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

Douglas Bates
Department of Statistics
University of Wisconsin – Madison

May 6, 2025

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.

For our purposes, we will assume a "spherical" multivariate normal conditional distribution of $\boldsymbol{\mathcal{Y}}$, given $\boldsymbol{\mathcal{B}}$. That is, we assume the variance-covariance matrix of $\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{B}}$ is simply $\sigma^2\boldsymbol{I}_n$, where $\boldsymbol{I}_n$ denotes the identity matrix of order

n. (The term "spherical" refers to the fact that contours of the conditional density are concentric spheres.)

The conditional mean, $\mathrm{E}[\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{B}} = \boldsymbol{b}]$, is a linear function of $\boldsymbol{b}$ and the p-dimensional fixed-effects parameter, $\boldsymbol{\beta}$,

$$\mathrm{E}[\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{B}} = \boldsymbol{b}] = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}, \tag{1}$$

where $\boldsymbol{X}$ and $\boldsymbol{Z}$ are known model matrices of sizes $n \times p$ and $n \times q$, respectively. Thus

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

The marginal distribution of the random effects

$$\boldsymbol{\mathcal{B}} \sim \mathcal{N}\left(\boldsymbol{0}, \sigma^2\boldsymbol{\Sigma}(\boldsymbol{\theta})\right) \tag{3}$$

is also multivariate normal, with mean $\boldsymbol{0}$ and variance-covariance matrix $\sigma^2\boldsymbol{\Sigma}(\boldsymbol{\theta})$. The scalar, σ^2, in (3) is the same as the σ^2 in (2). As described in the next section, the relative variance-covariance matrix, $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, is a $q \times q$ positive semidefinite matrix depending on a parameter vector, $\boldsymbol{\theta}$. Typically the dimension of $\boldsymbol{\theta}$ is much, much smaller than q.

1.1 Variance-covariance of the random effects

The relative variance-covariance matrix, $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, 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 estimate of a variance component can be zero, it is important to allow for a semidefinite $\boldsymbol{\Sigma}$. We do not assume that $\boldsymbol{\Sigma}$ is positive definite (i.e. $\boldsymbol{x}'\boldsymbol{\Sigma}\boldsymbol{x} > 0, \forall \boldsymbol{x} \in \mathbb{R}^q, \boldsymbol{x} \neq \boldsymbol{0}$) and, hence, we cannot 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 slight modification of this form,

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{T}(\boldsymbol{\theta})\boldsymbol{S}(\boldsymbol{\theta})\boldsymbol{S}(\boldsymbol{\theta})\boldsymbol{T}(\boldsymbol{\theta})', \tag{4}$$

where $\boldsymbol{T}(\boldsymbol{\theta})$ 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 that act as scale factors. (They are the relative standard deviations of certain linear combinations of the random effects.) Thus, $\boldsymbol{T}$ is a triangular matrix and $\boldsymbol{S}$ is a scale matrix.

Both $\boldsymbol{T}$ and $\boldsymbol{S}$ are highly patterned.

1.2 Orthogonal random effects

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

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

and, for a given value of $\boldsymbol{\theta}$, express $\mathcal{B}$ as a linear transformation of $\mathcal{U}$,

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

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

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

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

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

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 $\boldsymbol{\theta}$, $\boldsymbol{A}$ is also a function of $\boldsymbol{\theta}$,

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

In applications, the matrix $\boldsymbol{Z}$ is derived from indicator columns of the levels of one or more factors in the data and is a *sparse* matrix, in the sense that most of its elements are zero. The matrix $\boldsymbol{A}$ is also sparse. In fact, the structure of $\boldsymbol{T}$ and $\boldsymbol{S}$ are such that pattern of nonzeros in $\boldsymbol{A}$ is that same as that in $\boldsymbol{Z}'$.

1.3 Sparse matrix methods

The reason for defining $\boldsymbol{A}$ as the transpose of a model matrix is because $\boldsymbol{A}$ is stored and manipulated as a sparse matrix. In the compressed column-oriented storage form that we use 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, $\boldsymbol{L}(\boldsymbol{\theta})$, a sparse 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 $\boldsymbol{A}(\boldsymbol{\theta})$.

In (9) the $q \times q$ matrix $\boldsymbol{P}$ is a "fill-reducing" permutation matrix determined from the pattern of nonzeros in $\boldsymbol{Z}$. $\boldsymbol{P}$ does not affect the statistical theory (if $\boldsymbol{\mathcal{U}} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ then $\boldsymbol{P}'\boldsymbol{\mathcal{U}}$ also has a $\mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ 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 $\boldsymbol{L}$ from $\boldsymbol{A}$. 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. Several slots in this class are matrices corresponding directly to the matrices in the preceding equations. The `A` slot contains the sparse matrix $\boldsymbol{A}(\boldsymbol{\theta})$ and the `L` slot contains the sparse Cholesky factor, $\boldsymbol{L}(\boldsymbol{\theta})$. The `RZX` and `RX` slots contain $\boldsymbol{R_{ZX}}(\boldsymbol{\theta})$ and $\boldsymbol{R_X}(\boldsymbol{\theta})$, respectively, stored 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, which is the log-likelihood evaluated $\boldsymbol{\theta}$ and the conditional estimates of the other parameters, $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ 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, from the penalized least squares problem (12) and the determinant $|\boldsymbol{AA}' + \boldsymbol{I}| = |\boldsymbol{L}|^2$. Because $\boldsymbol{L}$ is triangular, its determinant is easily evaluated as the product of its diagonal elements. Furthermore, $|\boldsymbol{L}|^2 > 0$ because it is equal to $|\boldsymbol{AA}' + \boldsymbol{I}|$, which is the determinant of a positive definite matrix. Thus $\log(|\boldsymbol{L}|^2)$ is both well-defined and easily calculated from $\boldsymbol{L}$.

The profiled deviance (negative twice the profiled log-likelihood), as a function of $\boldsymbol{\theta}$ only ($\boldsymbol{\beta}$ and σ^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 of $\boldsymbol{\beta}$ is evaluated from (13) and the mle of σ^2 as $\widehat{\sigma^2}(\boldsymbol{\theta}) = r^2/n$.

Note that nothing has been said about the form of the sparse model matrix, $\boldsymbol{Z}$, other than the fact that it is sparse. In contrast to other methods for linear mixed models, these results apply to models where $\boldsymbol{Z}$ is derived from crossed or partially crossed grouping factors, in addition to models with multiple, nested grouping factors.

The system (13) is similar to Henderson's "mixed-model equations" (reference?). One important difference between (13) and Henderson's formulation is that Henderson represented his system of equations in terms of $\boldsymbol{\Sigma}^{-1}$ and, in important practical examples, $\boldsymbol{\Sigma}^{-1}$ does not exist at the parameter estimates. Also, Henderson assumed that equations like (13) would need to be solved explicitly and, as we have seen, only the decomposition of the system matrix is needed for evaluation of the profiled log-likelihood. The same is

true of the profiled the logarithm of the REML criterion, which we define later.

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 $\mathrm{E}[\boldsymbol{\mathcal{Y}}] = \boldsymbol{X\beta}$ and the marginal variance-covariance matrix is

$$\mathrm{Var}(\boldsymbol{\mathcal{Y}}) = \sigma^2\left(\boldsymbol{I}_n + \boldsymbol{ZTSST}'\boldsymbol{Z}'\right) = \sigma^2\left(\boldsymbol{I}_n + \boldsymbol{A}'\boldsymbol{A}\right) = \sigma^2\boldsymbol{V}(\boldsymbol{\theta}), \tag{17}$$

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

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

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y} \tag{18}$$

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 n is always larger than q — sometimes orders of magnitude larger. Blithely writing a formula in terms of $\boldsymbol{V}^{-1}$ when $\boldsymbol{V}$ is $n\times n$, and n can be in the millions does not a computational formula make.

3.1 Relating the GLS approach to the Cholesky factor

We can use the fact that

$$\boldsymbol{V}^{-1}(\boldsymbol{\theta}) = \left(\boldsymbol{I}_n + \boldsymbol{A}'\boldsymbol{A}\right)^{-1} = \boldsymbol{I}_n - \boldsymbol{A}'\left(\boldsymbol{I}_q + \boldsymbol{AA}'\right)^{-1}\boldsymbol{A} \tag{19}$$

to relate the GLS problem to the PLS problem. One way to establish (19) is simply to show that the product

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

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

$$\begin{aligned}\boldsymbol{V}^{-1}(\boldsymbol{\theta}) =& \boldsymbol{I}_n - \boldsymbol{A}'\boldsymbol{P}'\boldsymbol{P}\left(\boldsymbol{I}_q + \boldsymbol{A}\boldsymbol{A}'\right)^{-1}\boldsymbol{P}'\boldsymbol{P}\boldsymbol{A} \\ =& \boldsymbol{I}_n - \boldsymbol{A}'\boldsymbol{P}'(\boldsymbol{L}\boldsymbol{L}')^{-1}\boldsymbol{P}\boldsymbol{A} \\ =& \boldsymbol{I}_n - \left(\boldsymbol{L}^{-1}\boldsymbol{P}\boldsymbol{A}\right)'\boldsymbol{L}^{-1}\boldsymbol{P}\boldsymbol{A}.\end{aligned} \tag{20}$$

Even in this form we would not want to routinely evaluate $\boldsymbol{V}^{-1}$. However, (20) does allow us to simplify many common expressions.

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

$$\begin{aligned}\sigma^2\left(\boldsymbol{X}'\boldsymbol{V}^{-1}(\boldsymbol{\theta})\boldsymbol{X}\right)^{-1} =& \sigma^2\left(\boldsymbol{X}'\boldsymbol{X} - \left(\boldsymbol{L}^{-1}\boldsymbol{P}\boldsymbol{A}\boldsymbol{X}\right)'\left(\boldsymbol{L}^{-1}\boldsymbol{P}\boldsymbol{A}\boldsymbol{X}\right)\right)^{-1} \\ =& \sigma^2\left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{R}_{\boldsymbol{Z}\boldsymbol{X}}'\boldsymbol{R}_{\boldsymbol{Z}\boldsymbol{X}}\right)^{-1} \\ =& \sigma^2\left(\boldsymbol{R}_{\boldsymbol{X}}'\boldsymbol{R}_{\boldsymbol{X}}\right)^{-1}.\end{aligned} \tag{21}$$

4 Trace of the "hat" matrix

Another calculation that is of interest to some is the the trace of the "hat" matrix, which can be written as

$$\begin{aligned}&\operatorname{tr}\left(\begin{bmatrix}\boldsymbol{A}' & \boldsymbol{X}\end{bmatrix}\left(\begin{bmatrix}\boldsymbol{A}' & \boldsymbol{X} \\ \boldsymbol{I} & \boldsymbol{0}\end{bmatrix}'\begin{bmatrix}\boldsymbol{A}' & \boldsymbol{X} \\ \boldsymbol{I} & \boldsymbol{0}\end{bmatrix}\right)^{-1}\begin{bmatrix}\boldsymbol{A} \\ \boldsymbol{X}'\end{bmatrix}\right) \\ &\qquad = \operatorname{tr}\left(\begin{bmatrix}\boldsymbol{A}' & \boldsymbol{X}\end{bmatrix}\left(\begin{bmatrix}\boldsymbol{L} & \boldsymbol{0} \\ \boldsymbol{R}_{\boldsymbol{Z}\boldsymbol{X}}' & \boldsymbol{R}_{\boldsymbol{X}}'\end{bmatrix}\begin{bmatrix}\boldsymbol{L}' & \boldsymbol{R}_{\boldsymbol{Z}\boldsymbol{X}} \\ \boldsymbol{0} & \boldsymbol{R}_{\boldsymbol{X}}\end{bmatrix}\right)^{-1}\begin{bmatrix}\boldsymbol{A} \\ \boldsymbol{X}'\end{bmatrix}\right)\end{aligned} \tag{22}$$