Theory and computational methods for mixed models

Douglas Bates
Department of Statistics
University of Wisconsin – Madison

June 20, 2007

Abstract

The `lme4` package provides R functions to fit and analyze 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.

The three types of mixed models – linear, generalized linear and nonlinear – share common characteristics in that the model is specified in whole or

in part by a *mixed model formula* that describes a *linear predictor* and a variance-covariance structure for the random effects. In the next section we describe the mixed model formula and the forms if these matrices. The following section presents a general formulation of the Laplace approximation to the log-likelihood of a mixed model.

In subsequent sections we describe computational methods for specific kinds of mixed models. In particular, we should how a profiled log-likelihood for linear mixed models and for some nonlinear mixed models can be evaluated.

2 Mixed-model formulas

The right-hand side of a mixed-model formula, as used in the `lme4` package, consists of one or more random-effects terms and zero or more fixed-effects terms separated by the '`+`' symbol. The fixed-effects terms generate the fixed-effects model matrix, $\boldsymbol{X}$, from the data. The random-effects terms generate the random-effects model matrix, $\boldsymbol{Z}$, and determine the structure of the relative variance-covariance matrix, $\boldsymbol{\Sigma}$.

The model matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ are of size $m \times p$ and $m \times q$ respectively. For linear and generalized linear mixed models m, the number of rows in $\boldsymbol{X}$ and $\boldsymbol{Y}$, is equal to n, the dimension of the response vector, $\boldsymbol{y}$. For nonlinear mixed models m is a multiple of n, $m = ns$, where s is the number of nonlinear model parameters.

The dimension of the fixed-effects parameter vector $\boldsymbol{\beta}$ is p and the dimension of the random effects vector $\boldsymbol{b}$ is q. Together with the matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ these vectors determine the linear predictor

$$\boldsymbol{\eta_b}(\boldsymbol{\beta}, \boldsymbol{b}) = \boldsymbol{X\beta} + \boldsymbol{Zb}. \tag{1}$$

The elements of $\boldsymbol{\beta}$ are parameters in the model. Strictly speaking, the elements of $\boldsymbol{b}$ are not parameters – they are unobserved random variables. In the models we will consider, the conditional distribution of the observed responses, $\boldsymbol{y}$, depends on $\boldsymbol{b}$ and $\boldsymbol{\beta}$ only through the linear predictor. The conditional density, for continuous $\boldsymbol{y}$, or the conditional probability mass function, for discrete $\boldsymbol{y}$, can be written in the form

$$f_{\boldsymbol{y}|\boldsymbol{b}}(\boldsymbol{y}|\boldsymbol{b}, \boldsymbol{\beta}, \sigma^2) = k(\sigma^2, \boldsymbol{y}) e^{-d(\boldsymbol{\mu}(\boldsymbol{X\beta}+\boldsymbol{Zb}), \boldsymbol{y})/(2\sigma^2)} = k(\sigma^2, \boldsymbol{y}) e^{-d(\boldsymbol{\mu}(\boldsymbol{\eta_b}(\boldsymbol{\beta}, \boldsymbol{b})), \boldsymbol{y})/\left(2\sigma^2\right)} \tag{2}$$

In (2) the *discrepancy function*, $d(\boldsymbol{\mu}, \boldsymbol{y})$, defines the "squared distance" between the *conditional mean*,

$$\boldsymbol{\mu}(\boldsymbol{\eta}) = \boldsymbol{\mu}\left(\boldsymbol{\eta}_{\boldsymbol{b}}(\boldsymbol{\beta}, \boldsymbol{b})\right) = \boldsymbol{\mu}\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}\right) \tag{3}$$

and the observed data, $\boldsymbol{y}$, according to the form of the model.

In fact, for linear mixed models and for nonlinear mixed models, the discrepancy is exactly the square of the usual Euclidean distance, $d(\boldsymbol{\mu}, \boldsymbol{y}) = \|\boldsymbol{y} - \boldsymbol{\mu}\|^2$.

The scale factor, σ^2, if it is used in the model, only determines the variance-covariance of the conditional distribution of $\boldsymbol{y}$; it does not affect the conditional mean. Some mixed models, such as generalized linear mixed models for which the conditional distribution of $\boldsymbol{y}$ is Bernoulli or binomial or Poisson, do not have a separate scale factor because the mean of the conditional distribution completely determines the variance. In such cases the conditional density or conditional probability mass function can be written $f_{\boldsymbol{y}|\boldsymbol{b}}(\boldsymbol{y}|\boldsymbol{b}, \boldsymbol{\beta}) = k(\boldsymbol{y})e^{-d(\boldsymbol{\mu}(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{b}),\boldsymbol{y})/2}$.

The normalization factor $k(\sigma^2, \boldsymbol{y})$ depends only on the scale factor, σ^2, if it is used in the model, and the observed response, $\boldsymbol{y}$.

The marginal distribution of $\boldsymbol{b}$ is modelled as a multivariate normal (or Gaussian) distribution of the form

$$\boldsymbol{b} \sim \mathcal{N}\left(\boldsymbol{0}, \sigma^2\boldsymbol{\Sigma}(\boldsymbol{\theta})\right) \tag{4}$$

where σ^2 is the same scale factor as in (2) and the $q \times q$ symmetric, positive-semidefinite matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, called the *relative variance-covariance matrix* of the random effects $\boldsymbol{b}$, is a function of a parameter $\boldsymbol{\theta}$. The condition that $\boldsymbol{\Sigma}$ is positive-semidefinite means that $\boldsymbol{v}'\boldsymbol{\Sigma}\boldsymbol{v} \geq 0, \forall \boldsymbol{v} \in \mathbb{R}^q$.

For linear and generalized linear mixed models the model matrix $\boldsymbol{X}$ is constructed from the data and the fixed-effects terms in the model formula according to the usual rules for model matrices in the S language (Chambers and Hastie, 1992, Chapter 2). For nonlinear mixed models these rules are modified somewhat, as described in §5.2.

2.1 Random-effects terms

A simple random-effects term is of the form '*form*|*factor*' where *form* is a linear model formula and *factor* is an expression that can be evaluated as a factor, called the *grouping factor* for the term. Such factors isolate the effect

of certain components of the random effects vector, $\boldsymbol{b}$, to a specific group of elements in the conditional mean $\boldsymbol{\mu}(\boldsymbol{X\beta} + \boldsymbol{Zb})$.

Typically a random-effects term is enclosed by parentheses so that the extent of *form* is clearly defined.

Let k be the number of random-effects terms in the formula and $n_i, i = 1, \ldots, k$ be the number of levels in the ith grouping factor, $\boldsymbol{f}_i$.

The linear model formula in the ith random-effects term determines the $m \times q_i$ model matrix $\boldsymbol{Z}_i$ according to the usual rules for model matrices, in the case of linear or generalized linear models, and according to slightly modified rules, as described in §5.2, for nonlinear mixed models.

Together $\boldsymbol{f}_i$ and $\boldsymbol{Z}_i$ determine the *indicator interaction matrix* $\tilde{\boldsymbol{Z}}_i$ that is the horizontal concatenation of the interaction of the columns of $\boldsymbol{Z}_i$ with the matrix of indicators of the levels of $\boldsymbol{f}_i$. That is, the $m \times n_i q_i$ matrix $\tilde{\boldsymbol{Z}}_i$ consists of q_i vertical blocks, each of size $m \times n_i$, whose nonzeros are in the form of the indicator columns for $\boldsymbol{f}_i$. The values of the nonzeros in the jth vertical block in $\tilde{\boldsymbol{Z}}_i$ are the jth column of $\boldsymbol{Z}_i$.

In the not-uncommon case that the linear model formula in a random-effects term is '`1`' then $q_i = 1$, $\boldsymbol{Z}_i$ is an $m \times 1$ matrix all of whose elements are 1 and $\tilde{\boldsymbol{Z}}_i$ is the $m \times n_i$ matrix of indicators of the levels of $\boldsymbol{f}_i$.

Suppose, for example, that we wish to model data where three observations have been recorded on each of five subjects. A data frame with a "subject" factor could be constructed as

```
> str(dat <- data.frame(subj = gl(5, 3, labels = LETTERS[1:5])))
'data.frame':        15 obs. of  1 variable:
 $ subj: Factor w/ 5 levels "A","B","C","D",..: 1 1 1 2 2 2 3 3 3 4 ...
```

if the data are ordered by subject. A random effects term of the form `(1|subj)` generates a model matrix $\boldsymbol{Z}_i$ which has one column, all of whose elements are unity

```
> str(Zi <- model.matrix(~1, dat))
 num [1:15, 1] 1 1 1 1 1 1 1 1 1 1 ...
 - attr(*, "dimnames")=List of 2
  ..$ : NULL
  ..$ : chr "(Intercept)"
 - attr(*, "assign")= int 0
```

and the indicator interaction matrix, $\tilde{\boldsymbol{Z}}_i$, which, in this case, is simply the matrix of indicators,

```
15 x 5 sparse Matrix of class "dgCMatrix"

 [1,] 1 . . . .
 [2,] 1 . . . .
 [3,] 1 . . . .
 [4,] . 1 . . .
 [5,] . 1 . . .
 [6,] . 1 . . .
 [7,] . . 1 . .
 [8,] . . 1 . .
 [9,] . . 1 . .
[10,] . . . 1 .
[11,] . . . 1 .
[12,] . . . 1 .
[13,] . . . . 1
[14,] . . . . 1
[15,] . . . . 1
```

In the lme4 package these sparse model matrices are stored as compressed, column-oriented, sparse matrices (Davis, 2006) of class `"dgCMatrix"`. When such a matrix is printed, the systematic zeros are shown as '.'.

For a more general example, assume that each subject was observed at times 1, 2 and 3. If the `time` variable is

```
> dat$time <- rep(1:3, 5)
```

so that the first few rows of the model frame are

```
> head(dat, n = 5)

  subj time
1    A    1
2    A    2
3    A    3
4    B    1
5    B    2
```

(i.e. we have ordered the data first by subject then by time within subject) then the first few rows of $\boldsymbol{Z}_1$ for a random-effects term `(time|Subject)` are

```
> head(Z1 <- model.matrix(~time, dat), n = 7)
```

```
  (Intercept) time
1           1    1
2           1    2
3           1    3
4           1    1
5           1    2
6           1    3
7           1    1
```

and the matrix $\tilde{\boldsymbol{Z}}_i$ is

```
15 x 10 sparse Matrix of class "dgCMatrix"

 [1,] 1 . . . . 1 . . . .
 [2,] 1 . . . . 2 . . . .
 [3,] 1 . . . . 3 . . . .
 [4,] . 1 . . . . 1 . . .
 [5,] . 1 . . . . 2 . . .
 [6,] . 1 . . . . 3 . . .
 [7,] . . 1 . . . . 1 . .
 [8,] . . 1 . . . . 2 . .
 [9,] . . 1 . . . . 3 . .
[10,] . . . 1 . . . . 1 .
[11,] . . . 1 . . . . 2 .
[12,] . . . 1 . . . . 3 .
[13,] . . . . 1 . . . . 1
[14,] . . . . 1 . . . . 2
[15,] . . . . 1 . . . . 3
```

The $m \times q$ matrix $\boldsymbol{Z}$ is the horizontal concatenation of the $\tilde{\boldsymbol{Z}}_i, i = 1, \ldots, k$. Thus

$$q = \sum_{i=1}^{k} n_i q_i. \tag{5}$$

2.2 The relative variance-covariance matrix

The elements of the random-effects vector $\boldsymbol{b}$ are partitioned into groups in that same way that the columns of $\boldsymbol{Z}$ were partitioned. That is, they are divided into k groups, corresponding to the k random-effects terms, and the ith such group is subdivided into n_i groups of q_i elements, corresponding to the levels of the ith grouping factor.

This partitioning determines the structure of the variance-covariance matrix of $\boldsymbol{b}$ because we assume that random effects corresponding to different terms are uncorrelated, as are random effects corresponding to different levels of the same term. Furthermore, the variance-covariance structures of the n_i groups of size q_i from the ith term are identical.

Thus the relative variance-covariance matrix, $\boldsymbol{\Sigma}$, has the form

$$\boldsymbol{\Sigma} = \begin{bmatrix} \tilde{\boldsymbol{\Sigma}}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \tilde{\boldsymbol{\Sigma}}_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \tilde{\boldsymbol{\Sigma}}_k. \end{bmatrix} \tag{6}$$

with the ith diagonal block of the form

$$\tilde{\boldsymbol{\Sigma}}_i = \begin{bmatrix} \sigma_{1,1}\boldsymbol{I}_{n_i} & \sigma_{1,2}\boldsymbol{I}_{n_i} & \dots & \sigma_{1,q_i}\boldsymbol{I}_{n_i} \\ \sigma_{1,2}\boldsymbol{I}_{n_i} & \sigma_{2,2}\boldsymbol{I}_{n_i} & \dots & \sigma_{2,q_i}\boldsymbol{I}_{n_i} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1,q_i}\boldsymbol{I}_{n_i} & \sigma_{2,q_i}\boldsymbol{I}_{n_i} & \dots & \sigma_{q_i,q_i}\boldsymbol{I}_{n_i} \end{bmatrix} = \boldsymbol{\Sigma}_i \otimes \boldsymbol{I}_{n_i} \tag{7}$$

where

$$\boldsymbol{\Sigma}_i = \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} & \dots & \sigma_{1,q_i} \\ \sigma_{1,2} & \sigma_{2,2} & \dots & \sigma_{2,q_i} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1,q_i} & \sigma_{2,q_i} & \dots & \sigma_{q_i,q_i} \end{bmatrix} \tag{8}$$

is a $q_i \times q_i$ symmetric matrix. (The symbol $\otimes$ denotes the Kronecker product of matrices, which is a convenient shorthand for a structure like that shown in (7).)

The matrix $\boldsymbol{\Sigma}$ will be positive-semidefinite if all the symmetric matrices $\boldsymbol{\Sigma}_i, i = 1, \dots, k$ are positive-semidefinite. This occurs if and only if each of the $\boldsymbol{\Sigma}_i$ has an Cholesky factorization of the "LDL′" form where the left factor "L" is a unit lower triangular matrix and "D" is a diagonal matrix with non-negative diagonal elements.

In the "LDL′" form of a variance-covariance matrix the elements of "D" would be on the variance scale. Because it will be more convenient to work with elements on the standard deviation scale we write factorization as

$$\boldsymbol{\Sigma}_i = \boldsymbol{T}_i \boldsymbol{S}_i \boldsymbol{S}_i \boldsymbol{T}_i' \quad i = 1, \dots, k \tag{9}$$

where $\boldsymbol{T}_i$ is a unit lower triangular matrix of size $q_i \times q_i$ and $\boldsymbol{S}_i$ is a diagonal $q_i \times q_i$ matrix with non-negative diagonal elements.

We parameterize $\boldsymbol{\Sigma}_i$ according to the decomposition (9) by defining $\boldsymbol{\theta}_i$ to be the vector of length $q_i(q_i+1)/2$ consisting of the diagonal elements of $\boldsymbol{S}_i$ followed by the elements in the strict lower triangle of $\boldsymbol{T}_i$ in row-major order.

Finally, let $\boldsymbol{\theta}$ be the concatenation of the $\boldsymbol{\theta}_i, i = 1, \ldots, k$.

The unit lower-triangular and non-negative diagonal factors, $\boldsymbol{T}(\boldsymbol{\theta})$ and $\boldsymbol{S}(\boldsymbol{\theta})$, of $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ are constructed from the $\boldsymbol{T}_i$, $\boldsymbol{S}_i$ and n_i, $i = 1, \ldots, k$ according to the pattern of $\boldsymbol{\Sigma}$ illustrated in (6) and (7). That is, $\boldsymbol{T}(\boldsymbol{\theta})$ (respectively $\boldsymbol{S}(\boldsymbol{\theta})$) is block-diagonal with ith diagonal block $\tilde{\boldsymbol{T}}_i = \boldsymbol{T} \otimes \boldsymbol{I}_{n_i}$ (respectively $\tilde{\boldsymbol{S}}_i = \boldsymbol{S} \otimes \boldsymbol{I}_{n_i}$).

Although the number of levels of the ith factor, n_i, can be very large, the number of columns in $\boldsymbol{Z}_i$, q_i, is typically very small. Hence the dimension of the parameter $\boldsymbol{\theta}_i$, which depends on q_i but not on n_i, is also small and the structure of $\boldsymbol{T}_i$ and $\boldsymbol{S}_i$ is often very simple.

Consider our example of 3 observations on each of 5 subjects. For the random effects term `(1|subj)` $q_i = 1$ and $\boldsymbol{T}_i$, which is a 1×1 unit lower triangular matrix, must be $[1]$, the 1×1 identity matrix. Hence $\tilde{\boldsymbol{T}}_i = \boldsymbol{I}_5$ and the decomposition $\tilde{\boldsymbol{\Sigma}} = \tilde{\boldsymbol{T}}\tilde{\boldsymbol{S}}\tilde{\boldsymbol{S}}\tilde{\boldsymbol{T}}'$ reduces to $\tilde{\boldsymbol{\Sigma}} = \tilde{\boldsymbol{S}}\tilde{\boldsymbol{S}}$. Furthermore, $\boldsymbol{S}_i = [\theta_{i,1}]$ subject to $\theta_{i,1} \geq 0$,

$$\tilde{\boldsymbol{S}}_i = \theta_{i,1}\boldsymbol{I}_5$$

and

$$\tilde{\boldsymbol{\Sigma}}_i = \tilde{\boldsymbol{S}}_i\tilde{\boldsymbol{S}}_i = \theta_{i,1}^2\boldsymbol{I}_5.$$

For a term like `(time|subj)` in which $q_i = 2$ let us write $\boldsymbol{\theta}_i$ as $[a, b, c]'$. Then

$$\boldsymbol{S}_i = \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}$$

so that

$$\tilde{\boldsymbol{S}}_i = \begin{bmatrix} a\boldsymbol{I}_5 & \boldsymbol{0} \\ \boldsymbol{0} & b\boldsymbol{I}_5 \end{bmatrix}$$

and

$$\boldsymbol{T}_i = \begin{bmatrix} 1 & 0 \\ c & 1 \end{bmatrix}$$

so that

$$\tilde{\boldsymbol{T}}_i = \begin{bmatrix} \boldsymbol{I}_5 & \boldsymbol{0} \\ c\boldsymbol{I}_5 & \boldsymbol{I}_5 \end{bmatrix}$$

. The constraints on $\boldsymbol{\theta}_i$ are $a \geq 0$ and $b \geq 0$.

2.3 Orthogonal random effects

For a fixed value of $\boldsymbol{\theta}$ we can write $\boldsymbol{b}$ as

$$\boldsymbol{b} = \boldsymbol{T}(\boldsymbol{\theta})\boldsymbol{S}(\boldsymbol{\theta})\boldsymbol{u} \tag{10}$$

where $\boldsymbol{u}$ is a vector of *orthogonal random effects* with distribution $\boldsymbol{u} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. This provides the desired distribution $\boldsymbol{b} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{\Sigma})$ because

$$\mathrm{E}[\boldsymbol{b}] = \mathrm{E}[\boldsymbol{T}\boldsymbol{S}\boldsymbol{u}] = \boldsymbol{T}\boldsymbol{S}\mathrm{E}[\boldsymbol{u}] = \boldsymbol{0}$$

and

$$\mathrm{Var}(\boldsymbol{b}) = \mathrm{E}[\boldsymbol{b}\boldsymbol{b}'] = \boldsymbol{T}\boldsymbol{S}\mathrm{E}[\boldsymbol{u}\boldsymbol{u}']\boldsymbol{S}\boldsymbol{T}' = \boldsymbol{T}\boldsymbol{S}\mathrm{Var}(\boldsymbol{u})\boldsymbol{S}\boldsymbol{T}' = \sigma^2\boldsymbol{T}\boldsymbol{S}\boldsymbol{S}\boldsymbol{T}' = \sigma^2\boldsymbol{\Sigma}$$

Because $\boldsymbol{\mu}$, the conditional mean of $\boldsymbol{y}$ given $\boldsymbol{b}$ and $\boldsymbol{\beta}$, and the discrepancy function, $d(\boldsymbol{\mu}, \boldsymbol{y})$, depend on $\boldsymbol{b}$ only through the linear predictor and because we can rewrite the linear predictor as a function of $\boldsymbol{\beta}$ and $\boldsymbol{u}$

$$\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{T}(\boldsymbol{\theta})\boldsymbol{S}(\boldsymbol{\theta})\boldsymbol{u} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{V}(\boldsymbol{\theta})\boldsymbol{u}, \tag{11}$$

where

$$\boldsymbol{V}(\boldsymbol{\theta}) = \boldsymbol{Z}\boldsymbol{T}(\boldsymbol{\theta})\boldsymbol{S}(\boldsymbol{\theta}), \tag{12}$$

we can express the discrepancy as a function of $\boldsymbol{\beta}$ and $\boldsymbol{u}$. These two forms of the discrepancy are

$$d_{\boldsymbol{b}}(\boldsymbol{b}, \boldsymbol{\beta}, \boldsymbol{y}) = d(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}, \boldsymbol{y}) \tag{13}$$

and

$$d_{\boldsymbol{u}}(\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{y}) = d(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{V}(\boldsymbol{\theta})\boldsymbol{u}, \boldsymbol{y}). \tag{14}$$

Note that $d_{\boldsymbol{u}}$ depends on $\boldsymbol{\theta}$ but $d_{\boldsymbol{b}}$ does not.

In the next section we will evaluate an integral with respect to $\boldsymbol{b}$ by changing the variable of integration to $\boldsymbol{u}$. When performing the change of variable on an integral with respect to a vector we must incorporate the determinant of the Jacobian of the transformation. For the transformation $\boldsymbol{b} = \boldsymbol{T}\boldsymbol{S}\boldsymbol{u}$ this is

$$\left|\frac{d\boldsymbol{b}}{d\boldsymbol{u}}\right| = |\boldsymbol{T}\boldsymbol{S}| = |\boldsymbol{T}|\ |\boldsymbol{S}| = |\boldsymbol{S}|$$

because $|\boldsymbol{T}|$, which is the product of the diagonal elements of this unit triangular matrix, is unity.

The determinant $|\boldsymbol{S}|$, which is the product of the diagonal elements of this diagonal matrix, is easily evaluated and must be non-negative. Furthermore,

$$|\boldsymbol{\Sigma}|^{1/2} = \sqrt{|\boldsymbol{T}|^2|\boldsymbol{S}|^2} = |\boldsymbol{S}|. \tag{15}$$

3 Evaluating the likelihood

If the distribution of $\boldsymbol{y}$ is continuous, the likelihood of the parameters, $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and σ^2, given the observed data, $\boldsymbol{y}$, is equal to the marginal density of $\boldsymbol{y}$ given the parameters. If the distribution of $\boldsymbol{y}$ is discrete, the likelihood is equal to the marginal probability mass function of $\boldsymbol{y}$ given the parameters.

In either case we can write the likelihood as

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2|\boldsymbol{y}) = \int_{\boldsymbol{b}} f_{\boldsymbol{y}|\boldsymbol{b}}(\boldsymbol{y}|\boldsymbol{b}, \boldsymbol{\beta}, \sigma^2)\, f_{\boldsymbol{b}}(\boldsymbol{b}|\boldsymbol{\theta}, \sigma^2)\, d\boldsymbol{b} \tag{16}$$

where $f_{\boldsymbol{y}|\boldsymbol{b}}(\boldsymbol{y}|\boldsymbol{b}, \boldsymbol{\beta}, \sigma^2)$, defined in (2), is the conditional density or the conditional probability mass function of $\boldsymbol{y}$, as appropriate.

As described in §2.1, the unconditional distribution of $\boldsymbol{b}$ is $\boldsymbol{b} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{\Sigma})$, for which the density function is

$$f_{\boldsymbol{b}}(\boldsymbol{b}|\boldsymbol{\theta}, \sigma^2) = \frac{e^{-\boldsymbol{b}'\boldsymbol{\Sigma}^{-1}\boldsymbol{b}/(2\sigma^2)}}{(2\pi\sigma^2)^{q/2}\,|\boldsymbol{\Sigma}|^{1/2}}. \tag{17}$$

Substituting (2) and (17) into (16) and changing the variable of integration from $\boldsymbol{b}$ to $\boldsymbol{u}$ produces

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2|\boldsymbol{y}) &= \int_{\boldsymbol{b}} \frac{k(\sigma^2, \boldsymbol{y})}{(2\pi\sigma^2)^{q/2}\,|\boldsymbol{\Sigma}|^{1/2}} \exp\left[\frac{d_{\boldsymbol{b}}(\boldsymbol{b}, \boldsymbol{\beta}, \boldsymbol{y}) + \boldsymbol{b}'\boldsymbol{\Sigma}^{-1}\boldsymbol{b}}{-2\sigma^2}\right] d\boldsymbol{b} \\ &= \int_{\boldsymbol{u}} \frac{k(\sigma^2, \boldsymbol{y})}{(2\pi\sigma^2)^{q/2}} \exp\left[\frac{d_{\boldsymbol{u}}(\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{y}) + \boldsymbol{u}'\boldsymbol{u}}{-2\sigma^2}\right] d\boldsymbol{u} \end{aligned} \tag{18}$$

Note that the likelihood can be evaluated for a positive-semidefinite $\boldsymbol{\Sigma}$ when written as an integral with respect to $\boldsymbol{u}$.

3.1 The Laplace approximation

The numerator of the exponent in (18),

$$\delta(\boldsymbol{u}|\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{y}) = d_{\boldsymbol{u}}(\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{y}) + \boldsymbol{u}'\boldsymbol{u}, \tag{19}$$

is called the *penalized discrepancy*. We will see that, for the models we are considering, it is relatively straightforward to determine the minimizer of the penalized discrepancy

$$\tilde{\boldsymbol{u}}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{y}) = \arg\min_{\boldsymbol{u}} \delta(\boldsymbol{u}|\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{y}), \tag{20}$$

either directly, as the solution to a penalized least squares problem, or through an iterative algorithm in which each iteration requires the solution of a penalized least squares problem.

For a fixed value of σ^2 the minimizer of the penalized discrepancy maximizes the conditional density

$$f_{\boldsymbol{u}|\boldsymbol{y}}(\boldsymbol{u}|\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y}) = \frac{k(\sigma^2,\boldsymbol{y})e^{-\delta(\boldsymbol{u},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y})/(2\sigma^2)}}{(2\pi\sigma^2)^{q/2}}.$$

That is, $\tilde{\boldsymbol{u}}(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y})$ is the *conditional mode* of $\boldsymbol{u}$ given $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and $\boldsymbol{y}$.

Near the conditional mode the penalized discrepancy has a quadratic approximation

$$\delta(\boldsymbol{u}|\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y}) \approx \delta(\tilde{\boldsymbol{u}}|\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y}) + (\boldsymbol{u}-\tilde{\boldsymbol{u}})'\,\boldsymbol{L}\boldsymbol{L}'\,(\boldsymbol{u}-\tilde{\boldsymbol{u}}) \tag{21}$$

where $\boldsymbol{L}$ is the Cholesky factor of $\frac{1}{2}\nabla^2_{\boldsymbol{u}}\delta(\boldsymbol{u}|\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y})\big|_{\boldsymbol{u}=\tilde{\boldsymbol{u}}}$.

After substituting the quadratic approximation (21) into expression (18) for $L(\boldsymbol{\beta},\boldsymbol{\theta},\sigma^2|\boldsymbol{y})$, the only part of the integrand that depends on $\boldsymbol{u}$ will be the quadratic term in the exponent. To evaluate an integral of the form

$$I = \int_{\boldsymbol{u}} \frac{1}{(2\pi\sigma^2)^{q/2}} \exp\left[\frac{(\boldsymbol{u}-\tilde{\boldsymbol{u}})'\boldsymbol{L}\boldsymbol{L}'(\boldsymbol{u}-\tilde{\boldsymbol{u}})}{-2\sigma^2}\right] d\boldsymbol{u}$$

we change the variable variable of integration from $\boldsymbol{u}$ to

$$\boldsymbol{v} = \boldsymbol{L}'(\boldsymbol{u}-\tilde{\boldsymbol{u}})/\sigma. \tag{22}$$

The determinant of the Jacobian of the transformation (22) is

$$\left|\frac{d\boldsymbol{v}}{d\boldsymbol{u}}\right| = \frac{|\boldsymbol{L}|}{\sigma^q}.$$

and I becomes a multiple of the integral of the standard q-variate normal density

$$I = \int_{\boldsymbol{v}} \frac{e^{-\boldsymbol{v}'\boldsymbol{v}/2}}{(2\pi)^{q/2}} \frac{d\boldsymbol{v}}{|\boldsymbol{L}|} = \frac{1}{|\boldsymbol{L}|}\int_{\boldsymbol{v}} \frac{e^{-\boldsymbol{v}'\boldsymbol{v}/2}}{(2\pi)^{q/2}}\, d\boldsymbol{v} = |\boldsymbol{L}|^{-1}. \tag{23}$$

Expression (23) requires that $\boldsymbol{L}$ is non-singular, which will be the case when $\boldsymbol{L}$ is the Cholesky factor of $\frac{1}{2}\nabla^2_{\boldsymbol{u}}\delta(\boldsymbol{u}|\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y})\big|_{\boldsymbol{u}=\tilde{\boldsymbol{u}}}$.

Returning to expression (18), we can now express the Laplace approximation to the likelihood function or the log-likelihood,

$$\ell(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2|\boldsymbol{y}) = \log L(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2|\boldsymbol{y}), \tag{24}$$

which is more commonly used as the optimization criterion when determining maximum likelihood estimates. (Because the logarithm function is monotonic, the maximizer of the log-likelihood function also maximizes the likelihood function and generally the quadratic approximation to the log-likelihood is much better than the quadratic approximation to the likelihood.)

On the deviance scale (twice the negative log-likelihood) the Laplace approximation is

$$-2\ell(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2|\boldsymbol{y}) \approx -2\log[k(\sigma^2)] + \frac{\delta(\tilde{\boldsymbol{u}}|\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{y})}{\sigma^2} + 2\log|\boldsymbol{L}|. \tag{25}$$

Expression (25) will be an exact expression for the log-likelihood and not just an approximation whenever the penalized discrepancy $\delta(\boldsymbol{u}|\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{y})$ is a quadratic function of $\boldsymbol{u}$.

4 Linear mixed models

A linear mixed model can be expressed as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}\left(\boldsymbol{0}, \sigma^2\boldsymbol{I}\right), \quad \boldsymbol{b} \sim \mathcal{N}\left(\boldsymbol{0}, \sigma^2\boldsymbol{\Sigma}\right), \quad \boldsymbol{\epsilon} \perp \boldsymbol{b} \tag{26}$$

where the symbol $\perp$ denotes independence of random variables. This model implies that the mean of $\boldsymbol{y}$ is the linear predictor, the discrepancy function is the residual sum of squares and the normalizing factor is $(2\pi\sigma^2)^{-n/2}$. That is,

$$\boldsymbol{\mu}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{V}(\boldsymbol{\theta})\boldsymbol{u} \tag{27}$$

$$d(\boldsymbol{\mu}, \boldsymbol{y}) = \|\boldsymbol{\mu} - \boldsymbol{y}\|^2 \tag{28}$$

$$k(\sigma^2) = \left(2\pi\sigma^2\right)^{-n/2}. \tag{29}$$

The penalized discrepancy is

$$\begin{aligned}
\delta(\boldsymbol{u}|\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y}) &= d(\boldsymbol{\mu},\boldsymbol{y}) + \boldsymbol{u}'\boldsymbol{u} \\
&= \|\boldsymbol{\mu}-\boldsymbol{y}\|^2 + \boldsymbol{u}'\boldsymbol{u} \\
&= \|\boldsymbol{V}(\boldsymbol{\theta})\boldsymbol{u} + \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2 + \boldsymbol{u}'\boldsymbol{u} \\
&= \left\| \begin{bmatrix} \boldsymbol{V}(\boldsymbol{\theta}) & \boldsymbol{X} & \boldsymbol{y} \end{bmatrix} \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{\beta} \\ -1 \end{bmatrix} \right\|^2 + \boldsymbol{u}'\boldsymbol{u} \\
&= \begin{bmatrix} \boldsymbol{u}' & \boldsymbol{\beta}' & -1 \end{bmatrix} \begin{bmatrix} \boldsymbol{V}(\boldsymbol{\theta})'\boldsymbol{V}(\boldsymbol{\theta}) + \boldsymbol{I} & \boldsymbol{V}(\boldsymbol{\theta})'\boldsymbol{X} & \boldsymbol{V}(\boldsymbol{\theta})'\boldsymbol{y} \\ \boldsymbol{X}'\boldsymbol{V}(\boldsymbol{\theta}) & \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{y} \\ \boldsymbol{y}'\boldsymbol{V}(\boldsymbol{\theta}) & \boldsymbol{y}'\boldsymbol{X} & \boldsymbol{y}'\boldsymbol{y} \end{bmatrix} \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{\beta} \\ -1 \end{bmatrix}.
\end{aligned} \tag{30}$$

In this form it is obvious that $\delta(\boldsymbol{u}|\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y})$ is a quadratic function of $\boldsymbol{u}$ and that

$$\frac{\nabla^2_{\boldsymbol{u}}\delta(\boldsymbol{u}|\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y})}{2} = \boldsymbol{V}(\boldsymbol{\theta})'\boldsymbol{V}(\boldsymbol{\theta}) + \boldsymbol{I} \tag{31}$$

is positive definite. This expression for $\nabla^2_{\boldsymbol{u}}\delta$ depends on $\boldsymbol{\theta}$ but not on $\boldsymbol{\beta}$ or $\boldsymbol{u}$. Thus the Cholesky factor $\boldsymbol{L}$ required for (25) depends only on $\boldsymbol{\theta}$.

The conditional mode of the orthogonal random effects, $\tilde{\boldsymbol{u}}(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y})$ can be expressed as the solution to

$$\left(\boldsymbol{V}(\boldsymbol{\theta})'\boldsymbol{V}(\boldsymbol{\theta}) + \boldsymbol{I}\right)\tilde{\boldsymbol{u}}(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y}) = \boldsymbol{L}(\boldsymbol{\theta})\boldsymbol{L}(\boldsymbol{\theta})'\tilde{\boldsymbol{u}}(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y}) = \boldsymbol{V}(\boldsymbol{\theta})'\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right) \tag{32}$$

which can easily be calculated from the Cholesky factor $\boldsymbol{L}(\boldsymbol{\theta})$ of $\boldsymbol{V}(\boldsymbol{\theta})'\boldsymbol{V}(\boldsymbol{\theta}) + \boldsymbol{I}$.

We could use $\boldsymbol{L}(\boldsymbol{\theta})$ and $\tilde{\boldsymbol{u}}(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y})$ from (32) to evaluate the log-likelihood using (25), which is exact because the penalized discrepancy is quadratic for linear mixed models. However, we can take advantage of the fact that $\delta(\boldsymbol{u}|\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y})$ is a quadratic function of both $\boldsymbol{u}$ and $\boldsymbol{\beta}$ to minimize δ with respect to $\boldsymbol{u}$ and $\boldsymbol{\beta}$ simultaneously. For a given value of $\boldsymbol{\theta}$ we evaluate the Cholesky factorization

$$\begin{bmatrix} \boldsymbol{V}'\boldsymbol{V} + \boldsymbol{I} & \boldsymbol{V}'\boldsymbol{X} & \boldsymbol{V}'\boldsymbol{y} \\ \boldsymbol{X}'\boldsymbol{V} & \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{y} \\ \boldsymbol{y}'\boldsymbol{V} & \boldsymbol{y}'\boldsymbol{X} & \boldsymbol{y}'\boldsymbol{y} \end{bmatrix} = \begin{bmatrix} \boldsymbol{L} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{L}_{VX} & \boldsymbol{L}_X & \boldsymbol{0} \\ \boldsymbol{\ell}'_{Vy} & \boldsymbol{\ell}'_{Xy} & r \end{bmatrix} \begin{bmatrix} \boldsymbol{L}' & \boldsymbol{L}'_{VX} & \boldsymbol{\ell}_{Vy} \\ \boldsymbol{0} & \boldsymbol{L}'_X & \boldsymbol{\ell}_{Xy} \\ \boldsymbol{0} & \boldsymbol{0} & r \end{bmatrix} \tag{33}$$

where $\boldsymbol{\ell}_{Vy}$ and $\boldsymbol{\ell}_{Xy}$ are column vectors of dimensions q and p, respectively. The lower right element, r, is a scalar.

With the factorization (33) we can write

$$\begin{aligned}\delta(\boldsymbol{u}|\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y}) &= \left\|\begin{bmatrix}\boldsymbol{L}' & \boldsymbol{L}'_{VX} & \boldsymbol{\ell}_{Vy}\\ \mathbf{0} & \boldsymbol{L}'_X & \boldsymbol{\ell}_{Xy}\\ \mathbf{0} & \mathbf{0} & r\end{bmatrix}\begin{bmatrix}\boldsymbol{u}\\ \boldsymbol{\beta}\\ -1\end{bmatrix}\right\|^2\\ &= r^2 + \|\boldsymbol{L}'_X\boldsymbol{\beta} - \boldsymbol{\ell}_{Xy}\|^2 + \|\boldsymbol{L}'\boldsymbol{u} + \boldsymbol{L}'_{VX}\boldsymbol{\beta} - \boldsymbol{\ell}_{Vy}\|^2\\ &= r^2 + \left\|\boldsymbol{L}_X(\boldsymbol{\theta})'\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})\right)\right\|^2 + \|\boldsymbol{L}(\boldsymbol{\theta})'\left(\boldsymbol{u} - \widehat{\boldsymbol{u}}(\boldsymbol{\theta})\right)\|^2\end{aligned} \tag{34}$$

where $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$, the conditional estimate of $\boldsymbol{\beta}$ given $\boldsymbol{\theta}$, and $\widehat{\boldsymbol{u}}(\boldsymbol{\theta})$, the conditional mode of $\boldsymbol{u}$ given $\boldsymbol{\theta}$ and $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$, are the solutions to

$$\boldsymbol{L}_X(\boldsymbol{\theta})'\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \boldsymbol{\ell}_{Xy}(\boldsymbol{\theta}) \tag{35}$$

$$\boldsymbol{L}(\boldsymbol{\theta})'\widehat{\boldsymbol{u}}(\boldsymbol{\theta}) = \boldsymbol{\ell}_{Vy}(\boldsymbol{\theta}) - \boldsymbol{L}_{VX}(\boldsymbol{\theta})'\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}). \tag{36}$$

Furthermore, the minimum of the penalized discrepancy, conditional on $\boldsymbol{\theta}$, is

$$\min_{\boldsymbol{u}} \delta(\boldsymbol{u}|\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}),\boldsymbol{\theta},\boldsymbol{y}) = r^2(\boldsymbol{\theta}). \tag{37}$$

The deviance function, $-2\ell(\boldsymbol{\beta},\boldsymbol{\theta},\sigma^2|\boldsymbol{y})$, at the conditional estimate, $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$, is

$$-2\ell(\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}),\boldsymbol{\theta},\sigma^2|\boldsymbol{y}) = n\log\left(2\pi\sigma^2\right) + \frac{r^2(\boldsymbol{\theta})}{\sigma^2} + 2\log|\boldsymbol{L}(\boldsymbol{\theta})|. \tag{38}$$

Differentiating $-2\ell(\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}),\boldsymbol{\theta},\sigma^2|\boldsymbol{y})$ as a function of σ^2 and setting the derivative to zero provides the conditional estimate

$$\widehat{\sigma^2}(\boldsymbol{\theta}) = \frac{r^2(\boldsymbol{\theta})}{n}. \tag{39}$$

Substituting this estimate into (38) provides the *profiled deviance* function

$$\begin{aligned}-2\ell(\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}),\boldsymbol{\theta},\widehat{\sigma^2}(\boldsymbol{\theta})|\boldsymbol{y}) &= n\log\left(\frac{2\pi r^2(\boldsymbol{\theta})}{n}\right) + n + 2\log|\boldsymbol{L}(\boldsymbol{\theta})|\\ &= n\left[1 + \log\left(\frac{2\pi}{n}\right)\right] + n\log r^2(\boldsymbol{\theta}) + 2\log|\boldsymbol{L}(\boldsymbol{\theta})|\end{aligned}. \tag{40}$$

That is, the maximum likelihood estimate (mle) of $\boldsymbol{\theta}$ is

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} n\left[1 + \log\left(\frac{2\pi}{n}\right)\right] + n\log r^2(\boldsymbol{\theta}) + 2\log|\boldsymbol{L}(\boldsymbol{\theta})|. \tag{41}$$

The mle's of the other parameters are determined from $\widehat{\boldsymbol{\theta}}$ using (39) and (35). The conditional modes of the orthogonal random effects, $\widehat{\boldsymbol{u}}(\widehat{\boldsymbol{\theta}})$, evaluated using (36), and the corresponding conditional modes of the untransformed random effects,

$$\widehat{\boldsymbol{b}}(\widehat{\boldsymbol{\theta}}) = \boldsymbol{T}(\widehat{\boldsymbol{\theta}})\boldsymbol{S}(\widehat{\boldsymbol{\theta}})\widehat{\boldsymbol{u}}(\widehat{\boldsymbol{\theta}}), \tag{42}$$

are called the *Best Linear Unbiased Predictor* (BLUPs) of the random effects.

The three terms in the objective function being minimized in (41) are, respectively, a constant, $n\,[1 + \log(2\pi/n)]$, a measure of the fidelity of the fitted values to the observed data, $n \log r^2(\widehat{\boldsymbol{\theta}})$, and a measure of model complexity, $2\log|\boldsymbol{L}(\widehat{\boldsymbol{\theta}})|$. Thus we can consider maximum likelihood estimation of the parameters in a linear mixed model to be balancing fidelity to the data against model complexity.

4.1 REML estimates

The maximum likelihood estimate of σ^2, $\widehat{\sigma^2} = r^2/n$, is the penalized residual sum of squares divided by the number of observations. It has a form like the maximum likelihood estimate of the variance from a single sample, $\widehat{\sigma^2} = \sum_{i=1}^{n}(y_i - \bar{y})^2/n$ or the maximum likelihood estimate of the variance in a linear regression model with p coefficients in the predictor, $\widehat{\sigma^2} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2/n$.

Generally these variance estimates are not used because they are biased downward. This is, on average they will underestimate the variance in the model. Instead we use $\widehat{\sigma^2}_R = \sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1)$ for the variance estimate from a single sample or $\widehat{\sigma^2}_R = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-p)$ for the variance estimate in a linear regression model. These estimates are based on the residuals, $y_i - \hat{y}_i, i = 1, \ldots, n$ which satisfy p linear constraints and thus are constrained to an $(n-p)$-dimensional subspace of the n-dimensional sample space. In other words, the residuals have only $n - p$ degrees of freedom.

In a linear mixed model we often prefer to estimate the variance components, σ^2 and $\boldsymbol{\Sigma}$, according to the *residual maximum likelihood* (REML) criterion (sometimes called the *restricted* maximum likelihood criterion) which compensates for the estimation of the fixed-effects parameters when estimating the random effects.

The REML criterion can be expressed as

$$\begin{aligned} L_R(\boldsymbol{\theta}, \sigma^2|\boldsymbol{y}) &= \int_{\boldsymbol{\beta}} L(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2|\boldsymbol{y})\, d\boldsymbol{\beta} \\ &= \frac{e^{-r^2/(2\sigma^2)}}{|\boldsymbol{L}|(2\pi\sigma^2)^{(n-p)/2}} \int_{\boldsymbol{\beta}} \frac{e^{-(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})'\boldsymbol{L}_X\boldsymbol{L}_X'(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})/(2\sigma^2)}}{(2\pi\sigma^2)^{p/2}}\, d\boldsymbol{\beta} \\ &= \frac{e^{-r^2/(2\sigma^2)}}{|\boldsymbol{L}||\boldsymbol{L}_X|(2\pi\sigma^2)^{(n-p)/2}} \end{aligned} \tag{43}$$

or, on the deviance scale,

$$-2\ell_R(\boldsymbol{\theta}, \sigma^2|\boldsymbol{y}) = (n-p)\log\left(2\pi\sigma^2\right) + \frac{r^2(\boldsymbol{\theta})}{\sigma^2} + 2|\boldsymbol{L}(\boldsymbol{\theta})| + 2|\boldsymbol{L}_X(\boldsymbol{\theta})| \tag{44}$$

from which we can see that the REML estimate of σ^2 is

$$\widehat{\sigma}_R(\boldsymbol{\theta}) = \frac{r^2(\boldsymbol{\theta})}{n-p} \tag{45}$$

and the profiled REML deviance is

$$\begin{aligned} &-2\ell_R(\boldsymbol{\theta}, \widehat{\sigma}^2(\boldsymbol{\theta})|\boldsymbol{y}) = \\ &\quad (n-p)\left[1 + \log\left(\frac{2\pi}{n-p}\right)\right] + (n-p)\log r^2 + 2\log|\boldsymbol{L}| + 2\log|\boldsymbol{L}_X| \end{aligned} \tag{46}$$

5 Nonlinear mixed models

Like the linear mixed model, the nonlinear mixed model is based on a multivariate normal (or Gaussian) distribution of the response $\boldsymbol{y}$ given $\boldsymbol{\mu}$. That is,

$$\boldsymbol{y} = \boldsymbol{\mu}(\boldsymbol{\beta}, \boldsymbol{b}) + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}), \quad \boldsymbol{b} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{\Sigma}(\boldsymbol{\theta}), \quad \boldsymbol{\epsilon} \perp \boldsymbol{b} \tag{47}$$

and the discrepancy function and normalizing factor are the same as for the linear mixed model

$$d(\boldsymbol{\mu}, \boldsymbol{y}) = \|\boldsymbol{\mu} - \boldsymbol{y}\|^2 \tag{48}$$

$$k(\sigma^2) = \left(2\pi\sigma^2\right)^{-n/2}. \tag{49}$$

The mean, $\boldsymbol{\mu}$, however, is no longer equal to the linear predictor. Each element of $\boldsymbol{\mu}$ is the value of a nonlinear model function $g(\boldsymbol{x}, \boldsymbol{\phi})$ that depends on covariates, $\boldsymbol{x}$, and a nonlinear model parameter, $\boldsymbol{\phi}$, of length s. When fitting a model the values of the covariates at each observation are known so we can regard $\boldsymbol{\mu}_i$ as a function of $\boldsymbol{\phi}_i$ only and write

$$\boldsymbol{\mu} = \boldsymbol{g}(\boldsymbol{\Phi}) \tag{50}$$

where $\boldsymbol{\Phi}$ is the $n \times s$ matrix whose ith row is $\boldsymbol{\phi}_i, i = 1, \ldots, n$ and the vector-valued function $\boldsymbol{g}$ applies the scalar function g rowwise to $\boldsymbol{\Phi}$ using covariates $\boldsymbol{x}_i, i = 1, \ldots, n$ for the ith row.

The linear predictor determines $\boldsymbol{\Phi}$ as

$$\text{vec}(\boldsymbol{\Phi}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{V}(\boldsymbol{\theta})\boldsymbol{u} \tag{51}$$

where the vec operator concatenates the columns of $\boldsymbol{\Phi}$ to form a vector of length $m = ns$. Thus the matrix $\boldsymbol{X}$ is $ns \times p$ while $\boldsymbol{Z}$ and $\boldsymbol{V}$ are $ns \times q$.

5.1 Optimizing the penalized discrepancy

As for a linear mixed model, the problem of determining $\tilde{\boldsymbol{u}}(\boldsymbol{\beta}, \boldsymbol{\theta})$, the optimizer of the penalized discrepancy function, can be written as a penalized least squares problem

$$\tilde{\boldsymbol{u}}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \arg\min_{\boldsymbol{u}} \delta(\boldsymbol{u}|\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{y}) = \arg\min_{\boldsymbol{u}} \left[\|\boldsymbol{\mu}(\boldsymbol{\beta}, \boldsymbol{u}) - \boldsymbol{y}\|^2 + \boldsymbol{u}'\boldsymbol{u} \right]. \tag{52}$$

Unlike the case of the linear mixed model this is generally a penalized nonlinear least squares problem that requires an iterative solution.

Given $\boldsymbol{u}^{(i)}$ (the parenthesized superscripts denote the number of the iteration at which a quantity is evaluated) we evaluate

$$\begin{aligned} \left.\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{u}'}\right|_{\boldsymbol{u}=\boldsymbol{u}^{(i)}} &= \boldsymbol{M}^{(i)} \\ &= \begin{bmatrix} \boldsymbol{I} & \boldsymbol{I} & \ldots & \boldsymbol{I} \end{bmatrix} \text{diag}\left(\text{vec}\left.\frac{d\boldsymbol{\mu}}{d\boldsymbol{\Phi}}\right|_{\boldsymbol{\Phi}=\boldsymbol{\Phi}^{(i)}} \right) \boldsymbol{V}, \end{aligned} \tag{53}$$

where the matrix on the left is the horizontal concatenation of s copies of the $n \times n$ identity matrix. The proposed updated vector of orthogonal random

effects, $\boldsymbol{u}^{(i+1)}$, minimizes the approximate penalized discrepancy

$$\begin{aligned}\boldsymbol{u}^{(i+1)} &= \arg\min_{\boldsymbol{u}} \left\|\boldsymbol{y} - \boldsymbol{\mu}\left(\boldsymbol{\beta}, \boldsymbol{u}^{(i)}\right) - \boldsymbol{M}^{(i)}\left(\boldsymbol{u} - \boldsymbol{u}^{(i)}\right)\right\|^2 + \boldsymbol{u}'\boldsymbol{u} \\ &= \arg\min_{\boldsymbol{u}} \left\|\boldsymbol{y} - \boldsymbol{\mu}^{(i)} + \boldsymbol{M}^{(i)}\boldsymbol{u}^{(i)} - \boldsymbol{M}^{(i)}\boldsymbol{u}\right\|^2 + \boldsymbol{u}'\boldsymbol{u} \\ &= \arg\min_{\boldsymbol{u}} \left\|\begin{bmatrix}\boldsymbol{y} - \boldsymbol{\mu}^{(i)} + \boldsymbol{M}^{(i)}\boldsymbol{u}^{(i)} \\ \boldsymbol{0}\end{bmatrix} - \begin{bmatrix}\boldsymbol{M}^{(i)} \\ \boldsymbol{I}\end{bmatrix}\boldsymbol{u}\right\|^2.\end{aligned} \tag{54}$$

That is, $\boldsymbol{u}^{(i+1)}$ is the solution to a linear least squares problem for which the normal equations are

$$\left(\boldsymbol{M}^{(i)'}\boldsymbol{M}^{(i)} + \boldsymbol{I}\right)\boldsymbol{u}^{(i+1)} = \boldsymbol{L}^{(i)}\boldsymbol{L}^{(i)'}\boldsymbol{u}^{(i+1)} = \boldsymbol{M}^{(i)'}\left(\boldsymbol{y} - \boldsymbol{\mu}^{(i)} + \boldsymbol{M}^{(i)}\boldsymbol{u}^{(i)}\right). \tag{55}$$

At convergence the Laplace approximation to the deviance is

$$-2\ell(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2|\boldsymbol{y}) = n\log\left(2\pi\sigma^2\right) + \frac{\delta(\tilde{\boldsymbol{u}}|\boldsymbol{\theta}, \boldsymbol{\beta})}{\sigma^2} + 2\log|\boldsymbol{L}(\boldsymbol{\beta}, \boldsymbol{\theta})| \tag{56}$$

where $\boldsymbol{L}(\boldsymbol{\beta}, \boldsymbol{\theta})$ is the Cholesky factor of $\boldsymbol{M}'\boldsymbol{M} + \boldsymbol{I}$ evaluated at $\tilde{\boldsymbol{u}}(\boldsymbol{\beta}, \boldsymbol{\theta})$, $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. As for the linear mixed model we can form the conditional estimate of σ^2

$$\widehat{\sigma^2}(\boldsymbol{\theta}, \boldsymbol{\beta}) = \frac{\delta(\tilde{\boldsymbol{u}}|\boldsymbol{\theta}, \boldsymbol{\beta})}{n}. \tag{57}$$

Substituting this estimate into (56) produces the Laplace approximation to the profiled deviance

$$-2\ell(\boldsymbol{\beta}, \boldsymbol{\theta}, \widehat{\sigma^2}(\boldsymbol{\beta}, \boldsymbol{\theta})|\boldsymbol{y}) = n\left[1 + \log\left(\frac{2\pi}{n}\right)\right] + n\log\delta(\tilde{\boldsymbol{u}}|\boldsymbol{\theta}, \boldsymbol{\beta}) + 2\log|\boldsymbol{L}|. \tag{58}$$

5.2 Constructing model matrices for nonlinear mixed models

In our previous example involving three measurements at times 1, 2 and 3 on each of five subjects, the conditional mean $\boldsymbol{\mu}(\boldsymbol{\beta}, \boldsymbol{b})$ was linear in the parameters $\boldsymbol{\beta}$ and the in random effects $\boldsymbol{b}$ and also linear with respect to time. Suppose instead that we felt that the trajectory of each subject's response with respect to time was more appropriately modelled as

$$\boldsymbol{\phi}_1\left(1 - e^{-\boldsymbol{\phi}_2 x_{i,j}}\right) \quad i = 1, \ldots, 5;\ j = 1, \ldots, 3 \tag{59}$$

where $x_{i,j}$ is the time of the jth observation on the ith subject while $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ are subject-specific parameters representing the asymptotic value for subject i (i.e. the value predicted for large values of the time, x) and the rate constant for subject i, respectively.

The model formula used in the `nlmer` function is a three-part formula in which the left hand side determines the response, the middle part is the expression of the nonlinear model involving the parameters $\boldsymbol{\phi}$ and any covariates and the right hand side is a mixed model formula that can (in fact, must) involve the names of parameters from the nonlinear model.

In our example, if subject-specific parameters are modelled as population means, $\boldsymbol{\beta} = [\beta_1, \beta_2]\prime$ plus a subject-specific random effect for each parameter, allowing for correlation of the random effects within each subject, the formula would be written

```
y ~ A * (1 - exp(-rc * time)) ~ (A + rc | subj)
```

The `vec` of the 15×2 parameter matrix $\boldsymbol{\Phi}$ is a vector of length 30 where the first 15 elements are values of `A` and the last 15 elements are values of `rc`. In the mixed-model formula the names `A` and `rc` represent indicator variables for the first 15 and the last 15 positions, respectively. In the general case of a nonlinear model with s parameters there will be s indicator variables named according to the model parameters and determining the positions in $\mathrm{vec}(\boldsymbol{\Phi})$ that correspond to each parameter.

For the model matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ the implicit intercept term generated by the standard S language rules for model matrices would not make sense. The intercept term is suppressed in the random-effects terms and is replaced by the sum of the parameter name indicators in the fixed-effects terms. Thus the formula shown above is equivalent to

```
y ~ A * (1 - exp(-rc * time)) ~ A + rc + (0 + A + rc | subj)
```

The matrix $\boldsymbol{X}$ will be 30×2 with the two columns being the indicator for `A` and the indicator for `rc`.

5.3 Random effects for conditionally linear parameters only

There is a special case of a nonlinear mixed model where the Laplace approximation is the deviance and where the iterative algorithm to determine $\tilde{\boldsymbol{u}}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{y})$ will converge in one iteration. Frequently some of the elements of

the parameter vector $\boldsymbol{\phi}$ occur linearly in the nonlinear model $g(\boldsymbol{x}, \boldsymbol{\phi})$. These elements are said to be *conditionally linear* parameters because, conditional on the values of the other parameters, the model function is a linear function of these.

If the random effects determine only conditionally linear parameters then $\boldsymbol{\mu}$ is linear in $\boldsymbol{u}$ and the matrix $\boldsymbol{M}$ depends on $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ but not on $\boldsymbol{u}$. We can rewrite the mean function as

$$\boldsymbol{\mu}(\boldsymbol{\beta}, \boldsymbol{u}) = \boldsymbol{\mu_0}(\boldsymbol{\beta}) + \boldsymbol{M}\boldsymbol{u} \tag{60}$$

where $\boldsymbol{\mu_0}(\boldsymbol{\beta}) = \boldsymbol{\mu}(\boldsymbol{\beta}, \boldsymbol{0}) = \boldsymbol{\mu}(\boldsymbol{X}\boldsymbol{\beta})$. The penalized least squares problem (54) for the updated $\boldsymbol{u}$ can be rewritten as

$$\tilde{\boldsymbol{u}}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{y}) = \min_{\boldsymbol{u}} \left\| \begin{bmatrix} \boldsymbol{y} - \boldsymbol{\mu_0}(\boldsymbol{\beta}) \\ \boldsymbol{0} \end{bmatrix} - \begin{bmatrix} \boldsymbol{M} \\ \boldsymbol{I} \end{bmatrix} \boldsymbol{u} \right\|^2 . \tag{61}$$

That is, $\tilde{\boldsymbol{u}}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{y})$ is the solution to

$$(\boldsymbol{M}'\boldsymbol{M} + \boldsymbol{I})\,\tilde{\boldsymbol{u}} = \boldsymbol{M}'(\boldsymbol{y} - \boldsymbol{\mu_0}(\boldsymbol{\beta})) \tag{62}$$

6 Generalized linear mixed models

A generalized linear mixed model differs from a linear mixed model in the form of the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{\beta}$, $\boldsymbol{b}$ and, possibly, σ^2, which determines the discrepancy function $d(\boldsymbol{\mu}, \boldsymbol{y})$, and in the mapping from the linear predictor, $\boldsymbol{\eta}$, to the conditional mean, $\boldsymbol{\mu}$. This mapping between $\boldsymbol{\eta}$ and $\boldsymbol{\mu}$ is assumed to be one-to-one and to enforce any constraints on the elements of $\boldsymbol{\mu}$, such as the mean of a Bernoulli or binomial random variable being in the range $0 \le \{\boldsymbol{\mu}\}_k \le 1, k = 1, \ldots, n$ or the mean of a Poisson random variable being positive. By convention, it is the mapping from $\boldsymbol{\mu}$ to $\boldsymbol{\eta} = \boldsymbol{g}(\boldsymbol{\mu})$ that is called the *link function*, so the inverse mapping, $\boldsymbol{\mu} = \boldsymbol{g}^{-1}(\boldsymbol{\eta})$, is called the *inverse link*.

Although we have written the link and the inverse link as functions of vectors, they are defined in terms of scalar functions, so that

$$\begin{aligned} \eta_k &= \{\boldsymbol{\eta}\}_k = \{\boldsymbol{g}(\boldsymbol{\eta})\}_k = g(\{\boldsymbol{\eta}\}) = g(\mu_k) \quad k = 1, \ldots, n \\ \mu_k &= \{\boldsymbol{\mu}\}_k = \left\{\boldsymbol{g}^{-1}(\boldsymbol{\mu})\right\}_k = g^{-1}(\{\boldsymbol{\mu}\}) \quad k = 1, \ldots, n. \end{aligned} \tag{63}$$

where $g(\mu)$ and $g^{-1}(\eta)$ are the scalar link and inverse link functions, respectively. Furthermore, the elements of $\boldsymbol{y}$ are assumed to be conditionally

independent, given $\boldsymbol{\mu}$, and for $k = 1, \ldots, n$ the distribution of y_k depends only on μ_k and, possibly, $sigma^2$. That is, the discrepancy function can be written

$$d(\boldsymbol{\mu}, \boldsymbol{y}) = \sum_{k=1}^{n} r_D^2(\mu_k, y_k) \tag{64}$$

where r_D is the *deviance residual* function. For many models the discrepancy defines

6.1 Examples of deviance residual and link functions

If the $y_k, k = 1, \ldots, n$ are binary responses (i.e. each y_k is either 0 or 1) and they are conditionally independent given $\boldsymbol{\mu}$, then the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{\mu}$ has probability mass function

$$f_{\boldsymbol{y}|\boldsymbol{\mu}}(\boldsymbol{y}, \boldsymbol{\mu}) = \prod_{k=1}^{n} \mu_k^{y_k} \left(1 - \mu_k\right)^{(1-y_k)} \tag{65}$$

Because the distribution of y_k is completely determined by μ_k there is no need for a separate scale factor, σ^2, and expression (2) for the conditional density in terms of the discrepancy can be written

$$f_{\boldsymbol{y}|\boldsymbol{\mu}}(\boldsymbol{y}|\boldsymbol{\mu}) = ke^{-d(\boldsymbol{\mu},\boldsymbol{y})/2}. \tag{66}$$

Thus the discrepancy function must be

$$d(\boldsymbol{\mu}, \boldsymbol{y}) = \tag{67}$$

References

John M. Chambers and Trevor J. Hastie. *Statistical Models in S*. Chapman & Hall, London, 1992.

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