Computational methods for mixed models

Douglas Bates
Department of Statistics
University of Wisconsin – Madison

March 14, 2014

Abstract

The lme4 package provides R functions to fit and analyze several different types of mixed-effects models, including linear mixed models, generalized linear mixed models and nonlinear mixed models. In this vignette we describe the formulation of these models and the computational approach used to evaluate or approximate the log-likelihood of a model/data/parameter value combination.

1 Introduction

The lme4 package provides R functions to fit and analyze linear mixed models, generalized linear mixed models and nonlinear mixed models. These models are called *mixed-effects models* or, more simply, *mixed models* because they incorporate both *fixed-effects* parameters, which apply to an entire population or to certain well-defined and repeatable subsets of a population, and *random effects*, which apply to the particular experimental units or observational units in the study. Such models are also called *multilevel* models because the random effects represent levels of variation in addition to the per-observation noise term that is incorporated in common statistical models such as linear regression models, generalized linear models and nonlinear regression models.

We begin by describing common properties of these mixed models and the general computational approach used in the lme4 package. The estimates of the parameters in a mixed model are determined as the values that optimize

an objective function — either the likelihood of the parameters given the observed data, for maximum likelihood (ML) estimates, or a related objective function called the REML criterion. Because this objective function must be evaluated at many different values of the model parameters during the optimization process, we focus on the evaluation of the objective function and a critical computation in this evalution — determining the solution to a penalized, weighted least squares (PWLS) problem.

The dimension of the solution of the PWLS problem can be very large, perhaps in the millions. Furthermore, such problems must be solved repeatedly during the optimization process to determine parameter estimates. The whole approach would be infeasible were it not for the fact that the matrices determining the PWLS problem are sparse and we can use sparse matrix storage formats and sparse matrix computations (Davis, 2006). In particular, the whole computational approach hinges on the extraordinarily efficient methods for determining the Cholesky decomposition of sparse, symmetric, positive-definite matrices embodied in the CHOLMOD library of C functions (Davis, 2005).

In the next section we describe the general form of the mixed models that can be represented in the lme4 package and the computational approach embodied in the package. In the following section we describe a particular form of mixed model, called a linear mixed model, and the computational details for those models. In the fourth section we describe computational methods for generalized linear mixed models, nonlinear mixed models and generalized nonlinear mixed models.

2 Formulation of mixed models

A mixed-effects model incorporates two vector-valued random variables: the n-dimensional response vector, $\boldsymbol{\mathcal{Y}}$, and the q-dimensional random effects vector, $\boldsymbol{\mathcal{B}}$. We observe the value, $\boldsymbol{y}$, of $\boldsymbol{\mathcal{Y}}$. We do not observe the value of $\boldsymbol{\mathcal{B}}$.

The random variable $\boldsymbol{\mathcal{Y}}$ may be continuous or discrete. That is, the observed data, $\boldsymbol{y}$, may be on a continuous scale or they may be on a discrete scale, such as binary responses or responses representing a count. In our formulation, the random variable $\boldsymbol{\mathcal{B}}$ is always continous.

We specify a mixed model by describing the unconditional distribution of $\boldsymbol{\mathcal{B}}$ and the conditional distribution $(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{B}} = \boldsymbol{b})$.

2.1 The unconditional distribution of $\mathcal{B}$

In our formulation, the unconditional distribution of $\mathcal{B}$ is always a q-dimensional multivariate Gaussian (or "normal") distribution with mean $\mathbf{0}$ and with a parameterized covariance matrix,

$$\mathcal{B} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \Lambda(\theta)\Lambda'(\theta)\right). \tag{1}$$

The scalar, σ, in (1), is called the *common scale parameter*. As we will see later, not all types of mixed models incorporate this parameter. We will include σ^2 in the general form of the unconditional distribution of $\mathcal{B}$ with the understanding that, in some models, $\sigma \equiv 1$.

The $q \times q$ matrix $\Lambda(\theta)$, which is a left factor of the covariance matrix (when $\sigma = 1$) or the relative covariance matrix (when $\sigma \neq 1$), depends on an m-dimensional parameter θ. Typically $m \ll q$; in the examples we show below it is always the case that $m < 5$, even when q is in the thousands. The fact that m is very small is important because, as we shall see, determining the parameter estimates in a mixed model can be expressed as an optimization problem with respect to θ only.

The parameter θ may be, and typically is, subject to constraints. For ease of computation, we require that the constraints be expressed as "box" constraints of the form $\theta_{iL} \leq \theta_i \leq \theta_{iU}, i = 1, \ldots, m$ for constants θ_{iL} and $\theta_{iU}, i = 1, \ldots, m$. We shall write the set of such constraints as $\theta_L \leq \theta \leq \theta_R$. The matrix $\Lambda(\theta)$ is required to be non-singular (i.e. invertible) when θ is not on the boundary.

2.2 The conditional distribution, $(\mathcal{Y}|\mathcal{B} = b)$

The conditional distribution, $(\mathcal{Y}|\mathcal{B} = b)$, must satisfy:

1. The conditional mean, $\mu_{\mathcal{Y}|\mathcal{B}}(b) = \mathrm{E}[\mathcal{Y}|\mathcal{B} = b]$, depends on b only through the value of the *linear predictor*, $Zb + X\beta$, where β is the p-dimensional *fixed-effects* parameter vector and the *model matrices*, Z and X, are fixed matrices of the appropriate dimension. That is, the two model matrices must have the same number of rows and must have q and p columns, respectively. The number of rows in Z and X is a multiple of n, the dimension of y.

2. The scalar distributions, $(\mathcal{Y}_i|\mathcal{B} = b), i = 1, \ldots, n$, all have the same form and are completely determined by the conditional mean, $\mu_{\mathcal{Y}|\mathcal{B}}(b)$

and, at most, one additional parameter, σ, which is the common scale parameter.

3. The scalar distributions, $(\mathcal{Y}_i|\boldsymbol{\mathcal{B}} = \boldsymbol{b}), i = 1, \ldots, n$, are independent. That is, the components of $\boldsymbol{\mathcal{Y}}$ are *conditionally independent* given $\boldsymbol{\mathcal{B}}$.

An important special case of the conditional distribution is the multivariate Gaussian distribution of the form

$$(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{B}} = \boldsymbol{b}) \sim \mathcal{N}(\boldsymbol{Z}\boldsymbol{b} + \boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n) \tag{2}$$

where $\boldsymbol{I}_n$ denotes the identity matrix of size n. In this case the conditional mean, $\boldsymbol{\mu}_{\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{B}}}(\boldsymbol{b})$, is exactly the linear predictor, $\boldsymbol{Z}\boldsymbol{b} + \boldsymbol{X}\boldsymbol{\beta}$, a situation we will later describe as being an "identity link" between the conditional mean and the linear predictor. Models with conditional distribution (2) are called *linear mixed models*.

2.3 A change of variable to "spherical" random effects

Because the conditional distribution $(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{B}} = \boldsymbol{b})$ depends on $\boldsymbol{b}$ only through the linear predictor, it is easy to express the model in terms of a linear transformation of $\boldsymbol{\mathcal{B}}$. We define the linear transformation from a q-dimensional "spherical" Gaussian random variable, $\boldsymbol{\mathcal{U}}$, to $\boldsymbol{\mathcal{B}}$ as

$$\boldsymbol{\mathcal{B}} = \boldsymbol{\Lambda}(\boldsymbol{\theta})\boldsymbol{\mathcal{U}}, \quad \boldsymbol{\mathcal{U}} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_q). \tag{3}$$

(The term "spherical" refers to the fact that contours of constant probability density for $\boldsymbol{\mathcal{U}}$ are spheres centered at the mean — in this case, $\boldsymbol{0}$.)

When $\boldsymbol{\theta}$ is not on the boundary this is an invertible transformation. When $\boldsymbol{\theta}$ is on the boundary the transformation can fail to be invertible. However, we will only need to be able to express $\boldsymbol{\mathcal{B}}$ in terms of $\boldsymbol{\mathcal{U}}$ and that transformation is well-defined, even when $\boldsymbol{\theta}$ is on the boundary.

The linear predictor, as a function of $\boldsymbol{u}$, is

$$\boldsymbol{\gamma}(\boldsymbol{u}) = \boldsymbol{Z}\boldsymbol{\Lambda}(\boldsymbol{\theta})\boldsymbol{u} + \boldsymbol{X}\boldsymbol{\beta}. \tag{4}$$

When we wish to emphasize the role of the model parameters, $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, in the formulation of $\boldsymbol{\gamma}$, we will write the linear predictor as $\boldsymbol{\gamma}(\boldsymbol{u}, \boldsymbol{\theta}, \boldsymbol{\beta})$.

2.4 The conditional density $(\mathcal{U}|\mathcal{Y} = \boldsymbol{y})$

Because we observe $\boldsymbol{y}$ and do not observe $\boldsymbol{b}$ or $\boldsymbol{u}$, the conditional distribution of interest, for the purposes of statistical inference, is $(\mathcal{U}|\mathcal{Y} = \boldsymbol{y})$ (or, equivalently, $(\mathcal{B}|\mathcal{Y} = \boldsymbol{y})$). This conditional distribution is always a continuous distribution with conditional probability density $f_{\mathcal{U}|\mathcal{Y}}(\boldsymbol{u}|\boldsymbol{y})$.

We can evaluate $f_{\mathcal{U}|\mathcal{Y}}(\boldsymbol{u}|\boldsymbol{y})$, up to a constant, as the product of the unconditional density, $f_{\mathcal{U}}(\boldsymbol{u})$, and the conditional density (or the probability mass function, whichever is appropriate), $f_{\mathcal{Y}|\mathcal{U}}(\boldsymbol{y}|\boldsymbol{u})$. We write this unnormalized conditional density as

$$h(\boldsymbol{u}|\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma) = f_{\mathcal{Y}|\mathcal{U}}(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma) f_{\mathcal{U}}(\boldsymbol{u}|\sigma). \tag{5}$$

We say that h is the "unnormalized" conditional density because all we know is that the conditional density is proportional to $h(\boldsymbol{u}|\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma)$. To obtain the conditional density we must normalize h by dividing by the value of the integral

$$L(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma|\boldsymbol{y}) = \int_{\mathbb{R}^q} h(\boldsymbol{u}|\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma)\, d\boldsymbol{u}. \tag{6}$$

We write the value of the integral (6) as $L(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma|\boldsymbol{y})$ because it is exactly the *likelihood* of the parameters $\boldsymbol{\theta}$, $\boldsymbol{\beta}$ and σ, given the observed data $\boldsymbol{y}$. The *maximum likelihood (ML) estimates* of these parameters are the values that maximize L.

2.5 Determining the ML estimates

The general problem of maximizing $L(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma|\boldsymbol{y})$ with respect to $\boldsymbol{\theta}$, $\boldsymbol{\beta}$ and σ can be formidable because each evaluation of this function involves a potentially high-dimensional integral and because the dimension of $\boldsymbol{\beta}$ can be large. However, this general optimization problem can be split into manageable subproblems. Given a value of $\boldsymbol{\theta}$ we can determine the *conditional mode*, $\tilde{\boldsymbol{u}}(\boldsymbol{\theta})$, of $\boldsymbol{u}$ and the *conditional estimate*, $\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$ simultaneously using *penalized, iteratively re-weighted least squares* (PIRLS). The conditional mode and the conditional estimate are defined as

$$\begin{bmatrix} \tilde{\boldsymbol{u}}(\boldsymbol{\theta}) \\ \tilde{\boldsymbol{\beta}}(\boldsymbol{\theta}) \end{bmatrix} = \arg\max_{\boldsymbol{u},\boldsymbol{\beta}} h(\boldsymbol{u}|\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma). \tag{7}$$

(It may look as if we have missed the dependence on σ on the left-hand side but it turns out that the scale parameter does not affect the location of the optimal values of quantities in the linear predictor.)

As is common in such optimization problems, we re-express the conditional density on the *deviance scale*, which is negative twice the logarithm of the density, where the optimization becomes

$$\begin{bmatrix}\tilde{\boldsymbol{u}}(\boldsymbol{\theta})\\ \tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})\end{bmatrix} = \arg\min_{\boldsymbol{u},\boldsymbol{\beta}} -2\log\left(h(\boldsymbol{u}|\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\beta},\sigma)\right). \tag{8}$$

It is this optimization problem that can be solved quite efficiently using PIRLS. In fact, for linear mixed models, which are described in the next section, $\tilde{\boldsymbol{u}}(\boldsymbol{\theta})$ and $\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$ can be directly evaluated.

The second-order Taylor series expansion of $-2\log h$ at $\tilde{\boldsymbol{u}}(\boldsymbol{\theta})$ and $\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$ provides the Laplace approximation to the profiled deviance. Optimizing this function with respect to $\boldsymbol{\theta}$ provides the ML estimates of $\boldsymbol{\theta}$, from which the ML estimates of $\boldsymbol{\beta}$ and σ (if used) are derived.

3 Methods for linear mixed models

As indicated in the introduction, a critical step in our methods for determining the maximum likelihood estimates of the parameters in a mixed model is solving a penalized, weighted least squares (PWLS) problem. We will motivate the general form of the PWLS problem by first considering computational methods for linear mixed models that result in a penalized least squares (PLS) problem.

Recall from §2.2 that, in a linear mixed model, both the conditional distribution, $(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{U}}=\boldsymbol{u})$, and the unconditional distribution, $\boldsymbol{\mathcal{U}}$, are spherical Gaussian distributions and that the conditional mean, $\boldsymbol{\mu}_{\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{U}}}(\boldsymbol{u})$, is the linear predictor, $\boldsymbol{\gamma}(\boldsymbol{u})$. Because all the distributions determining the model are continuous distributions, we consider their densities. On the deviance scale these are

$$\begin{aligned}
-2\log(f_{\boldsymbol{\mathcal{U}}}(\boldsymbol{u})) &= q\log(2\pi\sigma^2) + \frac{\|\boldsymbol{u}\|^2}{\sigma^2}\\
-2\log(f_{\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{U}}}(\boldsymbol{y}|\boldsymbol{u})) &= n\log(2\pi\sigma^2) + \frac{\|\boldsymbol{y}-\boldsymbol{Z}\boldsymbol{\Lambda}(\boldsymbol{\theta})\boldsymbol{u}-\boldsymbol{X}\boldsymbol{\beta}\|^2}{\sigma^2}\\
-2\log(h(\boldsymbol{u}|\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\beta},\sigma)) &= (n+q)\log(2\pi\sigma^2) + \frac{\|\boldsymbol{y}-\boldsymbol{\gamma}(\boldsymbol{u},\boldsymbol{\theta},\boldsymbol{\beta})\|^2+\|\boldsymbol{u}\|^2}{\sigma^2}\\
&= (n+q)\log(2\pi\sigma^2) + \frac{d(\boldsymbol{u}|\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\beta})}{\sigma^2}
\end{aligned} \tag{9}$$

In (9) the *discrepancy* function,

$$d(\boldsymbol{u}|\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{\gamma}(\boldsymbol{u},\boldsymbol{\theta},\boldsymbol{\beta})\|^2 + \|\boldsymbol{u}\|^2 \tag{10}$$

has the form of a penalized residual sum of squares in that the first term, $\|\boldsymbol{y} - \boldsymbol{\gamma}(\boldsymbol{u},\boldsymbol{\theta},\boldsymbol{\beta})\|^2$ is the residual sum of squares for $\boldsymbol{y}$, $\boldsymbol{u}$, $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ and the second term, $\|\boldsymbol{u}\|^2$, is a penalty on the size of $\boldsymbol{u}$. Notice that the discrepancy does not depend on the common scale parameter, σ.

3.1 The canonical form of the discrepancy

Using a so-called "pseudo data" representation, we can write the discrepancy as a residual sum of squares for a regression model that is linear in both $\boldsymbol{u}$ and $\boldsymbol{\beta}$

$$d(\boldsymbol{u}|\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\beta}) = \left\| \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{bmatrix} - \begin{bmatrix} \boldsymbol{Z}\boldsymbol{\Lambda}(\boldsymbol{\theta}) & \boldsymbol{X} \\ \boldsymbol{I}_q & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{\beta} \end{bmatrix} \right\|^2 . \tag{11}$$

The term "pseudo data" reflects the fact that we have added q "pseudo observations" to the observed response, $\boldsymbol{y}$, and to the linear predictor, $\boldsymbol{\gamma}(\boldsymbol{u},\boldsymbol{\theta},\boldsymbol{\beta}) = \boldsymbol{Z}\boldsymbol{\Lambda}(\boldsymbol{\theta})\boldsymbol{u} + \boldsymbol{X}\boldsymbol{\beta}$, in such a way that their contribution to the overall residual sum of squares is exactly the penalty term in the discrepancy.

In the form (11) we can see that the discrepancy is a quadratic form in both $\boldsymbol{u}$ and $\boldsymbol{\beta}$. Furthermore, because we require that $\boldsymbol{X}$ has full column rank, the discrepancy is a positive-definite quadratic form in $\boldsymbol{u}$ and $\boldsymbol{\beta}$ that is minimized at $\tilde{\boldsymbol{u}}(\boldsymbol{\theta})$ and $\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$ satisfying

$$\begin{bmatrix} \boldsymbol{\Lambda}'(\boldsymbol{\theta})\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Lambda}(\theta) + \boldsymbol{I}_q & \boldsymbol{\Lambda}'(\boldsymbol{\theta})\boldsymbol{Z}'\boldsymbol{X} \\ \boldsymbol{X}'\boldsymbol{Z}\boldsymbol{\Lambda}(\theta) & \boldsymbol{X}'\boldsymbol{X} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{u}}(\boldsymbol{\theta}) \\ \tilde{\boldsymbol{\beta}}(\boldsymbol{\theta}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Lambda}'(\boldsymbol{\theta})\boldsymbol{Z}'\boldsymbol{y} \\ \boldsymbol{X}'\boldsymbol{y} \end{bmatrix} \tag{12}$$

An effective way of determining the solution to a sparse, symmetric, positive definite system of equations such as (12) is the sparse Cholesky decomposition (Davis, 2006). If $\boldsymbol{A}$ is a sparse, symmetric positive definite matrix then the sparse Cholesky factor with fill-reducing permutation $\boldsymbol{P}$ is the lower-triangular matrix $\boldsymbol{L}$ such that

$$\boldsymbol{L}\boldsymbol{L}' = \boldsymbol{P}\boldsymbol{A}\boldsymbol{P}'. \tag{13}$$

(Technically, the factor $\boldsymbol{L}$ is only determined up to changes in the sign of the diagonal elements. By convention we require the diagonal elements to be positive.)

The fill-reducing permutation represented by the permutation matrix $\boldsymbol{P}$, which is determined from the pattern of nonzeros in $\boldsymbol{A}$ but does not depend on particular values of those nonzeros, can have a profound impact on the number of nonzeros in $\boldsymbol{L}$ and hence on the speed with which $\boldsymbol{L}$ can be calculated from $\boldsymbol{A}$.

In most applications of linear mixed models the matrix $\boldsymbol{Z}\boldsymbol{\Lambda}(\boldsymbol{\theta})$ is sparse while $\boldsymbol{X}$ is dense or close to it so the permutation matrix $\boldsymbol{P}$ can be restricted to the form

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{P_Z} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{P_X} \end{bmatrix} \tag{14}$$

without loss of efficiency. In fact, in most cases we can set $\boldsymbol{P_X} = \boldsymbol{I}_p$ without loss of efficiency.

Let us assume that the permutation matrix is required to be of the form (14) so that we can write the Cholesky factorization for the positive definite system (12) as

$$\begin{bmatrix} \boldsymbol{L_Z} & \boldsymbol{0} \\ \boldsymbol{L_{XZ}} & \boldsymbol{L_X} \end{bmatrix} \begin{bmatrix} \boldsymbol{L_Z} & \boldsymbol{0} \\ \boldsymbol{L_{XZ}} & \boldsymbol{L_X} \end{bmatrix}' = \begin{bmatrix} \boldsymbol{P_Z} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{P_X} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Lambda}'(\boldsymbol{\theta})\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Lambda}(\theta) + \boldsymbol{I}_q & \boldsymbol{\Lambda}'(\boldsymbol{\theta})\boldsymbol{Z}'\boldsymbol{X} \\ \boldsymbol{X}'\boldsymbol{Z}\boldsymbol{\Lambda}(\theta) & \boldsymbol{X}'\boldsymbol{X} \end{bmatrix} \begin{bmatrix} \boldsymbol{P_Z} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{P_X} \end{bmatrix}'. \tag{15}$$

The discrepancy can now be written in the canonical form

$$d(\boldsymbol{u}|\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\beta}) = \tilde{d}(\boldsymbol{y}, \boldsymbol{\theta}) + \left\| \begin{bmatrix} \boldsymbol{L}'_{\boldsymbol{Z}} & \boldsymbol{L}'_{\boldsymbol{XZ}} \\ \boldsymbol{0} & \boldsymbol{L}'_{\boldsymbol{X}} \end{bmatrix} \begin{bmatrix} \boldsymbol{P_Z}(\boldsymbol{u} - \tilde{\boldsymbol{u}}) \\ \boldsymbol{P_X}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) \end{bmatrix} \right\|^2 \tag{16}$$

where

$$\tilde{d}(\boldsymbol{y}, \boldsymbol{\theta}) = d(\tilde{\boldsymbol{u}}(\boldsymbol{\theta})|\boldsymbol{y}, \boldsymbol{\theta}, \tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})) \tag{17}$$

is the minimum discrepancy, given $\boldsymbol{\theta}$.

3.2 The profiled likelihood for linear mixed models

Substituting (16) into (9) provides the unnormalized conditional density $h(\boldsymbol{u}|\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma)$ on the deviance scale as

$$-2\log(h(\boldsymbol{u}|\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma)) = (n+q)\log(2\pi\sigma^2) + \frac{\tilde{d}(\boldsymbol{y}, \boldsymbol{\theta}) + \left\| \begin{bmatrix} \boldsymbol{L}'_{\boldsymbol{Z}} & \boldsymbol{L}'_{\boldsymbol{XZ}} \\ \boldsymbol{0} & \boldsymbol{L}'_{\boldsymbol{X}} \end{bmatrix} \begin{bmatrix} \boldsymbol{P_Z}(\boldsymbol{u} - \tilde{\boldsymbol{u}}) \\ \boldsymbol{P_X}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) \end{bmatrix} \right\|^2}{\sigma^2}. \tag{18}$$

As shown in Appendix B, the integral of a quadratic form on the deviance scale, such as (18), is easily evaluated, providing the log-likelihood, $\ell(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma|\boldsymbol{y})$, as

$$\begin{aligned} -2\ell(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma|\boldsymbol{y}) & \\ &= -2\log\left(L(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma|\boldsymbol{y})\right) \\ &= n\log(2\pi\sigma^2) + \log(|\boldsymbol{L_Z}|^2) + \frac{\tilde{d}(\boldsymbol{y}, \boldsymbol{\theta}) + \left\|\boldsymbol{L}'_{\boldsymbol{X}}\boldsymbol{P_X}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})\right\|^2}{\sigma^2}, \end{aligned} \quad (19)$$

from which we can see that the conditional estimate of $\boldsymbol{\beta}$, given $\boldsymbol{\theta}$, is $\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$ and the conditional estimate of σ, given $\boldsymbol{\theta}$, is

$$\tilde{\sigma^2}(\boldsymbol{\theta}) = \frac{\tilde{d}(\boldsymbol{\theta}|\boldsymbol{y})}{n}. \quad (20)$$

Substituting these conditional estimates into (19) produces the *profiled likelihood*, $\tilde{L}(\boldsymbol{\theta}|\boldsymbol{y})$, as

$$-2\tilde{\ell}(\boldsymbol{\theta}|\boldsymbol{y})) = \log(|\boldsymbol{L_Z}(\boldsymbol{\theta})|^2) + n\left(1 + \log\left(\frac{2\pi\tilde{d}(\boldsymbol{y}, \boldsymbol{\theta})}{n}\right)\right). \quad (21)$$

The maximum likelihood estimate of $\boldsymbol{\theta}$ can then be expressed as

$$\widehat{\boldsymbol{\theta}}_L = \arg\min_{\boldsymbol{\theta}}\left(-2\tilde{\ell}(\boldsymbol{\theta}|\boldsymbol{y})\right). \quad (22)$$

from which the ML estimates of σ^2 and $\boldsymbol{\beta}$ are evaluated as

$$\widehat{\sigma^2_L} = \frac{\tilde{d}(\widehat{\boldsymbol{\theta}}_L, \boldsymbol{y})}{n} \quad (23)$$

$$\widehat{\boldsymbol{\beta}}_L = \tilde{\boldsymbol{\beta}}(\widehat{\boldsymbol{\theta}}_L). \quad (24)$$

The important thing to note about optimizing the profiled likelihood, (21), is that it is a m-dimensional optimization problem and typically m is very small.

3.3 The REML criterion

In practice the so-called REML estimates of variance components are often preferred to the maximum likelihood estimates. ("REML" can be considered

to be an acronym for "restricted" or "residual" maximum likelihood, although neither term is completely accurate because these estimates do not maximize a likelihood.) We can motivate the use of the REML criterion by considering a linear regression model,

$$\boldsymbol{\mathcal{Y}} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n), \tag{25}$$

in which we typically estimate σ^2 by

$$\widehat{\sigma^2_R} = \frac{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2}{n-p} \tag{26}$$

even though the maximum likelihood estimate of σ^2 is

$$\widehat{\sigma^2_L} = \frac{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2}{n}. \tag{27}$$

The argument for preferring $\widehat{\sigma^2_R}$ to $\widehat{\sigma^2_L}$ as an estimate of σ^2 is that the numerator in both estimates is the sum of squared residuals at $\widehat{\boldsymbol{\beta}}$ and, although the residual vector $\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$ is an n-dimensional vector, the residual at $\widehat{\boldsymbol{\theta}}$ satisfies p linearly independent constraints, $\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = \boldsymbol{0}$. That is, the residual at $\widehat{\boldsymbol{\theta}}$ is the projection of the observed response vector, $\boldsymbol{y}$, into an $(n-p)$-dimensional linear subspace of the n-dimensional response space. The estimate $\widehat{\sigma^2_R}$ takes into account the fact that σ^2 is estimated from residuals that have only $n-p$ *degrees of freedom*.

The REML criterion for determining parameter estimates $\widehat{\boldsymbol{\theta}}_R$ and $\widehat{\sigma^2_R}$ in a linear mixed model has the property that these estimates would specialize to $\widehat{\sigma^2_R}$ from (26) for a linear regression model. Although not usually derived in this way, the REML criterion can be expressed as

$$c_R(\boldsymbol{\theta}, \boldsymbol{\sigma}|\boldsymbol{y}) = -2\log\int_{\mathbb{R}^p} L(\boldsymbol{u}|\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma)\, d\boldsymbol{\beta} \tag{28}$$

on the deviance scale. The REML estimates $\widehat{\boldsymbol{\theta}}_R$ and $\widehat{\sigma^2_R}$ minimize $c_R(\boldsymbol{\theta}, \boldsymbol{\sigma}|\boldsymbol{y})$.

The profiled REML criterion, a function of $\boldsymbol{\theta}$ only, is

$$\tilde{c}_R(\boldsymbol{\theta}|\boldsymbol{y}) = \log(|\boldsymbol{L}_{\boldsymbol{Z}}(\boldsymbol{\theta})|^2|\boldsymbol{L}_{\boldsymbol{X}}(\boldsymbol{\theta})|^2) + (n-p)\left(1 + \log\left(\frac{2\pi\tilde{d}(\boldsymbol{\theta}|\boldsymbol{y})}{n-p}\right)\right) \tag{29}$$

and the REML estimate of $\boldsymbol{\theta}$ is

$$\widehat{\boldsymbol{\theta}}_R = \arg\min_{\boldsymbol{\theta}} \tilde{c}_R(\boldsymbol{\theta}, \boldsymbol{y}). \tag{30}$$

The REML estimate of σ^2 is $\widehat{\sigma^2_R} = \tilde{d}(\widehat{\boldsymbol{\theta}}_R|\boldsymbol{y})/(n-p)$.

It is not entirely clear how one would define a "REML estimate" of $\boldsymbol{\beta}$ because the REML criterion, $c_R(\boldsymbol{\theta}, \boldsymbol{\sigma}|\boldsymbol{y})$, defined in (28), does not depend on $\boldsymbol{\beta}$. However, it is customary (and not unreasonable) to use $\widehat{\boldsymbol{\beta}}_R = \tilde{\boldsymbol{\beta}}(\widehat{\boldsymbol{\theta}}_R)$ as the REML estimate of $\boldsymbol{\beta}$.

Note that the profiled REML criterion can be evaluated from a sparse Cholesky decomposition like that in (15) but without the requirement that the permutation can be applied to the columns of $\boldsymbol{Z\Lambda}(\boldsymbol{\theta})$ separately from the columnns of $\boldsymbol{X}$. That is, we can use a general fill-reducing permutation rather than the specific form (14) with separate permutations represented by $\boldsymbol{P_Z}$ and $\boldsymbol{P_X}$. This can be useful in cases where both $\boldsymbol{Z}$ and $\boldsymbol{X}$ are large and sparse.

3.4 Summary for linear mixed models

A linear mixed model is characterized by the conditional distribution

$$(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{U}} = \boldsymbol{u}) \sim \mathcal{N}(\boldsymbol{\gamma}(\boldsymbol{u}, \boldsymbol{\theta}, \boldsymbol{\beta}), \sigma^2 \boldsymbol{I}_n) \text{ where } \boldsymbol{\gamma}(\boldsymbol{u}, \boldsymbol{\theta}, \boldsymbol{\beta}) = \boldsymbol{Z\Lambda}(\boldsymbol{\theta})\boldsymbol{u} + \boldsymbol{X\beta} \tag{31}$$

and the unconditional distribution $\boldsymbol{\mathcal{U}} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_q)$. The discrepancy function,

$$d(\boldsymbol{u}|\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{\gamma}(\boldsymbol{u}, \boldsymbol{\theta}, \boldsymbol{\beta})\|^2 + \|\boldsymbol{u}\|^2,$$

is minimized at the conditional mode, $\tilde{\boldsymbol{u}}(\boldsymbol{\theta})$, and the conditional estimate, $\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$, which are the solutions to the sparse, positive-definite linear system

$$\begin{bmatrix} \boldsymbol{\Lambda}'(\boldsymbol{\theta})\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Lambda}(\theta) + \boldsymbol{I}_q & \boldsymbol{\Lambda}'(\boldsymbol{\theta})\boldsymbol{Z}'\boldsymbol{X} \\ \boldsymbol{X}'\boldsymbol{Z}\boldsymbol{\Lambda}(\theta) & \boldsymbol{X}'\boldsymbol{X} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{u}}(\boldsymbol{\theta}) \\ \tilde{\boldsymbol{\beta}}(\boldsymbol{\theta}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Lambda}'(\boldsymbol{\theta})\boldsymbol{Z}'\boldsymbol{y} \\ \boldsymbol{X}'\boldsymbol{y} \end{bmatrix}.$$

In the process of solving this system we create the sparse left Cholesky factor, $L_{\boldsymbol{Z}}(\boldsymbol{\theta})$, which is a lower triangular sparse matrix satisfying

$$\boldsymbol{L_Z}(\boldsymbol{\theta})\boldsymbol{L_Z}(\boldsymbol{\theta})' = \boldsymbol{P_Z}\left(\boldsymbol{\Lambda}'(\boldsymbol{\theta})\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Lambda}(\theta) + \boldsymbol{I}_q\right)\boldsymbol{P}_{\boldsymbol{Z}}'$$

where $\boldsymbol{P_Z}$ is a permutation matrix representing a fill-reducing permutation formed from the pattern of nonzeros in $\boldsymbol{Z\Lambda}(\boldsymbol{\theta})$ for any $\boldsymbol{\theta}$ not on the boundary

of the parameter region. (The values of the nonzeros depend on $\boldsymbol{\theta}$ but the pattern doesn't.)

The profiled log-likelihood, $\tilde{\ell}(\boldsymbol{\theta}|\boldsymbol{y})$, is

$$-2\tilde{\ell}(\boldsymbol{\theta}|\boldsymbol{y}) = \log(|\boldsymbol{L}_{\boldsymbol{Z}}(\boldsymbol{\theta})|^2) + n\left(1 + \log\left(\frac{2\pi\tilde{d}(\boldsymbol{y},\boldsymbol{\theta})}{n}\right)\right)$$

where $\tilde{d}(\boldsymbol{y},\boldsymbol{\theta}) = d(\tilde{\boldsymbol{u}}(\boldsymbol{\theta})|\boldsymbol{y},\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta}),\boldsymbol{\theta})$.

4 Generalizing the discrepancy function

Because one of the factors influencing the choice of implementation for linear mixed models is the extent to which the methods can also be applied to other mixed models, we describe several other classes of mixed models before discussing the implementation details for linear mixed models. At the core of our methods for determining the maximum likelihood estimates (MLEs) of the parameters in the mixed model are methods for minimizing the discrepancy function with respect to the coefficients $\boldsymbol{u}$ and $\boldsymbol{\beta}$ in the linear predictor $\boldsymbol{\gamma}(\boldsymbol{u},\boldsymbol{\theta},\boldsymbol{\beta})$.

In this section we describe the general form of the discrepancy function that we will use and a penalized iteratively reweighted least squares (PIRLS) algorithm for determining the conditional modes $\tilde{\boldsymbol{u}}(\boldsymbol{\theta})$ and $\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$. We then describe several types of mixed models and the form of the discrepancy function for each.

4.1 A weighted residual sum of squares

As shown in §3.1, the discrepancy function for a linear mixed model has the form of a penalized residual sum of squares from a linear model (11). In this section we generalize that definition to

$$d(\boldsymbol{u}|\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\beta}) = \left\|\boldsymbol{W}^{1/2}(\boldsymbol{\mu})\left[\boldsymbol{y} - \boldsymbol{\mu}_{\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{U}}}(\boldsymbol{u},\boldsymbol{\theta},\boldsymbol{\beta})\right]\right\|^2 + \|\boldsymbol{0} - \boldsymbol{u}\|^2. \tag{32}$$

where $\boldsymbol{W}$ is an $n\times n$ diagonal matrix, called the *weights matrix*, with positive diagonal elements and $\boldsymbol{W}^{1/2}$ is the diagonal matrix with the square roots of the weights on the diagonal. The ith weight is inversely proportional to the conditional variances of $(\mathcal{Y}|\boldsymbol{\mathcal{U}} = \boldsymbol{u})$ and may depend on the conditional mean, $\boldsymbol{\mu}_{\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{U}}}$.

We allow the conditional mean to be a nonlinear function of the linear predictor, but with certain restrictions. We require that the mapping from $\boldsymbol{u}$ to $\boldsymbol{\mu}_{\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{U}}=\boldsymbol{u}}$ be expressed as

$$\boldsymbol{u} \;\rightarrow\; \boldsymbol{\gamma} \;\rightarrow\; \boldsymbol{\eta} \;\rightarrow\; \boldsymbol{\mu} \tag{33}$$

where $\boldsymbol{\gamma} = \boldsymbol{Z}\boldsymbol{\Lambda}(\boldsymbol{\theta})\boldsymbol{u} + \boldsymbol{X}\boldsymbol{\theta}$ is an ns-dimensional vector ($s > 0$) while $\boldsymbol{\eta}$ and $\boldsymbol{\mu}$ are n-dimensional vectors.

The map $\boldsymbol{\eta} \rightarrow \boldsymbol{\mu}$ has the property that μ_i depends only on η_i, $i = 1, \ldots, n$. The map $\boldsymbol{\gamma} \rightarrow \boldsymbol{\eta}$ has a similar property in that, if we write $\boldsymbol{\gamma}$ as an $n \times s$ matrix $\boldsymbol{\Gamma}$ such that

$$\boldsymbol{\gamma} = \operatorname{vec} \boldsymbol{\Gamma} \tag{34}$$

(i.e. concatenating the columns of $\boldsymbol{\Gamma}$ produces $\boldsymbol{\gamma}$) then η_i depends only on the ith row of $\boldsymbol{\Gamma}$, $i = 1, \ldots, n$. Thus the Jacobian matrix $\frac{d\boldsymbol{\mu}}{d\boldsymbol{\eta}'}$ is an $n \times n$ diagonal matrix and the Jacobian matrix $\frac{d\boldsymbol{\eta}}{d\boldsymbol{\gamma}'}$ is the horizontal concatenation of s diagonal $n \times n$ matrices.

For historical reasons, the function that maps η_i to μ_i is called the *inverse link* function and is written $\mu = g^{-1}(\eta)$. The *link function*, naturally, is $\eta = g(\mu)$. When applied component-wise to vectors $\boldsymbol{\mu}$ or $\boldsymbol{\eta}$ we write these as $\boldsymbol{\eta} = \boldsymbol{g}(\boldsymbol{\mu})$ and $\boldsymbol{\mu} = \boldsymbol{g}^{-1}(\boldsymbol{\eta})$.

Recall that the conditional distribution, $(\mathcal{Y}_i|\boldsymbol{\mathcal{U}} = \boldsymbol{u})$, is required to be independent of $(\mathcal{Y}_j|\boldsymbol{\mathcal{U}} = \boldsymbol{u})$ for $i, j = 1, \ldots, n$, $i \neq j$ and that all the component conditional distributions must be of the same form and differ only according to the value of the conditional mean.

Depending on the family of the conditional distributions, the allowable values of the μ_i may be in a restricted range. For example, if the conditional distributions are Bernoulli then $0 \leq \mu_i \leq 1, i = 1, \ldots, n$. If the conditional distributions are Poisson then $0 \leq \mu_i, i = 1, \ldots, n$. A characteristic of the link function, g, is that it must map the restricted range to an unrestricted range. That is, a link function for the Bernoulli distribution must map $[0, 1]$ to $[-\infty, \infty]$ and must be invertible within the range.

The mapping from $\boldsymbol{\gamma}$ to $\boldsymbol{\eta}$ is defined by a function $m : \mathbb{R}^s \rightarrow \mathbb{R}$, called the *nonlinear model* function, such that $\eta_i = m(\boldsymbol{\gamma}_i), i = 1, \ldots, n$ where $\boldsymbol{\gamma}_i$ is the ith row of $\boldsymbol{\Gamma}$. The vector-valued function is $\boldsymbol{\eta} = \boldsymbol{m}(\boldsymbol{\gamma})$.

Determining the conditional modes, $\tilde{\boldsymbol{u}}(\boldsymbol{y}|\boldsymbol{\theta})$, and $\tilde{\boldsymbol{\beta}}(\boldsymbol{y}|\boldsymbol{\theta})$, that jointly minimize the discrepancy,

$$\begin{bmatrix} \tilde{\boldsymbol{u}}(\boldsymbol{y}|\boldsymbol{\theta}) \\ \tilde{\boldsymbol{\beta}}(\boldsymbol{y}|\boldsymbol{\theta}) \end{bmatrix} = \arg\min_{\boldsymbol{u},\boldsymbol{\beta}} \left[(\boldsymbol{y} - \boldsymbol{\mu})' \boldsymbol{W} (\boldsymbol{y} - \boldsymbol{\mu}) + \|\boldsymbol{u}\|^2 \right] \tag{35}$$

becomes a weighted, nonlinear least squares problem except that the weights, $\boldsymbol{W}$, can depend on $\boldsymbol{\mu}$ and, hence, on $\boldsymbol{u}$ and $\boldsymbol{\beta}$.

In describing an algorithm for linear mixed models we called $\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$ the *conditional estimate*. That name reflects that fact that this is the maximum likelihood estimate of $\boldsymbol{\beta}$ for that particular value of $\boldsymbol{\theta}$. Once we have determined the MLE, $\widehat{(\boldsymbol{\theta})}_L$ of $\boldsymbol{\theta}$, we have a "plug-in" estimator, $\widehat{\boldsymbol{\beta}}_L = \tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$ for $\boldsymbol{\beta}$.

This property does not carry over exactly to other forms of mixed models. The values $\tilde{\boldsymbol{u}}(\boldsymbol{\theta})$ and $\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$ are conditional modes in the sense that they are the coefficients in $\boldsymbol{\gamma}$ that jointly maximize the unscaled conditional density $h(\boldsymbol{u}|\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma)$. Here we are using the adjective "conditional" more in the sense of conditioning on $\boldsymbol{\mathcal{Y}} = \boldsymbol{y}$ than in the sense of conditioning on $\boldsymbol{\theta}$, although these values are determined for a fixed value of $\boldsymbol{\theta}$.

4.2 The PIRLS algorithm for $\tilde{u}$ and $\tilde{\beta}$

The penalized, iteratively reweighted, least squares (PIRLS) algorithm to determine $\tilde{\boldsymbol{u}}(\boldsymbol{\theta})$ and $\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$ is a form of the Fisher scoring algorithm. We fix the weights matrix, $\boldsymbol{W}$, and use penalized, weighted, nonlinear least squares to minimize the penalized, weighted residual sum of squares conditional on these weights. Then we update the weights to those determined by the current value of $\boldsymbol{\mu}$ and iterate.

To describe this algorithm in more detail we will use parenthesized superscripts to denote the iteration number. Thus $\boldsymbol{u}^{(0)}$ and $\boldsymbol{\beta}^{(0)}$ are the initial values of these parameters, while $\boldsymbol{u}^{(i)}$ and $\boldsymbol{\beta}^{(i)}$ are the values at the ith iteration. Similarly $\boldsymbol{\gamma}^{(i)} = \boldsymbol{Z}\boldsymbol{\Lambda}(\boldsymbol{\theta})\boldsymbol{u}^{(i)} + \boldsymbol{X}\boldsymbol{\beta}^{(i)}$, $\boldsymbol{\eta}^{(i)} = \boldsymbol{m}(\boldsymbol{\gamma}^{(i)})$ and $\boldsymbol{\mu}^{(i)} = \boldsymbol{g}^{-1}(\boldsymbol{\eta}^{(i)})$.

We use a penalized version of the Gauss-Newton algorithm (Bates and Watts, 1988, ch. 2) for which we define the weighted Jacobian matrices

$$\boldsymbol{U}^{(i)} = \boldsymbol{W}^{1/2} \left.\frac{d\boldsymbol{\mu}}{d\boldsymbol{u}'}\right|_{\boldsymbol{u}=\boldsymbol{u}^{(i)},\boldsymbol{\beta}=\boldsymbol{\beta}^{(i)}} = \boldsymbol{W}^{1/2} \left.\frac{d\boldsymbol{\mu}}{d\boldsymbol{\eta}'}\right|_{\boldsymbol{\eta}^{(i)}} \left.\frac{d\boldsymbol{\eta}}{d\boldsymbol{\gamma}'}\right|_{\boldsymbol{\gamma}^{(i)}} \boldsymbol{Z}\boldsymbol{\Lambda}(\boldsymbol{\theta}) \tag{36}$$

$$\boldsymbol{V}^{(i)} = \boldsymbol{W}^{1/2} \left.\frac{d\boldsymbol{\mu}}{d\boldsymbol{\beta}'}\right|_{\boldsymbol{u}=\boldsymbol{u}^{(i)},\boldsymbol{\beta}=\boldsymbol{\beta}^{(i)}} = \boldsymbol{W}^{1/2} \left.\frac{d\boldsymbol{\mu}}{d\boldsymbol{\eta}'}\right|_{\boldsymbol{\eta}^{(i)}} \left.\frac{d\boldsymbol{\eta}}{d\boldsymbol{\gamma}'}\right|_{\boldsymbol{\gamma}^{(i)}} \boldsymbol{X} \tag{37}$$

of dimension $n \times q$ and $n \times p$, respectively. The increments at the ith iteration, $\boldsymbol{\delta}_{\boldsymbol{u}}^{(i)}$ and $\boldsymbol{\delta}_{\boldsymbol{\beta}}^{(i)}$, are the solutions to

$$\begin{bmatrix} \boldsymbol{U}^{(i)\prime}\boldsymbol{U}^{(i)} + \boldsymbol{I}_q & \boldsymbol{U}^{(i)\prime}\boldsymbol{V}^{(i)} \\ \boldsymbol{V}^{(i)\prime}\boldsymbol{U}^{(i)} & \boldsymbol{V}^{(i)\prime}\boldsymbol{V}^{(i)} \end{bmatrix} \begin{bmatrix} \boldsymbol{\delta}_{\boldsymbol{u}}^{(i)} \\ \boldsymbol{\delta}_{\boldsymbol{\beta}}^{(i)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{U}^{(i)\prime}\boldsymbol{W}^{1/2}(\boldsymbol{y} - \boldsymbol{\mu}^{(i)}) - \boldsymbol{u}^{(i)} \\ \boldsymbol{U}^{(i)\prime}\boldsymbol{W}^{1/2}(\boldsymbol{y} - \boldsymbol{\mu}^{(i)}) \end{bmatrix} \tag{38}$$

providing the updated parameter values

$$\begin{bmatrix} \boldsymbol{u}^{(i+1)} \\ \boldsymbol{\beta}^{(i+1)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{u}^{(i)} \\ \boldsymbol{\beta}^{(i)} \end{bmatrix} + \lambda \begin{bmatrix} \boldsymbol{\delta}_{\boldsymbol{u}}^{(i)} \\ \boldsymbol{\delta}_{\boldsymbol{\beta}}^{(i)} \end{bmatrix} \tag{39}$$

where $\lambda > 0$ is a step factor chosen to ensure that

$$(\boldsymbol{y}-\boldsymbol{\mu}^{(i+1)})'\boldsymbol{W}(\boldsymbol{y}-\boldsymbol{\mu}^{(i+1)})+\|\boldsymbol{u}^{(i+1)}\|^2 < (\boldsymbol{y}-\boldsymbol{\mu}^{(i)})'\boldsymbol{W}(\boldsymbol{y}-\boldsymbol{\mu}^{(i)})+\|\boldsymbol{u}^{(i)}\|^2. \tag{40}$$

In the process of solving for the increments we form the sparse, lower triangular, Cholesky factor, $\boldsymbol{L}^{(i)}$, satisfying

$$\boldsymbol{L}^{(i)}\boldsymbol{L}^{(i)'} = \boldsymbol{P}_{\boldsymbol{Z}} \left(\boldsymbol{U}^{(i)'}\boldsymbol{U}^{(i)} + \boldsymbol{I}_n \right) \boldsymbol{P}_{\boldsymbol{Z}}'. \tag{41}$$

After each successful iteration, determining new values of the coefficients, $\boldsymbol{u}^{(i+1)}$ and $\boldsymbol{\beta}^{(i+1)}$, that reduce the penalized, weighted residual sum of squqres, we update the weights matrix to $\boldsymbol{W}(\boldsymbol{\mu}^{(i+1)})$ and the weighted Jacobians, $\boldsymbol{U}^{(i+1)}$ and $\boldsymbol{V}^{(i+1)}$, then iterate. Convergence is determined according to the orthogonality convergence criterion (Bates and Watts, 1988, ch. 2), suitably adjusted for the weights matrix and the penalty.

4.3 Weighted linear mixed models

One of the simplest generalizations of linear mixed models is a weighted linear mixed model where $s = 1$, the link function, g, and the nonlinear model function, m, are both the identity, the weights matrix, $\boldsymbol{W}$, is constant and the conditional distribution family is Gaussian. That is, the conditional distribution can be written

$$(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{U}} = \boldsymbol{u}) \sim \mathcal{N}(\boldsymbol{\gamma}(\boldsymbol{u}, \boldsymbol{\theta}, \boldsymbol{\beta}), \sigma^2\boldsymbol{W}^{-1}) \tag{42}$$

with discrepancy function

$$d(\boldsymbol{u}|\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\beta}) = \left\|\boldsymbol{W}^{1/2}(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\Lambda}(\boldsymbol{\theta})\boldsymbol{u} - \boldsymbol{X}\boldsymbol{\theta})\right\|^2 + \|\boldsymbol{u}\|^2. \tag{43}$$

The conditional mode, $\tilde{\boldsymbol{u}}(\boldsymbol{\theta})$, and the conditional estimate, $\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$, are the solutions to

$$\begin{bmatrix} \boldsymbol{\Lambda}'(\boldsymbol{\theta})\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z}\boldsymbol{\Lambda}(\theta) + \boldsymbol{I}_q & \boldsymbol{\Lambda}'(\boldsymbol{\theta})\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{X} \\ \boldsymbol{X}'\boldsymbol{W}\boldsymbol{Z}\boldsymbol{\Lambda}(\theta) & \boldsymbol{X}'\boldsymbol{W}\boldsymbol{X} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{u}}(\boldsymbol{\theta}) \\ \tilde{\boldsymbol{\beta}}(\boldsymbol{\theta}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Lambda}'(\boldsymbol{\theta})\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{y} \\ \boldsymbol{X}'\boldsymbol{W}\boldsymbol{y} \end{bmatrix}, \tag{44}$$

which can be solved directly, and the Cholesky factor, $\boldsymbol{L_Z}(\boldsymbol{\theta})$, satisfies

$$\boldsymbol{L_Z}(\boldsymbol{\theta})\boldsymbol{L_Z}(\boldsymbol{\theta})' = \boldsymbol{P_Z}\left(\boldsymbol{\Lambda}'(\boldsymbol{\theta})\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z}\boldsymbol{\Lambda}(\theta) + \boldsymbol{I}_q\right)\boldsymbol{P}_{\boldsymbol{Z}}'. \tag{45}$$

The profiled log-likelihood, $\tilde{\ell}(\boldsymbol{\theta}|\boldsymbol{y})$, is

$$-2\tilde{\ell}(\boldsymbol{\theta}|\boldsymbol{y}) = \log\left(\frac{|\boldsymbol{L_Z}(\boldsymbol{\theta})|^2}{|\boldsymbol{W}|}\right) + n\left(1 + \log\left(\frac{2\pi\tilde{d}(\boldsymbol{y},\boldsymbol{\theta})}{n}\right)\right). \tag{46}$$

If the matrix $\boldsymbol{W}$ is fixed then we can ignore the term $|\boldsymbol{W}|$ in (46) when determining the MLE, $\widehat{\boldsymbol{\theta}}_L$. However, in some models, we use a parameterized weight matrix, $\boldsymbol{W}(\boldsymbol{\phi})$, and wish to determine the MLEs, $\widehat{\boldsymbol{\phi}}_L$ and $\widehat{\boldsymbol{\theta}}_L$ simultaneously. In these cases we must include the term involving $|\boldsymbol{W}(\boldsymbol{\phi})|$ when evaluating the profiled log-likelihood.

Note that we must define the parameterization of $\boldsymbol{W}(\boldsymbol{\phi})$ such that σ^2 and $\boldsymbol{\phi}$ are not a redundant parameterization of $\sigma^2\boldsymbol{W}(\boldsymbol{\phi})$. For example, we could require that the first diagonal element of $\boldsymbol{W}$ be unity.

4.4 Nonlinear mixed models

In an unweighted, nonlinear mixed model the conditional distribution is Gaussian, the link, g, is the identity and the weights matrix, $\boldsymbol{W} = \boldsymbol{I}_n$. That is,

$$(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{U}} = \boldsymbol{u}) \sim \mathcal{N}(\boldsymbol{m}(\boldsymbol{\gamma}), \sigma^2\boldsymbol{I}_n) \tag{47}$$

with discrepancy function

$$d(\boldsymbol{u}|\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{\mu}\|^2 + \|\boldsymbol{u}\|^2. \tag{48}$$

For a given value of $\boldsymbol{\theta}$ we determine the conditional modes, $\tilde{\boldsymbol{u}}(\boldsymbol{\theta})$ and $\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$, as the solution to the penalized nonlinear least squares problem

$$\begin{bmatrix}\tilde{\boldsymbol{u}}(\boldsymbol{\theta})\\ \tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})\end{bmatrix} = \arg\min_{\boldsymbol{u},\boldsymbol{\theta}} d(\boldsymbol{u}|\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\beta}) \tag{49}$$

and we write the minimum discrepancy, given $\boldsymbol{y}$ and $\boldsymbol{\theta}$, as

$$\tilde{d}(\boldsymbol{y},\boldsymbol{\theta}) = d(\tilde{\boldsymbol{u}}(\boldsymbol{\theta})|\boldsymbol{y},\boldsymbol{\theta},\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})). \tag{50}$$

Let $\tilde{\boldsymbol{L}}_Z(\boldsymbol{\theta})$ and $\tilde{\boldsymbol{L}}_X(\boldsymbol{\theta})$ be the Cholesky factors at $\boldsymbol{\theta}$, $\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$ and $\tilde{\boldsymbol{u}}(\boldsymbol{\theta})$. Then the *Laplace approximation* to the log-likelihood is

$$-2\ell_P(\boldsymbol{\theta},\boldsymbol{\beta},\sigma|\boldsymbol{y}) \approx n\log(2\pi\sigma^2)+\log(|\tilde{\boldsymbol{L}}_{\boldsymbol{Z}}|^2)+\frac{\tilde{d}(\boldsymbol{y},\boldsymbol{\theta})+\left\|\tilde{\boldsymbol{L}}'_{\boldsymbol{X}}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})\right\|^2}{\sigma^2}, \quad (51)$$

producing the approximate profiled log-likelihood, $\tilde{\ell}_P(\boldsymbol{\theta}|\boldsymbol{y})$,

$$-2\tilde{\ell}_P(\boldsymbol{\theta}|\boldsymbol{y}) \approx \log(|\tilde{\boldsymbol{L}}_{\boldsymbol{Z}}|^2) + n\left(1+\log(2\pi\tilde{d}(\boldsymbol{y},\boldsymbol{\theta})/n)\right). \quad (52)$$

4.4.1 Nonlinear mixed model summary

In a nonlinear mixed model we determine the parameter estimate, $\widehat{\boldsymbol{\theta}}_P$, from the Laplace approximation to the log-likelihood as

$$\widehat{\boldsymbol{\theta}}_P = \arg\max_{\boldsymbol{\theta}} \tilde{\ell}_P(\boldsymbol{\theta}|\boldsymbol{y}) = \arg\min_{\boldsymbol{\theta}} \log(|\tilde{\boldsymbol{L}}_{\boldsymbol{Z}}|^2) + n\left(1+\log(2\pi\tilde{d}(\boldsymbol{y},\boldsymbol{\theta})/n)\right). \quad (53)$$

Each evaluation of $\tilde{\ell}_P(\boldsymbol{\theta}|\boldsymbol{y})$ requires a solving the penalized nonlinear least squares problem (49) simultaneously with respect to both sets of coefficients, $\boldsymbol{u}$ and $\boldsymbol{\beta}$, in the linear predictor, $\boldsymbol{\gamma}$.

For a weighted nonlinear mixed model with fixed weights, $\boldsymbol{W}$, we replace the unweighted discrepancy function $d(\boldsymbol{u}|\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\beta})$ with the weighted discrepancy function,

5 Details of the implementation

5.1 Implementation details for linear mixed models

The crucial step in implementing algorithms for determining ML or REML estimates of the parameters in a linear mixed model is evaluating the factorization (15) for any $\boldsymbol{\theta}$ satisfying $\boldsymbol{\theta}_L \le \boldsymbol{\theta} \le \boldsymbol{\theta}_U$. We will assume that $\boldsymbol{Z}$ is sparse as is $\boldsymbol{Z\Lambda}(\boldsymbol{\theta})$.

When $\boldsymbol{X}$ is not sparse we will use the factorization (15) setting $\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{I}_p$ and storing $\boldsymbol{L}_{\boldsymbol{XZ}}$ and $\boldsymbol{L}_{\boldsymbol{X}}$ as dense matrices. The permutation matrix $\boldsymbol{P}_{\boldsymbol{Z}}$ is determined from the pattern of non-zeros in $\boldsymbol{Z\Lambda}(\boldsymbol{\theta})$ which is does not depend on $\boldsymbol{\theta}$, as long as $\boldsymbol{\theta}$ is not on the boundary. In fact, in most cases the pattern of non-zeros in $\boldsymbol{Z\Lambda}(\boldsymbol{\theta})$ is the same as the pattern of non-zeros in $\boldsymbol{Z}$. For many

models, in particular models with scalar random effects (described later), the matrix $\boldsymbol{\Lambda}(\boldsymbol{\theta})$ is diagonal.

Given a value of $\boldsymbol{\theta}$ we determine the Cholesky factor $\boldsymbol{L_Z}$ satisfying

$$\boldsymbol{L_Z L_Z'} = \boldsymbol{P_Z}(\boldsymbol{\Lambda}'(\boldsymbol{\theta})\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Lambda}(\theta) + \boldsymbol{I}_q)\boldsymbol{P_Z'}. \tag{54}$$

The CHOLMOD package allows for $\boldsymbol{L_Z}$ to be calculated directly from $\boldsymbol{\Lambda}'(\boldsymbol{\theta})\boldsymbol{Z}'$ or from $\boldsymbol{\Lambda}'(\boldsymbol{\theta})\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Lambda}(\theta)$. The choice in implementation is whether to store $\boldsymbol{Z}'$ and update it to $\boldsymbol{\Lambda}'(\boldsymbol{\theta})\boldsymbol{Z}$ or to store $\boldsymbol{Z}'\boldsymbol{Z}$ and use it to form $\boldsymbol{\Lambda}'(\boldsymbol{\theta})\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Lambda}(\theta)$ at each evaluation.

In the lme4 package we store $\boldsymbol{Z}'$ and use it to form $\boldsymbol{\Lambda}'(\boldsymbol{\theta})\boldsymbol{Z}'$ from which $\boldsymbol{L_Z}$ is evaluated. There are two reasons for this choice. First, the calculations for the more general forms of mixed models cannot be reduced to calculations involving $\boldsymbol{Z}'\boldsymbol{Z}$ and by expressing these calculations in terms of $\boldsymbol{\Lambda}(\boldsymbol{\theta})\boldsymbol{Z}'$ for linear mixed models we can reuse the code for the more general models. Second, the calculation of $\boldsymbol{\Lambda}(\boldsymbol{\theta})'\left(\boldsymbol{Z}'\boldsymbol{Z}\right)\boldsymbol{\Lambda}(\boldsymbol{\theta})$ from $\boldsymbol{Z}'\boldsymbol{Z}$ is complicated compared to the calculation of $\boldsymbol{\Lambda}(\boldsymbol{\theta})'\boldsymbol{Z}'$ from $\boldsymbol{Z}'$.

This choice is disadvantageous when $n \gg q$ because $\boldsymbol{Z}'$ is much larger than $\boldsymbol{Z}'\boldsymbol{Z}$, even when they are stored as sparse matrices. Evaluation of $\boldsymbol{L_Z}$ directly from $\boldsymbol{Z}'$ requires more storage and more calculation that evaluating $\boldsymbol{L_Z}$ from $\boldsymbol{Z}'\boldsymbol{Z}$.

Next we evaluate $\boldsymbol{L'_{XZ}}$ as the solution to

$$\boldsymbol{L_Z L'_{XZ}} = \boldsymbol{P_Z}\boldsymbol{\Lambda}'(\boldsymbol{\theta})\boldsymbol{Z}'\boldsymbol{X}. \tag{55}$$

Again we have the choice of calculating and storing $\boldsymbol{Z}'\boldsymbol{X}$ or storing $\boldsymbol{X}$ and using it to reevaluate $\boldsymbol{Z}'\boldsymbol{X}$. In the lme4 package we store $\boldsymbol{X}$, because the calculations for the more general models cannot be expressed in terms of $\boldsymbol{Z}'\boldsymbol{X}$.

Finally $\boldsymbol{L_X}$ is evaluated as the (dense) solution to

$$\boldsymbol{L_X L'_X} = \boldsymbol{X}'\boldsymbol{X} - \boldsymbol{L_{XZ}}\boldsymbol{L_{XZ}}. \tag{56}$$

from which $\tilde{\boldsymbol{\beta}}$ can be determined as the solution to dense system

$$\boldsymbol{L_X L_X}\tilde{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y} \tag{57}$$

and $\tilde{\boldsymbol{u}}$ as the solution to the sparse system

$$\boldsymbol{L_Z L_Z}\tilde{u} = \boldsymbol{\Lambda}'\boldsymbol{Z}'\boldsymbol{y} \tag{58}$$

For many models, in particular models with scalar random effects, which are described later, the matrix $\boldsymbol{\Lambda}(\boldsymbol{\theta})$ is diagonal. For such a model, if both $\boldsymbol{Z}$ and $\boldsymbol{X}$ are sparse and we plan to use the REML criterion then we create and store

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{Z}'\boldsymbol{Z} & \boldsymbol{Z}'\boldsymbol{X} \\ \boldsymbol{X}'\boldsymbol{Z} & \boldsymbol{X}'\boldsymbol{X} \end{bmatrix} \quad \text{and} \quad \boldsymbol{c} = \begin{bmatrix} \boldsymbol{Z}'\boldsymbol{y} \\ \boldsymbol{X}'\boldsymbol{y} \end{bmatrix} \tag{59}$$

and determine a fill-reducing permutation, $\boldsymbol{P}$, for $\boldsymbol{A}$. Given a value of $\boldsymbol{\theta}$ we create the factorization

$$\boldsymbol{L}(\boldsymbol{\theta})\boldsymbol{L}(\boldsymbol{\theta})' = \boldsymbol{P}\left(\begin{bmatrix} \boldsymbol{\Lambda}(\boldsymbol{\theta}) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_p \end{bmatrix} \boldsymbol{A} \begin{bmatrix} \boldsymbol{\Lambda}(\boldsymbol{\theta}) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_p \end{bmatrix} + \begin{bmatrix} \boldsymbol{I}_q & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}\right)\boldsymbol{P}' \tag{60}$$

solve for $\tilde{\boldsymbol{u}}(\boldsymbol{\theta})$ and $\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$ in

$$\boldsymbol{L}\boldsymbol{L}'\boldsymbol{P}\begin{bmatrix} \tilde{\boldsymbol{u}}(\boldsymbol{\theta}) \\ \tilde{\boldsymbol{\beta}}(\boldsymbol{\theta}) \end{bmatrix} = \boldsymbol{P}\begin{bmatrix} \boldsymbol{\Lambda}(\boldsymbol{\theta}) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_p \end{bmatrix}\boldsymbol{c} \tag{61}$$

then evaluate $\tilde{d}(\boldsymbol{y}|\boldsymbol{\theta})$ and the profiled REML criterion as

$$\tilde{d}_R(\boldsymbol{\theta}|\boldsymbol{y}) = \log(|\boldsymbol{L}(\boldsymbol{\theta})|^2) + (n-p)\left(1 + \log\left(\frac{2\pi\tilde{d}(\boldsymbol{y}|\boldsymbol{\theta})}{n-p}\right)\right). \tag{62}$$

References

Douglas M. Bates and Donald G. Watts. *Nonlinear Regression Analysis and Its Applications*. Wiley, 1988.

Tim Davis. CHOLMOD: sparse supernodal Cholesky factorization and update/downdate. http://www.cise.ufl.edu/research/sparse/cholmod, 2005.

Timothy A. Davis. *Direct Methods for Sparse Linear Systems*. Fundamentals of Algorithms. SIAM, 2006.

A Notation

A.1 Random variables in the model

$\boldsymbol{\mathcal{B}}$ Random-effects vector of dimension q, $\boldsymbol{\mathcal{B}} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{V}(\boldsymbol{\theta})\boldsymbol{V}(\boldsymbol{\theta})')$.

$\boldsymbol{U}$ "Spherical" random-effects vector of dimension q, $\boldsymbol{\mathcal{U}} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_q)$, $\boldsymbol{\mathcal{B}} = \boldsymbol{V}(\boldsymbol{\theta})\boldsymbol{\mathcal{U}}$.

$\boldsymbol{\mathcal{Y}}$ Response vector of dimension n.

A.2 Parameters of the model

$\boldsymbol{\beta}$ Fixed-effects parameters (dimension p).

$\boldsymbol{\theta}$ Parameters determining the left factor, $\boldsymbol{\Lambda}(\boldsymbol{\theta})$ of the relative covariance matrix of $\boldsymbol{\mathcal{B}}$ (dimension m).

σ the common scale parameter - not used in some generalized linear mixed models and generalized nonlinear mixed models.

A.3 Dimensions

m dimension of the parameter $\boldsymbol{\theta}$.

n dimension of the response vector, $\boldsymbol{y}$, and the random variable, $\boldsymbol{\mathcal{Y}}$.

p dimension of the fixed-effects parameter, $\boldsymbol{\beta}$.

q dimension of the random effects, $\boldsymbol{\mathcal{B}}$ or $\boldsymbol{\mathcal{U}}$.

s dimension of the parameter vector, $\boldsymbol{\phi}$, in the nonlinear model function.

A.4 Matrices

$\boldsymbol{L}$ Left Cholesky factor of a positive-definite symmetric matrix. $\boldsymbol{L_Z}$ is $q \times q$; $\boldsymbol{L_X}$ is $p \times p$.

$\boldsymbol{P}$ Fill-reducing permutation for the random effects model matrix. (Size $q \times q$.)

$\boldsymbol{V}$ Left factor of the relative covariance matrix of the random effects. (Size $q \times q$.)

$\boldsymbol{X}$ Model matrix for the fixed-effects parameters, $\boldsymbol{\beta}$. (Size $(ns) \times p$.)

$\boldsymbol{Z}$ Model matrix for the random effects. (Size $(ns) \times q$.)

B Integrating a quadratic deviance expression

In (6) we defined the likelihood of the parameters given the observed data as

$$L(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma|\boldsymbol{y}) = \int_{\mathbb{R}^q} h(\boldsymbol{u}|\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma)\, d\boldsymbol{u}.$$

which is often alarmingly described as "an intractable integral". In point of fact, this integral can be evaluated exactly in the case of a linear mixed model and can be approximated quite accurately for other forms of mixed models.