Journal of Statistical Software

MMMMMM YYYY, Volume˜VV, Issue˜II. *http://www.jstatsoft.org/*

Fitting linear mixed-effects models using lme4

Douglas Bates
University of Wisconsin - Madison

Martin Mächler
ETH Zurich

Ben Bolker
McMaster University

Abstract

Maximum likelihood or REML estimates of the parameters in linear mixed-effects models can be determined using the `lmer` function in the **lme4** package for R. As in most model-fitting functions, the model is described in an `lmer` call by a formula, in this case including both fixed-effects terms and random-effects terms. The formula and data together determine a numerical representation of the model from which the profiled deviance or the profiled REML criterion can be evaluated as a function of some of the model parameters. The appropriate criterion is optimized, using one of the constrained optimization functions in R, to provide the parameter estimates. We describe the structure of the model, the steps in evaluating the profiled deviance or REML criterion and the structure of the S4 class that represents such a model. Sufficient detail is included to allow specialization of these structures by those who wish to write functions to fit specialized linear mixed models, such as models incorporating pedigrees or smoothing splines, that aren't easily expressible in the formula language used by `lmer`.

Keywords:˜sparse matrix methods, linear mixed models, penalized least squares, Cholesky decomposition.

1. Introduction

The `lme4` package for R provides functions to fit and analyze linear mixed models (LMMs), generalized linear mixed models (GLMMs) and nonlinear mixed models (NLMMs). In each of these names, the term "mixed" or, more fully, "mixed-effects", denotes a model that incorporates both fixed-effects terms and random-effects terms in a linear predictor expression from which the conditional mean of the response can be evaluated. In this paper we describe the formulation and representation of linear and generalized linear mixed models. The techniques used for nonlinear mixed models will be described separately.

Just as a linear model can be described in terms of the distribution of $\mathcal{Y}$, the vector-valued random variable whose observed value is $\boldsymbol{y}_{\text{obs}}$, the observed response vector, a linear mixed

model can be described by the distribution of two vector-valued random variables: $\mathcal{Y}$, the response and $\mathcal{B}$, the vector of random effects. In a linear model the distribution of $\mathcal{Y}$ is multivariate normal,

$$\mathcal{Y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n), \tag{1}$$

where n is the dimension of the response vector, $\boldsymbol{I}_n$ is the identity matrix of size n, $\boldsymbol{\beta}$ is a p-dimensional coefficient vector and $\boldsymbol{X}$ is an $n \times p$ model matrix. The parameters of the model are the coefficients, $\boldsymbol{\beta}$, and the scale parameter, σ.

In a linear mixed model it is the *conditional* distribution of $\mathcal{Y}$ given $\mathcal{B} = \boldsymbol{b}$ that has such a form,

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

where $\boldsymbol{Z}$ is the $n \times q$ model matrix for the q-dimensional vector-valued random effects variable, $\mathcal{B}$, whose value we are fixing at $\boldsymbol{b}$. The unconditional distribution of $\mathcal{B}$ is also multivariate normal with mean zero and a parameterized $q \times q$ variance-covariance matrix, $\boldsymbol{\Sigma}$,

$$\mathcal{B} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}). \tag{3}$$

As a variance-covariance matrix, $\boldsymbol{\Sigma}$ must be positive semidefinite. It is convenient to express the model in terms of a *relative covariance factor*, $\boldsymbol{\Lambda}_\theta$, which is a $q \times q$ matrix, depending on the *variance-component parameter*, $\boldsymbol{\theta}$, and generating the symmetric $q \times q$ variance-covariance matrix, $\boldsymbol{\Sigma}$, according to

$$\boldsymbol{\Sigma}_\theta = \sigma^2\boldsymbol{\Lambda}_\theta\boldsymbol{\Lambda}_\theta', \tag{4}$$

where σ is the same scale factor as in the conditional distribution (2).

Although q, the number of columns in $\boldsymbol{Z}$ and the size of $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$, can be very large indeed, the dimension of $\boldsymbol{\theta}$ is small, frequently less than 10.

In calls to the `lm` function for fitting linear models the form of the model matrix $\boldsymbol{X}$ is determined by the `formula` and `data` arguments. The right-hand side of the formula consists of one or more terms that each generate one or more columns in the model matrix, $\boldsymbol{X}$. For `lmer` the formula language is extended to allow for random-effects terms that generate the model matrix $\boldsymbol{Z}$ and the mapping from $\boldsymbol{\theta}$ to $\boldsymbol{\Lambda}_{\boldsymbol{\theta}}$.

To understand why the formulation in equations 2 and 3 is particularly useful, we first show that the profiled deviance (negative twice the log-likelihood) and the profiled REML criterion can be expressed as a function of $\boldsymbol{\theta}$ only. Furthermore these criteria can be evaluated quickly and accurately.

2. Profiling the deviance and the REML criterion

As stated above, $\boldsymbol{\theta}$ determines the $q \times q$ matrix $\boldsymbol{\Lambda}_\theta$ which, together with a value of σ^2, determines $\mathrm{Var}(\mathcal{B}) = \boldsymbol{\Sigma}_\theta = \sigma^2\boldsymbol{\Lambda}_\theta\boldsymbol{\Lambda}_\theta'$. If we define a *spherical*[1] *random effects* variable, $\mathcal{U}$, with distribution

$$\mathcal{U} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_q), \tag{5}$$

and set

$$\mathcal{B} = \boldsymbol{\Lambda}_\theta\mathcal{U}, \tag{6}$$

[1] $\mathcal{N}(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$ distributions are called "spherical" because contours of the probability density are spheres.

then $\mathcal{B}$ will have the desired $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_\theta)$ distribution.

Although it may seem more natural to define $\mathcal{U}$ in terms of $\mathcal{B}$ we must write the relationship as in eqn.˜6 because $\mathbf{\Lambda}_\theta$ may be singular. In fact, it is important to allow for $\mathbf{\Lambda}_\theta$ to be singular because situations where the parameter estimates, $\widehat{\boldsymbol{\theta}}$, produce a singular $\mathbf{\Lambda}_{\widehat{\theta}}$ do occur in practice. And even if the parameter estimates do not correspond to a singular $\mathbf{\Lambda}_\theta$, it may be necessary to evaluate the estimation criterion at such values during the course of the numerical optimization of the criterion.

The model can now be defined in terms of

$$(\mathcal{Y}|\mathcal{U} = \boldsymbol{u}) \sim \mathcal{N}(\boldsymbol{Z}\mathbf{\Lambda}_\theta\boldsymbol{u} + \boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n) \tag{7}$$

producing the joint density function

$$\begin{aligned} f_{\mathcal{Y},\mathcal{U}}(\boldsymbol{y}, \boldsymbol{u}) &= f_{\mathcal{Y}|\mathcal{U}}(\boldsymbol{y}|\boldsymbol{u})\, f_{\mathcal{U}}(\boldsymbol{u}) \\ &= \frac{\exp(-\frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\mathbf{\Lambda}_\theta\boldsymbol{u}\|^2)}{(2\pi\sigma^2)^{n/2}} \, \frac{\exp(-\frac{1}{2\sigma^2}\|\boldsymbol{u}\|^2)}{(2\pi\sigma^2)^{q/2}} \\ &= \frac{\exp(-\left[\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\mathbf{\Lambda}_\theta\boldsymbol{u}\|^2 + \|\boldsymbol{u}\|^2\right]/(2\sigma^2))}{(2\pi\sigma^2)^{(n+q)/2}}. \end{aligned} \tag{8}$$

The *likelihood* of the parameters, $\boldsymbol{\theta}$, $\boldsymbol{\beta}$ and σ^2, given the observed data is the value of the marginal density of $\mathcal{Y}$, evaluated at $\boldsymbol{y}_{\text{obs}}$. That is

$$L(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma^2|\boldsymbol{y}_{\text{obs}}) = \int_{\mathbb{R}^q} f_{\mathcal{Y},\mathcal{U}}(\boldsymbol{y}_{\text{obs}}, \boldsymbol{u})\, d\boldsymbol{u}. \tag{9}$$

The integrand of eqn.˜9 is the *unscaled conditional density* of $\mathcal{U}$ given $\mathcal{Y} = \boldsymbol{y}_{\text{obs}}$. The conditional density of $\mathcal{U}$ given $\mathcal{Y} = \boldsymbol{y}_{\text{obs}}$ is

$$f_{\mathcal{U}|\mathcal{Y}}(\boldsymbol{u}|\boldsymbol{y}_{\text{obs}}) = \frac{f_{\mathcal{Y},\mathcal{U}}(\boldsymbol{y}_{\text{obs}}, \boldsymbol{u})}{\int f_{\mathcal{Y},\mathcal{U}}(\boldsymbol{y}_{\text{obs}}, \boldsymbol{u})\, d\boldsymbol{u}} \tag{10}$$

which is, up to a scale factor, the joint density, $f_{\mathcal{Y},\mathcal{U}}(\boldsymbol{y}_{\text{obs}}, \boldsymbol{u})$. The unscaled conditional density will be, up to a scale factor, a q-dimensional multivariate Gaussian with an integral that is easily evaluated if we know the mean and variance-covariance of the conditional density.

The conditional mean, $\boldsymbol{\mu}_{\mathcal{U}|\mathcal{Y}=\boldsymbol{y}_{\text{obs}}}$, is also the mode of the conditional distribution. Because a constant factor in a function does not affect the location of the optimum, we can determine the conditional mode, and hence the conditional mean, by maximizing the unscaled conditional density. This is in the form of a *penalized linear least squares* problem,

$$\boldsymbol{\mu}_{\mathcal{U}|\mathcal{Y}=\boldsymbol{y}_{\text{obs}}} = \arg\min_{\boldsymbol{u}} \left(\|\boldsymbol{y}_{\text{obs}} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\mathbf{\Lambda}_\theta\boldsymbol{u}\|^2 + \|\boldsymbol{u}\|^2\right). \tag{11}$$

2.1. Solving the penalized least squares problem

In the so-called "pseudo-data" approach to penalized least squares problems we write the objective as a residual sum of squares for an extended response vector and model matrix

$$\|\boldsymbol{y}_{\text{obs}} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\mathbf{\Lambda}_\theta\boldsymbol{u}\|^2 + \|\boldsymbol{u}\|^2 = \left\| \begin{bmatrix} \boldsymbol{y}_{\text{obs}} - \boldsymbol{X}\boldsymbol{\beta} \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \boldsymbol{Z}\mathbf{\Lambda}_\theta \\ \boldsymbol{I}_q \end{bmatrix} \boldsymbol{u} \right\|^2. \tag{12}$$

The contribution to the residual sum of squares from the "pseudo" observations appended to $\boldsymbol{y}_{\text{obs}} - \boldsymbol{X}\boldsymbol{\beta}$, is exactly the penalty term, $\|\boldsymbol{u}\|^2$.

From eqn.˜12 we can see that the conditional mean satisfies

$$\left(\boldsymbol{\Lambda}_\theta'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Lambda}_\theta + \boldsymbol{I}_q\right)\boldsymbol{\mu}_{\mathcal{U}|\mathcal{Y}=\boldsymbol{y}_{\text{obs}}} = \boldsymbol{\Lambda}_\theta'\boldsymbol{Z}'(\boldsymbol{y}_{\text{obs}} - \boldsymbol{X}\boldsymbol{\beta}), \tag{13}$$

which would be interesting, but not terribly useful, were it not for the fact that we can determine the solution to eqn.˜13 quickly and accurately, even when q, the size of the system to solve, is very large indeed. (We have done so in cases where q is in the millions.)

The key to solving eqn.˜13 is the *sparse Cholesky factor*, $\boldsymbol{L}_\theta$, which is a sparse, lower-triangular matrix such that

$$\boldsymbol{L}_\theta\boldsymbol{L}_\theta' = \boldsymbol{P}\left(\boldsymbol{\Lambda}_\theta'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Lambda}_\theta + \boldsymbol{I}_q\right)\boldsymbol{P}', \tag{14}$$

where $\boldsymbol{P}$ is a permutation matrix representing a fill-reducing permutation˜(Davis 2006, Ch.˜7).

As for most sparse matrix methods, the sparse Cholesky factorization can be split into two phases: a symbolic phase in which the positions of the non-zero elements in the result are determined and a numeric phase in which the actual numeric values in these positions are determined. Determining the fill-reducing permutation represented by $\boldsymbol{P}$ is part of the symbolic phase, which often takes much longer than the numeric phase. During the course of determining the maximum likelihood or REML estimates of the parameters in a linear mixed-effects model we may need to evaluate $\boldsymbol{L}_\theta$ for many different values of $\boldsymbol{\theta}$, but each evaluation after the first requires only the numeric phase. Changing $\boldsymbol{\theta}$ can change the values of the non-zero elements in $\boldsymbol{L}$ but does not change their positions. Hence, the symbolic phase must be done only once.

The `Cholesky` function in the **Matrix** package for R performs both the symbolic and numeric phases of the factorization to produce $\boldsymbol{L}_\theta$ from $\boldsymbol{\Lambda}_\theta'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Lambda}_\theta$. The resulting object has S4 class `"CHMsuper"` or `"CHMsimp"` depending on whether it is in the supernodal˜(Davis 2006, §˜4.8) or simplicial form. Both these classes inherit from the virtual class `"CHMfactor"`. Optional arguments to the `Cholesky` function control determination of a fill-reducing permutation and addition of multiple of the identity to the symmetric matrix before factorization. Once the factor has been determined for the initial value, $\boldsymbol{\theta}_0$, it can be updated for new values of $\boldsymbol{\theta}$ in a single call to the `update` method.

Although the `solve` method for the `"CHMfactor"` class has an option to evaluate $\boldsymbol{\mu}_{\mathcal{U}|\mathcal{Y}=\boldsymbol{y}_{\text{obs}}}$ directly as the solution to

$$\boldsymbol{P}'\boldsymbol{L}_\theta\boldsymbol{L}_\theta'\boldsymbol{P}\boldsymbol{\mu}_{\mathcal{U}|\mathcal{Y}=\boldsymbol{y}_{\text{obs}}} = \boldsymbol{\Lambda}_\theta'\boldsymbol{Z}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}). \tag{15}$$

we will express the solution in two stages:

1. Solve $\boldsymbol{L}\boldsymbol{c}_{\boldsymbol{u}} = \boldsymbol{P}\boldsymbol{\Lambda}_\theta'\boldsymbol{Z}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$ for $\boldsymbol{c}_{\boldsymbol{u}}$.
2. Solve $\boldsymbol{L}'\boldsymbol{P}\boldsymbol{\mu}_{\mathcal{U}|\mathcal{Y}=\boldsymbol{y}_{\text{obs}}} = \boldsymbol{c}_{\boldsymbol{u}}$ for $\boldsymbol{P}\boldsymbol{\mu}_{\mathcal{U}|\mathcal{Y}=\boldsymbol{y}_{\text{obs}}}$ and then for $\boldsymbol{\mu}_{\mathcal{U}|\mathcal{Y}=\boldsymbol{y}_{\text{obs}}} = \boldsymbol{P}'\left(\boldsymbol{P}\boldsymbol{\mu}_{\mathcal{U}|\mathcal{Y}=\boldsymbol{y}_{\text{obs}}}\right)$.

2.2. Evaluating the likelihood

After solving for $\boldsymbol{\mu}_{\mathcal{U}|\mathcal{Y}=\boldsymbol{y}_{\text{obs}}}$ the exponent in $f_{\mathcal{Y},\mathcal{U}}(\boldsymbol{y}_{\text{obs}}, \boldsymbol{u})$ can be written

$$\|\boldsymbol{y}_{\text{obs}} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{\Lambda}_\theta\boldsymbol{u}\|^2 + \|\boldsymbol{u}\|^2 = r^2(\boldsymbol{\theta}, \boldsymbol{\beta}) + \|\boldsymbol{L}'\boldsymbol{P}(\boldsymbol{u} - \boldsymbol{\mu}_{\mathcal{U}|\mathcal{Y}=\boldsymbol{y}_{\text{obs}}})\|^2. \tag{16}$$

where $r^2(\boldsymbol{\theta},\boldsymbol{\beta}) = \|\boldsymbol{y}_{\text{obs}} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{\Lambda}_\theta\boldsymbol{\mu}_{\mathcal{U}|\mathcal{Y}=\boldsymbol{y}_{\text{obs}}}\|^2 + \|\boldsymbol{\mu}_{\mathcal{U}|\mathcal{Y}=\boldsymbol{y}_{\text{obs}}}\|^2$, is the minimum penalized residual sum of squares for these values of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$.

With expression (16) and the change of variable $\boldsymbol{v} = \boldsymbol{L}'\boldsymbol{P}(\boldsymbol{u} - \boldsymbol{\mu}_{\mathcal{U}|\mathcal{Y}=\boldsymbol{y}_{\text{obs}}})$, for which $d\boldsymbol{v} = \text{abs}(|\boldsymbol{L}||\boldsymbol{P}|)\, d\boldsymbol{u}$, we have

$$\int \frac{\exp\left(\frac{-\|\boldsymbol{L}'\boldsymbol{P}(\boldsymbol{u}-\boldsymbol{\mu}_{\mathcal{U}|\mathcal{Y}})\|^2}{2\sigma^2}\right)}{(2\pi\sigma^2)^{q/2}}\, d\boldsymbol{u} = \int \frac{\exp\left(\frac{-\|\boldsymbol{v}\|^2}{2\sigma^2}\right)}{(2\pi\sigma^2)^{q/2}} \frac{d\boldsymbol{v}}{\text{abs}(|\boldsymbol{L}||\boldsymbol{P}|)} = \frac{1}{\text{abs}(|\boldsymbol{L}||\boldsymbol{P}|)} = \frac{1}{|\boldsymbol{L}|} \tag{17}$$

because $\text{abs}\,|\boldsymbol{P}| = 1$ (one property of a permutation matrix is $|\boldsymbol{P}| = \pm 1$) and $|\boldsymbol{L}|$, which, because $\boldsymbol{L}$ is triangular, is the product of its diagonal elements, all of which are positive, is positive.

Using this expression we can write the deviance (negative twice the log-likelihood) as

$$-2\ell(\boldsymbol{\theta},\boldsymbol{\beta},\sigma^2|\boldsymbol{y}_{\text{obs}}) = -2\log L(\boldsymbol{\theta},\boldsymbol{\beta},\sigma^2|\boldsymbol{y}_{\text{obs}}) = n\log(2\pi\sigma^2) + \frac{r^2(\boldsymbol{\theta},\boldsymbol{\beta})}{\sigma^2} + \log(|\boldsymbol{L}_\theta|^2) \tag{18}$$

Because the dependence of eqn.˜18 on σ^2 is straightforward, we can form the conditional estimate

$$\widehat{\sigma^2}(\boldsymbol{\theta},\boldsymbol{\beta}) = \frac{r^2(\boldsymbol{\theta},\boldsymbol{\beta})}{n} m \tag{19}$$

producing the *profiled deviance*

$$-2\tilde{\ell}(\boldsymbol{\theta},\boldsymbol{\beta}|\boldsymbol{y}_{\text{obs}}) = \log(|\boldsymbol{L}_\theta|^2) + n\left[1 + \log\left(\frac{2\pi r^2(\boldsymbol{\theta},\boldsymbol{\beta})}{n}\right)\right] \tag{20}$$

However, observing that eqn.˜20 depends on $\boldsymbol{\beta}$ only through $r^2(\boldsymbol{\theta},\boldsymbol{\beta})$ provides a much greater simplification because it allows us to "profile out" the fixed-effects parameter, $\boldsymbol{\beta}$, from the evaluation of the deviance. The conditional estimate, $\widehat{\boldsymbol{\beta}}_\theta$, is the value of $\boldsymbol{\beta}$ at the solution of the joint penalized least squares problem

$$r_\theta^2 = \min_{\boldsymbol{u},\boldsymbol{\beta}} \left(\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{\Lambda}_\theta\boldsymbol{u}\|^2 + \|\boldsymbol{u}\|^2\right), \tag{21}$$

producing the profiled deviance,

$$-2\tilde{\ell}(\boldsymbol{\theta}) = \log(|\boldsymbol{L}_\theta|^2) + n\left[1 + \log\left(\frac{2\pi r_\theta^2}{n}\right)\right], \tag{22}$$

which is a function of $\boldsymbol{\theta}$ only. Eqn.˜22 is a remarkably compact expression for the deviance.

2.3. Solving the joint penalized least squares problem

The solutions, $\boldsymbol{\mu}_{\mathcal{U}|\mathcal{Y}=\boldsymbol{y}_{\text{obs}}}$ and $\widehat{\boldsymbol{\beta}}_\theta$, of the joint penalized least squares problem (21) satisfy

$$\begin{bmatrix} \boldsymbol{\Lambda}_\theta'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Lambda}_\theta + \boldsymbol{I}_q & \boldsymbol{\Lambda}_\theta'\boldsymbol{Z}'\boldsymbol{X} \\ \boldsymbol{X}'\boldsymbol{Z}\boldsymbol{\Lambda}_\theta & \boldsymbol{X}'\boldsymbol{X} \end{bmatrix} \begin{bmatrix} \boldsymbol{\mu}_{\mathcal{U}|\mathcal{Y}=\boldsymbol{y}_{\text{obs}}} \\ \widehat{\boldsymbol{\beta}}_\theta \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Lambda}_\theta'\boldsymbol{Z}'\boldsymbol{y}_{\text{obs}} \\ \boldsymbol{X}'\boldsymbol{y}_{\text{obs}}. \end{bmatrix} \tag{23}$$

As before we will use the sparse Cholesky decomposition producing, $\boldsymbol{L}_\theta$, the sparse Cholesky factor, and $\boldsymbol{P}$, the permutation matrix, satisfying $\boldsymbol{L}_\theta\boldsymbol{L}_\theta' = \boldsymbol{P}(\boldsymbol{\Lambda}_\theta'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Lambda}_\theta + \boldsymbol{I})\boldsymbol{P}'$ and $\boldsymbol{c}_{\boldsymbol{u}}$, the solution to $\boldsymbol{L}_\theta\boldsymbol{c}_{\boldsymbol{u}} = \boldsymbol{P}\boldsymbol{\Lambda}_\theta'\boldsymbol{Z}'\boldsymbol{y}_{\text{obs}}$.

We extend the decomposition with the $q \times p$ matrix $\boldsymbol{R}_{ZX}$, the upper triangular $p \times p$ matrix $\boldsymbol{R}_X$, and the p-vector $\boldsymbol{c}_{\boldsymbol{\beta}}$ satisfying

$$\begin{aligned}\boldsymbol{L}\boldsymbol{R}_{ZX} &= \boldsymbol{P}\boldsymbol{\Lambda}_\theta'\boldsymbol{Z}'\boldsymbol{X}\\ \boldsymbol{R}_X'\boldsymbol{R}_X &= \boldsymbol{X}'\boldsymbol{X} - \boldsymbol{R}_{ZX}'\boldsymbol{R}_{ZX}\\ \boldsymbol{R}_X'\boldsymbol{c}_{\boldsymbol{\beta}} &= \boldsymbol{X}'\boldsymbol{y}_{\text{obs}} - \boldsymbol{R}_{ZX}'\boldsymbol{c}_{\boldsymbol{u}}\end{aligned}$$

so that

$$\begin{bmatrix}\boldsymbol{P}'\boldsymbol{L} & \boldsymbol{0}\\ \boldsymbol{R}_{ZX}' & \boldsymbol{R}_X'\end{bmatrix}\begin{bmatrix}\boldsymbol{L}'\boldsymbol{P} & \boldsymbol{R}_{ZX}\\ \boldsymbol{0} & \boldsymbol{R}_X\end{bmatrix} = \begin{bmatrix}\boldsymbol{\Lambda}_\theta'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Lambda}_\theta + \boldsymbol{I} & \boldsymbol{\Lambda}_\theta'\boldsymbol{Z}'\boldsymbol{X}\\ \boldsymbol{X}'\boldsymbol{Z}\boldsymbol{\Lambda}_\theta & \boldsymbol{X}'\boldsymbol{X}\end{bmatrix}, \tag{24}$$

and the solutions, $\boldsymbol{\mu}_{\mathcal{U}|\mathcal{Y}=\boldsymbol{y}_{\text{obs}}}$ and $\widehat{\boldsymbol{\beta}}_\theta$, satisfy

$$\boldsymbol{R}_X\widehat{\boldsymbol{\beta}}_\theta = \boldsymbol{c}_{\boldsymbol{\beta}} \tag{25}$$

$$\boldsymbol{L}'\boldsymbol{P}\boldsymbol{\mu}_{\mathcal{U}|\mathcal{Y}=\boldsymbol{y}_{\text{obs}}} = \boldsymbol{c}_{\boldsymbol{u}} - \boldsymbol{R}_{ZX}\widehat{\boldsymbol{\beta}}_\theta. \tag{26}$$

2.4. The profiled REML criterion

Laird and Ware (1982) show that the criterion to be optimized by the REML estimates can be expressed as

$$L_R(\boldsymbol{\theta}, \sigma^2|\boldsymbol{y}_{\text{obs}}) = \int L(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma^2|\boldsymbol{y}_{\text{obs}})\, d\boldsymbol{\beta}. \tag{27}$$

Because the joint solutions, $\boldsymbol{\mu}_{\mathcal{U}|\mathcal{Y}=\boldsymbol{y}_{\text{obs}}}$ and $\widehat{\boldsymbol{\beta}}_\theta$, to the penalized least squares problem allow us to express

$$\begin{aligned}\|\boldsymbol{y}_{\text{obs}} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{\Lambda}_\theta\boldsymbol{u}\|^2 + \|\boldsymbol{u}\|^2 =\\ r_\theta^2 + \left\|\boldsymbol{L}'\boldsymbol{P}\left[\boldsymbol{u} - \boldsymbol{\mu}_{\mathcal{U}|\mathcal{Y}=\boldsymbol{y}_{\text{obs}}} - \boldsymbol{R}_{ZX}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_\theta)\right]\right\|^2 + \left\|\boldsymbol{R}_X(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_\theta)\right\|^2\end{aligned} \tag{28}$$

we can use a change of variable, similar to that in eqn.˜17, to evaluate the profiled REML criterion. On the deviance scale the criterion can be evaluated as

$$-2\tilde{\ell}_R(\boldsymbol{\theta}) = \log(|\boldsymbol{L}|^2) + \log(|\boldsymbol{R}_X|^2) + (n-p)\left[1 + \log\left(\frac{2\pi r_\theta^2}{n-p}\right)\right]. \tag{29}$$

The structures in **lme4** for representing mixed-models are somewhat more general than is required for linear mixed models. In the remainder of this section we briefly describe some of the computational requirements for generalized linear mixed models, to show why these more general structures are employed.

2.5. Definition of GLMMs

The generalized linear mixed models (GLMMs) that can be fit by the **lme4** package preserve the multivariate Gaussian unconditional distribution of the random effects, $\mathcal{B}$ (eqn.˜3). Because most families used for the conditional distribution, $\mathcal{Y}|\mathcal{B} = \boldsymbol{b}$, do not incorporate a separate scale factor, σ, we remove it from the definition of $\boldsymbol{\Sigma}$ and from the distribution of the spherical random effects, $\mathcal{U}$. That is

$$\mathcal{U} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_q) \tag{30}$$

and

$$\boldsymbol{\Sigma}_\theta = \boldsymbol{\Lambda}_\theta \boldsymbol{\Lambda}_\theta'. \tag{31}$$

The conditional distributions, $\mathcal{Y}|\mathcal{B} = \boldsymbol{b}$ and $\mathcal{Y}|\mathcal{U} = \boldsymbol{u}$, preserve the properties that the components of $\mathcal{Y}$ are conditionally independent and that the mean, $\boldsymbol{\mu}_{\mathcal{Y}|\mathcal{U}=\boldsymbol{u}}$, depends on $\boldsymbol{u}$ only through the linear predictor,

$$\boldsymbol{\eta} = \boldsymbol{Z}\boldsymbol{\Lambda}_\theta \boldsymbol{u} + \boldsymbol{X}\boldsymbol{\beta}. \tag{32}$$

The mapping from $\boldsymbol{\mu}_{\mathcal{Y}|\mathcal{U}=\boldsymbol{u}}$ to $\boldsymbol{\eta}$, which is called the *link function* and written

$$\boldsymbol{Z}\boldsymbol{\Lambda}_\theta \boldsymbol{u} + \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\eta} = \boldsymbol{g}\left(\boldsymbol{\mu}_{\mathcal{Y}|\mathcal{U}=\boldsymbol{u}}\right), \tag{33}$$

is a *diagonal mapping* in the sense that there is a scalar function, g, such that the ith component of $\boldsymbol{\eta}$ is g applied to the ith component of $\boldsymbol{\mu}_{\mathcal{Y}|\mathcal{U}=\boldsymbol{u}}$. (The name "diagonal" reflects the fact that the Jacobian matrix, $\frac{d\eta}{d\mu'}$, of such a mapping will be diagonal.)

The scalar link function must be invertible over its range. The vector-valued *inverse link* function, $\boldsymbol{g}^{-1}$, will be the scalar inverse link, g^{-1}, applied component-wise to $\boldsymbol{\eta}$.

Common forms of the conditional distribution are Bernoulli, for binary responses, binomial for binary responses that are recorded as the number of trials and the number of successes, and Poisson, for count data. The combination of a distributional form and a link function is called a *family*. For distributional forms in the exponential family there is a *canonical link*. For Bernoulli or binomial forms the canonical link is the *logit* link function

$$\eta_i = \log\left(\frac{\mu_i}{1-\mu_i}\right); \tag{34}$$

for the Poisson distribution the canonical link is the natural logarithm.

The form of the distribution determines the conditional variance, $\mathrm{Var}(\mathcal{Y}|\mathcal{U} = \boldsymbol{u})$, as a function of the conditional mean and, possibly, a separate scale factor. (In most cases the conditional variance is completely determined by the conditional mean.)

The likelihood of the parameters, given the observed data, is now

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}|\boldsymbol{y}_{\text{obs}}) = \int_{\mathbb{R}^q} f_{\mathcal{Y},\mathcal{U}}(\boldsymbol{y}_{\text{obs}}, \boldsymbol{u})\, d\boldsymbol{u} \tag{35}$$

where, as in the case of linear mixed models, $f_{\mathcal{Y},\mathcal{U}}(\boldsymbol{y}_{\text{obs}}, \boldsymbol{u})$ is the unscaled conditional density of $\mathcal{U}$ given $\mathcal{Y} = \boldsymbol{y}_{\text{obs}}$. The notation here is a bit blurred because, although the joint distribution of $\mathcal{Y}$ and $\mathcal{U}$ is always continuous with respect to $\mathcal{U}$, it can be (and often is) discrete with respect to $\mathcal{Y}$. However, when we condition on the observed value $\mathcal{Y} = \boldsymbol{y}_{\text{obs}}$, the resulting function is continuous with respect to $\boldsymbol{u}$ so the unscaled conditional density is indeed well-defined as a density, up to a scale factor.

2.6. Determining the conditional mode

As for linear mixed models, we simplify evaluation of the integral (35) by determining the value, $\tilde{\boldsymbol{u}}_{\beta,\theta}$, that maximizes the unscaled conditional density. When the conditional density, $\mathcal{U}|\mathcal{Y} = \boldsymbol{y}_{\text{obs}}$, is multivariate Gaussian, this conditional mode will also be the conditional mean. However, for most families used in GLMMs, the mode and the mean need not coincide so use the more general term and call $\tilde{\boldsymbol{u}}_{\beta,\theta}$ the *conditional mode*.

The iteratively reweighted least squares (IRLS) algorithm is an incredibly efficient method of determining the maximum likelihood estimates of the coefficients in a generalized linear model. We extend it to a *penalized iteratively reweighted least squares* (PIRLS) algorithm for determining the conditional mode, $\tilde{\boldsymbol{u}}_{\beta,\theta}$. This algorithm has the form

1. Given parameter values, $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, and starting estimates, $\boldsymbol{u}_0$, evaluate the linear predictor, $\boldsymbol{\eta}$, the corresponding conditional mean, $\boldsymbol{\mu}_{\mathcal{Y}|\mathcal{U}=\boldsymbol{u}}$, and the conditional variance. Establish the weights as the inverse of the variance. We write these weights in the form of a diagonal weight matrix, $\boldsymbol{W}$, although they are stored and manipulated as a vector.

2. Solve the penalized, weighted, nonlinear least squares problem

$$\arg\min_{\boldsymbol{u}} \left(\left\| \boldsymbol{W}^{1/2} \left(\boldsymbol{y}_{\text{obs}} - \boldsymbol{\mu}_{\mathcal{Y}|\mathcal{U}=\boldsymbol{u}} \right) \right\|^2 + \|\boldsymbol{u}\|^2 \right) \tag{36}$$

3. Update the weights, $\boldsymbol{W}$, and check for convergence. If not converged, go to step 2.

We use a Gauss-Newton algorithm with an orthogonality convergence criterion˜(Bates and Watts 1988, §2.2.3) to solve the penalized, weighted, nonlinear least squares problem in step 2. At the ith iteration we determine an increment, $\boldsymbol{\delta}_i$, as the solution to the penalized, weighted, linear least squares problem

$$\boldsymbol{\delta}_i = \arg\min_{\boldsymbol{\delta}} \left\| \begin{bmatrix} \boldsymbol{W}^{1/2} \left(\boldsymbol{y}_{\text{obs}} - \boldsymbol{\mu}_i \right) \\ \boldsymbol{u}_i \end{bmatrix} - \begin{bmatrix} \boldsymbol{W}^{1/2} \boldsymbol{M}_i \boldsymbol{Z} \boldsymbol{\Lambda}_\theta \\ \boldsymbol{I}_q \end{bmatrix} \boldsymbol{u} \right\|^2 \tag{37}$$

where $\boldsymbol{u}_i$ is current value of $\boldsymbol{u}$, $\boldsymbol{\mu}_i$ is the corresponding conditional mean of $\mathcal{Y}|\mathcal{U} = \boldsymbol{u}_i$ and $\boldsymbol{M}_i$ is the Jacobian matrix of the vector-valued inverse link, evaluated at $\boldsymbol{\mu}_i$. That is

$$\boldsymbol{M}_i = \left. \frac{d\boldsymbol{\mu}}{d\boldsymbol{\eta}'} \right|_{\boldsymbol{\eta}_i}, \tag{38}$$

which will be a diagonal matrix so, as for the weights, we store and manipulate the Jacobian as a vector.

The minimizer, $\boldsymbol{\delta}_i$, of (37) satisfies

$$\boldsymbol{P} \left(\boldsymbol{\Lambda}_\theta' \boldsymbol{Z}' \boldsymbol{M}_i \boldsymbol{W} \boldsymbol{M}_i \boldsymbol{Z} \boldsymbol{\Lambda}_\theta + \boldsymbol{I}_q \right) \boldsymbol{P}' \boldsymbol{\delta}_i = \boldsymbol{\Lambda}_\theta' \boldsymbol{Z}' \boldsymbol{M}_i \boldsymbol{W} (\boldsymbol{y}_{\text{obs}} - \boldsymbol{\mu}_i) - \boldsymbol{u}_i \tag{39}$$

which we solve using the sparse Cholesky factor. At convergence, the factor, $\boldsymbol{L}_{\beta,\theta}$, satisfies

$$\boldsymbol{L}_{\beta,\theta} \boldsymbol{L}_{\beta,\theta}' = \boldsymbol{P} \left(\boldsymbol{\Lambda}_\theta' \boldsymbol{Z}' \boldsymbol{M} \boldsymbol{W} \boldsymbol{M} \boldsymbol{Z} \boldsymbol{\Lambda}_\theta + \boldsymbol{I}_q \right) \boldsymbol{P}' \tag{40}$$

The integrand in the likelihood (35) is approximately a constant times the density of the $\mathcal{N}(\tilde{\boldsymbol{u}}, \boldsymbol{L}\boldsymbol{L}')$ distribution. The *Laplace approximation* to the deviance is

$$d(\boldsymbol{\beta}, \boldsymbol{\theta} | \boldsymbol{y}) = d_g(\boldsymbol{y}_{\text{obs}}, \boldsymbol{\mu}(\tilde{\boldsymbol{u}})) + \|\tilde{\boldsymbol{u}}\|^2 + \log(|\boldsymbol{L}|^2) \tag{41}$$

where $d_g(\boldsymbol{y}_{\text{obs}}, \boldsymbol{\mu}(\tilde{\boldsymbol{u}}))$ is the GLM deviance for $\boldsymbol{y}_{\text{obs}}$ and $\boldsymbol{\mu}(\tilde{\boldsymbol{u}})$.

References

Bates DM, Watts DG (1988). *Nonlinear Regression Analysis and Its Applications*. Wiley, Hoboken, NJ. ISBN 0-471-81643-4.

Davis T (2006). *Direct Methods for Sparse Linear Systems*. SIAM, Philadelphia, PA.

Laird NM, Ware JH (1982). "Random-Effects Models for Longitudinal Data." *Biometrics*, **38**, 963–974.

Affiliation:

Douglas Bates
Department of Statistics, University of Wisconsin
1300 University Ave.
Madison, WI 53706, U.S.A.
E-mail: bates@stat.wisc.edu

Martin Mächler
Seminar für Statistik, HG G˜16
ETH Zurich
8092 Zurich, Switzerland
E-mail: maechler@stat.math.ethz.ch

Benjamin M. Bolker
Departments of Mathematics & Statistics and Biology
McMaster University
1280 Main Street W
Hamilton, ON L8S 4K1, Canada
E-mail: bolker@mcmaster.ca

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume˜VV, Issue˜II *Submitted:* yyyy-mm-dd
MMMMMM YYYY *Accepted:* yyyy-mm-dd