows the average estimates of the variance components for the ordinal and binomial models. There is some estimation biases for the ordinal model, but less than for the binomial model with a denominator of $m = 4$. Estimation biases for γ_R and γ_V for all three models were negative, which is as expected from the results of previous simulation studies in this thesis. In contrast, estimation biases for the components associated with the spatial trend, γ_S, $\gamma_{\rho R}$ and $\gamma_{\rho C}$, tended to be positively biased. The estimates for the spatial variance γ_S are somewhat positively biased for the ordinal models due to the convergence problem with spatial correlations $\gamma_{\rho R}$ and $\gamma_{\rho C}$ noted above (estimates for the binary $m = 4$ model are also affected, resulting in apparently lower bias for the spatial parameters than for the $m = 32$ model). This positive bias thus resulted from a problem with the default modified Fisher scoring algorithm implemented in ASReml, rather than a problem with PQL – possibly, the use of EM updates of the spatial parameters would have removed this convergence problem, but this was not tested.

	γ_R (0.23)	γ_V (2.19)	γ_S (0.58)	$\gamma_{\rho R}$ (0.84)	$\gamma_{\rho C}$ (0.94)
Ordinal	0.203±0.011	1.951±0.015	1.021±0.161	0.839±0.005	0.935±0.003
Bin (4)	0.189±0.009	1.525±0.008	0.589±0.052	0.849±0.003	0.942±0.001
Bin (32)	0.223±0.010	1.934±0.009	0.555±0.006	0.834±0.001	0.934±0.001

Table 5.10: Average estimates of variance components from 500 simulations from the ordinal model (5.8) and the corresponding binomial model (5.10) with denominator $m = 4$ and 32 respectively.

5.4.3.2 Use of Laplace approximation with PQL estimation for an approximate LRT

As noted above, the second aim of the simulation study was to determine the null distribution of the (first order) Laplace approximation using the PQL estimates, for testing the alternative ordinal model (5.9) against the generative ordinal model (5.8). The details of the calculation of the Laplace approximation for each model is given in the Appendix A.4.

Using the 431 simulations which "converged", the distribution of the LRT was very similar to a χ^2_7 distribution, as shown by the quantiles in Table 5.11.

	Mean	5%	25%	50%	75%	95%
Laplace LRT	6.846	1.881	3.96	5.805	8.234	14.18
True χ^2_7	7	2.167	4.255	6.346	9.037	14.07

Table 5.11: Mean and sample quantiles of the Laplace approximated LRT compared against the corresponding true quantities for a χ^2_7 distribution.

5.5 Discussion

The studies in this chapter show that, with respect to estimation biases, the relative performance of the approximate likelihood approaches against two prominent alternative approaches, Bayesian and GHQ approaches, was mixed. However, there are some interesting trends and issues revealed by these case studies.

Firstly, we summarise the relative performance of the different approaches for each case study in turn. For the paired binary simulation study (section 5.2.2), the second order HGLM approaches performed as well as Bayesian and AGHQ approaches,

and PQL performed poorly as expected. For the extra paired binary and Poisson examples (section 5.2.3), where the variance component was smaller, some negative estimation bias for the Bayesian approach was also found.. In this case, the variance of the AGHQ estimator was higher than PQL, and it consequently had a larger MSE, despite the estimation bias of PQL. For the Rodriguez/Goldman examples (section 5.2.4), the Bayesian and GHQ estimators clearly performed better than either PQL or HGLM approaches, with the HGLM approaches performing especially poorly, with no finite maximum found using the second order approaches. However, for the spatial case study (section 5.3), the PQL approach performed as equally well as the Bayesian `geoRglm` approach with much smaller computational requirements. And for the RCBD case study, PQL surprisingly performed equally well as for the AGHQ approach, despite the poor performance of PQL demonstrated in chapter 3, and the HGLM and Bayesian approaches showed some positive biases.

Apart from the expected estimation bias problems for the other studies, PQL performed well for the Landis case study, the spatial case study and the RCBD simulation study. For the Landis dataset (section 5.2.1), the group (clinic) sizes were relatively large and there was a small number of groups (clinics), both of which conditions mediated towards lower biases, compared to simulation studies in Chapter 3 and other studies of this chapter. For the spatial case study (section 5.3), there was some bias observable for one of the variance parameters, but the Bayesian approach also suffered from similar estimation bias as well. Finally, for the RCBD study (section 5.2.5), where, like the Landis dataset, the group size was not too small and the number of groups small, the PQL approach gave very similar estimates to the GHQ approach. This similarity was perhaps somewhat of a fluke, since it appears that the positive correction to the estimates from using the REML-like correction in PQL offset the negative bias of the PQL approach in this case, as can be seen by comparing to the non-REML PQL estimators. Nevertheless, PQL performed relatively well in this simulation study compared to the other approaches, regardless of whether a REML-like correction was used or not. For the ordinal case study (section 5.4), PQL estimation biases were not severe either, and the use of a Laplace approximation to enable

likelihood-ratio testing of variance parameters appears promising.

The performance of the HGLM approaches was somewhat disappointing in these studies. As noted above, the second order approaches performed well for the first paired binary simulation study (section 5.2.2), but quite as well in the second paired study (section 5.2.3). The divergence problem for the Rodriguez-Goldman datasets for the second order approaches was disappointing.

It was interesting to note the concordance between the (second order) HGLM and Bayesian estimators of the variance components in both the first paired binary simulation study (section 5.2.2) and in the RCBD example (section 5.2.5). Similarly, there was a concordance between the non-REML (second order) HGLM estimators and GHQ estimators in both studies. The non-REML and GHQ estimators, which don't include a REML-like correction, performed better in both studies than the standard HGLM or Bayesian approaches. This possibly suggests that the generalization of a REML-like correction to GLMMs may lead to overcorrection of the biases for the variance parameters.

Chapter 6

Conclusions

Examples of non-normal data with clustering or other sources of correlation are abundant in the agricultural and biological sciences. GLMMs offer an appealing way of modelling multiple sources of clustering and correlation for non-normal data in a probabilistic framework, unlike marginal approaches such as GEEs (section 1.4.3.1). One disadvantage of the GLMM framework is the assumption of normally distributed random effects, however the normality assumption is probably a sensible choice as a default distribution for random effects, since the random effects are on the scale of the linear predictor, unless the data provides sufficient evidence otherwise. Alternatives to the normality assumption include the generalisation to non-normal random effects as in the HGLMs of Lee & Nelder (1996) or using a nonparametric distribution for the random effects, like Aitkin (1999). However, Aitkin's approach removes the ability to make predictions involving the random effects, which can be important, especially for estimating spatial trends. Therefore, GLMMs have some appealing advantages over competing techniques for modelling clustering and correlation for non-normal data.

The intractability of the expression for the GLMM likelihood (section 1.4.2) has motivated a variety of alternative approaches for inference. In this thesis, these approaches have broadly been divided into approximate likelihood approaches (section 2.1), including PQL and the HGLM approaches, and other approaches, of which the most prominent are Bayesian approaches (section 2.3.2) and Gauss-Hermite quadrature

(GHQ, section 2.2).

The most well-known approximate likelihood approach is penalized quasi-likelihood (PQL). The main appeal of the PQL approach, and one that is often re-iterated in the literature, is that it can fit virtually any type of GLMM with relatively light computational requirements. (However, as indicated in the computational issues of the HGLM (section 4.2.1), it is necessary that sparse matrix techniques are employed for PQL to have light computational requirements for larger models.) However, PQL can suffer from severe estimation biases, as demonstrated in previous literature and reviewed in the simulations of Chapter 3. This was particularly true for binary GLMMs, confirming results of previous literature, but was also found for sparse Poisson data with low average rates, the latter which have been less explored. Estimation biases were generally much larger in magnitude for the variance components than for the fixed effects. However, fixed effects and variance components which are "orthogonal" to the random effects, such as fixed effects corresponding to covariates varying within groups (e.g. β_1 in the model of section 3.1.4.1) and correlation parameters (e.g. ρ in the correlated AR model of section 3.1.5.2), may be subject to less bias. Despite the estimation bias problems, the simulation studies of section 3.2.2 suggest PQL performs adequately for hypothesis testing of fixed effects when the null hypothesis is a zero effect (which is the null hypothesis most commonly of interest). PQL can also be used to test variance components if the (first order) Laplace approximation is computed, however the test is expected to be generally conservative since the estimation biases for variance components are generally negative (section 3.2.1). Despite this conservativeness, it is recommended that PQL applications calculate and produce the (first order) Laplace approximation to allow the user to conduct approximate testing of variance components.

The HGLM approach of Lee and Nelder (chapter 4) is another prominent approximate likelihood approach. Its proponents claim it admits the best of both worlds, in reducing the estimation biases suffered by PQL whilst enjoying the relatively light computational requirements of PQL. Simulation studies in section 4.2.2 showed the

first order HGLM approaches did have lower magnitudes of estimation bias compared to PQL, and appeared to provide adequate estimators, that is estimators with minimal estimation bias, for sparse Poisson data and for the fixed coefficients of binary models. One important advantage of using the first order HGLM approaches, compared to PQL, was that the estimation biases did not increase with the magnitude of the variance components. However, for binary data with proportions well away from 0.5, the HGLM first order estimators were unstable, and diverged for many of the simulated datasets, that is, converged to unusually high values. Some intuitive arguments in section 4.2.4 suggest that these divergence problems may be restricted to binary GLMMs, or perhaps binomial GLMMs with small denominators. Since the use of first order HGLM approaches resulted in non-trivial estimation biases for the variance components in binary models, second order HGLM approaches were examined. Further simulation studies (section 4.3.1) showed that the use of second order HGLM approaches generally resulted in small estimation bias for the variance parameters in the binary one-way classification, but there were still some estimation biases for the nested two-way classification, with instability and divergence also a problem.

Both the PQL and the HGLM approaches are compared with Bayesian and GHQ approaches for studies in chapter *5*. The relative performance of PQL and HGLM approaches is mixed. For the infamous Rodriguez-Goldman dataset (section 5.2.4) both PQL and HGLM approaches performed badly. However, for the other case studies, the approximate likelihood techniques performed reasonably well. The second order HGLM approaches had little or no bias in the paired binary study (section 5.2.2). Although the PQL approach incurred non-trivial estimation bias for the extra paired binary and Poisson examples (section 5.2.3), it performed better with respect to the MSE than GHQ for these examples. PQL also performed very well for the RCBD simulation study (section 5.2.5) relative to GHQ and Bayesian approaches. In some of these case studies, for instance, the paired binary study of section 5.2.2, both the Bayesian and HGLM estimators of the variance component were positively biased, whereas the GHQ estimator, and the non-REML HGLM estimator, was not. This may suggest that a generalization of REML to non-normal models may not work

well in general.

Some broad recommendations on the use of approximate likelihood approaches will be ventured. It is clear that approximate likelihood techniques will probably be less suitable for GLMMs where estimation of the variance components is the main interest, especially in cases where estimation bias issues are likely to be prominent. The most obvious application where estimation of variance components is important is in quantitative genetics or breeding studies, where the variance components are required to estimate heritability of traits in populations. Although the use of HGLM approaches can significantly reduce the biases, they could be prone to instability for binary GLMMs as shown in the studies of chapters 4 and 5.

Where prediction of random effects is the main interest, such as in spatial modelling, estimation biases may also be a problem. In the spatial case study of section 5.3, however, PQL performed well with respect to prediction error against a much more computationally intensive Bayesian approach. In many agricultural studies, especially data arising from designed experiments, the fixed effects are often of most interest, representing, for instance, the treatment factors in the study. Since PQL appears to be adequate for the hypothesis testing of fixed effects, it is probably adequate for detecting treatment differences and contrasts in data from designed experiments.

Note that the studies in this thesis only examined the most prominent GLMM approaches, PQL, HGLM approaches, GHQ and Bayesian approaches. Other GLMM approaches may be no less worthy of attention, but were not examined here, particularly because little "off the shelf" implementations of these were available. However, many GLMM approaches simply cannot fit the broad range of GLMMs that approximate approaches like PQL can. In particular, a number of relatively new and promising approaches have not been explored, , such as the stochastic EM approach of Delyon *et al.* (1999) and the quasi-Monte Carlo (QMC) approach (Pan & Thompson, 2000, Kuo *et al.*, 2008). . They both are Monte Carlo based approaches, and so would not appear to be prone to estimation bias problems that approximate likelihood approaches have, but appear to require lighter computational requirements than

standard Monte Carlo approaches. In addition, instability problems with the second order HGLM approach could be alleviated by using an even higher order Laplace approximation, such as in Raudenbush *et al.* (2000). Finally, one could pursue hybrid approaches, using a combination of the Laplace approximation and Monte Carlo approaches, such as described in Kuk (1999) or Lai & Shih (2003).

This thesis has concentrated on estimation methods for GLMMs. Other important issues relevant for the application of GLMMs to applied work have not been investigated, such as model fitting and diagnostics. Extensions to GLMMs, such as the HGLM and DHGLM models discussed in Lee & Nelder (2006), or a factor analytical extension to GLMMs called GLLAMMs, were outside the scope of this thesis but appear to be also very useful extensions and worthy of further investigation.

Bibliography

Note: URL addresses are provided for some references (preprints) for convenience. These were correct at the time of writing.

Abramowitz, M. & Stegun, I.A. (Editors) (1972). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York.

Ainsworth, L.M & Dean, C.B. (2006). Approximate inference for disease mapping. *Computational Statistics and Data Analysis* **50**, 2552–2770.

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics* **55**, 117–128.

Albert, J.A. & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88**, 669–679.

Andersen, E.B. (1973). Conditional inference for multiple-choice questionnaires. *British Journal of Mathematical and Statistical Psychology* **26**, 31–41.

Barndorff-Neilsen, O.E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika* **70**, 343–365.

Barndorff-Nielsen, O.E. (1978). *Information and Exponential Families in Statistical Theory*. Wiley, New York.

Bates, D. (2007). Linear mixed model implementation in `lme4`. Technical report, University of Wisconsin, Madison.
URL `http://cran.r-project.org/doc/vignettes/lme4/Implementation.pdf`

Bates, D. & Sarkar, D. (2006). *lme4: Linear mixed-effects models using S4 classes*. R package version 0.9975-7.
URL http://cran.R-project.org/

Beitler, P.J. & Landis, J.R. (1985). A mixed-effects model for categorical data. *Biometrics* **41**, 991–1000.

Berger, J.O., Liseo, B. & Wolpert, R.L. (1999). Integrated likelihood methods for eliminating nuisance parameters. *Statistical Science* **14**, 1–28.
URL http://citeseer.ist.psu.edu/berger97integrated.html

Booth, J.G. & Hobert, J.P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society B – Methodological* **61**, 265–285.

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

Breslow, N.E. (2003). Whither PQL? Technical Report 192, UW Biostatistics Working Paper Series, University of Washington.
URL http://www.bepress.com/uwbiostat/paper192/

Breslow, N.E. & Lin, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika* **82**, 81–91.

Browne, W.J. & Draper, D. (2000). Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics* **15**, 391–420.
URL http://www.maths.nott.ac.uk/personal/pmzwjb/materials/wbcs.pdf

Browne, W.J. & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis* **1**, 473–514.
URL http://www.maths.nott.ac.uk/personal/pmzwjb/materials/wbrssa.pdf

Callens, M. & Croux, C. (2005). Performance of likelihood-based estimation methods for multilevel binary regression models. *Journal of Statistical Computation*

and Simulation **75**, 1003–1017.

URL `http://www.econ.kuleuven.be/public/ndbae06/`

Candy, S.G. (1997). Estimation in forest yield models using composite link functions with random effects. *Biometrics* **53**, 146–160.

Carlin, J.B, Wolfe, R., Brown, C.R. & Gelman, A. (2001). A case study on the choice, interpretation and checking of multilevel models for longitudinal binary outcomes. *Biostatistics* **2**, 397–416.

URL `http://www.stat.columbia.edu/~gelman/research/published/397.pdf`

Carroll, R.J. & Ruppert, D. (1988). *Transformation and weighting in regression*. Chapman and Hall, London.

Chen, Y-H. (2006). Computationally efficient Monte Carlo EM algorithms for generalized linear mixed models. *Journal of Statistical Computation and Simulation* **76**, 817–828.

Chib, S. & Carlin, B. (1999). On MCMC sampling in hierarchical longitudinal models. *Statistics and Computing* **9**, 17–26.

URL `http://citeseer.ist.psu.edu/chib98mcmc.html`

Christensen, O.F. & Ribeiro Jr., P.J. (2002). `geoRglm`: A package for generalised linear spatial models. *R-NEWS* **2**, 26–28.

URL `http://cran.R-project.org/doc/Rnews`

Clarkson, D.B. & Zhan, Y. (2002). Using spherical-radial quadrature to fit generalized linear mixed effect models. *Journal of Computational and Graphical Statistics* **11**, 639–659.

Clayton, D.G. (1996). Generalized linear mixed models. In W.R. Gilks, S. Richardson & D.J. Spiegelhalter, (editors), *Markov Chain Monte Carlo in Practice*, chapter 16, pages 275–301. Chapman and Hall, London.

Cox, D.R. (1975). Partial likelihood. *Biometrika* **62**, 269–276.

Cox, D.R & Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *Journal of the Royal Statistical Society B – Methodological* **49**, 1–39.

Damien, P., Wakefield, J. & Walker, S. (1999). Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *Journal of the Royal Statistical Society B – Methodological* **61**, 331–344.

Delyon, B., Lavielle, M. & Moulines, E. (1999). Convergence of a stochastic approximation version of the EM Algorithm. *Annals of Statistics* **27**, 94–128.

Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood from incomplete observations. *Journal of the Royal Statistical Society, Series B – Methodological* **39**, 1–38.

Diggle, P., Moyeed, R., Rawlingson, B. & Thomson, M. (2002). Childhood malaria in the Gambia: A case study in model-based geostatistics. *Applied Statistics* **51**, 493–506.

Draper, N.R. & Smith, H. (1998). *Applied Regression Analysis*. Wiley, New York, 3rd edition.

Drum, M. & McCullagh, P. (1993). REML estimation with exact covariance in the logistic mixed model. *Biometrics* **49**, 677–689.

Engel, B. (1998). A simple illustration of the failure of PQL, IRREML and APHL as approximate ML methods for mixed models for binary data. *Biometrical Journal* **2**, 141–154.

Engel, B. & Buist, W. (1996). Analysis of a generalized linear mixed model: a case study and simulation results. *Biometrical Journal* **38**, 61–80.

Engel, B. & Buist, W. (1998). Bias reduction of approximate maximum likelihood estimates for heritability in threshold models. *Biometrics* **54**, 1155–1164.

Engel, B., Buist, W. & Visscher, A. (1995). Inference for threshold models with variance components from the generalized linear mixed model perspective. *Genetics, Selection and Evolution* **27**, 15–32.

Engel, B. & Keen, A. (1994). A simple approach for the analysis of generalized linear mixed models. *Statistica Neerlandica* **48**(1), 1–22.

Engel, B. & Keen, B. (1996). An introduction to generalized linear mixed models. In *XVIIIth International Biometrics Conference*. Amsterdam.

Gelfand, A. E. & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398–409.

Gelfand, A.E., Sahu, S.K. & Carlin, B.P. (1995). Efficient parametrizations for normal linear mixed models. *Biometrika* **82**, 479–488.

Gelman, A. (2005). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* **1**, 1–19.
URL `http://www.stat.columbia.edu/~gelman/research/published/`

Ghosh, M. & Rao, J. N. K. (1994). Small area estimation: An appraisal (with discussion). *Statistical Science* **9**, 55–93.

Gilks, W.R. (1992). Derivative-free adaptive rejection sampling for Gibbs sampling. In J.M. Bernardo, Berger J.O., A.P. Dawid & Smith A.F.M., (editors), *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting*, pages 641–649. Clarendon Press, Oxford.

Gilks, W.R. & Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics* **41**, 337–348.

Gilmour, A.R. (1983). *The estimation of genetic parameters for categorical traits.* Ph.D. thesis, School of Animal Science, Massey University.

Gilmour, A.R, Anderson, B.D. & Rae, A.L. (1985). The analysis of binomial data by a generalized linear mixed model. *Biometrika* **72**, 593–599.

Gilmour, A.R., Cullis, B.R. & Verbyla, A.P. (1997). Accounting for natural and extraneous variation in the analysis of field experiments. *Journal of Agricultural, Biological, and Environmental Statistics* **2**, 269–293.

Gilmour, A.R., Gogel, B.J., Cullis, B., Welham, S.J. & Thompson, R. (2006). *ASReml User Guide Release 2.0*. VSN International Limited, Hemel Hampstead.
URL `http://www.asreml.com`

Gilmour, A.R., Thompson, R. & Cullis, B.R. (1995). Average information REML: An efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics* **51**, 1440–1450.

Godambe, V.P & Heyde, C.C (1987). Quasi-likelihood and optimal estimating equations. *Biometrika* **55**, 231–244.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika* **73**, 43–56.

Goldstein, H. (1989). Restricted unbiased iteratively reweighted least squares estimation. *Biometrika* **76**, 622–623.

Goldstein, H. (1991). Non-linear multilevel models, with an application to discrete response data. *Biometrika* **78**, 45–51.

Goldstein, H. (1995). *Multilevel statistical models*. Edward Arnold, London, 1st edition.
URL `http://www.soziologie.uni-halle.de/langer/multilevel/`

Goldstein, H. (1996). Consistent estimators for multilevel generalized linear models using an iterated bootstrap. *Multilevel modelling newsletter* **8**, 3–6.
URL `http://www.cmm.bristol.ac.uk/learning-training/`

Goldstein, H. & Rasbash, J. (1996). Improved approximations for multilevel models with binary response. *Journal of the Royal Statistical Society A – General*

159, 505–513.

URL http://www.mlwin.com/team/materials/iammbr.pdf

Goldstein, H., Rasbash, J., Plewis, I., Draper, D., Browne, W., Yang, M., Woodhouse, G & Healy, M (1998). *A user's guide to MLwiN*. Institute of Education, London.

URL http://multilevel.ioe.ac.uk/

Harville, D.A. (1977). Maximum-likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association* **72**, 320–340.

Harville, D.A. & Mee, R.W. (1984). A mixed model procedure for analysing ordered categorical data. *Biometrics* **40**, 393–408.

Hastings, W.K. (1970). Monte Carlo sampling methods using Markov Chains and their applications. *Biometrika* **57**, 97–109.

Hedeker, D. & Gibbons, R.D. (1996). MIXOR: A computer program for mixed effects ordinal regression modelling. *Computer Methods and Programs in Biomedicine* **49**, 157–176.

URL http://tigger.uic.edu/~hedeker/mix.html

Henderson, C.R., Kempthorne, O., Searle, S.R. & Von Krisig, C.N. (1959). Estimation of environmental and genetic trends from records subject to culling. *Biometrics* **15**, 192–218.

Hobert, J.P. (2000). Hierarchical models: A current computational perspective. *Journal of the American Statistical Association* **95**, 1312–1316.

Hobert, J.P. & Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association* **91**, 1461–1479.

Jiang, J. (1998). Consistent estimators in generalized linear mixed models. *Journal of the American Statistical Association* **93**, 720–729.

Karim, M.R. & Zeger, S.L. (1992). Generalized linear models with random effects; salamander mating revisited. *Biometrics* **48**, 631–644.

Kass, R. & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the Americal Statistical Association* **91**, 1343–1370.

Kenward, M.G. & Roger, J.H. (1997). Small sample inference for fixed effect estimators from restricted maximum likelihood. *Biometrics* **53**, 983–997.

Kneib, T. & Fahrmeir, L. (2004). Structured additive regression for multicategorical space-time data: A mixed model approach. Discussion Paper 377, Ludwig Maximilians University, SFB 386, Munich.
URL `http://www.stat.uni-muenchen.de/~fahrmeir/sfbpapers-e.html`

Kuk, A.Y.C. (1995). Asymptotically unbiased estimation in generalized linear mixed models with random effects. *Journal of the Royal Statistical Society B – Methodological* **57**, 395–407.

Kuk, A.Y.C. (1999). Laplace importance sampling for generalized linear mixed models. *Journal of Statistical Computation and Simulation* **63**, 143–158.

Kuk, A.Y.C. & Cheng, Y.W. (1999). Pointwise and functional approximations in Monte Carlo maximum likelihood estimation. *Statistics and Computing* **9**, 91–99.

Kuo, F.W., Dunsmuir, W.T.M, Sloan, I.H., Wand, M.P. & Womersley, R.S.W. (2008). Quasi-Monte Carlo for Highly Structured Generalised Response Models. *Methodology and Computing in Applied Probability* **10**, 239–275.
URL `http://ro.uow.edu.au/infopapers/529/`

Lai, T.L. & Shih, M-C. (2003). A hybrid estimator in nonlinear and generalised linear mixed effects models. *Biometrika* **90**, 859–879.

Laird, N.M. (1978). Empirical Bayes methods for two-way contingency tables. *Biometrika* **65**, 581–590.

Lee, Y. & Nelder, J.A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society B – Methodological* **58**, 619–678.

Lee, Y. & Nelder, J.A. (2001). Hierarchical generalized linear models: A synthesis of generalized linear models, random effects models and structured dispersions. *Biometrika* **88**, 987–1006.

Lee, Y. & Nelder, J.A. (2003). Extended-REML estimators. *Journal of Applied Statistics* **30**, 845–856.

Lee, Y. & Nelder, J.A. (2004). Conditional versus marginal models - another view. *Statistical Science* **19**, 219–238.

Lee, Y. & Nelder, J.A. (2005). Likelihood for random effects (with discussion). *Statistics and Operations Research Transactions* **29**, 141–178.

Lee, Y. & Nelder, J.A. (2006). Double hierarchical generalized linear models (with discussion). *Applied Statistics* **55**, 139–185.

Lee, Y., Nelder, J.A. & Pawitan, Y. (2006). *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood*. CRC Press, London.

Lee, Y.D., Yun, S. & Lee, Y. (2003). Analyzing weather effects on airborne particulate matter with HGLM. *Environmetrics* **14**, 687–697.

Lesaffre, E. & Spiessens, B. (2001). On the effect of number of quadrature points in a logistic random-effects model: an example. *Applied Statistics* **50**, 325–335.

Liang, K.Y. & Zeger, S.L (1986). Longitudinal data analysis using generalised linear models. *Biometrika* **73**, 13–22.

Lillard, L.A. & Panis, C.W.A. (2003). *AML Multilevel Multiprocess Statistical Software, Version 2.0*. EconWare, Los Angeles, California.

Lin, X. (1997). Variance component testing in generalized linear models with random effects. *Biometrika* **84**, 309–326.

Lin, X. & Breslow, N.E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association* **91**, 1007–1016.

Lindsey, J. (2004). On h-likelihood, random effects and penalized likelihood. Technical report, Biostatistics, Limburgs Universitair Centrum, Denmark.
URL http://popgen.unimaas.nl/~jlindsey/ms/hglm.ps

Lindsey, J.K. & Lambert, P. (1998). On the approximation of marginal models for repeated measurements in clinical trials. *Statistics in Medicine* **17**, 447–469.

Lindstrom, M.J. & Bates, D.M (1988). Newton-Raphson and EM algorithms for linear mixed effects models for repeated measures data. *Journal of the American Statistical Association* **83**, 1014–1022.

Lindstrom, M.J. & Bates, D.M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics* **46**, 673–687.

Liu, C.H., Rubin, D.B. & Wu, Y.N. (1998). Parameter expansion to accelerate EM – the PX-EM algorithm. *Biometrika* **85**, 755–770.
URL http://cm.bell-labs.com/stat/liu/docs/px-em.ps

Liu, J.S. & Wu, Y. (1999). Parameter expansion for data augmentation. *Journal of the American Statistical Association* **94**, 1264–1274.
URL http://www-stat.stanford.edu/~jliu/TechRept/97folder/

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

Longford, N.T. (1988). A quasi-likelihood adaption for variance component analysis. In *Proceedings of the Statistical Computing Section*. American Statistical Association.

MacKay, D. J. C. (1998). Introduction to Monte Carlo methods. In M. I. Jordan, (editor), *Learning in Graphical Models*, NATO Science Series, pages 175–204. Kluwer Academic Press, New York.
URL http://www.cs.ucsb.edu/~cs265/papers/erice.pdf

McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*. Chapman & Hall, London, 2nd edition.

McCullagh, P. & Tibshirani, R. (1990). A simple method for the adjustment of profile likelihoods. *Journal of the Royal Statistical Society B – Methodological* **52**, 325–344.

URL `http://www.stat.uchicago.edu/~pmcc/publications.html`

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

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

McGilchrist, C.A. (1994). Estimation in generalized mixed models. *Journal of the Royal Statistical Society B – Methodological* **56**, 61–69.

Meng, X.L & Rubin, D.B. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika* **80**, 267–278.

Meng, X.L. & van Dyk, D. (1997). The EM algorithm – an old folk song sung to the fast tune (with discussion). *Journal of the Royal Statistical Society B – Methodological* **59**, 511–567.

Mengersen, K.L, Robert, C.P. & Guihenneuc-Jouyaux, C. (1998). MCMC Convergence Diagnostics: A "RevieWWW". Technical report, Queensland University of Technology.

URL `http://citeseer.ist.psu.edu/78250.html`

Misztal, I. (1999). *SPARSEM: A collection of sparse matrix modules for Fortran 90 useful in animal breeding problems*. University of Georgia.

URL `http://nce.ads.uga.edu/~ignacy/newprograms.html`

Monahan, J & Genz, A. (1997). Spherical-radial integration rules for Bayesian computation. *Journal of the American Statistical Association* **92**, 664–674.

Naylor, J.C & Smith, A.F.M. (1982). Applications of a method for the efficient computation of posterior distributions. *Applied Statistics* **31**, 214–225.

Neal, R. (2003). Slice sampling. *Annals of Statistics* **31**, 705–767.

Nelder, J. A. (1954). The interpretation of negative components of variance. *Biometrika* **41**, 544–548.
URL `http://biomet.oxfordjournals.org/cgi/reprint/41/3-4/544.pdf`

Nelder, J.A. (1965a). The Analysis of Randomized Experiments with Orthogonal Block Structure. I. Block Structure and the Null Analysis of Variance. *Journal of the Royal Society* **283**, 147–162.

Nelder, J.A. (1965b). The Analysis of Randomized Experiments with Orthogonal Block Structure. II. Treatment Structure and the General Analysis of Variance. *Journal of the Royal Society* **283**, 163–178.

Nelder, J.A. & Mead, R. (1964). A simplex method for function minimization. *The Computer Journal,* **7**, 308–313.

Nelder, J.A. & Pregibon, D. (1987). An extended quasi-likelihood function. *Biometrika* **74**, 221–231.

Nelder, J.A. & Wedderburn, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society A – General* **135**, 370–384.

Neuhaus, J. M., Hauck, W. W. & Kalbfleish, J. D. Z. (1992). The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika* **79**, 755–762.

Ng, E.S.W., Carpenter, J.R., Goldstein, H. & Rasbash, J. (2006). Estimation in generalised linear mixed models with binary outcomes by simulated maximum likelihood. *Statistical Modelling* **6**, 23–42.

Noh, M. & Lee, Y. (2007). REML estimation for binary data in GLMMs. *Journal of Multivariate Analysis* **98**, 896–915.

Noh, M., Lee, Y. & Pawitan, Y. (2005). Robust ascertainment-adjusted parameter estimation. *Genetic Epidemiology* **29**, 68–75.

Noh, M., Yip, B., Lee, Y. & Pawitan, Y. (2006). Multicomponent variance estimation for binary traits in family-based studies. *Genetic Epidemiology* **30**, 37–47.

Paciorek, C.J. (2007). Computational techniques for spatial logistic regression with large data sets. *Computational Statistics and Data Analysis* **51**, 3631–3653.
URL `http://biosun1.harvard.edu/~paciorek/files/spatfit/paci.2007.pdf`

Pan, J. & Thompson, R. (2000). Generalized linear mixed models: an improved estimating procedure. In J.G. Bethlehem & P.G.M. van der Heijden, (editors), *Proceedings in Computational Statistics (COMPSTAT)*, pages 373–378. Physical-Verlag.

Pan, J. & Thompson, R. (2003). Gauss-Hermite quadrature approximation for estimation in generalized linear mixed models. *Computational Statistics* **18**, 57–78.

Papaspiliopoulos, O., Roberts, G. O. & Skold, M. (2003). Non-centered parameterisations for hierarchical models and data augmentation. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith & M. West, (editors), *Bayesian Statistics 7*, pages 307–326. Oxford University Press.

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

Payne, R.W., Harding, S.A., Murray, D.A., Soutar, D.M., Baird, D.B., Welham, S.J., Kane, A.F., Gilmour, A.R., Thompson, R., Webster, R. & Tunnicliffe Wilson, G. (2006). *The Guide to GenStat Release 9, Part 2: Statistics*. VSN International, Hemel Hempstead.
URL `http://www.genstat.com`

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL `http://www.R-project.org`

Rabe-Hesketh, S., Pickles, A. & Skrondal, A. (2001). `GLLAMM`: A class of models and a Stata program. *Multilevel Modelling Newsletter* **13**, 17–23.

Rabe-Hesketh, S., Skrondal, A. & Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive Gaussian quadrature. *Stata Journal* **2**(1), 1–21.

Rabe-Hesketh, S., Skrondal, A. & Pickles, A. (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics* **128**, 301–323.

Raudenbush, S.W., Yang, M.L. & Yosef, Y. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics* **9**, 141–157.

Reid, N. (1991). *Statistical theory and modelling*, chapter "Approximations and asymptotics", pages 287–305. Chapman and Hall, London.

Rodriguez, German & Goldman, Noreen (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society A – General* **158**, 73–89.

Rodriguez, German & Goldman, Noreen (2001). Improved estimation procedures for multilevel models with binary response: A case study. *Journal of the Royal Statistical Society A – General* **164**, 339–355.

SAS Institute Inc. (2000). *SAS 8 Help and Documentation*. SAS Institute Inc., Cary, NC.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika* **78**, 719–27.

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

Shun, Z. (1997). Another look at the salamander mating data: A modified Laplace approximation approach. *Journal of the American Statistical Association* **92**, 341–349.

Shun, Z. & McCullagh, P. (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society B – Methodological* **57**, 749–760.

Smyth, G. (1997). Optimisation and non-linear equations. Technical report, Walter and Eliza Research Institute of Medical Research, Parkville, Victoria.
URL `http://www.statsci.org/smyth/pubs/optimise.ps`

Smyth, G.K. (1989). Generalized linear models with varying dispersion. *Journal of the Royal Statistical Society B– Methodological* **51**, 47–60.

Spiegelhalter, D. J., Thomas, A., Best, N. G. & Gilks, W. R. (1995). *BUGS: Bayesian inference Using Gibbs Sampling, Version 0.50*. MRC Biostatistics Unit, Cambridge.

StataCorp (2007). *Stata Statistical Software: Release 10*. StataCorp LP., College Station, TX.

Statistics and Epidemiology Research Corporation (1993). *EGRET Reference Manual*. Seattle, revision 4 edition.
URL `http://www.cytel.com/Egret`

Steele, B.M (1996). A modified EM algorithm for estimation in generalized mixed models. *Biometrics* **52**, 1295–1310.

Stein, M. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, New York.

Stiratelli, R., Laird, N. & Ware, J.H. (1984). Random-effects models for serial observations with binary responses. *Biometrics* **40**, 961–971.

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

Thompson, R., Cullis, B., Smith, A. & Gilmour, A.R. (2003). A sparse implementation of the Average Information algorithm for factor-analytic models. *Australian and New Zealand Journal of Statistics* **45**, 445–459.

Tierney, L., Kass, R.E. & Kadane, J.B. (1989). Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *Journal of the American Statistical Association* **84**, 710–716.

Trottier, C. (1998). A quasi-score marginal approach in generalized linear mixed models. Report 3522, INRAI, St Martin, France.
URL `http://www.inria.fr/rrrt/rr-3522.html`

Tweedie, M.C.K. (1984). An index which distinguishes between some important exponential families. In Eds. J. K. Ghosh & J. Roy, (editors), *Statistics: Applications and New Directions. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference*, pages 579–604. Indian Statistical Institute, Calcutta.

Verbyla, A.P., Cullis, B.R., Kenward, M.G. & Welham, S.J. (1999). The analysis of designed experiments and longitudinal data by using smoothing splines (with discussion). *Applied Statistics* **48**, 269–311.

Wei, G.C.G. & Tanner, M.A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association* **85**, 699–704.

Wolfinger, R. (1999). Fitting nonlinear mixed models with the new NLMIXED procedure. In *Proceedings of the 24th Annual SAS User Group International (SUGI) conference*. SAS Institute Inc., Cary, NC.
URL `http://www.ats.ucla.edu/stat/sas/library/nlmixedsugi.pdf`

Wolfinger, R. & O'Connell, M. (1993). Generalized linear mixed models: A pseudo-likelihood approach. *Journal of Statistical Computation and Simulation* **48**, 233–243.

Yun, S. & Lee, Y. (2004). Comparison of hierarchical and marginal likelihood estimators for binary outcomes. *Computational Statistics and Data Analysis* **45**, 639–650.

Zeger, S.L. & Karim, M.R. (1991). Generalized linear models with random effects: A Gibbs sampling approach. *Journal of the American Statistical Association* **86**, 79–86.

Zeger, S.L., Liang, K.Y & Albert, P.S. (1988). Models for longitudinal data: A generalized estimating equation approach (corr: V45, p347). *Biometrics* **44**, 1049–1060.

Zhao, Y., Staudenmayer, J., Coull, B.A. & Wand, M.P. (2003). Towards general design Bayesian generalized linear mixed models. In *Proceedings of the ISI International Conference on Environmental Statistics and Health,*. International Statistical Institute, Santiago De Compostela, Spain.
URL `http://isi-eh.usi.es/trabajos/257_157_fullpaper.pdf`

Zhu, C., Byrd, R.H. & Nocedal, J. (1997). L-BFGS-B: Algorithm 778: L-BFGS-B, FORTRAN routines for large scale bound constrained optimization. *ACM transactions on Mathematical Software* **23**, 550–560.
URL `http://www.ece.northwestern.edu/~nocedal/lbfgsb.html`

Appendix A

Appendix

A.1 Expressions for implicit differentiation

The following uses the notation of section 4.2.3.

Since no closed form expression for $\tilde{\boldsymbol{u}}_{\tau,\gamma}$ is available, the derivatives $\partial\tilde{\boldsymbol{u}}_{\tau,\gamma}/\partial\boldsymbol{\tau}$ and $\partial\tilde{\boldsymbol{u}}_{\tau,\gamma}/\partial\boldsymbol{\gamma}$ are obtained using implicit differentiation. In the following, $\tilde{\boldsymbol{u}} = \tilde{\boldsymbol{u}}_{\tau,\gamma}$. Knowing that

$$\begin{aligned}\mathbf{0} &= \left.\frac{\partial h}{\partial \boldsymbol{u}}\right|_{\tilde{\boldsymbol{u}}} \\ &= \sum (y_i - \tilde{\mu}_i)\boldsymbol{z}_i - \boldsymbol{G}^{-1}\tilde{\boldsymbol{u}}\end{aligned}$$

and $\tilde{\mu}_i = \mu_i \mid_{\tilde{\boldsymbol{u}}}$, then

$$\begin{aligned}\mathbf{0} &= \left.\frac{\partial^2 h}{\partial \boldsymbol{u}\partial\tau_j}\right|_{\tilde{\boldsymbol{u}}} \\ &= -\boldsymbol{Z}^T\tilde{\boldsymbol{W}}(\boldsymbol{x}_j + \boldsymbol{Z}\frac{\partial\tilde{\boldsymbol{u}}}{\partial\tau_j}) - \boldsymbol{G}^{-1}\frac{\partial\tilde{\boldsymbol{u}}}{\partial\boldsymbol{\tau}_j},\end{aligned}$$

where $\boldsymbol{x}_j$ is the jth column of $\boldsymbol{X}$ and so

$$\frac{\partial\tilde{\boldsymbol{u}}}{\partial\tau_j} = -\left(\boldsymbol{Z}^T\tilde{\boldsymbol{W}}\boldsymbol{Z} + \boldsymbol{G}^{-1}\right)^{-1}\boldsymbol{Z}^T\tilde{\boldsymbol{W}}\boldsymbol{x}_j.$$

Assuming $\boldsymbol{G} = \boldsymbol{G}(\boldsymbol{\gamma})$, then

$$\begin{aligned}\mathbf{0} &= \left.\frac{\partial^2 h}{\partial \boldsymbol{u} \partial \gamma_j}\right|_{\tilde{\boldsymbol{u}}} \\ &= -\boldsymbol{Z}^T \tilde{\boldsymbol{W}} \boldsymbol{Z} \frac{\partial \tilde{\boldsymbol{u}}}{\partial \gamma_j} - \boldsymbol{G}^{-1} \frac{\partial \tilde{\boldsymbol{u}}}{\partial \gamma_j} + \boldsymbol{G}^{-1} \frac{\partial \boldsymbol{G}}{\partial \gamma_j} \boldsymbol{G}^{-1} \tilde{\boldsymbol{u}},\end{aligned}$$

and so

$$\frac{\partial \tilde{\boldsymbol{u}}}{\partial \gamma_j} = \left(\boldsymbol{Z}^T \tilde{\boldsymbol{W}} \boldsymbol{Z} + \boldsymbol{G}^{-1}\right)^{-1} \boldsymbol{G}^{-1} \frac{\partial \boldsymbol{G}}{\partial \gamma_j} \boldsymbol{G}^{-1} \tilde{\boldsymbol{u}}.$$

A.2 The second order Laplace approximation of an integral

The second order Laplace approximation will be derived for a univariate integral for simplicity.

A.2.1 Higher order Laplace approximations for a univariate integral

As in section 1.3.2.1, the integral to be approximated is $\int_{-\infty}^{\infty} \exp\{g(x)\}\, dx$.

A Taylor series expansion around the mode of $g(x)$, $\hat{x}$, gives

$$\begin{aligned}&\int_{-\infty}^{\infty} \exp\{g(x)\}\, dx \\ &\approx \int_{-\infty}^{\infty} \exp\left\{g(\hat{x}) + g'(\hat{x})(x-\hat{x}) + g''(\hat{x})(x-\hat{x})^2/2 + \sum_{i=3}^{\infty} g^{(i)}(x)(x-\hat{x})^i/i!\right\} dx \\ &= \exp\{g(\hat{x})\} \int_{-\infty}^{\infty} \exp\left\{g''(\hat{x})(x-\hat{x})^2/2\right\} \exp\left\{\sum_{i=3}^{\infty} g^{(i)}(\hat{x})(x-\hat{x})^i/i!\right\} dx\end{aligned}$$

where $g^{(i)}(\hat{x}) = \partial^i g/dx^i\big|_{x=\hat{x}}$, $i = 3, \ldots$ are the higher order derivatives of g evaluated at $\hat{x}$, and $g'(\hat{x}) = 0$ since $\hat{x}$ is the mode. Using

$$\exp\left\{\sum_{i=3}^{\infty} g^{(i)}(\hat{x})(x-\hat{x})^i/i!\right\} \approx 1 + \sum_{i=3}^{\infty} g^{(i)}(\hat{x})(x-\hat{x})^i/i! + \left(\sum_{i=3}^{\infty} g^{(i)}(\hat{x})(x-\hat{x})^i/i!\right)^2/2,$$

the expression becomes

$$\begin{aligned} \ldots &\approx \exp\{g(\hat{x})\} \int_{-\infty}^{\infty} \exp\left\{g''(\hat{x})(x-\hat{x})^2/2\right\} \\ &\quad + \exp\left\{g''(\hat{x})(x-\hat{x})^2/2\right\} \left\{G^{(3+)}(x) + \left[G^{(3+)}(x)\right]^2/2\right\} dx \end{aligned}$$

where

$$G^{(3+)}(x) = \sum_{i=3}^{\infty} g^{(i)}(\hat{x})(x-\hat{x})^i/i!.$$

Substituting

$$\int_{-\infty}^{\infty} \exp\left\{g''(\hat{x})(x-\hat{x})^2/2\right\} dx = \sqrt{\frac{2\pi}{-g''(\hat{x})}}$$

gives

$$\begin{aligned} \ldots &= \exp\{g(\hat{x})\} \sqrt{\frac{2\pi}{g''(\hat{x})}} \\ &\quad \left(1 + \int_{-\infty}^{\infty} \sqrt{\frac{-g''(\hat{x})}{2\pi}} e^{g''(\hat{x})(x-\hat{x})^2/2} \left\{G^{(3+)}(x) + \left[G^{(3+)}(x)\right]^2/2\right\} dx\right) \end{aligned}$$

The terms remaining within the integral represent expectations of the function $G^{(3+)}(x) + \left[G^{(3+)}(x)\right]^2/2$ over a normal distributed variable $x \sim N(\hat{x}, 1/-g''(\hat{x}))$, and so the approximation can be written

$$\int \exp\{g(x)\}\, dx \approx \exp\{g(\hat{x})\} \sqrt{\frac{2\pi}{-g''(\hat{x})}} \left[1 + E_x\left\{G^{(3+)}(x)\right\} + E_x\left\{\left[G^{(3+)}(x)\right]^2/2\right\}\right]. \tag{A.1}$$

The expectations of the terms in $E_x\left\{G^{(3+)}(x)\right\}$ can be shown to be

$$E\left\{g^{(k)}(\hat{x})(x-\hat{x})^k/k!\right\} = \begin{cases} 0, & \text{if k is odd} \\ \frac{(k-1)(k-3)\ldots 3}{k!} \frac{g^k(\hat{x})}{(-g''(\hat{x}))^{k/2}}, & \text{if k is even} \end{cases}.$$

Similarly for components of $E\left\{\left[G^{(3+)}(x)\right]^2/2\right\}$, the expectations are

$$E_x\left\{g^{(k)}(\hat{x})(x-\hat{x})^k g^{(l)}(\hat{x})(x-\hat{x})^l/k!l!\right\}$$
$$= \begin{cases} 0, & \text{if } (k+l) \text{ is odd} \\ \frac{(k+l-1)(k+l-3)\ldots 3}{k!l!}\frac{g^k(\hat{x})g^l(\hat{x})}{(-g''(\hat{x}))^{(k+l)/2}}, & \text{if } (k+l) \text{ is even} \end{cases}.$$

A.2.2 The "second order" Laplace approximation

The first order Laplace approximation (section 1.3.2.1) is defined as

$$\int \exp\{g(x)\}\,dx = \exp\{g(\hat{x})\}\sqrt{\frac{2\pi}{-g''(\hat{x})}} + O(n^{-1}).$$

Inclusion of the next two lowest order terms in (A.1) gives a "second order" Laplace approximation,

$$\begin{aligned}\int \exp\{g(x)\}\,dx &= \exp\{g(\hat{x})\}\sqrt{\frac{2\pi}{-g''(\hat{x})}} \\ &\quad \left[1 + E_x\left\{g^{(4)}(\hat{x})(x-\hat{x})^4/4!\right\} + \frac{1}{2}E_x\left\{g^{(3)}(\hat{x})(x-\hat{x})^3/3!\right\}^2\right] \\ &\quad + O\left(n^{-2}\right).\end{aligned}$$

Applying the approximation $1 + R \approx \exp(R)$ to the terms in square brackets gives

$$\exp\{g(\hat{x})\}\sqrt{\frac{2\pi}{g''(\hat{x})}}\left[1 + E_x\left\{g^{(4)}(\hat{x})(x-\hat{x})^4/4!\right\} + \frac{1}{2}E_x\left\{g^{(3)}(\hat{x})(x-\hat{x})^3/3!\right\}^2\right] \approx$$
$$e^{g(\hat{x})}\sqrt{\frac{2\pi}{g''(\hat{x})}}\exp\left[\frac{3}{4!}\frac{g^{(4)}(\hat{x})}{g''(\hat{x})^2} + \frac{5\times 3\times (g^{(3)}(\hat{x}))^2}{2!3!3!g''(\hat{x})^3}\right] \quad \text{(A.2)}$$

This is the univariate form of the expression that is used by Lee & Nelder (2001, 2006) for the second order Laplace approximation. Note that others have different definitions of a "second order" Laplace approximation. Breslow & Lin (1995), for instance, omit $E_x\left\{g^{(3)}(\hat{x})(x-\hat{x})^3/3!\right\}^2$. What Raudenbush *et al.* (2000) call a "sixth order" Laplace approximation is, in fact, only a third order Laplace approximation, with the addition of the subsequent term $E_x\left\{g^{(6)}(\hat{x})(x-\hat{x})^6/6!\right\}$.

A.3 Delta method of calculating SEs for PQL spatial predictions

This material outlines the "delta method" used in the spatial case study of section 5.3.

The "delta" method of Ainsworth & Dean (2006) was used for calculating the standard errors and confidence intervals of the predicted trend $\boldsymbol{\eta}^*$. Let $\boldsymbol{G} = cov(\boldsymbol{S})$ and $\hat{\boldsymbol{\psi}} = (\hat{\psi}_1 \dots \hat{\psi}_{400})'$ be the estimated working variable $\hat{\psi}_i = \log(\hat{\mu}_i) + (y_i - \hat{\mu}_i)/\hat{\mu}_i$. Then $\boldsymbol{\Sigma} = "var(\hat{\boldsymbol{\psi}})" = \boldsymbol{W}^{-1} + \boldsymbol{G}$ where $\boldsymbol{W}$ is a diagonal matrix of GLM weights with ith diagonal entry $w_i = 1/\mu_i = \exp(-\beta - S(\boldsymbol{\ell}_i))$.

Let $\boldsymbol{V}_{12} = cov(\boldsymbol{S}, \boldsymbol{S}^*)$. The conditional density of the spatial trend $\boldsymbol{\eta}^*$ given the parameters is

$$\boldsymbol{\eta}^* | \boldsymbol{\psi}, \beta, \sigma^2, \rho \sim N\left(\beta + \boldsymbol{V}_{12}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\psi} - \beta), \boldsymbol{V}_{22} - \boldsymbol{V}_{12}'\boldsymbol{\Sigma}^{-1}\boldsymbol{V}_{12}\right)$$

Accordingly, the predictions for the extra locations were calculated using the standard kriging formula

$$\hat{\boldsymbol{\eta}}^* = \hat{\beta} + \hat{\boldsymbol{V}}_{12}\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\psi}} - \hat{\beta}).$$

As in Ainsworth & Dean (2006) and letting $\boldsymbol{\delta} = (\beta, \sigma^2, \rho)'$,

$$\begin{aligned}
\text{var}\left(\hat{\boldsymbol{\eta}}^*\right) &= \text{E}\left\{\text{var}\left(\hat{\boldsymbol{\eta}}^* | \hat{\boldsymbol{\delta}}\right)\right\} + \text{var}\left\{\text{E}\left(\hat{\boldsymbol{\eta}}^* | \hat{\boldsymbol{\delta}}\right)\right\} \\
&= \text{E}\left(\hat{\boldsymbol{V}}_{22} - \hat{\boldsymbol{V}}_{12}'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{V}}_{12}\right) + \text{var}\left(\hat{\boldsymbol{\eta}}^*\right) \\
&\simeq \hat{\boldsymbol{V}}_{22} - \hat{\boldsymbol{V}}_{12}'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{V}}_{12} + \left(\frac{\partial \hat{\boldsymbol{\eta}}^*}{\partial \hat{\boldsymbol{\delta}}}\right)' \text{var}(\hat{\boldsymbol{\delta}}) \left(\frac{\partial \hat{\boldsymbol{\eta}}^*}{\partial \hat{\boldsymbol{\delta}}}\right)
\end{aligned}$$

where

$$\partial \hat{\boldsymbol{\eta}}^* / \partial \hat{\beta} = 1 + \hat{\boldsymbol{V}}_{12}\hat{\boldsymbol{\Sigma}}^{-1},$$

$$\frac{\partial \hat{\boldsymbol{\eta}}^*}{\partial \hat{\sigma}^2} \simeq \left\{\left(\hat{\boldsymbol{V}}_{12}/\hat{\sigma}^2\right)'\hat{\boldsymbol{\Sigma}}^{-1} - \hat{\boldsymbol{V}}_{12}\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{V}_{11}/\hat{\sigma}^2)\hat{\boldsymbol{\Sigma}}^{-1}\right\}(\hat{\boldsymbol{\psi}} - \hat{\beta}),$$

$$\frac{\partial \hat{\boldsymbol{\eta}}^*}{\partial \hat{\phi}} = \left\{\left(\frac{\partial \hat{\boldsymbol{V}}_{12}}{\partial \hat{\phi}}\right)\hat{\boldsymbol{\Sigma}}^{-1} - \hat{\boldsymbol{V}}_{12}\hat{\boldsymbol{\Sigma}}^{-1}\left(\frac{\partial \hat{\boldsymbol{V}}_{11}}{\partial \hat{\phi}}\right)\hat{\boldsymbol{\Sigma}}^{-1}\right\}(\hat{\boldsymbol{\psi}} - \hat{\beta}),$$

where the ith, jth element of $\partial\hat{\boldsymbol{V}}_{12}/\partial\hat{\phi}$ and $\partial\hat{\boldsymbol{V}}_{11}/\partial\hat{\phi}$, $i = 1\ldots400$, $j = 1\ldots425$ are

$$-\sigma^2\left(\frac{d_{ij}}{\phi^2}\right)e^{-d_{ij}/\phi} - \sigma^2\left(\frac{d_{ij}^2}{\phi^3}\right)e^{-d_{ij}/\phi}.$$

The 95% confidence intervals for $\hat{\mu}_i^* = \exp(\hat{\eta}_i^*)$ were calculated using a simple back-transformation as

$$\left(\exp\left\{\hat{\eta}_i^* - 2\sqrt{var\left(\hat{\eta}_i^*\right)}\right\}, \exp\left\{\hat{\eta}_i^* + 2\sqrt{var\left(\hat{\eta}_i^*\right)}\right\}\right).$$

A.4 Laplace approximations for the ordinal (5.8) and ordinal factor analytic (5.9) models

Since the only fixed effects in models (5.8) and (5.9) are the thresholds τ_k, the REML correction is ignored. The first order Laplace approximation of the likelihood for a GLMM is

$$p_u(h) = \left(h - \frac{1}{2}\log\left|-\frac{\partial h^2}{\partial\boldsymbol{u}\partial\boldsymbol{u}^T}\right|\right)_{\hat{\boldsymbol{u}}_{\tau,\gamma}}$$

where

$$h = \log f_{Y|U} + \log f_U = \sum_i\sum_j y_{ij}\log\left(\mu_{ij} - \mu_{i,j-1}\right) - \frac{1}{2}\log|\boldsymbol{G}| - \frac{1}{2}\boldsymbol{u}^T\boldsymbol{G}^{-1}\boldsymbol{u}$$

for both models and $y_{i,j} = I(y_i = j)$ and $I(\cdot)$ is the indicator function. As usual, $\boldsymbol{u}$ represents the vector of random effects in the model – for model (5.8), $\boldsymbol{u} = [u_{v1},\ldots,u_{v182},u_{r1},\ldots u_{t4},S(\boldsymbol{\ell}_{11}),\ldots,S(\boldsymbol{\ell}_{24,40})]^T$, and $\tilde{\boldsymbol{u}}_{\tau,\gamma}$ is the mode of h with respect to $\boldsymbol{u}$ for given $\boldsymbol{\gamma}$ and $\boldsymbol{\tau}$.

An expression for the $\left|-\partial h^2/\partial\boldsymbol{u}\partial\boldsymbol{u}^T\right|$ is therefore required for both models. Now

$$\frac{\partial h^2}{\partial\boldsymbol{u}\partial\boldsymbol{u}^T} = \frac{\partial^2 h_{Y|U}}{\partial\boldsymbol{u}\partial\boldsymbol{u}^T} + \boldsymbol{G}^{-1}$$

where $h_{Y|U} = \log f_{Y|U} = \sum_i\sum_j y_{ij}\log\left(\mu_{ij} - \mu_{i,j-1}\right)$. Let $\mu'_{ij} = \partial\mu_{ij}/\partial\eta_{ij} = \phi(\eta_{ij})$ and $\mu''_{ij} = \partial^2\mu_{ij}/\partial\eta_{ij}^2 = -\eta_{ij}\phi(\eta_{ij})$, and z_{is} be the i,sth element of $\boldsymbol{Z}$.

We consider $\partial h^2/\partial \boldsymbol{u}\partial \boldsymbol{u}^T$ for model (5.8) first. For this model,

$$\begin{aligned}\frac{\partial^2 h_{Y|U}}{\partial u_s \partial u_t} &= \sum_i \sum_j y_{ij} \left(\frac{(\mu_{ij} - \mu_{ij-1}) \left(\mu''_{ij} - \mu''_{ij-1}\right) - \left(\mu'_{ij} - \mu'_{i,j-1}\right)^2}{(\mu_{ij} - \mu_{i,j-1})^2} \right) z_{is} z_{it} \\ &= \boldsymbol{Z}^T \boldsymbol{W}_1 \boldsymbol{Z}\end{aligned}$$

for $s, t \in [1, \ldots, \dim(\boldsymbol{u})]$, where $\boldsymbol{W}_1$ is an $n \times n$ diagonal matrix with ith diagonal element

$$\sum_j y_{ij} \left(\frac{(\mu_{ij} - \mu_{ij-1}) \left(\mu''_{ij} - \mu''_{ij-1}\right) - \left(\mu'_{ij} - \mu'_{i,j-1}\right)^2}{(\mu_{ij} - \mu_{i,j-1})^2} \right).$$

For the XFA model (5.9), a singular $\boldsymbol{G}$ is obtained which cannot be inverted. The solution adopted here is to absorb the factor loadings λ_k into $\boldsymbol{Z}$. Let

$$\boldsymbol{u}_1 = [u_{R1}, \ldots u_{R4}, S(\boldsymbol{\ell}_{11}), \ldots, S(\boldsymbol{\ell}_{24,40})]^T$$

be the non-variety random effects and $\boldsymbol{u}_2 = [u^*_{V1}, \ldots, u^*_{V182}]^T$ be the variety factor scores. Let $\boldsymbol{Z}_1$ be the corresponding design matrix corresponding to $\boldsymbol{u}_1$ with i, kth element z_{1ik} for the ith observation and kth random effect and $\boldsymbol{Z}_2$ for $\boldsymbol{u}_2$ with i, j, kth element $z_{2ijk} = \lambda_j$ for the ith observation, jth cutoff ($j = 1, \ldots, 8$) and kth random effect. The second derivatives are

$$\frac{\partial^2 h_{Y|U}}{\partial u_{1s} \partial u_{1t}} = \sum_i \sum_j y_{ij} \left(\frac{(\mu_{ij} - \mu_{ij-1}) \left(\mu''_{ij} - \mu''_{ij-1}\right) - \left(\mu'_{ij} - \mu'_{i,j-1}\right)^2}{(\mu_{ij} - \mu_{i,j-1})^2} \right) z_{1is} z_{1it},$$

$$\begin{aligned}\frac{\partial^2 h_{Y|U}}{\partial u_{1s} \partial u_{2t}} &= \sum_i \sum_j y_{ij} \Bigg(\frac{(\mu_{ij} - \mu_{ij-1}) \left(\mu''_{ij} z_{2ijt} - \mu''_{ij-1} z_{2i,j-1,t}\right)}{(\mu_{ij} - \mu_{i,j-1})^2} \\ &\quad - \frac{\left(\mu'_{ij} z_{2ijt} - \mu'_{i,j-1} z_{2i,j-1,t}\right) \left(\mu'_{ij} - \mu'_{i,j-1}\right)}{(\mu_{ij} - \mu_{i,j-1})^2} \Bigg) z_{1is},\end{aligned}$$

and

$$\frac{\partial^2 h_{Y|U}}{\partial u_{2s}\partial u_{2t}} = \sum_i \sum_j y_{ij} \left(\frac{(\mu_{ij} - \mu_{ij-1}) \left(\mu''_{ij} z_{2ijs} z_{2ijt} - \mu''_{ij-1} z_{2i,j-1,s} z_{2i,j-1,t} \right)}{(\mu_{ij} - \mu_{i,j-1})^2} - \frac{\left(\mu'_{ij} z_{2ijs} - \mu'_{i,j-1} z_{2i,j-1,s} \right) \left(\mu'_{ij} z_{2ijt} - \mu'_{i,j-1} z_{2i,j-1,t} \right)}{(\mu_{ij} - \mu_{i,j-1})^2} \right).$$