A sequential reduction method for inference in generalized linear mixed models

Helen Ogden

University of Warwick, Coventry, UK

Abstract

The likelihood for the parameters of a generalized linear mixed model involves an integral which may be of very high dimension. Because of this intractability, many approximations to the likelihood have been proposed, but 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 described in this paper exploits the dependence structure of the posterior distribution of the random effects to reduce the cost of finding a good approximation to the likelihood in models with sparse structure.

Keywords: Graphical model, Intractable likelihood, Laplace approximation, Pairwise comparisons, Sparse grid interpolation

1 Introduction

Generalized linear mixed models are a natural and widely used class of models, but one in which the likelihood often involves an integral of very high dimension. Because of this apparent intractability, many alternative methods have been developed for inference in these models.

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 data are binary.

If there are n random effects in total, the likelihood may always be written as an n-dimensional integral over these random effects. If there are a large number of random effects, then it will be computationally infeasible to obtain an accurate approximation to this n-dimensional integral by direct numerical integration. However, it is not always necessary to compute this n-dimensional integral to find the likelihood. For example, in a two-level random intercept model, independence between clusters may be exploited to write the likelihood as a product of n one-dimensional integrals, so it is relatively easy to obtain a good

This work was supported by the Engineering and Physical Sciences Research Council [grant numbers EP/P50578X/1, EP/K014463/1]

CRiSM Paper No. 13-21, www.warwick.ac.uk/go/crism

approximation to the likelihood, even for large n. In more complicated situations it is often not immediately obvious whether any such simplification exists. The 'sequential reduction' method developed in this paper may be applied to check for a simplification of the likelihood in any generalized linear mixed model.

To use the method in practice, it is necessary to have some method to store an approximate representation of a function. Such a representation may be constructed by specifying a set of points at which to evaluate the function, and a method for interpolation between those points. By using a sparse grid interpolation method to store a modifier to the normal approximation used to construct the Laplace approximation, it is possible to obtain a sufficiently good approximation to the likelihood for use in a wide range of models.

Several examples are given to demonstrate the new method, including a real-data example of a tournament among lizards observed by Whiting et al. (2006).

2 The generalized linear mixed model

A generalized linear model (Nelder and 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 $g(.)$.

An assumption implicit in the generalized 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 generalized linear mixed model does this by modelling the linear predictor as

$$\eta = X\beta + Z\mathbf{b}, \tag{1}$$

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.

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

$$\mathbf{b} = A(\psi)\mathbf{u}, \tag{2}$$

CRiSM Paper No. 13-21, www.warwick.ac.uk/go/crism

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 (1) and (2), we write

$$\eta = X\beta + Z(\psi)\mathbf{u}, \tag{3}$$

where $u_i \sim F_U(., \psi)$ and $Z(\psi) = ZA(\psi)$. This paper 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 Section 5 could be extended easily for use with other random effect distributions of the form (2).

The non-zero elements of the columns of $Z(\psi)$ give us the observations which involve each random effect. We will say the generalized 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 data are binary, because the amount of information available on each random effect is small.

3 Examples of generalized linear mixed models

3.1 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

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

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.

CRiSM Paper No. 13-21, www.warwick.ac.uk/go/crism

3.2 Pairwise competition models

Consider a tournament among n players, consisting of contests between pairs of players. Let y_{ijt} record the outcome of the tth 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_{ijt} are independent, with distribution depending on the difference in abilities of the players i and j, so that

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

for some link function $g(.)$.

The examples used in this paper are all pairwise competition models in which the observations Y_{ijt} are binary, 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 and 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 generalized linear mixed model form by specifying that $\mathbb{E}(Y_{ijt}) = g^{-1}(\eta_{ijt})$, where

$$\eta_{ijt} = \lambda_i - \lambda_j = \beta^T(\mathbf{x}_i - \mathbf{x}_j) + b_i - b_j. \tag{4}$$

To write (4) in the form (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

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

CRiSM Paper No. 13-21, www.warwick.ac.uk/go/crism

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 generalized linear mixed model with sparse structure, since we only observe a total of 100 contests on the 77 lizards.

3.3 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 generalized linear mixed models which have been noted to have intractable likelihoods fall into this class. For example, models for the salamander mating data given in McCullagh and Nelder (1989) are of this form.

4 Inference in generalized linear mixed models

4.1 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

$$L(\beta, \psi) = \int_{\mathbb{R}^n} \prod_{i=1}^{m} f\left(y_i | \eta_i = X_i^T\beta + Z_i(\psi)^T\mathbf{u}\right) \prod_{j=1}^{n} \phi(u_j)du_j, \tag{5}$$

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 (5), there are some occasions where it may be simplified, so that computation of an n-dimensional integral is not necessary. The sequential reduction method developed in Section 5 gives a systematic way to check for such simplifications to the likelihood.

4.2 Laplace approximation to the likelihood

Many of the 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.

CRiSM Paper No. 13-21, www.warwick.ac.uk/go/crism

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-normalized version of the posterior density for $\mathbf{u}$, given $\mathbf{y}$, β and ψ.

Pinheiro and Bates (1995) suggest using a Laplace approximation to this integral. For each fixed $\theta = (\beta, \psi)$, this approach relies on a normal approximation to the posterior density of $\mathbf{u}$, given $\mathbf{y}$ and θ. To find this normal approximation, let $\hat{\mathbf{u}}_\theta$ maximize $\log g(\mathbf{u}|\mathbf{y}, \theta)$ over $\mathbf{u}$, and write $\Sigma_\theta = -H_\theta^{-1}$, where H_θ is the Hessian resulting from this optimization. The normal approximation to $g(.|\mathbf{y}, \theta)$ will be proportional to a $N_n(\hat{\mathbf{u}}_\theta, \Sigma_\theta)$ density. Writing $g^{\text{na}}(.|\mathbf{y}, \theta)$ for the normal approximation to $g(.|\mathbf{y}, \theta)$,

$$g^{\text{na}}(\mathbf{u}|\mathbf{y}, \theta) = \frac{g(\hat{\mathbf{u}}_\theta|\mathbf{y}, \theta)}{\phi_n(\hat{\mathbf{u}}_\theta; \hat{\mathbf{u}}_\theta, \Sigma_\theta)} \phi_n(\mathbf{u}; \hat{\mathbf{u}}_\theta, \Sigma_\theta),$$

where we write $\phi_n(.; \mu, \Sigma)$ for the $N_n(\mu, \Sigma)$ density. When we integrate over $\mathbf{u}$, only the normalizing constant remains, so that

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

In the case of a linear mixed model, the approximating normal 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 may be large.

4.2.1 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 an 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 and 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 Section 6.1), and where we consider a model with $n = O(m^{1/2})$ to have dense structure (see Section 6.2).

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 log-likelihood surface appears to be sufficiently good to give accurate inference, even

CRiSM Paper No. 13-21, www.warwick.ac.uk/go/crism

in cases where the approximation to the likelihood at each point has large relative error. See Section 6.2 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 and Kadane (1986) in the context of computing posterior moments. However, in models with very sparse structure, even the shape of the Laplace approximation to the log-likelihood surface may be inaccurate, so another method is required.

4.3 Importance sampling approximation to the likelihood

In cases where the Laplace approximation fails, Pinheiro and Bates (1995) suggest constructing an importance sampling approximation to the likelihood, based on samples from the normal distribution $N_n(\mu_\theta, \Sigma_\theta)$ which has density proportional to the approximation $g^{\text{na}}(.|\theta)$ used to construct the Laplace approximation. Writing

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

the likelihood may be approximated by

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

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

Unfortunately, there is no guarantee that the variance of the importance weights $w(\mathbf{u}^{(i)};\theta)$ will be finite. In such a situation, 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.

5 The sequential reduction method

5.1 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. If there is no observation involving both u_i and u_j, u_i and u_j will be 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.

CRiSM Paper No. 13-21, www.warwick.ac.uk/go/crism

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, so the posterior distribution of the random effects has the pairwise Markov property 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. For example, in a four-level random intercept model, each observation relies on the level 1, 2 and 3 grouping of each item, and so each observation will be represented by a triangle on the vertices representing the three groups to which that item belongs.

The problem of computing the likelihood has now been transformed to that of finding a normalizing constant of a density associated with an undirected graphical model. This problem has been well studied, and we now review some of the key ideas.

5.2 Factorizing the posterior density

In order to see how the conditional dependence structure can be used to enable a simplification of the likelihood, 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 $\mathcal{G}$, and a clique is said to be maximal if it is not itself contained within a larger clique. 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.

The Hammersley-Clifford theorem (Hammersley and Clifford (1971), Besag (1974)) implies that $g(.|\mathbf{y}, \theta)$ factorizes over the maximal cliques of $\mathcal{G}$, so that we may write

$$g(\mathbf{u}|\mathbf{y}, \theta) = \prod_{C \in M(\mathcal{G})} g_C(\mathbf{u}_C)$$

for some functions $g_C(.)$. A condition needed to obtain this result using the Hammersley-Clifford theorem is that $g(\mathbf{u}|\mathbf{y}, \theta) > 0$ for all $\mathbf{u}$. This will hold in this case because $\phi(u_i) > 0$ for all u_i. In fact, we may show that such a factorization exists directly. One particular such factorization is constructed in Section 5.4.2, and would be valid even if we assumed a random effects density $f_u(.)$ such that $f(u_i) = 0$ for some u_i.

5.3 Exploiting the clique factorization

Jordan (2004) reviews some methods to find the marginals of a density factorized over the maximal cliques of a graph. While these methods are well known, their use is typically 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

CRiSM Paper No. 13-21, www.warwick.ac.uk/go/crism

the distribution with density proportional to $g(.|\mathbf{y},\theta)$, and so approximate the likelihood

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

We take an iterative approach to the problem, first integrating out u_1 to find the non-normalized marginal posterior density of $\{u_2,\ldots,u_n\}$. We start with a factorization 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 $\mathcal{G}_i$.

Factorizing $g(.|\mathbf{y},\theta)$ 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 $\{g^1_C(.) : C \in M_1\}$. 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

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

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 factorization 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.

5.4 The sequential reduction method for likelihood approximation

5.4.1 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

CRiSM Paper No. 13-21, www.warwick.ac.uk/go/crism

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 5.6 discusses how to choose a good order, with the aim of minimizing the cost of approximating the likelihood. Reorder the random effects so that we integrate out $u_1, \ldots, u_n$ in that order.

2. Factorize $g(\mathbf{u}|\mathbf{y}, \theta)$ 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 factorization is not unique, so we must choose one particular factorization $\{g_C^1(.) : C \in M_1\}$. Section 5.4.2 gives the factorization we use in practice.

3. Once $u_1, \ldots u_{i-1}$ have been integrated out (using some approximate method), we have the factorization
$$\tilde{g}(u_i, \ldots, u_n|\mathbf{y}, \theta) = \prod_{C \in M_i} g_C^i(\mathbf{u}_C),$$
of the (approximated) non-normalized posterior for $u_i, \ldots, u_n$. Write
$$g_{N_i}^i(\mathbf{u}_{N_i}) = \prod_{C \in M_i : C \subset N_i} g_C^i(\mathbf{u}_C).$$
We first store an approximate representation $\tilde{g}_{N_i}^i(.)$ of $g_{N_i}^i(.)$, then integrate over u_i, to give an approximate representation $\tilde{g}_{N_i \setminus i}^i(.)$ of $g_{N_i \setminus i}^i(.)$. In Section 5.4.3 we discuss the construction of this approximate representation.

4. Write
$$\tilde{g}(u_{i+1}, \ldots, u_n|\mathbf{y}, \theta) = \tilde{g}_{N_i \setminus i}^i(\mathbf{u}_{N_i \setminus i}) \prod_{C \in M_i : C \not\subset N_i} g_C^i(\mathbf{u}_C),$$
defining a factorization of the (approximated) non-normalized 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.

5.4.2 A specific clique factorization

The general method described in Section 5.4.1 is valid for an arbitrary factorization 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 factorization 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

CRiSM Paper No. 13-21, www.warwick.ac.uk/go/crism

$\mathbf{y}$ 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

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

Then

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

so $g_C^1(.)$ does define a factorization of $g(.|y)$.

Using this factorization, at stage i, the non-normalized 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 $C \in M_i$,

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

where $g(\mathbf{u}_C|\mathbf{y}_C)$ is the non-normalized posterior distribution of $\mathbf{u}_C$ given $\mathbf{y}_C$.

5.4.3 An approximate representation of $g_{N_i}^i(.)$

We now construct an approximate representation of $g_{N_i}^i(.)$ at each stage i. We already have a normal approximation for the whole non-normalised posterior $g(.|\mathbf{y},\theta)$, and we base the approximate representation of $g_{N_i}^i(.)$ on this normal approximation.

First 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

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

CRiSM Paper No. 13-21, www.warwick.ac.uk/go/crism

and factorize as before as

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

where

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

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 $D^{-1}\Sigma_\theta$.

At each stage i, we store $g^i_{N_i}(\mathbf{z}_{N_i})$ for each $\mathbf{z}_{N_i}$ in some pre-specified standard set of 'storage' points $S_{|N_i|}$. To complete the representation, we need a method of interpolation between those points. To do this, we find a 'modifier' function $r^i_{N_i}(.)$ which will be constant over all $\mathbf{z}_{N_i}$ if the normal approximation is exact, and interpolate values of $\log r^i_{N_i}(\mathbf{z}_{N_i})$ for $\mathbf{z}_{N_i} \notin S_{|N_i|}$ using cubic splines.

We note that the conditional distribution of z_i given $z_{N_i\setminus i}$ and $\mathbf{y}$ is identical to the conditional distribution given only $\mathbf{y}_{N_i}$, so that all the information about this conditional distribution is contained in $g^i_{N_i}(.)$. If the approximating normal distribution is exact, $\mathbf{z} \sim N_n(0,\Omega)$, so that $z_i|\mathbf{z}_{N_i\setminus i} \sim N(\mu_i^c(\mathbf{z}_{N_i\setminus i}),\sigma_i^c)$, where

$$\mu_i^c(\mathbf{z}_{N_i\setminus i}) = \Omega_{i,N_i\setminus i}\left(\Omega_{N_i\setminus i,N_i\setminus i}\right)^{-1}\mathbf{z}_{N_i\setminus i}$$

and

$$\sigma_i^c = \Omega_{i,i} - \Omega_{i,N_i\setminus i}\Omega^{-1}_{N_i\setminus i,N_i\setminus i}\Omega_{N_i\setminus i,i}.$$

This means that if the normal approximation is exact

$$h^i_{N_i}(\mathbf{z}_{N_i}) = \frac{g^i_{N_i}(\mathbf{z}_{N_i})}{\phi(z_i;\mu_i^c(\mathbf{z}_{N_i\setminus i}),\sigma_i^c)}$$

will be constant in z_i, but may still vary in $\mathbf{z}_{N_i\setminus i}$. To remove the variation in $\mathbf{z}_{N_i\setminus i}$, write

$$r^i_{N_i}(\mathbf{z}_{N_i}) = \frac{h^i_{N_i}(z_i,\mathbf{z}_{N_i\setminus i})}{h^i_{N_i}(0,\mathbf{z}_{N_i\setminus i})},$$

which will be equal to one for all $\mathbf{z}_{N_i}$ if the normal approximation is exact. In cases where the normal approximation is not exact, $\log r^i_z(\mathbf{z}_{N_i})$ will vary with $\mathbf{z}_{N_i}$, and by capturing this variation, we can improve on the Laplace approximation to the likelihood.

The sets of storage points are chosen in such a way that $S_d = \{\mathbf{z} : (0,\mathbf{z}) \in S_{d+1}\}$. For each new storage point $\mathbf{z}_{N_i\setminus i} \in S_{|N_i|-1}$, we then seek to approximate

$$\begin{aligned}\int g^i_{N_i}(z_i,\mathbf{z}_{N_i\setminus i})dz_i &= \int h^i_{N_i}(\mathbf{z}_{N_i})\phi(z_i;\mu_i^c(\mathbf{z}_{N_i\setminus i}),\sigma_i^c)dz_i \\ &= \int r^i_{N_i}(\mathbf{z}_{N_i})h^i_{N_i}(0,\mathbf{z}_{N_i\setminus i})\phi(z_i;\mu_i^c(\mathbf{z}_{N_i\setminus i}),\sigma_i^c)dz_i \\ &= h^i_{N_i}(0,\mathbf{z}_{N_i\setminus i})\int r^i_{N_i}(\mathbf{z}_{N_i})\phi(z_i;\mu_i^c(\mathbf{z}_{N_i\setminus i}),\sigma_i^c)dz_i.\end{aligned}$$

CRiSM Paper No. 13-21, www.warwick.ac.uk/go/crism

In order to compute $r_z^i(\mathbf{z}_{N_i})$ at each storage point, we will already have found $h_{N_i}^i(0, \mathbf{z}_{N_i \setminus i})$ at each new storage point $\mathbf{z}_{N_i \setminus i}$. To compute the integral over z_i, we need to interpolate between storage points of $\log r_{N_i}^i(.)$, which we do using cubic splines with knots at each of the storage points. The choice of the storage points, and the method of interpolation between these points, is discussed in Section 5.5.

The Integrated Nested Laplace Approximations (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.

5.5 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 $\log r_z^i(.)$. 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.

5.5.1 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

$$\mathcal{U}^l(f) = \sum_{j=1}^{m_l} f(s_j) a_j^l, \tag{6}$$

where the a_j^l are basis functions. The approximate interpolated value of $f(.)$ at any point x is then given by $\mathcal{U}^l(f)(x)$. In Section 5.5.3, 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,

CRiSM Paper No. 13-21, www.warwick.ac.uk/go/crism

and compute 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.

5.5.2 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.

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

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

for $l > 1$, and $\Delta^1 = U^1$. Then Δ^i gives the quantity we should add 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

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

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.

5.5.3 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 (6). We first consider how to do this

CRiSM Paper No. 13-21, www.warwick.ac.uk/go/crism

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.

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 the 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 5.5.4.

5.5.4 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 equally spaced 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.

5.6 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.

CRiSM Paper No. 13-21, www.warwick.ac.uk/go/crism

The random effects may be removed in any order, so it makes sense to use an 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. In face, 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, there are special structures of graph which have known treewidth, and algorithms exist to find upper and lower bounds on the treewidth in reasonable time (see Bodlaender and Koster (2008) and Bodlaender and Koster (2010)). 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.

5.7 Using the sequential reduction method in practice

5.7.1 A program for the sequential reduction method

Code to implement the sequential reduction method in R (R Core Team, 2012) is provided as supplementary material, along with code to reproduce the examples of Section 6. The program only currently allows models with binary data, and with $Z(\sigma) = \sigma\bar{Z}$ for some matrix $\bar{Z}$ and scalar σ. The program allows a good approximation to the likelihood to be found in most cases, but does fail for some very extreme parameter values, for example in some cases where $\sigma > 2.5$. This is because of the very poor quality of the baseline normal approximation in those cases. A better selection of knots for storage could improve the approximation at these extreme parameter values.

5.7.2 Maximizing 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 maximizing the Laplace approximation to the likelihood), and use the resulting estimator as a starting point for the optimization of the approximation to the likelihood with $k = 2$ (we skip $k = 1$ since it always seems to give similar results to $k = 0$). We continue in this manner, increasing k and using the previous maximum as

CRiSM Paper No. 13-21, www.warwick.ac.uk/go/crism

the starting point for the optimization, until the location of the maximum of the approximated likelihood has stabilized, or until some maximum permissable level of approximation $k_{\max}$ has been reached. This basic method could be improved by making more use of the information gathered about the likelihood surface using a lower levels of approximation to tune the optimization procedure at each level k.

5.7.3 Hypothesis testing and confidence intervals

It will generally be of interest not just to obtain point estimates for the parameters of a generalized linear mixed model, but to test hypotheses about these parameters, or to construct confidence intervals for them. We will consider Wald and likelihood ratio tests to perform such tasks.

If the parameter 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 generalized 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. Self and Liang (1987) consider this type of problem under an asymptotic regime with independent and identically distributed observations, and find the correct asymptotic distribution of the likelihood ratio test statistic in this case. In a more realistic setting, Crainiceanu and Ruppert (2004) demonstrate that this modification itself may not be correct. Throughout the examples of this paper, the unadjusted chi-squared distributed will be used. This will result in a slight reluctance to reject the hypothesis that ψ is small.

The Wald test statistic is usually considerably easier to compute than the likelihood ratio test statistic, and the two test statistics are asymptotically equivalent. However, in the examples of Section 6, the Wald statistic is far from the likelihood ratio statistic, because the log-likelihood surface close to the maximum likelihood estimator is poorly approximated by a quadratic function. We conclude that a likelihood ratio test should be used instead of a Wald test wherever possible.

5.7.4 Penalized versions of the likelihood

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

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

to shrink the parameter estimates towards zero. Even when separation does not occur, using such a penalized likelihood could improve the statistical properties of the resulting estimator.

CRiSM Paper No. 13-21, www.warwick.ac.uk/go/crism

In the case of a generalized linear model, where there are no random effects, Firth (1993) demonstrates how to choose a penalty to remove the $O(n^{-1})$ asymptotic bias in β. In a logistic regression model, Heinze and Schemper (2002) advocate the use of this bias-reduction penalty to solve the problem of separation, since the resulting penalized likelihood is guaranteed to have a unique maximizer in this setting.

In the examples of Section 6, we will use a penalty which we call the independence bias reduction (IBR) penalty, which is equal to the bias-reduction penalty constructed under the assumption of no random effects in the model. Writing

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

where $\psi = 0$ corresponds to the case of no random effects, the independence bias reduction penalty is given by

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

We do not claim that this penalty will remove the $O(n^{-1})$ asymptotic bias in $\theta = (\beta, \psi)$. For any fixed ψ, this penalized likelihood has a unique maximizer over β. However, there are a few cases in which for fixed β, the penalized likelihood is monotone in ψ, so the use of this independence bias-reduction penalty does not guarantee a finite parameter estimate. These problems are not encountered in the examples of Section 6, but further work is required to find a better penalty, to ensure that the penalized likelihood has a unique maximizer in all cases.

We may use the penalized 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.

6 Examples

We give some examples to compare the performance of the proposed sequential reduction method with existing methods to approximate the likelihood. All the examples we consider are Thurstone-Mosteller models, as described in Section 3.2. That is, we only observe which player wins in a contest between two players, and suppose that $Pr(i \text{ beats } j|\lambda) = \Phi^{-1}(\lambda_i - \lambda_j)$. The various tournament structures we use are shown in Figure 1.

6.1 Tree tournament

Consider observing a tree tournament, with structure as shown in the first panel of Figure 1. Suppose that there is a single observed covariate x_i for each player, where $\lambda_i = \beta x_i + \sigma u_i$ and $u_i \sim N(0, 1)$.

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.

CRiSM Paper No. 13-21, www.warwick.ac.uk/go/crism

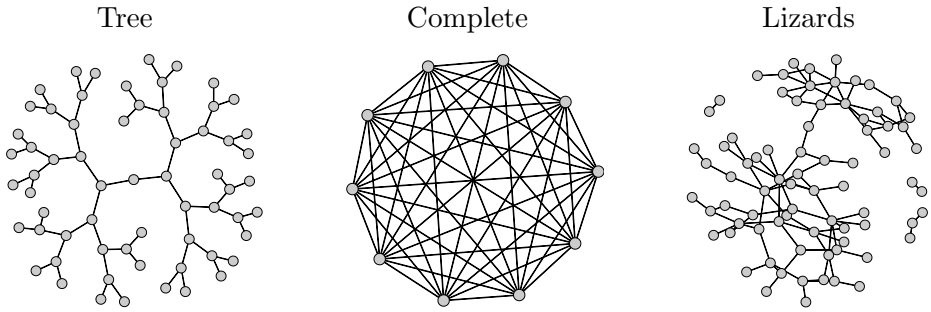

Figure 1: The tournament designs used in examples

Method	Time taken	$\hat{\beta}$	$\hat{\sigma}$
Laplace	4 seconds	1.008	0.771
IS ($N = 10^3$)	10 seconds	1.141	0.975
IS ($N = 10^4$)	37 seconds	1.160	1.000
IS ($N = 10^5$)	6.2 minutes	1.217	1.080
IS ($N = 5 \times 10^5$)	29 minutes	1.202	1.058
IS ($N = 10^6$)	72 minutes	1.203	1.059
SR ($k = 2$)	32 seconds	1.212	1.077
SR ($k = 3$)	1.0 minutes	1.204	1.060
SR ($k = 4$)	1.7 minutes	1.204	1.061

Table 1: Parameter estimates for the tree tournament

We aim to maximize the likelihood, penalized by the IBR penalty described in Section 5.7.4. We consider importance sampling approximations, for $N = 10^3$, 10^4, 10^5, 5×10^5, and 10^6, and sequential reduction approximations, for $k = 2$, 3 and 4. The posterior dependence graph of a tree tournament is a tree, which has treewidth 2. Using the sequential reduction method with sparse grid storage at level k, the cost of approximating the likelihood at each point will therefore be $O(n2^k)$. Approximating the likelihood at each point took about 0.5 seconds for $k = 2$, 0.7 seconds for $k = 3$ and 1.3 seconds for $k = 4$.

Optimization of both likelihood approximations was performed by using the algorithm described in Section 5.7.2. To check that our implementation of the importance sampling method is reasonably efficient, the maximum of an importance sampling approximation to the likelihood was also found using the ADMB package of Fournier et al. (2012). The maximum of the approximated likelihood was found in slightly less time by our method than by ADMB.

Table 1 gives the estimates of β and σ resulting from each approximation to the likelihood. The time taken to find the estimator using each approximation method is also given. The sequential reduction method finds an accurate estimate of the parameters more quickly than the importance sampling method, although in this case the estimates from both methods converge to a value fairly close to the maximizer of the true penalized likelihood.

However, we are interested in the shape of the log-likelihood surface, not just the location of its maximum. We consider approximations to the difference be-

CRiSM Paper No. 13-21, www.warwick.ac.uk/go/crism

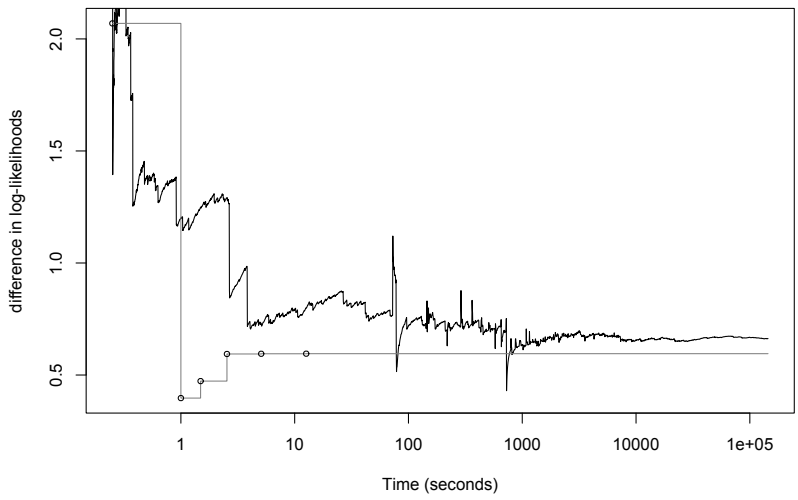

Figure 2: Approximations to $\ell^p(1.20, 1.06) - \ell^p(2.00, 1.50)$, plotted against the time taken to find the approximation.

tween the penalized log-likelihood at two points: the maximum $(1.20, 1.06)$, and the point $(2.00, 1.50)$, and consider the quality of each approximation relative to the time taken to compute it. Figure 2 shows the trace plots of importance sampling and sequential reduction approximations to this difference in penalized log-likelihoods, plotted against the length of time taken to find each approximation, on a log scale. In a few seconds, the sequential reduction approximation attains a greater accuracy than the importance sampling approximation taking over 24 hours to compute.

6.2 Complete tournament

We now examine the performance of the Laplace approximation in a tournament with dense structure. A complete tournament is one in which every pair from a set of n players competes exactly once. We consider tournaments consisting of R_n independent sub-tournaments, each of which is a complete tournament between 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 $(12, 3)$, $(8, 7)$ and $(5, 19)$, so that the total number of contests is approximately constant. We simulate a single binary observed covariate x_i for each player, which is 1 with probability 0.5, and 0 otherwise, and model $\lambda_i = \beta x_i + \sigma u_i$, where $u_i \sim N(0, 1)$. We simulate sample tournaments from the model, with $\beta = 1$ and $\sigma = 0.5$.

There are $m = \binom{n}{2}$ contests in a complete tournament between n players, 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 tends to 0 as $n \to \infty$ (see the discussion in Section 4.2). 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

CRiSM Paper No. 13-21, www.warwick.ac.uk/go/crism

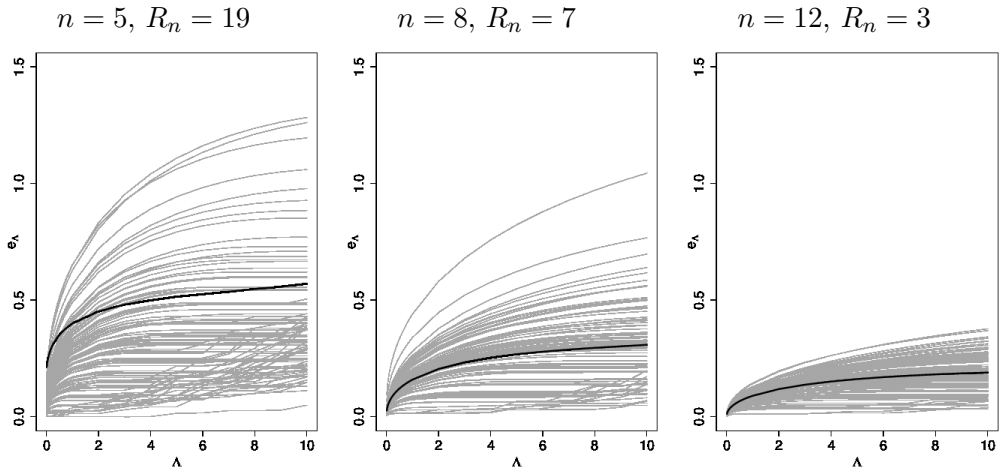

Figure 3: The error e_Λ plotted against Λ, for 100 simulated tournaments, each consisting of R_n complete sub-tournaments between n players.

is close to the inference obtained from the true likelihood.

When n is small, 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 that $\theta = \theta^*$ for various values of θ^*, 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.

For any fixed value Λ, those values of θ with $\Lambda^L(\theta) \leq \Lambda$ give a confidence region for θ, with approximate coverage $p(\Lambda) = Pr(\chi^2_2 \leq \Lambda)$. So, for each fixed Λ, 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

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

Figure 3 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.

The good performance of the Laplace approximation in models with dense structure means that it is only necessary to use a low level of approximation

CRiSM Paper No. 13-21, www.warwick.ac.uk/go/crism

k in the sequential reduction method for approximating the likelihood in such models. Furthermore, it is possible to detect the appropriate level of approximation to use in any given case, by increasing k until estimators and confidence intervals for parameters converge. This approach will succeed because of the stable nature of convergence of the sequential reduction approximation towards the true likelihood.

6.3 Application to the flat-lizards data

Whiting et al. (2006) conducted an experiment to determine the factors affecting the fighting ability of male flat lizards. They observe a tournament consisting of 100 contests between $n = 77$ lizards. Various covariate information is collected on each lizard, and the aim of the study is to create a model for the fighting ability of a lizard, based on these covariates. The tournament structure is shown on the right of Figure 1. The data are available in R as part of the BradleyTerry2 package (Turner and 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 Penalized Quasi Likelihood (PQL) of Breslow and 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 and 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 $\sigma = 0$. The data are binary, and we assume a Thurstone-Mosteller model, so that $Pr(i \text{ beats } j | \lambda_i, \lambda_j) = \Phi(\lambda_i - \lambda_j)$.

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 and Firth (2012), and introduce a new covariate for each of those lizards, to allow their abilities to be modelled separately. The first column of Table 2 gives the estimators under the assumption of no random effects, and the first column of Table 3 the Wald-type p-values for testing the hypothesis that each parameter is zero. The second columns provide 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 and 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 minimize the cost of the algorithm. 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 1.2 seconds for $k = 2$, 2.4 seconds for $k = 3$ and 6.7 seconds for $k = 4$.

CRiSM Paper No. 13-21, www.warwick.ac.uk/go/crism

				Sequential Reduction		
	$\sigma = 0$	PQL	Laplace	$k = 2$	$k = 3$	$k = 4$
throat.PC1	-0.055	-0.054	-0.070	-0.071	-0.071	-0.071
throat.PC3	0.21	0.23	0.25	0.25	0.25	0.25
head.length	-0.70	-0.84	-0.87	-0.86	-0.87	-0.87
SVL	0.11	0.11	0.13	0.14	0.14	0.14
lizard096	5.7	20.5	1.6	1.6	1.6	1.6
lizard099	0.52	0.60	0.51	0.51	0.52	0.52
σ	0 (fixed)	0.63	0.71	0.74	0.75	0.75

Table 2: Parameter estimates for the lizards tournament

				Sequential Reduction		
	$\sigma = 0$	PQL	Laplace	$k = 2$	$k = 3$	$k = 4$
throat.PC1	0.00085	0.021	0.00044	0.00052	0.00052	0.00052
throat.PC3	0.00099	0.0083	0.0017	0.0021	0.0022	0.0022
head.length	0.016	0.045	0.028	0.030	0.030	0.030
SVL	0.051	0.17	0.069	0.073	0.074	0.074
lizard096	0.98	1.00	0.13	0.14	0.14	0.14
lizard099	0.43	0.42	0.49	0.50	0.50	0.50
σ	-	0.00073	0.098	0.090	0.088	0.088

Table 3: p-values for testing each $\theta_i = 0$, in the lizards tournament

The estimators from maximizing the sequential reduction approximation to the likelihood, with IBR penalty, are given in Table 2. The p-values from a (penalized) likelihood ratio test for the presence of each parameter are given in Table 3. 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.

7 Conclusions

Many common approaches to inference in generalized linear mixed models rely on approximations to the likelihood, which may be of poor quality if there is little information available on each random effect. There are many situations in which it is unclear how good an approximation to the likelihood will be, and how much impact the error in the approximation will have on the statistical properties of the resulting estimator. It is therefore very useful to be able to obtain an accurate approximation to the likelihood at reasonable cost.

The sequential reduction method outlined in this paper allows a good approximation to the likelihood to be found in many models with sparse structure — precisely the situation where currently-used approximation methods perform worst. Little modification to the normal approximation used to in the Laplace approximation is required in models with dense structure, so by using sparse grid storage to store modifications to that normal approximation, it is possible to get

CRiSM Paper No. 13-21, www.warwick.ac.uk/go/crism

a sufficiently good approximation to the likelihood to use for reliable inference in a wide range of models.

References

Barthelmann, V., E. Novak, and K. Ritter (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. and A. Koster (2008). Treewidth computations I. Upper bounds. Technical Report, Department of Information and Computing Sciences, Utrecht University.

Bodlaender, H. and A. Koster (2010). Treewidth computations II. Lower bounds. Technical Report, Department of Information and Computing Sciences, Utrecht University.

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

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

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

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

Heinze, G. and M. Schemper (2002). A solution to the problem of separation in logistic regression. *Statistics in medicine 21*, 2409–2419.

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

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models* (Second ed.). Monographs on statistics and applied probability. Chapman and Hall.

CRiSM Paper No. 13-21, www.warwick.ac.uk/go/crism

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. and R. W. M. Wedderburn (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General) 135*(3), 370–384.

Pinheiro, J. C. and D. M. Bates (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*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rue, H., S. Martino, and N. Chopin (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.

Self, S. G. and K.-Y. Liang (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. and P. McCullagh (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society. Series B (Methodological) 57*(4), pp. 749–760.

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

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

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

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

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

CRiSM Paper No. 13-21, www.warwick.ac.uk/go/crism