Linear mixed model implementation in lme4

Douglas Bates
Department of Statistics
University of Wisconsin – Madison

January 31, 2007

Abstract

We describe the form of the linear mixed-effects and generalized linear mixed-effects models fit by `lmer` and give details of the representation and the computational techniques used to fit such models. These techniques are illustrated on several examples.

1 A simple example

The `Rail` data set from the `nlme` package is described in `?` as consisting of three measurements of the travel time of a type of sound wave on each of six sample railroad rails. We can examine the structure of these data with the `str` function

```
> str(Rail)
'data.frame':        18 obs. of  2 variables:
 $ travel: num  55 53 54 26 37 32 78 91 85 92 ...
 $ Rail  : Ord.factor w/ 6 levels "2"<"5"<"1"<"6"<..: 3 3 3 1 1 1 5 5 5 6 ...
```

Because there are only three observations on each of the rails a dotplot (Figure 1) shows the structure of the data well.

```
> print(dotplot(Rail ~ travel, Rail, xlab = "Travel time (ms)",
+     ylab = "Rail number"))
```

In building a model for these data

```
> Rail
```

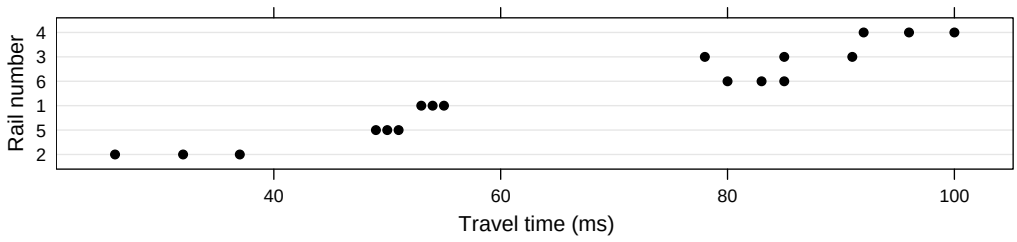

Figure 1: Travel time of sound waves in a sample of six railroad rails. There were three measurements of the travel time on each rail. The rail numbers are sorted by increasing mean travel time.

	travel	Rail
1	55	1
2	53	1
3	54	1
4	26	2
5	37	2
6	32	2
7	78	3
8	91	3
9	85	3
10	92	4
11	100	4
12	96	4
13	49	5
14	51	5
15	50	5
16	80	6
17	85	6
18	83	6

we wish to characterize a typical travel time, say μ, for the population of such railroad rails and the deviations, say $b_i, i = 1, \ldots, 6$ of the individual rails from this population mean. Because these specific rails are not of interest by themselves as much as the variation in the population we model the b_i, which are called the "random effects" for the rails, as having a normal (Gaussian) distribution of the form $\mathcal{N}(0, \sigma_b^2)$. The jth measurement on the ith rail is expressed as

$$y_{ij} = \mu + b_i + \epsilon_{ij} \quad b_i \sim \mathcal{N}(0, \sigma_b^2), \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \quad i = 1, \ldots, 6 \; j = 1, \ldots, 3 \tag{1}$$

The parameters of this model are μ, σ_b^2 and σ^2. Technically the $b_i, i = 1, \ldots, 6$ are not parameters but instead are considered to be unobserved random variables for which we form "predictions" instead of "estimates".

To express generalizations of models like (1) more conveniently we switch to a matrix/vector representation in which the 18 observations of the travel time form the response vector $\boldsymbol{y}$, the fixed-effect parameter μ forms a 1-dimensional column vector $\boldsymbol{\beta}$ and the six random effects $b_i, i = 1, \ldots, 6$ form the random effects vector $\boldsymbol{b}$. The structure of the data and the values of any covariates (none are used in this model) are used to create model matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$.

Using these vectors and matrices and the 18-dimensional vector $\boldsymbol{\epsilon}$ that represents the per-observation noise terms the model becomes

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

In the general form we write p for the dimension of $\boldsymbol{\beta}$, the fixed-effects parameter vector, and q for the dimension of $\boldsymbol{b}$, the vector of random effects. Thus the model matrix $\boldsymbol{X}$ has dimension $n \times p$, the model matrix $\boldsymbol{Z}$ has dimension $n \times q$ and the relative variance-covariance matrix, $\boldsymbol{\Sigma}$, for the random-effects has dimension $q \times q$. The symbol $\perp$ indicates independence of random variables and $\mathcal{N}$ denotes the multivariate normal (Gaussian) distribution.

We say that matrix $\boldsymbol{\Sigma}$ is the relative variance-covariance matrix of the random effects in the sense that it is the variance of $\boldsymbol{b}$ relative to σ^2, the scalar variance of the per-observation noise term $\boldsymbol{\epsilon}$. Although it size, q, can be very large, $\boldsymbol{\Sigma}$ is highly structured. It is symmetric, positive semi-definite and zero except for the diagonal elements and certain elements close to the diagonal.

1.1 Fitting the model and examining the results

The maximum likelihood estimates for parameters in model (1) fit to the `Rail` data are obtained as

```
> Rm1ML <- lmer2(travel ~ 1 + (1 | Rail), Rail, method = "ML",
+     control = list(msVerbose = 1))
  0      149.289: 0.942809
  1      137.531:  1.94281
  2      132.389:  2.85077
  3      129.942:  3.73815
  4      128.945:  4.52610
  5      128.629:  5.12723
  6      128.566:  5.47713
  7      128.560:  5.60451
  8      128.560:  5.62581
  9      128.560:  5.62686
 10      128.560:  5.62686
```

In this fit we have set the control parameter `msVerbose` to 1 indicating that information on the progress of the iterations should be printed after every iteration. Each line gives the iteration number, the value of the deviance (negative twice the log-likelihood) and the value of the parameter s which is the standard deviation of the random effects relative to the standard deviation of the residuals.

The printed form of the model

```
> Rm1ML
Linear mixed-effects model fit by maximum likelihood
Formula: travel ~ 1 + (1 | Rail)
   Data: Rail
   AIC   BIC logLik MLdeviance REMLdeviance
 132.6 134.3 -64.28      128.6        122.2
Random effects:
 Groups   Name        Variance Std.Dev.
 Rail     (Intercept) 541.971  23.2803
 Residual              17.118   4.1373
Number of obs: 18, groups: Rail, 6

Fixed effects:
            Estimate Std. Error t value
(Intercept)   66.500      9.285   7.162
```

provides additional information about the parameter estimates and some of the measures of the fit such as the log-likelihood (-64.28), the deviance for the maximum likelihood criterion (128.6), the deviance for the REML criterion (122.2), Akaike's Information Criterion (AIC= 132.6) and Schwartz's Bayesian Information Criterion (BIC= 134.3).

The model matrices $\boldsymbol{Z}$ and $\boldsymbol{X}$ and the negative of the response vector $-\boldsymbol{y}$ are stored in the `ZXyt` slot in the transposed form. Extracting the transpose of this slot

```
> t(Rm1ML@ZXyt)
18 x 8 sparse Matrix of class "dgCMatrix"

 [1,] . . 1 . . . 1  -55
 [2,] . . 1 . . . 1  -53
 [3,] . . 1 . . . 1  -54
 [4,] 1 . . . . . 1  -26
 [5,] 1 . . . . . 1  -37
 [6,] 1 . . . . . 1  -32
 [7,] . . . . 1 . 1  -78
 [8,] . . . . 1 . 1  -91
 [9,] . . . . 1 . 1  -85
[10,] . . . . . 1 1  -92
[11,] . . . . . 1 1 -100
[12,] . . . . . 1 1  -96
[13,] . 1 . . . . 1  -49
[14,] . 1 . . . . 1  -51
[15,] . 1 . . . . 1  -50
```

```
[16,] . . . 1 . . 1  -80
[17,] . . . 1 . . 1  -85
[18,] . . . 1 . . 1  -83
```

The first 6 columns of this matrix are $\boldsymbol{Z}$, the seventh column is $\boldsymbol{X}$ and the eighth and final column is $-\boldsymbol{y}$. As indicated in the display of the matrix, it is stored as a sparse matrix. The elements represented as '.' are known to be zero and are not stored explicitly.

The columns of $\boldsymbol{Z}$ are indicator columns (that is, the ith column has a 1 in row j if the jth observation is on rail i, otherwise it is zero) but they are not in the usual ordering. This is because the levels of the `Rail` factor have been reordered according to increasing mean response for Figure 1.

The crossproduct of the columns of this matrix are stored as a symmetric, sparse matrix in the `A` slot.

```
> Rm1ML@A
```

```
8 x 8 sparse Matrix of class "dsCMatrix"

[1,]   3    .    .    .    .    .     3   -95
[2,]   .    3    .    .    .    .     3  -150
[3,]   .    .    3    .    .    .     3  -162
[4,]   .    .    .    3    .    .     3  -248
[5,]   .    .    .    .    3    .     3  -254
[6,]   .    .    .    .    .    3     3  -288
[7,]   3    3    3    3    3    3    18 -1197
[8,] -95 -150 -162 -248 -254 -288 -1197 89105
```

The `L` component of this fitted model is a Cholesky factorization of a matrix $\boldsymbol{A}^*(\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is a parameter vector determining $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. This matrix can be factored as $\boldsymbol{\Sigma} = \boldsymbol{T}\boldsymbol{S}\boldsymbol{S}\boldsymbol{T}^\mathsf{T}$, where $\boldsymbol{T}$ is a unit, lower triangular matrix (that is, all the elements above the diagonal are zero and all the elements on the diagonal are unity) and $\boldsymbol{S}$ is a diagonal matrix with non-negative elements on the diagonal. The matrix $\boldsymbol{A}^*(\boldsymbol{\theta})$ is

$$
\begin{aligned}
\boldsymbol{A}^*(\boldsymbol{\theta}) &= \begin{bmatrix} \boldsymbol{Z}^{*\mathsf{T}}\boldsymbol{Z}^* + \boldsymbol{I} & \boldsymbol{Z}^{*\mathsf{T}}\boldsymbol{X} & -\boldsymbol{Z}^{*\mathsf{T}}\boldsymbol{y} \\ \boldsymbol{X}^\mathsf{T}\boldsymbol{Z}^* & \boldsymbol{X}^\mathsf{T}\boldsymbol{X} & -\boldsymbol{X}^\mathsf{T}\boldsymbol{y} \\ -\boldsymbol{y}^\mathsf{T}\boldsymbol{Z}^* & -\boldsymbol{y}^\mathsf{T}\boldsymbol{X} & \boldsymbol{y}^\mathsf{T}\boldsymbol{y} \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{T}^\mathsf{T}\boldsymbol{S} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & 1 \end{bmatrix} \boldsymbol{A} \begin{bmatrix} \boldsymbol{S}\boldsymbol{T} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & 1 \end{bmatrix} + \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & 0 \end{bmatrix}.
\end{aligned}
\tag{3}
$$

For model (1) the matrices $\boldsymbol{T}$ and $\boldsymbol{S}$ are particularly simple, $\boldsymbol{T} = \boldsymbol{I}_6$, the 6×6 identity matrix and $\boldsymbol{S} = s_{1,1}\boldsymbol{I}_6$ where $s_{1,1} = \sigma_b/\sigma$ is the standard

deviation of the random effects relative to the standard deviation of the per-observation noise term $\boldsymbol{\epsilon}$.

The Cholesky decomposition of $\boldsymbol{A}^*$ is a lower triangular sparse matrix $\boldsymbol{L}$

```
> as(Rm1ML@L, "sparseMatrix")
8 x 8 sparse Matrix of class "dtCMatrix"

[1,]   9.797   .       .       .       .       .       .       .
[2,]   .       9.797   .       .       .       .       .       .
[3,]   .       .       9.797   .       .       .       .       .
[4,]   .       .       .       9.797   .       .       .       .
[5,]   .       .       .       .       9.797   .       .       .
[6,]   .       .       .       .       .       9.797   .       .
[7,]   1.723   1.723   1.723   1.723   1.723   1.723   0.4330  .
[8,] -54.562 -86.150 -93.042 -142.435 -145.881 -165.408 -28.7977 17.06
```

As explained in later sections the matrix $\boldsymbol{L}$ provides all the information needed to evaluate the ML deviance or the REML deviance as a function of $\boldsymbol{\theta}$. The components of the deviance are given in the `deviance` slot of the fitted model

```
> Rm1ML@deviance
        ML       REML        ldZ        ldX        lr2
128.560037 122.237086  27.385123  -1.673815   5.673323
```

The element labelled `ldZ` is the logarithm of the square of the determinant of the upper left 6×6 section of $\boldsymbol{L}$. This corresponds to $\log \left| \boldsymbol{Z}^{*\mathsf{T}} \boldsymbol{Z}^* + \boldsymbol{I}_q \right|$ where $\boldsymbol{Z}^* = \boldsymbol{ZTS}$. We can verify that the value 27.38292 can indeed be calculated in this way.

```
> L <- as(Rm1ML@L, "sparseMatrix")
> 2 * sum(log(diag(L)[1:6]))
[1] 27.38512
```

The `lr2` element of the `deviance` slot is the logarithm of the penalized residual sum of squares. It can be calculated as the logarithm of the square of the last diagonal element in $\boldsymbol{L}$.

```
> 2 * log(L[8, 8])
[1] 5.673323
```

For completeness we mention that the `ldX` element of the `deviance` slot is the logarithm of the product of the squares of the diagonal elements of $\boldsymbol{L}$ corresponding to columns of $\boldsymbol{X}$.

```
> 2 * log(L[7, 7])
[1] -1.673815
```

This element is used in the calculation of the REML criterion.

Another slot in the fitted model object is `dims`, which contains information on the dimensions of the model and some of the characteristics of the fit.

```
> Rm1ML@dims
  nf    n    p    q REML glmm
   1   18    1    6    0    0
```

We can reconstruct the ML estimate of the residual variance as the penalized residual sum of squares divided by the number of observations.

```
> exp(Rm1ML@deviance["lr2"])/Rm1ML@dims["n"]
     lr2
16.16667
```

The *profiled deviance* function

$$\begin{aligned}\tilde{\mathcal{D}}(\boldsymbol{\theta}) &= \log\left|\boldsymbol{Z}^{*\mathsf{T}}\boldsymbol{Z}^{*}+\boldsymbol{I}_q\right| + n\log\left(1+\frac{2\pi r^2}{n}\right)\\ &= n\left[1+\log\left(\frac{2\pi}{n}\right)\right] + \log\left|\boldsymbol{Z}^{*\mathsf{T}}\boldsymbol{Z}^{*}+\boldsymbol{I}_q\right| + n\log r^2\end{aligned} \tag{4}$$

is a function of $\boldsymbol{\theta}$ only. In this case $\boldsymbol{\theta} = \sigma_1$, the relative standard deviation of the random effects, is one-dimensional. The maximum likelihood estimate (mle) of $\boldsymbol{\theta}$ minimizes the profiled deviance. The mle's of all the other parameters in the model can be derived from the estimate of this parameters.

The term $n\,[1+\log(2\pi/n)]$ in (4) does not depend on $\boldsymbol{\theta}$. The other two terms, $\log\left|\boldsymbol{Z}^{*\mathsf{T}}\boldsymbol{Z}^{*}+\boldsymbol{I}_q\right|$ and $n\log r^2$, measure the complexity of the model and the fidelity of the fitted values to the observed data, respectively. We plot the value of each of the varying terms versus σ_1 in Figure 2.

The component $\log\left|\boldsymbol{S}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}\boldsymbol{S}+\boldsymbol{I}\right|$ has the value 0 at $\sigma_1 = 0$ and increases as σ_1 increases. It is unbounded as $\sigma_1 \to \infty$. The component $n\log(r^2)$ has a finite value at $\sigma_1 = 1$ from which it decreases as σ_1 increases. The value at $\sigma_1 = 0$ corresponds to the residual sum of squares for the regression of $\boldsymbol{y}$ on the columns of $\boldsymbol{X}$.

```
> 18 * log(deviance(lm(travel ~ 1, Rail)))
[1] 164.8714
```

As $\sigma_1 \to \infty$, $n\log(r^2)$ approaches the value corresponding to the residual sum of squares for the regression of $\boldsymbol{y}$ on the columns of $\boldsymbol{X}$ and $\boldsymbol{Z}$. For this model that is

```
> 18 * log(deviance(lm(travel ~ Rail, Rail)))
[1] 94.82145
```

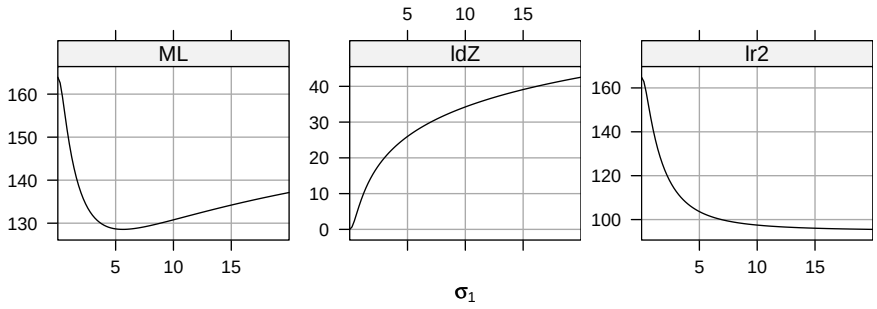

Figure 2: The profiled deviance, and those components of the profiled deviance that vary with $\boldsymbol{\theta}$, as a function of $\boldsymbol{\theta}$ in model `Rm1ML` for the `Rail` data. In this model the parameter $\boldsymbol{\theta}$ is the scalar σ_1, the standard deviation of the random effects relative to the standard deviation of the per-observation noise.

2 Structure of $\boldsymbol{\Sigma}$ and $\boldsymbol{Z}$

The columns of $\boldsymbol{Z}$ and the rows and columns of $\boldsymbol{\Sigma}$ are associated with the levels of one or more grouping factors in the data. For example, a common application of linear mixed models is the analysis of students' scores on the annual state-wide performance tests mandated by the No Child Left Behind Act. A given score is associated with a student, a teacher, a school and a school district. These could all be grouping factors in a model.

We write the grouping factors as $\boldsymbol{f}_i, i = 1, \dots k$. The number of levels of the ith factor, $\boldsymbol{f}_i$, is n_i and the number of random effects associated with each level is q_i. For example, if $\boldsymbol{f}_1$ is "student" then n_1 is the number of students in the study. If we have a simple additive random effect for each student then $q_1 = 1$. If we have a random effect for both the intercept and the slope with respect to time for each student then $q_1 = 2$. The $q_i, i = 1, \dots, k$ are typically very small whereas the $n_i, i = 1, \dots, k$ can be very large.

In the statistical model we assume that random effects associated with different grouping factors are independent, which implies that $\boldsymbol{\Sigma}$ is block diagonal with k diagonal blocks of sizes $n_i q_i \times n_i q_i, i = 1, \dots, k$. That is

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{\Sigma}_k \end{bmatrix} \tag{5}$$

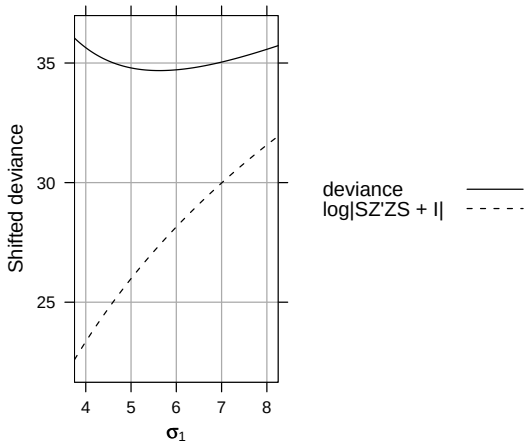

Figure 3: The part of the deviance that varies with σ_1 as a function of σ_1 near the optimum. The component $\log\left|\boldsymbol{S}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}\boldsymbol{S}+\boldsymbol{I}\right|$ is shown at the bottom of the frame. This is the component of the deviance that increases with σ_1. Added to this component is $n\log\left[r^2(\sigma_1)\right]-n\log\left[r^2(\infty)\right]$, the comonent of the deviance that decreases as σ_1 increases. Their sum is minimized at $\widehat{\sigma}_1 = 5.626$.

Furthermore, random effects associated with different levels of the same grouping factor are assumed to be independent and identically distributed, which implies that $\boldsymbol{\Sigma}_i$ is itself block diagonal in n_i blocks and that each of these blocks is a copy of a $q_i \times q_i$ matrix $\tilde{\boldsymbol{\Sigma}}_i$. That is

$$\boldsymbol{\Sigma}_i = \begin{bmatrix} \tilde{\boldsymbol{\Sigma}}_i & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \tilde{\boldsymbol{\Sigma}}_i & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \tilde{\boldsymbol{\Sigma}}_i \end{bmatrix} = \boldsymbol{I}_{n_i} \otimes \tilde{\boldsymbol{\Sigma}}_i \quad i = 1, \dots, k \tag{6}$$

where $\otimes$ denotes the Kronecker product.

The condition that $\boldsymbol{\Sigma}$ is positive semi-definite holds if and only if the $\tilde{\boldsymbol{\Sigma}}_i, i = 1, \dots, k$ are positive semi-definite. To ensure that the $\tilde{\boldsymbol{\Sigma}}_i$ are positive semi-definite, we express them as

$$\tilde{\boldsymbol{\Sigma}}_i = \tilde{\boldsymbol{T}}_i \tilde{\boldsymbol{S}}_i \tilde{\boldsymbol{S}}_i \tilde{\boldsymbol{T}}_i^{\mathsf{T}}, \quad i = 1, \dots, k \tag{7}$$

where $\tilde{\boldsymbol{T}}_i$ is a $q_i \times q_i$ unit lower-triangular matrix (i.e. all the elements above

the diagonal are zero and all the diagonal elements are unity) and $\tilde{\boldsymbol{S}}_i$ is a $q_i \times q_i$ diagonal matrix with non-negative elements on the diagonal.

This is the "LDL" form of the Cholesky decomposition of positive semi-definite matrices except that we express the diagonal matrix $\boldsymbol{D}$, which is on the variance scale, as the square of the diagonal matrix $\boldsymbol{S}$, which is on the standard deviation scale. The profiled deviance behaves more like a quadratic on the standard deviation scale than it does on the variance scale so the use of the standard deviation scale enhances convergence.

The $n_i q_i \times n_i q_i$ matrices $\boldsymbol{S}_i, \boldsymbol{T}_i$, $i = 1, \ldots, k$ and the $q \times q$ matrices $\boldsymbol{S}$ and $\boldsymbol{T}$ are defined analogously to (6) and (5). In particular,

$$\boldsymbol{S}_i = \boldsymbol{I}_{n_i} \otimes \tilde{\boldsymbol{S}}_i, \quad i = 1, \ldots, k \tag{8}$$

$$\boldsymbol{T}_i = \boldsymbol{I}_{n_i} \otimes \tilde{\boldsymbol{T}}_i, \quad i = 1, \ldots, k \tag{9}$$

Note that when $q_i = 1$, $\tilde{\boldsymbol{T}}_i = \boldsymbol{I}$ and hence $\boldsymbol{T}_i = \boldsymbol{I}$. Furthermore, $\boldsymbol{S}_i$ is a multiple of the identity matrix in this case.

The parameter vector $\boldsymbol{\theta}_i, i = 1, \ldots, k$ consists of the q_i diagonal elements of $\tilde{\boldsymbol{S}}_i$, which are constrained to be non-negative, followed by the $q_i(q_i - 1)/2$ elements in the strict lower triangle of $\tilde{\boldsymbol{T}}_i$ (in column-major ordering). These last $q_i(q_i - 1)/2$ elements are unconstrained. The $\boldsymbol{\theta}_i$ are combined as

$$\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \\ \vdots \\ \boldsymbol{\theta}_k \end{bmatrix}.$$

Each of the $q \times q$ matrices $\boldsymbol{S}$, $\boldsymbol{T}$ and $\boldsymbol{\Sigma}$ in the decomposition $\boldsymbol{\Sigma} = \boldsymbol{T}\boldsymbol{S}\boldsymbol{S}\boldsymbol{T}^{\mathsf{T}}$ is a function of $\boldsymbol{\theta}$.

As a unit triangular matrix $\boldsymbol{T}$ is non-singular. That is, $\boldsymbol{T}^{-1}$ exists and is easily calculated from the $\tilde{\boldsymbol{T}}_i^{-1}, i = 1, \ldots, k$. When $\boldsymbol{\theta}$ is not on the boundary defined by the constraints, $\boldsymbol{S}$ is a diagonal matrix with strictly positive elements on the diagonal, which implies that $\boldsymbol{S}^{-1}$ exists and that $\boldsymbol{\Sigma}$ is non-singular with $\boldsymbol{\Sigma}^{-1} = \boldsymbol{T}^{-\mathsf{T}}\boldsymbol{S}^{-1}\boldsymbol{S}^{-1}\boldsymbol{T}^{-1}$.

When $\boldsymbol{\theta}$ is on the boundary the matrices $\boldsymbol{S}$ and $\boldsymbol{\Sigma}$ exist but are not invertible. We say that $\boldsymbol{\Sigma}$ is a *degenerate* variance-covariance matrix in the sense that one or more linear combinations of the vector $\boldsymbol{b}$ are defined to have zero variance. That is, the distribution of these linear combinations is a point mass at 0.

The maximum likelihood estimates of $\boldsymbol{\theta}$ (or the restricted maximum likelihood estimates, defined below) can be located on the boundary. That is, they can correspond to a degenerate variance-covariance matrix and we must be careful to allow for this case. However, to begin we consider the non-degenerate case.

3 Methods for non-singular $\boldsymbol{\Sigma}$

When $\boldsymbol{\theta}$ is not on the boundary we can define a standardized random effects vector

$$\boldsymbol{b}^* = \boldsymbol{S}^{-1}\boldsymbol{T}^{-1}\boldsymbol{b} \tag{10}$$

with the properties

$$\begin{aligned}
\mathsf{E}[\boldsymbol{b}^*] &= \boldsymbol{S}^{-1}\boldsymbol{T}^{-1}\mathsf{E}[\boldsymbol{b}] && (11)\\
\mathsf{Var}[\boldsymbol{b}^*] &= \mathsf{E}[\boldsymbol{b}^*\boldsymbol{b}^{*\mathsf{T}}] && (12)\\
&= \boldsymbol{S}^{-1}\boldsymbol{T}^{-1}\mathsf{Var}[\boldsymbol{b}]\boldsymbol{T}^{-\mathsf{T}}\boldsymbol{S}^{-1}\\
&= \sigma^2\boldsymbol{S}^{-1}\boldsymbol{T}^{-1}\boldsymbol{\Sigma}\boldsymbol{T}^{-\mathsf{T}}\boldsymbol{S}^{-1}\\
&= \sigma^2\boldsymbol{S}^{-1}\boldsymbol{T}^{-1}\boldsymbol{T}\boldsymbol{S}\boldsymbol{S}\boldsymbol{T}^{\mathsf{T}}\boldsymbol{T}^{-\mathsf{T}}\boldsymbol{S}^{-1}\\
&= \sigma^2\boldsymbol{I}.
\end{aligned}$$

Thus, the unconditional distribution of the q elements of $\boldsymbol{b}^*$ is $\boldsymbol{b}^* \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, like that of the n elements of $\boldsymbol{\epsilon}$.

Obviously the transformation from $\boldsymbol{b}^*$ to $\boldsymbol{b}$ is

$$\boldsymbol{b} = \boldsymbol{T}\boldsymbol{S}\boldsymbol{b}^* \tag{13}$$

and the $n \times q$ model matrix for $\boldsymbol{b}^*$ is

$$\boldsymbol{Z}^* = \boldsymbol{Z}\boldsymbol{T}\boldsymbol{S} \tag{14}$$

so that

$$\boldsymbol{Z}^*\boldsymbol{b}^* = \boldsymbol{Z}\boldsymbol{T}\boldsymbol{S}\boldsymbol{S}^{-1}\boldsymbol{T}^{-1}\boldsymbol{b} = \boldsymbol{Z}\boldsymbol{b}. \tag{15}$$

Notice that $\boldsymbol{Z}^*$ can be evaluated even when $\boldsymbol{\theta}$ is on the boundary. Also, if we have a value of $\boldsymbol{b}^*$ in such a case, we can evaluated $\boldsymbol{b}$ from $\boldsymbol{b}^*$.

Given the data $\boldsymbol{y}$ and values of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, the mode of the conditional distribution of $\boldsymbol{b}^*$ is the solution to a penalized least squares problem

$$\begin{aligned}\tilde{\boldsymbol{b}}^*(\boldsymbol{\theta},\boldsymbol{\beta}|\boldsymbol{y}) &= \arg\min_{\boldsymbol{b}^*}\left[\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}^*\boldsymbol{b}^*\|^2+\boldsymbol{b}^{*\mathsf{T}}\boldsymbol{b}^*\right]\\ &= \arg\min_{\boldsymbol{b}^*}\left\|\begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{0}\end{bmatrix}-\begin{bmatrix}\boldsymbol{Z}^* & \boldsymbol{X}\\ \boldsymbol{I} & \boldsymbol{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{b}^*\\ \boldsymbol{\beta}\end{bmatrix}\right\|^2.\end{aligned} \tag{16}$$

In fact, if we optimize the penalized least squares expression in (16) with respect to both $\boldsymbol{b}$ and $\boldsymbol{\beta}$ we obtain the conditional estimates $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}|\boldsymbol{y})$ and the conditional modes $\tilde{\boldsymbol{b}}^*(\boldsymbol{\theta},\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})|\boldsymbol{y}))$ which we write as $\widehat{\boldsymbol{b}}^*(\boldsymbol{\theta})$. That is,

$$\begin{aligned}\begin{bmatrix}\widehat{\boldsymbol{b}}^*(\boldsymbol{\theta})\\ \widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})\end{bmatrix} &= \arg\min_{\boldsymbol{b}^*,\boldsymbol{\beta}}\left\|\begin{bmatrix}\boldsymbol{Z}^* & \boldsymbol{X} & -\boldsymbol{y}\\ \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{b}^*\\ \boldsymbol{\beta}\\ 1\end{bmatrix}\right\|^2\\ &= \arg\min_{\boldsymbol{b}^*,\boldsymbol{\beta}}\begin{bmatrix}\boldsymbol{b}^*\\ \boldsymbol{\beta}\\ 1\end{bmatrix}^{\mathsf{T}}\boldsymbol{A}^*(\boldsymbol{\theta})\begin{bmatrix}\boldsymbol{b}^*\\ \boldsymbol{\beta}\\ 1\end{bmatrix}\end{aligned} \tag{17}$$

where the matrix $\boldsymbol{A}^*(\boldsymbol{\theta})$ is as shown in (3) and

$$\boldsymbol{A}=\begin{bmatrix}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z} & \boldsymbol{Z}^{\mathsf{T}}\boldsymbol{X} & -\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{y}\\ \boldsymbol{X}^{\mathsf{T}}\boldsymbol{Z} & \boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} & -\boldsymbol{X}^{\mathsf{T}}\boldsymbol{y}\\ -\boldsymbol{y}^{\mathsf{T}}\boldsymbol{Z} & -\boldsymbol{y}^{\mathsf{T}}\boldsymbol{X} & \boldsymbol{y}^{\mathsf{T}}\boldsymbol{y}\end{bmatrix}. \tag{18}$$

Note that $\boldsymbol{A}$ does not depend upon $\boldsymbol{\theta}$. Furthermore, the nature of the model matrices $\boldsymbol{Z}$ and $\boldsymbol{X}$ ensures that the pattern of nonzeros in $\boldsymbol{A}^*(\boldsymbol{\theta})$ is the same as that in $\boldsymbol{A}$.

Let the $q\times q$ permutation matrix $\boldsymbol{P}_Z$ represent a fill-reducing permutation for $\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}$ and $\boldsymbol{P}_X$, of size $p\times p$, represent a fill-reducing permutation for $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}$. These could be determined, for example, using the *approximate minimal degree* (AMD) algorithm described in Davis (2006) and Davis (1996) and implemented in both the `Csparse` (Davis, 2005b) and the `CHOLMOD` (Davis, 2005a) libraries of C functions. (In many cases $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}$ is dense, but of small dimension compared to $\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}$, and $\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{X}$ is nearly dense so $\boldsymbol{P}_X$ can be $\boldsymbol{I}_p$, the $p\times p$ identity matrix.)

Let the permutation matrix $\boldsymbol{P}$ be

$$\boldsymbol{P}=\begin{bmatrix}\boldsymbol{P}_Z & \boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{P}_X & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0} & 1\end{bmatrix} \tag{19}$$

and $\boldsymbol{L}(\boldsymbol{\theta})$ be the sparse Cholesky decomposition of $\boldsymbol{A}^*(\boldsymbol{\theta})$ relative to this permutation. That is, $\boldsymbol{L}(\boldsymbol{\theta})$ is a sparse lower triangular matrix with the property that

$$\boldsymbol{L}(\boldsymbol{\theta})\boldsymbol{L}(\boldsymbol{\theta})^\mathsf{T} = \boldsymbol{P}\boldsymbol{A}^*(\boldsymbol{\theta})\boldsymbol{P}^\mathsf{T} \tag{20}$$

For $\boldsymbol{L}(\boldsymbol{\theta})$ to exist we must ensure that $\boldsymbol{A}^*(\boldsymbol{\theta})$ is positive definite. Examination of (17) shows that this will be true if $\boldsymbol{X}$ is of full column rank and $\boldsymbol{y}$ does not lie in the column span of $\boldsymbol{X}$ (or, in statistical terms, if we can't fit $\boldsymbol{y}$ perfectly using only the fixed effects).

Let $r > 0$ be the last element on the diagonal of $\boldsymbol{L}$. Then the minumum penalized residual sum of squares in (17) is r^2 and it occurs at $\widehat{\boldsymbol{b}^*}(\boldsymbol{\theta})$ and $\hat{\beta}(\boldsymbol{\theta})$, the solutions to the sparse triangular system

$$\boldsymbol{L}(\boldsymbol{\theta})^\mathsf{T}\boldsymbol{P}\begin{bmatrix}\widehat{\boldsymbol{b}^*}(\boldsymbol{\theta})\\ \widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})\\ 1\end{bmatrix} = \begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{0}\\ r\end{bmatrix} \tag{21}$$

(Technically we should not write the 1 in the solution; it should be an unknown. However, for $\boldsymbol{L}$ lower triangular with r as the last element on the diagonal and $\boldsymbol{P}$ a permutation that does not move the last row, the solution for this "unknown" will always be 1.) Furthermore, $\log|\boldsymbol{Z}^{*\mathsf{T}}\boldsymbol{Z} + \boldsymbol{I}|$ can be evaluated as the sum of the logarithms of the first q diagonal elements of $\boldsymbol{L}(\boldsymbol{\theta})$.

The *profiled deviance function*, $\tilde{\mathcal{D}}(\boldsymbol{\theta})$, which is negative twice the log-likelihood of model (2) evaluated at $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ and $\hat{\sigma}^2(\boldsymbol{\theta})$, can be expressed as

$$\tilde{\mathcal{D}}(\boldsymbol{\theta}) = \log\left|\boldsymbol{Z}^{*\mathsf{T}}\boldsymbol{Z}^* + \boldsymbol{I}\right| + n\left(1 + \log\frac{2\pi r^2}{n}\right). \tag{22}$$

Notice that it is not necessary to solve for $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ or $\widehat{\boldsymbol{b}^*}(\boldsymbol{\theta})$ or $\widehat{\boldsymbol{b}}(\boldsymbol{\theta})$ to be able to evaluate $d(\boldsymbol{\theta})$. All that is needed is to update $\boldsymbol{A}$ to form $\boldsymbol{A}^*$ from which the sparse Cholesky decomposition $\boldsymbol{L}(\boldsymbol{\theta})$ can be calculated and $\tilde{\mathcal{D}}(\boldsymbol{\theta})$ evaluated.

Once $\hat{\boldsymbol{\theta}}$ is determined we can solve for $\widehat{\boldsymbol{\beta}}(\hat{\boldsymbol{\theta}})$ and $\widehat{\boldsymbol{b}^*}(\boldsymbol{\theta})$ using (21) and for

$$\widehat{\sigma}^2(\hat{\boldsymbol{\theta}}) = \frac{r^2(\hat{\boldsymbol{\theta}})}{n}. \tag{23}$$

Furthermore, $\widehat{\boldsymbol{b}}(\hat{\boldsymbol{\theta}}) = \boldsymbol{S}\boldsymbol{T}\widehat{\boldsymbol{b}^*}(\hat{\boldsymbol{\theta}})$.

4 Methods for singular $\boldsymbol{\Sigma}$

When $\boldsymbol{\theta}$ is on the boundary, corresponding to a singular $\boldsymbol{\Sigma}$, some of the columns of $\boldsymbol{Z}^*$ are zero. However, the matrix $\boldsymbol{A}^*$ is non-singular and elements of $\boldsymbol{b}^*$ corresponding to the zeroed columns in $\boldsymbol{Z}^*$ approach zero smoothly as $\boldsymbol{\theta}$ approaches the boundary. Thus $r(\boldsymbol{\theta})$ and $\left|\boldsymbol{Z}^{*\mathsf{T}}\boldsymbol{Z}+\boldsymbol{I}\right|$ are well-defined, as are $\tilde{\mathcal{D}}(\boldsymbol{\theta})$ and the conditional modes $\widehat{\boldsymbol{b}}(\boldsymbol{\theta})$.

In other words, (3) and (20) can be used to define $\tilde{\mathcal{D}}(\boldsymbol{\theta})$ whether or not $\boldsymbol{\theta}$ is on the boundary.

5 REML estimates

It is common to estimate the per-observation noise variance σ^2 in a fixed-effects linear model as $\hat{\sigma}^2 = r^2/(n-p)$ where r^2 is the (unpenalized) residual sum-of-squares, n is the number of observations and p is the number of fixed-effects parameters. This is not the maximum likelihood estimate of σ^2, which is r^2/n. It is the "restricted" or "residual" maximum likelihood (REML) estimate, which takes into account that the residual vector $\boldsymbol{y}-\hat{\boldsymbol{y}}$ is constrained to a linear subspace of dimension $n-p$ in the response space. Thus its squared length, $\|\boldsymbol{y}-\hat{\boldsymbol{y}}\|^2 = r^2$, has only $n-p$ *degrees of freedom* associated with it.

The profiled REML deviance for a linear mixed model can be expressed as

$$\tilde{\mathcal{D}}_R(\boldsymbol{\theta}) = \log\left|\boldsymbol{Z}^{*\mathsf{T}}\boldsymbol{Z}^*+\boldsymbol{I}\right| + \log|\boldsymbol{L}_{\boldsymbol{X}}|^2 + (n-p)\left(1+\log\frac{2\pi r^2}{n-p}\right). \quad (24)$$

References

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

Tim Davis. CSparse: a concise sparse matrix package. http://www.cise.ufl.edu/research/sparse/CSparse, 2005b.

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.