Inference for generalised linear mixed models with sparse structure

by

Helen Elizabeth Ogden

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Department of Statistics

February 2014

THE UNIVERSITY OF WARWICK

Contents

List of Tables

List of Figures

Acknowledgments

I would like to thank my supervisor, David Firth, for giving me advice when I needed it, for allowing me the time and freedom to develop my own ideas, and for careful reading of this thesis.

I am grateful to Allesandra Salvan for allowing me to visit the Department of Statistics at the University of Padua in November 2012, and to Cristiano Varin for his hospitality during this visit, and for many useful discussions. The ADMB software used for comparison in Chapter 4 was introduced to me by Cristiano Varin and Manuela Cattelan at this time.

My thanks go to the many friends I have made here at Warwick over the course of my PhD, for making this such an enjoyable time. I am hugely grateful to my family, and in particular my husband Tom, for many years of love, support and encouragement.

I acknowledge funding from the Engineering and Physical Sciences Research Council.

Declarations

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree.

The work presented was carried out by the author except where explicitly indicated otherwise by references.

Abstract

The likelihood for the parameters of a generalised linear mixed model involves an integral which may be of very high dimension. Because of this apparent intractability, many alternative methods have been proposed for inference in these models, but it is shown that all can fail when the model is sparse, in that there is only a small amount of information available on each random effect.

The sequential reduction method developed in this thesis seeks to fill in this gap, by exploiting the dependence structure of the posterior distribution of the random effects to reduce dramatically the cost of approximating the likelihood in models with sparse structure. Examples are given to demonstrate the high quality of the new approximation relative to the available alternatives.

Finally, robustness of various estimators to misspecification of the random-effect distribution is considered. It is found that certain marginal composite likelihood estimators are not robust to such misspecification in situations in which the full maximum likelihood estimator is robust, providing a counterexample to the notion that composite likelihood estimators will always be at least as robust as the maximum likelihood estimator under model misspecification.

Chapter 1

Introduction

When constructing a regression model, it is rarely believed that the covariates recorded for each item really contain all the information about that item which might affect the distribution of the response. This may be acknowledged in the model by the addition of an error term in the regression for each item on its covariates. In the case that a single, separate, observation is made on each item, it is not possible to distinguish the error in the regression for each item from the uncertainty in the response given the predictor resulting from that regression. However, if multiple observations are made on some items, it is important to include an item-specific error term in the model. A model including these extra error terms, called random effects, is known as a mixed model.

If the original regression takes the form of a generalised linear model, then the addition of random effects leads to a generalised linear mixed model. For concreteness, it is these models which are studied in this thesis, although many of the same ideas could be transferred to more general mixed models. Chapter 2 gives a more in-depth introduction to generalised linear mixed models, including as an example a class of models for competitions between pairs of players, which are used as the primary example throughout the thesis.

It is easy to write down the conditional likelihood of the parameters given the value of the random effects. However, the random effects are treated as random variables, not fixed quantities, so in order to find the marginal likelihood, we must integrate this conditional likelihood over the assumed distribution of the random effects. Except in a few special cases, this integral has no analytical form, and direct numerical integration is not computationally feasible if there are a large number of random effects. For this reason, the likelihood is often referred to as intractable, except in cases in which the model has particularly simple structure, when it is

possible to simplify the integral.

In Chapter 2, a review of some of the alternative methods for inference in generalised linear mixed models is given. One class of approaches involves replacing the likelihood with some approximation, for example using Laplace's method or importance sampling. However, these approximations can fail badly in cases where the structure of the model is sparse, in that only a small amount of information is available on each random effect, especially when the response is discrete, and may only take a small number of values. The poor quality of approximation can result in estimators with very poor statistical properties. Composite likelihood methods for inference are also considered, and are shown to have low efficiency in many cases. Chapter 3 considers in more detail the situations in which each alternative method of inference may be expected to perform well, and those in which they will perform badly. A gap is discovered for some models with sparse structure, for which none of the currently available methods appears to work well.

Chapter 4 introduces a new method of approximating the likelihood, called the sequential reduction method, which is designed to bridge this gap. The main idea of the method is to fully exploit the structure of the model in order to simplify the computation of the likelihood. This is done using ideas from the field of graphical models, which are introduced at the start of the chapter. The new method is compared with existing approximation methods through numerical examples, and is found to perform well in many cases. In particular, it is shown that the sequential reduction method offers a great improvement over importance sampling approximations for models with sparse structure.

Chapters 2-4 of the thesis consider the question of how best to conduct inference in a generalised linear mixed model, under the assumption that the model is correctly specified. In Chapter 5, this assumption is partially dropped, and the robustness of various estimators to misspecification of the random-effects distribution is studied. It is shown that estimators from composite likelihoods constructed using small blocks of components are inconsistent under misspecification of the random-effects distribution in an asymptotic setting in which the maximum likelihood estimator is consistent. A major motivation for the use of composite likelihoods is their potential for increased robustness compared to the full likelihood, so this result is important as a warning that such increased robustness cannot be guaranteed. The estimator from the sequential reduction approximation to the likelihood inherits the same robustness as the true maximum likelihood estimator, thereby providing another justification for its use over the available alternatives.

Chapter 2

Generalised linear mixed models

2.1 Generalised linear mixed models

2.1.1 The linear mixed model

Suppose that we wish to construct a regression model for a response Y_i measured on an item i, based on some covariates $\mathbf{x}_i$ observed on that item. The simplest and most well-studied of all such regression models is the linear model, in which it is assumed that

$$Y_i = \beta^T \mathbf{x}_i + \epsilon_i,$$

where $\epsilon_i \sim N(0, \tau^2)$, and β and τ are unknown parameters. Writing $\eta_i = \beta^T \mathbf{x}_i$, so that we model

$$Y_i = \eta_i + \epsilon_i,$$

we could think of η_i as some underlying quantity about item i which is measured, with error, by Y_i.

Now suppose that instead of making a single observation on each item i, we make m_i repeated observations

$$Y_{ij},\ i = 1, \ldots, n,\ j = 1, \ldots, m_i.$$

Thinking of each Y_{ij} as an imperfect measurement of η_i, we could model

$$Y_{ij} = \eta_i + \epsilon_{ij},$$

where $\epsilon_{ij} \sim N(0, \tau^2)$. If we still assume that

$$\eta_i = \beta^T \mathbf{x}_i,$$

then the value of the observed covariate entirely determines the distribution of the response. Put another way, if two items i and j have the same covariate values, then the responses for those items must have the same distribution.

This assumption is clearly unrealistic: there will undoubtedly be information relevant to the response which is not encoded by the observed covariates. A more realistic model should acknowledge this extra uncertainty, letting

$$\eta_i = \beta^T \mathbf{x}_i + b_i,$$

where b_i represents the error in predicting η_i from the observed covariates. A common assumption is that b_i is also a normal error term, so that $b_i \sim N(0, \sigma^2)$ for some unknown parameter σ.

In the case that $m_i = 1$, so that there is only one observation on each item,

$$Y_i = \beta^T \mathbf{x}_i + b_i + \epsilon_i,$$

where $b_i \sim N(0, \sigma^2)$ and $\epsilon_i \sim N(0, \tau^2)$, and b_i and ϵ_i are independent. There is then no way to distinguish the error b_i in predicting η_i given $\mathbf{x}_i$, and the observation error ϵ_i. We may combine the two errors, to give $e_i = b_i + \epsilon_i \sim N(0, \sigma^2 + \tau^2)$. The parameters (σ, τ) are not separately identifiable. Instead, it is only possible to identify the combination $\omega^2 = \sigma^2 + \tau^2$. So in fact,

$$Y_i = \eta_i + e_i,$$

where $e_i \sim N(0, \omega^2)$, and we have returned to the original linear model.

The case of repeated observations on each item is not the only one in which it is important to recognise the heterogeneity between items with the same covariate value. Section 2.2 gives some more examples of settings in which it is possible to detect the difference between the two types of error. In all of these settings, it is important to include the extra error term b_i, usually called a random effect, in the model.

Writing the linear model in vector form, we have

$$\mathbf{Y} = X\beta + \epsilon,$$

where X is a known design matrix, and $\epsilon_i \sim N(0, \tau^2)$. To extend this to allow heterogeneity between items with the same covariate values, we write

$$\mathbf{Y} = X\beta + Z\mathbf{b} + \epsilon,$$

where Z is a known design matrix for the random effects, $\mathbf{b} \sim N(0, D(\psi))$, and ψ is an unknown parameter. This extension of the linear model, to include random effects $\mathbf{b}$ as well as fixed effects β, is known as a linear mixed model. West *et al.* (2006) give a brief history of the development of the linear mixed model. It turns out that full likelihood inference for the linear mixed model remains relatively straightforward, as discussed in Section 2.3.2.

2.1.2 Mixed models in general

The same arguments for the addition of random effects to the linear model apply equally to any regression model, and lead to the definition of mixed models in a more general setting. The class of generalised linear models provide a natural extension to the linear model, and are widely used in practice. We now consider the addition of random effects to generalised linear models, which will lead us to define the generalised linear mixed model. For concreteness, the focus of this thesis will be these generalised linear mixed models, although many of the ideas are also applicable to mixed models outside of this framework.

2.1.3 The generalised linear mixed model

A generalised linear model (Nelder & Wedderburn, 1972) allows the distribution of a response $\mathbf{Y} = (Y_1, \ldots, Y_m)$ to depend on observed covariates through a linear predictor η, where

$$\eta = X\beta,$$

for some known design matrix X. Conditional on knowledge of the linear predictor, and possibly an unknown dispersion parameter τ, the components of $\mathbf{Y}$ are independent, and the distribution of $\mathbf{Y}$ is fixed. The distribution of $\mathbf{Y}$ is assumed to have exponential family form, with mean

$$\mu = \mathbb{E}(\mathbf{y}|\eta) = g^{-1}(\eta),$$

for some known link function $g(.)$.

As in the linear model, an assumption implicit in the generalised linear model is that the distribution of the response is entirely determined by the values of the observed covariates. In practice, this assumption is rarely believed: in fact, there may be other information not encoded in the observed covariates which may affect the response. If multiple observations are made on some items, or if each observation involves more than one item, it is important to allow for this extra heterogeneity.

A generalised linear mixed model does this by modelling the linear predictor as

$$\eta = X\beta + Z\mathbf{b}, \tag{2.1}$$

where X and Z are known design matrices, and $\mathbf{b}$ is a sample from a distribution known up to a parameter vector ψ. In most cases, it is assumed that $\mathbf{b} \sim N_n(0, D(\psi))$, and some methods rely on this assumption. The idea of adding random effects to the the linear predictor in particular examples of generalised linear models first appeared in the 1980s, for example for logistic regression for binary data in Williams (1982) and for log-linear regression for Poisson data in Breslow (1984). Schall (1991) discusses the addition of random effects to an arbitrary generalised linear model.

There are only a relatively small number of named multivariate distributions to choose from for the distribution of $\mathbf{b}$. Instead, non-normal $\mathbf{b}$ could be constructed by taking

$$\mathbf{b} = A(\psi)\mathbf{u}, \tag{2.2}$$

where the components of $\mathbf{u}$ are independent of one another, and may have any univariate distribution $F_U(., \psi)$, possibly depending on the unknown parameter ψ. Combining (2.1) and (2.2), we write

$$\eta = X\beta + Z(\psi)\mathbf{u}, \tag{2.3}$$

where $u_i \sim F_U(., \psi)$ and $Z(\psi) = ZA(\psi)$. This thesis concentrates on the case in which $u_i \sim N(0, 1)$, which allows $\mathbf{b}$ to have any multivariate normal distribution with mean zero. The techniques described in Chapter 4 could be extended easily for use with other random effect distributions of the form (2.2).

The non-zero elements of the columns of $Z(\psi)$ give us the observations which involve each random effect. We will say the generalised linear mixed model has 'sparse structure' if most of these columns have few non-zero elements, so that most random effects are only involved in a few observations. These sparse models are particularly problematic for inference, especially when the response may only take a small number of values, because the amount of information available on each random effect is small.

2.2 Examples of generalised linear mixed models

2.2.1 Models with nested structure

Suppose that observations are recorded on items which are clustered into groups, so that we have m_i observations for each group $i = 1, \ldots, n$. Consider the model

$$\eta_{ij} = \beta_0 + \beta_1 x_{ij}^{(1)} + \ldots + \beta_p x_{ij}^{(p)} + b_{i0},$$

where ij denotes the jth item in group i, for $j = 1, \ldots, m_i$, $i = 1, \ldots n$, and $b_{i0} \sim N(0, \sigma^2)$ is a group-level random effect. The addition of b_{i0} to the model allows for the fact that there may be some error in prediction of η_{ij} from the observed covariates. The items contained within each group may share characteristics which are not observed, and b_{i0} may be thought of as representing those unobserved shared characteristics of group i. Write $m = \sum_{i=1}^{n} m_i$ for the total number of items. Written in vector form,

$$\eta = X\beta + Z(\sigma)\mathbf{u},$$

where X is an $m \times (p+1)$ matrix with rows

$$X_r = (1, x_{i_r j_r}^{(1)}, \ldots, x_{i_r j_r}^{(p)})$$

where r is the j_rth item in group i_r. The $m \times n$ matrix $Z(\sigma)$ has components

$$Z_{rs}(\sigma) = \begin{cases} \sigma & \text{if } i_r = s \\ 0 & \text{otherwise} \end{cases}. \tag{2.4}$$

This model is called a two-level random intercept model. This is a sparse model if the number of observations per group, m_i, is small for most i.

In a three-level model, the groups themselves are clustered within larger groups. Models with even more levels can be built by repeatedly clustering the top-level group within larger groups. It is also possible to allow for interaction between the observed covariates and the random effects, leading to a model with random slopes, in addition to a random intercept. These are all examples of a wider class of multilevel models, and many examples of such models may be found in Goldstein (2011).

2.2.2 Pairwise competition models

Consider a tournament among n players, consisting of contests between pairs of players. Let y_{ij} record the outcome of a contest between players i and j. We suppose that each player i has some ability λ_i, and that conditional on all the abilities, the outcomes Y_{ij} are independent, with distribution depending on the difference in abilities of the players i and j, so that

$$\mathbb{E}(Y_{ij}|\lambda) = g^{-1}(\lambda_i - \lambda_j)$$

for some link function $g(.)$.

The pairwise competition models used as examples in this thesis all have binary outcomes, so we observe only which player wins each contest. We consider these binary models because each observation only provides a small amount of information about the random effects, so approximations to the likelihood are most likely to fail. If $g(x) = \text{logit}(x)$, then this describes a Bradley-Terry model (Bradley & Terry, 1952). If $g(x) = \Phi^{-1}(x)$ (the probit link), then it describes a Thurstone-Mosteller model (Thurstone (1927), Mosteller (1951)).

If covariate information $\mathbf{x}_i$ is available for each player, then interest may lie in the effect of the observed covariates on ability, rather than the individual abilities λ_i themselves. For example, Whiting *et al.* (2006), conducted an experiment to determine the effect of covariates on the fighting ability of Augrabies flat lizards, *Platysaurus broadleyi*. The scientific hypothesis of interest was whether the spectrum of the throat of the lizard had an effect on fighting ability. To investigate this, Whiting *et al.* (2006) captured $n = 77$ lizards, recorded various measurements on each, and then released them and recorded the outcomes of fights between pairs of animals.

To model situations of this sort, suppose that the ability of a player may be modelled as a linear function of their covariates, plus an error term, so that

$$\lambda_i = \beta^T \mathbf{x}_i + b_i$$

where the b_i are independent samples from a $N(0, \sigma^2)$ distribution.

This may be written in generalised linear mixed model form by specifying that $\mathbb{E}(Y_{ij}) = g^{-1}(\eta_{ij})$, where

$$\eta_{ij} = \lambda_i - \lambda_j = \beta^T(\mathbf{x}_i - \mathbf{x}_j) + b_i - b_j. \tag{2.5}$$

To write (2.5) in the form (2.3), we let X be an $m \times p$ matrix with components

$$X_{rs} = x_{p_1(r)s} - x_{p_2(r)s}$$

where $p_1(r)$ gives the first player involved in contest r, and $p_2(r)$ the second, and where x_{is} is the sth component of the vector of observed covariates for player i. Let $\bar{Z}$ be an $m \times n$ matrix with components

$$\bar{Z}_{rs} = \begin{cases} 1 & \text{if } p_1(r) = s \\ -1 & \text{if } p_2(r) = s \\ 0 & \text{otherwise} \end{cases},$$

and write $Z(\sigma) = \sigma \bar{Z}$. Then

$$\eta = X\beta + Z(\sigma)\mathbf{u},$$

where $\mathbf{u} \sim N_n(0, I)$.

Notice that the non-zero components of the sth column give the contests involving player s, so the tournament will have sparse structure if most players only compete in only a small number of contests.

Such a model for the flat-lizards tournament gives us an example of a generalised linear mixed model with sparse structure, since we only observe a total of 100 contests among the 77 lizards.

2.2.3 Other pairwise interaction models

There are many other models with a similar structure to these pairwise competition models, in that the outcome is determined by some interaction between pairs of items. In fact, many of the generalised linear mixed models which have been noted to have intractable likelihoods fall into this class. For example, McCullagh & Nelder (1989) give data gathered on pairs of salamanders from two separate regions. The experiment consisted of attempts to mate each salamander with other salamanders from the same region, and with salamanders from the other region. The aim was to determine whether two salamanders from different regions were less likely to mate with one another than two salamanders from the same region.

In this case, the linear predictor is determined by sum of the individual effects of each salamander, plus a term for the interaction between the region of the salamanders. For instance, if we attempt to mate male i with female j, the linear

predictor might be given by

$$\eta_{ij} = \beta^T x_{ij} + b_i^M + b_j^F,$$

where $b_i^M \sim N(0, \sigma_M^2)$ and $b_j^F \sim N(0, \sigma_F^2)$, and $\mathbf{x}_{ij}$ contains the interaction between the regions of the two salamanders. Written in vector form,

$$\eta = X\beta + Z(\sigma_M, \sigma_F)\mathbf{u},$$

where

$$X_r = \mathbf{x}_{m(r)f(r)},$$

where $m(r)$ and $f(r)$ are the male and female salamanders involved in mating attempt r, and

$$Z_{rs}(\sigma_M, \sigma_F) = \begin{cases} \sigma_M & \text{if } m(r) = s \\ \sigma_F & \text{if } f(r) = s \\ 0 & \text{otherwise} \end{cases}.$$

2.3 Inference in generalised linear mixed models

2.3.1 The likelihood

Let $f(.|\eta_i)$ be the density of Y_i, conditional on knowledge of the value of η_i. Conditional on η, the components of $\mathbf{Y}$ are independent, so that

$$L(\beta, \psi|\mathbf{y}) = \int_{\mathbb{R}^n} \prod_{i=1}^{m} f\left(y_i|\eta_i = X_i^T\beta + Z_i(\psi)^T\mathbf{u}\right) \prod_{j=1}^{n} \phi(u_j)du_j, \tag{2.6}$$

where X_i is the ith row of X, and $Z_i(\psi)$ is the ith row of $Z(\psi)$.

By using a product of K-point quadrature rules, an n-dimensional integral may be approximated at cost $O(K^n)$, where the error in the approximation tends to 0 as $K \to \infty$. Unless n is very small, it will therefore not be possible to approximate the likelihood well by direct computation of this integral using a product of quadrature rules. However, while the likelihood may always be written in form (2.6), there are some occasions where it may be simplified, so that computation of an n-dimensional integral is not necessary. Example 3.1 demonstrates such a simplification in one particular pairwise competition model. The sequential reduction method developed in Chapter 4 gives a systematic way to check for such simplifications to the likelihood.

2.3.2 Laplace approximation to the likelihood

Many alternative methods of inference work by replacing the likelihood with some approximation to it. The success of these methods depends on the quality of that approximation. The Laplace approximation is a particularly simple approximation method.

Write

$$g(u_1, \ldots, u_n|\mathbf{y}, \beta, \psi) = \prod_{i=1}^{m} f\left(y_i|\eta_i = X_i^T\beta + Z_i(\psi)^T\mathbf{u}\right) \prod_{j=1}^{n} \phi(u_j)$$

for the integrand of the likelihood. This may be thought of as a non-normalised version of the posterior density for $\mathbf{u}$, given $\mathbf{y}$, β and ψ.

Pinheiro & Bates (1995) suggest using a Laplace approximation to this integral. For each fixed $\theta = (\beta, \psi)$, the Laplace approximation approach relies on a normal approximation to the posterior density of $\mathbf{u}$, given $\mathbf{y}$ and θ. To find this normal approximation, let $\hat{\mathbf{u}}_\theta$ maximise

$$\log g(\mathbf{u}|\mathbf{y}, \theta)$$

over $\mathbf{u}$, and let $H_\theta = H_\theta(\hat{\mathbf{u}})$ be the Hessian resulting from this optimisation, where

$$H_\theta(\mathbf{u}) = \nabla_{\mathbf{u}}\nabla_{\mathbf{u}}^T \log g(\mathbf{u}|\mathbf{y}, \theta).$$

The normal approximation to $g(.|\mathbf{y}, \theta)$ will be proportional to a $N_n(\hat{\mathbf{u}}_\theta, H_\theta^{-1})$ density. Writing $g^{\mathrm{na}}(.|\mathbf{y}, \theta)$ for the normal approximation to $g(.|\mathbf{y}, \theta)$,

$$g^{\mathrm{na}}(\mathbf{u}|\mathbf{y}, \theta) = \frac{g(\hat{\mathbf{u}}_\theta|\mathbf{y}, \theta)}{\phi_n(\hat{\mathbf{u}}_\theta; \hat{\mathbf{u}}_\theta, H_\theta^{-1})}\phi_n(\mathbf{u}; \hat{\mathbf{u}}_\theta, H_\theta^{-1}),$$

where we write $\phi_n(.; \mu, \Sigma)$ for the $N_n(\mu, \Sigma)$ density.

When we integrate over $\mathbf{u}$, only the normalising constant remains, so that

$$\tilde{L}(\theta|\mathbf{y}) = \frac{g(\hat{\mathbf{u}}_\theta|\mathbf{y}, \theta)}{\phi_n(\hat{\mathbf{u}}_\theta; \hat{\mathbf{u}}_\theta, H_\theta^{-1})} = (2\pi)^{-\frac{n}{2}}(\det H_\theta)^{-\frac{1}{2}} g(\hat{\mathbf{u}}_\theta|\mathbf{y}, \theta).$$

In the case of a linear mixed model,

$$g(\mathbf{u}|\mathbf{y}, \theta) = \prod_{i=1}^{m} \phi\left(\frac{X_i^T\beta + Z_i(\psi)^T\mathbf{u} - y_i}{\tau}\right) \prod_{j=1}^{n} \phi(u_j),$$

which is proportional to a normal density. So in this case, the approximating nor-

mal density is precise, and there is no error in the Laplace approximation to the likelihood. In other cases, and particularly when the response is discrete and may only take a few values, the error in the Laplace approximation can be large.

Validity of the Laplace approximation

Recall that we have n random effects, and m observations, so that the likelihood may be written as an n-dimensional integral over a integrand containing a product of m terms. In the case that n is fixed, and $m \to \infty$, the relative error in the Laplace approximation may be shown to tend to zero. However, in the type of model we consider here, n is not fixed, but grows with m. The validity of the Laplace approximation depends upon the rate of this growth. Shun & McCullagh (1995) study this problem, and conclude that the Laplace approximation should be reliable, provided that $n = o(m^{1/3})$. If n grows with m more quickly than $o(m^{1/3})$, the relative error in the Laplace approximation may be $O(1)$. These rates are rather slower than those which are typical of the type of models considered here, in which a sparse model may have $n = O(m)$ (see Example 3.2), and where we consider a model with $n = O(m^{1/2})$ to have dense structure (see Example 3.5).

However, the Laplace approximation to the difference in the log-likelihood at two nearby points tends to be much more accurate than the approximation to the log-likelihood itself, and in denser models, where there is more information available per random effect, the Laplace approximation to the shape of the likelihood appears to be sufficiently good to give accurate inference, even in cases where the approximation to the likelihood itself has large relative error. See Example 3.5 for a demonstration of this, in a situation with $n = O(m^{1/2})$. The effect that ratios of Laplace approximations to similar functions tend to be more accurate than each Laplace approximation individually has been noted before, for example by Tierney & Kadane (1986) in the context of computing posterior moments. However, in models with very sparse structure, even the shape of the Laplace approximation may be inaccurate, so another method is required. The behaviour of the Laplace approximation in models with differing levels of sparsity is studied in Chapter 3.

Other methods based on the Laplace approximation

Many alternative methods for inference in a generalised linear mixed model are based on a Laplace approximation to the likelihood. For instance, Breslow & Clayton (1993) base their Penalised Quasi Likelihood (PQL) on a further approximation to the Laplace approximation to the likelihood. It is well acknowledged that PQL

is not suitable for inference in models for binary data, with a small number of observations for each random effect. Browne & Draper (2006) consider an example of such a model, and show that PQL gives highly biased estimators and confidence intervals with low coverage in that case.

2.3.3 Importance sampling approximation to the likelihood

Recall that the likelihood may be written as an integral

$$L(\theta|\mathbf{y}) = \int_{\mathbb{R}^n} g(\mathbf{u}|\mathbf{y}, \theta) d\mathbf{u}, \tag{2.7}$$

where $g(.|\mathbf{y}, \theta)$ may be approximated by a function $g^{\text{na}}(.|\mathbf{y}, \theta)$ proportional to the density of a $N_n(\mu_\theta, \Sigma_\theta)$ distribution. Replacing $g(.|\mathbf{y}, \theta)$ with $g^{\text{na}}(.|\mathbf{y}, \theta)$ in (2.7) yields the Laplace approximation to the likelihood.

In cases where the Laplace approximation fails, Pinheiro & Bates (1995) suggest constructing an importance sampling approximation to (2.7), based on samples from the approximating normal distribution $N_n(\mu_\theta, \Sigma_\theta)$. Writing

$$w(\mathbf{u}; \theta) = \frac{g(\mathbf{u}|\mathbf{y}, \theta)}{\phi_n(\mathbf{u}; \mu_\theta, \Sigma_\theta)},$$

the likelihood may be written as

$$\begin{aligned} L(\theta|\mathbf{y}) &= \int_{\mathbb{R}^n} w(\mathbf{u}; \theta) \phi_n(\mathbf{u}; \mu_\theta, \Sigma_\theta) d\mathbf{u} \\ &= \mathbb{E}\left[w(\mathbf{U}; \theta)\right], \end{aligned}$$

where $\mathbf{U} \sim N(\mu_\theta, \Sigma_\theta)$. This may be approximated by the importance sampling approximation

$$L^{IS}(\theta|\mathbf{y}) = \frac{1}{N} \sum_{i=1}^{N} w(\mathbf{u}^{(i)}; \theta),$$

where $\mathbf{u}^{(i)} \sim N(\mu_\theta, \Sigma_\theta)$.

By the weak law of large numbers,

$$L^{IS}(\theta|\mathbf{y}) \to^p L(\theta|\mathbf{y})$$

as $N \to \infty$. Provided that $\sigma_W^2 = \text{Var}\left(w(\mathbf{U}; \theta)\right) < \infty$, the central limit theorem implies that

$$\sqrt{N}\left(L^{IS}(\theta|\mathbf{y}) - L(\theta|\mathbf{y})\right) \to^d N(0, \sigma_W^2)$$

as $N \to \infty$. So, if the importance weights $w(\mathbf{U};\theta)$ have finite variance, the error in the importance sampling approximation shrinks at rate $\sqrt{N}$, irrespective of the dimension n of the integral.

Unfortunately, there is no guarantee that the variance of the importance weights will be finite. It is difficult to check this theoretically in most practical applications, but in some situations where the normal approximation to $g(.|\theta)$ is poor, it appears that the variance is not finite (see Example 3.2). In such a situation, the weak law of large numbers still holds, so the importance sampling approximation will still converge to the true likelihood, but the convergence may be slow and erratic, and estimates of the variance of the approximation may be unreliable.

Interest lies in approximating the likelihood surface, rather than just the likelihood at a single point θ. If a new random sample $\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(N)}$ from $N_n(\mu_\theta, \Sigma_\theta)$ is used for each θ, the resulting approximation to the likelihood surface will be quite rough. To ensure a smooth surface, Pinheiro & Bates (1995) suggest sampling $\mathbf{t}^{(i)} \sim N_n(0, I)$, and letting

$$\mathbf{u}^{(i)} = \mu_\theta + A_\theta \mathbf{t}^{(i)}$$

for each θ, where A_θ is the Cholesky decomposition of Σ_θ.

2.3.4 Composite likelihood

If we had only some small subset $\mathbf{y}_{(s)}$ of the observations, where $\mathbf{y}_{(s)}$ only depends on a small number, d_s, of random effects, then we could evaluate the likelihood $L(\theta|\mathbf{y}_{(s)})$ via an integral of dimension d_s.

The idea of a marginal composite likelihood (Lindsay, 1988) is to split the data up into blocks $\{\mathbf{y}_{(s)}\}_{s=1}^K$ so that the likelihood based on each block is reasonably easy to compute. A marginal composite likelihood based on these blocks may be defined as

$$L^C(\theta) = \prod_{s=1}^{K} L\left(\theta|\mathbf{y}_{(s)}\right)^{w_s},$$

where w_s is a weight given to block s.

For instance, for inference in a pairwise competition model, Cattelan & Varin (2010) suggest using the blocks $y_{ijk} = (y_{ij}, y_{ik})$ of pairs of contests which share a common player, with weights $w_s = 1$ for all s.

For an appropriate choice of blocks, this method is fairly fast, and results on the consistency and asymptotic normality of the estimators are available in some settings. A review of these results is given in Chapter 3. However, in many cases the marginal composite likelihood estimators will be much less efficient than the

maximum likelihood estimator. This is demonstrated in Example 3.1. It is also often difficult to decide which blocks of components $\mathbf{y}_{(s)}$ and weights w_s to use to construct the composite likelihood in any given setting. Cox & Reid (2004) and Lindsay *et al.* (2011) give some initial ideas on answering these questions, but there is currently no general-purpose method available for selection of good blocks of components and weights.

Composite likelihood is usually used mainly for computational reasons, but a secondary motivation is given by the fact that only the marginal distributions of the blocks used in the composite likelihood need to be specified, rather than requiring specification of the full joint distribution of the response vector. Varin *et al.* (2011) and Xu & Reid (2011) discuss the potential this gives for increased robustness of the composite likelihood estimator relative to the full maximum likelihood estimator. However, in Chapter 5 we find that in the context of generalised linear mixed models, certain marginal composite likelihood estimators will be less robust to misspecification of the random effects distribution than the full maximum likelihood estimator.

2.3.5 Bayesian inference

Given a prior distribution $\pi(\theta)$ for the parameter-vector θ, the posterior distribution of θ is given by

$$\pi(\theta|y) \propto L(\theta|\mathbf{y})\pi(\theta).$$

If we were able to obtain a good approximation to the likelihood, then this could be used to approximate the posterior distribution. However, in the case when no such approximation is readily available, Markov chain Monte Carlo methods may be used to find samples $\theta_{(i)}$ whose distribution converges to the posterior, once a sufficiently large number of samples have been taken. Zeger & Karim (1991) describe how such samples may be obtained by Gibbs sampling in the case of a generalised linear mixed model.

Such methods have their problems, primarily that it can sometimes take a long time for the convergence to the posterior distribution to occur, and it is difficult to check when this convergence has taken place.

To construct the posterior distribution, it is necessary to choose a prior distribution for θ. In models with binary response and sparse structure, the choice of prior distribution for the parameters of the random effects distribution often has a large effect on the posterior distribution. Care should be taken before using an improper prior in such a setting, since such a prior may result in an improper

posterior distribution (Hobert & Casella, 1996).

There is certainly much more to be said about Bayesian methods for inference for generalised linear mixed models, but they are not directly discussed further in this thesis. The focus is instead on approximations and alternatives to the likelihood, which may be used in place of the likelihood when conducting inference.

Chapter 3

Performance of existing methods of inference

3.1 Asymptotics under independent replication

3.1.1 Estimating equations

In the simple asymptotic framework of an increasing number, R, of independent replications, the maximum likelihood estimator is consistent and normally distributed in the limit as $R \to \infty$, as long as certain regularity conditions hold. Provided that analogues of these conditions continue to hold, and in particular that the parameter of interest is identifiable from the composite likelihood, the composite likelihood estimator is also consistent and asymptotically normal as $R \to \infty$. The variance of the limiting distribution of the composite likelihood estimator is larger than that of the maximum likelihood estimator, except in cases where the two estimators are identical.

These results apply in a general setting, not just in the case of generalised linear mixed models. They can be shown by applying the theory of estimating equations, which is now reviewed. These results will be applied several times in the thesis, for example to provide results on the asymptotic behaviour of composite likelihood estimators, and of estimators found by maximising an approximation to the likelihood, rather than the likelihood itself.

Suppose that we have R independent, identically distributed observations $\mathbf{y}_{(1)}, \ldots, \mathbf{y}_{(R)}$, which we model as being samples from a distribution depending on a p-dimensional parameter θ, taking values in some set Θ. Suppose that $\hat{\theta}_R$ is the

estimator of θ solving the estimating equation

$$v_R(\theta, \mathbf{y}) = \sum_{i=1}^{R} v_{(1)}(\theta, \mathbf{y}_{(i)}).$$

For instance, in the case of full likelihood, $v_{(1)}(\theta, \mathbf{y}_{(i)})$ would be the score function associated with observing $\mathbf{y}_{(i)}$, and $v_R(\theta, \mathbf{y})$ the overall score function for $\mathbf{y}$. Let

$$\bar{v}(\theta) = \mathbb{E}[v_{(1)}(\theta; \mathbf{Y})],$$

where the expectation is taken over the true distribution of each $\mathbf{Y}_{(i)}$.

Theorem 1. *Suppose that $\bar{v}(.)$ is a continuous function of θ, with unique root θ^*. Suppose that Θ is compact, that $v_{(1)}(\theta, \mathbf{y})$ is a continuous function of θ for each $\mathbf{y}$, and that*

$$\left| v_{(1)}(\theta, \mathbf{y}) \right| \leq g(\mathbf{y})$$

for some integrable function g, for all θ in a neighbourhood of θ^*. Then

$$\hat{\theta}_R \to^p \theta^*$$

as $R \to \infty$, where $\to^p$ denotes convergence in probability.

Proof. See van der Vaart (1998, pp. 44–46)

With a few extra conditions, results on the asymptotic normality of $\hat{\theta}_R$ may also be obtained.

Let

$$H(\theta) = \mathbb{E}\left[-\nabla_\theta v_{(1)}(\theta, \mathbf{Y})\right]$$

and

$$J(\theta) = \mathbb{E}\left[v_{(1)}(\theta, \mathbf{Y}) \left[v_{(1)}(\theta, \mathbf{Y})\right]^T\right].$$

Theorem 2. *Suppose that the conditions assumed in Theorem 1 hold, and in addition that $v_{(1)}(\theta, \mathbf{y})$ has continuous second derivatives for each $\mathbf{y}$, and*

$$\left| \frac{\partial^2 v_{(1)}(\theta, \mathbf{y})}{\partial \theta_i \partial \theta_j} \right| \leq h(\mathbf{y})$$

for some integrable function h, for all θ in a neighbourhood of θ^*. Then

$$\sqrt{R}(\hat{\theta}_R - \theta^*) \to^d N_p(0, H(\theta^*)^{-1} J(\theta^*) \left[H(\theta^*)^T\right]^{-1})$$

as $R \to \infty$, where $\to^d$ denotes convergence in distribution.

Proof. See van der Vaart (1998, pp. 51–52)

Note. The regularity conditions given in the statement of Theorems 1 and 2 are stronger than are needed. See van der Vaart (1998, pp. 46–47, 52–53) for some weaker conditions which could be used instead.

3.1.2 Maximum likelihood estimator

We may show the consistency and asymptotic normality of the maximum likelihood estimator $\hat{\theta}_R$ by application of these results to the score function

$$u(\theta|\mathbf{y}) = \nabla_\theta \ell(\theta|\mathbf{y}).$$

Let

$$\bar{u}(\theta) = \mathbb{E}[u(\theta|\mathbf{Y})].$$

Suppose that the model for the data is correct, so that $\mathbf{y}_{(1)}, \ldots, \mathbf{y}_{(R)}$ really are samples from the assumed distribution, for some value θ_0 of θ. Assuming sufficient further regularity conditions to allow interchange of differentiation and integration, the first Bartlett identity

$$\bar{u}(\theta_0) = 0 \tag{3.1}$$

may then be shown to hold. Assuming that this root is unique, and that the other regularity conditions of Theorems 1 and 2 hold with $v_{(1)}(\theta, \mathbf{y}) = u(\theta|\mathbf{y})$, we obtain that

$$\hat{\theta}_R \to^p \theta_0,$$

so the maximum likelihood estimator is consistent, and

$$\sqrt{R}(\hat{\theta}_n - \theta_0) \to^d N_p(0, H(\theta_0)^{-1} J(\theta_0) \left[H(\theta_0)^T\right]^{-1}) \tag{3.2}$$

as $R \to \infty$. In this case, the second Bartlett identity gives that $H(\theta_0) = J(\theta_0)$, and we call this quantity the Fisher information matrix, denoted by $I(\theta_0)$. Equation (3.2) then simplifies to the familiar

$$\sqrt{R}(\hat{\theta}_n - \theta_0) \to^d N_p(0, I(\theta_0)^{-1}).$$

3.1.3 Composite likelihood estimator

The results can also be applied to show consistency and asymptotic normality of the composite likelihood estimator $\hat{\theta}_R^C$. Writing

$$u^C(\theta|\mathbf{y}) = \nabla_\theta \ell^C(\theta|\mathbf{y}).$$

for the composite score function, let

$$\bar{u}^C(\theta) = \mathbb{E}[u^C(\theta|\mathbf{Y})].$$

Note that

$$\begin{aligned}\bar{u}^C(\theta) &= \mathbb{E}\left[\nabla_\theta \ell^C(\theta;\mathbf{Y})\right] \\ &= \mathbb{E}\left[\sum_{s=1}^K \nabla_\theta \ell(\theta;\mathbf{Y}_{(s)})\right] \\ &= \sum_{s=1}^K \mathbb{E}\left[\nabla_\theta \ell(\theta;\mathbf{Y}_{(s)})\right].\end{aligned}$$

So if the model is correct with $\theta = \theta_0$, then $u^C(\theta_0) = 0$, since the expectation of each score in the sum is 0, by (3.1). If this solution is unique, and the other regularity conditions of Theorems 1 and 2 hold with $v_{(1)}(\theta, \mathbf{y}) = u^C(\theta|\mathbf{y})$, then

$$\hat{\theta}_R^C \to^p \theta_0,$$

and

$$\sqrt{R}(\hat{\theta}_R^C - \theta_0) \to^d N_p(0, H_C(\theta_0)^{-1} J_C(\theta_0) \left[H_C(\theta_0)^T\right]^{-1})$$

as $R \to \infty$. This time, there is no simplification in the variance term, since the second Bartlett equality does not hold. We write $G_C(\theta_0) = H_C(\theta_0)^T J_C(\theta_0)^{-1} H_C(\theta_0)$. $G_C(.)$ is known as the Godambe or sandwich information matrix. The asymptotic variance of the composite likelihood estimator is in general larger than that of the full likelihood estimator. In the scalar-parameter case $p = 1$, it is sometimes useful to consider the **asymptotic relative efficiency**

$$r = \frac{\text{avar}(\hat{\theta}_R)}{\text{avar}(\hat{\theta}_R^C)} = \frac{[H_C(\theta_0)]^2}{I(\theta_0) J_C(\theta_0)} \in (0, 1].$$

Example 3.1. To demonstrate these asymptotic results in the context of the generalised linear mixed model, we consider a very simple binary pairwise competition

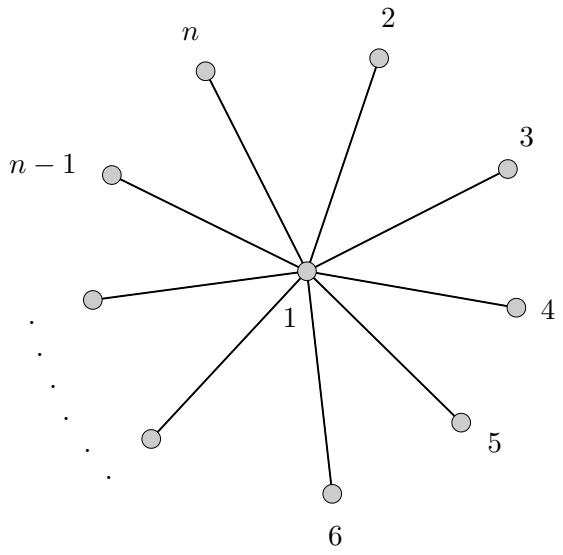


Figure 3.1: A star tournament on n players

model, in which the ability of a player is modelled as

$$\lambda_i = \sigma u_i,$$

where $u_i \sim N(0,1)$, and σ is the unknown parameter of interest. That is, we are interested in the overall spread of abilities among players. Conditional on the abilities, we assume that $Pr(i \text{ beats } j) = \Phi(\lambda_i - \lambda_j)$.

This model is simpler than most pairwise competition models of practical interest, in which we model the effect of some observed covariates on the ability of a player. However, because of the simplicity of this model, it is relatively easy to study some statistical properties of the various estimators considered in Section 2.3. If an estimator of σ in this simple model has poor properties, there is little hope that the corresponding estimator will be successful in a more complex model.

In order for the asymptotic regime described above to apply, we assume that the tournament consists of R repetitions of a smaller tournament with fixed structure, and consider what happens in the limit at $R \to \infty$. We assume that the small tournaments each have an 'star' structure, as shown in Figure 3.1. We allow the number of players n in each star to vary, and consider how the the asymptotic relative efficiency of the pairwise likelihood estimator changes as it does so.

Figure 3.2 is a plot of the asymptotic relative efficiency of the pairwise likelihood estimator relative to the maximum likelihood estimator, for a repeated star tournament for various values of n. In the case $n = 3$, the pairwise likelihood is identical to the full likelihood, and so the pairwise likelihood estimator is fully efficient. As n increases, the asymptotic relative efficiency of the pairwise likelihood

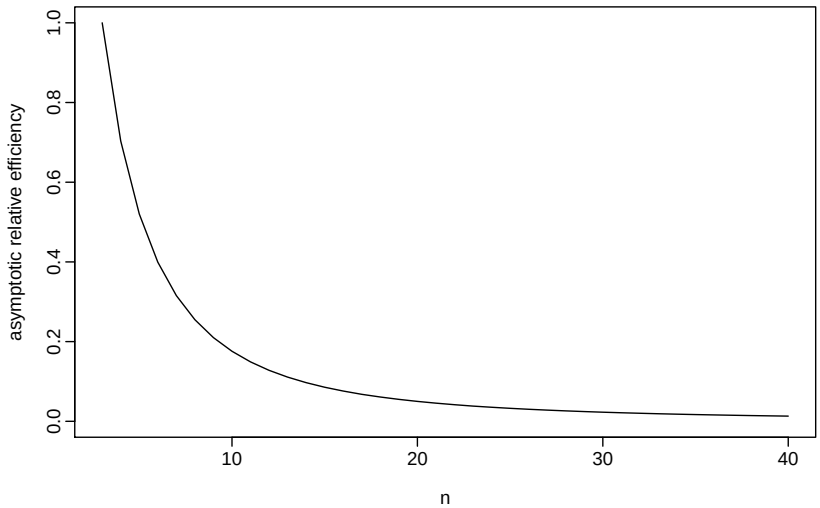


Figure 3.2: The asymptotic relative efficiency of the pairwise likelihood estimator of σ, in a repeated star tournament for various values of n.

estimator drops quickly. For large n, there is a large drop in efficiency incurred by using this pairwise likelihood instead of the full likelihood, and it seems sensible to try using some approximation to the likelihood instead.

Note. The structure of the tournament in Example 3.1 was chosen to make it easy to obtain the limiting distributions of the various estimators. In the case of the star tournament, it is possible to simplify the likelihood, by noting that conditional on the value of the random effect for the central player in the star, the contests are independent. This means that the likelihood for a star on n players may be written as

$$L(\sigma|\mathbf{y}) = \int_{-\infty}^{\infty} \left[\prod_{i=2}^{n} \int_{-\infty}^{\infty} \Phi\left(\sigma(u_i - u_1)(-1)^{y_{i1}+1}\right) \phi(u_i) du_i \right] \phi(u_1) du_1.$$

In fact,

$$\int_{-\infty}^{\infty} \Phi\left(\sigma(u_i - u_1)(-1)^{y_{i1}+1}\right) \phi(u_i) du_i = \Phi\left(\frac{\sigma u_1 (-1)^{y_{i1}}}{\sqrt{1+\sigma^2}}\right),$$

so the likelihood can be computed by evaluating a single one-dimensional integral. By differentiating under the integral, the first and second derivatives of the likelihood with respect to σ are also easy to compute.

The example also has a lot of symmetry, and a sufficient statistic for σ is W_1, the number of times the central player in the star wins. In fact, we would make the same inference about σ whether we observed the central player to win w_1 times, and lose the remaining $n-1-w_1$ contests, or to lose w_1 times, and win the remainder.

So

$$S_1 = \min\{W_1, n-1-W_1\} \in \left\{0, 1, \ldots, \left\lfloor \frac{n-1}{2} \right\rfloor\right\}$$

is a sufficient statistic for σ. It is straightforward to enumerate all the possibilities for S_1, and to find the probability of each occurring under the model. By computing the likelihood and its derivatives, and the pairwise likelihood and its derivatives, for a tournament corresponding to each choice of S_1, it is therefore possible to find the asymptotic variance of the full likelihood estimator, and that of the pairwise likelihood estimator.

3.1.4 Estimators maximising an approximation to the likelihood

Many of the methods for inference considered in Chapter 2 involve replacing the likelihood with some approximation to it. The estimators obtained by maximising this approximated likelihood need no longer be consistent as the number of independent replications tends to infinity. Suppose that the approximation to the log-likelihood is smooth, so that we have an approximation $\tilde{u}(\theta|\mathbf{y})$ to the score function given data $\mathbf{y}$. The resulting estimator $\tilde{\theta}_R$ maximising the approximation to the likelihood solves

$$\sum_{i=1}^{R} \tilde{u}(\theta|\mathbf{y}_{(i)}) = 0.$$

Let $\tilde{\theta}$ be the solution to

$$\mathbb{E}\left[\tilde{u}(\theta|\mathbf{Y})\right] = 0,$$

where the expectation is taken under the true distribution. Then, applying Theorem 1,

$$\tilde{\theta}_R \to^p \tilde{\theta}$$

as $R \to \infty$.

Applying Theorem 2 gives that

$$\sqrt{R}(\tilde{\theta}_R - \tilde{\theta}) \to^d N_p\left(0, \tilde{H}(\tilde{\theta})^{-1}\tilde{J}(\tilde{\theta})\left[\tilde{H}(\tilde{\theta})^T\right]^{-1}\right).$$

However, by treating the approximation to the likelihood as if it were the true likelihood, we will assume that, for large R, $\sqrt{R}(\tilde{\theta}_n - \theta_0)$ is well approximated by a normal distribution, with variance the inverse of the observed information matrix of the approximated likelihood at its maximum $\tilde{\theta}_R$. So, as the estimator tends towards $\tilde{\theta}$, we are more and more sure that the estimator is close to θ_0. If $\tilde{\theta} \neq \theta_0$, this will eventually cause confidence regions for θ to have coverage much lower than the

nominal level.

If the approximation to the likelihood is sufficiently good, $\tilde{\theta}$ will be very close to θ_0, and the confidence regions for θ will have approximately correct coverage provided that R is not extremely large. However, poor-quality approximations can result in an estimator with a very large asymptotic bias, as Example 3.2 shows.

Example 3.2. We return to the star tournaments of Example 3.1. Recall that the ability of player i is modelled as

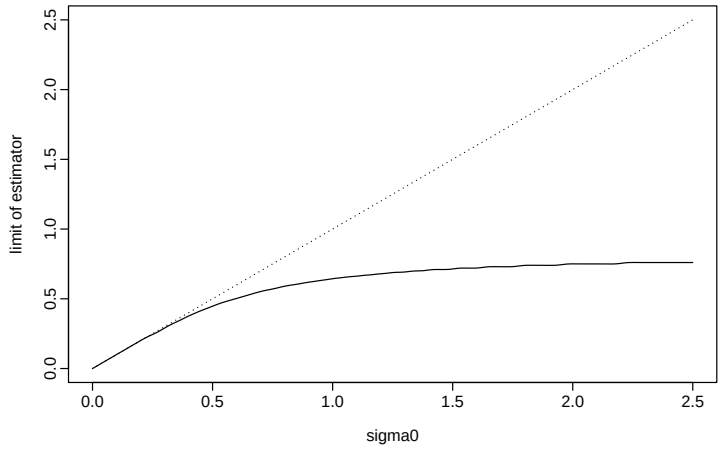
$$\lambda_i = \sigma u_i,$$

where σ is an unknown parameter of interest. Now we suppose that we do inference by replacing the likelihood with an approximation, found either by using Laplace's method or by importance sampling. Recall that in the limit as the number of repetitions, R, tends to infinity, the maximum likelihood estimator and the pairwise likelihood estimator are both consistent. This will not be true for the estimators found by maximising approximations to the likelihood, so we are interested in how far the limits of the estimators are from the true value in each case.

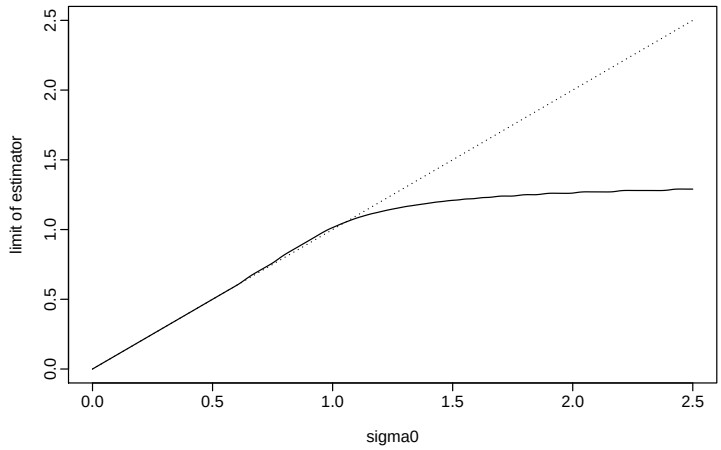
Suppose that the number of players in each tournament, n, is fixed at 50. First, we consider the effect of using the Laplace approximation to the likelihood in place of the true likelihood. Figure 3.3a shows the limit of the estimator maximising the Laplace approximation to the likelihood, plotted against the true value of σ. Even for quite small values of σ, there is a large negative asymptotic bias in the Laplace estimator of σ.

The Laplace approximation performs badly in this case, so we might instead choose to use an importance sampling approximation to the likelihood. This requires N evaluations of the function $g(.|\mathbf{y}, \theta)$, in addition to those required to find the normal approximation. Thus the method will be considerably more expensive than using a quadrature rule to compute the one dimensional integral in the simplified version of the likelihood, even for small N. Nonetheless, the method is still reasonably fast for small N, and importantly it is possible to use the method for inference in any generalised linear mixed model, where the simplification afforded by the special structure of this case is not available.

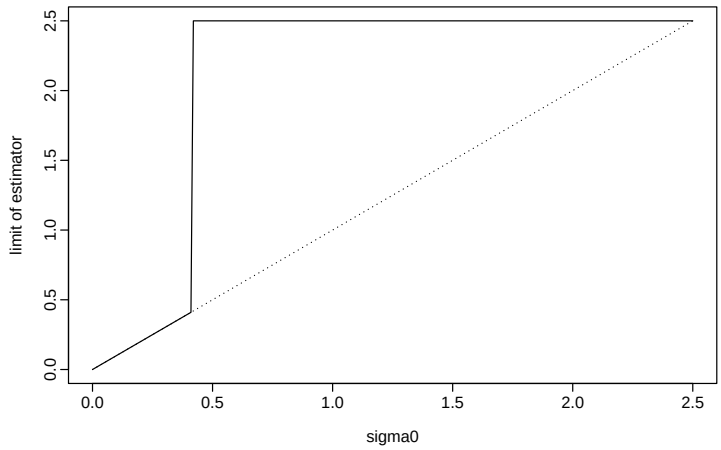
Figure 3.3b shows the limit of an importance sampling estimator, where the approximation is constructed based on one particular random sample of size $N = 10^4$, using the same samples to approximate the likelihood for each possible observed tournament. The limit of the estimator would change if a different random sample were used, but this gives an indication of the level of asymptotic bias incurred by using an importance sampling approximation with $N = 10^4$ samples.



(a) Laplace

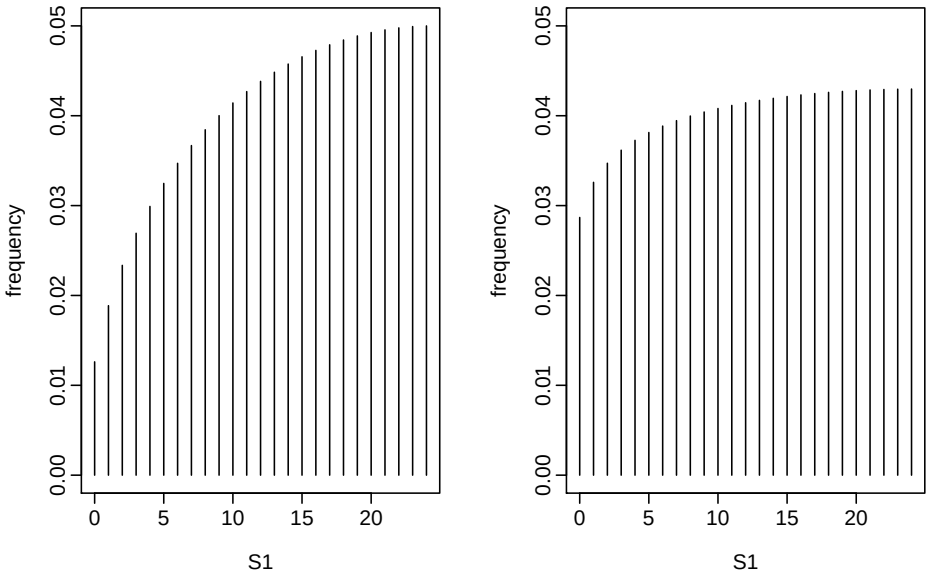


(b) Importance sampling (Using one sample of size $N = 10^4$)



(c) Importance sampling (Using one sample of size $N = 10^5$)

Figure 3.3: The limit of various estimators in a repeated star tournament with $n = 50$, as $R \to \infty$. In each case the dotted line is $y = x$, representing the limit of a consistent estimator.



(a) $\sigma = 1.3$ (b) $\sigma = 2.5$

Figure 3.4: The distribution of S_1 in a star tournament on $n = 50$ players, for different values of σ.

The importance sampling approximation seems to offer a real improvement over the Laplace approximation, but there is still a large asymptotic bias for moderate to large values of σ.

When $\sigma_0 = 2.5$, the limit of this importance sampling estimator is 1.3. This limiting value itself is moderately large, and since the distribution of tournaments will be very similar for all σ sufficiently large, it seems possible that the practical effects of this asymptotic bias will be small. However, Figure 3.4 shows that there is a noticeable difference in behaviour between the distribution of S_1 for $\sigma = 1.3$ and $\sigma = 2.5$, so the difference is an important one.

It is worth seeing what happens when we use an even larger number of samples N to construct the importance sampling approximation. The limit of the estimator for one particular sample of size $N = 10^5$ is shown in Figure 3.3c. The asymptotic bias of the estimator is now much larger than in the $N = 10^4$ case. Examining the shape of the approximation for $N = 10^5$ for $S_1 = 15$ in Figure 3.5, we can see that the approximated log-likelihood for large σ is much too large. Figure 3.6 shows a trace plot of the log-likelihood approximation for $S_1 = 15$ and $\sigma = 2$, using from one up to 10^6 samples. There is a large jump in the likelihood approximation at around $N = 10^5$, after the approximation has already appeared to converge. This behaviour suggests that the importance weights do not have

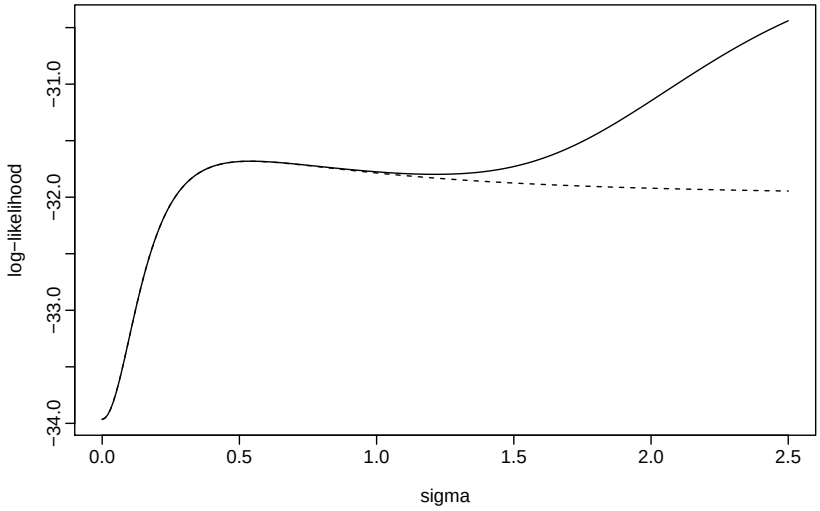


Figure 3.5: An importance sampling approximation to the log-likelihood for a star tournament with $S_1 = 15$, using $N = 10^5$ samples. The dotted line gives the true log-likelihood.

finite variance, which, as discussed in Section 2.3.3, can lead to slow and unstable convergence of the approximation.

3.1.5 Hypothesis testing and confidence intervals

It will generally be of interest not just to obtain point estimates for the parameters of a generalised linear mixed model, but to test hypotheses about these parameters, or to construct confidence intervals for them.

Suppose that we have a p-dimensional parameter vector θ, and consider testing the hypothesis $\phi = \phi_0$ for some p_0-dimensional subset ϕ of θ.

We will consider two test statistics for testing such hypotheses, first assuming that the true likelihood and the corresponding maximum likelihood estimator $\hat{\theta}$ are available. Write $\hat{\phi}$ for maximum likelihood estimator of ϕ, and $I^{\phi\phi}(\hat{\theta}) = \left[I(\hat{\theta})^{-1}_{\phi\phi}\right]^{-1}$, where $I(\hat{\theta})^{-1}_{\phi\phi}$ is the submatrix of $I(\hat{\theta})^{-1}$ corresponding to ϕ.

The Likelihood ratio test statistic is defined as

$$\Lambda(\phi_0) = 2\left[\sup_{\theta} \ell(\theta|\mathbf{y}) - \sup_{\theta:\phi=\phi_0} \ell(\theta|\mathbf{y})\right],$$

and Wald test statistic as

$$W(\phi_0) = (\hat{\phi} - \phi_0)^T I^{\phi\phi}(\hat{\theta})(\hat{\phi} - \phi_0).$$

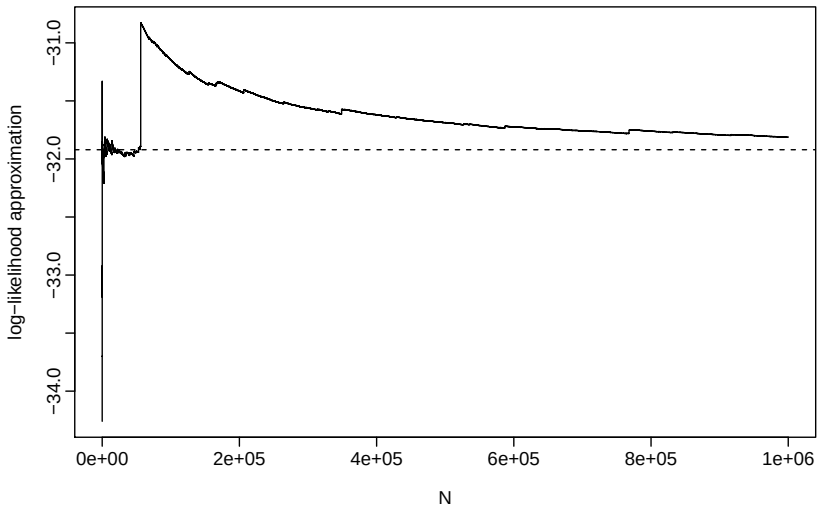


Figure 3.6: A trace of the importance sampling approximation to the log-likelihood at $\sigma = 2$, for a star tournament with $S_1 = 15$. The dotted line gives the true log-likelihood.

Under standard regularity conditions, it can be shown that under the null hypothesis that $\phi = \phi_0$, both $\Lambda(\phi_0) \to^d \chi^2_{p-p0}$ and $W(\phi_0) \to^d \chi^2_{p-p0}$ as $R \to \infty$.

However, one of the regularity conditions needed to prove these results is sometimes violated in a generalised linear mixed model. If ϕ_0 is on the boundary of the parameter space, the standard asymptotic distribution of the test statistics need not apply. Recall that in our definition of a generalised linear mixed model, there is a parameter ψ controlling how the random effects enter into the linear predictor. We suppose that $\psi = 0$ corresponds to the case in which there are no random effects in the model. If we are interested in testing $\psi = 0$, then the parameter value of interest is on the boundary of the parameter space. In this sort of situation, Self & Liang (1987) show that, if there are an increasing number of independent and identically distributed observations, Λ has limiting distribution which is a mixture between the usual χ^2_{p-p0} distribution and a point mass at 0. In practice, this means that some adjustment should be made to the assumed null distribution of the test statistic if the distribution under the null is such that the maximum likelihood estimator will be on the boundary of the parameter space a non-negligible proportion of the time. In the more realistic setting in which the replications are not identically distributed, Crainiceanu & Ruppert (2004) show that using a mixture between χ^2_{p-p0} distribution and a point mass at 0 may itself be incorrect. They demonstrate how to find the correct distribution of the likelihood ratio test statistic in a linear mixed model, but it remains difficult to obtain a good correction for other generalised linear mixed

models.

Throughout the remainder of the thesis, the unadjusted χ^2_{p-p0} distributed will be used. This will result in a slight over-reluctance to reject the hypothesis that ψ is small.

To construct a confidence interval (or region, if $p_0 > 1$) for ϕ with approximate coverage $1-\alpha$, we can invert the hypothesis test, giving a Wald-type confidence region of

$$I_W^{(1-\alpha)} = \{\phi : W(\phi) \leq c_\alpha\}$$

and a likelihood ratio confidence region of

$$I_\Lambda^{(1-\alpha)} = \{\phi : \Lambda(\phi) \leq c_\alpha\},$$

where c_α is chosen so that $Pr(\chi^2_{p-p_0} > c_\alpha) = \alpha$. Writing $\ell^p(\phi_0|\mathbf{y}) = \sup_{\theta:\phi=\phi_0} \ell(\theta|\mathbf{y})$ for the profile log-likelihood of ϕ_0, the the likelihood ratio statistic may be written as

$$\Lambda(\phi_0) = 2\left[\ell^p(\hat{\phi}|\mathbf{y}) - \ell^p(\phi_0|\mathbf{y})\right],$$

and the likelihood ratio confidence region

$$I_\Lambda^{(1-\alpha)} = \left\{\phi : \ell^p(\phi|\mathbf{y}) \geq \ell^p(\hat{\phi}|\mathbf{y}) - \frac{c_\alpha}{2}\right\}$$

is just a set of points with sufficiently large profile likelihood.

The Wald test statistic is usually considerably easier to compute than the likelihood ratio test statistic. However, it exhibits worrying behaviour in some settings. Hauck & Donner (1977) demonstrate the phenomenon that, as the true value of a regression parameter in a logistic regression model increases, the power of the hypothesis test for testing that it is zero eventually starts to fall. Example 3.3 demonstrates similar behaviour for the variance parameter of the random effects in a generalised linear mixed model.

Example 3.3. Suppose that we have 3 replications of a star tournament on 50 players, as described in Example 3.1, and consider testing $\sigma = \sigma_0$, for various values of σ_0. We construct Wald and likelihood ratio tests of nominal size $\alpha = 0.05$. The true sizes of these tests are given in Figure 3.7. The Wald test has size much larger than the nominal size for most values of σ_0. On the other hand, the likelihood ratio test has size smaller than 0.05 for σ very small (since no adjustment was made for the fact that the parameter is close to the boundary), but for larger σ has the correct size.

The power curves of the two tests for rejecting $\sigma = 0$ are given in Figure 3.8.

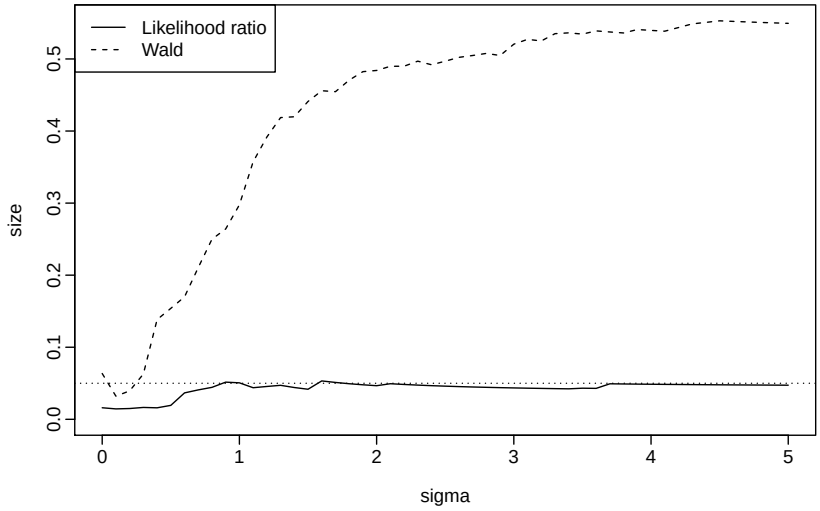


Figure 3.7: The true size of hypothesis tests for σ, of nominal size 0.05, in a repeated star tournament with $n = 50$, $R = 3$.

A similar phenomenon to that observed by Hauck & Donner (1977) may be seen here: the power of the Wald test for testing $\sigma = 0$ diminishes for large σ.

We conclude that the use of the Wald test statistic may be very misleading, and it is far better to use the likelihood ratio test statistic.

All of the above discussion relates to the situation in which the likelihood itself is available. When the likelihood is replaced by an approximation, the quality of inference from the likelihood ratio test will depend on the quality of that approximation. Figure 3.9 shows the true size of the approximate likelihood ratio test constructed by using the Laplace approximation, in the repeated star tournament described in Example 3.3. As expected, the size of the test for large σ is much larger than the nominal level.

When a composite likelihood is used in place of the full likelihood, the asymptotic distribution of the test statistics is no longer χ^2_{p-p0}. A replacement composite likelihood Wald test statistic could be constructed as

$$W_C = (\hat{\phi}^C - \phi_0)^T G_C^{\phi\phi}(\hat{\theta}^C)(\hat{\phi}^C - \phi_0),$$

where $G_C^{\phi\phi}(\hat{\theta}^C) = \left[G_C(\hat{\theta}^C)^{-1}_{\phi\phi}\right]^{-1}$ and $G_C(\hat{\theta}^C)^{-1}_{\phi\phi}$ is the submatrix of the estimated variance matrix $\left[G_C(\hat{\theta}^C)\right]^{-1}$ corresponding to ϕ. Varin *et al.* (2011) review some of the different adjustments to the likelihood ratio test statistic which have been proposed. We do not discuss these further here, having already concluded that the

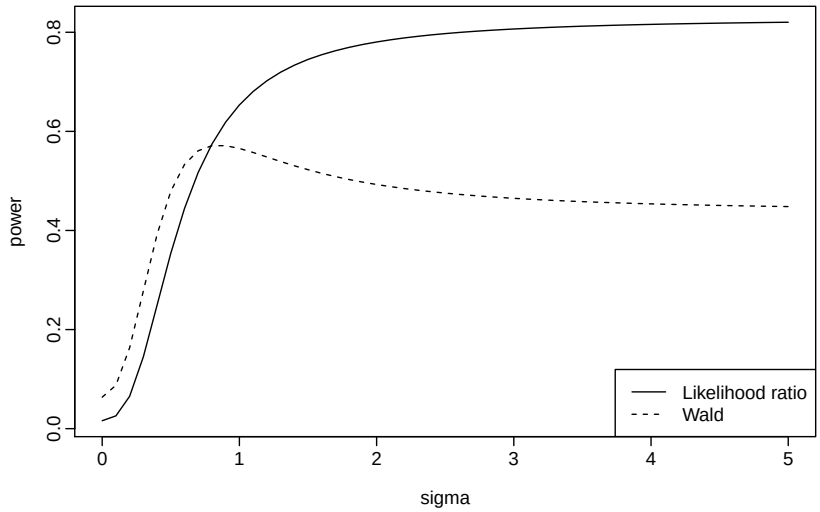


Figure 3.8: The power of hypothesis tests of $\sigma = 0$, of nominal size 0.05, in a repeated star tournament with $n = 50$, $R = 3$.

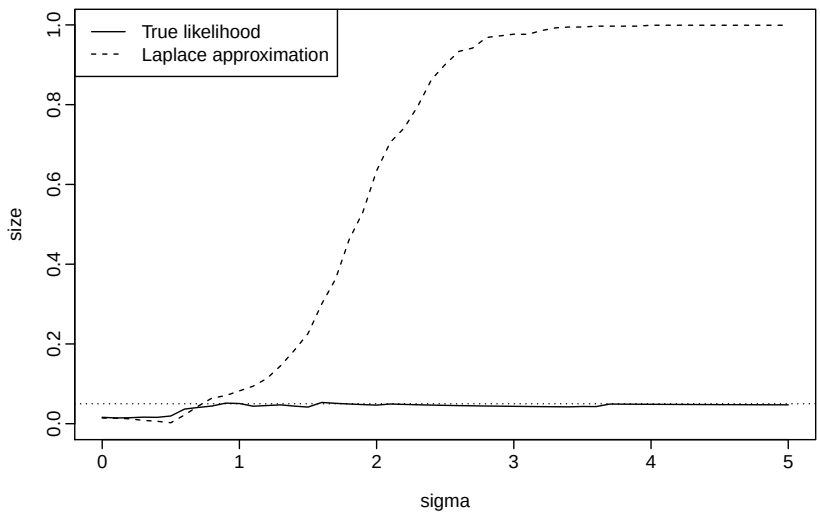


Figure 3.9: The true size of likelihood ratio tests for σ, of nominal size 0.05, based on different approximations to the likelihood.

drop in efficiency in the pairwise likelihood estimator is unacceptably large in some circumstances.

3.1.6 Penalised forms of the likelihood

In sparse models with binary data, it is fairly common that the maximum likelihood estimator is not finite. To prevent such problems, it seems sensible to impose some penalty on the parameters.

In the case of a generalised linear model, where there are no random effects, Firth (1993) demonstrates how to choose a penalty to remove the first-order asymptotic bias in $\hat{\beta}$. The penalty suggested here is equal to that bias reduction penalty, chosen under the assumption of no random effects.

Write

$$I_0(\beta) = \mathbb{E}\left[-\nabla_\beta^T \nabla_\beta \ell(\beta, \psi = 0)\right],$$

where $\psi = 0$ corresponds to the case of no random effects, and consider the penalised likelihood

$$\ell^p(\beta, \psi) = \ell(\beta, \psi) - p_0(\beta),$$

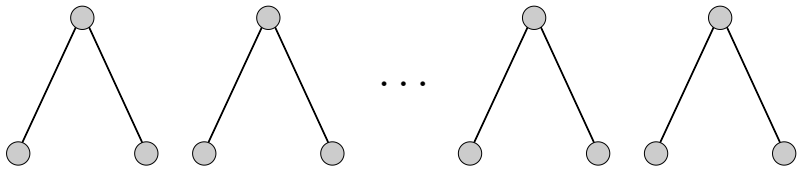
where

$$p_0(\beta) = -\frac{1}{2}\log|I_0(\beta)|.$$

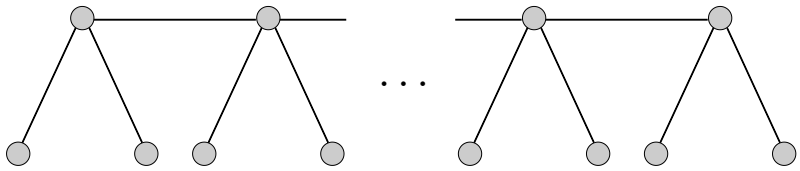
We call $p_0(.)$ the bias reduction penalty in all cases, although it removes the first-order asymptotic bias only when $\psi = 0$.

We may use the penalised likelihood in place of the full likelihood to test hypotheses using the Wald or likelihood ratio tests, because as the amount of information on the parameters in the data increases, the influence of the penalty term shrinks, and the test statistics retain the same limiting distributions.

We will use $p_0(.)$ as a penalty in some of the examples in the thesis, and it avoids infinite parameter estimates in those cases, but we do not claim that this penalty is optimal in any way. It may be useful to additionally impose some penalty on ψ, or to construct a joint penalty on β and ψ, so that a larger penalty is given to parameter values where β is small but ψ is large. One idea for an improved penalty function is given in Section 6.2.2, but some further work is required to check whether the inference using this suggested penalty has good properties.



(a) Repeated tournament



(b) Joined tournament

Figure 3.10: Tournament designs for Example 3.4

3.2 Asymptotics without independent replication

In reality, we cannot rely on having a large number of independent replications of the data. It is common to have few, or no, independent replications. Instead, we often have a large number of observations, all dependent on one another. We still want to use the sort of results deduced under $R \to \infty$ asymptotics, such as the consistency and asymptotic normality of the maximum likelihood estimator. Assuming that there is no independent replication, will these results still hold?

Example 3.4. Consider R repetitions of a star tournament on 3 players, as shown in Figure 3.10a. Suppose again that

$$\lambda_i = \sigma u_i,$$

where $u_i \sim N(0,1)$ and σ is an unknown parameter. We have already shown that the maximum likelihood estimator of σ will be consistent in this setting as $R \to \infty$. For large R, the maximum likelihood estimator of σ will be approximately normally distributed, with variance $\frac{V}{R}$, for some constant V. The total number of contests is such a tournament is $m = 2R$, and we could equally well speak of asymptotics as $m \to \infty$.

Now consider a new tournament is which an additional contest is played between the central players of consecutive star tournaments, as shown in Figure

3.10b, so that there are $m = 3R - 1$ contests in total. Now there is no independent replication in the tournament. However, knowledge of the outcome of the additional contests increases the amount of information available about σ, so the maximum likelihood estimator must remain consistent as $R \to \infty$. For large R, the maximum likelihood estimator of σ should also remain approximately normally distributed, although the variance will be slightly smaller than $\frac{V}{R}$, because of the additional information provided by the extra contests.

This type of argument can be used more generally to deduce whether it is valid to rely on the usual asymptotic results. Suppose that it is possible to find some subset of the observations $\mathbf{Y}_{(S)} = \{Y_j : j \in S\}$ so that $\mathbf{Y}_{(S)}$ consists of R independent blocks of data. If R is large, and the parameter θ is identifiable from the distribution of each block, then we can expect that the maximum likelihood estimator will be close to the true value θ_0, and that it will be approximately normally distributed. This means that the usual asymptotic distributions of test statistics for testing hypotheses about θ may be assumed to be approximately valid.

3.3 The effect of sparsity on the quality of the Laplace approximation

The special structure of the repeated star tournament of Examples 3.1 and 3.2 allowed us to study the asymptotic behaviour of the various estimators in that model. In this sparse setting, the Laplace and importance sampling approximations give estimators with poor statistical properties. However, the performance of these estimators is unlikely to be this bad in all cases. In dense models, where substantial information is available on each random effect, we anticipate that the Laplace approximation may provide inference much closer to the true likelihood inference. A numerical example is given to provide some insight into situations in which the inference gained using Laplace approximation may be sufficiently accurate, and when it is necessary to use some other method.

Example 3.5. A complete tournament is one in which every pair from a set of n players competes exactly once. This is a relatively common tournament design in practice, for example in sports tournaments where a contest is played between each pair of teams. There are $m = \binom{n}{2}$ contests in such a tournament, so $n = O(m^{1/2})$. For large n, this a relatively dense tournament, although the amount of information available per random effect is not sufficiently large that the relative error in Laplace approximation to the likelihood (or equivalently the absolute error in the Laplace

approximation to the log-likelihood) tends to 0 as $n \to \infty$ (see the discussion in Section 2.3.2). However, the accuracy of the approximation to the likelihood itself is unimportant: what matters is that the approximation to the log-likelihood surface is sufficiently accurate that the resulting inference is close to the inference obtained from the true likelihood.

To investigate this, we consider tournaments consisting of R_n independent subtournaments, where each subtournament is a complete tournament among n players. In such a tournament, there are nR_n players and $\binom{n}{2}R_n$ contests overall. We choose pairs (n, R_n) of $(20, 1)$, $(12, 3)$, $(8, 7)$ and $(5, 19)$, so that the total number of contests is approximately constant. We do this so that each tournament provides a similar level of information about the parameters. We simulate a single binary observed covariate x_i for each player, which is 1 with probability 0.5, and 0 otherwise. We assume a binary pairwise competition model with probit link, so that

$$Pr(i \text{ beats } j|\lambda_i, \lambda_j) = \Phi(\lambda_i - \lambda_j),$$

where

$$\lambda_i = \beta x_i + \sigma u_i,$$

and $u_i \sim N(0, 1)$, and simulate sample tournaments from the model, with $\beta = 1$ and $\sigma = 0.5$.

For the sparse tournaments, with small n, we anticipate that the Laplace approximation to the likelihood will not give inference close to the true likelihood. In order to study the quality of inference from the Laplace approximation in denser tournaments, we consider testing the hypothesis

$$H_0 : \theta = \theta^*$$

for each value of θ^* on a large two-dimensional grid, based on either the Laplace approximation to the likelihood, or an importance sampling approximation to the likelihood based on $N = 10^4$ samples. To do this, we construct a likelihood ratio statistic based on each approximation to the likelihood, letting

$$\Lambda^L(\theta) = 2\left(\left\{\max_\theta \ell^L(\theta)\right\} - \ell^L(\theta)\right)$$

and

$$\Lambda(\theta) = 2\left(\left\{\max_\theta \ell^{IS}(\theta, 10^4)\right\} - \ell^{IS}(\theta, 10^4)\right),$$

where $\ell^L(\theta)$ and $\ell^{IS}(\theta, N)$ are respectively the Laplace and importance sampling

approximations to the log-likelihood for θ.

Although there is no guarantee that inference based on the importance sampling approximation will be fully accurate, if the inference is very similar to that from the Laplace approximation, then it seems reasonable to conclude that the inference from both methods is accurate. If, however, there is a large difference in the inference, then the Laplace approximation is not sufficiently accurate. For this reason, for the purposes of this example, the importance sampling approximation to the likelihood is treated as if it were the true likelihood.

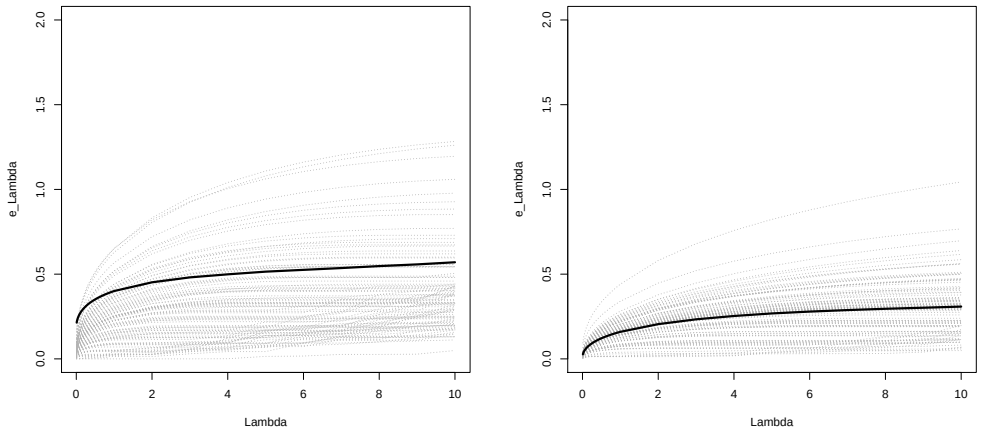
For any fixed value Λ, those values of θ with $\Lambda^L(\theta) \leq \Lambda$ give a confidence region for θ, with approximate coverage $p(\Lambda) = Pr(\chi^2_2 \leq \Lambda)$, since θ has dimension 2. So, for each fixed Λ (or equivalently, each fixed coverage $p(\Lambda)$), we look at the maximal error in the likelihood ratio statistic for all θ contained in the corresponding confidence region. That is, we consider the error

$$e_\Lambda = \sup_{\theta:\Lambda_L(\theta)\leq\Lambda} |\Lambda^L(\theta) - \Lambda(\theta)|.$$

Since we only compute $\Lambda^L(\theta^*)$ and $\Lambda(\theta^*)$ at each value θ^* on a grid of points, we compute e_Λ by interpolating values for $\Lambda^L(\theta)$ and $\Lambda(\theta)$ for θ not on this grid. We use the cubic spline interpolation methods described Section 4.6 to do this. Figure 3.11 shows plots of e_Λ against Λ for each tournament structure, for 100 simulated examples in each case. We see that this error diminishes with n: the inference from the Laplace approximation become more similar to the inference from the true likelihood as the tournament becomes more dense.

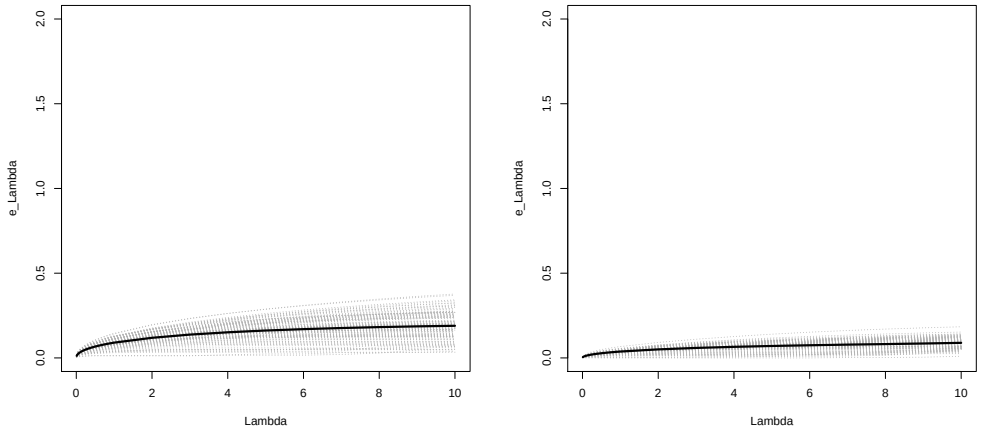
3.4 Conclusions

In the repeated star tournament of Examples 3.1 and 3.2, the pairwise likelihood estimator was inefficient, and the estimators obtained by maximising the Laplace and importance sampling approximations to the likelihood had large asymptotic bias. A better method is clearly required here, and as it happens, a simplification of the likelihood is available which makes full likelihood inference possible in this case. However, if just a few edges were added to the star structure, the likelihood would no longer simplify to a single one-dimensional integral, and to find the likelihood we would again be forced to evaluate an n-dimensional integral. The addition of these few edges would be unlikely to have much impact on the properties of the existing methods of inference: for a large number of players, the pairwise likelihood would remain inefficient, and the Laplace and importance sampling approximations would



(a) $n = 5$, $R_n = 19$

(b) $n = 8$, $R_n = 7$



(c) $n = 12$, $R_n = 3$

(d) $n = 20$, $R_n = 1$

Figure 3.11: The error e_Λ plotted against Λ, for 100 simulations from each of various complete tournaments on n players, repeated R_n times. The bold line gives the average of e_Λ across all 100 simulations.

be of poor quality.

The aim of the sequential reduction method of Chapter 4 is to find simplifications to the likelihood for any generalised linear mixed model. For models with sparse structure, this method will reduce dramatically the cost of finding a good approximation to the likelihood.

In models with dense structure, the new sequential reduction method of likelihood approximation will not provide much simplification of the n-dimensional integral. However, in such models, the Laplace approximation appears to be sufficiently good that replacing the likelihood with its Laplace approximation yields close to full likelihood inference (see Example 3.5). The new method will reduce the cost of likelihood computation in precisely the situations where such a simplification is needed most.

Chapter 4

A new method for approximating the likelihood

4.1 Introduction

The aim of this chapter is to develop a method for simplifying the likelihood. Recall that the likelihood may be written as the integral of

$$g(\mathbf{u}|\mathbf{y},\theta) = f(\mathbf{y}|\mathbf{u},\theta)\prod_{i=1}^{n}\phi(u_i)$$

over $\mathbf{u}$. By considering $\phi(u_i)$ as a prior for u_i, we may think of $g(.|\mathbf{y},\theta)$ as a non-normalised posterior density for the random effects $\mathbf{u}$, given $\mathbf{y}$ and θ.

Consider once more the example of a star tournament on n players, as shown in Figure 3.1 on page 21. The discussion in Example 3.1 describes how to simplify the likelihood in this case. The simplification is possible because, in the posterior distribution of the random effects, $\{u_2,\ldots,u_n\}$ are conditionally independent, given u_1. It therefore seems worthwhile to study the dependence structure of the posterior distribution of the random effects in the general case, to see if a similar simplification may be obtained.

4.2 The posterior dependence graph

Before observing the data $\mathbf{y}$, the random effects $\mathbf{u}$ are independent. The information provided by $\mathbf{y}$ about the value of combinations of those random effects induces dependence between them. Suppose, for example, that we have a model with three random effects, $\{u_1,u_2,u_3\}$, and make two observations, the first involving u_1 and u_2,

and the second involving u_2 and u_3. The first observation provides some information about the value of some combination of u_1 and u_2, so u_1 and u_2 are dependent in the posterior. Similarly, u_2 and u_3 are dependent in the posterior. This induces a dependence between u_1 and u_3 in the posterior distribution, because being told the value of u_1 tells us something about u_2, which in turn provides information about the likely value of u_3. However, if the value of u_2 were known, being told the value of u_1 would provide no additional information about the value of u_3. That is, u_1 and u_3 are conditionally independent in the posterior distribution, given u_2.

We can generalise this, and say that if there is no observation involving both u_i and u_j, then u_i and u_j are conditionally independent in the posterior distribution, given the values of all the other random effects.

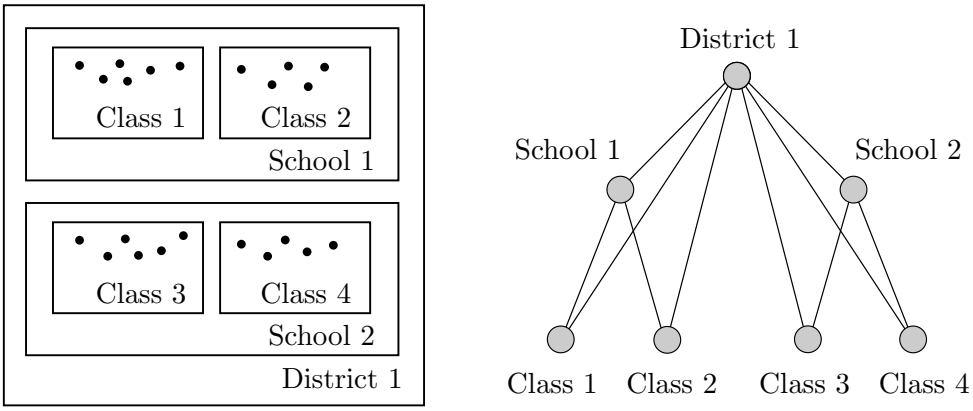
It is possible to represent this conditional independence structure graphically. Consider a graph $\mathcal{G}$ constructed to have:

1. A vertex for each random effect
2. An edge between two vertices if there is at least one observation involving both of the corresponding random effects.

By construction of $\mathcal{G}$, there is an edge between i and j in $\mathcal{G}$ only if $\mathbf{y}$ contains an observation involving both u_i and u_j. So if there is no edge between i and j in $\mathcal{G}$, u_i and u_j are conditionally independent in the posterior distribution, given the values of all the other random effects. This property is called the **pairwise Markov property** of the posterior distribution of the random effects with respect to $\mathcal{G}$. We call $\mathcal{G}$ the **posterior dependence graph** for $\mathbf{u}$ given $\mathbf{y}$.

In a pairwise competition model, the posterior dependence graph simply consists of a vertex for each player, with an edge between two vertices if those players compete in at least one contest. For models in which each observation relies on more than two random effects, an observation will not be represented by a single edge in the graph.

Example 4.1. Consider a four-level random intercept model. A typical example might be of observations being recorded on schoolchildren, who are grouped into classes, which themselves are contained within schools, and then within larger districts. Figure 4.1a gives a small example of this sort of structure, where the points represent the individual children in each class. An observation on a child depends on the random effects associated with the class, school and district to which that child belongs. The posterior dependence graph will contain a triangle on the vertices representing those three groups, for each possible combination of the groups. Figure



(a) Model structure (b) Posterior dependence graph

Figure 4.1: A four-level model

4.1b shows the posterior dependence graph for this example. If there were more than one district, the posterior dependence graph would contain separate components for each district, with no edges between these components.

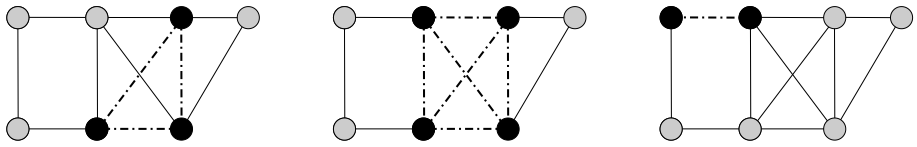
The posterior dependence graph provides a useful visualisation of the conditional independence structure of the posterior distribution of the random effects, but does not immediately gives us a simplification of the likelihood. However, the problem of computing the likelihood has now been transformed to that of finding a normalising constant of a density associated with an undirected graphical model. Methods are available to simplify the computation of such a normalising constant, which we now review.

4.3 Factorising the posterior density

The Hammersley-Clifford theorem (Hammersley & Clifford (1971), Besag (1974)) gives a relationship between the pairwise Markov property and the form of $g(.|\mathbf{y}, \theta)$.

To state the result, we first need a few definitions. A **complete** graph is one in which there is an edge from each vertex to every other vertex. A **clique** of a graph $\mathcal{G}$ is a complete subgraph of $\mathcal{G}$. A clique is said to be **maximal** if it is not itself contained within a larger clique. Figure 4.2 illustrates these definitions. For any graph $\mathcal{G}$, the set of all maximal cliques of $\mathcal{G}$ is unique, and we write $M(\mathcal{G})$ for this set.

Theorem 3 (Hammersley-Clifford theorem)**.** *Suppose* $\mathbf{u} = (u_1, \ldots, u_n)$ *is a random vector, with density* $f(.)$*. Suppose that* $f(\mathbf{u}) > 0$ *for all* $\mathbf{u} \in \mathbb{R}^n$*. Then* $\mathbf{u}$ *has the*



(a) A non-maximal clique (b) A maximal clique (c) A maximal clique

Figure 4.2: Illustration of some graph theory definitions

pairwise Markov property with respect to $\mathcal{G}$ if and only if $f(.)$ factorises over the maximal cliques of $\mathcal{G}$, so that we may write

$$f(\mathbf{u}) = \prod_{C \in M(\mathcal{G})} f_C(\mathbf{u}_C)$$

for some functions $f_C(.)$.

Proof. See Besag (1974)

Note. If $f(.)$ factorises over the maximal cliques of $\mathcal{G}$, it immediately follows that any non-normalised density $g(\mathbf{u}) \propto f(\mathbf{u})$ must also factorise in this way.

We may apply this result to the non-normalised posterior density of $\mathbf{u}$. Since $\phi(u_i) > 0$ for all u_i, we do indeed have that $g(\mathbf{u}|\mathbf{y}, \theta) > 0$ for all $\mathbf{u}$, so $g(.|\mathbf{y}, \theta)$ must factorise over the maximal cliques of the posterior dependence graph. In fact, we may show that such a factorisation exists directly, without the requirement that $g(.|\mathbf{y}, \theta)$ is positive everywhere. One particular such factorisation is constructed in Section 4.5.2, and would be valid even if we assigned zero prior density to some values of u_i.

4.4 Exploiting the clique factorisation

Jordan (2004) reviews how to use a find marginal density of a density factorised over the maximal cliques of a graph. These methods are well known in the graphical modelling literature, although their use is limited to certain special classes of distribution, such as discrete or normal distributions. We will use the same ideas, combined with a method for approximate storage of functions, to approximate the marginals of the continuous distribution with density proportional to $g(.|\mathbf{y}, \theta)$, and so find an approximation to the likelihood.

Recall that our aim is to find the likelihood

$$L(\theta|\mathbf{y}) = \int_{\mathbb{R}^n} g(\mathbf{u}|\mathbf{y}, \theta) d\mathbf{u}.$$

We take an iterative approach to the problem, first integrating out u_1 to find the non-normalised marginal posterior density of $\{u_2, \ldots, u_n\}$. We start with a factorisation of $g(.|\mathbf{y}, \theta)$ over the maximal cliques of the posterior dependence graph of $\{u_1, \ldots, u_n\}$, and the idea will be to write the marginal posterior density of $\{u_2, \ldots, u_n\}$ as a product over the maximal cliques of a new marginal posterior dependence graph. Once this is done, the process may be repeated n times to find the likelihood. We will write $\mathcal{G}_i$ for the posterior dependence graph of $\{u_i, \ldots, u_n\}$, so we start with posterior dependence graph $\mathcal{G}_1 = \mathcal{G}$. Write $M_i = M(\mathcal{G}_i)$ for the maximal cliques of $\mathcal{G}_i$.

Factorising $g(.|\mathbf{y}, \theta)$ over the maximal cliques of $\mathcal{G}_1$ gives

$$g(\mathbf{u}|\mathbf{y}, \theta) = \prod_{C \in M_1} g^1_C(\mathbf{u}_C),$$

for some functions $g^1_C(.)$. To integrate over u_1, note that it is only necessary to integrate over maximal cliques containing vertex 1, leaving the functions on other cliques unchanged. Let N_1 be the set of neighbours of vertex 1 in $\mathcal{G}$ (including vertex 1 itself). Then

$$\begin{aligned}\int g(\mathbf{u}|\mathbf{y}, \theta) du_1 &= \int \prod_{C \in M_1 : C \subseteq N_1} g^1_C(\mathbf{u}_C) du_1 \prod_{\tilde{C} \in M_1 : \tilde{C} \not\subseteq N_1} g^1_{\tilde{C}}(\mathbf{u}_{\tilde{C}}) \\ &= \int g^1_{N_1}(\mathbf{u}_{N_1}) du_1 \prod_{\tilde{C} \in M_1 : \tilde{C} \not\subseteq N_1} g^1_{\tilde{C}}(\mathbf{u}_{\tilde{C}}).\end{aligned}$$

Thus $g^1_{N_1}(.)$ is obtained by multiplication of all the functions on cliques which are subsets of N_1. This is then integrated over u_1, to give

$$g^2_{N_1 \setminus 1}(\mathbf{u}_{N_1 \setminus 1}) = \int g^1_{N_1}(u_1, \mathbf{u}_{N_1 \setminus \{1\}}) du_1.$$

The functions on all cliques $\tilde{C}$ which are not subsets of N_1 remain unchanged, with $g^2_{\tilde{C}}(\mathbf{u}_{\tilde{C}}) = g^1_{\tilde{C}}(\mathbf{u}_{\tilde{C}})$.

This defines a new factorisation of $g(u_2, \ldots u_n|\mathbf{y}, \theta)$ over the maximal cliques M_2 of the posterior dependence graph for $\{u_2, \ldots, u_n\}$, where M_2 contains $N_1 \setminus 1$, and all the remaining cliques in M_1 which are not subsets of N_1. The same process may then be followed to remove each u_i in turn.

4.5 The sequential reduction method for likelihood approximation

4.5.1 A general algorithm

We now give the general form of a sequential reduction method for approximating the likelihood. We highlight the places where choices must be made to use this method in practice. The following sections then discuss each of these choices in detail.

1. The u_i may be integrated out in any order. Section 4.7 discusses how to choose a good order, with the aim of minimising the cost of approximating the likelihood. Reorder the random effects so that we integrate out $u_1, \ldots, u_n$ in that order.

2. Factorise $g(\mathbf{u}|\mathbf{y}, \theta)$ over the maximal cliques M_1 of the posterior dependence graph, as
$$g(\mathbf{u}|\mathbf{y}, \theta) = \prod_{C \in M_1} g_C^1(\mathbf{u}_C).$$
This factorisation is not unique, so we must choose one particular factorisation $\{g_C^1(.) : C \in M_1\}$. Section 4.5.2 gives the factorisation we use in practice.

3. Once $u_1, \ldots u_{i-1}$ have been integrated out (using some approximate method), we have the factorisation
$$\tilde{g}(u_i, \ldots, u_n|\mathbf{y}, \theta) = \prod_{C \in M_i} g_C^i(\mathbf{u}_C),$$
of the (approximated) non-normalised posterior for $u_i, \ldots, u_n$. Write
$$g_{N_i}(\mathbf{u}_{N_i}) = \prod_{C \in M_i : C \subseteq N_i} g_C^i(\mathbf{u}_C).$$
We then must integrate $g_{N_i}(.)$ over u_i. We need to store an approximate representation of the resulting integrated function $g_{N_i \setminus i}(.)$. Assuming that it is not possible to compute the integral analytically, for each fixed $\mathbf{u}_{N_i \setminus i}$, we could use a numerical integration technique to do the integration. We could then store the result of the numerical integration at some fixed points for $\mathbf{u}_{N_i \setminus i}$, and specify some means of interpolating the value of $g_{N_i \setminus i}(.)$ between those points.

In fact, to combine the two approximations of numerical integration and interpolation into one, we first store an approximate representation $\tilde{g}_{N_i}(.)$ of $g_{N_i}(.)$. We need to ensure that it is easy to obtain the integral of this approximate representation over u_i, to give an approximate representation $\tilde{g}_{N_i \setminus i}(.)$ of $g_{N_i \setminus i}(.)$. In Section 4.5.3 we consider the type of error we wish to minimise when constructing $\tilde{g}_{N_i}(.)$. Motivated by this, in Section 4.5.4, we make a change of variable so that we can use combinations of a standard set of points to store each function. In Section 4.6 we consider the choice of such a set of standard points, and the method which may be used to interpolate between them.

4. Write
$$\tilde{g}(u_{i+1}, \ldots, u_n | \mathbf{y}, \theta) = \tilde{g}_{N_i \setminus i}(\mathbf{u}_{N_i \setminus i}) \prod_{C \in M_i : C \not\subseteq N_i} g_C^i(\mathbf{u}_C),$$
defining a factorisation of the (approximated) non-normalised posterior density of $\{u_{i+1}, \ldots, u_n\}$ over the maximal cliques M_{i+1} of the new posterior dependence graph $\mathcal{G}_{i+1}$.

5. Repeat steps 3 and 4 for $i = 1, \ldots, n-1$, then integrate $\tilde{g}(u_n | \mathbf{y}, \theta)$ over u_n to give the approximation to the likelihood.

4.5.2 A specific clique factorisation

The general method described in Section 4.5.1 is valid for an arbitrary factorisation of $g(\mathbf{u}|\mathbf{y}, \theta)$ over the maximal cliques M_1 of the posterior dependence graph. To use the method in practice, we must first define the factorisation used.

Given an ordering of the vertices, order the cliques in M_1 lexicographically according to the set of vertices contained within them. The observation vector $\mathbf{y}$ is partitioned over the cliques in M_1 by including in $\mathbf{y}_C$ all the observations only involving items in the clique C, which have not already been included in $\mathbf{y}_B$ for some earlier clique in the ordering, B. Write $a(C)$ for the set of vertices appearing for the first time in clique C. Let

$$\begin{aligned} g_C^1(\mathbf{u}_C) &= f(\mathbf{y}_C | \mathbf{u}_C) \prod_{j \in a(C)} \phi(u_j) \\ &= \frac{g(\mathbf{u}_C | \mathbf{y}_C)}{\prod_{j \in \{C \setminus a(C)\}} \phi(u_j)} \end{aligned}$$

Then

$$g(\mathbf{u}|\mathbf{y}, \theta) = \prod_{C \in M_1} g_C^1(\mathbf{u}_C),$$

so $\left\{g_C^1(.) : C \in M_1\right\}$ does define a factorisation of $g(.|\mathbf{y}, \theta)$.

4.5.3 Minimising the error in the likelihood approximation

Suppose we use the factorisation described in Section 4.5.2. At stage i, the non-normalised posterior distribution of the remaining random effects is

$$\begin{aligned} g_i(u_i, \ldots, u_n|\mathbf{y}, \theta) &= \prod_{C \in M_i} g_C^i(\mathbf{u}_C) \\ &= g_{N_i}^i(\mathbf{u}_{N_i}) \prod_{\tilde{C} \in M_i : \tilde{C} \not\subseteq N_i} g_{\tilde{C}}^i(\mathbf{u}_{\tilde{C}}). \end{aligned}$$

Redefine the partition of the observation vector by

$$\mathbf{y}_{N_i} = \bigcup_{C \in M_i : C \subseteq N_i} \{\mathbf{y}_C\},$$

and let

$$a(N_i) = \bigcup_{C \in M_i : C \subseteq N_i} a(C),$$

keeping $\mathbf{y}_{\tilde{C}}$ and $a(\tilde{C})$ unchanged for all $\tilde{C} \not\subseteq N_i$.

Then, for each $\tilde{C} \in M_i$,

$$g_{\tilde{C}}^i(\mathbf{u}_{\tilde{C}}) = \frac{g(\mathbf{u}_{\tilde{C}}|\mathbf{y}_{\tilde{C}})}{\prod_{j \in \{\tilde{C} \setminus a(\tilde{C})\}} \phi(u_j)},$$

where $g(\mathbf{u}_C|\mathbf{y}_C)$ is the non-normalised posterior distribution of $\mathbf{u}_C$ given $\mathbf{y}_C$. Write N_i^I for 'internal' nodes only contained in $C \subseteq N_i$, N_i^E for 'external' nodes contained both in some $C \subseteq N_i$ and in some $\tilde{C} \in M_i \not\subseteq N_i$, and $R_i = \{2, \ldots, n\} \setminus N_i^I$ for the nodes which remain after removal of the internal nodes. Notice that

$$\bigcup_{\tilde{C} \in M_i : \tilde{C} \not\subseteq N_i} a(\tilde{C}) = \{i, \ldots, n\} \setminus a(N_i) = R_i \setminus \{a(N_i) \cap N_i^E\}.$$

Writing $a(N_i^E) = a(N_i) \cap N_i^E$, we have that

$$\prod_{\tilde{C} \in M_i : \tilde{C} \not\subseteq N_i} g_{\tilde{C}}^i(\mathbf{u}_{\tilde{C}}) = \frac{g(\mathbf{u}_{R_i}|\mathbf{y} \setminus \mathbf{y}_{N_i})}{\prod_{j \in a(N_i^E)} \phi(u_j)}.$$

Then, noting that $\{N_i, R_i \setminus N_i^E\}$ is a partition of $\{i, \ldots, n\}$, the likelihood is

given by

$$
\begin{aligned}
L(\theta|\mathbf{y}) &= \int g^i_{N_i}(\mathbf{u}_{N_i}) \frac{g(\mathbf{u}_{R_i}|\mathbf{y} \setminus \mathbf{y}_{N_i})}{\prod_{j \in a(N_i^E)} \phi(u_j)} d\mathbf{u}_{R_i \setminus N_i^E} d\mathbf{u}_{N_i} \\
&= \int g^i_{N_i}(\mathbf{u}_{N_i}) \left[\frac{\int g(\mathbf{u}_{R_i}|\mathbf{y} \setminus \mathbf{y}_{N_i}) d\mathbf{u}_{R_i \setminus N_i^E}}{\prod_{j \in a(N_i^E)} \phi(u_j)} \right] d\mathbf{u}_{N_i} \\
&= \int g^i_{N_i}(\mathbf{u}_{N_i}) h^i(\mathbf{u}_{N_i^E}) d\mathbf{u}_{N_i},
\end{aligned}
$$

where

$$
h^i(\mathbf{u}_{N_i^E}) = \frac{g(\mathbf{u}_{N_i^E}|\mathbf{y} \setminus \mathbf{y}_{N_i})}{\prod_{j \in a(N_i^E)} \phi(u_j)}.
$$

So a representation $\tilde{g}^i_{N_i}(.)$ of $g^i_{N_i}(.)$ should be found so that, given the function $h^i(\mathbf{u}_{N_i^E})$ exactly, the error

$$
\int \left(\tilde{g}^i_{N_i}(\mathbf{u}_{N_i}) - g^i_{N_i}(\mathbf{u}_{N_i}) \right) h^i(\mathbf{u}_{N_i^E}) d\mathbf{u}_{N_i}
$$

is as small as possible. However, at stage i, $h^i(\mathbf{u}_{N_i^E})$ will not be known exactly. Since the role of $h^i(.)$ is to determine the location of points $\tilde{\mathbf{u}}_{N_i}$ where it is important to store $g^i(\tilde{\mathbf{u}}_{N_i})$ accurately, a rough approximation to $h^i(.)$ will suffice.

Let $g^{\text{na}}(\mathbf{u}_{N_i^E}|\mathbf{y} \setminus \mathbf{y}_{N_i})$ be a normal approximation to $g(\mathbf{u}_{N_i^E}|\mathbf{y} \setminus \mathbf{y}_{N_i})$. Writing

$$
\tilde{h}^i(u_{N_i^E}) = \frac{g^{\text{na}}(u_{N_i^E}|\mathbf{y} \setminus \mathbf{y}_{N_i})}{\prod_{j \in a(N_i^E)} \phi(u_j)}
$$

and $r_i(\mathbf{u}_{N_i}; \tilde{h}^i) = g^i_{N_i}(\mathbf{u}_{N_i}) \tilde{h}^i(\mathbf{u}_{N_i^E})$, we will aim to make the error

$$
\int \tilde{r}^i(\mathbf{u}_{N_i}; \tilde{h}^i) - r^i(\mathbf{u}_{N_i}; \tilde{h}^i) du_{N_i} = \int \left(\tilde{g}^i_{N_i}(\mathbf{u}_{N_i}) - g^i_{N_i}(\mathbf{u}_{N_i}) \right) \tilde{h}^i(\mathbf{u}_{N_i^E}) d\mathbf{u}_{N_i}
$$

as small as possible.

To do this, we will store a (sufficiently accurate) representation $\tilde{r}^i(.; \tilde{h}^i)$ of $r^i(.; \tilde{h}^i)$, then extract the representation of $g^i_{N_i}(.)$ by setting

$$
\tilde{g}^i_{N_i}(\mathbf{u}_{N_i}) = \frac{\tilde{r}^i(\mathbf{u}_{N_i}; \tilde{h}^i)}{\tilde{h}^i(\mathbf{u}_{N_i^E})}.
$$

4.5.4 Transformation to a new basis

Recall that the Laplace approximation to the likelihood is formed by approximating $g(\mathbf{u}|\mathbf{y},\theta)$ with a function proportional to the $N(\mu_\theta, \Sigma_\theta)$ density.

If we transform to a new basis $\mathbf{z} = D^{-1}(\mathbf{u} - \mu)$, where D is a diagonal matrix with $D_{ii} = \sqrt{[\Sigma_\theta]_{ii}}$, then we may write

$$g_z(\mathbf{z}|\mathbf{y},\theta) = g_u(D\mathbf{z} + \mu|\mathbf{y},\theta)\det(D),$$

and factorise as before as

$$g_z(\mathbf{z}|\mathbf{y},\theta) = \prod_{C \in M_1} g_C^1(\mathbf{z}_C)$$

where

$$g_C^1(\mathbf{z}_C) = f(\mathbf{y}_C|\mathbf{u}_C = D_C\mathbf{z}_C + \mu_C) \prod_{j \in a(C)} D_{jj}\phi(D_{jj}z_j + \mu_j).$$

Then the likelihood is just $\int g_z(\mathbf{z}|\mathbf{y},\theta)d\mathbf{z}$. The new normal approximation to $g_z(\mathbf{z}|\mathbf{y},\theta)$ is $N(0,\Omega)$, where Ω is the correlation matrix $D^{-1}\Sigma_\theta$. We write

$$r_z^i(\mathbf{z}_{N_i}) = g_z^i(\mathbf{z}_{N_i})\tilde{h}^i(D_{N_i^E}\mathbf{z}_{N_i^E} + \mu_{N_i^E}),$$

so that $r_z^i(.)$ may be roughly approximated by using the $N(0,\Omega_{N_i})$ density.

4.5.5 Storing a modifier to the normal approximation

After the transformation described in Section 4.5.4 has taken place, at stage i we want to store an approximate representation of $r_z^i(.)$, so that the error in $\int r_z^i(\mathbf{z}_{N_i})d\mathbf{z}_{N_i}$ is as small as possible. A rough normal approximation to $r_z^i(.)$ is already available, so we use this as a starting point for our representation of $r_z^i(.)$.

We will store the modifiers

$$\log c_z^i(\mathbf{z}_{N_i}) = \log r_z^i(\mathbf{z}_{N_i}) - \log \phi_{|N_i|}(\mathbf{z}_{N_i}, 0, \Omega_{N_i})$$

at some fixed grid of points for $\mathbf{z}_{N_i}$, and use cubic splines to interpolate between the grid points.

The construction of this grid of points, and the method for interpolating between them are discussed in Section 4.6. The grid of points will be constructed using combinations of a fixed set of knots in each direction. Since we will always have a standard normal approximation for each direction, this set of knots will remain

unaltered throughout the algorithm, and interpolation between grid points will only be required when we have to integrate over z_i. Furthermore, the knots which are used do not vary with θ, so we can reuse some computations when we approximate the likelihood at each new value of θ.

Note. The INLA method of Rue *et al.* (2009) provides a way to approximate the marginal posterior density of each random effect in a latent Gaussian model, by integrating out all the other random effects using the Laplace approximation, and storing the resulting approximated marginal density using a spline modification to a normal approximation. This has some similarities to the approach taken here, although INLA only uses a one-dimensional modification to a normal approximation, whereas we consider storage of a function of arbitrary dimension.

4.6 Interpolation methods

Suppose that $f(.)$ is a function on $\mathbb{R}^d$, for which we want to store an approximate representation. In the case of the sequential reduction method, we take $f(.)$ to be $\log c_z^i(.)$. We now give a brief overview of the interpolation methods based on full and sparse grids of evaluation points. Some of the notation we use is taken from Barthelmann *et al.* (2000), although there are some differences: notably that we assume $f(.)$ to be a function on $\mathbb{R}^d$, rather than on the d-dimensional hypercube $[-1,1]^d$, and we will use cubic splines, rather than (global) polynomials for interpolation.

4.6.1 Full grid interpolation

First we consider a method for interpolation for a one-dimensional function $f : \mathbb{R} \to \mathbb{R}$. We evaluate $f(.)$ at m_l points $s_1, \ldots, s_{m_l}$ and write

$$\mathcal{U}^l(f) = \sum_{j=1}^{m_l} f(s_j) a_j^l, \tag{4.1}$$

where the a_j^l are basis functions. The approximate interpolated value of $f(.)$ at any point x is then given by $\mathcal{U}^l(f)(x)$. A very simple case is given by piecewise constant interpolation, in which case $a_j^l(x)$ is an indicator that the closest interpolation point to x is s_j. In Section 4.6.3, we describe how cubic splines may be written in this form.

Here l denotes the level of approximation, and we suppose that the set of evaluation points is nested so that at level l, we simply use the first m_l points of a

fixed set of evaluation points

$$S = \{s_1, s_2, \ldots\}.$$

We assume that $m_1 = 1$, so at the first level of approximation, only one point is used, and $m_l = 2^l - 1$ for $l > 1$, so there is an approximate doubling of the number of points when the level of approximation is increased by one.

The full grid method of interpolation is to take m_{l_j} points in dimension j, and compute $f(.)$ at each possible combination of those points. We write

$$(\mathcal{U}^1 \otimes \ldots \otimes \mathcal{U}^d)(f) = \sum_{j_1=1}^{m_{l_1}} \ldots \sum_{j_d=1}^{m_{l_d}} f(s_{j_1}, ..., s_{j_d}) \left(a_{j_1}^{l_1} \otimes \ldots \otimes a_{j_d}^{l_d} \right),$$

where

$$(a_{j_1}^{l_1} \otimes \ldots \otimes a_{j_d}^{l_d})(x_1, \ldots, x_d) = a_{j_1}^{l_1}(x_1) \times \ldots \times a_{j_d}^{l_d}(x_d).$$

Thus, in the full grid method, we must evaluate $f(.)$ at

$$\prod_{j=1}^{d} m_{l_j} = O\left(\prod_{j=1}^{d} 2^{l_j} \right) = O\left(2^{\sum l_j} \right)$$

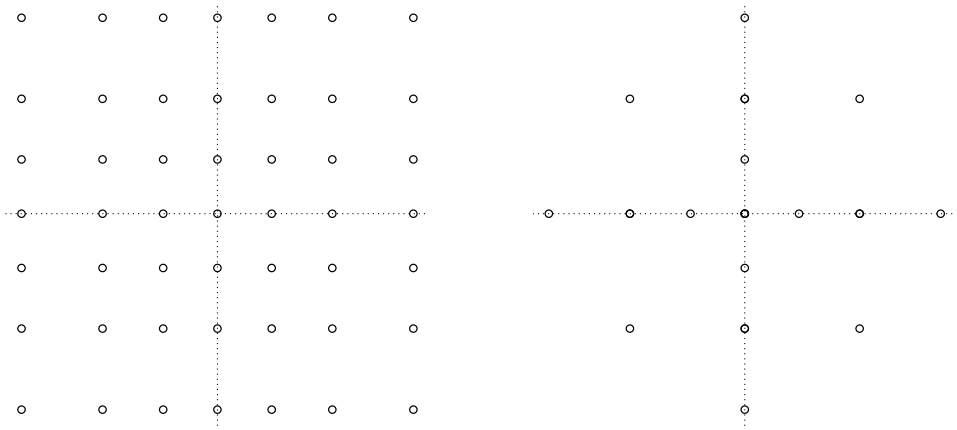
points. This will not be possible if $\sum_{j=1}^{d} l_j$ is too large.

4.6.2 Sparse grid interpolation

In order to construct an approximate representation of $f(.)$ in reasonable time, we could limit the sum $\sum_{j=1}^{d} l_j$ used in a full grid to be at most q, for some $q \geq d$. If $q > d$, there are many possibilities for 'small full grids' indexed by the levels $\mathbf{l} = (l_1, \ldots, l_d)$ which satisfy this constraint. A natural question is how to combine the information given by each of these small full grids to give a good representation overall. Such a representation may be built up sequentially, by specifying a method to update an approximate representation of $f(.)$, given the values of $f(.)$ evaluated at each new small full grid of points.

For a univariate function $f(.)$, let

$$\begin{aligned} \Delta^l(f) &= \mathcal{U}^l(f) - \mathcal{U}^{l-1}(f) \\ &= \sum_{j=1}^{m_{l-1}} f(s_j) \left[a_l^j - a_{l-1}^j \right] + \sum_{j=m_{l-1}+1}^{m_l} f(s_j) a_l^j, \end{aligned}$$



(a) Full grid (b) Sparse grid

Figure 4.3: A full and sparse grid, in two dimensions.

for $l > 1$, and $\Delta^1 = \mathcal{U}^1$. Then $\Delta^i(f)$ gives the quantity we should add to the approximate storage of $f(.)$ at level $l-1$, to incorporate the new information given by the knots added at level l.

Writing $k = q - d$, the sparse grid interpolation of $f(.)$ is given by

$$[f]_k = \sum_{\mathbf{l}:\sum l_i \leq d+k} (\Delta^{l_1} \otimes \ldots \otimes \Delta^{l_d})(f).$$

In the sparse grid method using all small full grids with $\sum l_i \leq d+k$, we must evaluate $f(.)$ at $O\left(d^{k+1}\right)$ points, which allows approximate storage for much larger dimension d than is possible using a full grid method.

Figure 4.3 gives a two-dimensional example of a full grid with $l_1 = l_2 = 3$, and the corresponding sparse grid with $k = 2$, which contains full grids indexed by levels (l_1, l_2) such that $l_1 + l_2 \leq 4$.

4.6.3 Interpolation using cubic splines

In order to make use of these interpolation methods, it is first necessary to write the one-dimensional interpolant in the form (4.1). We first consider how to do this for the cubic spline interpolant. This requires a little work, since the obvious form

of cubic splines is

$$f_{\text{interp}}(x) = \begin{cases} \alpha_0 + \beta_0(x - s_1) + \gamma_0(x - s_1)^2 + \delta_0(x - s_1)^3 & \text{if } x < s_1 \\ \alpha_i + \beta_i(x - s_i) + \gamma_i(x - s_i)^2 + \delta_i(x - s_i)^3 & \text{if } x \in [s_i, s_{i+1}) \\ \alpha_n + \beta_n(x - s_n) + \gamma_n(x - s_n)^2 + \delta_n(x - s_n)^3 & \text{if } x \geq s_n. \end{cases}$$

Stacking the coefficients as $\mathbf{c} = (\alpha_0, \beta_0, \gamma_0, \delta_0, \alpha_1, \beta_1, \gamma_1, \delta_1, \ldots, \alpha_n, \beta_n, \gamma_n, \delta_n)$, we may write $\mathbf{c} = D\mathbf{y}$, where $y_i = f(s_i)$ for a $4(n+1) \times n$ matrix D. The form of D varies according to the choice of end conditions. The method used to construct D for natural cubic splines is given in Appendix A.

Writing $b_{ij}(x) = \mathbf{1}\{x \in [s_i, s_{i+1})\}(x - s_i)^j$, for $i = 2, \ldots, n-1$, and $b_{1j}(x) = \mathbf{1}\{x < s_1\}(x - s_i)^j$, $b_{nj}(x) = \mathbf{1}\{x \geq s_n\}(x - s_i)^j$, we may stack the original basis functions as $\mathbf{b}(x) = (b_{00}, b_{01}, b_{02}, b_{03}, b_{10}, b_{11}, b_{12}, b_{13}, \ldots, b_{n0}, b_{n1}, b_{n2}, b_{n3})$. Then

$$f_{\text{interp}}(x) = \mathbf{c}^T\mathbf{b}(x) = [D\mathbf{y}]^T\mathbf{b}(x) = \mathbf{y}^T[D^T\mathbf{b}(x)] = \sum_{i=1}^{n} f(s_i)a_i(x)$$

if we write $\mathbf{a}(x) = D^T\mathbf{b}(x)$, so $\mathbf{a}$ is the set of basis functions we require for sparse grid interpolation.

Barthelmann *et al.* (2000) use global polynomial interpolation for a function defined on a hypercube, with Chebyshev knots. We prefer to use a spline-based approach, since the positioning of the knots is less critical. The choice of knots is discussed briefly in Section 4.6.4.

4.6.4 Choice of knots

We use sparse grid interpolation to store the function $\log c_z^i(.)$, which is a modifier to a $N(0, \Omega)$ density. The function will be stored using a sparse grid at level k, composed of small full grids with $\sum_i l_i \leq |N_i| + k$. Each small full grid will be constructed using a set of m_{l_i} knots in each direction i. Since we have a standard normal approximation for each direction, we use the same knots in each direction, and choose these standard knots $\mathbf{s}_l$ at level l to be m_l quantiles of a $N(0, \tau_k^2)$ distribution. As k increases, we choose larger τ_k, so that the size of the region covered by the sparse grid increases with k. However, the rate at which τ_k increases should be sufficiently slow to ensure that the distance between the knots $\mathbf{s}_k$ decreases with k. Somewhat arbitrarily, we choose

$$\tau_k = 1 + \frac{k}{2},$$

which appears to work reasonably well in practice. Appendix B gives a small numerical investigation of the impact of the choice of τ_k on the errors in the sequential reduction approximation to the likelihood.

4.7 Computational complexity of the sequential reduction algorithm

If the storage is done on a full grid with level $l = k + 1$ in each direction, or $m = 2^{k+1} - 1$ points in each direction, the cost of stage i of the algorithm is $O\left(m^{|N_i|}\right) = O\left(2^{k|N_i|}\right)$. The total cost of finding the likelihood is therefore $O\left(\sum_{i=1}^{n} 2^{k|N_i|}\right)$.

If the storage is done using a sparse grid composed of small full grids satisfying $\sum_i l_i \leq d + k$, the cost of stage i reduces to $O(|N_i|^k)$. In either case, the cost will be large if $\max_i |N_i|$ is large.

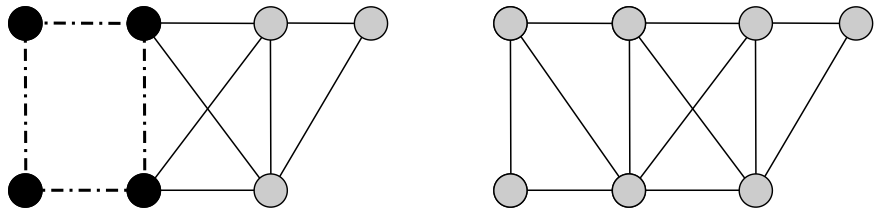
The random effects may be removed in any order, so it makes sense to use the ordering that allows approximation of the likelihood at minimal cost. This problem may be reduced to a problem in graph theory: to find an ordering of the vertices of a graph, such that when these nodes are removed in order, joining together all neighbours of the vertex to be removed at each stage, the largest clique obtained at any stage is as small as possible. This is known as the triangulation problem, and the smallest possible value, over all possible orderings, of the largest clique obtained at some stage is known as the treewidth of the graph.

Note. Treewidth is usually defined as *one less* than the size of this maximal clique. We follow Jordan (2004) in our definition of treewidth.

Unfortunately, algorithms available to calculate the treewidth of a graph on n vertices can take at worst $O(2^n)$ operations, so to find the exact treewidth may be too costly for n at all large. However,

1. There are special structures of graph which have known treewidth. The subgraph formed by taking some set of nodes and all edges involving only those nodes is called an **induced subgraph**. A **hole** in a graph is an induced subgraph which is a cycle of length 4 or greater. A graph is called **chordal** if it does not contain any holes. See Figure 4.4 for an illustration of these properties. The treewidth of a chordal graph is the size of its maximal clique. Furthermore, an elimination ordering achieving this treewidth can be found quickly, in $O(n + e)$ time, where e is the number of edges in the graph.

2. Upper bounds can be found on the treewidth by finding elimination orderings by certain heuristic algorithms, and finding the corresponding maximal clique



(a) A hole, so the graph is not chordal (b) A chordal graph

Figure 4.4: Illustration of the notion of a chordal graph

in the graph at any step. See, for example, Bodlaender & Koster (2008) for details of some of these algorithms.

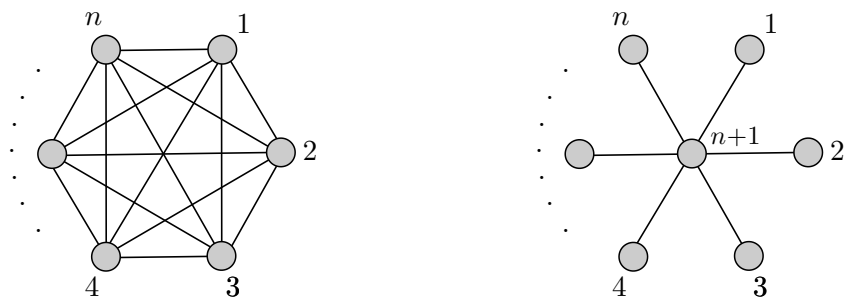
3. Lower bounds can be found on the treewidth. See, for example, Bodlaender & Koster (2010) for details of some algorithms.

In a typical case, where there is no special structure of the posterior dependence graph which may be exploited, we first find a lower bound and an upper bound on the treewidth. If this upper bound equals the lower bound, then we have found the treewidth exactly, and take the elimination ordering corresponding to that bound. If not, we might try some different methods for finding an upper bound to see if a better bound can be found, before using an elimination ordering corresponding to the best upper bound available.

Example 4.2. We return to the four-level random intercept model of Example 4.1. The posterior dependence graph of one such model is shown in Figure 4.1b. The graph is chordal, and its maximal cliques are triangles, so it has treewidth 3. This result may be extended to an l-level random intercept model, which has posterior dependence graph with treewidth $l - 1$.

Example 4.3. We return to a few of the examples of pairwise competition models we have already seen. A star tournament on n players has treewidth 2. A star is a special case of a **tree**: a connected graph with no cycles. Any tree has treewidth 2. An elimination ordering of the vertices which achieves this treewidth is given by removing a **leaf** vertex at each stage — a vertex with degree 1 in the remaining graph.

A complete tournament among n players has posterior dependence graph which is a complete graph on n vertices. The treewidth of this graph is n, and so the sequential reduction method does not provide a simplification of the likelihood in this case.



(a) Using the Cholesky decomposition (b) Adding an extra random effect

Figure 4.5: The posterior dependence graph after two different transformations in Example 4.4

Example 4.4. Suppose that we have a number of observations each of n items, where the distribution of an observation on item i depends on η_i, where

$$\eta_i = \beta^T \mathbf{x}_i + b_i.$$

We suppose that $\mathbf{b} = (b_1, \ldots, b_n) \sim N(0, \sigma^2\Omega(\rho))$, where $\Omega(\rho)$ is a correlation matrix with off-diagonal elements ρ. In all the examples given up to this point, the model has been specified with independent random effects. In this case, in order to use the sequential reduction method, we must first transform to a setting in which the random effects are independent.

To do this, we look for a matrix $A = A(\sigma, \rho)$, and a random effects vector $\mathbf{u} \sim N(0, I)$ so that $A\mathbf{u} \sim N(0, \sigma^2\Omega(\rho))$. Then we may write

$$\eta = X\beta + A\mathbf{u}.$$

The non-zero components of the rows of A determine which random effects are involved in each observation, which in turn determines the posterior dependence graph for $\mathbf{u}$. There are multiple choices for A and $\mathbf{u}$, and the aim of the example is to demonstrate that the choice of this transformation matters.

An obvious choice is to take A to be the Cholesky decomposition of $\sigma^2\Omega(\rho)$, so that $A^T A = \sigma^2\Omega(\rho)$, and A is upper triangular. In this case, if $\sigma > 0$ and $\rho > 0$, the first row of A has no non-zero components, so every component of $\mathbf{u}$ is involved in the first observation. The posterior dependence graph for $\mathbf{u}$ is therefore a complete graph on n vertices, as shown in Figure 4.5a, which has treewidth n.

Now we consider a different transformation, using a vector of independent random effects $\mathbf{u}$ of length $n+1$. We let $b_i = \sigma_1 u_i + \sigma_2 u_{n+1}$ for $i = 1, \ldots, n$, or

equivalently let A be the $(n+1) \times n$ matrix with components

$$A_{ij} = \begin{cases} \sigma_1 & \text{if } i = j \\ \sigma_2 & \text{if } j = n+1 \\ 0 & \text{otherwise,} \end{cases}$$

where $\sigma_1^2 = 2\rho\sigma^2$ and $\sigma_2^2 = (1-2\rho)\sigma^2$ are chosen to given $A\mathbf{u}$ the correct distribution. Now η_i only involves two random effects: u_i and u_{n+1}. The new posterior dependence graph is shown in Figure 4.5b, and has treewidth 2.

Notice that Figure 4.5a gives the posterior dependence graph of Figure 4.5b, after integrating out u_{n+1}. However, the optimal ordering is to remove u_{n+1} only after removing all the other random effects. This example shows that it is important to be careful about the method used to transform a model given with correlated random effects into one with independent random effects. Often, models with correlated random effects disguise a nested structure of a larger number of independent random effects. In such situations, the model should be written directly in terms of the full set of these independent random effects.

4.8 Using the sequential reduction method in practice

4.8.1 A program for the sequential reduction method

We have written code to implement one version of the sequential reduction method in R (R Core Team, 2012). In order to do this, it was also necessary to write code for sparse grid interpolation using cubic splines.

Example 4.6 gives a comparison of the sequential reduction method with an importance sampling approximation. Part of that comparison is the total time taken to maximise the likelihood approximation: using the ADMB package of Fournier *et al.* (2012) to maximise an importance sampling approximation, and the program developed here to maximise a sequential reduction approximation. As a comparison of the methods themselves, as apposed to any particular implementation of them, this is slightly unfair to the sequential reduction method, since the ADMB package has been carefully designed to be as computationally efficient as possible. An equally efficient implementation of the sequential reduction method may well be faster than the program that we have developed and used here. Nonetheless, the comparison turns out to be a favourable one for the sequential reduction method.

4.8.2 Maximising the approximated likelihood

It is faster to obtain a sequential reduction approximation to the likelihood using small k than it is with large k. For this reason, we first find the maximum of the approximation to the likelihood with $k = 0$ (equivalent to maximising the Laplace approximation to the likelihood), and then use the resulting estimator as a starting point for the optimisation of the approximation to the likelihood with $k = 2$ (we skip $k = 1$ since it typically gives similar results to $k = 0$). We continue in this manner, increasing k and using the previous maximum as the starting point for the optimisation, until the location of the maximum of the approximated likelihood has stabilised, or until some maximum permissable level of approximation $k_{\max}$ has been reached. Some ideas for modifications to this basic method are discussed in Section 4.10, which could potentially speed up the process of obtaining an approximation to the maximum likelihood estimator.

The sequential reduction method as implemented in the program we have developed and used here gives an approximation to the likelihood which is slightly rough for parameter values which are very close together. For this reason, we use the Nelder-Mead method for optimisation (Nelder & Mead, 1965), rather than a quasi-Newton method, since numerical differentiation methods are unreliable in such a setting.

4.8.3 Approximating the Hessian at the maximum

If we want to construct Wald-type hypothesis tests and confidence intervals, it is necessary to have an approximation to the Hessian matrix of the log-likelihood at its maximum. Since the sequential reduction approximation to the likelihood is not entirely smooth, it will not be possible to obtain a good approximation the second derivatives of the log-likelihood at its maximum using numerical differentiation methods directly. Instead, we make use of having a program to do sparse grid interpolation. The log-likelihood is evaluated at a sparse grid of points surrounding the maximum, and a cubic spline is fitted through these points. The Hessian is then approximated with the second derivatives of this spline at its maximum.

It was shown in Section 3.1.5 that Wald-type hypothesis tests have poor properties in some generalised linear mixed models, so we prefer to use likelihood ratio tests where possible, which do not require this Hessian matrix.

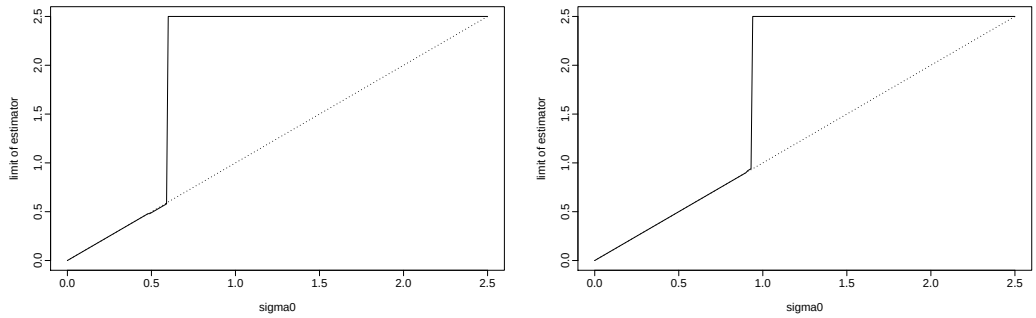
4.9 Examples

Example 4.5. We return to Example 3.2, and consider the asymptotic bias of estimators resulting from various likelihood approximations in a repeated star tournament. Recall Figure 3.3, showing that the Laplace and importance sampling approximations (using a fixed sample $\mathbf{t}$ to find the approximation for each possible observed tournament) both give estimators with a large asymptotic bias. Figure 4.6 shows the asymptotic bias in the estimators maximising the sequential reduction approximation to the likelihood, for various values of k. As k increases, the range of σ_0 for which asymptotic bias is close to 0 becomes progressively larger. The sequential reduction approximation gets gradually better with k, which is far preferable to the unstable convergence of the importance sampling approximation. Nonetheless, the method is not as accurate as using the 'manual' simplification of the likelihood described in Example 3.1. On the computer used here (a PC with a 3.4GHz Intel Core i7-2600 processor), the sequential reduction approximation of the likelihood at each point takes about 0.9 seconds for $k = 2$, 1.2 seconds for $k = 3$, 1.8 seconds for $k = 4$ and 3.6 seconds for $k = 5$, compared with just 0.03 seconds for the manual method. This suggests that improvements to the current version of the sequential reduction method may be possible, to create a general method which behaves more like the manual method in this special case.

Example 4.6. Consider observing a tournament with binary outcomes and structure as shown in Figure 4.7a. Suppose that there is a single observed covariate x_i for each player. A reasonable model for this tournament might be $Pr(i \text{ beats } j) = \Phi(\lambda_i - \lambda_j)$, where $\lambda_i = \beta x_i + \sigma u_i$, and $u_i \sim N(0, 1)$. This is a Thurstone-Mosteller model, as described in Section 2.2.2.

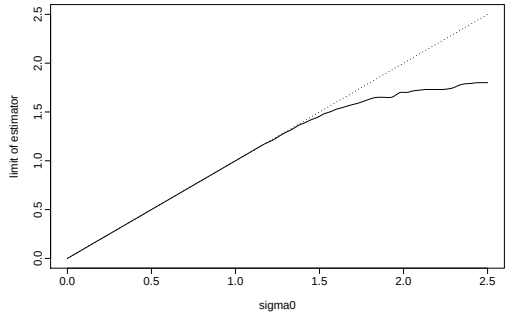
We consider one particular tournament with this tree structure, simulated from the model with the moderately large parameter values $\beta = 1.5$ and $\sigma = 1.5$. The covariates x_i are independent draws from a Bernoulli($\frac{1}{2}$) distribution. We aim to maximise the likelihood, penalised by the bias reduction penalty described in Section 3.1.6.

We compare the available options for approximating the likelihood. One possibility is to use an importance sampling approximation. We use the ADMB package (Fournier *et al.*, 2012) to maximise the importance sampling approximation to the penalised likelihood, and to output the Hessian at the maximum. From this, we can construct approximated standard errors for the parameters. Table 4.1 shows the output of this process, for different numbers of samples, N, used in the importance sampling approximation. The time taken by ADMB to maximise the

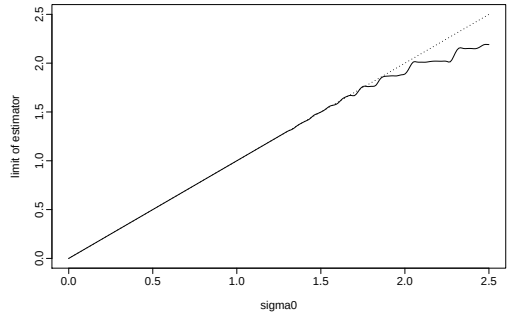


(a) $k = 2$

(b) $k = 3$

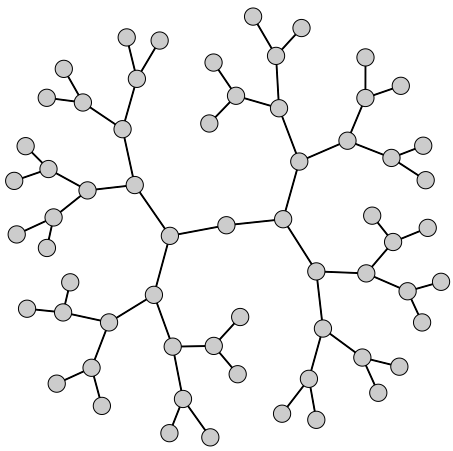


(c) $k = 4$

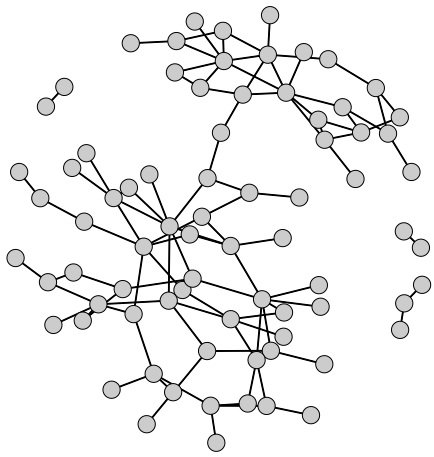


(d) $k = 5$

Figure 4.6: The limit of estimators maximising sequential reduction approximations to the likelihood, for various k, in a repeated star tournament with $n = 50$, as $R \to \infty$. In each case the dotted line is $y = x$, representing the limit of a consistent estimator.



(a) Tree tournament



(b) Lizards tournament

Figure 4.7: Some tournament designs

approximation to the penalised likelihood is also given.

Another option is to use the sequential reduction method to approximate the likelihood. The posterior dependence graph of a tree tournament is a tree, so has treewidth 2, as described in Example 4.3. Using sparse grid storage at level k, the cost of approximating the likelihood will therefore be $O(n2^k)$. The results of optimising the sequential reduction approximation to the penalised likelihood for this observed tournament, using sparse grid storage for various levels of approximation k, are also given in Table 4.1. Table 4.2 shows the average time required for each likelihood approximation, for each k.

We also consider testing for $\beta = 0$ or $\sigma = 0$ by using a likelihood ratio test, using the Laplace and sequential reduction approximations with bias reduction penalty. The p-values for the test of $\sigma = 0$ have not been adjusted for the non-standard asymptotic null distribution of the likelihood ratio test — see the discussion in Section 3.1.5. Although a likelihood ratio test could also be constructed for the importance sampling approximation, this method is not available in the ADMB software package used to obtain the importance sampling estimators.

Figure 4.8 shows the trace of an importance sampling approximation to the likelihood at its maximum $(1.20, 1.06)$. Notice that the number of samples, N, used in the approximation is plotted on a logarithmic scale, so that the approximation at each value of N marked on the axis takes 10 times longer than the previous value to compute. The sequential reduction approximations for different k are overlaid on the same plot. The first point corresponds to $k = 2$, and its x-coordinate N_2 is chosen so that the time to maximise the likelihood using sequential reduction with $k = 2$ is approximately the same as using ADMB with an importance sampling approximation using N_2 samples. The steps in the plot indicate an increase in the value of k, from $k = 2$ to $k = 5$, and the positions are multiples of N_2, where the multiple is chosen according to time taken for each likelihood approximation, as given in Table 4.2. There is no visible change at this scale in the sequential reduction approximations for $k = 4$ and $k = 5$, so we do not increase k any further.

Example 4.7. Recall that in Example 3.5 it was shown that for a complete tournament with a large number of players, the inference obtained from the Laplace approximation is quite accurate. We return to that example again, this time to demonstrate that we can use the sequential reduction method to detect the level of adjustment of the Laplace approximation which is required in any given case.

First, we simulate a complete tournament with $n = 20$ players, with $\beta = 1$ and $\sigma = 0.5$. The is a relatively dense model, and we have already demonstrated that the Laplace approximation gives inference close to the true likelihood in this setting.

Method	Time taken	Estim.	Approx. s.e.	p-values (Wald)	p-values (LR)
Laplace approximation	2 seconds	$(1.000, 0.759)$	$(0.526, 0.405)$	$(0.057, 0.061)$	$(0.0073, 0.055)$
Importance sampling ($N = 10^3$)	17 seconds	$(1.259, 1.124)$	$(0.722, 0.678)$	$(0.081, 0.097)$	-
Importance sampling ($N = 10^4$)	1.5 minutes	$(1.186, 1.034)$	$(0.648, 0.595)$	$(0.067, 0.082)$	-
Importance sampling ($N = 10^5$)	15 minutes	$(1.213, 1.072)$	$(0.710, 0.706)$	$(0.088, 0.13)$	-
Importance sampling ($N = 2 \times 10^5$)	31 minutes	$(1.206, 1.063)$	$(0.691, 0.672)$	$(0.081, 0.11)$	-
Importance sampling ($N = 5 \times 10^5$)	83 minutes	$(1.205, 1.061)$	$(0.691, 0.668)$	$(0.081, 0.11)$	-
Importance sampling ($N = 10^6$)	164 minutes	$(1.205, 1.061)$	$(0.693, 0.672)$	$(0.082, 0.11)$	-
Sequential reduction ($k = 2$)	1.3 minutes	$(1.215, 1.078)$	$(0.723, 0.728)$	$(0.093, 0.14)$	$(0.0074, 0.032)$
Sequential reduction ($k = 3$)	2.3 minutes	$(1.205, 1.060)$	$(0.711, 0.694)$	$(0.090, 0.13)$	$(0.0073, 0.032)$
Sequential reduction ($k = 4$)	3.9 minutes	$(1.206, 1.062)$	$(0.702, 0.682)$	$(0.086, 0.12)$	$(0.0073, 0.032)$
Sequential reduction ($k = 5$)	6.1 minutes	$(1.205, 1.061)$	$(0.705, 0.686)$	$(0.087, 0.12)$	$(0.0073, 0.032)$

Table 4.1: Parameter estimates for the tree tournament described in Example 4.6.

k	Time for each likelihood approximation	Number of evaluations at level k
2	1.3 seconds	61
3	1.6 seconds	37
4	2.4 seconds	37
5	4.2 seconds	33

Table 4.2: The average time to approximate the likelihood, for each value of k, in the tree tournament described in Example 4.6.

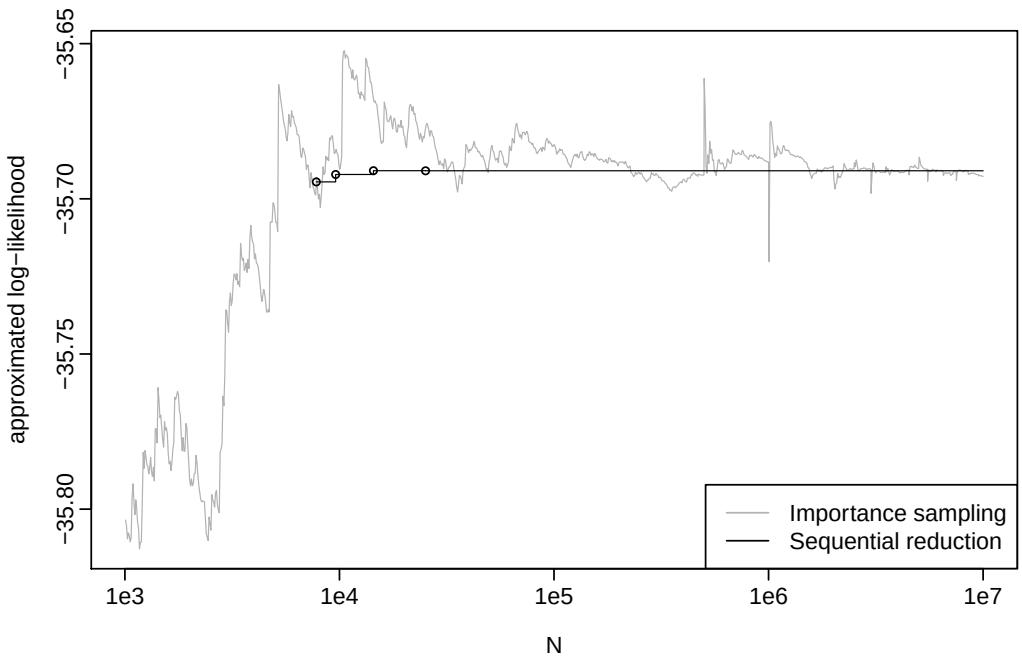


Figure 4.8: An importance sampling approximations to $\ell(1.20, 1.06)$, plotted against the number of samples, N, on a logarithmic scale. The sequential reduction approximations for different k are overlaid, with x-coordinates chosen to give a time comparison between the two methods.

We now show that it is possible to detect this using the sequential reduction method, and so deduce that it is not necessary to increase k above 2 in this case. We choose an example which is typical of the 100 cases simulated in Example 3.5, in terms of the level of error in the likelihood ratio test statistic. We first maximise the Laplace approximation to the likelihood, and find the corresponding confidence intervals for β and σ. The Laplace estimate is $(0.79, 0.45)$, with 95% confidence interval for β of $(0.03, 1.59)$, and for σ of $(0.16, 0.92)$. We use the sequential reduction method with $k = 2$ to improve the approximation to the likelihood, and maximise the approximated likelihood once more, using the Laplace estimator as a starting point. The resulting estimator is $(0.79, 0.45)$, which is equal to the Laplace estimator to two decimal places. This suggests that it may not be necessary to increase k any further to give good inference, but the location of the estimate alone is not sufficient to deduce this. Instead, we also recalculate 95% confidence intervals using the sequential reduction method with $k = 2$. This gives a confidence interval for β of $(0.03, 1.59)$, and for σ of $(0.16, 0.93)$, which again are very close to those obtained using the Laplace approximation. This allows us to conclude that there is no need to increase k any further.

Now we consider a sparser tournament: 19 repetitions of a complete tournament with $n = 5$ players. This example is less dense than the previous one, but not as sparse as the star tournament, where the Laplace approximation failed badly. It is therefore unclear how much modification to the Laplace approximation will be required. We repeat a similar procedure to that described above. The estimate maximising the Laplace approximation is $(1.03, 0.51)$. When we use the sequential reduction method with $k = 2$, the new estimate is $(1.04, 0.55)$. There has been a moderate change in the estimator, so we now increase k to 3 to see if the estimate stabilises. The resulting estimate is $(1.04, 0.55)$, which is the same as the estimate from $k = 2$ to two decimal places. We then check that the confidence intervals for each parameter have also stabilised. For $k = 2$, we get a 95% confidence interval for β of $(0.45, 1.81)$, and for σ of $(0, 1.09)$. For $k = 3$, we get a 95% confidence interval for β of $(0.45, 1.81)$, and for σ of $(0, 1.08)$. We may deduce that the inference has stabilised by $k = 3$, so it is unnecessary to increase k any further.

Example 4.8. Whiting *et al.* (2006) conducted an experiment to determine the factors affecting the fighting ability of male flat lizards. They observed a tournament consisting of 100 contests between $n = 77$ lizards. Various covariate information was collected on each lizard, and the aim of the study was to create a model for the fighting ability of a lizard, based on these covariates. The data are available in R as part of the BradleyTerry2 package (Turner & Firth, 2010). This package allows

analysis of pairwise competition models, both with and without random effects. When random effects are present, inference is conducted by using the Penalised Quasi Likelihood (PQL) of Breslow & Clayton (1993).

Whiting *et al.* (2006) assume a model with no random effects, so that the ability of lizard i,

$$\lambda_i = \beta^T \mathbf{x}_i,$$

is entirely determined by the value of some observed covariates, $\mathbf{x}_i$. This assumption is unrealistic in practice, so Turner & Firth (2012) suggest the introduction of a random effect for each lizard, letting

$$\lambda_i = \beta^T \mathbf{x}_i + \sigma u_i,$$

where $u_i \sim N(0, 1)$. The model without random effects is a special case of this model, where $\sigma = 0$.

Conditional on the abilities λ, Whiting *et al.* (2006) assume

$$Pr(i \text{ beats } j|\lambda_i, \lambda_j) = \text{logit}^{-1}(\lambda_i - \lambda_j).$$

After the addition of a normal random effect, it seems slightly more natural to assume a probit link, so that

$$Pr(i \text{ beats } j|\lambda_i, \lambda_j) = \Phi(\lambda_i - \lambda_j).$$

The choice of link function is unlikely to make any real difference to the results, after appropriate transformation of the parameter estimates.

The covariates included in the final model are the first (PC1) and third (PC3) principal components of the spectrum of the throat, the head length and the snout vent length (SVL) of each lizard. We consider fitting the model above using those covariates. Two lizards have missing values in some of these covariates, so we follow the suggestion of Turner & Firth (2012), and introduce a new covariate for each of those lizards, to allow their abilities to be modelled separately. The first row of Table 4.3 gives the estimators under the assumption of no random effects, and Wald-type p-values for testing the hypothesis that each parameter is zero. The second row provides the PQL estimators and the corresponding Wald type p-values, found using the BradleyTerry2 package. This is a reproduction of the analysis given in Turner & Firth (2012).

In order to use the sequential reduction method, we must first attempt to find an ordering in which to remove the players, an ordering which will minimise

the cost of the algorithm. The posterior dependence graph is not chordal, so more work is required to find the treewidth. Methods to find upper and lower bounds for the treewidth give that the treewidth is either 4 or 5. The upper bound gives us an ordering which may be used to evaluate the likelihood at cost $O(n5^k)$, using sequential reduction with sparse grid storage. The likelihood approximations took on average 2.38 seconds for $k = 2$, 3.81 seconds for $k = 3$ and 9.42 seconds for $k = 4$. The estimators from maximising the sequential reduction approximation to the likelihood, with bias-reduction penalty, are given in Table 4.3. The p-values from a (penalised) likelihood ratio test for the presence of each parameter are also given. The estimators and p-values are both quite stable for all $k \geq 2$. Even the Laplace approximation gives reasonably good inference in this case. However, the p-values from Wald tests based on PQL are highly inaccurate.

4.10 Some ideas for improving the method

The version of the sequential reduction method described in the last chapter, as programmed in R by the author, gives a substantial improvement on other approximation methods in sparse models, and also works reasonably well for denser models. There are, however, some areas in which the method, and the implementation of it, could be improved. Getting rid of inefficiencies in the code to approximate the likelihood has the potential to speed up the resulting program significantly. Some more specific ideas for improvements to the method are now considered.

Problems for large variance parameters

The current version of the sequential reduction method to approximate the likelihood sometimes fails completely at parameter values for which a variance parameter of the random effect is very large. For instance, in the tree tournament of Example 4.6, the method often fails for $\sigma > 3$. For such large values of σ, the normal density used as a basis for the sequential reduction method is a very poor approximation to $g(.|\mathbf{y}, \theta)$. An alternative baseline approximation may be useful in such cases.

Knot choice

Section 4.6.4 gives the method used to select knots to use for approximate storage. This method uses the quantiles of a $N(0, \tau_k^2)$ distribution, where the default choice for $\tau_k = 1 + \frac{k}{2}$. These choice was fairly arbitrary, and the numerical investigation in Appendix B suggests that the default choice of τ_k may grow slightly too quickly

	throat.PC1	throat.PC3	head.length	SVL	lizard096	lizard099	σ
Estim ($\sigma = 0$)	-0.055	0.21	-0.70	0.11	5.74	0.52	0 (fixed)
p-value ($\sigma = 0$, Wald)	0.00085	0.00099	0.016	0.051	0.98	0.43	-
Estim (PQL)	-0.054	0.23	-0.84	0.11	20.5	0.60	0.63
p-value (PQL, Wald)	0.021	0.0083	0.045	0.17	1.00	0.42	0.00073
Estim (Laplace, BR)	-0.089	0.32	-1.13	0.17	6.15	0.92	0.98
p-value (Laplace, BR, LR)	0.00026	0.00095	0.020	0.040	0.056	0.35	0.040
Estim ($k = 2$, BR)	-0.093	0.34	-1.20	0.18	5.85	0.99	1.12
p-value ($k = 2$, BR, LR)	0.00034	0.0011	0.021	0.060	0.057	0.36	0.035
Estim ($k = 3$, BR)	-0.096	0.35	-1.22	0.19	5.85	1.02	1.16
p-value ($k = 3$, BR, LR)	0.00034	0.0012	0.021	0.065	0.058	0.35	0.033
Estim ($k = 4$, BR)	-0.096	0.35	-1.23	0.19	5.85	1.01	1.16
p-value ($k = 4$, BR, LR)	0.00034	0.0012	0.021	0.065	0.057	0.35	0.033

Table 4.3: Parameter estimates for the lizards tournament

with k. It would also be worth investigating knots which are not the quantiles of a normal distribution. A better choice of knots could also help avoid problems in approximating the likelihood at extreme parameter values. If a variance parameter is large, it seems that the knots used should be closer together than those used for approximating the likelihood for smaller values of the variance parameter. Further work is needed to construct an improved default choice of knots, which may vary smoothly with θ.

Maximising the approximated likelihood

Even if we had an entirely efficient version of the sequential reduction method to approximate the likelihood, the question of how best to use the method to approximate the location of the maximum of the true likelihood would remain. Section 4.8.2 describes how this is done in the current version of the program developed for this work, using the estimate maximising the sequential reduction approximation to the likelihood at level k as a starting point for optimisation of the approximation at level $k+1$. This sort of strategy seems sensible, but there are some inefficiencies in the current method:

1. It should not be necessary to run the optimisation algorithm to completion at each stage k, since the location of the maximiser of the approximation only needs to be known roughly, to be used as a starting point for the next stage. As k gets larger, we should seek the location of the maximum with higher precision.

2. We should make more use of what we have learnt about the shape of the true likelihood surface based on previous iterations. For instance, it would be useful to be able to identify regions of the parameter space which we may be confident have very low likelihood compared to the maximum. If such a region had been identified at level k, it would not be necessary to revisit this part of the parameter space again at level $k+1$. To determine these regions, we would need an estimate of the error in the approximation at level k of the difference in the likelihood between any two points.

Profile likelihood confidence intervals

We would like to use the profile likelihood to construct more reliable confidence intervals of a parameter θ_i. To find the profile likelihood, we hold θ_i fixed, and seek to maximise the likelihood over the remaining components of θ. The maximiser of this

constrained likelihood, $\hat{\theta}(\theta_i)$, may be found in a similar way to the unconstrained maximum likelihood estimator, starting by maximising the Laplace approximation, and gradually increasing the level of approximation, using the estimate from the previous level as a starting point. In the method currently implemented, we specify a level of approximation, k, used to construct the profile likelihood confidence interval, and stop the algorithm to approximate $\hat{\theta}(\theta_i)$ when that level is reached. The approximation to the profile likelihood is then given by the level-k sequential reduction approximation to the likelihood at $\hat{\theta}(\theta_i)$.

We do not necessarily need a good approximation to the profile likelihood value at any given point, but only to the shape of profile likelihood surface. The method currently used in our program does not use any information about the shape of the profile likelihood which could be gained using lower levels of approximation. Even for small k, there are many points θ_i which we could confidently place either inside or outside a confidence region of some given coverage. It is only those points around the boundary for which we need a more accurate approximation of the profile likelihood surface. Even then, if the set of points which we cannot confidently classify as either inside or outside the confidence region is small, simply including them in the region will only result in a slightly conservative confidence region for θ_i. Again, the availability of an estimate of the error in the sequential reduction approximation would be very useful here.

It is no longer the precise location of $\hat{\theta}(\theta_i)$ which concerns us, but rather the value of the likelihood at that point. This means that it is unnecessary for the optimisation procedure to spend a long time finding the precise location of the maximum, if the maximum value of the approximated likelihood has already converged to within an acceptable tolerance.

In the current version of our program, it is too time consuming to find profile likelihood confidence intervals if the number of parameters is moderately large. For example, in the lizards tournament of Example 4.8, it was only possible to compute the profile likelihood at a single point for each parameter, in order to test for the necessity of including each term in the model. It would be useful to obtain a full confidence interval for each parameter. Improvements to the methods both for approximating the likelihood, and for finding profile-likelihood confidence intervals based on those approximations, could help to achieve this goal.

In cases where finding a profile-likelihood confidence interval for each parameter remains prohibitively expensive, an alternative would be to store the log-likelihood at a sparse grid of parameter values, and use an approximated log-likelihood surface, found by interpolating between these points, to compute ap-

proximate profile-likelihood confidence intervals.

Error estimation in the sequential reduction method

In order to improve the methods for maximising the approximated likelihood, it would be useful to have a method of estimating the error in the approximation to the likelihood. We are interested in the shape of the log-likelihood surface, rather than the value of the log-likelihood at any one point, so we might estimate the error in the approximation of the difference $\ell(\theta)-\ell(\tilde{\theta})$ between the log-likelihood at θ, and at some reference parameter value $\tilde{\theta}$. To estimate the error in the approximation to $\ell(\theta)-\ell(\tilde{\theta})$ found by using the sequential reduction method at level k, we could also approximate the same quantity at level $k-1$, and use the difference between the two approximations as a (conservative) estimate of the error of the level-k approximation.

4.11 Conclusions

The sequential reduction method described in this chapter allows a good approximation to the likelihood to be found in many models with sparse structure — precisely the situation in which currently used approximation methods perform the worst. This is achieved by exploiting the dependence structure of the random effects in the posterior distribution, and making full use of the factorisation of the non-normalised posterior density which is implied by this structure. By rephrasing the problem of approximating the likelihood as that of finding the normalising constant of a non-normalised density with conditional independence structure represented by an undirected graphical model, we are able to apply methods from graphical modelling to approximate the likelihood.

We have written code in R to implement one version of the sequential reduction method. This method uses sparse grid interpolation to store a modifier to the normal approximation used to construct the Laplace approximation. There was no package available in R for sparse grid interpolation, so we developed our own code, enabling cubic spline interpolation on a sparse grid in a fairly general setting. This has many potential uses, not only in approximating the likelihood for a generalised linear mixed model.

Little modification to the normal approximation used to construct the Laplace approximation is required in models with dense structure, so it is possible to use a sequential reduction method with sparse grid storage to get a sufficiently good approximation to the likelihood to use for inference in a wide range of models. Furthermore, it is possible to detect the level of modification, k, which is required in

any particular case, by checking the stability of inference from the sequential reduction method. There is some scope for further improvements to the method, in order reduce the time required to make reliable inference about the model parameters.

Chapter 5

Robustness to model misspecification

5.1 Introduction

In Chapter 2, generalised linear mixed models were introduced as an extension to generalised linear models, allowing for extra heterogeneity between items through the inclusion of random effects. Up until now, we have usually made the assumption that these random effects are normally distributed, but there is often no good reason to suppose that this should be true. It is therefore of interest to study the sensitivity of inference to deviations from this assumed random-effects distribution.

In Section 5.2.1, we apply the asymptotic results reviewed in Section 3.1.1, which provide a derivation of some well-known results on the limit and asymptotic distribution of the maximum likelihood estimator under model misspecification. These results tell us that the maximum likelihood estimator is typically not exactly consistent under misspecification of the random-effects distribution, but they do not give an indication of how large the asymptotic bias might be in practical settings. A substantial amount of research has been conducted into these practical consequences of random-effects misspecification, and in Section 5.2.2 we review some of the relevant literature.

In Section 5.3 we consider how the level of sparsity of a model affects the susceptibility of the maximum likelihood estimator to random-effect misspecification. We conclude that estimators of regression coefficients in models with sparse structure are more sensitive to the assumed random-effect distribution than those in dense models.

Moving to a more general framework, in Section 5.4 we consider the robust-

ness of composite likelihood estimators (as introduced in Section 2.3.4) to model misspecification. One motivation for the use of low-order marginal composite likelihoods is that they only require the specification of the distributions of some small subsets of the data, rather than the entire distribution of the data. This means that in some cases composite likelihood estimators have potential for increased robustness over the full maximum likelihood estimator. In Section 5.4, we consider the sort of setting in which such increased robustness holds.

We then return to the issue of robustness to misspecification of the random-effects distribution in a generalised linear mixed model. It is shown that if a low-order marginal composite likelihood is used for inference in a model with dense structure, it can be inconsistent even though the maximum likelihood estimator is consistent. This provides a warning that it should not be assumed that a composite likelihood estimator will always be at least as robust to model misspecification as the full maximum likelihood estimator.

5.2 The impact of misspecification of the random-effects distribution

5.2.1 Asymptotics under independent replication

Suppose that we have a generalised linear mixed model, where the distribution of the observations $\mathbf{Y} = (Y_1, \ldots, Y_m)$ depends on the linear predictor

$$\eta = X\beta + Z(\psi)\mathbf{u},$$

where β and ψ are unknown parameters, $\mathbf{u} = (u_1, \ldots, u_n)$, and we model the u_i as independent $N(0, 1)$ variables.

Now suppose that each u_i does not actually have $N(0, 1)$ distribution, but instead has some other distribution, with density $f_u(.)$, where $\mathbb{E}(u_i) = 0$ and $\mathrm{Var}(u_i) = 1$. We assume that the rest of the model is correctly specified, and that we are primarily interested in the quality of inference made about β, which has true value β_0.

In order to study the asymptotics of the maximum likelihood estimator under misspecification, we return to the results reviewed in Section 3.1.1, which give the limit and asymptotic distribution of an estimator found by solving an estimating equation, under independent replication. We can use these general results to derive the limiting distribution of the maximum likelihood estimator under model misspecification, giving the well-known results of White (1982).

If we write $\tilde{\ell}(\theta|\mathbf{y})$ for the log-likelihood of θ under the (incorrect) assumption of normal random effects, and $\tilde{u}(\theta|\mathbf{y}) = \nabla_\theta \tilde{\ell}(\theta|\mathbf{y})$ for the corresponding score function, then Theorem 1 tells us that, as the number of independent replications $R \to \infty$, the maximum (misspecified) likelihood estimator $\tilde{\theta}_R$, tends towards a value θ^* which solves

$$\mathbb{E}\left[\tilde{u}(\theta|\mathbf{Y})\right] = 0,$$

where the expectation is taken over the true distribution of $\mathbf{Y}$.

If we write

$$\tilde{H}(\theta) = \mathbb{E}\left[-\nabla_\theta \tilde{u}(\theta|\mathbf{Y})\right]$$

and

$$\tilde{J}(\theta) = \mathbb{E}\left[\tilde{u}(\theta|\mathbf{Y})[\tilde{u}(\theta|\mathbf{Y})]^T\right],$$

then Theorem 2 tells us that

$$\sqrt{R}(\tilde{\theta}_R - \theta^*) \to^d N_p\left(0, \tilde{H}(\theta^*)^{-1}\tilde{J}(\theta^*)[\tilde{H}(\theta^*)^T]^{-1}\right)$$

as $R \to \infty$. Under misspecification, the second Bartlett identity no longer holds, so we may not simplify the variance of this limiting distribution.

When can as estimator of β, or inference about β more generally, said to be robust against misspecification of the random-effects distribution? There are various properties which may be of interest:

1. Consistency. Does the estimator of β still tend towards β_0 as the amount of information in the data increases? The estimator will be consistent if $\beta^* = \beta_0$.

2. Efficiency. Does the estimator of β have the same asymptotic variance as if we had correctly specified the random-effects distribution? This will be the case if $[\tilde{G}(\theta^*)^{-1}]_{\beta\beta} = [I(\theta_0)^{-1}]_{\beta\beta}$, where $I(\theta)$ is the Fisher information matrix under the correctly specified model, and $\tilde{G}(\theta^*) = \tilde{H}(\theta^*)^T\tilde{J}(\theta^*)^{-1}\tilde{H}(\theta^*)$ is the Godambe information matrix under the misspecified model.

3. Correct distribution of test statistics. Is the assumed asymptotic distribution of test statistics for β still valid under the misspecification of the random-effects distribution? For this to be true, in addition to requiring that $\beta^* = \beta_0$, we need the estimate of the variance of $\hat{\beta}$ (found by using the misspecified likelihood) to be accurate. This will be the case if the relevant component of the inverse of the Godambe information matrix, $[\tilde{G}(\theta^*)^{-1}]_{\beta\beta}$, is equal to the respective component of the inverse information matrix of the misspecified likelihood, $[\tilde{H}(\theta_0)^{-1}]_{\beta\beta}$.

Often, under misspecification of the random effects, these properties will not hold exactly. Instead, we check whether the impact of misspecification is small. For instance, if an estimator is inconsistent, but has small asymptotic bias, the practical consequences of this bias will often be small.

In the case in which assuming the misspecified model leads to an incorrect estimate of the variance of $\hat{\beta}$, standard test statistics do not have their usual asymptotic distributions, and various modified versions have been proposed. For instance, Kent (1982) studies the properties of the likelihood ratio test under model misspecification, and proposes a modified version which has the usual chi-squared distribution. However, this modification requires estimation of the asymptotic variance of the maximum likelihood estimator under misspecification, $\tilde{G}(\theta^*)^{-1}$. If the data consists of a large number of independent replications, it will be fairly easy to obtain such an estimate. However, we also consider some cases with no independent replication, in which $\tilde{G}(\theta^*)$ becomes difficult to estimate. This is related to the problem of how to construct versions of test statistics based on a composite likelihood, as discussed briefly in Section 3.1.5.

5.2.2 Review: sensitivity to random-effects distribution

Neuhaus *et al.* (1992) consider the impact of misspecification of the random-effect distribution on inference in a two-level random intercept model, with binary response. In this case, with a fixed number of observations in each cluster, the maximum likelihood estimator is shown to be inconsistent as the number of items tends to infinity, but an approximation to the limiting value shows that the asymptotic bias is usually small. Many other authors have studied the impact of misspecification of the random-effect distribution in a two-level model. McCulloch & Neuhaus (2011) give a review of a large number of such studies, and conclude that the asymptotic bias under misspecification of the random-effect distribution results is typically small, unless the misspecification is quite severe. Some examples of cases when the asymptotic bias may be moderately large are if the true random-effect distribution is binary, with a large variance (Agresti *et al.*, 2004), or when the variance of the true random-effect distribution varies with the observed covariates (Heagerty & Kurland, 2001). An exception to this is when estimating an intercept, when the asymptotic bias can be quite large, for only moderate departures from the assumed random-effects distribution (Heagerty & Kurland, 2001).

Some authors have also considered the impact of misspecification on efficiency or on the validity of standard hypothesis tests, although the body of work in these areas is far smaller than that for the asymptotic bias. Zhang & Davidian (2001)

investigate the impact of misspecification on efficiency in a linear mixed model, and find that although the asymptotic bias of the maximum likelihood estimator is very small, there is some increase in the variance of the estimator due to misspecification.

Litière *et al.* (2008) discuss problems with hypothesis testing, concluding that although the asymptotic bias of regression coefficients is typically small, the power and type I error rate of tests are severely affected by misspecification. Verbeke & Lesaffre (1997) consider modified test statistics based on an approximation to $\tilde{G}(\theta^*)$, which are shown to offer a great improvement on standard test statistics. However, if there is no independent replication in the data, it may not be easy to obtain a good approximation to $\tilde{G}(\theta^*)$.

Most of these studies use a fixed model structure, and examine the effect of varying the true random-effect distribution. In Section 5.3 we consider how the model structure itself could affect the sensitivity of inference to fixed changes in the random-effects distribution.

5.2.3 Review: testing for non-normality of random effects

If it is possible that the incorrect specification of the random-effects distribution could have serious consequences for inference, then it makes sense to check if there is any evidence in the data to suggest that such a misspecification has occurred. In the context of a linear mixed model, Lange & Ryan (1989) suggested estimating the values of each of the random effects $u_1, \ldots, u_n$, after fixing θ at the maximum likelihood estimate $\hat{\theta}$, and checking the distribution of these fitted values against the supposed normal random-effects distribution. In the case where there is a different amount of information available on each random effect, they suggest weighting the estimated random effects to reflect the variance of each estimated random effect. However, Verbeke & Lesaffre (1996) consider the quality of the estimates of the random effects in a linear mixed model, and show that there is often substantial shrinkage towards the supposed distribution, making it hard to detect deviations from the assumed distribution by examining these fitted values alone. Jiang (2001) explicitly discusses the role of the sparsity of the model in such methods, and concludes that they can only be expected to be successful if the amount of information available in the data on each random effect is large.

Instead of testing for non-normality, it is possible to fit a model with a more flexible random-effect distribution. For instance, Verbeke & Lesaffre (1996) suggest using a finite mixture of normal distributions in place of a single normal distribution. There is a tradeoff between robustness and consistency inherent in all such approaches, since the more parameters there are in the model, the more

difficult it becomes to estimate those parameters well. Thus, if there is no evidence in the data to suppose that the normal random-effects distribution is a poor fit, it seems preferable to use this instead of some more complicated distribution.

Verbeke & Molenberghs (2013) suggest checking for non-normality of the random-effect distribution using the gradient function, a statistic which may be used to check whether using a finite mixture of normal distributions would substantially improve the model fit. In the case where a single normal distribution fits well, the gradient function is close to 1, otherwise it is substantially larger than 1. One difficulty in this approach is that there is no formal test provided to determine how much larger than 1 the gradient function must be in order to reject the hypothesis that the random effects have a normal distribution.

5.2.4 Review: linear models with non-normal errors

We will show later than in a model with dense structure, misspecifying the random-effects distribution is asymptotically equivalent to misspecifying the distribution of the error term in a linear model. It is therefore useful to study the robustness of inference in a linear model to such misspecification.

Suppose that we have n observations $\eta_1, \ldots, \eta_n$, with

$$\eta_i = \beta^T \mathbf{x}_i + \sigma u_i,$$

where u_i are uncorrelated errors, with mean $\mathbb{E}(u_i) = 0$, and variance $\mathrm{Var}(u_i) = 1$. Making no further assumption about the form of u_i, the Gauss-Markov theorem states that the ordinary least squares estimator of β is unbiased, and has the lowest variance of all unbiased estimators of β which are linear in η. If we make the assumption that $u_i \sim N(0, 1)$, the maximum likelihood estimator of β is this ordinary least squares estimator.

So, if assume that the error distribution is normal, then even if this assumption is false, the maximum likelihood estimator is consistent as $n \to \infty$. The Gauss-Markov theorem does not guarantee that the estimator will be entirely efficient, only that it is at least as efficient as any estimator which is linear in the data. However, in practice the estimator often retains high efficiency.

It is not necessarily the case that an estimate of the variance of $\hat{\beta}$, constructed assuming the misspecified model were correct, is equal to the true variance of β. This means that standard test statistics no longer have their assumed asymptotic distributions, and modified versions are required.

Alternatively, we may check the assumed distribution of the error u_i by

examining the fitted residuals

$$\hat{u}_i = \eta_i - \hat{\beta}^T \mathbf{x}_i$$

to check that they appear to be normally distributed, and that their distribution does not vary with $\mathbf{x}_i$.

5.3 The impact of model structure on robustness

5.3.1 Two-level models

To investigate the effect of sparsity on the sensitivity of inference to the assumed random-effect distribution, we first consider a simple two-level model. Suppose that there are m_i observations $\mathbf{y}_{(i)} = (y_{i1}, \ldots, y_{im_i})$ on each of n items $i = 1, \ldots, n$. The distribution of Y_{ij} depends on $\eta_i = \beta^T \mathbf{x}_i + \sigma u_i$, where u_i are independent random effects, with density $f_u(.)$. We suppose that $\mathbb{E}(u_i) = 0$ and $\mathrm{Var}(u_i) = 1$. The model becomes more dense as the number of observations on each item, m_i, increases. We want to study the degree of sensitivity of the inference about β to the specification of $f_u(.)$, and the impact of the sparsity of the model on this sensitivity.

The likelihood for $\theta = (\beta, \sigma)$ is given by

$$L(\theta|\mathbf{y}) = \prod_{i=1}^{n} L_i(\theta|\mathbf{y}_{(i)}),$$

where

$$L_i(\theta|\mathbf{y}_{(i)}) = \int_{-\infty}^{\infty} \left[\prod_{j=1}^{m_i} f_y(y_{ij}|\eta_i = \beta^T \mathbf{x}_i + \sigma u_i) \right] f_u(u_i) du_i.$$

We now consider the behaviour of $L_i(\theta|\mathbf{y}_{(i)})$ as m_i increases.

Intuitively, for large m_i, it should possible to obtain an estimate of the value of the linear predictor η_i from the data $\mathbf{y}_{(i)}$, which will be close to the true value η_i^0, whatever the assumed distribution for u_i. This means that for sufficiently large m_i, inference given the data $\mathbf{y}_{(i)}$ should be similar to the inference we would obtain if we were given the true value η_i^0 of each linear predictor η_i. Thus, for large m_i, the problem is reduced to studying the impact of incorrectly assuming that the errors in a linear model are normally distributed, as reviewed in Section 5.2.4. We now provide some more detail to further justify this argument.

We first reparameterize the integral, to write the likelihood as an integral

over η_i. Let

$$g_i(\eta_i|\mathbf{y}_{(i)},\theta) = \left[\prod_{j=1}^{m_i} f_y(y_{ij}|\eta_i)\right] \frac{1}{\sigma} f_u\left(\frac{\eta_i - \beta^T\mathbf{x}_i}{\sigma}\right),$$

so that

$$L_i(\theta|\mathbf{y}_{(i)}) = \int_{-\infty}^{\infty} g_i(\eta_i|\mathbf{y}_{(i)},\theta)d\eta_i.$$

We may think of $g_i(.|\mathbf{y}_{(i)},\theta)$ as a non-normalised posterior density for η_i, given a prior $\frac{1}{\sigma}f_u(\frac{\eta_i-\beta^T\mathbf{x}_i}{\sigma})$ which shrinks η_i towards $\beta^T\mathbf{x}_i$. Provided that $\sigma > 0$, as m_i increases, the $\prod_{j=1}^{m_i} f_y(y_{ij}|\eta_i)$ term, which does not depend on θ, dominates the prior, so that the maximiser of $g_i(.|\mathbf{y}_{(i)},\theta)$ over η_i, $\hat{\eta}_i(\theta)$, loses its dependence on θ, and tends towards its true value $\eta_i^0 = \beta_0^T\mathbf{x}_i + \sigma_0 u_i^0$.

As m_i increases, $g_i(.|\mathbf{y}_{(i)},\theta)$ becomes well approximated by a normal density about $\hat{\eta}_i(\theta)$, and the relative error in the Laplace approximation tends to zero. Writing $\ell_i(\theta|\mathbf{y}_{(i)}) = \log L_i(\theta|\mathbf{y}_{(i)})$, as $m_i \to \infty$,

$$\ell_i(\theta|\mathbf{y}_{(i)}) = \log g_i(\hat{\eta}_i(\theta)|\mathbf{y}_{(i)},\theta) + \frac{1}{2}\log H_\theta(\hat{\eta}_i(\theta)) - \frac{1}{2}\log 2\pi + o(1),$$

where

$$H_\theta(\eta_i) = \frac{\partial^2}{\partial\eta_i^2}\log g_i(\eta_i|\mathbf{y}_{(i)},\theta).$$

So, for any two distinct θ_1, θ_2, the difference in log-likelihoods $\ell_i(\theta_1|\mathbf{y}_{(i)}) - \ell_i(\theta_2|\mathbf{y}_{(i)})$ is equal to

$$\begin{aligned}
&\log g_i(\hat{\eta}_i(\theta_1)|\mathbf{y}_{(i)},\theta_1) - \log g_i(\hat{\eta}_i(\theta_2)|\mathbf{y}_{(i)},\theta_2) + \frac{1}{2}\log H_{\theta_1}(\hat{\eta}_i(\theta_1)) - \frac{1}{2}\log H_{\theta_2}(\hat{\eta}_i(\theta_2)) + o(1)\\
&\quad = \log g_i(\eta_i^0|\mathbf{y}_{(i)},\theta_1) - \log g_i(\eta_i^0|\mathbf{y}_{(i)},\theta_2) + \frac{1}{2}\log H_{\theta_1}(\eta_i^0) - \frac{1}{2}\log H_{\theta_2}(\eta_i^0) + o(1)
\end{aligned}$$

since for any θ, $\hat{\eta}_i(\theta) \to^p \eta_i^0$ as $m_i \to \infty$. But

$$\frac{H_{\theta_2}(\eta_i)}{H_{\theta_1}(\eta_i)} = \frac{\frac{1}{m_i}\left\{\sum_{j=1}^{m_i}\frac{\partial^2}{\partial\eta_i^2}\log f_y(y_{ij}|\eta_i) + \frac{\partial^2}{\partial\eta_i^2}\log\left[\frac{1}{\sigma_2}f_u\left(\frac{\eta_i-\beta_2^T\mathbf{x}_i}{\sigma_2}\right)\right]\right\}}{\frac{1}{m_i}\left\{\sum_{j=1}^{m_i}\frac{\partial^2}{\partial\eta_i^2}\log f_y(y_{ij}|\eta_i) + \frac{\partial^2}{\partial\eta_i^2}\log\left[\frac{1}{\sigma_1}f_u\left(\frac{\eta_i-\beta_1^T\mathbf{x}_i}{\sigma_1}\right)\right]\right\}} \to^p 1$$

as $m_i \to \infty$, so

$$\begin{aligned}
\ell_i(\theta_1|\mathbf{y}_{(i)}) - \ell_i(\theta_2|\mathbf{y}_{(i)}) &= \log g_i(\eta_i^0|\mathbf{y}_{(i)},\theta_1) - \log g_i(\eta_i^0|\mathbf{y}_{(i)},\theta_2) + o(1)\\
&= \log\left[\frac{1}{\sigma_1}f_u\left(\frac{\eta_i^0-\beta_1^T\mathbf{x}_i}{\sigma_1}\right)\right] - \log\left[\frac{1}{\sigma_2}f_u\left(\frac{\eta_i^0-\beta_2^T\mathbf{x}_i}{\sigma_2}\right)\right] + o(1).
\end{aligned}$$

Up to the $o(1)$ term, this is exactly the difference in log-likelihoods which

would be obtained if we had observed η_i directly, from the linear model $\eta_i = \beta^T \mathbf{x}_i + \sigma u_i$, where u_i is assumed to be a sample from a distribution with density $f_u(.)$. In this linear model setting, the impact of the distribution of the error term has been well studied, and was reviewed in Section 5.2.4. The maximum likelihood estimator of β remains consistent as $n \to \infty$, even if $f_u(.)$ is misspecified. Returning to the two-level model, this means that if m_i is the same for all items i, the asymptotic bias in the estimator of β will tend to 0 as $m_i \to \infty$.

5.3.2 A numerical example

We now give a simple numerical example to demonstrate the impact of increasing m_i on the maximum likelihood estimator in a two-level model.

Example 5.1. Consider making m_i repeated observations on each of n items $i = 1, \ldots, n$, each of which has a binary covariate x_i associated with it, where $x_i \sim$ Bernoulli($\frac{1}{2}$). Suppose that each observation is binary, and that

$$Pr(Y_{ij} = 1 | \alpha, \beta, \sigma, u_i, x_i) = \Phi(\alpha + \beta x_i + \sigma u_i),$$

where we suppose $u_i \sim N(0, 1)$. We start with a slightly unrealistic example, which nonetheless provides some insight. Instead of treating σ as an unknown parameter of interest, we fix it at some constant value $\tilde{\sigma}$. We suppose that in truth, $u_i \sim N(0, 1)$ as assumed, but that $\sigma = \sigma_0 \neq \tilde{\sigma}$, and consider the impact of this misspecification.

Suppose that $\tilde{\sigma} = 1$, $\sigma_0 = 0.5$, $\alpha_0 = 0.5$, and $\beta_0 = 1$. The limits of $\hat{\alpha}$ for different m_i are given in Table 5.1a, and the limits of $\hat{\beta}$ in Table 5.1b. The asymptotic variance of the estimators of each parameter are also given, both under correct model specification $\sigma = \sigma_0$ and under misspecification $\sigma = \tilde{\sigma}$. We also give the value of the asymptotic variance which would be obtained if the misspecified model were assumed to be true. If this is different from the true variance under misspecification, the conclusions of hypothesis tests may by invalid. As m_i increases, the asymptotic bias in the estimator diminishes, and the variance of the estimator under misspecification converges towards that under correct model specification. However, the estimated variance is larger than the true variance of the estimator.

In Section 5.5, we consider the model used in Example 5.1 again, but with some more realistic deviations from the assumed random-effects distribution.

m_i	Correct model		Misspecified model		
	Limit α_0	Variance $\left[I^{-1}(\theta_0)\right]_{11}$	Limit α_*	Variance $\left[\tilde{G}^{-1}(\theta_*)\right]_{11}$	Estimated variance $\left[\tilde{H}^{-1}(\theta_*)\right]_{11}$
2	-0.50	7.24	-0.57	9.33	10.69
5	-0.50	3.43	-0.56	4.32	6.40
10	-0.50	2.16	-0.54	2.58	4.94
20	-0.50	1.53	-0.52	1.71	4.22
50	-0.50	1.14	-0.51	1.21	3.81
100	-0.50	1.02	-0.51	1.04	3.68
200	-0.50	0.95	-0.50	0.96	3.61

(a) Inference for α

m_i	Correct model		Misspecified model		
	Limit β_0	Variance $\left[I^{-1}(\theta_0)\right]_{22}$	Limit β_*	Variance $\left[\tilde{G}^{-1}(\theta_*)\right]_{22}$	Estimated variance $\left[\tilde{H}^{-1}(\theta_*)\right]_{22}$
2	1.00	14.48	1.14	18.33	21.18
5	1.00	6.87	1.11	8.69	12.82
10	1.00	4.33	1.08	5.15	9.88
20	1.00	3.05	1.05	3.42	8.45
50	1.00	2.29	1.02	2.41	7.62
100	1.00	2.03	1.01	2.09	7.35
200	1.00	1.90	1.01	1.93	7.22

(b) Inference for β

Table 5.1: Inference under misspecification of σ in a two-level model

5.3.3 Other generalised linear mixed models

The same sort of argument will apply to other generalised linear mixed models. As the amount of information available on each random effect increases, the impact of the specification of $f_u(.)$ on the maximum likelihood estimator shrinks. The maximum likelihood estimator will be consistent in the limit as both the number of random effects, and the amount of information available per random effect, simultaneously tend to infinity. In sufficiently dense models, the impact of model misspecification on the maximum likelihood estimator should therefore be small. However, in sparse models, the maximum likelihood estimator may be considerably more sensitive to the random-effects distribution. We demonstrate this sensitivity using a pairwise competition model with sparse structure.

Example 5.2. Consider a binary repeated star tournament, with three players in each star. The structure of this model is shown in Figure 3.10a on page 33. Notice that there are only 2 observations for every 3 random effects, so this model is sparser than the two-level model, where there were $m_i \geq 2$ observations for each random effect i. Suppose that each player has a binary covariate, simulated as independent draws from a Bernoulli $\left(\frac{1}{2}\right)$ distribution. We assume that $Pr(i \text{ beats } j) = \Phi(\lambda_i - \lambda_j)$, where

$$\lambda_i = \beta x_i + \sigma u_i,$$

and $u_i \sim N(0,1)$. As in Example 5.1, suppose that u_i really are $N(0,1)$ samples, but we fix σ at some constant value $\tilde{\sigma} \neq \sigma_0$. Table 5.2 gives the limit of the maximum likelihood estimator of β, when $\beta_0 = 1$ and $\sigma_0 = 0.5$, for various values of $\tilde{\sigma}$. It also gives the asymptotic variance of the estimator, which grows with $\tilde{\sigma}$. The estimated asymptotic variance is also given. In this model, this estimated variance is a good approximation to the true variance of the estimator.

$\tilde{\sigma}$	Limit β_*	Variance $\tilde{G}^{-1}(\beta_*)$	Estimated variance $\tilde{H}^{-1}(\beta_*)$
0.1	0.82	2.14	2.06
0.2	0.85	2.26	2.19
0.4	0.94	2.75	2.73
$0.5 = \sigma_0$	1.00	3.12	3.12
0.6	1.07	3.58	3.60
1	1.40	6.23	6.32
1.5	1.88	11.45	11.53

Table 5.2: Inference under misspecification of σ in a repeated star tournament.

Example 5.3. We consider exactly the same type of misspecification as used in Example 5.2, but now suppose that we have R repetitions of a complete tournament among n players, for various values of $n \geq 10$. These tournaments are much denser than the repeated star tournament of Example 5.2, and we want to investigate how this change in model structure affects the asymptotic bias of the maximum likelihood estimator of β. We recall from Example 3.5 that inference from the true likelihood is very well approximated by inference from the Laplace approximation to the likelihood in a complete tournament between a large number of players, so we use the Laplace approximation in place of the full likelihood here.

We simulate N_n tournaments for each $n \in \{10, 20, 40, 100\}$, with the ability of each player simulated as

$$\lambda_i = \beta_0 x_i + \sigma_0 u_i,$$

where $x_i \sim \text{Bernoulli}\left(\frac{1}{2}\right)$ and $u_i \sim N(0, 1)$. We (wrongly) suppose that

$$\lambda_i = \beta x_i + \tilde{\sigma} u_i,$$

and try to estimate β. We take the number of simulations N_n large enough that we can be confident of the limiting value of the maximum likelihood estimator of β, as $R \to \infty$.

Suppose that $\beta_0 = 1$ and $\sigma_0 = 0.5$, but $\tilde{\sigma} = 1$. Under the same misspecification, the limit of $\hat{\beta}$ in the repeated star tournament of Example 5.2 was 1.40. The limits of the Laplace estimator as $R \to \infty$ for each repeated complete tournament are given in Table 5.3. For $n = 10$, there is a small asymptotic bias in the estimator even in the correct model. This is because the estimator is computed by maximising the Laplace approximation to the likelihood, rather than the true likelihood. However, the magnitude of this bias is negligible compared to that induced by the misspecification of σ.

As n increases, the limit of maximum likelihood estimator of β becomes closer to β_0. In this case, the variance of the maximum likelihood estimator under misspecification tends towards the variance under correct model specification as n increases. However, the estimated variance under misspecification is too large in each case, which would make standard hypothesis tests too conservative.

Note. We have demonstrated that if there are R independent replications of a complete tournament among n players, the asymptotic bias in $\hat{\beta}$ tends to zero as n grows. In fact, we do not need a growing number of independent replications in order for this result to hold. To see this, note that a complete tournament among n players contains within it $\sqrt{n}$ tournaments, each with $\sqrt{n}$ players. For a single replication

n	Correct model		Misspecified model			N_n
	Limit β_0	Variance $I^{-1}(\beta_0)$	Limit β_*	Variance $\tilde{G}^{-1}(\beta_*)$	Estimated variance $\tilde{H}^{-1}(\beta_*)$	
10	1.01	0.23	1.24	0.36	0.59	10000
20	1.00	0.08	1.13	0.10	0.23	5000
40	1.00	0.03	1.07	0.03	0.11	100
100	1.00	0.01	1.04	0.01	0.04	100

Table 5.3: Inference under misspecification of σ in a repeated complete tournament

$R = 1$, by an argument similar to those given in Section 3.2, the maximum likelihood estimator of β will therefore be consistent as $n \to \infty$, even under misspecification of the random-effects distribution.

Note. In dense models, the Laplace, importance sampling and sequential reduction approximations to the likelihood will all give inference very close to the true likelihood (see the discussion of Section 3.3). This means that the robustness of the maximum likelihood estimator in a dense model is inherited by the estimators maximising any of these approximations.

5.4 Robustness of composite likelihood estimators

5.4.1 Robustness in a marginal framework

Suppose that we observe a vector of (dependent) observations $\mathbf{y} = (y_1, \ldots, y_m)$. We might be confident in specifying a good model for the marginal distributions of some subsets of data, $\left\{\mathbf{Y}_{(s)}\right\}_{s=1}^{K}$, but be unsure about how to model the dependence between these subsets. If we have to specify a full model for $\mathbf{Y}$, we may end up with a correct model for the marginal distribution of each $\mathbf{Y}_{(s)}$, depending on a 'marginal' parameter θ, but an incorrect model for the whole vector of observations, $\mathbf{Y}$.

In such a setting, if we construct a marginal composite likelihood based on $\left\{\mathbf{y}_{(s)}\right\}_{s=1}^{K}$, then the corresponding estimator of θ will be consistent as the amount of information on θ increases, since the model distribution for the subsets used to construct the composite likelihood is correctly specified. There is, however, no guarantee that the maximum likelihood estimator will be consistent, since the likelihood relies on the full, misspecified, distribution of $\mathbf{Y}$. Such potential for increased robustness of a composite likelihood estimator is discussed by Varin *et al.* (2011) and Xu & Reid (2011).

Example 5.4. Suppose that Y_i are binary responses, and that, marginally, we are confident that

$$Pr(Y_i = 1|\mathbf{b}, \mathbf{x_i}) = \Phi(\mathbf{b}^T\mathbf{x_i})$$

for some parameter $\mathbf{b}$, but that we are unsure about the dependence between Y_i. In such a setting, we could use the onewise likelihood

$$L^{\text{onewise}}(\mathbf{b}|\mathbf{y}) = \prod_{i=1}^{n} Pr(Y_i = y_i|\mathbf{b}, \mathbf{x_i})$$

to give a consistent estimator of $\mathbf{b}$, irrespective of the true dependence structure of $\mathbf{Y}$.

A marginal composite likelihood gives potential for increased robustness in situations where certain marginal distributions are correctly specified by the model, but the full distribution is not. If the marginal distributions are themselves misspecified, the full distribution of $\mathbf{Y}$ must also be incorrect. From this, it is tempting to conclude that a marginal composite likelihood estimator must always be at least as robust to model misspecification as the full likelihood estimator, but this is not the case. We will demonstrate this using a generalised linear mixed model, under misspecification of the random-effects distribution.

5.4.2 Another view of marginal composite likelihood

Recall that we construct a marginal composite likelihood by taking a product of the densities of some subsets of the observed data. It is possible to construct a new data vector $\tilde{\mathbf{y}}$, with model density $\tilde{f}(., \theta)$, so that the likelihood for θ given $\tilde{\mathbf{y}}$ is identical to the composite likelihood for θ given $\mathbf{y}$. To do this, we write $\tilde{\mathbf{y}} = (\mathbf{y}_{(s)}, s = 1, \ldots, K)$, and model the components indexed by different s as independent, with $\mathbf{Y}_{(s)}$ having the same marginal distribution as in the original model. The composite likelihood estimator on observing $\mathbf{y}$ under the original model will be identical to the full likelihood estimator on observing $\tilde{\mathbf{y}}$ under the alternative model.

Note. It is not claimed that the inference about θ based on a composite likelihood given $\mathbf{y}$ should be the same as the inference based on the full likelihood given $\tilde{\mathbf{y}}$. Recall from Section 3.1.5 that adjustments must be made to standard test statistics before they may be used for inference from a composite likelihood.

Example 5.5. Suppose we observe a complete tournament on n players, and construct the pairwise likelihood based on this tournament. Then $\tilde{\mathbf{y}}$ would consist of all

$3\binom{n}{3}$ pairs of matches involving a common player. Figure 5.1 gives $\tilde{\mathbf{y}}$ for a complete tournament with 4 players.

Example 5.6. We return to misspecification in a range of complete tournaments, as in Example 5.3. Recall that the asymptotic bias in the maximum likelihood estimator became smaller with n. Under the same misspecification, the limit of $\hat{\beta}$ in the repeated star tournament of Example 5.2 was 1.40. By the construction described in Example 5.5, this will also be the limit of the pairwise likelihood estimator in a repeated complete tournament, irrespective of the number of players, n, in each complete sub-tournament.

For a complete tournament, as $n \to \infty$, the maximum likelihood estimator of β is consistent under the misspecification of σ, whereas the pairwise likelihood estimator is inconsistent.

5.5 Some more realistic examples

All of the examples so far have involved fixing the parameter σ at an incorrect value $\tilde{\sigma}$. In practice, this is unrealistic, and instead σ will be treated as an unknown parameter, and estimated from the data. We now consider the impact of some more realistic deviations from the assumed random-effects distribution in each case.

We first review the different model structures we will use.

Two-level model

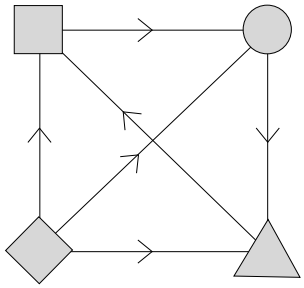
Suppose that we have m_i binary observations $y_{i1}, \ldots, y_{im_i}$ on each item $i = 1, \ldots, n$, and that $Pr(Y_{ij} = 1) = \Phi(\eta_i)$, where $\eta_i = \alpha + \beta x_i + \sigma u_i$. We suppose that $u_i \sim N(0, 1)$, and α, β and σ are unknown parameters, which we must estimate. We consider various different values of $m_i \in \{2, 5, 10, 20, 50\}$.

Pairwise competition models

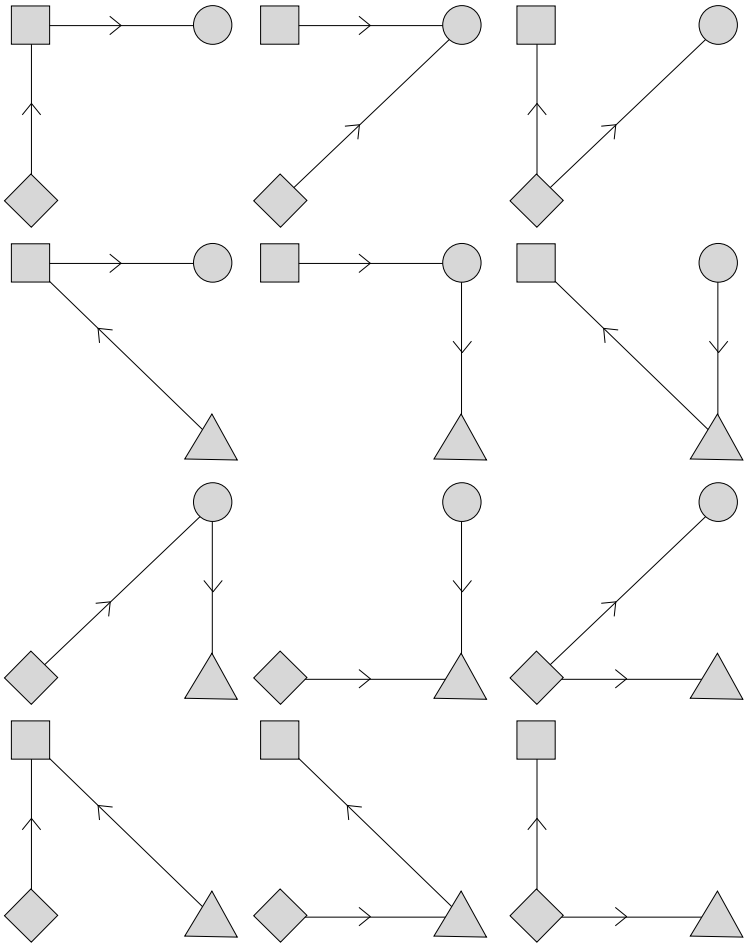
Suppose that we observe a tournament between players as in Examples 5.2 and 5.3, where there is a single covariate x_i, taking values in $\{0, 1\}$, for each player. We suppose that $Pr(Y_{ij} = 1|\lambda_i, \lambda_j) = \Phi(\lambda_i - \lambda_j)$, where we model the ability of each player as

$$\lambda_i = \beta x_i + \sigma u_i,$$

where we suppose that $u_i \sim N(0, 1)$, and β and σ are unknown parameters, which we must estimate.



(a) $\mathbf{y}$



(b) $\tilde{\mathbf{y}}$

Figure 5.1: Construction of $\tilde{\mathbf{y}}$ for a complete tournament among 4 players

We consider various different tournament structures. The sparse example is that of a repeated star tournament, with each sub-tournament being between three players. We also consider repeated complete tournaments, with $n \in \{10, 20, 40, 100\}$ players in each sub-tournament.

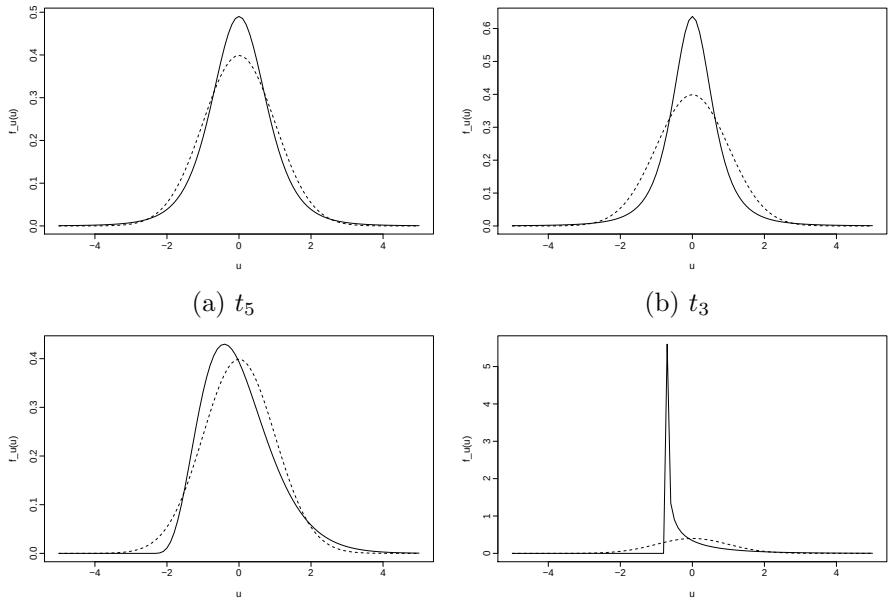
5.5.1 Non-normal random effects

Suppose that, in truth, u_i are independent and identically distributed, but have some non-normal distribution. We consider the same alternative random-effects distributions as used in the simulations of Neuhaus *et al.* (1992). These are transformed versions of t distributions with 5 or 3 degrees of freedom, and gamma distributions with scale parameter 1, and shape parameter 6 or 0.5. The transformation is so that in each case u_i has mean 0 and variance 1. The densities of the various distributions are shown in Figure 5.2.

For the two-level model, we suppose that α, β and σ take true values $\alpha_0 = -0.5$, $\beta_0 = 1$ and $\sigma_0 = 1$. The (approximate) limits of the estimators of α, β and σ for different values of m_i are given in Table 5.4. The limits may be slightly inaccurate for large m_i, because of the accumulation of small inaccuracies in numerical integration. The asymptotic bias in the estimator of β is extremely small for all m_i. In the cases in which the random effects have a non-symmetric distribution, some asymptotic bias is present in $\hat{\alpha}$, which shrinks as m_i grows. Finally, $\hat{\sigma}$ has a moderate bias for the t_3 and $\Gamma(0.5, 1)$ cases. Again, this bias shrinks with m_i, but only very slowly.

In the repeated 3-star tournament, with $\beta_0 = 1$ and $\sigma_0 = 1$, the limit of $\hat{\beta}$ was 1.01 for t_5, 1.00 for t_3, 0.98 for $\Gamma(6, 1)$ and 0.87 for $\Gamma(0.5, 1)$. In the $\Gamma(0.5, 1)$ case, which has the largest asymptotic bias of all the cases in the sparse tournament, the limit of the maximum likelihood estimator in a repeated complete tournament with 10 players is 0.98. In all the other cases, the asymptotic bias is very small even in the sparse repeated 3-star tournament, so we do not consider the limits of the estimator in the complete tournament.

In the cases we have considered so far, the asymptotic bias of $\hat{\beta}$ has been quite small. In order to find an example in which there is a moderately large bias in $\hat{\beta}$ for sparse models, we consider an example in which σ may vary with the observed covariate x_i.



(a) t_5 (b) t_3

(c) $\Gamma(6,1)$ (d) $\Gamma(0.5,1)$

Figure 5.2: The densities of the true distributions for u_i. The dotted line gives the assumed $N(0,1)$ density.

m_i	α_*	β_*	σ_*
2	-0.50	1.00	0.90
5	-0.50	1.00	0.90
10	-0.50	0.99	0.90
20	-0.49	0.98	0.91
50	-0.50	1.00	0.91
100	-0.51	1.02	0.93

(a) t_5

m_i	α_*	β_*	σ_*
2	-0.50	1.00	0.76
5	-0.50	1.00	0.76
10	-0.50	0.99	0.77
20	-0.49	0.99	0.78
50	-0.49	0.98	0.80
100	-0.49	0.99	0.81

(b) t_3

m_i	α_*	β_*	σ_*
2	-0.56	1.00	0.97
5	-0.55	1.00	0.97
10	-0.55	1.00	0.97
20	-0.54	1.00	0.97
50	-0.53	0.98	0.99
100	-0.51	1.02	0.99

(c) $\Gamma(6,1)$

m_i	α_*	β_*	σ_*
2	-0.63	1.00	0.75
5	-0.63	1.00	0.75
10	-0.61	1.00	0.76
20	-0.58	0.99	0.78
50	-0.52	0.99	0.80
100	-0.51	1.00	0.83

(d) $\Gamma(0.5,1)$

Table 5.4: The limits of estimators for a repeated star tournament, with various non-normal random-effect distributions. The true parameter values are $\alpha_0 = -0.5$, $\beta_0 = 1$ and $\sigma_0 = 1$.

5.5.2 A binary, heteroscedastic random-effects distribution

Suppose that in truth

$$u_i = \begin{cases} 1 & \text{with probability } 0.5 \\ -1 & \text{with probability } 0.5 \end{cases}.$$

Recall that u_i enters the linear predictor through σu_i, where in the model σ is assumed to be constant for all values of the observed covariate x_i. We now suppose that in truth $\sigma = \sigma(x_i)$ varies with x_i, and study the case in which $\sigma(1) = 1$, $\sigma(0) = 0.1$.

For the pairwise competition models, we assume that β has true value $\beta_0 = 1$. For a repeated star tournament with 3 players in each star, the limit of the maximum likelihood estimator of β in this case is 0.85.

We also consider the limit of the maximum likelihood estimator of β in a repeated complete tournament, with various choices for the number of players, n, in each tournament. For $n = 10$, $\hat{\beta}$ tends towards 0.89, for $n = 20$, 0.89, for $n = 40$, 0.96, and for $n = 100$, 0.97. As n increases, the asymptotic bias in the estimator of β becomes smaller, although the convergence is quite slow. The limit of the pairwise likelihood estimator remains fixed at 0.85, irrespective of n.

We also consider the two-level model with the same type of misspecification, where α and β again have true values $\alpha_0 = -0.5$ and $\beta_0 = 1$. The maximum likelihood estimator for β is approximately unbiased for all m_i in this case, but the estimator for α (which has true value -0.5) tends towards -0.61 for $m_i = 2$, -0.58 for $m_i = 5$, -0.55 for $m_i = 10$, -0.53 for $m_i = 20$, and -0.51 for $m_i = 50$. The asymptotic bias in $\hat{\alpha}$ again shrinks as m_i grows.

5.6 Conclusions

The level of sparsity of a generalised linear mixed model affects how sensitive the maximum likelihood estimator is to the assumed random-effects distribution. In models with sufficiently dense structure, a lot of information is provided by the data about the true value of each random effect, and the assumed distribution from which these random effects were drawn has little impact. The maximum likelihood estimator has small asymptotic bias, and the variance of the estimator is often similar to that of the estimator which would have been obtained using the true random-effects distribution. However, if we assume that the model is correctly specified, and use standard hypothesis tests, the resulting inference may be incorrect.

In a model with dense structure, the maximum likelihood estimator itself is quite robust to misspecification of the random-effects distribution, but the inference found using the misspecified likelihood need not be. It is therefore worthwhile to attempt to test for evidence of non-normal random effects, even for models with dense structure. Methods for testing assumptions about the random-effects distribution which rely on estimates of the values of the random effects will be reliable if the model is sufficiently dense.

In a sparse model, the maximum likelihood estimator itself need not be robust to misspecification, since in this case the data only provide a small amount of information on each random effect, and the assumed distribution of the random effects still has a sizeable impact on the likelihood. Furthermore, it is more difficult to check for departures from the assumed normal random-effects distribution in a sparse model.

These results have implications for the robustness of low-order composite likelihood estimators. We can construct a new sparse dataset so that the composite likelihood estimator given the original data is identical to the full maximum likelihood estimator for the new data. In a model with sufficiently dense structure, the maximum likelihood estimator is robust to model misspecification, but the composite likelihood estimator need not be. By only making use of marginal distributions of small sets of components, a low-order composite likelihood typically discards a large amount of available information on each random effect. This provides a counterexample to the notion that a composite likelihood estimator must always be at least as robust to model misspecification as the maximum likelihood estimator.

These findings have little practical impact for many possible deviations from the assumed normal random-effects distribution, since in many cases the asymptotic bias of the maximum likelihood estimator is very small, even in sparse models. However, in some slightly more extreme settings, such as if the true random-effect distribution may depend on the value of the observed covariates, the asymptotic bias of the maximum likelihood estimator may be non-negligible in some sparse models.

Chapter 6

Concluding remarks and further work

6.1 The thesis in brief: an overview of the main findings

In this thesis we have discussed various aspects of inference in generalised linear mixed models. Much of the thesis has focussed on alternatives and approximations to the likelihood, for use in models in which there is no obvious simplification to the high-dimensional integral involved in the likelihood. In Chapter 3, the performance of some of these alternatives was examined, with particular attention paid to the role that the sparsity of the model has on the success of each method. The Laplace approximation to the likelihood was shown to fail in very sparse models, in which little information is available on each random effect. However, the empirical evidence of Example 3.5 suggests that inference based on the Laplace approximation will be close to that from the true likelihood in models with $m = O(n^2)$ observations and n random effects, even though in that setting the relative error in the approximation to the likelihood at each point need not tend to zero as $n \to \infty$.

The other alternatives we considered for inference were also shown to fail in some cases: in Example 3.2 the convergence of the importance sampling approximation was slow and unstable, and in Example 3.1 the pairwise likelihood estimator which has been suggested for use in generalised linear mixed models suffered a large loss of efficiency compared to the full maximum likelihood estimator.

Motivated by these results, in Chapter 4 a method was developed for approximating the likelihood of the parameters of a generalised linear mixed model, designed to give a good approximation to the likelihood in sparse models, where other approximation methods fail. This sequential reduction method exploits the

dependence structure of the model by expressing the problem of finding the likelihood in terms of finding a normalising constant of a density with dependence structure represented by an undirected graphical model. The method suggested uses sparse grid interpolation to store modifiers to the baseline normal approximation used to obtain the Laplace approximation, in order to make approximating the likelihood feasible in models with both sparse and dense structure. Several examples were given to demonstrate the new method. Example 4.6 compared the performance of the various alternatives in one model with sparse structure, and the sequential reduction approximation to the likelihood was shown to have much faster and more stable convergence than the importance sampling approximation in this case.

Despite the emphasis of the thesis on models in which the likelihood is intractable, some of the findings remain relevant if the likelihood is easy to compute. Wald-type hypothesis tests are commonly used in practice, yet we demonstrated in Example 3.3 that they perform very badly in some cases. This finding is especially relevant to situations in which the likelihood is tractable, in which it will typically be easy to use likelihood ratio tests instead of Wald tests for hypothesis testing. In sparse models, penalised forms of the likelihood have potential to substantially improve the inference, and prevent infinite parameter estimates, although the penalty described in Section 3.1.6 places insufficient penalty on some parameter values. In Section 6.2.2 we give an idea for an improved penalty.

In Chapter 5, the robustness of inference to deviations from the assumed random-effects distribution was studied. We concluded that the maximum likelihood estimator will be robust to such deviations in models which are sufficiently dense, but might not retain this robustness in sparse models. By using the equivalence between a low-order composite likelihood estimator in a dense model and the maximum likelihood estimator in an alternative sparse model, we were able to provide an example of a situation in which a low-order composite likelihood estimator is not robust to misspecification of the random-effects distribution, even though the maximum likelihood estimator is robust in the same setting.

6.2 Suggestions for further work

6.2.1 Quality of inference from the Laplace approximation

The numerical results of Example 3.5 show that the inference based on the Laplace approximation to the likelihood is close to that from the true likelihood in one particular model with $m = O(n^2)$ observations and n random effects. It is claimed that this finding will extend to any model with dense structure, so that the inference

from the Laplace approximation will become more similar to that from the true likelihood as the number of random effects increases, provided that the number of observations m increases more quickly than n. However, theoretical results are still required to support this claim.

6.2.2 Penalised likelihoods for generalised linear mixed models

In Section 3.1.6, a penalised form of the log-likelihood for a generalised linear mixed model was introduced, which is identical to the bias reduction penalty of Firth (1993) in the case in which there are no random effects. However, we do not claim that this penalty removes the first-order asymptotic bias in the maximum likelihood estimator. In fact, the penalised likelihood estimator is no longer guaranteed to be finite. Too small a penalty is imposed on parameter values where β is small but the variance parameter ψ is large.

One idea for a better penalty is to subtract

$$\tilde{p}(\theta) = p_{\text{ext}}(\theta|\mu_\theta) - p_{\text{ext}}(\mathbf{0}|\mu_\theta)$$

from the log-likelihood, where $p_{\text{ext}}(\theta|\mathbf{u}) = -\frac{1}{2}\log\left|I_{\text{ext}}(\theta|\mathbf{u})\right|$ is the bias reduction penalty for the generalised linear model which would be obtained if $\mathbf{u}$ were known, and could be treated as an additional observed covariate. We fix $\mathbf{u}$ at μ_θ, the maximiser of $g(.|\theta, \mathbf{y})$ over $\mathbf{u}$. In some preliminary investigations, this new penalty appears to offer an improvement on the penalty introduced in Section 3.1.6, and the estimator maximising the new penalised likelihood is guaranteed to be finite. However, more investigation into the performance of the suggested penalty is required before its use should be advocated.

6.2.3 Improvements to the sequential reduction method

While the current version of the sequential reduction method has been shown to far outperform existing approximation methods in models with sparse structure, there is still potential for improvements to the method. Some ideas for such improvements were listed in Section 4.10. A better choice of the knots for the sparse grid storage could improve the accuracy of the approximation, particularly at extreme parameter values. Potential improvements to the algorithm for maximising the approximated likelihood were also given, which rely on having an estimate of the error in the approximated difference in the log-likelihood at two parameter values. A suggestion was given of how to estimate this error, using the difference in the approximated values at consecutive levels of approximation, k.

6.2.4 Sparse grid interpolation in R

In order to implement the sequential reduction method in R, it was necessary to write code to perform cubic spline interpolation between a sparse grid of points. Such code has many potential uses aside from approximating the likelihood, and it would be useful to create an R package for sparse grid interpolation, in order to make it easy for others to use these methods in practice.

6.2.5 Approximate likelihood ratio tests

In Chapter 3, we showed that Wald tests can behave very badly in some models with sparse structure, and concluded that a likelihood ratio test should be used instead wherever possible. It is more expensive to compute a likelihood ratio test statistic than a Wald test statistic, and with the methods currently used to find and optimise the sequential reduction approximation to the likelihood, it is infeasible to find profile likelihood confidence interval for each parameter if there are more than a few parameters in the model. For example, in the flat-lizards example (Example 4.8), it is not currently computationally feasible to find full profile-likelihood confidence intervals for all 7 parameters. Some ideas for improving the method to find profile-likelihood confidence intervals were given in Section 4.10. One idea worth further consideration is to find approximate profile-likelihood confidence intervals by using an approximation to the log-likelihood surface found by interpolating between values stored at a sparse grid of points. This is an example of a potential application of sparse grid interpolation outside of its use in the sequential reduction method for approximating the likelihood.

6.2.6 Asymptotic results on robustness

In Section 5.3.1, it was shown that the asymptotic bias in the maximum likelihood estimator in a two-level model tends to zero as the number of observations on each random effect increases, even if the distribution of the random effects is incorrectly specified. Intuition and numerical studies indicate that similar results will hold more generally, so that the maximum likelihood estimator will be consistent if the number of random effects n, and the number of observations m tend simultaneously to infinity, provided that m grows at a faster rate than n. However, no formal proof of this result has yet been achieved. In the same asymptotic setting, we would also like to study the accuracy of inference from the Laplace approximation to the likelihood under correct model specification, as discussed in Section 6.2.1, and it seems likely that the same ideas will be useful to prove results in both cases.

6.2.7 Hypothesis testing under model misspecification

Even in a dense model, inference is not fully robust to misspecification of the random-effects distribution, since test statistics typically do not have their standard asymptotic distributions under such misspecification. Further work is still required to construct modified versions of these test statistics, which account for possible model misspecification. The difficulty lies in the computation of the Godambe information matrix, which would allow us to find an accurate approximation of the variance of the maximum likelihood estimator under misspecification. In cases where there is no independent replication, it is not obvious how to estimate this matrix. This problem is linked to that of estimating the variance of a composite likelihood estimator, which is one of the largest barriers to the routine use of composite likelihoods in practice.

Appendix A

The form of D for natural cubic splines

Suppose that are give the value $y_i = f(s_i)$ of $f(.)$ at each of n knots s_i, and wish to construct a cubic spline $f_{\text{interp}}(.)$ to interpolate between those knots. To do this, we want to find the matrix D so that the stacked coefficients $\mathbf{c}$ of the cubic polynomial in each segment of a cubic spline may be written as $\mathbf{c} = D\mathbf{y}$. The form of D varies according to the choice of end conditions. All of the results used below are well known (see, for example Knott (2000)), but usually only the coefficients $\mathbf{c}$ are needed, rather than the explicit form of the matrix D, which we use to extend to interpolation on a sparse grid. The coefficients $\mathbf{c}$ are usually computed in several steps, first solving a linear system for the second derivatives of the cubic spline at its knots, then using these second derivatives to construct the coefficients of the cubic spline. To find D we need to combine those steps.

Write $m_i = f''_{\text{interp}}(s_i)$ for the second derivative at knot i, and $h_i = s_{i+1} - s_i$ for the distance between knots i and $i+1$. For each of the interior knots $i = 2, \ldots, n-1$, we have

$$h_{i-1}m_{i-1} + 2(h_i - h_{i-1})m_i + h_i m_{i+1} = \frac{6(y_{i+1} - y_i)}{h_i} - \frac{6(y_i - y_{i-1})}{h_{i-1}}. \tag{A.1}$$

To be able to solve this system for $\mathbf{m} = \{m_1, \ldots, m_n\}$, we must specify an extra two end conditions. If we specify that $m_n = m_1 = 0$, so that the cubic spline will be linear beyond the range of the knots, this gives a natural cubic spline.

Write

$$r_i = \frac{6(y_{i+1} - y_i)}{h_i} - \frac{6(y_i - y_{i-1})}{h_{i-1}}$$

for the right hand side of (A.1), for $i = 2, \ldots, n-1$, and let $r_1 = r_n = 0$, for the

end conditions. Then the linear system of equations may be written

$$A\mathbf{m} = \mathbf{r},$$

where A is the $n \times n$ matrix with elements

$$A_{ij} = \begin{cases} 1 & \text{if } i = j = 1 \text{ or } i = j = n, \\ 2(h_i - h_{i-1}) & \text{if } i = j = 2, \ldots, n-1, \\ h_{i-1} & \text{if } i = 3, \ldots, n-1 \text{ and } j = i-1, \\ h_i & \text{if } i = 2, \ldots, n-2 \text{ and } j = i+1, \\ 0 & \text{otherwise.} \end{cases}$$

We may write $\mathbf{r} = B\mathbf{y}$, where B is the $n \times n$ matrix with elements

$$B_{ij} = 6 \begin{cases} h_i^{-1} & \text{if } j = i,\ i = 2, \ldots, n-1 \\ h_i^{-1} - h_{i+1}^{-1} & \text{if } j = i+1,\ i = 2, \ldots, n-1 \\ h_{i+1}^{-1} & \text{if } j = i+2,\ i = 2, \ldots, n-1 \\ 0 & \text{otherwise.} \end{cases}$$

Combining these, we get that

$$\mathbf{m} = A^{-1}B\mathbf{y}.$$

Write $\mathbf{c}^{(i)} = (\alpha_i, \beta_i, \gamma_i, \delta_i)$ for the coefficients of the cubic between knots s_i and s_{i+1}. The coefficients may be obtained as a linear combination of the second derivatives m_i and m_{i+1}, and the values y_i and y_{i+1}, for $i = 1, \ldots, n-1$. Define the $4 \times n$ matrices $Q^{(i)}$ and $R^{(i)}$, so that the columns $(i, i+1)$ of $Q^{(i)}$ are given by

$$Q^{(i)}\Big|_{i,i+1} = \begin{pmatrix} 0 & 0 \\ \frac{-h_i}{3} & \frac{-h_i}{6} \\ \frac{1}{2} & 0 \\ \frac{-1}{6h_i} & \frac{1}{6h_i} \end{pmatrix},$$

the columns $(i, i+1)$ of $R^{(i)}$ are

$$R^{(i)}\Big|_{i,i+1} = \begin{pmatrix} 1 & 0 \\ \frac{-1}{h_i} & \frac{1}{h_i} \\ 0 & 0 \\ 0 & 0 \end{pmatrix},$$

and the remaining components of $Q^{(i)}$ and $R^{(i)}$ are zero. Then

$$\begin{aligned}\mathbf{c}^{(i)} &= Q^{(i)}\mathbf{m} + R^{(i)}\mathbf{y} \\ &= \left[Q^{(i)}A^{-1}B + R^{(i)}\right]\mathbf{y} \\ &= D^{(i)}\mathbf{y}.\end{aligned}$$

for $i = 1, \ldots, n-1$. It only remains to write the coefficients $\mathbf{c}^{(0)}$ and $\mathbf{c}^{(n)}$ for the linear functions outside of the range of the knots in terms of $\mathbf{y}$. By matching the gradients at the external knots, we get that

$$D_{ij}^{(0)} = \begin{cases} 1 & \text{if } i = 1,\ j = 1, \\ -D_{2,j}^{(1)} & \text{if } i = 2,\ j = 1, 2, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$D_{ij}^{(n)} = \begin{cases} 1 & \text{if } i = 1,\ j = n, \\ 3h_{n-1}^2 D_{4,j}^{(n-1)} + 2h_{n-1} D_{3,j}^{(n-1)} + D_{2,j}^{(n-1)} & \text{if } i = 2,\ j = n-1, n, \\ 0 & \text{otherwise.} \end{cases}$$

The required matrix D may then be obtained by stacking the $D^{(i)}$ on top of one another, so that $\mathbf{c} = D\mathbf{y}$.

Appendix B

Choice of knots for sparse grid interpolation

Recall from Section 4.6.4 that in the sequential reduction method at level k, we use knots which are quantiles of a $N(0, \tau_k^2)$ distribution to construct sparse grids for storage. For the same tree tournament which is used in Example 4.6, we approximate the log-likelihood using the sequential reduction method with $k \in \{0, 2, 3, 4, 5, 6\}$, using knots of this form, with τ_k fixed at each of several possible values. In each case, we look at the difference between the approximation for each k and the approximation at $k = 6$. This gives the approximate error in the log-likelihood approximation obtained for each k, although when τ_k is fixed at a very small value, the approximation has not quite converged by $k = 6$.

Table B.1 gives these errors at each combination of k and τ, for approximations to the likelihood at several different values of β and σ. The cells corresponding to $\tau_k = 1 + \frac{k}{2}$, which is the default choice in our program for the sequential reduction method, are highlighted in each table. In most cases, the error shrinks reasonably quickly for this choice. Table B.1c indicates that this default choice τ_k may be growing slightly too quickly with k, since the error for $k = 5$ is greater than that for $k = 4$ with this default choice for τ_k. However, the error for $k = 5$ is nonetheless very small (relative to the approximated log-likelihood, which is -36.9 in this case).

τ \ k	0	2	3	4	5
1	-2	-1.6	-0.53	-0.2	-0.066
1.25	-2.1	-1.2	-0.24	-0.068	-0.018
1.5	-2.1	-0.75	-0.028	-0.0049	-0.0024
1.75	-2.1	-0.26	0.079	0.011	0.0015
2	-2.1	0.19	0.11	0.0094	0.0015
2.25	-2.1	0.61	0.11	0.0059	0.00039
2.5	-2.1	1	0.12	0.0023	0.0002
2.75	-2.1	1.4	0.16	0.0011	0.0002
3	-2.1	1.8	0.23	0.0014	0.00002
3.25	-2.1	2.1	0.33	0.0022	0.00011
3.5	-2.1	2.5	0.62	0.0037	0.00038

(a) $\beta = 1.5$, $\sigma = 2$

τ \ k	0	2	3	4	5
1	-0.61	-0.35	-0.15	-0.063	-0.021
1.25	-0.63	-0.25	-0.08	-0.027	-0.0084
1.5	-0.64	-0.15	-0.035	-0.0096	-0.0022
1.75	-0.64	-0.063	-0.013	-0.0026	-0.0005
2	-0.64	-0.0034	-0.0044	-0.00058	0.000046
2.25	-0.64	0.032	-0.0015	0.00013	-0.000013
2.5	-0.64	0.048	-0.0014	0.000075	0.0000031
2.75	-0.64	0.051	-0.0029	-0.000029	-0.000003
3	-0.64	0.045	-0.0059	-0.000019	0.000006
3.25	-0.64	0.035	-0.01	-0.00008	0.0000014
3.5	-0.64	0.023	-0.016	-0.00017	-0.0000067

(b) $\beta = 1$, $\sigma = 1$

τ \ k	0	2	3	4	5
1	-0.82	-0.66	-0.27	-0.11	-0.039
1.25	-0.86	-0.54	-0.16	-0.05	-0.014
1.5	-0.87	-0.35	-0.069	-0.015	-0.0036
1.75	-0.87	-0.16	-0.02	-0.0021	-0.00065
2	-0.87	0.017	-0.0003	0.00055	0.00023
2.25	-0.87	0.17	0.006	0.0011	-0.0000089
2.5	-0.87	0.3	0.011	0.00029	0.00000029
2.75	-0.87	0.42	0.021	-0.0000088	0.0000069
3	-0.87	0.53	0.039	0.000085	-0.0000032
3.25	-0.87	0.63	0.066	-0.000099	-0.000062
3.5	-0.87	0.73	0.1	-0.0006	-0.00014

(c) $\beta = 2$, $\sigma = 1.5$

Table B.1: Approximate errors in the sequential reduction approximation to the log-likelihood, for different k and τ.

Bibliography

Agresti, A., Caffo, B., & Ohman-Strickland, P. 2004. Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. *Computational Statistics and Data Analysis*, **47**(3), 639–653.

Barthelmann, V., Novak, E., & Ritter, K. 2000. High dimensional polynomial interpolation on sparse grids. *Advances in Computational Mathematics*, **12**(4), 273–288.

Besag, J. 1974. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, **36**(2), 192–236.

Bodlaender, H., & Koster, A. 2008. *Treewidth computations I. Upper bounds.* Technical report, Department of Information and Computing Sciences, Utrecht University.

Bodlaender, H., & Koster, A. 2010. *Treewidth computations II. Lower bounds.* Technical report, Department of Information and Computing Sciences, Utrecht University.

Bradley, R. A., & Terry, M. E. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, **39**(3/4), 324–345.

Breslow, N. E. 1984. Extra-Poisson variation in log-linear models. *Applied Statistics*, **33**(1), 38–44.

Breslow, N. E., & Clayton, D. G. 1993. Approximate inference in generalised linear mixed models. *Journal of the American Statistical Association*, **88**(421), 9–25.

Browne, W. J., & Draper, D. 2006. A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, **1**(3), 463–514.

Cattelan, M., & Varin, C. 2010. A model for correlated paired comparison data. *Working paper series, Department of Statistical Sciences, University of Padua.*

Cox, D. R., & Reid, N. 2004. A note on pseudolikelihood constructed from marginal densities. *Biometrika*, **91**(3), 729–737.

Crainiceanu, C. M., & Ruppert, D. 2004. Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **66**(1), 165–185.

Firth, D. 1993. Bias reduction of maximum likelihood estimates. *Biometrika*, **80**(1), 27–38.

Fournier, D. A., Skaug, H. J., Ancheta, J., Ianelli, J., Magnusson, A., Maunder, M. N., Nielsen, A., & Sibert, J. 2012. AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. *Optimisation Methods and Software*, **27**(2), 233–249.

Goldstein, H. 2011. *Multilevel Statistical Models.* 4 edn. Vol. 922. Wiley.

Hammersley, J. M., & Clifford, P. 1971. *Markov fields on finite graphs and lattices.* Unpublished.

Hauck, W. W., & Donner, A. 1977. Wald's test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association*, **72**(360a), 851–853.

Heagerty, P. J., & Kurland, B. F. 2001. Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika*, **88**(4), 973–985.

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**(436), 1461–1473.

Jiang, J. 2001. Goodness-of-fit tests for mixed model diagnostics. *The Annals of Statistics*, **29**(4), 1137–1164.

Jordan, M. I. 2004. Graphical models. *Statistical Science*, **19**(1), 140–155.

Kent, J. T. 1982. Robust properties of likelihood ratio tests. *Biometrika*, **69**(1), pp. 19–27.

Knott, G. D. 2000. *Interpolating Cubic Splines.* Progress in Computer Science and Applied Logic, vol. 18. Springer.

Lange, N., & Ryan, L. 1989. Assessing normality in random effects models. *The Annals of Statistics*, **17**(2), 624–642.

Lindsay, B. G. 1988. Composite likelihood methods. *Contemporary Mathematics*, **80**, 221–239.

Lindsay, B. G., Yi, G. Y., & Sun, J. 2011. Issues and strategies in the selection of composite likelihoods. *Statistica Sinica*, **21**, 71–105.

Litière, S., Alonso, A., & Molenberghs, G. 2008. The impact of a misspecified random-effects distribution on the estimation and the performance of inferential procedures in generalised linear mixed models. *Statistics in Medicine*, **27**(16), 3125–3144.

McCullagh, P., & Nelder, J. A. 1989. *Generalised Linear Models*. Second edn. Monographs on statistics and applied probability. Chapman and Hall.

McCulloch, C. E., & Neuhaus, J. M. 2011. Misspecifying the shape of a random effects distribution: why getting it wrong may not matter. *Statistical Science*, **26**(3), 388–402.

Mosteller, F. 1951. Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations. *Psychometrika*, **16**(1), 3–9.

Nelder, J. A., & Mead, R. 1965. A simplex method for function minimization. *The Computer Journal*, **7**(4), 308–313.

Nelder, John A., & Wedderburn, Robert W. M. 1972. Generalised linear models. *Journal of the Royal Statistical Society. Series A (General)*, **135**(3), 370–384.

Neuhaus, J. M., Hauck, W. W., & Kalbfleisch, J. D. 1992. The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika*, **79**(4), 755–762.

Pinheiro, J. C., & Bates, D. M. 1995. Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, **4**(1), 12–35.

R Core Team. 2012. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rue, H., Martino, S., & Chopin, N. 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71**(2), 319–392.

Schall, R. 1991. Estimation in generalised linear models with random effects. *Biometrika*, **78**(4), 719–727.

Self, S. G., & Liang, K-Y. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, **82**(398), 605–610.

Shun, Z., & McCullagh, P. 1995. Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**(4), 749–760.

Thurstone, L. L. 1927. A law of comparative judgment. *Psychological review*, **34**(4), 273–286.

Tierney, L., & Kadane, J. B. 1986. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, **81**(393), 82–86.

Turner, H., & Firth, D. 2010. *Bradley-Terry models in R: The BradleyTerry2 package*. R package version 0.9-4.

Turner, H. L., & Firth, D. 2012. Bradley-Terry models in R: the BradleyTerry2 package. *Journal of Statistical Software*, **48**(9).

van der Vaart, A. W. 1998. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Varin, C., Reid, N., & Firth, D. 2011. An overview of composite likelihood methods. *Statistica Sinica*, **21**(1), 5–42.

Verbeke, G., & Lesaffre, E. 1996. A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, **91**(433), 217–221.

Verbeke, G., & Lesaffre, E. 1997. The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational Statistics and Data Analysis*, **23**(4), 541–556.

Verbeke, G., & Molenberghs, G. 2013. The gradient function as an exploratory goodness-of-fit assessment of the random-effects distribution in mixed models. *Biostatistics*, **14**(3), 477–490.

West, B., Welch, K. B., & Galecki, A. T. 2006. *Linear mixed models: a practical guide using statistical software*. CRC Press.

White, H. 1982. Maximum likelihood estimation of misspecified models. *Econometrica*, **50**, 1–25.

Whiting, M. J., Stuart-Fox, D. M., O'Connor, D., Firth, D., Bennett, N. C., & Blomberg, S. P. 2006. Ultraviolet signals ultra-aggression in a lizard. *Animal Behaviour*, **72**(2), 353–363.

Williams, D. A. 1982. Extra-binomial variation in logistic linear models. *Applied Statistics*, **31**(2), 144–148.

Xu, X., & Reid, N. 2011. On the robustness of maximum composite likelihood estimate. *Journal of Statistical Planning and Inference*, **141**(9), 3047–3054.

Zeger, S. L., & Karim, M. R. 1991. Generalised linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association*, **86**(413), 79–86.

Zhang, D., & Davidian, M. 2001. Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics*, **57**(3), 795–802.