Sparse Matrix Representations of Linear Mixed Models

Douglas Bates
R Development Core Team
Douglas.Bates@R-project.org

June 1, 2004

Abstract

We describe a representation of linear mixed-effects models using positive semidefinite, symmetric, compressed, column-oriented, sparse matrices. This representation provides for efficient evaluation of the profiled log-likelihood or profiled restricted log-likelihood of the model, given the relative precision parameters of the random effects. The evaluation is based upon the $\boldsymbol{LDL}^\mathsf{T}$ form of the Cholesky decomposition of the augmented sparse representation. Additionally, we can use information from this representation to evaluate ECME updates and the gradient of the criterion being optimized.

The sparse matrix methods that we employ have both a symbolic phase, in which the number and positions of the off-diagonal elements are determined, and a numeric phase, in which the actual numeric values are determined. The symbolic phase need only be done once and it can be accomplished knowing only the grouping factors with which the random effects are associated. An important part of the symbolic phase is determination of a fill-minimizing permutation of the rows and columns of a sparse semi-definite matrix. This matrix have a special structure in the linear mixed-effects problem and we provide a new fill-minimizing algorithm tuned to this structure.

1 Introduction

Mixed-effects models, also called multilevel models, panel data models, and frailty models, are widely used in many areas of statistical applications (Pinheiro and Bates, 2000). The basic form of the model, the linear mixed model, also serves as an approximation in iterative estimation of the parameters in more general forms such as the generalized linear mixed model (GLMM) and the nonlinear mixed model (NMM).

In §2 we define a general form of a linear mixed model using grouping factors and model matrices that are associated with the grouping factors. This form, which can be used for multiple levels of random effects in either nested or crossed configurations, can be represented and manipulated using a sparse, symmetric, semidefinite matrix and several dense matrices. We show that a profiled log-likelihood can be evaluated from that solution of a penalized least squares problem and that this solution can be obtained from the Cholesky decomposition of an augmented form of the sparse, symmetric matrix.

Many implementations of the Cholesky decomposition of sparse, symmetric, semidefinite matrices have both a symbolic phase, in which the number and positions of the off-diagonal elements are determined, and a numeric phase, in which the actual numeric values are determined. In §3 we show that the symbolic analysis for the matrices we consider need only be done once and can be accomplished knowing only the grouping factors. An important part of the symbolic phase is determination of a fill-reducing permutation of the rows and columns of the symmetric matrix. We show that by suitably ordering the grouping factors and by restricting ourselves to permutations that correspond to reorderings of the levels within the grouping factors we can determine effective fill-reducing orderings.

Finally, in §**??** we show how these methods can be used to implement general penalized least squares approaches to models such as the GLMM and the NMM and then to implement more accurate approximations to the marginal likelihood using Laplacian integration or adaptive Gauss-Hermite integration.

2 Linear mixed models

We describe the form of the linear mixed-effects model that we consider and restate some of the formulas from Bates and DebRoy (2004) using the LDL form of the Cholesky decomposition of a sparse, semi-definite matrix.

2.1 Form of the model

We consider linear mixed-effects models that can be written as

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

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 $\boldsymbol{b}$ that has a Gaussian distribution with mean $\boldsymbol{0}$ and relative precision matrix $\boldsymbol{\Omega}$ (i.e., $\boldsymbol{\Omega}$ is the precision of $\boldsymbol{b}$ relative to the precision of $\boldsymbol{\epsilon}$), and $\boldsymbol{\epsilon}$ is the random noise assumed to have a spherical Gaussian distribution. The symbol $\perp$ indicates independence of random variables. We assume that $\boldsymbol{X}$ has full column rank and that $\boldsymbol{\Omega}$, which is a function of an (unconstrained) parameter vector $\boldsymbol{\theta}$, is positive definite.

2.1.1 Grouping factors for the random effects

Although q, the dimension of the vector $\boldsymbol{b}$ (and, correspondingly, the number of columns in $\boldsymbol{Z}$ and the number of rows and columns in $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}$ and $\boldsymbol{\Omega}$) can be very large, these vectors and matrices can be divided into components associated with grouping factors $\boldsymbol{f}_i, i = 1, \ldots, k$ in the data. Each grouping factor is of length n, the same as the length of $\boldsymbol{y}$. The number of distinct values in $\boldsymbol{f}_i$, also called the number of *levels* of $\boldsymbol{f}_i$, is $m_i, i = 1, \ldots, k$. In the general form of the model, a model matrix $\boldsymbol{Z}_i$ of size $n \times q_i$ is associated with grouping factor $\boldsymbol{f}_i$, $i = 1, \ldots, k$. Typically the q_i are very small. In fact, in one common form of the model, called a *variance components* model, $q_1 = q_2 = \cdots = q_k = 1$ and each of the $\boldsymbol{Z}_i, i = 1, \ldots, k$ consist of a single column of 1's.

In the general form, the random effects vector $\boldsymbol{b}$, of length $q = \sum_{i=1}^{k} m_i q_i$, is partitioned into k "outer blocks" where the i'th outer block is of size $m_i q_i, i = 1, \ldots, k$. The columns of $\boldsymbol{Z}$ and the rows and columns of $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}$ and $\boldsymbol{\Omega}$ are similarly partitioned. Each of the outer blocks is further subdivided into m_i inner blocks of size q_i. The grouping factors determine the outer blocks and the levels of each grouping factor determine the inner blocks.

In the models that we will consider, the random effects associated with different grouping factors are independent. That is, $\mathbf{\Omega}$ is block-diagonal in k blocks of sizes $m_i q_i \times m_i q_i, i = 1, \ldots, k$. Furthermore, the random effects associated with the levels of a given blocking factor are independent and identically distributed. Thus the i'th diagonal block in $\mathbf{\Omega}$ is itself block diagonal and these diagonal blocks are m_i repetitions of a $q_i \times q_i$ matrix $\mathbf{\Omega}_i$, $i = 1, \ldots, k$, providing

$$|\mathbf{\Omega}| = \sum_{i=1}^{k} m_i |\mathbf{\Omega}_i| \tag{2}$$

For a variance components model the matrices $\mathbf{\Omega}_i, i = 1, \ldots, k$ are 1×1 positive definite matrices which we can consider to be positive scalars $\omega_i, i = 1, \ldots, k$. The matrix $\mathbf{\Omega}$ is block-diagonal of size $\sum_{i=1}^{k} m_i$ and the diagonal blocks are $\omega_i \boldsymbol{I}_{m_i}$ where $\boldsymbol{I}_{m_i}$ is the $m_i \times m_i$ identity matrix. Thus $|\mathbf{\Omega}| = \sum_{i=1}^{k} m_i \omega_i$. The k-dimensional vector $\boldsymbol{\theta}$ where $\theta_i = \log \omega_i, i = 1, \ldots, k$ can be used as the unconstrained parameter vector.

The columns of the matrix $\boldsymbol{Z}$ are similarly divided into blocks. For the variance components model the ith block is the set of indicator columns for the m_i levels of $\boldsymbol{f}_i, i = 1, \ldots, k$. Because each block is a set of indicators, the diagonal blocks of $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}$ are themselves diagonal. However, unlike the corresponding blocks in $\mathbf{\Omega}$, these blocks are not necessarily a multiple of the identity. The diagonal elements of the ith diagonal block are the m_i frequencies of occurence of each the levels of the ith grouping factor in the data. (Because all the elements of $\boldsymbol{Z}$ are zero or one, the diagonals of $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}$ are simply the counts of the number of ones in the corresponding column of $\boldsymbol{Z}$.)

The off-diagonal blocks of $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}$ in a variance components model are the pairwise cross-tabulations of the corresponding grouping factors.

2.2 The Scottish secondary school example

An example may help to clarify these descriptions.

Data on achievement scores of Scottish secondary school students are described in Paterson (1991) and are analyzed in Rasbash et al. (2002, ch. 18) and other references. In the `Matrix` package for R these data are available as the data set *ScotsSec* containing the achievement scores (*attain*), some demographic data (*sex* and *social* class), a *verbal* reasoning score based on

tests taken at entry to secondary school, and the *primary* and secondary (*second*) schools attended by 3435 students.

The grouping factors for the random effects are *primary* (148 distinct schools) and *second* (19 distinct schools). the locations of the non-zeros in the 167×167 matrix $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}$ are shown in Figure 1. for a variance components model with these grouping factors. In the figure darker greys indicate larger magnitudes.

2.3 General structure of the sparse matrix

For the variance components model $bZ^\mathsf{T}\boldsymbol{Z}$ is based on the pairwise cross-tabulation of the grouping factors. In the more general model where some of the $\boldsymbol{Z}_i$ can have multiple columns, the structure of $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}$ can be derived from the structure of the pairwise cross-tabulation matrix. Both $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}$ and the pairwise cross-tabulation can be divided into a $k \times k$ grid of blocks. The pattern of non-zeros in the (i, j) block of $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}$ is obtained by replacing each non-zero in the (i, j) block of the cross-tabulation by a $q_i \times q_j$ matrix. Notice that we can determine the patterns of non-zeros in $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}$ knowing only the $q_i, i = 1, \ldots, k$ and the cross-tabulation of the grouping factors.

2.4 Crossed and nested grouping factors

In the Scottish secondary school example if all the students from a given primary school attended the same secondary school we would say that *primary* is *nested within second*. That is not the case. We can see in Figure 1 that there is a moderate amount of *crossing* of these two grouping factors. If there was at least one student in the study from each combination of primary school and secondary school we would describe the grouping factors *primary* and *second* as being *fully crossed*. Again, that is not the case for the Scottish secondary data. Grouping factors like these, which are neither nested nor fully crossed, are said to be *partially crossed*.

2.5 Estimation criteria and related quantities

For ease of reference we restate some of the results from Bates and DebRoy (2004) in the form in which they will be calculated.

Given the observed responses $\boldsymbol{y}$ and the model matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$, we wish to determine either the maximum likelihood (ML) or the restricted maximum

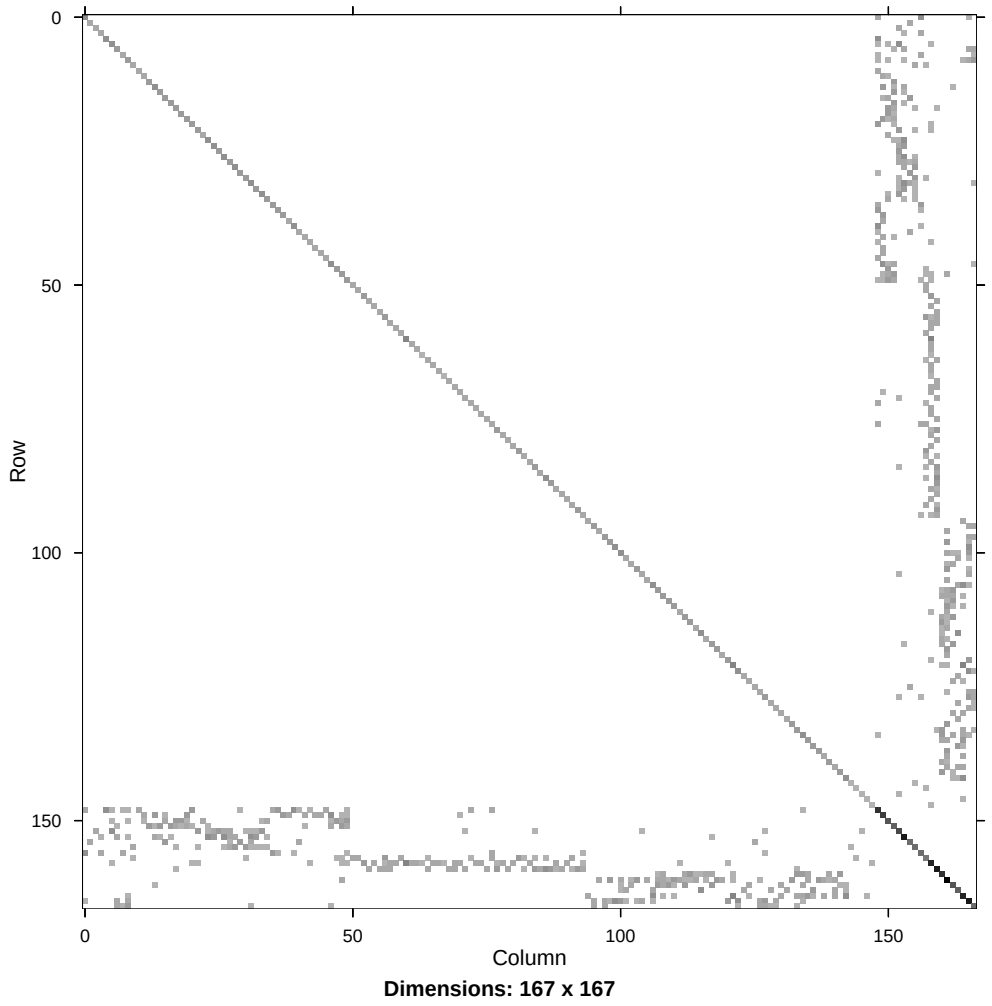

Figure 1: Location of non-zero elements in $\boldsymbol{Z}^\top \boldsymbol{Z}$ for model *fm1* fit to the *ScotsSec* data. Darker squares indicate larger magnitudes. Rows and columns are numbered from zero.

likelihood (REML) estimates of the parameters $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, and σ^2. Because the conditional estimates of $\boldsymbol{\beta}$ and σ^2, given a value of $\boldsymbol{\theta}$, for either criterion can be determined from the solution to penalized least squares problem, we can reduce the optimization problem to one involving $\boldsymbol{\theta}$ only. This reduction of the dimension of the optimization problem is called *profiling* the objective.

The conditional, penalized least squares problem can be solved using the Cholesky decomposition

$$\begin{bmatrix} \boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\boldsymbol{\Omega} & \boldsymbol{Z}^\mathsf{T}\boldsymbol{X} & \boldsymbol{Z}^\mathsf{T}\boldsymbol{y} \\ \boldsymbol{X}^\mathsf{T}\boldsymbol{Z} & \boldsymbol{X}^\mathsf{T}\boldsymbol{X} & \boldsymbol{X}^\mathsf{T}\boldsymbol{y} \\ \boldsymbol{y}^\mathsf{T}\boldsymbol{Z} & \boldsymbol{y}^\mathsf{T}\boldsymbol{X} & \boldsymbol{y}^\mathsf{T}\boldsymbol{y} \end{bmatrix} = \boldsymbol{R}^\mathsf{T}\boldsymbol{R} \quad \text{where} \quad \boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}_{ZZ} & \boldsymbol{R}_{ZX} & \boldsymbol{r}_{Zy} \\ \boldsymbol{0} & \boldsymbol{R}_{XX} & \boldsymbol{r}_{Xy} \\ \boldsymbol{0} & \boldsymbol{0} & r_{yy} \end{bmatrix}. \tag{3}$$

where the matrices $\boldsymbol{R}_{ZZ}$ and $\boldsymbol{R}_{XX}$ are upper triangular of dimension $q\times q$ and $p\times p$ respectively. The corresponding vectors, $\boldsymbol{r}_{Zy}$ and $\boldsymbol{r}_{Xy}$, are of dimension q and p, and r_{yy} is a scalar. The conditions that $\boldsymbol{\Omega}$ be positive definite and $\boldsymbol{X}$ have full column rank ensure that $\boldsymbol{R}_{ZZ}$ and $\boldsymbol{R}_{XX}$ are nonsingular.

In our implementation we do not form the upper triangular Cholesky factor $\boldsymbol{R}_{ZZ}$. Instead we use Tim Davis's LDL package (Davis, 2004) to factor

$$\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\boldsymbol{\Omega} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^\mathsf{T} \tag{4}$$

where $\boldsymbol{L}$ is a sparse, unit, lower triangular matrix and $\boldsymbol{D}$ is diagonal with positive diagonal elements. Because the diagonal elements of the unit triangular matrix $\boldsymbol{L}$ are, by definition, unity, they are not explicitly stored.

In general the matrices $\boldsymbol{Z}^\mathsf{T}\boldsymbol{X}$ and $\boldsymbol{X}^\mathsf{T}\boldsymbol{X}$ are dense. We use functions from the LDL package to solve for $\boldsymbol{R}_{ZX}$ in

$$\boldsymbol{D}^{1/2}\boldsymbol{L}^\mathsf{T}\boldsymbol{R}_{ZX} = \boldsymbol{Z}^\mathsf{T}\boldsymbol{X} \tag{5}$$

Having solved for $\boldsymbol{R}_{ZX}$ we can downdate $\boldsymbol{X}^\mathsf{T}\boldsymbol{X}$ and determine the dense Cholesky factor $\boldsymbol{R}_{XX}$ in

$$\boldsymbol{X}^\mathsf{T}\boldsymbol{X} - \boldsymbol{R}_{ZX}^\mathsf{T}\boldsymbol{R}_{ZX} = \boldsymbol{R}_{XX}^\mathsf{T}\boldsymbol{R}_{XX} \tag{6}$$

Similar relationships are used to determine $\boldsymbol{r}_{Zy}$, $\boldsymbol{r}_{Xy}$, and r_{yy}. In fact, in our implementation we append $\boldsymbol{y}$ to $\boldsymbol{X}$ when forming $\boldsymbol{Z}^\mathsf{T}\boldsymbol{X}$ and $\boldsymbol{X}^\mathsf{T}\boldsymbol{X}$ so that (**??**) provides both $\boldsymbol{R}_{ZX}$ and $\boldsymbol{r}_{Zy}$ and (6) provides $\boldsymbol{R}_{XX}$, $\boldsymbol{r}_{Xy}$, and r_{yy}.

The conditional estimates of $\boldsymbol{\beta}$ satisfy

$$\boldsymbol{R}_{XX}\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \boldsymbol{r}_{Xy} \tag{7}$$

and the conditional modes of the random effects satisfy

$$\boldsymbol{D}^{1/2}\boldsymbol{L}^{\mathsf{T}}\widehat{\boldsymbol{b}}(\boldsymbol{\theta}) = \boldsymbol{r}_{Zy} - \boldsymbol{R}_{ZX}\widehat{\boldsymbol{\beta}}. \tag{8}$$

The conditional ML estimate of σ^2 is $\widehat{\sigma^2}(\boldsymbol{\theta}) = r_{yy}^2/n$ and the conditional REML estimate is $\widehat{\sigma^2}_R(\boldsymbol{\theta}) = r_{yy}^2/(n-p)$.

The profiled optimization problem, expressed in terms of the deviance, is

$$\begin{aligned}\widehat{\boldsymbol{\theta}} &= \arg\min_{\boldsymbol{\theta}} -2\tilde{\ell}(\boldsymbol{\theta}) \\ &= \arg\min_{\boldsymbol{\theta}} \left\{ \log\left(\frac{|\boldsymbol{D}|}{|\boldsymbol{\Omega}|}\right) + n\left[1 + \log\left(\frac{2\pi r_{yy}^2}{n}\right)\right]\right\}\end{aligned} \tag{9}$$

$$\begin{aligned}\widehat{\boldsymbol{\theta}_R} &= \arg\min_{\boldsymbol{\theta}} -2\tilde{\ell}_R(\boldsymbol{\theta}) \\ &= \arg\min_{\boldsymbol{\theta}} \left\{ \log\left(\frac{|\boldsymbol{D}|\,|\boldsymbol{R}_{XX}|^2}{|\boldsymbol{\Omega}|}\right) + (n-p)\left[1 + \log\left(\frac{2\pi r_{yy}^2}{n-p}\right)\right]\right\}\end{aligned} \tag{10}$$

for ML and REML estimation, respectively. The gradients of these criteria are

$$\nabla(-2\tilde{\ell}) = \operatorname{tr}\left[\mathsf{D}\,\boldsymbol{\Omega}\left((\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}+\boldsymbol{\Omega})^{-1} - \boldsymbol{\Omega}^{-1} + \frac{\widehat{\boldsymbol{b}}}{\widehat{\sigma}}\frac{\widehat{\boldsymbol{b}}^{\mathsf{T}}}{\widehat{\sigma}}\right)\right] \tag{11}$$

$$\nabla(-2\tilde{\ell}_R) = \operatorname{tr}\left[\mathsf{D}\,\boldsymbol{\Omega}\left(\boldsymbol{V}_{\boldsymbol{b}} - \boldsymbol{\Omega}^{-1} + \frac{\widehat{\boldsymbol{b}}}{\widehat{\sigma}_R}\frac{\widehat{\boldsymbol{b}}}{\widehat{\sigma}_R}^{\mathsf{T}}\right)\right] \tag{12}$$

where

$$\boldsymbol{V}_{\boldsymbol{b}} = \boldsymbol{L}^{-\mathsf{T}}\boldsymbol{D}^{-1/2}\left(\boldsymbol{I} + \boldsymbol{R}_{ZX}\boldsymbol{R}_{XX}^{-1}\boldsymbol{R}_{XX}^{-\mathsf{T}}\boldsymbol{R}_{ZX}^{\mathsf{T}}\right)\boldsymbol{D}^{-1/2}\boldsymbol{L}^{-1} \tag{13}$$

and D denotes the Frechet derivative.

If good starting estimates of $\boldsymbol{\theta}$ are not available, the initial Newton iterations for (9) or (10) can be unstable. We can refine our initial estimates with a moderate number of ECME steps for which $\boldsymbol{\theta}_{i+1}$ satisfies

$$\operatorname{tr}\left[\mathsf{D}\,\boldsymbol{\Omega}\left(\frac{\widehat{\boldsymbol{b}}(\boldsymbol{\theta}_i)}{\widehat{\sigma}(\boldsymbol{\theta}_i)}\frac{\widehat{\boldsymbol{b}}(\boldsymbol{\theta}_i)^{\mathsf{T}}}{\widehat{\sigma}(\boldsymbol{\theta}_i)} + \left(\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z} + \boldsymbol{\Omega}(\boldsymbol{\theta}_i)\right)^{-1} - \boldsymbol{\Omega}(\boldsymbol{\theta}_{i+1})^{-1}\right)\right] = \boldsymbol{0} \tag{14}$$

for ML estimates or

$$\operatorname{tr}\left[\mathsf{D}\,\boldsymbol{\Omega}\left(\frac{\widehat{\boldsymbol{b}}(\boldsymbol{\theta}_i)}{\widehat{\sigma}_R(\boldsymbol{\theta}_i)}\frac{\widehat{\boldsymbol{b}}(\boldsymbol{\theta}_i)^{\mathsf{T}}}{\widehat{\sigma}_R(\boldsymbol{\theta}_i)} + \boldsymbol{V}_{\boldsymbol{b}}(\boldsymbol{\theta}_i) - \boldsymbol{\Omega}(\boldsymbol{\theta}_{i+1})^{-1}\right)\right] = \boldsymbol{0} \tag{15}$$

for REML.

At this point it is easy to formulate a general method of obtaining ML or REML estimates for a linear mixed model:

1. Given the data $\boldsymbol{y}$ and the model matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$, formulate initial estimates $\boldsymbol{\theta}_0$. Some heuristics for doing so are given in Pinheiro and Bates (2000, ch. 3).

2. Use a moderate number of ECME steps, (14) or (15), to refine these starting estimates. Each ECME step requires evaluating $\boldsymbol{\Omega}(\boldsymbol{\theta})$ followed by the decomposition and solutions (3), (5), (**??**), (7) and (8).

3. Use a Newton method to optimize the criterion (9) or (10) with gradient (11) or (12). Each evaluation of the criterion requires evaluating $\boldsymbol{\Omega}(\boldsymbol{\theta})$ followed by the decomposition and solutions (3), (5), and (6). Gradient evaluations require the solutions to (7) and (8).

In Bates and DebRoy (2004) we show that similar calculations can be used to evaluate the Hessian of the profiled criteria and that the deviance forms of the criteria are bounded below throughout the parameter space. Reasonable starting values determined by the ECME iterations and analytic expressions for the gradients and Hessians help to make (9) and (10) very well controlled optimization problems. The most difficult computational step in the ECME or Newton iterations is the sparse Cholesky decomposition (3).

3 Symbolic analysis

Although the decomposition (3) will be performed many times for different trial values of $\boldsymbol{\theta}$, the structure of $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\boldsymbol{\Omega}$ – in particular, the number and positions of the non-zeros in $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\boldsymbol{\Omega}$ and in $\boldsymbol{L}$ – will be the same for each evaluation. The LDL package provides one C function that performs the symbolic analysis to determine the number and position of the non-zeros in $\boldsymbol{L}$ and another C function to determine the numerical values in the decomposition.

The number and positions of the nonzeros in $\boldsymbol{L}$ depends on the positions of the nonzeros in $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}$. There will potentially be a nonzero in $\boldsymbol{L}$ anywhere there is a nonzero in the lower triangle of $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}$ but other nonzeros in $\boldsymbol{L}$ can be induced during the course of the decomposition. This is called "fill-in".

The extent of the fill-in can be changed by reordering the components of $\boldsymbol{b}$ and, correspondingly, the columns of $\boldsymbol{Z}$.

Although there are general approaches, such as approximate minimal degree (Davis, 1996) or approaches based on partitions of the adjacency graph (Karapis, 2003), to determining a fill-minimizing permutation, it is more effective for us to exploit the special structure of $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}$ in searching for a permutation.

As mentioned above, when considering the structure of $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\boldsymbol{\Omega}$ we only need to consider the structure for the variance components model because the structure for the general model is obtained from the structure for the variance components model by replacing each nonzero in the (i, j) block of the variance components model by a $q_i \times q_j$ nonzero matrix. Similarly we can derive the structure of the $\boldsymbol{L}$ matrix for the general model from that of the variance components model we restrict our attention to permutations that do not mix levels from different grouping factors. That is, we consider only those fill-reducing permutations that consists of a permutation of the grouping factors and permutations of the levels within each grouping factor.

In what follows we restrict our attention to the variance components model.

Fill-in is determined by the elimination tree (Liu, 1990) for the symmetric matrix. We can determine the Cholesky decomposition, and hence the elimination tree and the extent of the fill-in, column-wise starting with the first column. We know that there will be "original" nonzeros in $\boldsymbol{L}$ wherever there are nonzeros in the lower triangle of $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\boldsymbol{\Omega}$ and, possibly, some additional, "induced" nonzeros. At column j if there are nonzeros, either original or induced, below the diagonal in rows i and k then a nonzero is induced in the (i, k) position of $\boldsymbol{L}$. Consider again the division of $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\boldsymbol{\Omega}$ and $\boldsymbol{L}$ into a $k \times k$ array of blocks determined by the grouping factors. For a variance components model, the diagonal blocks are themselves diagonal. Because the $(1, 1)$ block is diagonal the row numbers of any nonzeros below the diagonal must be greater than m_1. That is, there will not be any induced nonzeros in the first m_1 columns and we choose the first grouping factor to make m_1 as large as possible.

In most examples there are just one or two grouping factors. For the rare examples with three or more grouping factors we order the factors so that $m_1 \geq m_2 \geq \cdots \geq m_k$.

Because the columns corresponding to the first grouping factor do not experience any fill-in, there is no need to permute the levels of the first

grouping factor. This can be a considerable savings in the effort required to determine a fill-reducing permutation. For the Scottish secondary school example we can leave the first 148 columns in their original order and consider only permutations of the last 19 columns.

For the purposes of determining the induced nonzeros we can "project" the first m_1 columns onto the last $q - m_1$ columns as described above.

If the grouping factors are a nested sequence of factors there will be no fill-in. In fact, both $\boldsymbol{L}$ and its inverse will have exactly the same pattern of nonzeros as does the lower triangle of $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}$. We do not seek a fill-reducing permutation if the grouping factors form a nested sequence. The case $k = 1$ (the random effects are determined by a single grouping factor) is, trivially, a nested sequence.

4 Generalizations of linear mixed models

5 Further enhancements

6 Acknowledgements

This work was supported by U.S. Army Medical Research and Materiel Command under Contract No. DAMD17-02-C-0119. The views, opinions and/or findings contained in this report are those of the authors and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

I thank Deepayan Sarkar and Tim Davis for helpful discussions and suggestions and Harold Doran for his suggestion of using sparse matrix techniques for linear mixed-effects models.

The `Matrix` package for R is based on code from the LDL (Davis, 2004), TAUCS (Toledo, 2003), and Metis (Karapis, 2003) packages.

References

Douglas M. Bates and Saikat DebRoy. Linear mixed models and penalized least squares. *J. of Multivariate Analysis*, 2004. to appear. 3, 5, 9

Tim Davis. An approximate minimal degree ordering algorithm. *SIAM J. Matrix Analysis and Applications*, 17(4):886–905, 1996. 10

Timothy A. Davis. Algorithm 8xx: A concise sparse Cholesky factorization package. Technical report, Department of Computer and Information Science and Engineering, University of Florida, 2004. 7, 11

George Karapis. Metis: Family of multilevel partioning algorithms. http://www-users.cs.umn.edu/ karypis/metis/, 2003. 10, 11

J. W. H. Liu. The role of elimination trees in sparse factorization. *SIAM J. Matrix Analysis and Applications*, 11(1):134–172, 1990. 10

L. Paterson. Socio economic status and educational attainment: a multidimensional and multilevel study. *Evaluation and Research in Education*, 5: 97–121, 1991. 4

José C. Pinheiro and Douglas M. Bates. *Mixed-Effects Models in S and S-PLUS*. Springer, 2000. ISBN 0-387-98957-9. 2, 9

Jon Rasbash, William Browne, Harvey Goldstein, Min Yand, Ian Plewis, Michael Healy, Geoff Woodhouse, David Draper, Ian Longford, and Toby Lewis. *A User's Guide to MLwiN*. Institute of Education, University of London, version 2.1d edition, 2002. ISBN 0-85473-612-3. 4

Sivan Toledo. Taucs: A library of sparse linear solvers. http://www.tau.ac.il/ stoledo/taucs/, 2003. 11