Sparse Matrix Representations of Linear Mixed Models

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

June 18, 2004

Abstract

We describe a representation of linear mixed-effects models using a sparse semidefinite matrix. This representation provides for efficient evaluation of the profiled log-likelihood or profiled restricted log-likelihood of the model, given the relative precision parameters for 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 the positions of nonzero elements in the result 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 the sparse semidefinite matrix. This matrix has 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 associated with these 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 the solution to 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 the positions of nonzero elements in the result 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 §4 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 will consider and restate some of the formulas from Bates and DebRoy (2004) using the

LDL form of the Cholesky decomposition of a sparse, semidefinite 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 are highly structured. They are divided into components associated with grouping factors $\boldsymbol{f}_i, i = 1, \ldots, k$ (each of length n, the same as the length of $\boldsymbol{y}$) that are part of the data. 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 $\boldsymbol{Z}_i, i = 1, \ldots, k$ is 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. The ith outer block is further subdivided into m_i inner blocks of size q_i. Note that the grouping factors determine the outer blocks and the levels of the grouping factors determine the inner blocks.

In the models that we will consider, the random effects associated with different grouping factors are independent. That is, $\boldsymbol{\Omega}$ 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 $\boldsymbol{\Omega}$ is itself block diagonal and these diagonal blocks are m_i repetitions of a $q_i \times q_i$ matrix $\boldsymbol{\Omega}_i$, $i = 1, \ldots, k$, providing

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

For a variance components model the matrices $\boldsymbol{\Omega}_i, i = 1, \ldots, k$ are 1×1 positive definite matrices which we can consider to be positive scalars $\omega_i, i = 1, \ldots, k$. The matrix $\boldsymbol{\Omega}$ 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 $\log |\boldsymbol{\Omega}| = \sum_{i=1}^{k} m_i \log \omega_i$. 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 $\boldsymbol{\Omega}$, 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 crosstabulations of the corresponding grouping factors.

2.1.2 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 nonzeros in

the 167×167 matrix $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}$ are shown in Figure 1 for a variance components model using these grouping factors.

2.1.3 General structure of the sparse matrix

For the variance components model $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}$ is based on the pairwise crosstabulation 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 crosstabulation matrix. Both $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}$ and the pairwise crosstabulation can be divided into a $k \times k$ grid of blocks. The pattern of nonzeros in the (i, j) block of $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}$ is obtained by replacing each nonzero in the (i, j) block of the crosstabulation by a $q_i \times q_j$ matrix. Notice that we can determine the patterns of nonzeros in $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}$ knowing only the $q_i, i = 1, \ldots, k$ and the pairwise crosstabulation of the grouping factors.

2.1.4 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*.

2.2 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 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 a 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.

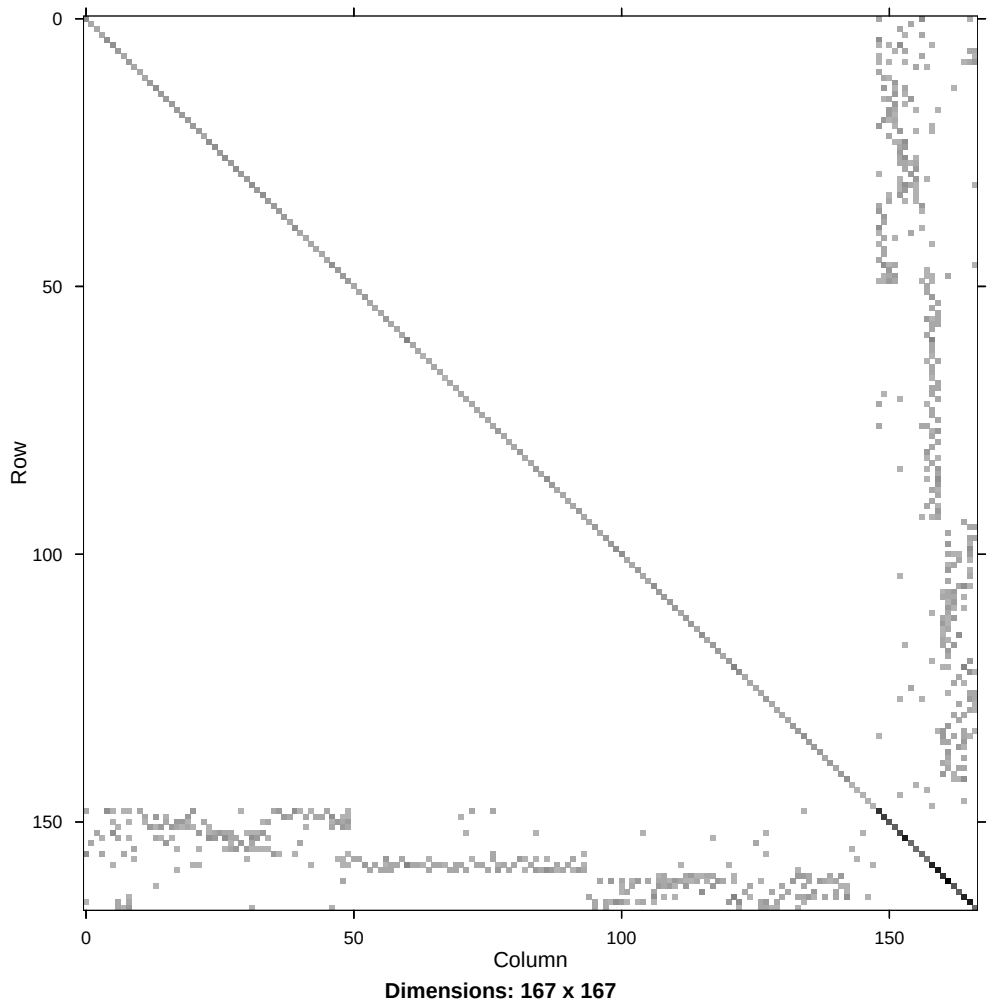

Figure 1: Location of nonzero elements in $\boldsymbol{Z}^\top\boldsymbol{Z}$ for a variance components model for the *ScotsSec* data. Darker squares indicate larger magnitudes. Rows and columns are numbered from zero. The first 148 rows and columns correspond to the levels of the *primary* grouping factor and the last 19 rows and columns correspond to levels of the *second* grouping factor.

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

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

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 upper triangular Cholesky factor $\boldsymbol{R}_{XX}$ satisfying

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

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 (5) 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 (3) and the solutions to (5), (6), (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 (3) and the solutions to (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 the positions of nonzeros in $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\boldsymbol{\Omega}$ and in $\boldsymbol{L}$ – will be the same for each evaluation. Because the LDL package provides one function to perform the symbolic analysis and another function to determine the numerical values in the decomposition, we can do the symbolic analysis separately.

The number and the positions of the nonzeros in $\boldsymbol{L}$ depend on the positions of the nonzeros in $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}$. Any nonzero position in the lower triangle of $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}$ can be nonzero in $\boldsymbol{L}$. However other positions in $\boldsymbol{L}$ can become nonzero during the course of the decomposition. This is called "fill-in". The extent of the fill-in can be altered 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 graph-partitioning algorithms (Karapis, 2003), for 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 such a permutation.

As mentioned above, when considering the structure of $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\boldsymbol{\Omega}$ we need only 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 provided 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 consist of, at most, a permutation of the grouping factors and permutations of the levels within each grouping factor. Note that we can determine such a permutation based only on the pairwise crosstabulation of the grouping factors. That is, the fill-reducing permutation for the variance components model provides the fill-reducing permutation for the general model. Hence, in what follows, we consider only 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. If there are nonzeros, either original or induced, below the diagonal in column j, say 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. Because this first block of m_1 rows and columns will not experience any fill-in, we choose the first grouping factor to have the greatest number of levels. In general we order the grouping factors so that $m_1 \geq m_2 \geq \cdots \geq m_k$.

There is no need to permute the levels of the first grouping factor, which can result in 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.

We obtain the matrix to determine the permutation of the second and subsequent groups by "projecting" the first m_1 columns onto the last $q - m_1$

columns. If there are only two grouping factors we record as potentially nonzero the positions (i, k) with $i, k > m_1$ and both (i, j) and (k, j) nonzero for some $j \leq m_1$. If there are more than two grouping factors we record all these positions plus any of the original nonzeros below the m_1st row and to the right of the m_1st column.

In the case of two nested grouping factors there will only be one nonzero element below the diagonal in each of the first m_1 columns, hence there is no fill-in. For more than two nested grouping factors, any pair of nonzeros occuring in the first m_1 columns must be in rows associated with different grouping factors and, furthermore, the nonzero off-diagonal that would be generated in the projected matrix for that combination must already be nonzero. That is, nested grouping factors do not generate any fill-in. Not only can the matrix $\boldsymbol{L}$ be created "in place" (that is, with exactly the same positions of nonzeros as in the lower triangle of $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}$) but also $\boldsymbol{L}^{-1}$ has the same pattern of nonzeros. This is unusual. In most cases $\boldsymbol{L}^{-1}$ has many more nonzeros than does $\boldsymbol{L}$.

Notice that a single grouping factor is, trivially, a nested sequence.

3.1 Examples

In Figure 2 we show the projection of the symmetric matrix in Figure 1 onto the 19×19 block for the *second* factor. We also show the patterns of nonzeros in $\boldsymbol{L}$ from the LDL decomposition of this block (using the original ordering of the rows and columns); the symmetric block with its rows and columns permuted according to a fill-reducing permutation determined by Metis (Karapis, 2003); and the nonzeros in the $\boldsymbol{L}$ matrix from the decomposition of the permuted symmetric matrix.

In this case the fill-minimizing permutation does not produce a great savings in the amount of storage, even relative to dense storage of the matrix which would have 171 elements below the diagonal. Without the fill-minimizing permutation there are 154 off-diagonal nonzeros in $\boldsymbol{L}$. With the permutation there are 131.

Even so, only 470 floating point values are required to store $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}$ and a total of 601 for the decomposition (167 for $\boldsymbol{D}$ and 434 for $\boldsymbol{L}$). Dense storage of the $(2, 1)$ block of $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}$, as suggested in Rasbash et al. (2002), would require 2812 floating point locations for that block alone, plus a similar number for the decomposition. In this example it is the sparse storage, more than the fill-reducing permutation, that reduces both the space and the computational

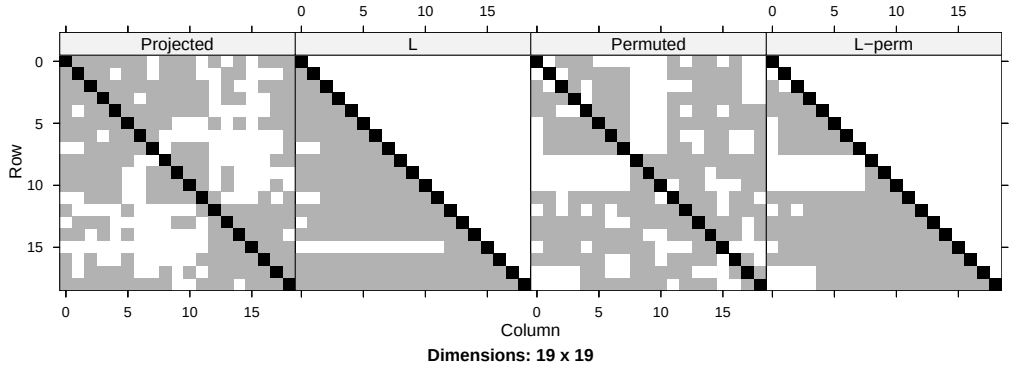

Figure 2: Projection of the pairwise crosstabulation for the Scottish secondary school data onto the 19×19 lower right block for the *second* factor. The "Projected" panel shows the original (black squares) and projected (gray squares) nonzero positions. The "L" panel shows the implicit diagonal (black squares) and the nonzero off-diagonal (gray squares) for this ordering of the levels of the factor. The "Permuted" panel shows the symmetric matrix after permuting the rows and columns according to the fill-reducing permutation determined by the Metis package and the "L-perm" panel shows the $\boldsymbol{L}$ matrix from this ordering.

time required, relative to previous methods.

In other cases, however, the storage savings from the fill-reducing permutation can be substantial. We have used these methods to analyze 378,047 test scores of 134,713 students in 3722 schools. There are 377,111 nonredundant nonzeros in the pairwise crosstabulation (138,435 diagonals and 238,676 off-diagonals). The projection of the $(2,1)$ block into the $(2,2)$ block produces a symmetric 3722×3722 matrix with 49,305 non-redundant nonzeros (3722 diagonals and 45,583 off-diagonals). Without the fill-reducing permutation the $\boldsymbol{L}$ matrix has 4,469,124 nonzero off-diagonals or 64.5% of the maximum possible number (6,924,781). With the permutation the number of nonzero off-diagonals is reduced to 193,562 or 2.8% of the maximum.

Fitting linear mixed models to data, such as these, that have a large number of levels in one or more of a set of partially crossed grouping factors has been very difficult, if not impossible, with existing multilevel modeling software. Using a dense representation of the $(2,1)$ block (over 5×10^8 floating point values or 4 GB per copy, in double precision) is impractical on most current computers. A sparse matrix representation combined with a fill-reducing permutation makes it practical to perform multilevel analysis of such data, which is becoming common (Lockwood et al., 2003).

4 Generalizations of linear mixed models

A generalized linear mixed model is similar to a linear mixed model except that the linear predictor

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} = g(\boldsymbol{\mu}) \tag{16}$$

is a function g, called the *link* function, of the mean vector, $\boldsymbol{\mu}$. Furthermore, the conditional density of the responses, $p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\theta})$, may be other than Gaussian.

Usually the conditional density is a member of the exponential family, such as the Bernoulli distribution for binary responses or the Poisson distribution for count data, in which case there is a canonical link function (the logit link for the Bernoulli and the log link for the Poisson).

In many cases, such as the Bernoulli and the Poisson, the conditional distribution of the response depends only on the mean. In other cases, such as the Gaussian and gamma distributions, there is a (common) scale parameter

in the distribution. If there is a scale parameter we incorporate it in the general dispersion parameter $\boldsymbol{\theta}$.

McCullagh and Nelder (1989) describe an iteratively reweighted least squares (IRLS) algorithm for determining the MLEs of the coefficients $\boldsymbol{\beta}$ and, if used, the scale parameter, in a generalized linear model without random effects. A penalized quasi-likelihood (PQL) algorithm for estimation of the parameters in a GLMM can be implemented by replacing the least squares problem in IRLS by the penalized least squares (PLS) problem represented by (3). Within each of the IRLS iterations, we perform some ECME and/or Newton iterations to optimize the log-likelihood (as a function of $\boldsymbol{\theta}$) represented by the PLS problem. Convergence of this PQL algorithm is indicated by its reaching a stationary point.

Thus the techniques described in §2 and §3 provide the inner optimization for the PQL algorithm. Although PQL is not guaranteed to produce the maximum likelihood estimators of the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, it will generally get close to the MLEs. After PQL has converged we switch to direct optimization of the marginal likelihood

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}) = \int_{\boldsymbol{b}} p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\theta}) p(\boldsymbol{b}|\boldsymbol{\theta})\, d\boldsymbol{b}. \tag{17}$$

The integral (17) does not, in general, have a closed form. Two effective ways of approximating this integral are Laplacian integration (Tierney and Kadane, 1986) and adaptive Gauss-Hermite integration (Pinheiro and Bates, 2000) for which the conditional modes of the random effects

$$\widehat{\boldsymbol{b}}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \arg\max_{\boldsymbol{b}} p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\theta}) p(\boldsymbol{b}, \boldsymbol{\theta}) \tag{18}$$

are required.

The conditional modes can be determined by a penalized IRLS algorithm implemented using the representations and decompositions described in §2 and §3.

4.1 Nonlinear mixed models

In a nonlinear mixed model element i of the expected response $\mathrm{E}[\boldsymbol{Y}]$ is a function of covariates $\boldsymbol{x}_i$ associated with that observed response and a r-dimensional parameter $\boldsymbol{\phi}_i$.

$$\{\mathrm{E}[\boldsymbol{Y}]\}_i = f(\boldsymbol{\phi}_i, \boldsymbol{x}_i) \quad i = 1, \ldots, n. \tag{19}$$

The values of $\boldsymbol{\phi}$ can be represented as an $n \times r$ matrix $\boldsymbol{\Phi}$. Each column of this matrix is a linear function of $\boldsymbol{\beta}$ and $\boldsymbol{b}$

$$\boldsymbol{\Phi}_j = \boldsymbol{X}_j\boldsymbol{\beta} + \boldsymbol{Z}_j\boldsymbol{b} \quad j = 1, \ldots, r \tag{20}$$

If the function f is linear in all the elements of $\boldsymbol{\phi}_i$ the NMM can be reexpressed as a LMM, hence we will assume that f is nonlinear in at least one of these elements.

For a fixed value of $\boldsymbol{\theta}$ we can express the conditional estimates of the fixed effects $\boldsymbol{\beta}$ and the condtional modes of the random effects as the solution to a penalized nonlinear least squares (PNLS) problem. The algorithm proposed in Lindstrom and Bates (1990) is essentially the the same as the PQL algorithm for generalized linear models but with the PIRLS step replaced by PNLS. Just like PQL this algorithm will generally get close to the MLEs but does not produce the exact MLEs except in special circumstances.

Estimates obtained by maximizing the Laplacian approximation or an adaptive Gauss-Hermite approximation to the marginal likelihood should be closer to the MLEs. As in the case of the GLMM these approximations require the conditional modes of the random effects. These are the solutions to a penalized nonlinear least squares problem for which the representation and decompositions described in §2 and §3 can be used.

5 Implementation

The techniques described in §2 and §3 for linear mixed models, and the PQL and Laplace methods for generalized linear mixed models are currently implemented in the `lme4` package for R (R Development Core Team, 2004). In future versions of this package we will incorporate the adaptive Gauss-Hermite method for GLMMs and the PQL, Laplace, and adaptive Gauss-Hermite methods for NMMs.

Because both R and the `lme4` package are open source software, they provide a reference implementation of these methods against which other methods can be compared. However, the package is more than a "reference implementation"; it is carefully implemented and is suitable as a production system. As part of the package we provide a "vignette" (a code/documentation combination for literate data analysis) that fits the models from the examples in the comparative reviews of multilevel modeling software (`http://multilevel.ioe.ac.uk/softrev/`). These examples show that the `lme`

function in the `lme4` package is fast and reliable on what would currently be considered typical multilevel modeling examples. However, the sparse matrix representation and the ease of model specification allows it to go far beyond currently available software. We know of no other software that can fit models with partially crossed grouping factors to data sets with 10^5 or more observations, such as the longitudinal analysis of over 300,000 test scores described in §3.1 incorporating random effects for student and school and allowing for student migration between schools.

6 Conclusions

We have presented a specification of linear mixed models using model matrices and grouping factors, a computational representation based this specification, computational methods using the representation, and, to a lesser extent, described an implementation of the specification, representation, and computational methods. Furthermore, we show that the specification, representation, methods, and implementation can be extended to generalizations of the linear mixed model including GLMMs and NMMs.

Specification of the model is important. In many software implementations of methods for fitting linear mixed models it can be awkward to specify the structure of the random effects. We contend that by concentrating on the grouping factors in the data and on the model matrices associated with these grouping factors, the specification of linear mixed models is made much easier. Nesting or crossing of grouping factors can be determined from the data, rather than having to be specified as part of the model. Common forms of linear mixed models (including all those in the comparative review of multilevel modeling software, `http://multilevel.ioe.ac.uk/softrev/`) can be specified without using specialized forms of the precision matrices $\mathbf{\Omega}_i, i = 1, \ldots, k$. We only require that these matrices are symmetric and positive definite, which makes the implementation much simpler because we do not need to work with esoteric forms of precision matrices, or variance-covariance matrices, for the random effects.

A pairwise crosstabulation of the grouping factors, which we store as a sparse symmetric matrix, is the basis of all the symbolic analysis. Even when some of the $q_i > 1$ we can do all the symbolic analysis on the crosstabulation without needing the model matrices. It is only when the numeric representation is being formed that we need the model matrices. Because the number of

nonzero elements in the LDL decomposition is determined during the symbolic analysis we can allocate all the storage needed for later calculations immediately after this stage.

We can "project out" the effect of the first grouping factor on the reordering of the columns of the matrix $\boldsymbol{Z}$. Because it is common for the number of levels in one of the grouping factors to be comparable to the number of observations, ordering the grouping factors so that $m1 \geq m2 \geq \cdots \geq m_k$ can simplify the symbolic analysis considerably.

The representation provided by the sparse symmetric storage of $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}$ and the dense storage of $\boldsymbol{Z}^\mathsf{T}\boldsymbol{X}$ and $\boldsymbol{X}^\mathsf{T}\boldsymbol{X}$ (where $\boldsymbol{X}$ actually contains both $\boldsymbol{X}$ and $\boldsymbol{y}$), provides a remarkably efficient means of evaluating the profiled log-likelihood and restricted log-likelihood. ECME iterative steps can be efficiently implemented with this representation. After a moderate number of ECME steps it is advantageous to switch to a Newton or quasi-Newton optimization method, for which the gradient and, optionally, the Hessian of the objective can be evaluated.

7 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.

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

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.

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

Mary J. Lindstrom and Douglas M. Bates. Nonlinear mixed effects models for repeated measures data. *Biometrics*, 46:673–687, 1990.

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

J.R. Lockwood, Harold Doran, and Daniel F. McCaffrey. Using R for estimating longitudinal student achievement models. *R News*, 3(3):17–23, December 2003. URL `http://CRAN.R-project.org/doc/Rnews/`.

P. McCullagh and J.A. Nelder. *Generalized Linear Models*. Chapman & Hall, 1989.

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

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

R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2004. URL `http://www.R-project.org`. ISBN 3-900051-00-3.

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.

Luke Tierney and Joseph B. Kadane. Accurate approximations for posterior moments and densities. 81(393):82–86, 1986.

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