# Penalized least squares versus generalized least squares representations of linear mixed models

Douglas Bates
Department of Statistics
University of Wisconsin – Madison

June 23, 2008

**Abstract**

The methods in the `lme4` package for `R` for fitting linear mixed models are based on sparse matrix methods, especially the Cholesky decomposition of sparse positive-semidefinite matrices, in a penalized least squares representation of the conditional model for the response given the random effects. The representation is similar to that in Henderson's mixed-model equations. An alternative representation of the calculations is as a generalized least squares problem. We describe the two representations, show the equivalence of the two representations and explain why we feel that the penalized least squares approach is more versatile and more computationally efficient.

## 1 Definition of the model

We consider linear mixed models in which the random effects are represented by a $q$-dimensional random vector, $\boldsymbol{\mathcal{B}}$, and the response is represented by an $n$-dimensional random vector, $\boldsymbol{\mathcal{Y}}$. We observe a value, $\boldsymbol{y}$, of the response. The random effects are unobserved.

The marginal distribution of the random effects is a multivariate normal distribution with mean $\mathbf{0}$ and a variance-covariance matrix, $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, that depends on a parameter vector, $\boldsymbol{\theta}$. Typically the dimension of $\boldsymbol{\theta}$ is much, much smaller than $q$. We will defer describing a particular parameterization until

later. For the time being we simply characterize the marginal distribution of $\boldsymbol{\mathcal{B}}$ as

$$\boldsymbol{\mathcal{B}} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta})\right) \tag{1}$$

The conditional distribution, $\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{B}}$, is also multivariate normal. The conditional mean, $\mathrm{E}[\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{B}} = \boldsymbol{b}]$, is a linear function of the $p$-dimensional fixed-effects parameter, $\boldsymbol{\beta}$, and the $q$-dimensional random effects vector, $\boldsymbol{b}$, defined by the $n \times p$ and $n \times q$ model matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ as

$$\mathrm{E}[\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{B}} = \boldsymbol{b}] = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}. \tag{2}$$

The conditional variance-covariance matrix of $\boldsymbol{\mathcal{Y}}$ is simply $\sigma^2\boldsymbol{I}_n$, where $\boldsymbol{I}_n$ denotes the identity matrix of order $n$. Thus

$$\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{B}} \sim \mathcal{N}\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}, \sigma^2\boldsymbol{I}_n\right) \tag{3}$$

## 1.1 Variance-covariance of the random effects

The variance-covariance matrix, $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, of the random effects, $\boldsymbol{\mathcal{B}}$, must be symmetric and positive semidefinite (i.e. $\boldsymbol{x}'\boldsymbol{\Sigma}\boldsymbol{x} \geq 0, \forall \boldsymbol{x} \in \mathbb{R}^q$). Because the maximum likelihood estimate of a variance component can be zero, it is important to allow for a semidefinite $\boldsymbol{\Sigma}$. That is, we do not assume that $\boldsymbol{\Sigma}$ is positive definite (i.e. $\boldsymbol{x}'\boldsymbol{\Sigma}\boldsymbol{x} > 0, \forall \boldsymbol{x} \in \mathbb{R}^q$) and we do not assume that $\boldsymbol{\Sigma}^{-1}$ exists.

A positive semidefinite matrix such as $\boldsymbol{\Sigma}$ has a Cholesky decomposition of the so-called "LDL′" form. We use a slightly modified version

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \sigma^2\boldsymbol{T}(\boldsymbol{\theta})\boldsymbol{S}(\boldsymbol{\theta})\boldsymbol{S}(\boldsymbol{\theta})\boldsymbol{T}(\boldsymbol{\theta})' \tag{4}$$

where $\sigma$ is the same scale parameter that occurs in the variance-covariance of $\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{B}}$, $\boldsymbol{T}(\boldsymbol{\theta})$ is a unit lower-triangular $q \times q$ matrix and $\boldsymbol{S}(\boldsymbol{\theta})$ is a diagonal $q \times q$ matrix with nonnegative diagonal elements.

## 1.2 Orthogonal random effects

Let us define a $q$-dimensional random vector, $\boldsymbol{\mathcal{U}}$, of orthogonal random effects with a marginal distribution of

$$\boldsymbol{\mathcal{U}} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\boldsymbol{I}_q\right) \tag{5}$$

and express $\mathcal{B}$ as a linear transformation of $\mathcal{U}$,

$$\mathcal{B} = \boldsymbol{T}\boldsymbol{S}\mathcal{U}. \tag{6}$$

Note that the transformation (6) gives the desired distribution of $\mathcal{B}$ in that $\mathrm{E}[\mathcal{B}] = \boldsymbol{T}\boldsymbol{S}\mathrm{E}[\mathcal{U}] = \mathbf{0}$ and

$$\mathrm{Var}(\mathcal{B}) = \mathrm{E}[\mathcal{B}\mathcal{B}'] = \boldsymbol{T}\boldsymbol{S}\mathrm{E}[\mathcal{U}\mathcal{U}']\boldsymbol{S}\boldsymbol{T}' = \sigma^2\boldsymbol{T}\boldsymbol{S}\boldsymbol{S}\boldsymbol{T}' = \Sigma.$$

The conditional distribution, $\mathcal{Y}|\mathcal{U}$, can be derived from $\mathcal{Y}|\mathcal{B}$ as

$$\mathcal{Y}|\mathcal{U} \sim \mathcal{N}\left(\boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{T}\boldsymbol{S}\boldsymbol{u}, \sigma^2\boldsymbol{I}\right) \tag{7}$$

We will write the transpose of $\boldsymbol{Z}\boldsymbol{T}\boldsymbol{S}$ as $\boldsymbol{A}$. Because the matrices $\boldsymbol{T}$ and $\boldsymbol{S}$ depend on the parameter $\boldsymbol{\theta}$, $\boldsymbol{A}(\boldsymbol{\theta})$ is also a function of $\boldsymbol{\theta}$. That is

$$\boldsymbol{A}'(\boldsymbol{\theta}) = \boldsymbol{Z}\boldsymbol{T}(\boldsymbol{\theta})\boldsymbol{S}(\boldsymbol{\theta}). \tag{8}$$

## 1.3 Sparse matrix methods

The reason for the peculiar definition of $\boldsymbol{A}$ as the transpose of the model matrix is because $\boldsymbol{A}$ is stored and manipulated as a sparse matrix. In the compressed column-oriented storage form for sparse matrices there are advantages to storing $\boldsymbol{A}$ as a matrix of $n$ columns and $q$ rows. In particular, the CHOLMOD sparse matrix library allows us to evaluate the sparse Cholesky factor, $\boldsymbol{L}$, a lower triangular matrix that satisfies

$$\boldsymbol{L}(\boldsymbol{\theta})\boldsymbol{L}(\boldsymbol{\theta})' = \boldsymbol{P}\left(\boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{A}(\boldsymbol{\theta})' + \boldsymbol{I}_q\right)\boldsymbol{P}', \tag{9}$$

directly from $\boldsymbol{A}$.

In (9) the $q \times q$ matrix $\boldsymbol{P}$ is a "fill-reducing" permutation matrix determined from the pattern of nonzeros in the sparse model matrix $\boldsymbol{Z}$. It does not affect the statistical theory (if $\mathcal{U} \sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{I})$ then $\boldsymbol{P}'\mathcal{U}$ also has a $\mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{I})$ distribution because $\boldsymbol{P}\boldsymbol{P}' = \boldsymbol{P}'\boldsymbol{P} = \boldsymbol{I}$) but, because it affects the number of nonzeros in $\boldsymbol{L}$, it can have a tremendous impact on the amount storage required for $\boldsymbol{L}$ and the time required to evaluate it. Indeed, it is precisely because $\boldsymbol{L}(\boldsymbol{\theta})$ can be evaluated quickly, even for complex models applied the large data sets, that the `lmer` function is effective in fitting such models.

# 2 The penalized least squares approach to linear mixed models

Given a value of $\boldsymbol{\theta}$ we form $\boldsymbol{A}(\boldsymbol{\theta})$ from which we evaluate $\boldsymbol{L}(\boldsymbol{\theta})$. We can then solve for the $q \times p$ matrix, $\boldsymbol{R_{ZX}}$, in the system of equations

$$\boldsymbol{L}(\theta)\boldsymbol{R_{ZX}} = \boldsymbol{P}\boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{X} \tag{10}$$

and for the $p \times p$ upper triangular matrix, $\boldsymbol{R_X}$, satisfying

$$\boldsymbol{R}'_{\boldsymbol{X}}\boldsymbol{R_X} = \boldsymbol{X}'\boldsymbol{X} - \boldsymbol{R}'_{\boldsymbol{ZX}}\boldsymbol{R_{ZX}} \tag{11}$$

The conditional mode, $\tilde{\boldsymbol{u}}(\boldsymbol{\theta})$, of the orthogonal random effects and the conditional mle, $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$, of the fixed-effects parameters can be determined simultaneously as the solutions to a penalized least squares problem

$$\begin{bmatrix}\tilde{\boldsymbol{u}}(\boldsymbol{\theta})\\ \widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})\end{bmatrix} = \arg\min_{\boldsymbol{u},\boldsymbol{\beta}} \left\| \begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{0}\end{bmatrix} - \begin{bmatrix}\boldsymbol{A}'\boldsymbol{P}' & \boldsymbol{X}\\ \boldsymbol{I}_q & \boldsymbol{0}\end{bmatrix} \begin{bmatrix}\boldsymbol{u}\\ \boldsymbol{\beta}\end{bmatrix} \right\|^2 \tag{12}$$

for which the solution satisfies

$$\begin{bmatrix}\boldsymbol{P}\left(\boldsymbol{A}\boldsymbol{A}'+\boldsymbol{I}\right)\boldsymbol{P}' & \boldsymbol{P}\boldsymbol{A}\boldsymbol{X}\\ \boldsymbol{X}'\boldsymbol{A}'\boldsymbol{P}' & \boldsymbol{X}'\boldsymbol{X}\end{bmatrix} \begin{bmatrix}\tilde{\boldsymbol{u}}(\boldsymbol{\theta})\\ \widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})\end{bmatrix} = \begin{bmatrix}\boldsymbol{P}\boldsymbol{A}\boldsymbol{y}\\ \boldsymbol{X}'\boldsymbol{y}\end{bmatrix} \tag{13}$$

The Cholesky factor of the system matrix for the PLS problem can be expressed using $\boldsymbol{L}$, $\boldsymbol{R_{ZX}}$ and $\boldsymbol{R_X}$ because

$$\begin{bmatrix}\boldsymbol{P}\left(\boldsymbol{A}\boldsymbol{A}'+\boldsymbol{I}\right)\boldsymbol{P}' & \boldsymbol{P}\boldsymbol{A}\boldsymbol{X}\\ \boldsymbol{X}'\boldsymbol{A}'\boldsymbol{P}' & \boldsymbol{X}'\boldsymbol{X}\end{bmatrix} = \begin{bmatrix}\boldsymbol{L} & \boldsymbol{0}\\ \boldsymbol{R}'_{\boldsymbol{ZX}} & \boldsymbol{R}'_{\boldsymbol{X}}\end{bmatrix} \begin{bmatrix}\boldsymbol{L}' & \boldsymbol{R_{ZX}}\\ \boldsymbol{0} & \boldsymbol{R_X}\end{bmatrix}. \tag{14}$$

In the `lme4` package the `"mer"` class is the representation of a mixed-effects model. The `A` slot contains the sparse matrix $\boldsymbol{A}(\boldsymbol{\theta})$ and the `L` slot contains the Cholesky factor $\boldsymbol{L}(\boldsymbol{\theta})$ satisfying (9). The `RZX` and `RX` slots contain $\boldsymbol{R_{ZX}}(\boldsymbol{\theta})$ and $\boldsymbol{R_X}(\boldsymbol{\theta})$ as dense matrices.

It is not necessary to solve for $\tilde{\boldsymbol{u}}(\boldsymbol{\theta})$ and $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ to evaluate the profiled log-likelihood as a function of $\boldsymbol{\theta}$, which is the log-likelihood evaluated at $\boldsymbol{\theta}$, $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ and $\widehat{\sigma^2}(\boldsymbol{\theta})$. All that is needed for evaluation of the profiled log-likelihood is the penalized residual sum of squares, $r^2$, and the determinant, $|\boldsymbol{A}\boldsymbol{A}'+\boldsymbol{I}| = |\boldsymbol{L}|^2$.

Because $\boldsymbol{L}$ is triangular, its determinant is simply the product of its diagonal elements and, because $\boldsymbol{A}\boldsymbol{A}' + \boldsymbol{I}$ is positive definite, $|\boldsymbol{L}|^2 > 0$.

The profiled deviance (negative twice the profiled log-likelihood), as a function of $\boldsymbol{\theta}$ only ($\boldsymbol{\beta}$ and $\sigma^2$ at their conditional estimates), is

$$d(\boldsymbol{\theta}|\boldsymbol{y}) = \log(|\boldsymbol{L}|^2) + n\left(1 + \log(r^2) + \frac{2\pi}{n}\right) \tag{15}$$

The maximum likelihood estimates, $\widehat{\boldsymbol{\theta}}$, satisfy

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} d(\boldsymbol{\theta}|\boldsymbol{y}) \tag{16}$$

Once the value of $\widehat{\boldsymbol{\theta}}$ has been determined, the mle's of the other parameters are evaluated from (13) and

$$\widehat{\sigma^2}(\boldsymbol{\theta}) = \frac{r^2}{n}. \tag{17}$$

## 2.1 Comments on the sparse matrix representation

Note that nothing has been said about the form of the sparse model matrix $\boldsymbol{Z}$ other than the fact that it is sparse. The computational methods outlined above can be applied to models with multiple random effects terms in which the factors determining the random effects are nested or crossed or partially crossed.

# 3 The generalized least squares approach to linear mixed models

Another common approach to linear mixed models is to derive the marginal variance-covariance matrix of $\boldsymbol{\mathcal{Y}}$ as a function of $\boldsymbol{\theta}$ and use that to determine the conditional estimates, $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$, as the solution of a generalized least squares (GLS) problem. In the notation of §1 the marginal mean of $\boldsymbol{\mathcal{Y}}$ is $\mathrm{E}[\boldsymbol{\mathcal{Y}}] = \boldsymbol{X}\boldsymbol{\beta}$ and the marginal variance-covariance matrix is

$$\mathrm{Var}(\boldsymbol{\mathcal{Y}}) = \sigma^2\left(\boldsymbol{I}_n + \boldsymbol{Z}\boldsymbol{T}\boldsymbol{S}\boldsymbol{S}\boldsymbol{T}'\boldsymbol{Z}'\right) = \sigma^2\left(\boldsymbol{I}_n + \boldsymbol{A}'\boldsymbol{A}\right) = \sigma^2\boldsymbol{V}(\boldsymbol{\theta}), \tag{18}$$

where $\boldsymbol{V}(\boldsymbol{\theta}) = \boldsymbol{I}_n + \boldsymbol{A}'\boldsymbol{A}$.

The conditional estimates of $\boldsymbol{\beta}$ are often written as

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y} \tag{19}$$

but, of course, this formula is not suitable for computation. The matrix $\boldsymbol{V}(\boldsymbol{\theta})$ is a symmetric $n \times n$ positive definite matrix and hence has a Cholesky factor. However, this factor is $n \times n$, not $q \times q$ and $q$ is always smaller than $n$ - sometimes orders of magnitude smaller.

## 3.1 Relating the GLS approach to the Cholesky factor $L$.

We can use the fact that

$$\boldsymbol{V}^{-1}(\boldsymbol{\theta}) = (\boldsymbol{I}_n + \boldsymbol{A}'\boldsymbol{A})^{-1} = \boldsymbol{I}_n - \boldsymbol{A}'\left(\boldsymbol{I}_q + \boldsymbol{A}\boldsymbol{A}'\right)^{-1}\boldsymbol{A} \tag{20}$$

to relate the GLS problem to the PLS problem. One way to establish (20) is simply to show that the product

$$\begin{aligned}
&\left(\boldsymbol{I} + \boldsymbol{A}'\boldsymbol{A}\right)\left(\boldsymbol{I} - \boldsymbol{A}'\left(\boldsymbol{I} + \boldsymbol{A}\boldsymbol{A}'\right)^{-1}\boldsymbol{A}\right) \\
&\qquad = \boldsymbol{I} + \boldsymbol{A}'\boldsymbol{A} - \boldsymbol{A}'\left(\boldsymbol{I} + \boldsymbol{A}\boldsymbol{A}'\right)\left(\boldsymbol{I} + \boldsymbol{A}\boldsymbol{A}'\right)^{-1}\boldsymbol{A} \\
&\qquad = \boldsymbol{I} + \boldsymbol{A}'\boldsymbol{A} - \boldsymbol{A}'\boldsymbol{A} \\
&\qquad = \boldsymbol{I}.
\end{aligned}$$

Incorporating the permutation matrix $\boldsymbol{P}$ we have

$$\begin{aligned}
\boldsymbol{V}^{-1}(\boldsymbol{\theta}) &= \boldsymbol{I}_n - \boldsymbol{A}'\boldsymbol{P}'\boldsymbol{P}\left(\boldsymbol{I}_q + \boldsymbol{A}\boldsymbol{A}'\right)^{-1}\boldsymbol{P}'\boldsymbol{P}\boldsymbol{A} \\
&= \boldsymbol{I}_n - \boldsymbol{A}'\boldsymbol{P}'(\boldsymbol{L}\boldsymbol{L}')^{-1}\boldsymbol{P}\boldsymbol{A} \\
&= \boldsymbol{I}_n - \left(\boldsymbol{L}^{-1}\boldsymbol{P}\boldsymbol{A}\right)'\boldsymbol{L}^{-1}\boldsymbol{P}\boldsymbol{A}.
\end{aligned} \tag{21}$$

Even in this form we would not want to evaluate such a matrix but (21) does allow us to simplify many common expressions.

For example, the variance-covariance of the estimator $\widehat{\boldsymbol{\beta}}$, conditional on $\boldsymbol{\theta}$ and $\sigma$, can be expressed as

$$\begin{aligned}
\sigma^2\left(\boldsymbol{X}'\boldsymbol{V}^{-1}(\boldsymbol{\theta})\boldsymbol{X}\right)^{-1} &= \sigma^2\left(\boldsymbol{X}'\boldsymbol{X} - \left(\boldsymbol{L}^{-1}\boldsymbol{P}\boldsymbol{A}\boldsymbol{X}\right)'\left(\boldsymbol{L}^{-1}\boldsymbol{P}\boldsymbol{A}\boldsymbol{X}\right)\right)^{-1} \\
&= \sigma^2\left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{R}'_{\boldsymbol{Z}\boldsymbol{X}}\boldsymbol{R}_{\boldsymbol{Z}\boldsymbol{X}}\right)^{-1} \\
&= \sigma^2\left(\boldsymbol{R}'_{\boldsymbol{X}}\boldsymbol{R}_{\boldsymbol{X}}\right)^{-1}.
\end{aligned} \tag{22}$$