Computational methods for mixed models

Douglas Bates
Department of Statistics
University of Wisconsin – Madison

July 3, 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 of 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 exactly.

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 variance-covariance of the random effects. As described in §2.1, random-effects terms in the model formula always include the vertical bar symbol, '`|`', which is sometimes read as "given" or "by". Any terms that do not include this symbol are fixed-effects terms.

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

The form of $\boldsymbol{Z}$ and the rules for constructing it from the data and the random-effects terms in the model formula are described in §2.1.

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 $\boldsymbol{Z}$, 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 parameters in the underlying nonlinear model, as described in §5.

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*

$$\boldsymbol{\eta}_{\boldsymbol{b}}(\boldsymbol{b}, \boldsymbol{\beta}) = \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{X}\boldsymbol{\beta} \tag{1}$$

The notation $\boldsymbol{\eta}_{\boldsymbol{b}}$ emphasizes that $\boldsymbol{\eta}$ is being expressed as a function of $\boldsymbol{b}$ (and $\boldsymbol{\beta}$). In §2.4 we will define orthogonal random effects, $\boldsymbol{u}$, which are a linear transformation of $\boldsymbol{b}$, and the corresponding expression, $\boldsymbol{\eta}_{\boldsymbol{u}}$, for the linear predictor.

The vector $\boldsymbol{\beta}$ is a parameter of the model. Strictly speaking, the vector $\boldsymbol{b}$ is not a parameter — it is a value of the unobserved random variable, $\boldsymbol{B}$. The observed responses, $\boldsymbol{y}$, are the value of the n-dimensional random variable, $\boldsymbol{Y}$. In the models we will consider, $\boldsymbol{b}$ and $\boldsymbol{\beta}$ determine the *conditional mean*, $\boldsymbol{\mu}_{\boldsymbol{Y}|\boldsymbol{B}}$, of $\boldsymbol{Y}$ through the linear predictor, $\boldsymbol{\eta}_{\boldsymbol{b}}(\boldsymbol{b}, \boldsymbol{\beta})$. That is,

$$\mathsf{E}[\boldsymbol{Y}|\boldsymbol{B}] = \boldsymbol{\mu}_{\boldsymbol{Y}|\boldsymbol{B}} = \boldsymbol{\mu}(\boldsymbol{\eta}_{\boldsymbol{b}}(\boldsymbol{b}, \boldsymbol{\beta})) = \boldsymbol{\mu}(\boldsymbol{Z}\boldsymbol{b} + \boldsymbol{X}\boldsymbol{\beta}). \tag{2}$$

The random variable $\boldsymbol{Y}$ can be continuous or discrete. In both cases we will write the conditional distribution as $f_{\boldsymbol{Y}|\boldsymbol{B}}$, representing the condition probability density or the conditional probability mass function, whichever is appropriate. This conditional distribution depends on the conditional mean, $\boldsymbol{\mu}_{\boldsymbol{Y}|\boldsymbol{B}}$, only through a *discrepancy function*, $d(\boldsymbol{\mu}_{\boldsymbol{Y}|\boldsymbol{B}}, \boldsymbol{y})$, that defines a "squared distance" between the conditional mean, $\boldsymbol{\mu}_{\boldsymbol{Y}|\boldsymbol{B}}$, and the observed data, $\boldsymbol{y}$. For linear mixed models and for nonlinear mixed models $d(\boldsymbol{\mu}_{\boldsymbol{Y}|\boldsymbol{B}}, \boldsymbol{y})$ is precisely the square of the Euclidean distance, $d(\boldsymbol{\mu}_{\boldsymbol{Y}|\boldsymbol{B}}, \boldsymbol{y}) = \left\|\boldsymbol{\mu}_{\boldsymbol{Y}|\boldsymbol{B}} - \boldsymbol{y}\right\|^2$. The more general form of the discrepancy function used for generalized linear mixed models is described in §6.

The discrepancy function is related to the *deviance* between the observed data and a conditional mean, in that the deviance is the discrepancy for the current model minus the discrepancy for what is called "the full model" (see (McCullagh and Nelder, 1989, §2.3) for details). For linear mixed models and for nonlinear mixed models the discrepancy for the full model is zero so that the deviance and the discrepancy coincide. For generalized linear mixed models the discrepancy for the full model can be nonzero. The deviance is defined in such a way that it must be non-negative and does, in fact, behave like a squared distance. The discrepancy, on the other hand, can be negative.

Because we will primarily be concerned with minimizing the discrepancy (or a related quantity, the penalized discrepancy, defined in §2.4), the additive term that distinguishes the discrepancy from the deviance is not important to us.

The conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{B}$ is completely determined by the conditional mean, $\boldsymbol{\mu}_{\boldsymbol{Y}|\boldsymbol{B}}$, and, possibly, a variance scale parameter that in part determines the conditional variance, $\mathsf{Var}(\boldsymbol{Y}|\boldsymbol{B})$, but does not affect the

conditional mean, $\boldsymbol{\mu}_{\boldsymbol{Y}|\boldsymbol{B}}$. We will write this variance scale parameter, when it is used, as σ^2.

The general form of the conditional distribution is

$$f_{\boldsymbol{Y}|\boldsymbol{B}}(\boldsymbol{y}|\boldsymbol{b},\boldsymbol{\beta},\sigma^2) = k(\boldsymbol{y},\sigma^2)e^{-d\left(\boldsymbol{\mu}_{\boldsymbol{Y}|\boldsymbol{B}},\boldsymbol{y}\right)/\left(2\sigma^2\right)}. \tag{3}$$

The quantity $k(\boldsymbol{y},\sigma^2)$ is the *normalizing factor* defined so that the $\int_{\boldsymbol{y}} f_{\boldsymbol{Y}|\boldsymbol{B}}(\boldsymbol{y}|\boldsymbol{b},\boldsymbol{\beta},\sigma^2)\,d\boldsymbol{y} = 1$.

In the models we will consider, the elements of $\boldsymbol{Y}$ are conditionally independent, given $\boldsymbol{B}$. That is, $f_{\boldsymbol{Y}|\boldsymbol{B}}$ can be written as a product of n factors, each involving just one element of $\boldsymbol{y}$ and the corresponding element of $\boldsymbol{\mu}_{\boldsymbol{Y}|\boldsymbol{B}}$. Consequently, the discrepancy can be written as a sum of n terms that also are evaluated component-wise on $\boldsymbol{\mu}_{\boldsymbol{Y}|\boldsymbol{B}}$ and $\boldsymbol{y}$.

For linear and generalized linear mixed models, where $\boldsymbol{\mu}$ and $\boldsymbol{\eta}$ both have dimension n, the *mean function*, $\boldsymbol{\mu}(\boldsymbol{\eta})$, is also evaluated component-wise. That is, the ith element of $\boldsymbol{\mu}_{\boldsymbol{Y}|\boldsymbol{B}}$ depends only on the ith element of $\boldsymbol{\eta}$. For nonlinear mixed models, the dimension of $\boldsymbol{\eta}$ is a multiple, $m = ns$, of the dimension of $\boldsymbol{\mu}$. If we convert $\boldsymbol{\eta}$ to an $n \times s$ matrix (using, say, column-major ordering), then the ith element of $\boldsymbol{\mu}$ depends only on the ith row of this matrix.

Generalized linear mixed models where the independent components of the conditional distribution, $f_{\boldsymbol{Y}|\boldsymbol{B}}$, have Bernoulli distributions or binomial distributions or Poisson distributions, do not require a separate scale parameter, σ^2, for the variance. The underlying scalar distributions are completely determined by their means. In such cases the conditional distribution, $f_{\boldsymbol{Y}|\boldsymbol{B}}$, can be written $f_{\boldsymbol{Y}|\boldsymbol{B}}(\boldsymbol{y}|\boldsymbol{b},\boldsymbol{\beta}) = k(\boldsymbol{y})e^{-d(\boldsymbol{\mu}_{\boldsymbol{Y}|\boldsymbol{B}}(\boldsymbol{b},\boldsymbol{\beta})),\boldsymbol{y})/2}$ and the normalization factor is $k(\boldsymbol{y})$.

The marginal distribution of $\boldsymbol{B}$ is the multivariate Gaussian (or "normal") distribution

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

where σ^2 is the same variance scale parameter used in (3). The $q \times q$ symmetric, positive-semidefinite matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is the *relative variance-covariance matrix* of $\boldsymbol{B}$. The form of $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ and the parameter, $\boldsymbol{\theta}$, that determines it are described in §2.2. (The condition that $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is positive-semidefinite means that $\boldsymbol{v}'\boldsymbol{\Sigma}(\boldsymbol{\theta})\boldsymbol{v} \geq 0, \forall \boldsymbol{v} \in \mathbb{R}^q$.)

2.1 Random-effects terms

A simple random-effects term is of the form '`(`*`formula`*`|`*`factor`*`)`' where *formula* is a linear model formula and *`factor`* is an expression that can be evaluated as a factor. This factor is called the *grouping factor* for the term because it partitions the elements of the conditional mean, $\boldsymbol{\mu}_{\boldsymbol{Y}|\boldsymbol{B}}$, into non-overlapping groups and isolates the effect of some elements of the random effects vector, $\boldsymbol{b}$, to a specific group.

A random-effects term is typically enclosed in parentheses so that the extent of *`formula`* is clearly defined. As stated earlier, it is the presence of the vertical bar, '`|`', that distinguishes a random-effects term from a fixed-effects term.

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 an $m \times q_i$ 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 an *indicator interaction matrix*, $\tilde{\boldsymbol{Z}}_i$, which is the horizontal concatenation of q_i matrices, each representing the interaction of the indicators of $\boldsymbol{f}_i$ with a column of $\boldsymbol{Z}_i$. 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 nonzeros in the jth vertical block in $\tilde{\boldsymbol{Z}}_i$ (exactly one nonzero per row) correspond to the jth column of $\boldsymbol{Z}_i$.

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

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

In the not-uncommon case of a random effects term of the form `(1|`*`factor`*`)`, where the formula '`1`' designates the "Intercept" column only, $q_i = 1$, $\boldsymbol{Z}_i = \mathbf{1}_m$, the $m \times 1$ matrix all of whose elements are unity, and $\tilde{\boldsymbol{Z}}_i$ becomes the $m \times n_i$ matrix of indicators of the levels of $\boldsymbol{f}_i$.

For example, suppose we wish to model data where three observations have been recorded on each of four subjects. A data frame containing just a "subject" factor, `subj`, could be constructed as

```
> dat <- data.frame(subj = gl(4, 3, labels = LETTERS[1:4]))
```

The first few rows of `dat` are

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

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

and a summary of the structure of `dat` is

```
> str(dat)
```

```
'data.frame':        12 obs. of  1 variable:
 $ subj: Factor w/ 4 levels "A","B","C","D": 1 1 1 2 2 2 3 3 3 4 ...
```

The 12×1 model matrix $\boldsymbol{Z}_i$ for the random-effects term `(1|subj)` can be generated and stored (as `Zi`) by

```
> Zi <- model.matrix(~1, dat)
```

The transpose of `Zi` is

```
> t(Zi)
```

```
            [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
(Intercept)    1    1    1    1    1    1    1    1    1     1
            [,11] [,12]
(Intercept)     1     1
attr(,"assign")
[1] 0
```

and the corresponding indicator interaction matrix, $\tilde{\boldsymbol{Z}}_i$, is

```
12 x 4 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
```

As stated earlier, $\tilde{\boldsymbol{Z}}_i$ for the random-effects term `(1|subj)` is simply the matrix of indicator columns for the levels of `subj`.

In the lme4 package the transposes of sparse model matrices like $\tilde{\boldsymbol{Z}}_i$ are stored as compressed column matrices (Davis, 2006, ch. 2) of class `"dgCMatrix"`. When a matrix of this class is printed, the systematic zeros are shown as '.'.

The transpose of the indicator matrix can be generated by coercing the factor to the virtual class `"sparseMatrix"`

```
> as(dat$subj, "sparseMatrix")
```

```
4 x 12 sparse Matrix of class "dgCMatrix"

A 1 1 1 . . . . . . . . .
B . . . 1 1 1 . . . . . .
C . . . . . . 1 1 1 . . .
D . . . . . . . . . 1 1 1
```

This display shows explicitly that rows of the transpose of the indicator matrix are associated with levels of the grouping factor.

For a more general example, assume that each subject is observed at times 1, 2 and 3. We can insert a `time` variable in the data frame as

```
> dat$time <- rep(1:3, 4)
```

so the first few rows of the data frame become

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

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

The term `(time|Subject)` (which is equivalent to `(1+time|Subject)` because linear model formulas have an implicit intercept term) generates a model matrix, $\boldsymbol{Z}_i$, with $q_i = 2$ columns, and whose first few rows are

```
> head(Zi <- model.matrix(~time, dat), n = 5)
```

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

The transpose of the indicator interaction matrix could be constructed as

```
> tt <- ii <- as(dat$subj, "sparseMatrix")
> tt@x <- as.numeric(dat$time)
> rBind(ii, tt)

8 x 12 sparse Matrix of class "dgCMatrix"

A 1 1 1 . . . . . . . . .
B . . . 1 1 1 . . . . . .
C . . . . . . 1 1 1 . . .
D . . . . . . . . . 1 1 1
A 1 2 3 . . . . . . . . .
B . . . 1 2 3 . . . . . .
C . . . . . . 1 2 3 . . .
D . . . . . . . . . 1 2 3
```

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 $\boldsymbol{Z}$ are partitioned. That is, $\boldsymbol{b}$ is divided into k groups, corresponding to the k random-effects terms, and the ith of these groups is subdivided into q_i groups of n_i elements. The q_i groups correspond to the q_i columns of the model matrix, $\boldsymbol{Z}_i$, and the n_i elements in each group correspond to the n_i levels of the ith grouping factor.

This partitioning determines the structure of the variance-covariance matrix, $\mathsf{Var}(\boldsymbol{B}) = \sigma^2\boldsymbol{\Sigma}(\boldsymbol{\theta})$, because random effects corresponding to different terms are assumed to be uncorrelated, as are random effects corresponding to different levels of the same term. Furthermore, the variance-covariance structure of each of the n_i groups of q_i possibly dependent elements within the ith "outer" group are identical.

Although this description may seem complicated, the structures are reasonably straightforward. The matrix $\boldsymbol{\Sigma}$ has the form

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

and the ith diagonal block, $\tilde{\boldsymbol{\Sigma}}_i$, has 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 ith diagonal block, $\tilde{\boldsymbol{\Sigma}}_i$, of size $n_i q_i \times n_i q_i$ is the relative variance-covariance of $\boldsymbol{B}_i$, the elements of $\boldsymbol{B}$ that are multiplied by $\tilde{\boldsymbol{Z}}_i$ in the linear predictor. The elements of $\boldsymbol{B}_i$ are ordered first by the column of $\boldsymbol{Z}_i$ then by the level of $\boldsymbol{f}_i$. It may be easier to picture the structure of $\tilde{\boldsymbol{\Sigma}}_i$ if we permute the elements of $\boldsymbol{B}_i$ so the ordering is first by level of $\boldsymbol{f}_i$ then by column of $\boldsymbol{Z}_i$. Let $\boldsymbol{P}_i$ be the matrix representing this permutation. Then

$$\boldsymbol{P}_i \tilde{\boldsymbol{\Sigma}}_i \boldsymbol{P}_i' = \begin{bmatrix} \boldsymbol{\Sigma}_i & \boldsymbol{0} & \dots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_i & \dots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \dots & \boldsymbol{\Sigma}_i \end{bmatrix} = \boldsymbol{I}_{n_i} \otimes \boldsymbol{\Sigma}_i. \tag{9}$$

The matrix $\boldsymbol{\Sigma}$ will be positive-semidefinite if and only 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.

Because we want to allow for $\boldsymbol{\Sigma}_i$ to be semidefinite and we also want to be able to write a "square root" of $\boldsymbol{\Sigma}_i$ (i.e. a matrix $\boldsymbol{K}$ such that $\boldsymbol{\Sigma}_i = \boldsymbol{K}\boldsymbol{K}'$), we write the factorization as

$$\boldsymbol{\Sigma}_i = \boldsymbol{T}_i \boldsymbol{S}_i \boldsymbol{S}_i \boldsymbol{T}_i', \quad i = 1, \dots, k \tag{10}$$

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. This is the "LDL′" form except that the diagonal elements of $\boldsymbol{S}_i$ are the square roots of the diagonal elements of the "D" factor in the "LDL′" form (and we have named the left, unit lower triangular factor $\boldsymbol{T}_i$ instead of "L").

If all of the diagonal elements of $\boldsymbol{S}_i$ are positive then $\boldsymbol{\Sigma}_i$ is positive-definite (i.e. $\boldsymbol{v}'\boldsymbol{\Sigma}(\boldsymbol{\theta})\boldsymbol{v} > 0, \forall \boldsymbol{v} \in \mathbb{R}^q, \boldsymbol{v} \neq \boldsymbol{0}$). If all the $\boldsymbol{\Sigma}_i, i = 1, \ldots, k$ are positive-definite then $\boldsymbol{\Sigma}$ is positive-definite.

We parameterize $\boldsymbol{\Sigma}_i$ according to the factorization (10). We define $\boldsymbol{\theta}_i$ 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 row-major order) of the strict lower triangle of $\boldsymbol{T}_i$. 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 for $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ illustrated in (6) and (7). That is, $\boldsymbol{T}(\boldsymbol{\theta})$ (respectively $\boldsymbol{S}(\boldsymbol{\theta})$) is block-diagonal with ith diagonal block $\tilde{\boldsymbol{T}}_i(\boldsymbol{\theta}) = \boldsymbol{T}(\boldsymbol{\theta}) \otimes \boldsymbol{I}_{n_i}$ (respectively $\tilde{\boldsymbol{S}}_i(\boldsymbol{\theta}) = \boldsymbol{S}(\boldsymbol{\theta}) \otimes \boldsymbol{I}_{n_i}$).

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.

In general, for a random-effects term `(1|factor)`, $q_i = 1$ and $\boldsymbol{T}_i$, which is a 1×1 unit lower triangular matrix, must be $\boldsymbol{I}_1$, the 1×1 identity matrix. Hence $\tilde{\boldsymbol{T}}_i = \boldsymbol{I}_{n_i}$ and the factorization $\tilde{\boldsymbol{\Sigma}}_i = \tilde{\boldsymbol{T}}_i\tilde{\boldsymbol{S}}_i\tilde{\boldsymbol{S}}_i\tilde{\boldsymbol{T}}_i'$ reduces to $\tilde{\boldsymbol{\Sigma}}_i = \tilde{\boldsymbol{S}}_i\tilde{\boldsymbol{S}}_i$. Furthermore, $\boldsymbol{S}_i$ is a 1×1 matrix $[\theta_i]$, subject to $\theta_i \geq 0$, and

$$\tilde{\boldsymbol{S}}_i = \theta_i \boldsymbol{I}_{n_i}$$

while

$$\tilde{\boldsymbol{\Sigma}}_i = \tilde{\boldsymbol{S}}_i\tilde{\boldsymbol{S}}_i = \theta_i^2 \boldsymbol{I}_{n_i}.$$

We see that the standard deviations of the elements of $\boldsymbol{B}_i$ are all equal to $\theta_i\sigma$, where, for linear mixed models and nonlinear mixed models, σ is the standard deviation of elements of $f_{\boldsymbol{Y}|\boldsymbol{B}}$. Similarly, the variance of the elements of $\boldsymbol{B}_i$, relative to the diagonal of the conditional variance, $\mathsf{Var}(\boldsymbol{Y}|\boldsymbol{B})$, is θ_i^2.

For the random-effects term like `(time|subj)`, for which $q_i = 2$, let us

write the $2(2+1)/2 = 3$-dimensional $\boldsymbol{\theta}_i$ as $[a, b, c]'$. Then

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

so that

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

and

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

so that

$$\tilde{\boldsymbol{T}}_i = \begin{bmatrix} \boldsymbol{I}_{n_i} & \mathbf{0} \\ c\boldsymbol{I}_{n_i} & \boldsymbol{I}_{n_i} \end{bmatrix}.$$

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

2.3 The fill-reducing permutation matrix, $\boldsymbol{P}$

We saw in §2.1 that the random-effects model matrix, $\boldsymbol{Z}$, is typically quite sparse (i.e. it is mostly zeros). Because $\boldsymbol{S}(\boldsymbol{\theta})$ is diagonal and because the pattern in $\boldsymbol{T}(\boldsymbol{\theta})$ is generated from the same partitioning of the elements of $\boldsymbol{b}$ that generates the pattern of the columns of $\boldsymbol{Z}$, the matrix

$$\boldsymbol{V}(\boldsymbol{\theta}) = \boldsymbol{Z}\boldsymbol{T}(\boldsymbol{\theta})\boldsymbol{S}(\boldsymbol{\theta}), \tag{11}$$

is also quite sparse. (In fact, the number and positions of the nonzeros in $\boldsymbol{V}(\boldsymbol{\theta})$ are the same as those for $\boldsymbol{Z}$, whenever $\boldsymbol{\theta}$ is not on the boundary.) We store $\boldsymbol{Z}$ and $\boldsymbol{V}(\boldsymbol{\theta})$ as sparse matrices. (To be more precise, we store $\boldsymbol{Z}'$ and $\boldsymbol{V}(\boldsymbol{\theta})'$ as compressed column matrices (Davis, 2006, ch. 2).)

As we will see in later sections, our techniques for determining the maximum likelihood estimates of the parameters, in any of the three kinds of mixed models we are considering, require evaluation of the Cholesky decomposition of sparse, symmetric, positive-definite matrices of the form

$$\boldsymbol{A}(\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{y}) = \boldsymbol{V}(\boldsymbol{\theta})'\boldsymbol{W}(\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{y})\boldsymbol{V}(\boldsymbol{\theta}) + \boldsymbol{I}_q \tag{12}$$

where $\boldsymbol{W}(\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{y})$ is a $q \times q$ diagonal matrix of positive weights and $\boldsymbol{u}$ is the orthogonal random-effects vector defined in the next section.

Evaluation of the Cholesky decomposition of $\boldsymbol{A}(\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{y})$ may be required for hundreds or even thousands of different combinations of $\boldsymbol{u}$, $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ during iterative optimization of the parameter estimates. Furthermore, the dimension, q, of $\boldsymbol{A}(\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{y})$ can be in the tens or hundreds of thousands for some of the data sets and models that are encountered in practice. Thus it is crucial that these Cholesky decompositions be evaluated efficiently.

Permuting (i.e. reordering) the columns of $\boldsymbol{V}(\boldsymbol{\theta})$ can affect, sometimes dramatically, the number of nonzero elements in the Cholesky factor of $\boldsymbol{A}(\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{y})$ and, consequently, the time required to perform the factorization. The number of nonzeros in the factor will always be at least as large as the number of nonzeros in the lower triangle of $\boldsymbol{A}$, but it can be larger — in which case we say that the factor has been "filled-in" relative to $\boldsymbol{A}$. Determining a fill-minimizing column permutation of $\boldsymbol{V}(\boldsymbol{\theta})$ is an extremely difficult and time-consuming operation when q is large. However, some heuristics, such as the approximate minimal degree ordering algorithm (Davis, 1996), can be used to rapidly determine a near-optimal, *fill-reducing permutation*. (See Davis (2006, ch. 7) for details.)

The symbolic analysis of the nonzero pattern in $\boldsymbol{V}(\boldsymbol{\theta})$ need only be done once (at $\boldsymbol{\theta}^{(0)}$) because the pattern of nonzeros in $\boldsymbol{A}(\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{y})$ depends only on the nonzero pattern of $\boldsymbol{V}(\boldsymbol{\theta})$, which is the same for all values of $\boldsymbol{\theta}$ not on the boundary. We will express the permutation as the $q \times q$ permutation matrix, $\boldsymbol{P}$, which is formed by applying the permutation to the rows of $\boldsymbol{I}_q$, and which has the property $\boldsymbol{P}'\boldsymbol{P} = \boldsymbol{P}\boldsymbol{P}' = \boldsymbol{I}_q$. The transpose, $\boldsymbol{P}'$, is also a permutation matrix. It represents the inverse to the permutation represented by $\boldsymbol{P}$.

2.4 Orthogonal random effects

For a fixed value of $\boldsymbol{\theta}$ we express the random variable $\boldsymbol{B}$ as

$$\boldsymbol{B} = \boldsymbol{T}(\boldsymbol{\theta})\boldsymbol{S}(\boldsymbol{\theta})\boldsymbol{P}'\boldsymbol{U} \tag{13}$$

where $\boldsymbol{U}$ is q-dimensional random variable representing *orthogonal random effects* having distribution $\boldsymbol{U} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ for which the density function, $f_{\boldsymbol{U}}$, is

$$f_{\boldsymbol{U}}(\boldsymbol{u}|\sigma^2) = (2\pi\sigma^2)^{-q/2} e^{-\boldsymbol{u}'\boldsymbol{u}/(2\sigma^2)}. \tag{14}$$

(When a generalized linear mixed model does not include the variance scale factor, σ^2, the distribution of $\boldsymbol{U}$ is the standard q-variate Gaussian distribu-

tion $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_q)$ with density $f_{\boldsymbol{U}}(\boldsymbol{u}) = (2\pi)^{-q/2} e^{-\boldsymbol{u}'\boldsymbol{u}/2}$.) The random effects $\boldsymbol{U}$ are "orthogonal" in the sense of being uncorrelated.

We note that (13) provides the desired distribution $\boldsymbol{B} \sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{\Sigma})$ because $\boldsymbol{B}$, as a linear transformation of $\boldsymbol{B}$, has a multivariate Gaussian distribution with mean

$$\mathsf{E}[\boldsymbol{B}] = \mathsf{E}[\boldsymbol{T}(\boldsymbol{\theta})\boldsymbol{S}(\boldsymbol{\theta})\boldsymbol{P}'\boldsymbol{U}] = \boldsymbol{T}(\boldsymbol{\theta})\boldsymbol{S}(\boldsymbol{\theta})\boldsymbol{P}'\mathsf{E}[\boldsymbol{U}] = \mathbf{0}$$

and variance-covariance matrix

$$\begin{aligned}
\mathsf{Var}(\boldsymbol{B}) = \mathsf{E}[\boldsymbol{B}\boldsymbol{B}'] &= \boldsymbol{T}(\boldsymbol{\theta})\boldsymbol{S}(\boldsymbol{\theta})\boldsymbol{P}'\mathsf{E}[\boldsymbol{U}\boldsymbol{U}']\boldsymbol{P}\boldsymbol{S}(\boldsymbol{\theta})\boldsymbol{T}(\boldsymbol{\theta})' \\
&= \boldsymbol{T}(\boldsymbol{\theta})\boldsymbol{S}(\boldsymbol{\theta})\boldsymbol{P}'\mathsf{Var}(\boldsymbol{U})\boldsymbol{P}\boldsymbol{S}(\boldsymbol{\theta})\boldsymbol{T}(\boldsymbol{\theta})' \\
&= \sigma^2\boldsymbol{T}(\boldsymbol{\theta})\boldsymbol{S}(\boldsymbol{\theta})\boldsymbol{P}'\boldsymbol{P}\boldsymbol{S}(\boldsymbol{\theta})\boldsymbol{T}(\boldsymbol{\theta})' \\
&= \sigma^2\boldsymbol{T}(\boldsymbol{\theta})\boldsymbol{S}(\boldsymbol{\theta})\boldsymbol{S}(\boldsymbol{\theta})\boldsymbol{T}(\boldsymbol{\theta})' \\
&= \sigma^2\boldsymbol{\Sigma}(\boldsymbol{\theta}).
\end{aligned}$$

Because $\boldsymbol{T}(\boldsymbol{\theta})$ is a unit lower triangular matrix its determinant, $|\boldsymbol{T}(\boldsymbol{\theta})|$, which is the product of the diagonal elements in the case of a triangular matrix, is unity. Hence $\boldsymbol{T}^{-1}(\boldsymbol{\theta})$ always exists. When $\boldsymbol{\theta}$ is not on the boundary of its constraint region, so that all the diagonal elements of $\boldsymbol{S}(\boldsymbol{\theta})$ are positive, then $\boldsymbol{S}^{-1}(\boldsymbol{\theta})$ exists, as does

$$\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}) = \boldsymbol{T}^{-1}(\boldsymbol{\theta})'\boldsymbol{S}^{-1}(\boldsymbol{\theta})\boldsymbol{S}^{-1}(\boldsymbol{\theta})\boldsymbol{T}^{-1}(\boldsymbol{\theta}). \tag{15}$$

That is, when $\boldsymbol{\theta}$ is not on the boundary $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ will be non-singular and we can express $\boldsymbol{U}$ as

$$\boldsymbol{U} = \boldsymbol{P}\boldsymbol{S}^{-1}(\boldsymbol{\theta})\boldsymbol{T}^{-1}(\boldsymbol{\theta})\boldsymbol{B}. \tag{16}$$

When $\boldsymbol{\theta}$ is on the boundary, meaning that one or more of the diagonal elements of the $\boldsymbol{S}_i, i = 1, \ldots, k$ is zero, $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is said to be a singular, or degenerate, variance-covariance matrix. In such cases there will be non-trivial linear combinations, $\boldsymbol{v}'\boldsymbol{B}$ where $\boldsymbol{v} \neq \mathbf{0}$, such that $\mathsf{Var}(\boldsymbol{v}'\boldsymbol{B}) = \sigma^2\boldsymbol{v}'\boldsymbol{\Sigma}(\boldsymbol{\theta})\boldsymbol{v} = 0$.

Because the conditional mean, $\boldsymbol{\mu}_{\boldsymbol{Y}|\boldsymbol{B}}$, depends on $\boldsymbol{b}$ only through the linear predictor, $\boldsymbol{\eta}_{\boldsymbol{b}}(\boldsymbol{b}, \boldsymbol{\beta})$, and because we can rewrite the linear predictor as a function of $\boldsymbol{\beta}$ and $\boldsymbol{u}$

$$\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{T}(\boldsymbol{\theta})\boldsymbol{S}(\boldsymbol{\theta})\boldsymbol{P}'\boldsymbol{u} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{V}(\boldsymbol{\theta})\boldsymbol{P}'\boldsymbol{u} = \boldsymbol{\eta}_{\boldsymbol{u}}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{u}) \tag{17}$$

we can form the conditional mean,

$$\boldsymbol{\mu}_{\boldsymbol{Y}|\boldsymbol{U}}(\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{\theta}) = \boldsymbol{\mu}(\boldsymbol{\eta}_{\boldsymbol{u}}(\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{\theta})), \tag{18}$$

with discrepancy, $d(\boldsymbol{\mu}(\boldsymbol{\eta}_{\boldsymbol{u}}(\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{\theta}), \boldsymbol{y})$. The conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{U}$ is

$$f_{\boldsymbol{Y}|\boldsymbol{U}}(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2) = k(\boldsymbol{y}, \sigma^2) e^{-d(\boldsymbol{\mu}(\boldsymbol{\eta}_{\boldsymbol{u}}(\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{\theta}), \boldsymbol{y}) / \left(2\sigma^2\right)}. \tag{19}$$

3 Evaluating the likelihood

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

Just as in (3), where we wrote the conditional density or the conditional probability mass function of $\boldsymbol{Y}$ given $\boldsymbol{B}$, whichever is appropriate, as $f_{\boldsymbol{Y}|\boldsymbol{B}}(\boldsymbol{y}|\boldsymbol{b}, \boldsymbol{\beta}, \sigma^2)$, we will write the unconditional (or marginal) density of $\boldsymbol{Y}$ or the unconditional probability mass function of $\boldsymbol{Y}$, whichever is appropriate, as $f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2)$. We can obtain $f_{\boldsymbol{Y}}$ by integrating $f_{\boldsymbol{Y}|\boldsymbol{B}}$ with respect to the marginal density $f_{\boldsymbol{B}}$ or by integrating $f_{\boldsymbol{Y}|\boldsymbol{U}}$ with respect to $f_{\boldsymbol{U}}$. Thus the likelihood can be expressed as

$$\begin{aligned}
L(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2|\boldsymbol{y}) &= f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2) \\
&= \int_{\boldsymbol{b}} f_{\boldsymbol{Y}|\boldsymbol{B}}(\boldsymbol{y}|\boldsymbol{b}, \boldsymbol{\beta}, \sigma^2)\, f_{\boldsymbol{B}}(\boldsymbol{b}|\boldsymbol{\theta}, \sigma^2)\, d\boldsymbol{b} \\
&= \int_{\boldsymbol{u}} f_{\boldsymbol{Y}|\boldsymbol{U}}(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2)\, f_{\boldsymbol{U}}(\boldsymbol{u}|\sigma^2)\, d\boldsymbol{u} \\
&= k(\boldsymbol{y}, \sigma^2) \int_{\boldsymbol{u}} \frac{e^{-(d(\boldsymbol{\mu}(\boldsymbol{\eta}_{\boldsymbol{u}}(\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{\theta})), \boldsymbol{y}) + \boldsymbol{u}'\boldsymbol{u})/(2\sigma^2)}}{\left(2\pi\sigma^2\right)^{q/2}}\, d\boldsymbol{u} \\
&= k(\boldsymbol{y}, \sigma^2) \int_{\boldsymbol{u}} \left(2\pi\sigma^2\right)^{-q/2} e^{-\delta(\boldsymbol{u}|\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{y})/(2\sigma^2)}\, d\boldsymbol{u}.
\end{aligned} \tag{20}$$

where

$$\delta(\boldsymbol{u}|\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{y}) = d(\boldsymbol{\mu}(\boldsymbol{\eta}_{\boldsymbol{u}}(\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{\theta})), \boldsymbol{y}) + \boldsymbol{u}'\boldsymbol{u} \tag{21}$$

is the *penalized discrepancy* function. It is composed of a "squared distance" between the conditional mean, $\boldsymbol{\mu}_{\boldsymbol{Y}|\boldsymbol{U}} = \boldsymbol{\mu}(\boldsymbol{\eta}_{\boldsymbol{u}}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{u}))$, and the observed data, $\boldsymbol{y}$, plus a "penalty", $\boldsymbol{u}'\boldsymbol{u}$, on the size of $\boldsymbol{u}$.

Note that the penalized discrepancy (21) and the likelihood (20) can be evaluated even when $\boldsymbol{\theta}$ on the boundary (and, hence, $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})$ does not exist).

It is important to be able to evaluate the likelihood for values of $\boldsymbol{\theta}$ on the boundary because the maximum likelihood estimates of $\boldsymbol{\theta}$ can (and do) occur on the boundary.

3.1 The Laplace approximation

In later sections we will see that, for the models that we are considering, it is relatively straightforward to determine the minimizer of the penalized discrepancy

$$\tilde{\boldsymbol{u}}(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y}) = \arg\min_{\boldsymbol{u}} \delta(\boldsymbol{u}|\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y}), \tag{22}$$

either directly, as the solution to a penalized linear least squares problem, or through an iterative algorithm in which each iteration requires the solution of a penalized linear least squares problem. Because the value that minimizes the penalized discrepancy will maximize the conditional density of $\boldsymbol{U}$, given $\boldsymbol{Y}$,

$$f_{\boldsymbol{U}|\boldsymbol{Y}}(\boldsymbol{u}|\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{\theta},\sigma^2) \propto k(\boldsymbol{y},\sigma^2)(2\pi\sigma^2)^{-q/2} e^{-\delta(\boldsymbol{u}|\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y})/(2\sigma^2)}, \tag{23}$$

$\tilde{\boldsymbol{u}}(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y})$ is called the *conditional mode* of $\boldsymbol{u}$ given the data, $\boldsymbol{y}$, and the parameters, $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. (The conditional density (23) depends on σ^2, in addition to the other parameters and the data, but the conditional mode (22) does not.)

Near the conditional mode, the quadratic approximation to the penalized discrepancy is

$$\delta(\boldsymbol{u}|\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y}) \approx \delta(\tilde{\boldsymbol{u}}|\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y}) + (\boldsymbol{u}-\tilde{\boldsymbol{u}})' \frac{\nabla^2_{\boldsymbol{u}}\delta(\boldsymbol{u}|\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y})|_{\boldsymbol{u}=\tilde{\boldsymbol{u}}}}{2} (\boldsymbol{u}-\tilde{\boldsymbol{u}}) \tag{24}$$

where $\nabla^2_{\boldsymbol{u}}\delta(\boldsymbol{u}|\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y})$ denotes the symmetric $q \times q$ *Hessian* matrix of the scalar function $\delta(\boldsymbol{u}|\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y})$. The (j,k)th element of $\nabla^2_{\boldsymbol{u}}\delta(\boldsymbol{u}|\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y})$ is

$$\frac{\partial^2\delta(\boldsymbol{u}|\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y})}{\partial u_j \partial u_k}. \tag{25}$$

One of the conditions for $\tilde{\boldsymbol{u}}(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y})$ to be the minimizer of the penalized discrepancy is that the Hessian at $\tilde{\boldsymbol{u}}$, $\nabla^2_{\boldsymbol{u}}\delta(\boldsymbol{u}|\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y})|_{\boldsymbol{u}=\tilde{\boldsymbol{u}}}$, must be positive-definite. We can, therefore, evaluate the Cholesky factor $\boldsymbol{L}(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y})$, which is the $q \times q$ lower triangular matrix with positive diagonal elements satisfying

$$\boldsymbol{L}(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y})\boldsymbol{L}(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y})' = \frac{\nabla^2_{\boldsymbol{u}}\delta(\boldsymbol{u}|\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y})|_{\boldsymbol{u}=\tilde{\boldsymbol{u}}(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y})}}{2}. \tag{26}$$

Substituting the quadratic approximation (24) into expression (20) for the likelihood, $L(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2|\boldsymbol{y})$, results in an integral in which the only part of the integrand that depends on $\boldsymbol{u}$ is the quadratic term in the exponent. To evaluate the non-constant part of the integral, which we can write as

$$I = \int_{\boldsymbol{u}} \frac{e^{-(\boldsymbol{u}-\tilde{\boldsymbol{u}})'\boldsymbol{L}(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y})\boldsymbol{L}(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y})'(\boldsymbol{u}-\tilde{\boldsymbol{u}})/(2\sigma^2)}}{(2\pi\sigma^2)^{q/2}}\, d\boldsymbol{u},$$

we change the variable of integration from $\boldsymbol{u}$ to $\boldsymbol{v} = \boldsymbol{L}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{y})'(\boldsymbol{u} - \tilde{\boldsymbol{u}})/\sigma$. The determinant of the Jacobian of this transformation is

$$\left|\frac{d\boldsymbol{v}}{d\boldsymbol{u}}\right| = \frac{|\boldsymbol{L}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{y})|}{\sigma^q},$$

implying that the differential, $d\boldsymbol{u}$, is

$$d\boldsymbol{u} = |\boldsymbol{L}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{y})|^{-1}\, \sigma^q d\boldsymbol{v}.$$

After the change of variable, I becomes a multiple of the integral of the standard q-variate Gaussian density

$$I = |\boldsymbol{L}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{y})|^{-1} \int_{\boldsymbol{v}} \frac{e^{-\boldsymbol{v}'\boldsymbol{v}/2}}{(2\pi)^{q/2}}\, d\boldsymbol{v}. \tag{27}$$

Finally, $I = |\boldsymbol{L}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{y})|^{-1}$ because the integral of a probability density over all $\boldsymbol{v} \in \mathbb{R}^q$ must be unity.

Returning to expression (20), we can now express the Laplace approximation to the likelihood function or, as more commonly used as the optimization criterion when determining maximum likelihood estimates, the log-likelihood,

$$\ell(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2|\boldsymbol{y}) = \log L(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2|\boldsymbol{y}). \tag{28}$$

(Because the logarithm function is monotonic, the maximizer of the log-likelihood also maximizes the likelihood. Generally the quadratic approximation to the log-likelihood is a better approximation than is the quadratic approximation to the likelihood.)

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

$$-2\ell(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2|\boldsymbol{y}) \approx -2\log[k(\boldsymbol{y}, \sigma^2)] + \frac{\delta(\tilde{\boldsymbol{u}}|\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{y})}{\sigma^2} + 2\log|\boldsymbol{L}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{y})|. \tag{29}$$

Expression (29) will be an exact evaluation of the log-likelihood, 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

$$\boldsymbol{Y} = \boldsymbol{X\beta}+\boldsymbol{ZB}+\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}\left(\boldsymbol{0}, \sigma^2\boldsymbol{I}\right), \quad \boldsymbol{B} \sim \mathcal{N}\left(\boldsymbol{0}, \sigma^2\boldsymbol{\Sigma}(\boldsymbol{\theta})\right), \quad \boldsymbol{\epsilon} \perp \boldsymbol{b} \tag{30}$$

where the symbol $\perp$ denotes independence of random variables. The conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{B}$ is

$$f_{\boldsymbol{Y}|\boldsymbol{B}}(\boldsymbol{y}|\boldsymbol{b}, \boldsymbol{\beta}, \sigma^2) = \left(2\pi\sigma^2\right)^{-n/2} e^{-\|\boldsymbol{Zb}+\boldsymbol{X\beta}-\boldsymbol{y}\|^2/(2\sigma^2)} \tag{31}$$

with conditional mean $\boldsymbol{\mu}_{\boldsymbol{Y}|\boldsymbol{B}}(\boldsymbol{b}, \boldsymbol{\beta}) = \boldsymbol{Zb} + \boldsymbol{X\beta} = \boldsymbol{\eta}_{\boldsymbol{b}}(\boldsymbol{b}, \boldsymbol{\beta})$.

We say that the conditional distribution of the response, $\boldsymbol{Y}$, given the random effects, $\boldsymbol{B}$, is a "spherical" Gaussian, $\boldsymbol{Y}|\boldsymbol{B} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{Y}|\boldsymbol{B}}, \sigma^2\boldsymbol{I})$, producing the discrepancy function and normalizing factor

$$d(\boldsymbol{\mu}, \boldsymbol{y}) = \|\boldsymbol{\mu} - \boldsymbol{y}\|^2 \tag{32}$$

$$k(\sigma^2) = \left(2\pi\sigma^2\right)^{-n/2}. \tag{33}$$

(A distribution of the form $\mathcal{N}(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$ is called a "spherical Gaussian" because contours of its density are spheres in $\mathbb{R}^n$.)

Furthermore, $\boldsymbol{\mu}_{\boldsymbol{Y}|\boldsymbol{B}}$ is exactly the linear predictor, $\boldsymbol{\eta}_{\boldsymbol{b}}(\boldsymbol{b}, \boldsymbol{\beta})$. That is, the "mean function", $\boldsymbol{\mu}(\boldsymbol{\eta})$, mapping the linear predictor, $\boldsymbol{\eta}$, to the conditional mean, $\boldsymbol{\mu}_{\boldsymbol{Y}|\boldsymbol{B}}$, is the identity, $\boldsymbol{\mu}(\boldsymbol{\eta}) = \boldsymbol{\eta}$.

The penalized discrepancy for this model is

$$\begin{aligned}
&\delta(\boldsymbol{u}|\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{y}) \\
&\qquad = d\left(\boldsymbol{\mu}(\boldsymbol{\eta}_{\boldsymbol{u}}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{u})), \boldsymbol{y}\right) + \boldsymbol{u}'\boldsymbol{u} \\
&\qquad = \|\boldsymbol{\eta}_{\boldsymbol{u}}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{u}) - \boldsymbol{y}\|^2 + \boldsymbol{u}'\boldsymbol{u} \\
&\qquad = \|\boldsymbol{VP}'\boldsymbol{u} + \boldsymbol{X\beta} - \boldsymbol{y}\|^2 + \boldsymbol{u}'\boldsymbol{u} \\
&\qquad = \left\| \begin{bmatrix} \boldsymbol{VP}' & \boldsymbol{X} & \boldsymbol{y} \end{bmatrix} \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{\beta} \\ -1 \end{bmatrix} \right\|^2 + \boldsymbol{u}'\boldsymbol{u} \\
&\qquad = \begin{bmatrix} \boldsymbol{u}' & \boldsymbol{\beta}' & -1 \end{bmatrix} \begin{bmatrix} \boldsymbol{PV}'\boldsymbol{VP}' + \boldsymbol{I} & \boldsymbol{PV}'\boldsymbol{X} & \boldsymbol{PV}'\boldsymbol{y} \\ \boldsymbol{X}'\boldsymbol{VP}' & \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{y} \\ \boldsymbol{y}'\boldsymbol{VP}' & \boldsymbol{y}'\boldsymbol{X} & \boldsymbol{y}'\boldsymbol{y} \end{bmatrix} \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{\beta} \\ -1 \end{bmatrix}.
\end{aligned} \tag{34}$$

(To save space we suppressed the dependence of $\boldsymbol{V}(\boldsymbol{\theta})$ on $\boldsymbol{\theta}$ and wrote it as $\boldsymbol{V}$.) In (34) it is obvious that $\delta(\boldsymbol{u}|\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y})$ is a quadratic function of $\boldsymbol{u}$, which means that expression (29) provides an exact evaluation of the log-likelihood. Furthermore, the Hessian

$$\nabla^2_{\boldsymbol{u}}\delta(\boldsymbol{u}|\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y}) = 2\left(\boldsymbol{P}\boldsymbol{V}(\boldsymbol{\theta})'\boldsymbol{V}(\boldsymbol{\theta})\boldsymbol{P}' + \boldsymbol{I}_q\right) = 2\boldsymbol{P}\left(\boldsymbol{V}(\boldsymbol{\theta})'\boldsymbol{V}(\boldsymbol{\theta}) + \boldsymbol{I}_q\right)\boldsymbol{P}', \tag{35}$$

is positive definite and depends only on $\boldsymbol{\theta}$. The Cholesky factor $\boldsymbol{L}(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y})$, defined in (26) and used in the log-likelihood evaluation (29), becomes $\boldsymbol{L}(\boldsymbol{\theta})$ and is the sparse lower triangular matrix with positive diagonal elements satisfying

$$\boldsymbol{L}(\boldsymbol{\theta})\boldsymbol{L}(\boldsymbol{\theta})' = \boldsymbol{P}\left(\boldsymbol{V}(\boldsymbol{\theta})'\boldsymbol{V}(\boldsymbol{\theta}) + \boldsymbol{I}_q\right)\boldsymbol{P}'. \tag{36}$$

Note that $\boldsymbol{V}(\boldsymbol{\theta})'\boldsymbol{V}(\boldsymbol{\theta})+\boldsymbol{I}_q$ is positive-definite, even when $\boldsymbol{\theta}$ is on the boundary, and thus the diagonal elements of $\boldsymbol{L}(\boldsymbol{\theta})$ are all positive, for any $\boldsymbol{\theta}$. The log-determinant, $2\log|\boldsymbol{L}(\boldsymbol{\theta})| = \log|\boldsymbol{V}(\boldsymbol{\theta})'\boldsymbol{V}(\boldsymbol{\theta}) + \boldsymbol{I}_q|$, required to evaluate (29), is simply twice the sum of the logarithms of these positive diagonal elements of $\boldsymbol{L}(\boldsymbol{\theta})$.

Determining the conditional mode, $\tilde{\boldsymbol{u}}(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y})$, as the solution to

$$\boldsymbol{L}(\boldsymbol{\theta})\boldsymbol{L}(\boldsymbol{\theta})'\tilde{\boldsymbol{u}}(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y}) = \boldsymbol{V}(\boldsymbol{\theta})'\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right) \tag{37}$$

is straightforward once the Cholesky factor, $\boldsymbol{L}(\boldsymbol{\theta})$, has been determined, thereby providing all the information needed to evaluate the log-likelihood from (29).

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. Given $\boldsymbol{\theta}$ we, in effect, evaluate a Cholesky factor for

$$\begin{bmatrix} \boldsymbol{P}(\boldsymbol{V}(\boldsymbol{\theta})'\boldsymbol{V}(\boldsymbol{\theta}) + \boldsymbol{I})\boldsymbol{P}' & \boldsymbol{P}\boldsymbol{V}'\boldsymbol{X} & \boldsymbol{P}\boldsymbol{V}(\boldsymbol{\theta})'\boldsymbol{y} \\ \boldsymbol{X}'\boldsymbol{V}(\boldsymbol{\theta})\boldsymbol{P}' & \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{y} \\ \boldsymbol{y}'\boldsymbol{V}(\boldsymbol{\theta})\boldsymbol{P}' & \boldsymbol{y}'\boldsymbol{X} & \boldsymbol{y}'\boldsymbol{y} \end{bmatrix}.$$

Because this factorization will involve combinations of sparse and dense matrices, we do it in stages, beginning with the evaluation of the sparse Cholesky factor, $\boldsymbol{L}(\boldsymbol{\theta})$, from (36). Next, solve for the $q \times p$ dense matrix $\boldsymbol{R}_{VX}(\boldsymbol{\theta})$ and the q-vector $\boldsymbol{r}_{Vy}(\boldsymbol{\theta})$ in

$$\boldsymbol{L}(\boldsymbol{\theta})\boldsymbol{R}_{VX}(\boldsymbol{\theta}) = \boldsymbol{P}\boldsymbol{V}(\boldsymbol{\theta})'\boldsymbol{X} \tag{38}$$

$$\boldsymbol{L}(\boldsymbol{\theta})\boldsymbol{r}_{Vy}(\boldsymbol{\theta}) = \boldsymbol{P}\boldsymbol{V}(\boldsymbol{\theta})'\boldsymbol{y}, \tag{39}$$

followed by the $p \times p$ upper triangular dense Cholesky factor $\boldsymbol{R}_X(\boldsymbol{\theta})$ satisfying

$$\boldsymbol{R}_X(\boldsymbol{\theta})'\boldsymbol{R}_X(\boldsymbol{\theta}) = \boldsymbol{X}'\boldsymbol{X} - \boldsymbol{R}_{VX}(\boldsymbol{\theta})'\boldsymbol{R}_{VX}(\boldsymbol{\theta}) \tag{40}$$

and the p-vector $\boldsymbol{r}_{Xy}(\boldsymbol{\theta})$ satisfying

$$\boldsymbol{R}_X(\boldsymbol{\theta})'\boldsymbol{r}_{Xy}(\boldsymbol{\theta}) = \boldsymbol{X}'\boldsymbol{y} - \boldsymbol{R}_{VX}(\boldsymbol{\theta})\boldsymbol{r}_{Vy}(\boldsymbol{\theta}). \tag{41}$$

Finally, evaluate the scalar

$$r(\boldsymbol{\theta}) = \sqrt{\|\boldsymbol{y}\|^2 - \|\boldsymbol{r}_{Xy}(\boldsymbol{\theta})\|^2 - \|\boldsymbol{r}_{Vy}(\boldsymbol{\theta})\|^2}. \tag{42}$$

(The astute reader may have noticed that the six steps, (36), (38), (39), (40)), (41) and (42), for evaluation of the log-likelihood, can be reduced to three, (36), (38) and (39), if we begin with the $n \times (p+1)$ matrix $[\boldsymbol{X} : \boldsymbol{y}]$ in place of the $n \times p$ matrix $\boldsymbol{X}$. We do so.)

Using these factors we can write

$$\begin{aligned}
&\delta(\boldsymbol{u}|\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{y}) \\
&= \left\| \begin{bmatrix} \boldsymbol{L}(\boldsymbol{\theta})' & \boldsymbol{R}_{VX}(\boldsymbol{\theta}) & \boldsymbol{r}_{Vy}(\boldsymbol{\theta}) \\ \boldsymbol{0} & \boldsymbol{R}_X(\boldsymbol{\theta}) & \boldsymbol{r}_{Xy}(\boldsymbol{\theta}) \\ \boldsymbol{0} & \boldsymbol{0} & r(\boldsymbol{\theta}) \end{bmatrix} \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{\beta} \\ -1 \end{bmatrix} \right\|^2 \\
&= r^2(\boldsymbol{\theta}) + \|\boldsymbol{R}_X(\boldsymbol{\theta})\boldsymbol{\beta} - \boldsymbol{r}_{Xy}(\boldsymbol{\theta})\|^2 + \|\boldsymbol{L}(\boldsymbol{\theta})'\boldsymbol{u} + \boldsymbol{R}_{VX}(\boldsymbol{\theta})\boldsymbol{\beta} - \boldsymbol{r}_{Vy}(\boldsymbol{\theta})\|^2 \\
&= r^2(\boldsymbol{\theta}) + \left\| \boldsymbol{R}_X(\boldsymbol{\theta}) \left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}) \right) \right\|^2 + \|\boldsymbol{L}(\boldsymbol{\theta})' \left(\boldsymbol{u} - \widehat{\boldsymbol{u}}(\boldsymbol{\theta}) \right)\|^2
\end{aligned} \tag{43}$$

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

$$\boldsymbol{R}_X(\boldsymbol{\theta})\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \boldsymbol{r}_{Xy}(\boldsymbol{\theta}) \tag{44}$$

$$\boldsymbol{L}(\boldsymbol{\theta})'\widehat{\boldsymbol{u}}(\boldsymbol{\theta}) = \boldsymbol{r}_{Vy}(\boldsymbol{\theta}) - \boldsymbol{R}_{VX}(\boldsymbol{\theta})'\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}). \tag{45}$$

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

$$\min_{\boldsymbol{u}} \delta(\boldsymbol{u}|\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \boldsymbol{\theta}, \boldsymbol{y}) = r^2(\boldsymbol{\theta}). \tag{46}$$

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

$$-2\ell(\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \boldsymbol{\theta}, \sigma^2|\boldsymbol{y}) = n \log\left(2\pi\sigma^2\right) + \frac{r^2(\boldsymbol{\theta})}{\sigma^2} + 2\log|\boldsymbol{L}(\boldsymbol{\theta})|. \tag{47}$$

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

$$\widehat{\sigma^2}(\boldsymbol{\theta})=\frac{r^2(\boldsymbol{\theta})}{n}. \tag{48}$$

Substituting this estimate into (47) provides the *profiled deviance* function

$$\begin{aligned}-2\ell(\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}),\boldsymbol{\theta},\widehat{\sigma^2}(\boldsymbol{\theta})|\boldsymbol{y}) &= n\log\left(\frac{2\pi r^2(\boldsymbol{\theta})}{n}\right)+n+2\log|\boldsymbol{L}(\boldsymbol{\theta})|\\ &= n\left[1+\log\left(2\pi/n\right)\right]+n\log r^2(\boldsymbol{\theta})+2\log|\boldsymbol{L}(\boldsymbol{\theta})|.\end{aligned} \tag{49}$$

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

$$\widehat{\boldsymbol{\theta}}=\arg\min_{\boldsymbol{\theta}}\left\{n\left[1+\log\left(2\pi/n\right)\right]+n\log r^2(\boldsymbol{\theta})+2\log|\boldsymbol{L}(\boldsymbol{\theta})|\right\}. \tag{50}$$

The mle's of the other parameters are determined from $\widehat{\boldsymbol{\theta}}$ using (48) and (44). The conditional modes of the orthogonal random effects, $\widehat{\boldsymbol{u}}(\widehat{\boldsymbol{\theta}})$, evaluated using (45), and the corresponding conditional modes of the untransformed random effects,

$$\widehat{\boldsymbol{b}}(\widehat{\boldsymbol{\theta}})=\boldsymbol{T}(\widehat{\boldsymbol{\theta}})\boldsymbol{S}(\widehat{\boldsymbol{\theta}})\widehat{\boldsymbol{u}}(\widehat{\boldsymbol{\theta}}), \tag{51}$$

are called the *empirical Best Linear Unbiased Predictors* (eBLUPs) of the random effects.

The three terms in the objective function being minimized in (50) are, respectively, a constant, $n\left[1+\log\left(2\pi/n\right)\right]$, a measure of the fidelity of the fitted values to the observed data, $n\log r^2(\boldsymbol{\theta})$, and a measure of model complexity, $2\log|\boldsymbol{L}(\boldsymbol{\theta})|$. 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 by an appropriate choice of $\boldsymbol{\theta}$.

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

$$\begin{aligned} L_R(\boldsymbol{\theta}, \sigma^2|\boldsymbol{y}) &= \int_{\boldsymbol{\beta}} L(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2|\boldsymbol{y})\, d\boldsymbol{\beta} \\ &= \frac{e^{-r^2(\boldsymbol{\theta})/(2\sigma^2)}}{|\boldsymbol{L}(\boldsymbol{\theta})|(2\pi\sigma^2)^{(n-p)/2}} \int_{\boldsymbol{\beta}} \frac{e^{-(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})'\boldsymbol{R}_X'\boldsymbol{R}_X(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})/(2\sigma^2)}}{(2\pi\sigma^2)^{p/2}}\, d\boldsymbol{\beta} \\ &= \frac{e^{-r^2(\boldsymbol{\theta})/(2\sigma^2)}}{|\boldsymbol{L}(\boldsymbol{\theta})||\boldsymbol{R}_X(\boldsymbol{\theta})|(2\pi\sigma^2)^{(n-p)/2}} \end{aligned} \tag{52}$$

or, on the deviance scale,

$$-2\ell_R(\boldsymbol{\theta}, \sigma^2|\boldsymbol{y}) = (n-p)\log\left(2\pi\sigma^2\right) + \frac{r^2(\boldsymbol{\theta})}{\sigma^2} + 2|\boldsymbol{L}(\boldsymbol{\theta})| + 2|\boldsymbol{R}_X(\boldsymbol{\theta})| \tag{53}$$

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

$$\widehat{\sigma}_R(\boldsymbol{\theta}) = \frac{r^2(\boldsymbol{\theta})}{n-p} \tag{54}$$

and the profiled REML deviance is

$$\begin{aligned} -2\ell_R(\boldsymbol{\theta}, \widehat{\sigma}^2(\boldsymbol{\theta})|\boldsymbol{y}) = (n-p)\left[1 + \log\left(2\pi/(n-p)\right)\right] + (n-p)\log r^2(\boldsymbol{\theta}) \\ + 2\log|\boldsymbol{L}(\boldsymbol{\theta})| + 2\log|\boldsymbol{R}_X(\boldsymbol{\theta})|. \end{aligned} \tag{55}$$

5 Nonlinear mixed models

The nonlinear mixed model can be expressed as

$$\boldsymbol{Y} = \boldsymbol{\mu}(\boldsymbol{\eta_b}(\boldsymbol{B}, \boldsymbol{\beta})) + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}), \quad \boldsymbol{B} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{\Sigma}(\boldsymbol{\theta})), \quad \boldsymbol{\epsilon} \perp \boldsymbol{B}, \tag{56}$$

which is very similar to the linear mixed model (30). In fact, these two types of mixed models differ only in the form of the mean function, $\boldsymbol{\mu}(\boldsymbol{\eta})$.

The discrepancy function, $d(\boldsymbol{\mu}_{\boldsymbol{Y}|\boldsymbol{B}}, \boldsymbol{y})$, and the normalizing factor, $k(\sigma^2)$, are the same as those for a linear mixed model, (32) and (33), respectively. However, for a nonlinear mixed model, the mean function, $\boldsymbol{\mu}(\boldsymbol{\eta})$, is not the identity. In the nonlinear model each element of $\boldsymbol{\mu}$ is the value of a scalar nonlinear model function, $g(\boldsymbol{x}, \boldsymbol{\phi})$, that depends on the observed values of some covariates, $\boldsymbol{x}$, and on a parameter vector, $\boldsymbol{\phi}$, of length s. This model function, $g(\boldsymbol{x}, \boldsymbol{\phi})$, can be nonlinear in some or all of the elements of $\boldsymbol{\phi}$.

When estimating the parameters in a model, the values of the covariates at each observation are known so we can regard the ith element of $\boldsymbol{\mu}_{\boldsymbol{Y}|\boldsymbol{B}}$ as a function of $\boldsymbol{\phi}_i$ only and write

$$\boldsymbol{\mu} = \boldsymbol{g}(\boldsymbol{\Phi}) \tag{57}$$

where $\boldsymbol{\Phi}$ is the $n \times s$ matrix with ith row $\boldsymbol{\phi}_i, i = 1, \ldots, n$ and the vector-valued function, $\boldsymbol{g}$, applies the scalar function $g(\boldsymbol{x}, \boldsymbol{\phi})$ to the rows of $\boldsymbol{\Phi}$ and the corresponding covariates $\boldsymbol{x}_i, i = 1, \ldots, n$.

The linear predictor, $\boldsymbol{\eta}$, is

$$\boldsymbol{\eta} = \mathrm{vec}(\boldsymbol{\Phi}) = \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{V}(\boldsymbol{\theta})\boldsymbol{P}'\boldsymbol{u} + \boldsymbol{X}\boldsymbol{\beta}. \tag{58}$$

(The vec operator concatenates the columns of a matrix to form a vector.) The matrix $\boldsymbol{\Phi}$ is $n \times s$, hence the dimension of $\mathrm{vec}(\boldsymbol{\Phi})$ is $m = ns$, so $\boldsymbol{X}$ is $ns \times p$ while $\boldsymbol{Z}$ and $\boldsymbol{V}(\boldsymbol{\theta})$ are $ns \times q$. Because the ith element of $\boldsymbol{\mu}$ depends only on the ith row of $\boldsymbol{\Phi}$, the $n \times ns$ gradient matrix,

$$\boldsymbol{W}(\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{\theta}) = \frac{d\boldsymbol{\mu}}{d\boldsymbol{\eta}'}, \tag{59}$$

is the horizontal concatenation of s diagonal $n \times n$ matrices. The ith diagonal element in the jth diagonal block of $\boldsymbol{W}$ is

$$\{\boldsymbol{W}(\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{\theta})\}_{i,i+(j-1)n} = \left.\frac{\partial g(\boldsymbol{x}, \boldsymbol{\phi})}{\partial \phi_j}\right|_{\boldsymbol{x}=\boldsymbol{x}_i, \boldsymbol{\phi}=\boldsymbol{\phi}_i}.$$

All other elements in $\boldsymbol{W}$ are zero.

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

$$\begin{aligned}\tilde{\boldsymbol{u}}(\boldsymbol{\beta},\boldsymbol{\theta}) &= \arg\min_{\boldsymbol{u}} \delta(\boldsymbol{u}|\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y}) \\ &= \arg\min_{\boldsymbol{u}} \left(\|\boldsymbol{\mu}(\boldsymbol{\eta}_{\boldsymbol{u}}(\boldsymbol{u},\boldsymbol{\beta},\boldsymbol{\theta})) - \boldsymbol{y}\|^2 + \boldsymbol{u}'\boldsymbol{u}\right).\end{aligned} \tag{60}$$

Generally (60) is a penalized nonlinear least squares problem requiring an iterative solution, not a penalized linear least squares problem like (37) with a direct solution.

To describe the general case of an iterative solution to (60) we will use parenthesized superscripts to denote the number of the iteration at which a quantity is evaluated. At $\boldsymbol{u}^{(i)}$, the value of the $\boldsymbol{u}$ at the ith iteration, the linear approximation to $\boldsymbol{\mu}$ as a function of $\boldsymbol{u}$ is

$$\begin{aligned}\boldsymbol{\mu}(\boldsymbol{\eta}_{\boldsymbol{u}}(\boldsymbol{u},\boldsymbol{\beta},\boldsymbol{\theta})) \approx& \boldsymbol{\mu}(\boldsymbol{\eta}_{\boldsymbol{u}}(\boldsymbol{u}^{(i)},\boldsymbol{\beta},\boldsymbol{\theta})) + \left.\frac{\partial\boldsymbol{\mu}}{\partial\boldsymbol{u}'}\right|_{\boldsymbol{u}=\boldsymbol{u}^{(i)}} (\boldsymbol{u}-\boldsymbol{u}^{(i)}) \\ &= \boldsymbol{\mu}^{(i)} + \boldsymbol{W}(\boldsymbol{u}^{(i)},\boldsymbol{\beta},\boldsymbol{\theta})\boldsymbol{V}(\boldsymbol{\theta})\boldsymbol{P}'(\boldsymbol{u}-\boldsymbol{u}^{(i)}) \\ &= \boldsymbol{\mu}^{(i)} + \boldsymbol{M}^{(i)}\boldsymbol{P}'(\boldsymbol{u}-\boldsymbol{u}^{(i)})\end{aligned} \tag{61}$$

where $\boldsymbol{\mu}^{(i)} = \boldsymbol{\mu}(\boldsymbol{\eta}_{\boldsymbol{u}}(\boldsymbol{u}^{(i)},\boldsymbol{\beta},\boldsymbol{\theta}))$ and the $n \times q$ matrix

$$\boldsymbol{M}^{(i)} = \boldsymbol{W}(\boldsymbol{u}^{(i)},\boldsymbol{\beta},\boldsymbol{\theta})\boldsymbol{V}(\boldsymbol{\theta}). \tag{62}$$

As described in §5.3, for some nonlinear models the conditional mean $\boldsymbol{\mu}_{\boldsymbol{Y}|\boldsymbol{U}}$ is a linear function of $\boldsymbol{u}$ (but a nonlinear function of one or more components of $\boldsymbol{\beta}$, so that the model cannot be written as a linear mixed model). For these *conditionally linear mixed models* (61) is exact and $\boldsymbol{M}$ does not depend upon $\boldsymbol{u}$.

In the general case, the proposed increment, $\boldsymbol{\delta}^{(i)} = \boldsymbol{u}^{(i+1)} - \boldsymbol{u}^{(i)}$, minimizes the approximate penalized discrepancy obtained from (61). That is,

$$\begin{aligned}\boldsymbol{\delta}^{(i)} &= \arg\min_{\boldsymbol{\delta}} \left\|\boldsymbol{\mu}^{(i)} + \boldsymbol{M}^{(i)}\boldsymbol{P}'\boldsymbol{\delta} - \boldsymbol{y}\right\|^2 + (\boldsymbol{\delta}+\boldsymbol{u}^{(i)})'(\boldsymbol{\delta}+\boldsymbol{u}^{(i)}) \\ &= \arg\min_{\boldsymbol{\delta}} \left\| \begin{bmatrix} \boldsymbol{M}^{(i)}\boldsymbol{P}'\boldsymbol{\delta} \\ \boldsymbol{\delta}+\boldsymbol{u}^{(i)} \end{bmatrix} - \begin{bmatrix} \boldsymbol{y}-\boldsymbol{\mu}^{(i)} \\ \boldsymbol{0} \end{bmatrix} \right\|^2 \\ &= \arg\min_{\boldsymbol{\delta}} \left\| \begin{bmatrix} \boldsymbol{M}^{(i)}\boldsymbol{P}' \\ \boldsymbol{I}_q \end{bmatrix} \boldsymbol{\delta} - \begin{bmatrix} \boldsymbol{y}-\boldsymbol{\mu}^{(i)} \\ -\boldsymbol{u}^{(i+1)} \end{bmatrix} \right\|^2,\end{aligned} \tag{63}$$

which implies that $\boldsymbol{\delta}^{(i)}$ satisfies the "normal equations"

$$\boldsymbol{P}\left(\boldsymbol{M}^{(i)\prime}\boldsymbol{M}^{(i)}+\boldsymbol{I}_q\right)\boldsymbol{P}'\boldsymbol{\delta}^{(i)}=\boldsymbol{P}\boldsymbol{M}^{(i)\prime}\left(\boldsymbol{y}-\boldsymbol{\mu}^{(i)}\right)-\boldsymbol{u}^{(i)}. \tag{64}$$

Let the Cholesky factor $\boldsymbol{L}(\boldsymbol{u}^{(i)},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y})$, which we will write as $\boldsymbol{L}^{(i)}$, be the sparse, lower triangular matrix satisfying

$$\boldsymbol{L}^{(i)}\boldsymbol{L}^{(i)\prime}=\boldsymbol{P}\left(\boldsymbol{M}^{(i)\prime}\boldsymbol{M}^{(i)}+\boldsymbol{I}_q\right)\boldsymbol{P}'. \tag{65}$$

The increment $\boldsymbol{\delta}^{(i)}$ is evaluated by successively solving the two sparse triangular systems in

$$\boldsymbol{L}^{(i)}\boldsymbol{L}^{(i)\prime}\boldsymbol{\delta}^{(i)}=\boldsymbol{P}\boldsymbol{M}^{(i)\prime}\left(\boldsymbol{y}-\boldsymbol{\mu}^{(i)}\right)-\boldsymbol{u}^{(i)}. \tag{66}$$

5.1.1 Step factor and convergence criterion

Some examination of the penalized least squares problem (63) will show that is it possible to write a penalized least squares problem for the updated random effects, $\boldsymbol{u}^{(i+1)}=\boldsymbol{u}^{(i)}+\boldsymbol{\delta}^{(i)}$, directly. We prefer to write the conditions in terms of the increment and to calculate the proposed increment, $\boldsymbol{\delta}^{(i)}$, using (66) for two reasons: to allow us to incorporate a step factor (Bates and Watts, 1988, §2.2.1) easily and to evaluate the relative offset convergence criterion, which is based on the extent to which the residual vector is orthogonal to the columns of the gradient matrix.

With highly nonlinear models it can happen that applying the proposed increment, $\boldsymbol{\delta}^{(i)}$, actually increases the penalized discrepancy rather than decreasing it. In these cases we use only a fraction, h, of the proposed step, $\boldsymbol{\delta}^{(i)}$, where $0<h_{min}\le h\le 1$ and h_{min} is a prespecified minimum step factor. We evaluate the penalized discrepancy at $\boldsymbol{u}=\boldsymbol{u}^{-1}+h\boldsymbol{\delta}^{(i)}$ for successively smaller values of h until we obtain a decrease in the penalized discrepancy or the minimum step factor is reached. Generally a simple strategy such as setting $h=1$ at the beginning of each iteration and successively halving h if necessary is sufficient.

We continue iterating until the increments become negligible whereupon we declare convergence. To design an algorithm, however, we must decide how to measure the size of the increment and when to declare that this size is "negligible".

As described in Bates and Watts (1988, §2.2.3) the numerical uncertainty in the value of the conditional mode, $\tilde{\boldsymbol{u}}(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y}$, relative to the statistical

uncertainty, can be assessed as the ratio of the length of two orthogonal components of the residual vector, evaluated at the current value of $\boldsymbol{u}$, $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. This "relative offset" or orthogonality convergence criterion is

$$\frac{\|\boldsymbol{L}^{(i)'}\boldsymbol{\delta}^{(i)}\|/\sqrt{q}}{\|\boldsymbol{P}\boldsymbol{M}^{(i)'}(\boldsymbol{y}-\boldsymbol{\mu}^{(i)})-\boldsymbol{u}^{(i)}\|/\sqrt{n-q}}$$

when the increment is calculated using (66).

Convergence is declared when this criterion drops below a threshold, typically a value on the order of 0.001. It is desirable to use a convergence criterion such as this relative offset because it is a true convergence criterion, not simply an indicator that the iterations are no longer making progress. In other words, this criterion depends only not the current position and not on the path taken by the algorithm to this position. See Bates and Watts (1988, §2.2.3) for details and McCullough (1999) for comparison of the behavior of software that uses this criterion (S-PLUS, in this comparison) versus other commercial software. The software using this criterion was one of only two packages that did not declare convergence to a spurious optimum on at least one of the problems in the test suite. (The convergence criterion used by the `nls` function in S-PLUS is misstated in McCullough (1999). That function uses the relative offset criterion.)

5.1.2 The Laplace approximation for nonlinear mixed models

At convergence the Laplace approximation to the deviance is

$$-2\ell(\boldsymbol{\beta},\boldsymbol{\theta},\sigma^2|\boldsymbol{y}) = n\log\left(2\pi\sigma^2\right) + \frac{\delta(\tilde{\boldsymbol{u}}|\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y})}{\sigma^2} + 2\log|\boldsymbol{L}(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y})| \tag{67}$$

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

$$\widehat{\sigma^2}(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y}) = \frac{\delta(\tilde{\boldsymbol{u}}|\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y})}{n}. \tag{68}$$

Substituting this estimate into (67) produces the Laplace approximation to the profiled deviance

$$\begin{aligned}-2\ell(\boldsymbol{\beta},\boldsymbol{\theta},\widehat{\sigma^2}(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y})|\boldsymbol{y}) = n\left[1+\log\left(2\pi/n\right)\right]\\ + n\log\delta(\tilde{\boldsymbol{u}}|\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y}) + 2\log|\boldsymbol{L}(\tilde{\boldsymbol{u}},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y})|.\end{aligned} \tag{69}$$

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

$$\boldsymbol{\phi}_1 \left(1 - e^{-\boldsymbol{\phi}_2 x_{i,j}}\right) \quad i = 1, \ldots, 4;\ j = 1, \ldots, 3 \tag{70}$$

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, $\boldsymbol{\beta} = [\beta_1, \beta_2]'$ plus a subject-specific random effect for each parameter, and 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 12×2 parameter matrix $\boldsymbol{\Phi}$ is a vector of length 24 where the first 12 elements are values of `A` and the last 12 elements are values of `rc`. In the mixed-model formula the names `A` and `rc` represent indicator variables for the first 12 and the last 12 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 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. In the random-effects terms the intercept is always removed. In the fixed effects it is replaced by the sum of the parameter name indicators. Thus the formula shown above is equivalent to

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

The matrix $\boldsymbol{X}$ will be 24×2 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

$$\boldsymbol{\mu}\left(\boldsymbol{\eta}_{\boldsymbol{u}}(\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{\theta})\right) = \boldsymbol{\mu}(\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{\theta}) = \boldsymbol{\mu}_{\mathbf{0}}(\boldsymbol{\beta}) + \boldsymbol{M}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{y})\boldsymbol{u} \tag{71}$$

where $\boldsymbol{\mu}_{\mathbf{0}}(\boldsymbol{\beta}) = \boldsymbol{\mu}_{\boldsymbol{Y}|\boldsymbol{U}}(\mathbf{0}, \boldsymbol{\beta}, \boldsymbol{\theta}) = \boldsymbol{\mu}\left(\boldsymbol{X}\boldsymbol{\beta}\right)$. The penalized least squares problem (**??**) for the updated $\boldsymbol{u}$ can be rewritten as

$$\tilde{\boldsymbol{u}}\left(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{y}\right) = \min_{\boldsymbol{u}} \left\| \begin{bmatrix} \boldsymbol{y} - \boldsymbol{\mu}_{\mathbf{0}}(\boldsymbol{\beta}) \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \boldsymbol{M}(\boldsymbol{\beta}, \boldsymbol{\theta}) \\ \boldsymbol{I}_q \end{bmatrix} \boldsymbol{u} \right\|^2 . \tag{72}$$

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

$$\left(\boldsymbol{M}(\boldsymbol{\beta}, \boldsymbol{\theta})'\boldsymbol{M}(\boldsymbol{\beta}, \boldsymbol{\theta}) + \boldsymbol{I}_q\right)\tilde{\boldsymbol{u}} = \boldsymbol{L}(\boldsymbol{\beta}, \boldsymbol{\theta})\boldsymbol{L}(\boldsymbol{\beta}, \boldsymbol{\theta})'\tilde{\boldsymbol{u}}\boldsymbol{M}(\boldsymbol{\beta}, \boldsymbol{\theta})'\left(\boldsymbol{y} - \boldsymbol{\mu}_{\mathbf{0}}(\boldsymbol{\beta})\right) \tag{73}$$

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

$$\begin{aligned} \eta_k &= \{\boldsymbol{\eta}\}_k = \{\boldsymbol{g}(\boldsymbol{\eta})\}_k = g(\{\boldsymbol{\eta}\}) = g(\mu_k) \quad k = 1, \ldots, n \\ \mu_k &= \{\boldsymbol{\mu}\}_k = \left\{\boldsymbol{g}^{-1}(\boldsymbol{\mu})\right\}_k = g^{-1}(\{\boldsymbol{\mu}\}) \quad k = 1, \ldots, n. \end{aligned} \tag{74}$$

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

$$d(\boldsymbol{\mu}, \boldsymbol{y}) = \sum_{k=1}^{n} r_D^2(\mu_k, y_k) \tag{75}$$

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

$$f_{\boldsymbol{Y}|\boldsymbol{\mu}}(\boldsymbol{y}, \boldsymbol{\mu}) = \prod_{k=1}^{n} \mu_k^{y_k} (1 - \mu_k)^{(1-y_k)} \tag{76}$$

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 (3) for the conditional density in terms of the discrepancy can be written

$$f_{\boldsymbol{Y}|\boldsymbol{\mu}}(\boldsymbol{y}|\boldsymbol{\mu}) = ke^{-d(\boldsymbol{\mu}, \boldsymbol{y})/2}. \tag{77}$$

Thus the discrepancy function must be

$$d(\boldsymbol{\mu}, \boldsymbol{y}) = \tag{78}$$

References

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

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

Tim Davis. An approximate minimal degree ordering algorithm. *SIAM J. Matrix Analysis and Applications*, 17(4):886–905, 1996.

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

Peter McCullagh and John Nelder. *Generalized Linear Models*. Chapman and Hall, 2nd edition, 1989.

B. D. McCullough. Assessing the reliability of statistical software: Part ii. *The American Statistician*, 53(2), May 1999.