Linear mixed model implementation in lme4

Douglas Bates
Department of Statistics
University of Wisconsin – Madison
Bates@wisc.edu

September 19, 2006

Abstract

Expressions for the evaluation of the profiled log-likelihood or profiled log-restricted-likelihood of a linear mixed model, the gradients and Hessians of these criteria, and update steps for an ECME algorithm to optimize these criteria are given in Bates and DebRoy (2004). In this paper we generalize those formulae and describe the representation of mixed-effects models using sparse matrix methods available in the Matrix package.

1 Introduction

General formulae for the evaluation of the profiled log-likelihood and profiled log-restricted-likelihood in a linear mixed model are given in Bates and DebRoy (2004). Here we describe a more general formulation of the model using sparse matrix decompositions and describe the implementation of these methods in the lmer function for R.

In §2 we describe the form and representation of the model. The calculation of the criteria to be optimized by the parameter estimates and related quantities is discussed in §3.

2 Form and representation of the model

We consider linear mixed models of the form

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zb} + \boldsymbol{\epsilon} \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}), \boldsymbol{b} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}), \boldsymbol{\epsilon} \perp \boldsymbol{b} \tag{1}$$

where $\boldsymbol{y}$ is the n-dimensional response vector, $\boldsymbol{X}$ is an $n \times p$ model matrix for the p dimensional fixed-effects vector $\boldsymbol{\beta}$, $\boldsymbol{Z}$ is the $n \times q$ model matrix for the q dimensional random-effects vector $\boldsymbol{b}$, which has a Gaussian distribution with mean $\boldsymbol{0}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, and $\boldsymbol{\epsilon}$ is the random noise assumed to have a spherical Gaussian distribution. The symbol $\perp$ indicates independence of random variables.

We will assume that $\boldsymbol{X}$ has full column rank and that $\boldsymbol{\Sigma}$ is positive definite.

2.1 Structure of the variance-covariance matrix

Components of the random effects vector $\boldsymbol{b}$ and portions of its variance-covariance matrix $\boldsymbol{\Sigma}$ are associated with k grouping factors $\boldsymbol{f}_i, i = 1, \ldots, k$, each of length n, and with the $n_i, i = 1, \ldots, k$ levels of each of the grouping factors. In general there are q_i components of $\boldsymbol{b}$ associated with each of the n_i levels the grouping factor $\boldsymbol{f}_i, i = 1, \ldots, k$. Thus

$$q = \sum_{i=1}^{k} n_i q_i \tag{2}$$

We assume that the components of $\boldsymbol{b}$ and the rows and columns of $\boldsymbol{\Sigma}$ are ordered according to the k grouping factors and, within the block for the ith grouping factor, according to the n_i levels of the grouping factor.

Random effects associated with different grouping factors are independent. This implies that $\boldsymbol{\Sigma}$ is block-diagonal with k diagonal blocks of orders $n_i q_i, i = 1, \ldots, k$.

Random effects associated with different levels of the same grouping factor are independent. This implies that the ith (outer) diagonal block of $\boldsymbol{\Sigma}$ is itself block diagonal in n_i blocks of order q_i. We say that the structure of $\boldsymbol{\Sigma}$ is block/block diagonal.

Finally, the variance-covariance matrix within each of the q_i-dimensional subvectors of $\boldsymbol{b}$ associated with one of the n_i levels of grouping factor $\boldsymbol{f}_i, i = 1, \ldots, k$ is a constant (but unknown) positive-definite symmetric $q_i \times q_i$ matrix

$\mathbf{\Sigma}_i, i = 1, \dots, k$. This implies that each of the n_i inner diagonal blocks of order q_i is a copy of $\mathbf{\Sigma}_i$. We say that $\mathbf{\Sigma}$ has a *repeated block/block diagonal* structure.

In the notation of the Kronecker product, the ith outer diagonal block of $\mathbf{\Sigma}$ is $\boldsymbol{I}_{n_i} \otimes \mathbf{\Sigma}_i$.

2.2 The relative precision matrix

Many of the computational formulae that follow are more conveniently expressed in terms of $\mathbf{\Sigma}^{-1}$, which is called the *precision* matrix of the random effects, than in terms of $\mathbf{\Sigma}$, the variance-covariance matrix. In fact, the formulae are most conveniently expressed in terms of the *relative precision matrix* $\sigma^2\mathbf{\Sigma}^{-1}$ which we write as $\mathbf{\Omega}$. That is,

$$\mathbf{\Omega} = \sigma^2\mathbf{\Sigma}^{-1} \tag{3}$$

This called the "relative" precision because it is precision of $\boldsymbol{b}$ (i.e. $\mathbf{\Sigma}^{-1}$) relative to the precision of $\boldsymbol{\epsilon}$ (i.e. $\sigma^{-2}\boldsymbol{I}$).

It is easy to establish that $\mathbf{\Omega}$ will have a repeated block/block diagonal structure like that of $\mathbf{\Sigma}$. That is, $\mathbf{\Omega}$ consists of k outer diagonal blocks of sizes $n_i q_i, i = 1, \dots, k$ and the ith outer diagonal block is itself block diagonal with n_i inner blocks of size $q_i \times q_i$. Furthermore, each of the inner diagonal blocks in the ith outer block is a copy of the $q_i \times q_i$ positive-definite, symmetric matrix $\mathbf{\Omega}_i$.

Because $\mathbf{\Omega}$ has a repeated block/block structure we can define the entire matrix by specifying the symmetric matrices $\mathbf{\Omega}_i, i = 1, \dots, k$ and, because of the symmetry, $\mathbf{\Omega}_i$ has at most $q_i(q_i + 1)/2$ distinct elements. We will write a parameter vector of length at most $\sum_{i=1}^{k} q_i(q_i + 1)/2$ that determines $\mathbf{\Omega}$ as $\boldsymbol{\theta}$. For example, we could define $\boldsymbol{\theta}$ to be the non-redundant elements in the $\mathbf{\Omega}_i$, although in the actual computations we use a different, but equivalent, parameterization for reasons to be discussed later.

We only need to store the matrices $\mathbf{\Omega}_i, i = 1, \dots, k$ and the number of levels in the grouping factors to be able to create $\mathbf{\Omega}$. The matrices $\mathbf{\Omega}_i$ are stored in the `Omega` slot of an object of class `"lmer"`. The values of k and $n_i, i = 1, \dots, k$ can be determined from the list of the grouping factors themselves, stored as the `flist` slot, or from the dimensions $q_i, i = 1, \dots, k$, stored in the `nc` slot, and the group pointers, stored in the `Gp` slot. The group pointers are the (0-based) indices of the first component of $\boldsymbol{b}$ associated with the ith grouping factor. The last element of `Gp` is the number of elements in $\boldsymbol{b}$.

Thus successive differences of the group pointers are the total number of components of $\boldsymbol{b}$ associated with the ith grouping factor. That is, these differences are $n_i q_i, i = 1, \ldots, k$. The first element of the `Gp` slot is always 0.

2.3 Examples

Consider the fitted models

```
> Sm1 <- lmer(Reaction ~ Days + (Days | Subject), sleepstudy)
> data(Chem97, package = "mlmRev")
> Cm1 <- lmer(score ~ gcsescore + (1 | school) + (1 | lea), Chem97,
+     control = list(niterEM = 0, gradient = FALSE))
> data(star, package = "mlmRev")
> Mm1 <- lmer(math ~ gr + sx * eth + cltype + (yrs | id) + (1 |
+     tch) + (yrs | sch), star, control = list(niterEM = 0, gradient = FALSE))
```

Model `Sm1` has a single grouping factor with 18 levels and $q_1 = 2$. The `Omega` slot is a list of length one containing a 2×2 symmetric matrix. There are 36 elements in $\boldsymbol{b}$.

```
> str(Sm1@flist)

List of 1
 $ Subject: Factor w/ 18 levels "308","309","310",..: 1 1 1 1 1 1 1 1 1 1 ...

> show(Sm1@Omega)

$Subject
2 x 2 Matrix of class "dpoMatrix"
            (Intercept)       Days
(Intercept)   1.0746169 -0.2943005
Days         -0.2943005 18.7549833

> show(Sm1@nc)

Subject
      2

> show(Sm1@Gp)

[1]  0 36

> diff(Sm1@Gp)/Sm1@nc

Subject
     18
```

Model `Cm1` has two grouping factors: the `school` factor with 2410 levels and the `lea` factor (local education authority - similar to a school district in the U.S.A.) with 131 levels. It happens that the `school` factor is nested within the `lea` factor, a property that we discuss below. The `Omega` slot is a list of length two containing two 1×1 symmetric matrices.

```
> str(Cm1@flist)
```

```
List of 2
 $ school: Factor w/ 2410 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ lea   : Factor w/ 131 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
> show(Cm1@Omega)
$school
1 x 1 Matrix of class "dpoMatrix"
            (Intercept)
(Intercept)    4.419665

$lea
1 x 1 Matrix of class "dpoMatrix"
            (Intercept)
(Intercept)    349.0649
> show(Cm1@nc)
school    lea
     1      1
> show(Cm1@Gp)
[1]    0 2410 2541
> diff(Cm1@Gp)/Cm1@nc
school    lea
  2410    131
```

Model `Mm1` has three grouping factors: `id` (student) with 10732 levels, `tch` (teacher) with 1374 levels and `sch` (school) with 80 levels. The `Omega` slot is a list of length three containing two 2×2 symmetric matrices and one 1×1 matrix.

```
> str(Mm1@flist)
List of 3
 $ id : Factor w/ 10732 levels "100017","100028",..: 1 2 3 3 3 4 5 5 6 6 ...
 $ tch: Factor w/ 1374 levels "1","2","3","4",..: 476 889 695 698 703 1097 676 681 349 357 ...
 $ sch: Factor w/ 80 levels "1","2","3","4",..: 28 52 41 41 41 64 40 40 22 22 ...
> show(Mm1@Omega)
$id
2 x 2 Matrix of class "dpoMatrix"
            (Intercept)       yrs
(Intercept)   0.3320385 0.4956119
yrs           0.4956119 8.1874244

$tch
1 x 1 Matrix of class "dpoMatrix"
            (Intercept)
(Intercept)    1.425591

$sch
2 x 2 Matrix of class "dpoMatrix"
            (Intercept)       yrs
(Intercept)    3.289128  6.069419
yrs            6.069419 18.655288
> show(Mm1@nc)
 id tch sch
  2   1   2
```

Name	n	p	k	n_1	q_1	n_2	q_2	n_3	q_3	q	$\#(\boldsymbol{\theta})$
Sm1	180	2	1	18	2					36	3
Cm1	31022	2	2	2410	1	131	1			2541	2
Mm1	24578	17	3	10732	2	1374	1	80	2	22998	7

Table 1: Dimensions of model matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ for example model fits. The model matrix $\boldsymbol{X}$ is $n \times p$ and dense. The model matrix $\boldsymbol{Z}$ is $n \times q$ and sparse. The variance-covariance matrix $\boldsymbol{\Sigma}$ of the random effects $\boldsymbol{b}$ is $q \times q$ and repeated block/block diagonal with k outer blocks of sizes $n_i q_i, i = 1, \dots, k$ each consisting of n_i inner blocks of size $q_i \times q_i$. The matrix $\boldsymbol{\Sigma}$ is determined by a parameter $\boldsymbol{\theta}$ whose length is shown in the table.

```
> show(Mm1@Gp)
[1]     0 21464 22838 22998
> diff(Mm1@Gp)/Mm1@nc
   id   tch   sch
10732  1374    80
```

The last element of the Gp slot is the dimension of $\boldsymbol{b}$. Notice that for model Mm1 the dimension of $\boldsymbol{b}$ is 22,998. This is also the order of the symmetric matrix $\boldsymbol{\Omega}$ although the contents of the matrix are determined by $\boldsymbol{\theta}$ which has a length of $3 + 1 + 3 = 7$ in this case.

Table 1 summarizes some of the dimensions in these examples.

2.4 Permutation of the random-effects vector

For most mixed-effects model fits, the model matrix $\boldsymbol{Z}$ for the random effects vector $\boldsymbol{b}$ is large and sparse. That is, most of the entries in $\boldsymbol{Z}$ are zero (by design, not by accident).

Numerical analysts have developed special techniques for representing and manipulating sparse matrices. Of particular importance to us are techniques for obtaining the left Cholesky factor $\boldsymbol{L}$ of large, sparse, positive-definite, symmetric matrices. In particular, we want to obtain the Cholesky factorization of $\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z} + \boldsymbol{\Omega}(\boldsymbol{\theta})$ for different values of $\boldsymbol{\theta}$.

Sparse matrix operations are typically performed in two phases: a *symbolic phase*, in which the number of non-zero elements in the result and their positions are determined, followed by a *numeric phase*, in which the actual numeric values are calculated. Advanced sparse Cholesky factorization software, such as the CHOLMOD library (Davis, 2005) that we use, allow for

			$\boldsymbol{Z}$		$\boldsymbol{Z}^\top\boldsymbol{Z}$		$\boldsymbol{L}$	
Name	n	q	nz	sp	nz	sp	nz	sp
`Sm1`	180	36	360	0.0556	54	0.0811	54	0.0811
`Cm1`	31022	2541	62044	0.0008	4951	0.0015	13021	0.0040
`Mm1`	24578	22998	1222890	0.0002	130138	0.0005	187959	0.0007

Table 2: Summary of the sparsity of the model matrix $\boldsymbol{Z}$, its crossproduct matrix $\boldsymbol{Z}^\top\boldsymbol{Z}$ and the left Cholesky factor $\boldsymbol{L}$ in the examples. The notation nz indicates the number of nonzeros in the matrix and sp indicates the sparsity index (the fraction of the elements in the matrix that are non-zero). Because $\boldsymbol{Z}^\top\boldsymbol{Z}$ is symmetric, only the nonzeros in the upper triangle are counted and the sparsity index is relative to the total number of elements in the upper triangle.

calculation of a fill-reducing permutation of the rows and columns during the symbolic phase. In fact the CHOLMOD code allows for evaluation of both a fill-reducing permutation and a post-ordering that groups together columns of $\boldsymbol{L}$ with identical patterns of nonzeros, thus allowing for dense matrix techniques to be used on these blocks of columns or "super-nodes". Such a decomposition is called a *supernodal* Cholesky factorization.

Because the number of nonzeros in $\boldsymbol{\Omega}(\boldsymbol{\theta})$ and their positions do not change with $\boldsymbol{\theta}$ and because the nonzeros in $\boldsymbol{\Omega}(\boldsymbol{\theta})$ are a subset of the nonzeros in $\boldsymbol{Z}^\top\boldsymbol{Z}$, we need only perform the symbolic phase once and we can do on $\boldsymbol{Z}^\top$ (the CHOLMOD library has a module that calculates the permutation for a super-nodal decomposition of $\boldsymbol{Z}^\top\boldsymbol{Z}$ from $\boldsymbol{Z}^\top$). That is, using $\boldsymbol{Z}^\top$ only we can determine the permutation matrix $\boldsymbol{P}$ for all supernodal decompositions of the form

$$\boldsymbol{P}\left[\boldsymbol{Z}^\top\boldsymbol{Z}+\boldsymbol{\Omega}(\boldsymbol{\theta})\right]\boldsymbol{P}^\top=\boldsymbol{L}(\boldsymbol{\theta})\boldsymbol{L}(\boldsymbol{\theta})^\top \tag{4}$$

We revise (1) by incorporating the permutation to obtain

$$\boldsymbol{y}=\boldsymbol{X\beta}+\boldsymbol{ZP}^\top\boldsymbol{Pb}+\boldsymbol{\epsilon}\quad \boldsymbol{\epsilon}\sim\mathcal{N}(\boldsymbol{0},\sigma^2\boldsymbol{I}),\boldsymbol{b}\sim\mathcal{N}(\boldsymbol{0},\sigma^2\boldsymbol{\Sigma}),\boldsymbol{\epsilon}\perp\boldsymbol{b} \tag{5}$$

2.5 Extent of the sparsity

Table 2 shows the extent of the sparsity of the matrices $\boldsymbol{Z}$, $\boldsymbol{Z}^\top\boldsymbol{Z}$ and $\boldsymbol{L}$ in our examples.

The matrix $\boldsymbol{L}$ is the supernodal representation of the left Cholesky factor of $\boldsymbol{P}\left(\boldsymbol{Z}^\top\boldsymbol{Z}+\boldsymbol{\Omega}\right)\boldsymbol{P}^\top$. Because the fill-reducing permutation $\boldsymbol{P}$ has been applied the number of nonzeros in $\boldsymbol{L}$ will generally be less than the number of nonzeros in the left Cholesky factor of $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\boldsymbol{\Omega}$. However, when any supernodes of $\boldsymbol{L}$ contain more than one column there will be elements above the diagonal of $\boldsymbol{L}$ stored and these elements are necessarily zero. They are stored in the supernodal factorization so that the diagonal block for a supernode can be treated as a dense rectangular matrix. Although these elements are stored in the structure they are never used because any calculations involving the diagonal blocks take into account its being a lower triangular matrix. We do not count these elements as nonzeros in computing the size of $\boldsymbol{L}$ or the sparsity index.

In model `Sm1` the number of nonzeros in $\boldsymbol{L}$ is equal to the number of nonzeros in $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}$. That is, there is no fill-in. In model `Mm1` the number of nonzeros in $\boldsymbol{L}$ is approximately 144% the number of nonzeros in $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}$ representing a modest amount of fill-in. For model `Cm1` the number of nonzeros in $\boldsymbol{L}$ is apparently 263% the number of nonzeros in $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}$, which is still not dramatic. However, it is misleading in that the extra "nonzeros" are, in fact, systematic zeros. Models based on a nested sequence of grouping factors do not generate any fill-in but the pattern in the factor $\boldsymbol{L}$ is not of the type that can be detected and accomodated by standard algorithms for sparse matrices.

3 Likelihood and restricted likelihood

In general the *maximum likelihood estimates* of the parameters in a statistical model are those values of the parameters that maximize the likelihood function, which is the same numerical value as the probability density of $\boldsymbol{y}$ given the parameters but regarded as a function of the parameters given $\boldsymbol{y}$, not as a function of $\boldsymbol{y}$ given the parameters.

For model (5) the parameters are $\boldsymbol{\beta}$, σ^2 and $\boldsymbol{\theta}$ (as described in §2.2, $\boldsymbol{\theta}$ and σ^2 jointly determine $\boldsymbol{\Sigma}$) so we evaluate the likelihood $L(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}|\boldsymbol{y})$ as

$$L(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}|\boldsymbol{y}) = f_{\boldsymbol{y}|\boldsymbol{\beta},\sigma^2,\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}) \tag{6}$$

where $f_{\boldsymbol{y}|\boldsymbol{\beta},\sigma^2,\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta})$ is the marginal probability density for $\boldsymbol{y}$ given the parameters.

Because we will need to write several different marginal and conditional probability densities in this section, and because expressions like $f_{\boldsymbol{y}|\boldsymbol{\beta},\sigma^2,\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta})$ are difficult to read, we will adopt a convention sometimes used in the

Bayesian inference literature that a conditional expression in square brackets indicates the probability density of the quantity on the left of the | given the quantities on the right of the |. That is

$$\left[\boldsymbol{y}|\boldsymbol{\beta},\sigma^2,\boldsymbol{\theta}\right] = f_{\boldsymbol{y}|\boldsymbol{\beta},\sigma^2,\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2,\boldsymbol{\theta}) \tag{7}$$

Model (5) specifies the conditional distributions

$$\left[\boldsymbol{y}|\boldsymbol{\beta},\sigma^2,\boldsymbol{b}\right] = \frac{\exp\left\{-\|\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{ZP}^\mathsf{T}\boldsymbol{Pb}\|^2/\left(2\sigma^2\right)\right\}}{\left(2\pi\sigma^2\right)^{n/2}} \tag{8}$$

and

$$\begin{aligned}\left[\boldsymbol{b}|\boldsymbol{\theta},\sigma^2\right] &= \frac{\exp\left\{-\boldsymbol{b}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{b}/2\right\}}{|\boldsymbol{\Sigma}|^{1/2}\left(2\pi\right)^{q/2}} \\ &= \frac{|\boldsymbol{\Omega}|^{1/2}\exp\left\{-\boldsymbol{b}^\mathsf{T}\boldsymbol{\Omega}\boldsymbol{b}/\left(2\sigma^2\right)\right\}}{\left(2\pi\sigma^2\right)^{q/2}}\end{aligned} \tag{9}$$

from which we can derive the marginal distribution

$$\begin{aligned}&\left[\boldsymbol{y}|\boldsymbol{\beta},\sigma^2,\boldsymbol{\theta}\right] = \int_{\boldsymbol{b}}\left[\boldsymbol{y}|\boldsymbol{\beta},\sigma^2,\boldsymbol{b}\right]\left[\boldsymbol{b}|\boldsymbol{\theta},\sigma^2\right]\,d\boldsymbol{b} \\ &= \frac{|\boldsymbol{\Omega}|^{1/2}}{\left(2\pi\sigma^2\right)^{n/2}}\int_{\boldsymbol{b}}\frac{\exp\left\{-\left[\|\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{ZP}^\mathsf{T}\boldsymbol{Pb}\|^2+\boldsymbol{b}^\mathsf{T}\boldsymbol{\Omega}\boldsymbol{b}\right]/\left(2\sigma^2\right)\right\}}{\left(2\pi\sigma^2\right)^{q/2}}\,d\boldsymbol{b}.\end{aligned} \tag{10}$$

3.1 A penalized least squares representation

To evaluate the integral in (10) we expand the expression in the numerator of the exponent

$$\begin{aligned}g(\boldsymbol{b},\boldsymbol{\beta}|\boldsymbol{Z},\boldsymbol{X},\boldsymbol{y},\boldsymbol{P}) &= \|\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{ZP}^\mathsf{T}\boldsymbol{Pb}\|^2+\boldsymbol{b}^\mathsf{T}\boldsymbol{P}^\mathsf{T}\boldsymbol{P\Omega P}^\mathsf{T}\boldsymbol{Pb} \\ &= \left\|\begin{bmatrix}\boldsymbol{ZP}^\mathsf{T} & \boldsymbol{X} & \boldsymbol{y}\end{bmatrix}\begin{bmatrix}-\boldsymbol{Pb}\\ -\boldsymbol{\beta}\\ -1\end{bmatrix}\right\|^2+\boldsymbol{b}^\mathsf{T}\boldsymbol{P}^\mathsf{T}\boldsymbol{P\Omega P}^\mathsf{T}\boldsymbol{Pb} \\ &= \begin{bmatrix}-\boldsymbol{Pb}\\ -\boldsymbol{\beta}\\ -1\end{bmatrix}^\mathsf{T}\begin{bmatrix}\boldsymbol{P}\left(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\boldsymbol{\Omega}\right)\boldsymbol{P}^\mathsf{T} & \boldsymbol{PZ}^\mathsf{T}\boldsymbol{X} & \boldsymbol{PZ}^\mathsf{T}\boldsymbol{y}\\ \boldsymbol{X}^\mathsf{T}\boldsymbol{ZP}^\mathsf{T} & \boldsymbol{X}^\mathsf{T}\boldsymbol{X} & \boldsymbol{X}^\mathsf{T}\boldsymbol{y}\\ \boldsymbol{y}^\mathsf{T}\boldsymbol{ZP}^\mathsf{T} & \boldsymbol{y}^\mathsf{T}\boldsymbol{X} & \boldsymbol{y}^\mathsf{T}\boldsymbol{y}\end{bmatrix}\begin{bmatrix}-\boldsymbol{Pb}\\ -\boldsymbol{\beta}\\ -1\end{bmatrix}\end{aligned} \tag{11}$$

from which we see that the expression is a quadratic form.

As we have already indicated, we simplify the quadratic form by taking a Cholesky decomposition of the positive-definite, symmetric matrix defining the form. We write this as

$$\begin{bmatrix} \boldsymbol{P}\left(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\boldsymbol{\Omega}\right)\boldsymbol{P}^\mathsf{T} & \boldsymbol{P}\boldsymbol{Z}^\mathsf{T}\boldsymbol{X} & \boldsymbol{P}\boldsymbol{Z}^\mathsf{T}\boldsymbol{y} \\ \boldsymbol{X}^\mathsf{T}\boldsymbol{Z}\boldsymbol{P}^\mathsf{T} & \boldsymbol{X}^\mathsf{T}\boldsymbol{X} & \boldsymbol{X}^\mathsf{T}\boldsymbol{y} \\ \boldsymbol{y}^\mathsf{T}\boldsymbol{Z}\boldsymbol{P}^\mathsf{T} & \boldsymbol{y}^\mathsf{T}\boldsymbol{X} & \boldsymbol{y}^\mathsf{T}\boldsymbol{y} \end{bmatrix} = \begin{bmatrix} \boldsymbol{L} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{R}_{ZX}^\mathsf{T} & \boldsymbol{R}_{XX}^\mathsf{T} & \boldsymbol{0} \\ \boldsymbol{r}_{Zy}^\mathsf{T} & \boldsymbol{r}_{Xy}^\mathsf{T} & r_{yy} \end{bmatrix} \begin{bmatrix} \boldsymbol{L}^\mathsf{T} & \boldsymbol{R}_{ZX} & \boldsymbol{r}_{Zy} \\ \boldsymbol{0} & \boldsymbol{R}_{XX} & \boldsymbol{r}_{Xy} \\ \boldsymbol{0} & \boldsymbol{0} & r_{yy} \end{bmatrix} \quad (12)$$

which gives

$$g(\boldsymbol{b},\boldsymbol{\beta}|\boldsymbol{Z},\boldsymbol{X},\boldsymbol{y},\boldsymbol{P}) = \|\boldsymbol{r}_{Zy}-\boldsymbol{R}_{ZX}\boldsymbol{\beta}-\boldsymbol{L}^\mathsf{T}\boldsymbol{P}\boldsymbol{b}\|^2+\|\boldsymbol{r}_{Xy}-\boldsymbol{R}_{XX}\boldsymbol{\beta}\|^2+r_{yy}^2. \quad (13)$$

The last two terms in (13) do not depend on $\boldsymbol{b}$ so the integral in (10) can be evaluated if we evaluate

$$\int_{\boldsymbol{b}} \frac{\exp\left\{-\|\boldsymbol{r}_{Zy}-\boldsymbol{R}_{ZX}\boldsymbol{\beta}-\boldsymbol{L}^\mathsf{T}\boldsymbol{P}\boldsymbol{b}\|^2/\left(2\sigma^2\right)\right\}}{\left(2\pi\sigma^2\right)^{q/2}}\,d\boldsymbol{b}$$

which we do with a change of variable

$$\boldsymbol{v} = \boldsymbol{L}\boldsymbol{P}\boldsymbol{b}$$

for which the Jacobian is

$$\left|\frac{d\boldsymbol{v}}{d\boldsymbol{b}}\right| = \sqrt{|\boldsymbol{L}\boldsymbol{P}|^2} = \sqrt{|\boldsymbol{L}|^2} = |\boldsymbol{L}\boldsymbol{L}^\mathsf{T}|^{1/2} = |\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\boldsymbol{\Omega}|^{1/2}$$

Thus

$$\begin{aligned} \int_{\boldsymbol{b}} & \frac{\exp\left\{-\|\boldsymbol{r}_{Zy}-\boldsymbol{R}_{ZX}\boldsymbol{\beta}-\boldsymbol{L}^\mathsf{T}\boldsymbol{P}\boldsymbol{b}\|^2/\left(2\sigma^2\right)\right\}}{\left(2\pi\sigma^2\right)^{q/2}}\,d\boldsymbol{b} \\ &= \frac{1}{|\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\boldsymbol{\Omega}|^{1/2}} \int_{\boldsymbol{v}} \frac{\exp\left\{-\|\boldsymbol{r}_{Zy}-\boldsymbol{R}_{ZX}\boldsymbol{\beta}-\boldsymbol{v}\|^2/\left(2\sigma^2\right)\right\}}{\left(2\pi\sigma^2\right)^{q/2}}\,d\boldsymbol{v} \\ &= \frac{1}{|\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\boldsymbol{\Omega}|^{1/2}} \end{aligned} \quad (14)$$

because the integral with respect to $\boldsymbol{v}$ is the integral of a q-dimensional multivariate normal density.

3.2 Likelihood results

Substituting (13) and (14) into (10) we can evaluate the likelihood $L(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}|\boldsymbol{y})$. As often happens, it is easier to write the *log-likelihood*

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

and even easier to write the result on the *deviance* scale as

$$\begin{aligned} &-2\ell(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}|\boldsymbol{y}) \\ &\qquad = \log\left(\frac{|\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} + \boldsymbol{\Omega}|}{|\boldsymbol{\Omega}|}\right) + \frac{r_{yy}^2}{\sigma^2} + \frac{\|\boldsymbol{r}_{Xy} - \boldsymbol{R}_{XX}\boldsymbol{\beta}\|^2}{\sigma^2} + n\log(2\pi\sigma^2) \end{aligned} \qquad (15)$$

The maximum likelihood estimators $[\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\beta}}, \widehat{\sigma^2}]$ minimize the deviance expression (15), which has some properties that can be used to simplify the optimization process. In particular,

1. The conditional estimates of $\boldsymbol{\beta}$ satisfy
$$\boldsymbol{R}_{XX}\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \boldsymbol{r}_{Xy}. \qquad (16)$$

2. The conditional modes (which are also the means) of the random effects $\boldsymbol{b}$ satisfy
$$\boldsymbol{L}^\mathsf{T}\boldsymbol{P}\widehat{\boldsymbol{b}}(\boldsymbol{\theta}, \boldsymbol{\beta}) = \boldsymbol{r}_{Zy} - \boldsymbol{R}_{ZX}\boldsymbol{\beta}. \qquad (17)$$
Usually we want to evaluate these at $\boldsymbol{\theta}$ and $\widehat{\boldsymbol{\beta}(\boldsymbol{\theta})}$, which we write as $\widehat{\boldsymbol{b}}(\boldsymbol{\theta}) = \widehat{\boldsymbol{b}}\,(\boldsymbol{\theta}, \boldsymbol{\beta}(\boldsymbol{\theta}))$.

3. The conditional ML estimate of σ^2 is
$$\widehat{\sigma^2}(\boldsymbol{\theta}) = r_{yy}^2/n. \qquad (18)$$

4. The profiled ML deviance, which is a function of $\boldsymbol{\theta}$ only produced by plugging in the conditional estimates for $\boldsymbol{\beta}$ and σ^2, is
$$\log\left(\frac{|\boldsymbol{L}|^2}{|\boldsymbol{\Omega}|}\right) + n\left[1 + \log\left(\frac{2\pi r_{yy}^2}{n}\right)\right] \qquad (19)$$

5. The profiled REML deviance is
$$\log\left(\frac{|\boldsymbol{D}|\,|\boldsymbol{R}_{XX}|^2}{|\boldsymbol{\Omega}|}\right) + (n-p)\left[1 + \log\left(\frac{2\pi r_{yy}^2}{n-p}\right)\right]$$

References

Douglas M. Bates and Saikat DebRoy. Linear mixed models and penalized least squares. *J. of Multivariate Analysis*, 2004. to appear.