

A sequential reduction method for inference in generalized linear mixed models

Helen E. Ogden

*Department of Statistics, University of Warwick
Coventry, CV4 7AL, United Kingdom*

e-mail: h.ogden@warwick.ac.uk; url: warwick.ac.uk/heogden

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 substantially the cost of finding an accurate approximation to the likelihood in models with sparse structure.

MSC 2010 subject classifications: Primary 62F10; secondary 62J15.

Keywords and phrases: Graphical model, intractable likelihood, Laplace approximation, pairwise comparison, sparse grid interpolation.

Received August 2014.

Contents

| | | |
|-----|---|-----|
| 1 | Introduction | 136 |
| 2 | The generalized linear mixed model | 136 |
| 2.1 | The model | 136 |
| 2.2 | Example: Pairwise competition models | 137 |
| 2.3 | The likelihood | 137 |
| 2.4 | Existing approximations to the likelihood | 138 |
| 2.5 | Bayesian inference | 139 |
| 3 | The sequential reduction method | 139 |
| 3.1 | Conditional independence structure | 139 |
| 3.2 | Exploiting the clique factorization | 140 |
| 3.3 | The sequential reduction method | 141 |
| 3.4 | A specific clique factorization | 142 |
| 3.5 | Approximate function representation | 142 |
| 3.6 | Computational complexity | 145 |
| 3.7 | An R package for sequential reduction | 146 |
| 4 | Examples | 146 |
| 4.1 | Tree tournament | 146 |

| | | |
|-----|---|-----|
| 4.2 | An animal behavior “tournament”: Augrabies Flat lizards | 148 |
| 4.3 | A three-level model | 149 |
| 5 | Conclusions | 150 |
| | Acknowledgements | 150 |
| | Supplementary Material | 150 |
| | References | 151 |

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 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 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.

The likelihood may be written as an integral over the random effects. If there are a large number of random effects, then it will be computationally infeasible to obtain an accurate approximation to this integral by direct numerical integration. However, it is not always necessary to compute this high-dimensional integral to find the likelihood. In a two-level random intercept model, independence between clusters may be exploited to write the likelihood as a product of one-dimensional integrals, so it is relatively easy to obtain a good approximation to the likelihood, even if there are large number of random effects. In more complicated situations it is often not immediately obvious whether any such simplification exists.

The ‘sequential reduction’ method developed in this paper exploits the structure of the integrand to simplify computation of the likelihood, and as a result allows a fast and accurate approximation to the likelihood to be found in many cases where existing approximation methods fail. Examples are given to demonstrate the new method, including pairwise competition models and a model with nested structure.

2. The generalized linear mixed model

2.1. The model

A generalized linear mixed model allows the distribution of a response $\mathbf{Y} = (Y_1, \dots, Y_m)$ to depend on observed covariates through a linear predictor η . Conditional on η , 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 $g(\cdot)$. We model

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

where $\beta = (\beta_1, \dots, \beta_p)$ are fixed effects, with fixed design matrix X , and $\mathbf{u} = (u_1, \dots, u_n)$ are random effects, with design matrix $Z(\psi)$ whose entries may depend on a parameter ψ . We assume that the u_i are independent standard normal variables. At first sight, this assumption of independent and identically distributed random effects appears restrictive, but by suitable choice of $Z(\psi)$, we may let $Z(\psi)\mathbf{u}$ have any multivariate normal distribution with mean zero, to coincide with the standard parameterization of a generalized linear mixed model (as in Breslow and Clayton, 1993, for example).

For each item i , we define the ‘active’ set $A_i = \{j : Z_{ji}(\psi) \neq 0\}$ to describe which components of the response involve the random effect u_i . We will say the generalized linear mixed model has ‘sparse structure’ if this active set is small for most items. 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.

2.2. Example: Pairwise competition models

Consider a tournament among n players, consisting of contests between pairs of players. For each contest, we observe a binary outcome: either i beats j or j beats i . We suppose that each player i has some ability λ_i , and that conditional on all the abilities, the outcomes of the contests are independent, with distribution depending on the difference in abilities of the players i and j , so that $\Pr(i \text{ beats } j | \lambda) = g^{-1}(\lambda_i - \lambda_j)$ for a link function $g(\cdot)$ such that $g^{-1}(-t) = 1 - g^{-1}(t)$. If $g(\cdot) = \text{logit}(\cdot)$, then this describes a Bradley-Terry model (Bradley and Terry, 1952). If $g(\cdot) = \Phi^{-1}(\cdot)$ (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. We allow the ability of player i to depend on the covariates \mathbf{x}_i through $\lambda_i = \beta^T \mathbf{x}_i + \sigma u_i$, where u_i are independent $N(0, 1)$ samples. This gives a generalized linear mixed model, depending on a linear predictor η with components $\eta_r = \lambda_{p_1(r)} - \lambda_{p_2(r)}$, where $p_1(r)$ and $p_2(r)$ are the first and second player involved in match r . The active set for each player is the set of matches that player competes in, so the model will have sparse structure if each player competes in only a small number of matches, a common scenario in practice.

2.3. The likelihood

Let $f(\cdot | \eta_i)$ be the density of Y_i , conditional on knowledge of the value of η_i , and write $\theta = (\beta, \psi)$ for the full set of model parameters. Conditional on η , the components of \mathbf{Y} are independent, so that

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

where X_i is the i th row of X , and $Z_i(\psi)$ is the i th row of $Z(\psi)$. Unless n is very small, it will not be possible to approximate the likelihood well by direct computation of this n -dimensional integral.

2.4. Existing approximations to the likelihood

Pinheiro and Bates (1995) suggest using a Laplace approximation to the integral (1). Write

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

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} and θ . For each fixed θ , the Laplace approximation relies on a normal approximation to this posterior density. To find this normal approximation, let μ_θ 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 $g(\cdot | \mathbf{y}, \theta)$ will be proportional to a $N_n(\mu_\theta, \Sigma_\theta)$ density. Writing $g^{\text{na}}(\cdot | \mathbf{y}, \theta)$ for the normal approximation to $g(\cdot | \mathbf{y}, \theta)$,

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

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

$$L^{\text{Laplace}}(\theta) = \frac{g(\mu_\theta | \mathbf{y}, \theta)}{\phi_n(\mu_\theta; \mu_\theta, \Sigma_\theta)} = (2\pi)^{-\frac{n}{2}} (\det \Sigma_\theta)^{-\frac{1}{2}} g(\mu_\theta | \mathbf{y}, \theta).$$

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. In the case that n is fixed, and $m \rightarrow \infty$, 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 $n = o(m^{1/3})$. 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. 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. Nonetheless, in models with very sparse structure (where we might have $n = O(m)$), even the shape of the Laplace approximation to the log-likelihood surface may be inaccurate, so another method is required.

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 $N_n(\mu_\theta, \Sigma_\theta)$. Writing

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

the likelihood may be approximated by $L^{IS}(\theta) = \sum_{i=1}^N w(\mathbf{u}^{(i)}; \theta)/N$, 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 as $N \rightarrow \infty$, but the convergence may be slow and erratic, and estimates of the variance of the approximation may be unreliable.

2.5. Bayesian inference

From a Bayesian perspective, Markov chain Monte Carlo methods could be used to sample from the posterior distribution. However, such methods are computationally intensive, and it can be difficult to detect whether the Markov chain has converged to the correct distribution. Rue, Martino and Chopin (2009) suggest the Integrated Nested Laplace Approximation (INLA) to approximate the marginal posterior distribution of each parameter. INLA is computationally efficient, but Fong, Rue and Wakefield (2010) note that the approximation may perform poorly in models for binary data. In situations where the Laplace approximation to the likelihood fails, INLA may be also unreliable.

We do not consider these methods further, and instead focus on those methods which provide a direct approximation to the marginal likelihood (1).

3. The sequential reduction method

3.1. Conditional independence structure

We now consider the distribution of the unknown random effects \mathbf{u} given \mathbf{y} , at fixed parameter value θ , which has density proportional to the integrand of the likelihood $g(\mathbf{u}|\mathbf{y}, \theta)$. We will refer to this distribution as the posterior distribution of \mathbf{u} given \mathbf{y} and θ . Despite this terminology, our aim is to approximate the likelihood, which is the normalizing constant associated with this distribution, rather than to conduct Bayesian inference.

Recall that for each item i we define the active set A_i to be the indices of the non-zero components of the i th column of $Z(\psi)$. If the active sets A_i and A_j are disjoint, 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 i and j if $A_i \cap A_j \neq \emptyset$.

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.

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. 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 (Besag, 1974) implies that $g(\cdot|\mathbf{y}, \theta)$ 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 $g_C(\cdot)$. 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 3.4, and would be valid even if we assumed a random effects density $f_u(\cdot)$ such that $f_u(u_i) = 0$ for some u_i .

3.2. 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 Gaussian distributions. We will use the same ideas, combined with a method for approximate storage of functions, to approximate the marginals of the distribution with density proportional to $g(\cdot|\mathbf{y}, \theta)$, 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 $\{u_2, \dots, u_n\}$. We start with a

factorization of $g(\cdot|\mathbf{y}, \theta)$ over the maximal cliques of the posterior dependence graph of $\{u_1, \dots, u_n\}$, and the idea will be to write the marginal posterior density of $\{u_2, \dots, u_n\}$ 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 $\{u_i, \dots, u_n\}$, 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 $g(\cdot|\mathbf{y}, \theta)$ over the maximal cliques of \mathcal{G}_1 gives

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

for some functions $\{g_C^1(\cdot) : C \in M_1\}$. To integrate over u_1 , 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 neighbors of vertex 1 in \mathcal{G} (including vertex 1 itself). Then

$$\begin{aligned} \int g(\mathbf{u}|\mathbf{y}, \theta) du_1 &= \int \prod_{C \in M_1: C \subseteq N_1} g_C^1(\mathbf{u}_C) du_1 \prod_{\tilde{C} \in M_1: \tilde{C} \not\subseteq N_1} g_{\tilde{C}}^1(\mathbf{u}_{\tilde{C}}) \\ &= \int g_{N_1}^1(u_1, \mathbf{u}_{N_1 \setminus 1}) du_1 \prod_{\tilde{C} \in M_1: \tilde{C} \not\subseteq N_1} g_{\tilde{C}}^1(\mathbf{u}_{\tilde{C}}). \end{aligned}$$

Thus $g_{N_1}^1(\cdot)$ 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

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

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

This defines a new factorization of $g(u_2, \dots, u_n|\mathbf{y}, \theta)$ over the maximal cliques M_2 of the posterior dependence graph for $\{u_2, \dots, u_n\}$, 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.

3.3. The sequential reduction method

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 3.6 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, \dots, 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 $\{g_C^1(\cdot) : C \in M_1\}$. Section 3.4 gives the factorization we use in practice.
3. Once u_1, \dots, u_{i-1} have been integrated out (using some approximate method), we have the factorization $\tilde{g}(u_i, \dots, u_n|\mathbf{y}, \theta) = \prod_{C \in M_i} g_C^i(\mathbf{u}_C)$, of the (approximated) non-normalized posterior for u_i, \dots, u_n . Write

$$g_{N_i}(\mathbf{u}_{N_i}) = \prod_{C \in M_i: C \subset N_i} g_C^i(\mathbf{u}_C).$$

We then integrate over u_i (using a quadrature rule), and store an approximate representation $\tilde{g}_{N_i \setminus i}(\cdot)$ of the resulting function $g_{N_i \setminus i}(\cdot)$. In Section 3.5 we discuss the construction of this approximate representation.

4. Write

$$\tilde{g}(u_{i+1}, \dots, u_n|\mathbf{y}, \theta) = \tilde{g}_{N_i \setminus i}(\mathbf{u}_{N_i \setminus i}) \prod_{C \in M_i: C \not\subset N_i} g_C^i(\mathbf{u}_C),$$

defining a factorization of the (approximated) non-normalized posterior density of $\{u_{i+1}, \dots, u_n\}$ over the maximal cliques M_{i+1} of the new posterior dependence graph \mathcal{G}_{i+1} .

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

3.4. A specific clique factorization

The general method described in Section 3.3 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 \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

$$g_C^1(\mathbf{u}_C) = f(\mathbf{y}_C|\mathbf{u}_C) \prod_{j \in a(C)} \phi(u_j).$$

Then $g(\mathbf{u}|\mathbf{y}) = \prod_{C \in M_1} g_C^1(\mathbf{u}_C)$, so $g_C^1(\cdot)$ does define a factorization of $g(\cdot|\mathbf{y})$.

3.5. Approximate function representation

3.5.1. A modified function for storage

A key choice in the sequential reduction algorithm is the method used to ‘store’ the function $g_{N_i \setminus i}(\cdot)$. The storage consists of a set of points S_i at which to

evaluate $g_{N_i \setminus i}(\cdot)$, and a method of interpolation between those points, which will be used later in the algorithm if we need to evaluate $g_{N_i \setminus i}(\mathbf{u}_{N_i \setminus i})$ for some $\mathbf{u}_{N_i \setminus i} \notin S_i$.

We would like to minimize the size of the absolute error in the interpolation for those points $\mathbf{u}_{N_i \setminus i}$ at which we will later interpolate. The quality of the interpolation may be far more important at some points $\mathbf{u}_{N_i \setminus i}$ than at others. We will transform to a new function $r_{N_i \setminus i}(\mathbf{u}_{N_i \setminus i}) = g_{N_i \setminus i}(\mathbf{u}_{N_i \setminus i})h_{N_i \setminus i}(\mathbf{u}_{N_i \setminus i})$, where we choose $h_{N_i \setminus i}(\cdot)$ so that the size of the absolute interpolation error for $r_{N_i \setminus i}(\cdot)$ is of roughly equal concern across the whole space. Given an interpolation method for $r_{N_i \setminus i}(\cdot)$, we obtain interpolated values for $g_{N_i \setminus i}(\cdot)$ through $g_{N_i \setminus i}^{\text{interp}}(\mathbf{u}_{N_i \setminus i}) = r_{N_i \setminus i}^{\text{interp}}(\mathbf{u}_{N_i \setminus i})/h_{N_i \setminus i}(\mathbf{u}_{N_i \setminus i})$, so we must ensure that $h_{N_i \setminus i}(\cdot)$ is easy to compute.

Recall that we may think of the original integrand $g(\cdot|\mathbf{y}, \theta)$ as being the non-normalized posterior density for $\mathbf{u}|\mathbf{y}, \theta$. The region where we will interpolate a large number of points corresponds to the region where the marginal posterior density of $\mathbf{u}_{N_i \setminus i}|\mathbf{y}, \theta$ is large. Ideally, we would choose $h_{N_i \setminus i}(\cdot)$ to make $r_{N_i \setminus i}(\cdot)$ proportional to the density of $\mathbf{u}_{N_i \setminus i}|\mathbf{y}, \theta$, but this density is difficult to compute.

To solve this problem, we make use of the normal approximation to $g(\cdot|\mathbf{y}, \theta)$ used to construct the Laplace approximation to the likelihood, which approximates the posterior distribution $\mathbf{u}|\mathbf{y}, \theta$ as $N_n(\mu, \Sigma)$. The marginal posterior distribution of $\mathbf{u}_{N_i \setminus i}|\mathbf{y}, \theta$ may therefore be approximated as $N_d(\mu_{N_i \setminus i}, \Sigma_{N_i \setminus i})$, where $d = |N_i \setminus i|$. We choose $h_{N_i \setminus i}(\cdot)$ to ensure that the normal approximation to $r_{N_i \setminus i}(\cdot)$ (computed as described in Section 2.4) is $N_d(\mu_{N_i \setminus i}, \Sigma_{N_i \setminus i})$. That is, we choose $\log h_{N_i \setminus i}(\cdot)$ to be a quadratic function, with coefficients chosen so that $\nabla \log h_{N_i \setminus i}(\mu_{N_i \setminus i}) = -\nabla \log g_{N_i \setminus i}(\mu_{N_i \setminus i})$ and $\nabla^T \nabla \log h_{N_i \setminus i}(\mu_{N_i \setminus i}) = -\Sigma_{N_i \setminus i}^{-1} - \nabla^T \nabla \log g_{N_i \setminus i}(\mu_{N_i \setminus i})$.

3.5.2. Storing a function with a normal approximation

Suppose that $f(\cdot)$ is a non-negative function on \mathbb{R}^d , for which we want to store an approximate representation, and that we may approximate $f(\cdot)$ with $f^{\text{na}}(\mathbf{x}) \propto \phi_d(\mathbf{x}, \mu, \Sigma)$, for some μ and Σ . In our case, the function $f(\cdot)$ which we store is $r_{N_i \setminus i}(\cdot)$, of dimension $d = |N_i \setminus i|$, and with normal approximation $N_d(\mu_{N_i \setminus i}, \Sigma_{N_i \setminus i})$.

We transform to a new basis. Let $\mathbf{z} = A^{-1}(\mathbf{x} - \mu)$, where A is chosen so that $AA^T = \Sigma$. More specifically, we choose $A = PD$, where P is a matrix whose columns are the normalized eigenvectors of Σ and D is a diagonal matrix with diagonal entries the square roots of the eigenvalues of Σ . Write $f_z(\mathbf{z}) = f(A\mathbf{z} + \mu)$, and let $c(\mathbf{z}) = \log f_z(\mathbf{z}) - \log \phi_d(\mathbf{z}, 0, I)$, so that $c(\cdot)$ will be constant if the normal approximation is precise. We store $c(\cdot)$ by evaluating at some fixed points for \mathbf{z} , and specifying the method of interpolation between them. The choice of these points and the interpolation method is discussed in the next section. Given the interpolation method for $c(\cdot)$, we may define $f^{\text{interp}}(\mathbf{x}) = \exp\{c^{\text{interp}}(A^{-1}(\mathbf{x} - \mu))\} \phi_d(A^{-1}(\mathbf{x} - \mu), 0, I)$, to give an interpolation method for $f(\cdot)$.

If $g(\mathbf{u}|\mathbf{y}, \theta) \propto \phi_n(\mathbf{u}, \mu, \Sigma)$, there will be no error in the Laplace approximation to the likelihood. In this situation, $c(\cdot)$ will be constant, and the sequential reduction approximation will also be exact. In situations where the normal approximation is imprecise, $c(\cdot)$ will no longer be constant, and we may improve on the baseline (Laplace) approximation to the likelihood by increasing the number of points used for storage.

3.5.3. Sparse grid interpolation

In order to store an approximate representation of the standardized modifier function $c(\cdot)$, we will compute values of $c(\cdot)$ at a fixed set of evaluation points, and specify a method of interpolation between these points. We now give a brief overview of the interpolation methods based on sparse grids of evaluation points. Some of the notation we use is taken from Barthelmann, Novak and Ritter (2000), although there are some differences: notably that we assume $c(\cdot)$ 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.

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

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

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

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 $S = \{s_1, s_2, \dots\}$. 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 at each possible combination of those points. We write

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

where

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

Thus, in the full grid method, we must evaluate $c(\cdot)$ at $\prod_{j=1}^d m_{l_j} = O(\prod_{j=1}^d 2^{l_j}) = O(2^{\sum_{j=1}^d l_j})$ points. This will not be possible if $\sum_{j=1}^d l_j$ is too large.

In order to construct an approximate representation of $c(\cdot)$ in reasonable time, we could limit the sum $\sum_{j=1}^d l_j$ used in a full grid to be at most $d + k$, for some $k \geq 0$. If $k > 0$, there are many possibilities for ‘small full grids’ indexed by the levels $\mathbf{l} = (l_1, \dots, l_d)$ 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 $c(\cdot)$, let

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

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

Returning to the multivariate case, the sparse grid interpolation of $c(\cdot)$ at level k is given by

$$c_k^{\text{interp}} = \sum_{\mathbf{l}: |\mathbf{l}| \leq d+k} (\Delta^{l_1} \otimes \dots \otimes \Delta^{l_d})(c).$$

To store $c(\cdot)$ on a sparse grid at level k , we must evaluate at $O(d^{k+1})$ points, which allows approximate storage for much larger dimension d than is possible using a full grid method.

Barthelmann, Novak and Ritter (2000) use global polynomial interpolation for a function defined on a hypercube, with the Chebyshev knots. We prefer to use cubic splines for interpolation, since the positioning of the knots is less critical. Since we have already standardized the function we wish to store, 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.

3.5.4. Bounded interpolation

To ensure that $g_{N_i}(\cdot)$ remains integrable at each stage, we impose an upper bound M on the interpolated value of $c(\cdot)$. In practice, we choose M to be the largest value of $c(\mathbf{z})$ observed at any of the evaluation points.

3.6. Computational complexity

Using sparse grid storage at level k , the cost of stage i of the sequential reduction algorithm is at most $O(|N_i|^{2k})$. The overall cost of approximating the likelihood 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 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 neighbors 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.

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, 2010). We use a constructive algorithm for finding an upper bound on the treewidth, which outputs an elimination ordering achieving that upper bound, to find a reasonably good (though not necessarily optimal) ordering.

3.7. An R package for sequential reduction

The sequential reduction method is implemented in R (R Core Team, 2014) by the package `glmmsr`, included as supplementary material (Ogden, 2015). The code for sparse grid interpolation is based on the efficient storage schemes suggested by Murarasu et al. (2011). Code to reproduce the examples of Section 4 is also provided.

4. Examples

We give some examples to compare the performance of the proposed sequential reduction method with existing methods to approximate the likelihood. The first two examples here are of pairwise competition models (a simple tree tournament with simulated data, and a more complex, real-data example); the third is a mixed logit model with two nested layers of random effects.

4.1. Tree tournament

Consider observing a tree tournament, with structure as shown in Figure 1a. 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 $\beta = 0.5$ and $\sigma = 1.5$. We suppose that we observe two matches between each pair of competing players. The covariates x_i are independent draws from a standard normal distribution.

We fit the model using the Laplace approximation, and the sequential reduction approximations, for $k = 1, 2, 3, 4$ and 5. The posterior dependence graph of a tree tournament is a tree, which has treewidth 2. Using the sequential reduc-

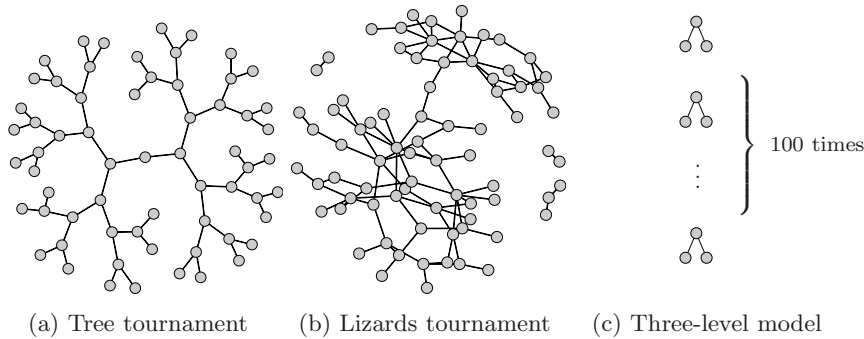


FIG 1. The posterior dependence graphs for the examples.

TABLE 1
The parameter estimates and standard errors for the tree tournament

| | | Laplace | k=1 | k=2 | k=3 | k=4 | k=5 |
|----------|----------|---------|------|------|------|------|------|
| β | estimate | 0.44 | 0.44 | 0.45 | 0.46 | 0.46 | 0.46 |
| | s.e. | 0.26 | 0.27 | 0.27 | 0.27 | 0.27 | 0.27 |
| σ | estimate | 1.13 | 1.26 | 1.29 | 1.30 | 1.30 | 1.30 |
| | s.e. | 0.31 | 0.36 | 0.37 | 0.38 | 0.38 | 0.38 |

tion method with sparse grid storage at level k , the cost of approximating the likelihood at each point will be $O(n4^k)$. In reality, the computation time does not quadruple each time k is increased, since the computation is dominated by fixed operations whose cost does not depend on k . To compute the approximation to the likelihood at a single point took about 0.02 seconds for the Laplace approximation, 0.22 seconds for $k = 1$, 0.24 seconds for $k = 2$, 0.24 seconds for $k = 3$, 0.27 seconds for $k = 4$ and 0.30 seconds for $k = 5$.

Table 1 gives the estimates of β and σ resulting from each approximation to the likelihood. The estimates of β are similar for all the approximations, but the estimate of σ found by maximizing the Laplace approximation to the likelihood is smaller than the true maximum likelihood estimator.

We also want to consider the quality of an importance sampling approximation to the log-likelihood, as described in Section 2.4. We are interested in the shape of the log-likelihood surface, rather than the pointwise quality of the approximation, so we consider approximations to the difference between the log-likelihood at two points: the maximum (0.46, 1.30), and the point (0.60, 2.00). We 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 log-likelihoods, plotted against the length of time taken to find each approximation, on a log scale. In well under a second, the sequential reduction approximation converges to such an extent that differences in the approximations are not visible on this scale. By contrast, after more than 14 hours, the importance sampling approximation has still not converged.

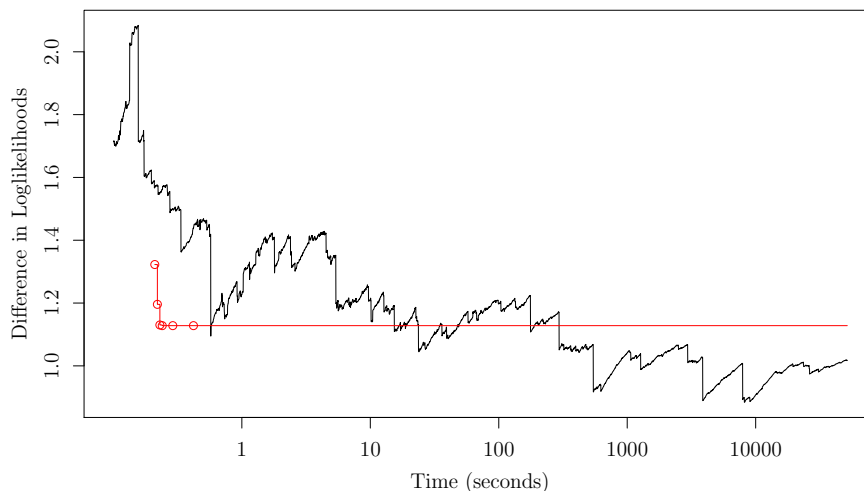


FIG 2. Importance sampling and sequential reduction approximations to $\ell(0.46, 1.30) - \ell(0.60, 2.00)$, plotted against the time taken to find the approximation, on a log scale. The sequential reduction approximation converges in less than a second, but the importance sampling approximation has still not converged after over 14 hours.

4.2. An animal behavior “tournament”: *Augrabies Flat lizards*

Whiting et al. (2006) conducted an experiment to determine the factors affecting the fighting ability of male Augrabies flat lizards, *Platysaurus broadleyi*. They captured $n = 77$ lizards, recorded various measurements on each, and then released them and recorded the outcomes of fights between pairs of animals. The tournament structure is shown in Figure 1b. The data are available in R as part of the BradleyTerry2 package (Turner and Firth, 2012).

There are several covariates \mathbf{x}_i available for each lizard. Turner and Firth (2012) suggest to model the ability of each lizard as $\lambda_i = \beta^T \mathbf{x}_i + \sigma u_i$, where $u_i \sim N(0, 1)$. The data are binary, and we assume a Thurstone-Mosteller model, so that $\Pr(i \text{ beats } j | \lambda) = \Phi(\lambda_i - \lambda_j)$.

In order to find the sequential reduction approximation to the likelihood, we must first 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, and we use an ordering corresponding to the upper bound.

To demonstrate the performance of the sequential reduction approximation, we consider the cut across the log-likelihood surface at $\beta = 0$, as σ varies. The various approximations to this curve are shown in Figure 3. It becomes harder to obtain a good approximation to the log-likelihood as σ increases. The case $k = 0$ corresponds to the Laplace approximation, and gives a poor-quality approximation for $\sigma > 0.5$. As k increases, the approximation improves. All values of $k \geq 3$ give an excellent approximation to the log-likelihood, and the approximations for $k = 4$ and $k = 5$ are indistinguishable at this scale.

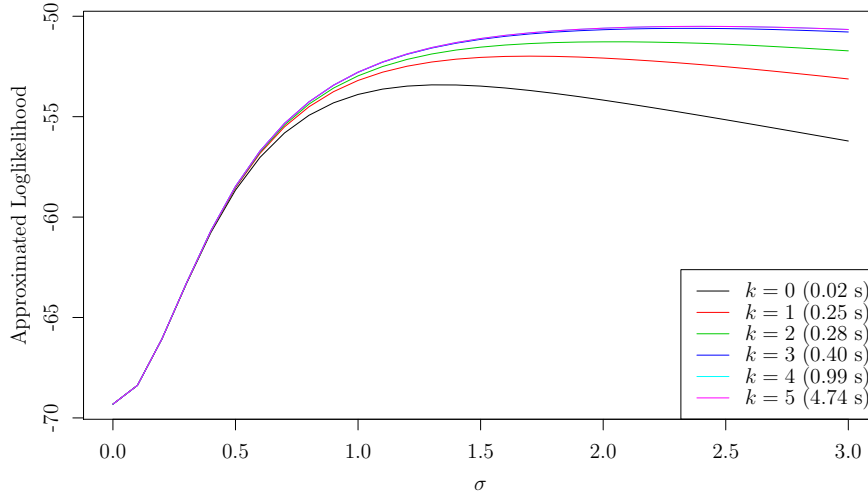


FIG 3. Sequential reduction approximations to $\ell(\beta = 0, \sigma)$, for various values of k . The curve for $k = 0$ (the Laplace approximation) is the lowest line, and the lines get higher as k increases. The curves for $k = 4$ and $k = 5$ are indistinguishable.

If we include all covariates suggested by Turner and Firth (2012) in the model, the maximum likelihood estimator is not finite. A penalized version of the likelihood could be used to obtain a finite estimate. In a generalized linear model, the bias-reduction penalty of Firth (1993) may be used for this purpose. Further work is required to obtain a good penalty for use with generalized linear mixed models.

4.3. A three-level model

Rabe-Hesketh, Skrondal and Pickles (2005) note that it is possible to simplify computation of the likelihood in models with nested random-effect structure. Using the sequential reduction method, there is no need to treat nested models as a special case. Their structure is automatically detected and exploited by the algorithm.

We demonstrate the method for a three-level model. Observations are made on items, where each item is contained within a level-1 group, and each level-1 group is itself contained in a level-2 group. The linear predictor is modeled as $\eta_i = \alpha + \beta x_i + \sigma_1 u_{g_1(i)} + \sigma_2 v_{g_2(i)}$, where $g_1(i)$ and $g_2(i)$ denote the first and second-level groups to which i belongs. We consider the case in which there are 100 second-level groups, each containing two first-level groups, which themselves each contain two items. The posterior dependence graph of this model is shown in Figure 1c, and has treewidth 2. The treewidth of the posterior dependence graph for a similarly defined L -level model is $L - 1$.

We suppose that $y_i \sim \text{Bernoulli}(p_i)$, where $p_i = \text{logit}^{-1}(\eta_i)$, and simulate from this model, with $\alpha = -0.5$, $\beta = 0.5$, $\sigma_1 = 1$ and $\sigma_2 = 0.5$. The fitted

TABLE 2
The parameter estimates and standard errors for the three-level model

| | | Laplace | k=1 | k=2 | k=3 | k=4 | k=5 |
|------------|----------|---------|-------|-------|-------|-------|-------|
| α | estimate | -0.46 | -0.50 | -0.50 | -0.50 | -0.50 | -0.50 |
| | s.e. | 0.17 | 0.19 | 0.19 | 0.19 | 0.19 | 0.19 |
| β | estimate | 0.45 | 0.49 | 0.49 | 0.49 | 0.49 | 0.49 |
| | s.e. | 0.23 | 0.25 | 0.25 | 0.25 | 0.25 | 0.25 |
| σ_1 | estimate | 0.54 | 0.92 | 0.90 | 0.89 | 0.89 | 0.89 |
| | s.e. | 0.38 | 0.33 | 0.35 | 0.35 | 0.35 | 0.35 |
| σ_2 | estimate | 0.54 | 0.53 | 0.57 | 0.58 | 0.58 | 0.58 |
| | s.e. | 0.27 | 0.32 | 0.32 | 0.32 | 0.32 | 0.32 |

values found using the sequential reduction method with various different values of k are shown in Table 2. The parameter estimates found from the Laplace approximation to the likelihood are some distance from the maximum likelihood estimator, especially for the variance parameter of the level-1 group.

5. 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. By using sparse grid interpolation methods to store modifications to the normal approximation used to construct the Laplace approximation, it is possible to get an accurate approximation to the likelihood for a wide range of models.

Acknowledgements

I am grateful to David Firth for useful discussions, and to a referee for helpful comments that improved the paper. This work was supported by the Engineering and Physical Sciences Research Council [grant numbers EP/P50578X/1, EP/K014463/1].

Supplementary Material

R code for sequential reduction

(doi: [10.1214/15-EJS991SUPP](https://doi.org/10.1214/15-EJS991SUPP); .zip). The R package `glmmsr` for sequential reduction, and code to reproduce the examples.

References

- BARTHELMANN, V., NOVAK, E. and RITTER, K. (2000). High Dimensional Polynomial Interpolation on Sparse Grids. *Advances in Computational Mathematics* **12** 273–288. [MR1768951](#)
- BESAG, J. (1974). Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society: Series B* **36** 192–236. [MR0373208](#)
- BODLAENDER, H. and KOSTER, A. (2008). Treewidth Computations I. Upper Bounds Technical Report, Department of Information and Computing Sciences, Utrecht University.
- BODLAENDER, H. and KOSTER, A. (2010). Treewidth Computations II. Lower Bounds Technical Report, Department of Information and Computing Sciences, Utrecht University. [MR2829452](#)
- BRADLEY, R. A. and TERRY, M. E. (1952). Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika* **39** 324–345. [MR0070925](#)
- BRESLOW, N. E. and CLAYTON, D. G. (1993). Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association* **88** 9–25.
- FIRTH, D. (1993). Bias Reduction of Maximum Likelihood Estimates. *Biometrika* **80** 27–38. [MR1225212](#)
- FONG, Y., RUE, H. and WAKEFIELD, J. (2010). Bayesian Inference for Generalized Linear Mixed Models. *Biostatistics* **11** 397–412.
- JORDAN, M. I. (2004). Graphical Models. *Statistical Science* **19** 140–155. [MR2082153](#)
- MOSTELLER, F. (1951). Remarks on the Method of Paired Comparisons: I. The Least Squares Solution Assuming Equal Standard Deviations and Equal Correlations. *Psychometrika* **16** 3–9.
- MURARASU, A., WEIDENDORFER, J., BUSE, G., BUTNARU, D. and PFLÜGER, D. (2011). Compact Data Structure and Scalable Algorithms for the Sparse Grid Technique. *SIGPLAN Notices* **46** 25–34.
- OGDEN, H. E. (2015). Supplement to “A sequential reduction method for inference in generalized linear mixed models”. DOI: [10.1214/15-EJS991SUPP](#).
- PINHEIRO, J. C. and BATES, D. M. (1995). Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model. *Journal of Computational and Graphical Statistics* **4** 12–35.
- RABE-HESKETH, S., SKRONDAL, A. and PICKLES, A. (2005). Maximum Likelihood Estimation of Limited and Discrete Dependent Variable Models With Nested Random Effects. *Journal of Econometrics* **128** 301–323. [MR2189555](#)
- RUE, H., MARTINO, S. and CHOPIN, N. (2009). Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations. *Journal of the Royal Statistical Society: Series B* **71** 319–392. [MR2649602](#)
- SHUN, Z. and MCCULLAGH, P. (1995). Laplace Approximation of High Dimensional Integrals. *Journal of the Royal Statistical Society: Series B* **57** 749–760. [MR1354079](#)

- R CORE TEAM (2014). R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.
- THURSTONE, L. L. (1927). A Law of Comparative Judgment. *Psychological Review* **34** 273–286.
- TIERNEY, L. and KADANE, J. B. (1986). Accurate Approximations for Posterior Moments and Marginal Densities. *Journal of the American Statistical Association* **81** 82–86. [MR0830567](#)
- TURNER, H. L. and FIRTH, D. (2012). Bradley-Terry Models in R: The BradleyTerry2 Package. *Journal of Statistical Software* **48**.
- WHITING, M. J., STUART-FOX, D. M., O’CONNOR, D., FIRTH, D., BENNETT, N. C. and BLOMBERG, S. P. (2006). Ultraviolet Signals Ultra-Aggression in a Lizard. *Animal Behaviour* **72** 353–363.