Computational methods for mixed models

Douglas Bates
Department of Statistics
University of Wisconsin – Madison

September 9, 2008

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 maxmimum 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 a particular form of mixed model, called a linear mixed model, and also the general form of the mixed models that can be represented in the lme4 package.

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 the following conditions:

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. Frequently, the number of rows in Z and X is n, the dimension of y, but not always.

2. The scalar distributions, $(\mathcal{Y}_i|\boldsymbol{\mathcal{B}} = \boldsymbol{b}), i = 1, \ldots, n$, are independent. We say that the components of $\boldsymbol{\mathcal{Y}}$ are *conditionally independent* given $\boldsymbol{\mathcal{B}}$. Furthermore, the scalar distributions must all have the same form. They differ only in the value of the conditional means, $\mu_i(\boldsymbol{b})$.

3. The scalar distributions, $(\mathcal{Y}_i|\boldsymbol{\mathcal{B}} = \boldsymbol{b}), i = 1, \ldots, n$, are completely determined by the conditional mean, $\boldsymbol{\mu}_{\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{B}}}(\boldsymbol{b})$ and, at most, one additional parameter, σ, which is the common scale parameter.

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 require that we 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, defined by the 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 estimates* of these parameters are the values that maximize L.

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, $(\mathcal{Y}|\mathcal{U} = \boldsymbol{u})$, and the unconditional distribution, $\mathcal{U}$, are spherical Gaussian distributions and that the conditional mean, $\boldsymbol{\mu}_{\mathcal{Y}|\mathcal{U}}(\boldsymbol{u})$, is the linear predictor, $\boldsymbol{\gamma}(\boldsymbol{u})$. Because all the distributions determining the model are continuous distributions, we consider their densities. It is convenient to express these densities on the scale of the logarithm of the density or, as

is common in many statistical expressions, on the *deviance scale* (negative twice the logarithm of the density),

$$\begin{aligned}
-2\log(f_{\mathcal{U}}(\boldsymbol{u})) &= q\log(2\pi\sigma^2) + \frac{\|\boldsymbol{u}\|^2}{\sigma^2} \\
-2\log(f_{\mathcal{Y}|\mathcal{U}}(\boldsymbol{y}|\boldsymbol{u})) &= n\log(2\pi\sigma^2) + \frac{\|\boldsymbol{y}-\boldsymbol{Z}\boldsymbol{\Lambda}(\boldsymbol{\theta})\boldsymbol{u}-\boldsymbol{X}\boldsymbol{\beta}\|^2}{\sigma^2} \\
-2\log(h(\boldsymbol{u}|\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\beta},\sigma)) &= (n+q)\log(2\pi\sigma^2) + \frac{\|\boldsymbol{y}-\boldsymbol{Z}\boldsymbol{\Lambda}(\boldsymbol{\theta})\boldsymbol{u}-\boldsymbol{X}\boldsymbol{\beta}\|^2+\|\boldsymbol{u}\|^2}{\sigma^2} \\
&= (n+q)\log(2\pi\sigma^2) + \frac{d(\boldsymbol{u}|\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\beta})}{\sigma^2}
\end{aligned} \tag{7}$$

In (7) the *discrepancy* function,

$$d(\boldsymbol{u}|\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\beta}) = \|\boldsymbol{y}-\boldsymbol{Z}\boldsymbol{\Lambda}(\boldsymbol{\theta})\boldsymbol{u}-\boldsymbol{X}\boldsymbol{\beta}\|^2 + \|\boldsymbol{u}\|^2 \tag{8}$$

has the form of a penalized residual sum of squares in that the first term, $\|\boldsymbol{y}-\boldsymbol{Z}\boldsymbol{\Lambda}(\boldsymbol{\theta})\boldsymbol{u}-\boldsymbol{X}\boldsymbol{\beta}\|^2$ 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 $\boldsymbol{\beta}$

$$d(\boldsymbol{u}|\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\beta}) = \left\| \begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{0}\end{bmatrix} - \begin{bmatrix}\boldsymbol{Z}\boldsymbol{\Lambda}(\boldsymbol{\theta}) & \boldsymbol{X}\\ \boldsymbol{I}_q & \boldsymbol{0}\end{bmatrix} \begin{bmatrix}\boldsymbol{u}\\ \boldsymbol{\beta}\end{bmatrix} \right\|^2. \tag{9}$$

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, $\boldsymbol{Z}\boldsymbol{\Lambda}(\boldsymbol{\theta})\boldsymbol{u} + \boldsymbol{X}\boldsymbol{\beta}$, 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 (9) 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

$$\begin{bmatrix}\boldsymbol{\Lambda}'(\boldsymbol{\theta})\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Lambda}(\theta)+\boldsymbol{I}_q & \boldsymbol{\Lambda}'(\boldsymbol{\theta})\boldsymbol{Z}\boldsymbol{X}\\ \boldsymbol{X}'\boldsymbol{Z}\boldsymbol{\Lambda}(\theta) & \boldsymbol{X}'\boldsymbol{X}\end{bmatrix} \begin{bmatrix}\tilde{\boldsymbol{u}}(\boldsymbol{\theta})\\ \tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})\end{bmatrix} = \begin{bmatrix}\boldsymbol{\Lambda}'(\boldsymbol{\theta})\boldsymbol{Z}'\boldsymbol{y}\\ \boldsymbol{X}\boldsymbol{y}\end{bmatrix} \tag{10}$$

An effective way of determining the solution to a sparse, symmetric, positive definite system of equations such as (10) is the sparse Cholesky decomposition (Davis, 2006). If $\boldsymbol{A}$ is a sparse, symmetric positive definite matrix then the sparse Cholesky decomposition with fill-reducing permutation $\boldsymbol{P}$ is the lower-triangular matrix $\boldsymbol{L}$ such that

$$\boldsymbol{L}\boldsymbol{L}' = \boldsymbol{P}\boldsymbol{A}\boldsymbol{P}' \tag{11}$$

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

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{P_Z} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{P_X} \end{bmatrix} \tag{12}$$

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 (12) so that we can write the Cholesky factorization for the positive definite system (10) as

$$\begin{bmatrix} \boldsymbol{L_Z} & \boldsymbol{0} \\ \boldsymbol{L_{XZ}} & \boldsymbol{L_X} \end{bmatrix} \begin{bmatrix} \boldsymbol{L_Z} & \boldsymbol{0} \\ \boldsymbol{L_{XZ}} & \boldsymbol{L_X} \end{bmatrix}' = \begin{bmatrix} \boldsymbol{P_Z} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{P_X} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Lambda}'(\boldsymbol{\theta})\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Lambda}(\theta) + \boldsymbol{I}_q & \boldsymbol{\Lambda}'(\boldsymbol{\theta})\boldsymbol{Z}\boldsymbol{X} \\ \boldsymbol{X}'\boldsymbol{Z}\boldsymbol{\Lambda}(\theta) & \boldsymbol{X}'\boldsymbol{X} \end{bmatrix} \begin{bmatrix} \boldsymbol{P_Z} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{P_X} \end{bmatrix}' \tag{13}$$

The diagonal elements of $\boldsymbol{L_Z}$ and $\boldsymbol{L_X}$ are determined only up to a change in sign. By convention we choose them to be positive.

The discrepancy can now be written in the canonical form

$$d(\boldsymbol{u}|\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\beta}) = \tilde{d}(\boldsymbol{y}, \boldsymbol{\theta}) + \left\| \begin{bmatrix} \boldsymbol{L}'_{\boldsymbol{Z}} & \boldsymbol{L}'_{\boldsymbol{XZ}} \\ \boldsymbol{0} & \boldsymbol{L}'_{\boldsymbol{X}} \end{bmatrix} \begin{bmatrix} \boldsymbol{P_Z}(\boldsymbol{u} - \tilde{\boldsymbol{u}}) \\ \boldsymbol{P_X}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) \end{bmatrix} \right\|^2 \tag{14}$$

where

$$\tilde{d}(\boldsymbol{y}, \boldsymbol{\theta}) = d(\tilde{\boldsymbol{u}}(\boldsymbol{\theta})|\boldsymbol{y}, \boldsymbol{\theta}, \tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})) \tag{15}$$

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

3.2 Profiled log-likelihood for linear mixed models

Substituting (14) into (7) provides the unnormalized conditional density $h(\boldsymbol{u}|\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\beta},\sigma)$ on the deviance scale as

$$\begin{aligned}&-2\log(h(\boldsymbol{u}|\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\beta},\sigma))\\&\quad=(n+q)\log(2\pi\sigma^2)+\frac{\tilde{d}(\boldsymbol{y},\boldsymbol{\theta})+\left\|\begin{bmatrix}\boldsymbol{L}_{\boldsymbol{Z}}' & \boldsymbol{L}_{\boldsymbol{XZ}}'\\ \boldsymbol{0} & \boldsymbol{L}_{\boldsymbol{X}}'\end{bmatrix}\begin{bmatrix}\boldsymbol{P}_{\boldsymbol{Z}}(\boldsymbol{u}-\tilde{\boldsymbol{u}})\\ \boldsymbol{P}_{\boldsymbol{X}}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})\end{bmatrix}\right\|^2}{\sigma^2}.\end{aligned}\tag{16}$$

We need to integrate h with respect to $\boldsymbol{u}$ and minimize the resulting function with respect to the parameter values. We first minimize h with respect to $\boldsymbol{\beta}$ by setting $\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$. from which we can evaluate the *profiled likelihood*, a function of $\boldsymbol{\theta}$ only. On the deviance scale the profiled likelihood is

$$-2\log(\tilde{L}(\boldsymbol{\theta}|\boldsymbol{y})=\log(|\boldsymbol{L}_{\boldsymbol{Z}}(\boldsymbol{\theta})|^2)+n\left(1+\log\left(\frac{2\pi\tilde{d}(\boldsymbol{y},\boldsymbol{\theta})}{n}\right)\right)\tag{17}$$

and the maximum likelihood estimate of $\boldsymbol{\theta}$ is

$$\widehat{\boldsymbol{\theta}}_L=\arg\min_{\boldsymbol{\theta}}\left(-2\log(\tilde{L}(\boldsymbol{\theta}|\boldsymbol{y}))\right).\tag{18}$$

The maximum likelihood estimates of σ^2 and $\boldsymbol{\beta}$ are then evaluated as

$$\widehat{\sigma_L^2}=\frac{\tilde{d}(\widehat{\boldsymbol{\theta}}_L,\boldsymbol{y})}{n}\tag{19}$$

$$\widehat{\boldsymbol{\beta}}_L=\tilde{\boldsymbol{\beta}}(\widehat{\boldsymbol{\theta}}_L).\tag{20}$$

3.3 The REML criterion

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

$$\boldsymbol{\mathcal{Y}}\sim\mathcal{N}(\boldsymbol{X}\boldsymbol{\beta},\sigma^2\boldsymbol{I}_n),\tag{21}$$

in which we typically estimate σ^2 by

$$\widehat{\sigma^2_R} = \frac{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2}{n-p} \tag{22}$$

even though the maximum likelihood estimate of σ^2 is

$$\widehat{\sigma^2_L} = \frac{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2}{n}. \tag{23}$$

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 (22) for a linear regression model. Although not usually derived in this way, the REML criterion can be expressed as

$$c_R(\boldsymbol{\theta}, \boldsymbol{\sigma}|\boldsymbol{y}) = -2\log \int_{\mathbb{R}^p} L(\boldsymbol{u}|\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma)\, d\boldsymbol{\beta} \tag{24}$$

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

$$\tilde{c}_R(\boldsymbol{\theta}|\boldsymbol{y}) = \log(|\boldsymbol{L}_{\boldsymbol{Z}}(\boldsymbol{\theta})|^2) + \log(|\boldsymbol{L}_{\boldsymbol{X}}(\boldsymbol{\theta})|^2) + (n-p)\left(1 + \log\left(\frac{2\pi\tilde{d}(\boldsymbol{\theta}|\boldsymbol{y})}{n-p}\right)\right) \tag{25}$$

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

$$\widehat{\boldsymbol{\theta}}_R = \arg\min_{\boldsymbol{\theta}} \tilde{c}_R(\boldsymbol{\theta}, \boldsymbol{y}). \tag{26}$$

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 (24), 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 (13) 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 (12) 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 Implementation details for linear mixed models

We determine the maximum likelihood (ML) or REML estimates of the parameters $\boldsymbol{\theta}$, $\boldsymbol{\beta}$ and σ in a linear mixed model by first minimizing the profiled log-likelihood or the profiled REML criterion with respect to $\boldsymbol{\theta}$. Thus the first implementation question is evaluating the profiled log-likelihood or the profiled REML criterion.

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

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{Z}'\boldsymbol{Z} & \boldsymbol{Z}'\boldsymbol{X} \\ \boldsymbol{X}'\boldsymbol{Z} & \boldsymbol{X}'\boldsymbol{X} \end{bmatrix} \quad \text{and} \quad \boldsymbol{c} = \begin{bmatrix} \boldsymbol{Z}'\boldsymbol{y} \\ \boldsymbol{X}'\boldsymbol{y} \end{bmatrix} \tag{27}$$

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

$$\boldsymbol{L}(\boldsymbol{\theta})\boldsymbol{L}(\boldsymbol{\theta})' = \boldsymbol{P}\left(\begin{bmatrix} \boldsymbol{\Lambda}(\boldsymbol{\theta}) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_p \end{bmatrix} \boldsymbol{A} \begin{bmatrix} \boldsymbol{\Lambda}(\boldsymbol{\theta}) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_p \end{bmatrix} + \begin{bmatrix} \boldsymbol{I}_q & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}\right)\boldsymbol{P}' \tag{28}$$

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

$$\boldsymbol{L}\boldsymbol{L}'\boldsymbol{P}\begin{bmatrix} \tilde{\boldsymbol{u}}(\boldsymbol{\theta}) \\ \tilde{\boldsymbol{\beta}}(\boldsymbol{\theta}) \end{bmatrix} = \boldsymbol{P}\begin{bmatrix} \boldsymbol{\Lambda}(\boldsymbol{\theta}) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_p \end{bmatrix}\boldsymbol{c} \tag{29}$$

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

$$\tilde{d}_R(\boldsymbol{\theta}|\boldsymbol{y}) = \log(|\boldsymbol{L}(\boldsymbol{\theta})|^2) + (n-p)\left(1 + \log\left(\frac{2\pi\tilde{d}(\boldsymbol{y}|\boldsymbol{\theta})}{n-p}\right)\right). \tag{30}$$

References

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, $\boldsymbol{\mathcal{B}} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{V}(\boldsymbol{\theta})\boldsymbol{V}(\boldsymbol{\theta})')$.

$\boldsymbol{U}$ "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.