Linear mixed models and penalized least squares

Douglas M. Bates

Department of Statistics, University of Wisconsin–Madison

Saikat DebRoy

Department of Biostatistics, Harvard School of Public Health

Abstract

Linear mixed-effects models are an important class of statistical models that are not only used directly in many fields of applications but also used as iterative steps in fitting other types of mixed-effects models, such as generalized linear mixed models. The parameters in these models are typically estimated by maximum likelihood (ML) or restricted maximum likelihood (REML). In general there is no closed form solution for these estimates and they must be determined by iterative algorithms such as EM iterations or general nonlinear optimization. Many of the intermediate calculations for such iterations have been expressed as generalized least squares problems. We show that an alternative representation as a penalized least squares problem has many advantageous computational properties including the ability to evaluate explicitly a profiled log-likelihood or log-restricted likelihood, the gradient and Hessian of this profiled objective, and an ECME update to refine this objective.

Key words: REML, gradient, Hessian, EM algorithm, ECME algorithm, maximum likelihood, profile likelihood, multilevel models

1 Introduction

We will establish some results for the penalized least squares representation of a general form of a linear mixed-effects model, then show how these results specialize for particular models. The general form we consider is

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

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 $\mathbf{0}$ 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 $\boldsymbol{\Omega}$ is positive definite. (If $\boldsymbol{X}$ is rank deficient or if $\boldsymbol{\Omega}$ is singular then the model can be transformed to an alternative model that fulfills the desired conditions.)

A relative precision factor, $\boldsymbol{\Delta}$, is any $q \times q$ matrix that satisfies $\boldsymbol{\Omega} = \boldsymbol{\Delta}^\mathsf{T}\boldsymbol{\Delta}$. One possible $\boldsymbol{\Delta}$ is the Cholesky factor of $\boldsymbol{\Omega}$ but others can be used. Because $\boldsymbol{\Omega}$ is positive definite, any $\boldsymbol{\Delta}$ will be non-singular. In general $\boldsymbol{\Delta}$ (and hence $\boldsymbol{\Omega}$) depend on a k-dimensional parameter vector $\boldsymbol{\theta}$. Typically $\boldsymbol{Z}$, $\boldsymbol{\Omega}$, and $\boldsymbol{\Delta}$ are very large and sparse (mostly zeros) while k, the dimension of $\boldsymbol{\theta}$, is small.

The likelihood for the parameters $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, and σ^2, given $\boldsymbol{y}$, in model (1) is

$$L\left(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2 | \boldsymbol{y}\right) = \int \frac{\sqrt{|\boldsymbol{\Omega}|}}{\left(2\pi\sigma^2\right)^{(n+q)/2}} \exp\left(\frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{b}\|^2 + \boldsymbol{b}^\mathsf{T}\boldsymbol{\Omega}\boldsymbol{b}}{-2\sigma^2}\right) d\boldsymbol{b}. \quad (2)$$

The restricted (or residual) maximum likelihood (REML) estimates of $\boldsymbol{\theta}$ and σ^2 optimize a related criterion that can be written

$$L_R(\boldsymbol{\theta}, \sigma^2) = \int L(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2)\, d\boldsymbol{\beta}. \quad (3)$$

This integral is not the typical way to derive or to justify REML estimates but, as we shall see, it provides a convenient form in which to evaluate L_R.

In the next section we show that the integrals in (2) and (3) can be expressed succinctly using the solution to a penalized least squares problem. In particular, we derive a profiled log-likelihood and a profiled log-restricted likelihood that depend on $\boldsymbol{\theta}$ only. Maximizing the profiled log-likelihood (or profiled log-restricted-likelihood) is generally a much smaller and more stable optimization problem than attempting to optimize the log-likelihood for the full parameter vector. In section 3 we derive new expressions for the gradient and the Hessian of these profiled log-likelihoods. Using these derivatives, the profiled log-likelihood can be rapidly optimized, say by Newton steps, once the neighbourhood of $\widehat{\boldsymbol{\theta}}$ has been determined.

Some heuristics provide reasonable starting estimates for $\boldsymbol{\theta}$ but these may not be sufficiently close to $\widehat{\boldsymbol{\theta}}$ to ensure stable Newton steps. The expectation-maximization (EM) algorithm is a robust algorithm that approaches the neighbourhood of $\widehat{\boldsymbol{\theta}}$ quickly but tends to converge slowly once it is in the neighbourhood. In section 4 we show how the penalized least squares results provide the update for a related algorithm called "expectation-conditional maximization-either" (ECME). This allows the starting estimates to be updated by a moderate number of ECME iterations before starting the Newton steps.

These results are derived for the general model (1). In sections 5 and 6 we consider several common special cases. Our implementation of some of these methods is discussed in section 7.

2 A Penalized Least-Squares Problem

For a fixed value of $\boldsymbol{\theta}$ we consider the penalized least squares problem defined by the augmented model matrix $\boldsymbol{\Phi}(\boldsymbol{\theta})$ and the augmented response vector $\tilde{\boldsymbol{y}}$;

$$\min_{\boldsymbol{b},\boldsymbol{\beta}} \left\| \tilde{\boldsymbol{y}} - \boldsymbol{\Phi}(\boldsymbol{\theta}) \begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{\beta} \end{bmatrix} \right\|^2 \text{ where } \boldsymbol{\Phi}(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{Z} & \boldsymbol{X} \\ \boldsymbol{\Delta}(\boldsymbol{\theta}) & \boldsymbol{0} \end{bmatrix} \text{ and } \tilde{\boldsymbol{y}} = \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{bmatrix}. \tag{4}$$

One way to solve problem (4) is to form $\boldsymbol{\Phi}_e = [\boldsymbol{\Phi}, \tilde{\boldsymbol{y}}]$ and let $\boldsymbol{R}_e$ be the Cholesky decomposition of $\boldsymbol{\Phi}_e^\mathsf{T}\boldsymbol{\Phi}_e$

$$\boldsymbol{\Phi}_e^\mathsf{T}\boldsymbol{\Phi}_e = \begin{bmatrix} \boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} + \boldsymbol{\Omega} & \boldsymbol{Z}^\mathsf{T}\boldsymbol{X} & \boldsymbol{Z}^\mathsf{T}\boldsymbol{y} \\ \boldsymbol{X}^\mathsf{T}\boldsymbol{Z} & \boldsymbol{X}^\mathsf{T}\boldsymbol{X} & \boldsymbol{X}^\mathsf{T}\boldsymbol{y} \\ \boldsymbol{y}^\mathsf{T}\boldsymbol{Z} & \boldsymbol{y}^\mathsf{T}\boldsymbol{X} & \boldsymbol{y}^\mathsf{T}\boldsymbol{y} \end{bmatrix} = \boldsymbol{R}_e^\mathsf{T}\boldsymbol{R}_e \text{ where } \boldsymbol{R}_e = \begin{bmatrix} \boldsymbol{R}_{ZZ} & \boldsymbol{R}_{ZX} & \boldsymbol{r}_{Zy} \\ \boldsymbol{0} & \boldsymbol{R}_{XX} & \boldsymbol{r}_{Xy} \\ \boldsymbol{0} & \boldsymbol{0} & r_{yy} \end{bmatrix}. \tag{5}$$

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 $\boldsymbol{\Phi}$ has full column rank, and hence that $\boldsymbol{R}_{ZZ}$ and $\boldsymbol{R}_{XX}$ are nonsingular.

Representation (5) is a particular form of the mixed model equations described in Henderson (1984) [1]. We write the blocks in the opposite order from which they are typically written because of computational advantages associated with this order.

Using $\boldsymbol{a} = \left[-\boldsymbol{b}^\mathsf{T}, -\boldsymbol{\beta}^\mathsf{T}, 1\right]^\mathsf{T}$ we can write the numerator of the exponent in the integral in (2) as

$$\begin{aligned} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{b}\|^2 + \boldsymbol{b}^\mathsf{T}\boldsymbol{\Omega}\boldsymbol{b} = \boldsymbol{a}^\mathsf{T}\boldsymbol{\Phi}_e^\mathsf{T}\boldsymbol{\Phi}_e\boldsymbol{a} = \boldsymbol{a}^\mathsf{T}\boldsymbol{R}_e^\mathsf{T}\boldsymbol{R}_e\boldsymbol{a} = \|\boldsymbol{R}_e\boldsymbol{a}\|^2 \\ = \|\boldsymbol{r}_{Zy} - \boldsymbol{R}_{ZX}\boldsymbol{\beta} - \boldsymbol{R}_{ZZ}\boldsymbol{b}\|^2 + \|\boldsymbol{r}_{Xy} - \boldsymbol{R}_{XX}\boldsymbol{\beta}\|^2 + r_{yy}^2 \end{aligned} \tag{6}$$

and a simple change of variable allows us to evaluate

$$\int \frac{1}{(2\pi\sigma^2)^{q/2}} \exp\left(\frac{\|\boldsymbol{r}_{Zy} - \boldsymbol{R}_{ZX}\boldsymbol{\beta} - \boldsymbol{R}_{ZZ}\boldsymbol{b}\|^2}{-2\sigma^2}\right) d\boldsymbol{b} = \frac{1}{\operatorname{abs}|\boldsymbol{R}_{ZZ}|} = \frac{1}{\sqrt{|\boldsymbol{Z}^\top\boldsymbol{Z} + \boldsymbol{\Omega}|}} \tag{7}$$

Combining (2), (6), and (7) and taking the logarithm produces the log-likelihood $\ell(\boldsymbol{\beta}, \theta, \sigma^2) = \log L(\boldsymbol{\beta}, \theta, \sigma^2)$, which, for convenience, we write in the form of a deviance

$$-2\ell(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2) = \log\left(\frac{\left|\boldsymbol{Z}^\top\boldsymbol{Z} + \boldsymbol{\Omega}\right|}{|\boldsymbol{\Omega}|}\right) + n\log(2\pi\sigma^2) + \frac{r_{yy}^2 + \|\boldsymbol{r}_{Xy} - \boldsymbol{R}_{XX}\boldsymbol{\beta}\|^2}{\sigma^2} \tag{8}$$

leading to the following results for the maximum likelihood estimates (mle's):

(1) The conditional mle of the fixed-effects, $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$, satisfies

$$\boldsymbol{R}_{XX}\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \boldsymbol{r}_{Xy} \tag{9}$$

(2) The conditional mle of the variance, $\widehat{\sigma^2}(\boldsymbol{\theta})$, is r_{yy}^2/n.

(3) The profiled log-likelihood, $\tilde{\ell}(\boldsymbol{\theta})$, a function of $\boldsymbol{\theta}$ only, is given by

$$\begin{aligned} -2\tilde{\ell}(\boldsymbol{\theta}) &= -2\ell\left(\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \boldsymbol{\theta}, \widehat{\sigma^2}(\boldsymbol{\theta})\right) \\ &= \log\left(\frac{\left|\boldsymbol{Z}^\top\boldsymbol{Z} + \boldsymbol{\Omega}\right|}{|\boldsymbol{\Omega}|}\right) + n\left[1 + \log\left(\frac{2\pi r_{yy}^2}{n}\right)\right] \end{aligned} \tag{10}$$

(4) The conditional expected value of $\boldsymbol{b}$, which we write as $\widehat{\boldsymbol{b}}(\boldsymbol{\beta}, \boldsymbol{\theta})$, satisfies

$$\boldsymbol{R}_{ZZ}\widehat{\boldsymbol{b}}(\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\theta}) = \boldsymbol{r}_{Zy} - \boldsymbol{R}_{ZX}\boldsymbol{\beta} \tag{11}$$

Typically we evaluate $\widehat{\boldsymbol{b}}(\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \boldsymbol{\theta})$, which we write as $\widehat{\boldsymbol{b}}(\boldsymbol{\theta})$.

(5) The conditional distribution of $\boldsymbol{b}$ is

$$\boldsymbol{b}|\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2 \sim \mathcal{N}\left(\widehat{\boldsymbol{b}}(\boldsymbol{\beta}, \boldsymbol{\theta}), \sigma^2\left(\boldsymbol{Z}^\top\boldsymbol{Z} + \boldsymbol{\Omega}\right)^{-1}\right) \tag{12}$$

2.1 REML results

As in (7), we can use a simple change of variable to obtain

$$\int \frac{1}{(2\pi\sigma^2)^{p/2}} \exp\left(\frac{\|\boldsymbol{r}_{Xy} - \boldsymbol{R}_{XX}\boldsymbol{\beta}\|^2}{-2\sigma^2}\right) d\boldsymbol{\beta} = \frac{1}{\operatorname{abs}|\boldsymbol{R}_{XX}|} \tag{13}$$

providing the log-restricted-likelihood, $\ell_R(\boldsymbol{\theta}, \sigma^2) = \log L_R(\boldsymbol{\theta}, \sigma^2)$, as

$$-2\ell_R(\boldsymbol{\theta}, \sigma^2) = \log\left(\frac{\left|\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} + \boldsymbol{\Omega}\right| |\boldsymbol{R}_{XX}|^2}{|\boldsymbol{\Omega}|}\right) + (n-p)\log(2\pi\sigma^2) + \frac{r_{yy}^2}{\sigma^2}. \quad (14)$$

Noting that $\left|\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} + \boldsymbol{\Omega}\right| |\boldsymbol{R}_{XX}|^2 = \left|\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\right|$ we have the following results:

(1) The conditional REML estimate of the variance, $\widehat{\sigma^2}_R(\boldsymbol{\theta})$, is $r_{yy}^2/(n-p)$.
(2) The profiled log-restricted-likelihood is given by

$$-2\tilde{\ell}_R(\boldsymbol{\theta}) = \log\left(\frac{\left|\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\right|}{|\boldsymbol{\Omega}|}\right) + (n-p)\left[1 + \log\left(\frac{2\pi r_{yy}^2}{n-p}\right)\right]. \quad (15)$$

(3) The conditional distribution of $\boldsymbol{b}$ is

$$\boldsymbol{b}|\boldsymbol{y}, \boldsymbol{\theta}, \sigma^2 \sim \mathcal{N}\left(\widehat{\boldsymbol{b}}(\boldsymbol{\theta}), \sigma^2\boldsymbol{V}_{\boldsymbol{b}}\right) \quad (16)$$

where $\boldsymbol{V}_{\boldsymbol{b}}$ is the upper-left $q \times q$ submatrix of $\left(\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\right)^{-1}$,

$$\boldsymbol{V}_{\boldsymbol{b}} = \boldsymbol{R}_{ZZ}^{-1}\left(\boldsymbol{I} + \boldsymbol{R}_{ZX}\boldsymbol{R}_{XX}^{-1}\boldsymbol{R}_{XX}^{-\mathsf{T}}\boldsymbol{R}_{ZX}^\mathsf{T}\right)\boldsymbol{R}_{ZZ}^{-\mathsf{T}} \quad (17)$$

2.2 A preliminary decomposition

Because we wish to evaluate the log-likelihood or log-restricted-likelihood for many different values of $\boldsymbol{\theta}$, we form a preliminary decomposition of the cross-products of the model matrices and the response,

$$\begin{bmatrix} \boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} & \boldsymbol{Z}^\mathsf{T}\boldsymbol{X} & \boldsymbol{Z}^\mathsf{T}\boldsymbol{y} \\ \boldsymbol{X}^\mathsf{T}\boldsymbol{Z} & \boldsymbol{X}^\mathsf{T}\boldsymbol{X} & \boldsymbol{X}^\mathsf{T}\boldsymbol{y} \\ \boldsymbol{y}^\mathsf{T}\boldsymbol{Z} & \boldsymbol{y}^\mathsf{T}\boldsymbol{X} & \boldsymbol{y}^\mathsf{T}\boldsymbol{y} \end{bmatrix} = \boldsymbol{T}^\mathsf{T}\boldsymbol{T} \text{ where } \boldsymbol{T} = \begin{bmatrix} \boldsymbol{T}_{ZZ} & \boldsymbol{T}_{ZX} & \boldsymbol{t}_{Zy} \\ \boldsymbol{0} & \boldsymbol{T}_{XX} & \boldsymbol{t}_{Xy} \\ \boldsymbol{0} & \boldsymbol{0} & t_{yy} \end{bmatrix} \quad (18)$$

so that the evaluation of (5) is equivalent to forming the orthogonal-triangular decomposition

$$\begin{bmatrix} \boldsymbol{\Delta} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{T}_{ZZ} & \boldsymbol{T}_{ZX} & \boldsymbol{t}_{Zy} \\ \boldsymbol{0} & \boldsymbol{T}_{XX} & \boldsymbol{t}_{Xy} \\ \boldsymbol{0} & \boldsymbol{0} & t_{yy} \end{bmatrix} = \boldsymbol{Q}\begin{bmatrix} \boldsymbol{R}_{ZZ} & \boldsymbol{R}_{ZX} & \boldsymbol{r}_{Zy} \\ \boldsymbol{0} & \boldsymbol{R}_{XX} & \boldsymbol{r}_{Xy} \\ \boldsymbol{0} & \boldsymbol{0} & r_{yy} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}. \quad (19)$$

To ensure well-defined estimates for the parameters we should have $n \geq q + p$ which means that (18) produces a reduction, sometimes a substantial reduc-

tion, in the amount of data that must be stored and manipulated for each evaluation of the log-likelihood.

Furthermore, the matrices $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}$ and $\boldsymbol{T}_{ZZ}$ are sparse and we store and manipulate these matrices taking the sparsity into account (see §5 and §6 for specific examples). We will assume that $\boldsymbol{Z}^\mathsf{T}\boldsymbol{X}$, $\boldsymbol{T}_{ZX}$, $\boldsymbol{X}^\mathsf{T}\boldsymbol{X}$, and $\boldsymbol{T}_{XX}$ are dense and are stored accordingly.

Equation (19) shows that t_{yy}^2 is a lower bound for r_{yy}^2. Similarly, the residual sum of squares from regressing $\boldsymbol{y}$ on $\boldsymbol{X}$ is an upper bound on r_{yy}^2. Thus the terms in (10 and 15) involving r_{yy}^2 are bounded.

The ratio $|\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} + \boldsymbol{\Omega}|/|\boldsymbol{\Omega}|$ is bounded below by unity (and approaches this bound as $\boldsymbol{\Omega}^{-1} \to \mathbf{0}$) so the profiled deviance (10) is bounded below by $n\left[1 + \log\left(2\pi t_{yy}^2/n\right)\right]$ and the profiled restricted deviance (15) is bounded below by $(n-p)\left[1 + \log\left(2\pi t_{yy}^2/(n-p)\right)\right]$.

As $\boldsymbol{\Omega}$ approaches singularity (say the minimum eigenvalue of $\boldsymbol{\Omega}$ approaches zero while the other eigenvalues are bounded above) $|\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} + \boldsymbol{\Omega}|/|\boldsymbol{\Omega}| \to \infty$. Thus we know that the ML or REML estimates will not occur on the boundary of the set of positive definite $\boldsymbol{\Omega}$.

It is possible that finite ML or REML estimates of $\boldsymbol{\Omega}$ will not exist. The minimum profiled deviance may correspond to an infinite precision (unbounded $\boldsymbol{\Omega}$), which is to say that the ML or REML estimates of $\sigma^2\boldsymbol{\Omega}^{-1}$, the variance-covariance of $\boldsymbol{b}$, are singular.

3 Derivatives of the profiled log-likelihood

Expressions (10) and (15) provide extremely efficient ways to determine ML or REML estimates for a linear mixed-effects model because we can optimize these expressions as a function of $\boldsymbol{\theta}$ only, instead of as a function of the complete parameter vector $\left(\boldsymbol{\theta}^\mathsf{T}, \boldsymbol{\beta}^\mathsf{T}, \sigma^2\right)^\mathsf{T}$. The reduction in the dimension of the parameter over which we are optimizing helps to improve the performance and reliability of numerical optimization routines. Another way in which we can improve performance and reliability in numerical optimization is by providing analytic derivatives of the objective, which we do using results from Golub and Pereyra [2].

As in [2] we will use the operator D to indicate the Frechet derivative of an array or a scalar. For example, $\mathsf{D}\,\boldsymbol{\Phi}(\boldsymbol{\theta})$ is an array of dimension $n \times (q+p) \times k$, which we treat as k matrices of size $n \times (q+p)$ when writing expressions involving matrices and arrays. When we need to indicate partial derivatives

with respect to particular parameters we will use the notation $\mathsf{D}_i\boldsymbol{\Phi}(\boldsymbol{\theta})$ for the $n \times (q+p)$ matrix $\partial\boldsymbol{\Phi}/\partial\theta_i$.

Golub and Pereyra [2] provide derivatives of the projection orthogonal to the column space of $\boldsymbol{\Phi}$, which, because $\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}$ is nonsingular, we can write $P^\perp = \boldsymbol{I} - \boldsymbol{\Phi}\left(\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^\mathsf{T}$; the pseudo-inverse of $\boldsymbol{\Phi}$, which is $\boldsymbol{\Phi}^+ = \left(\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^\mathsf{T}$; and the residual sum of squares, $r_{yy}^2(\boldsymbol{\theta}) = \left\|P^\perp\tilde{\boldsymbol{y}}\right\|^2$. These derivatives are

$$\mathsf{D}\,P^\perp = -P^\perp\,\mathsf{D}\,\boldsymbol{\Phi}\boldsymbol{\Phi}^+ - \left(P^\perp\,\mathsf{D}\,\boldsymbol{\Phi}\boldsymbol{\Phi}^+\right)^\mathsf{T}, \tag{20}$$

$$\tfrac{1}{2}\nabla r_{yy}^2(\boldsymbol{\theta}) = -\tilde{\boldsymbol{y}}^\mathsf{T}P^\perp\,\mathsf{D}\,\boldsymbol{\Phi}\boldsymbol{\Phi}^+\tilde{\boldsymbol{y}}, \tag{21}$$

$$\mathsf{D}\,\boldsymbol{\Phi}^+ = -\boldsymbol{\Phi}^+\,\mathsf{D}\,\boldsymbol{\Phi}\boldsymbol{\Phi}^+ + \boldsymbol{\Phi}^+\left(\boldsymbol{\Phi}^+\right)^\mathsf{T}\mathsf{D}\,\boldsymbol{\Phi}^\mathsf{T}P^\perp. \tag{22}$$

(Equation (22) is derived from equation (4.12) in [2], which has another term. However, that term is identically zero when $\boldsymbol{\Phi}$ has full column rank.)

For $\boldsymbol{A}$ a square, nonsingular matrix we have

$$\nabla(\log|\boldsymbol{A}(\boldsymbol{\theta})|) = \operatorname{tr}\left[\mathsf{D}(\boldsymbol{A})\boldsymbol{A}^{-1}\right] \tag{23}$$

$$\mathsf{D}\left(\boldsymbol{A}^{-1}\right) = -\boldsymbol{A}^{-1}\,\mathsf{D}(\boldsymbol{A})\boldsymbol{A}^{-1} \tag{24}$$

where tr denotes the trace of a matrix. Notice that (24) is a special case of (22).

In Appendix A we show that the gradient and Hessian terms for the penalized residual sum-of-squares, $r_{yy}^2(\boldsymbol{\theta})$, are

$$\nabla r_{yy}^2(\boldsymbol{\theta}) = \hat{\boldsymbol{b}}^\mathsf{T}\,\mathsf{D}\,\boldsymbol{\Omega}\hat{\boldsymbol{b}} = \operatorname{tr}\left(\mathsf{D}\,\boldsymbol{\Omega}\hat{\boldsymbol{b}}\hat{\boldsymbol{b}}^\mathsf{T}\right) \tag{25}$$

$$\mathsf{D}_j\mathsf{D}_i r_{yy}^2(\boldsymbol{\theta}) = \hat{\boldsymbol{b}}^\mathsf{T}\left[\mathsf{D}_j(\mathsf{D}_i\boldsymbol{\Omega}) - 2\mathsf{D}_j\boldsymbol{\Omega}\boldsymbol{V}_{\boldsymbol{b}}\mathsf{D}_i\boldsymbol{\Omega}\right]\hat{\boldsymbol{b}} \tag{26}$$

providing the gradients

$$\nabla(-2\tilde{\ell}) = \operatorname{tr}\left[\mathsf{D}\,\boldsymbol{\Omega}\left((\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\boldsymbol{\Omega})^{-1} - \boldsymbol{\Omega}^{-1} + \frac{\hat{\boldsymbol{b}}}{\widehat{\sigma}}\frac{\hat{\boldsymbol{b}}}{\widehat{\sigma}}^\mathsf{T}\right)\right] \tag{27}$$

$$\nabla(-2\tilde{\ell}_R) = \operatorname{tr}\left[\mathsf{D}\,\boldsymbol{\Omega}\left(\boldsymbol{V}_{\boldsymbol{b}} - \boldsymbol{\Omega}^{-1} + \frac{\hat{\boldsymbol{b}}}{\widehat{\sigma}_R}\frac{\hat{\boldsymbol{b}}}{\widehat{\sigma}_R}^\mathsf{T}\right)\right] \tag{28}$$

and Hessian terms of the form

$$\begin{aligned}
\mathsf{D}_j\mathsf{D}_i(-2\tilde{\ell}) = \operatorname{tr}&\left[\mathsf{D}_j(\mathsf{D}_i\boldsymbol{\Omega})\left((\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\boldsymbol{\Omega})^{-1}-\boldsymbol{\Omega}^{-1}+\frac{\hat{\boldsymbol{b}}}{\widehat{\sigma}}\frac{\hat{\boldsymbol{b}}}{\widehat{\sigma}}^\mathsf{T}\right)\right] \\
&-\operatorname{tr}\left[\mathsf{D}_j\boldsymbol{\Omega}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\boldsymbol{\Omega})^{-1}\mathsf{D}_i\boldsymbol{\Omega}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\boldsymbol{\Omega})^{-1}\right] \\
&+\operatorname{tr}\left(\mathsf{D}_j\boldsymbol{\Omega}\boldsymbol{\Omega}^{-1}\mathsf{D}_i\boldsymbol{\Omega}\boldsymbol{\Omega}^{-1}\right)-2\frac{\hat{\boldsymbol{b}}}{\widehat{\sigma}}^\mathsf{T}\mathsf{D}_j\boldsymbol{\Omega}\boldsymbol{V}_{\boldsymbol{b}}\mathsf{D}_i\boldsymbol{\Omega}\frac{\hat{\boldsymbol{b}}}{\widehat{\sigma}} \\
&-\frac{1}{n}\left(\frac{\hat{\boldsymbol{b}}}{\widehat{\sigma}}^\mathsf{T}\mathsf{D}_j\boldsymbol{\Omega}\frac{\hat{\boldsymbol{b}}}{\widehat{\sigma}}\right)\left(\frac{\hat{\boldsymbol{b}}}{\widehat{\sigma}}^\mathsf{T}\mathsf{D}_i\boldsymbol{\Omega}\frac{\hat{\boldsymbol{b}}}{\widehat{\sigma}}\right)
\end{aligned} \tag{29}$$

$$\begin{aligned}
\mathsf{D}_j\mathsf{D}_i(-2\tilde{\ell_R}) = \operatorname{tr}&\left[\mathsf{D}_j(\mathsf{D}_i\boldsymbol{\Omega})\left(\boldsymbol{V}_{\boldsymbol{b}}-\boldsymbol{\Omega}^{-1}+\frac{\hat{\boldsymbol{b}}}{\widehat{\sigma}_R}\frac{\hat{\boldsymbol{b}}}{\widehat{\sigma}_R}^\mathsf{T}\right)\right] \\
&-\operatorname{tr}\left[\mathsf{D}_j\boldsymbol{\Omega}\boldsymbol{V}_{\boldsymbol{b}}\mathsf{D}_i\boldsymbol{\Omega}\boldsymbol{V}_{\boldsymbol{b}}\right] \\
&+\operatorname{tr}\left(\mathsf{D}_j\boldsymbol{\Omega}\boldsymbol{\Omega}^{-1}\mathsf{D}_i\boldsymbol{\Omega}\boldsymbol{\Omega}^{-1}\right)-2\frac{\hat{\boldsymbol{b}}}{\widehat{\sigma}_R}^\mathsf{T}\mathsf{D}_j\boldsymbol{\Omega}\boldsymbol{V}_{\boldsymbol{b}}\mathsf{D}_i\boldsymbol{\Omega}\frac{\hat{\boldsymbol{b}}}{\widehat{\sigma}_R} \\
&-\frac{1}{n-p}\left(\frac{\hat{\boldsymbol{b}}}{\widehat{\sigma}_R}^\mathsf{T}\mathsf{D}_j\boldsymbol{\Omega}\frac{\hat{\boldsymbol{b}}}{\widehat{\sigma}_R}\right)\left(\frac{\hat{\boldsymbol{b}}}{\widehat{\sigma}_R}^\mathsf{T}\mathsf{D}_i\boldsymbol{\Omega}\frac{\hat{\boldsymbol{b}}}{\widehat{\sigma}_R}\right)
\end{aligned} \tag{30}$$

4 An ECME algorithm

The EM algorithm [3] is a general iterative algorithm for computing maximum likelihood estimates in the presence of missing data. For linear mixed-effects models we formulate an EM algorithm by considering the random effects $\boldsymbol{b}$ to be unobserved data. In the terminology of the EM algorithm, we call the observed data, $\boldsymbol{y}$, the *incomplete* data, and $\boldsymbol{y}$ augmented by $\boldsymbol{b}$, the *complete* data.

The EM algorithm has two steps: in the E step we compute Q, the expected log-likelihood (or deviance) for the complete data, and, in the M step, we maximize the expected log-likelihood (or minimize the expected deviance) with respect to the parameters in the model.

Liu and Rubin [4] derived the EM algorithm for linear mixed-effects models using $\boldsymbol{b}$ as the missing data. In same paper they introduced *expectation conditional maximization either* (ECME) algorithms, which are an extension of the EM algorithm. In ECME algorithms the M step is broken down into a number of conditional maximization steps and in each conditional maximization step either the original log-likelihood, ℓ, or its conditional expectation, Q, is maximized. The maximization in each step is done by placing constraints on the parameters in such a way that the collection of all the maximization steps

is with respect to the full parameter space.

In describing an EM algorithm we must distinguish between current values of parameters and updated values. We denote the current values of the parameters by $\boldsymbol{\beta}_0$, σ_0^2 and $\boldsymbol{\theta}_0$. These are either starting values or values obtained from the last E and M steps. The parameter estimates to be obtained after an E and an M step are $\boldsymbol{\beta}_1$, σ_1^2 and $\boldsymbol{\theta}_1$. The log-likelihood for the complete data is

$$-2\ell(\boldsymbol{\beta},\sigma^2,\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{b}) = (n+q)\log(2\pi\sigma^2) - \log|\boldsymbol{\Omega}| + \frac{\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{b}\|^2 + \boldsymbol{b}^\mathsf{T}\boldsymbol{\Omega}\boldsymbol{b}}{\sigma^2} \tag{31}$$

Because we can easily calculate $\widehat{\sigma^2}(\boldsymbol{\theta})$ and $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ we define an update step in an ECME algorithm to be:

(1) Given $\boldsymbol{\theta}_0$, set $\boldsymbol{\beta}_1 = \widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}_0)$ and $\sigma_1^2 = \widehat{\sigma^2}(\boldsymbol{\theta}_0)$. Then the conditional distribution of $\boldsymbol{b}$ is $\mathcal{N}\left(\widehat{\boldsymbol{b}}(\boldsymbol{\theta}), \sigma_1^2\left(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\boldsymbol{\Omega}\right)^{-1}\right)$.

(2) Choose $\boldsymbol{\theta}_1$ to minimize the conditional expectation of -2ℓ

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{y},\sigma_1^2,\boldsymbol{\beta}_1,\boldsymbol{\theta}_0) &= \mathrm{E}_{\boldsymbol{b}|\boldsymbol{\theta}_0}[-2\ell(\boldsymbol{\beta}_1,\sigma_1^2,\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{b})] \\ &= \mathrm{E}_{\boldsymbol{b}|\boldsymbol{\theta}_0}\left[c - \log|\boldsymbol{\Omega}| + \boldsymbol{b}^\mathsf{T}\boldsymbol{\Omega}\boldsymbol{b}/\sigma_1^2\right] \\ &= c - \log|\boldsymbol{\Omega}| + \frac{\hat{\boldsymbol{b}}^\mathsf{T}\boldsymbol{\Omega}\hat{\boldsymbol{b}}}{\sigma_1^2} + \operatorname{tr}\left[\boldsymbol{\Omega}\left(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\boldsymbol{\Omega}(\boldsymbol{\theta}_0)\right)^{-1}\right] \end{aligned} \tag{32}$$

Thus $\boldsymbol{\theta}_1$ satisfies

$$\nabla_{\boldsymbol{\theta}} Q = \operatorname{tr}\left[\mathsf{D}\,\boldsymbol{\Omega}\left(\frac{\hat{\boldsymbol{b}}(\boldsymbol{\theta}_0)}{\sigma_1}\frac{\hat{\boldsymbol{b}}(\boldsymbol{\theta}_0)^\mathsf{T}}{\sigma_1} + \left(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\boldsymbol{\Omega}(\boldsymbol{\theta}_0)\right)^{-1} - \boldsymbol{\Omega}(\boldsymbol{\theta}_1)^{-1}\right)\right] = \boldsymbol{0} \tag{33}$$

A similar derivation shows that, for the REML criterion, $\boldsymbol{\theta}_1$ satisfies

$$\nabla_{\boldsymbol{\theta}} Q_R = \operatorname{tr}\left[\mathsf{D}\,\boldsymbol{\Omega}\left(\frac{\hat{\boldsymbol{b}}(\boldsymbol{\theta}_0)}{\sigma_R}\frac{\hat{\boldsymbol{b}}(\boldsymbol{\theta}_0)^\mathsf{T}}{\sigma_R} + \boldsymbol{V}_{\boldsymbol{b}}(\boldsymbol{\theta}_0) - \boldsymbol{\Omega}(\boldsymbol{\theta}_1)^{-1}\right)\right] = \boldsymbol{0} \tag{34}$$

From the similarity of (33) to (27) (and of (34) to (28)) we can see that a stationary point of this ECME algorithm will be a critical value of the profiled log-likelihood.

5 Computational methods for a single grouping factor

The results in the previous sections provide concise expressions for the profiled log-likelihood (10), the profiled log-restricted-likelihood (15); the ECME

increments (33 or 34); and the gradient (27 or 28) and Hessian (29 or 30) of the profiled objective functions. All these expressions depend on being able to evaluate the initial decomposition (18) and, for several different $\boldsymbol{\theta}$, the decomposition (19), which can be a formidable computational problem because the matrices $\boldsymbol{Z}$ and $\boldsymbol{\Omega}$ can be very large. However, these matrices generally are sparse and, by exploiting the sparsity, we can provide computationally feasible methods for all these results.

The sparsity in $\boldsymbol{Z}$ and $\boldsymbol{\Omega}$ occurs when the random effects vector $\boldsymbol{b}$ is divided into small components associated with one or more factors that group the observations. In the simplest situation there is one grouping factor, or one set of *experimental units*, and the model can be written

$$\begin{aligned} \boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{\Omega}_1^{-1}), \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}), \quad i = 1, \ldots, m, \\ \boldsymbol{\epsilon}_i \perp \boldsymbol{\epsilon}_j, \quad \boldsymbol{b}_i \perp \boldsymbol{b}_j, \quad i \neq j; \qquad \boldsymbol{\epsilon}_i \perp \boldsymbol{b}_j, \quad \text{all } i, j \end{aligned} \tag{35}$$

where $\boldsymbol{y}_i$ is the vector of length n_i of responses for unit i; $\boldsymbol{X}_i$ is the $n_i \times p$ fixed-effects model matrix for unit i; and $\boldsymbol{Z}_i$ is the $n_i \times q_1$ model matrix for unit i and the random effects $\boldsymbol{b}_i$ corresponding to that unit.

Because we only have one grouping factor in this model we say that we have one "level" of random effects. However, this terminology is not universal. In particular, this model is called the "two-level" model in the multilevel modeling literature (e.g. [5]) because it has two levels of random variation.

To convert model (35) to the form (1) we would set

$$\boldsymbol{b} = \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \\ \vdots \\ \boldsymbol{b}_{m_1} \end{bmatrix}, \ \boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_{m_1} \end{bmatrix}, \ \boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_{m_1} \end{bmatrix}, \ \boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Z}_1 & \boldsymbol{0} & \ldots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Z}_2 & \ldots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \ldots & \boldsymbol{Z}_{m_1} \end{bmatrix}, \tag{36}$$

and

$$\boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\Omega}_1 & \boldsymbol{0} & \ldots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Omega}_1 & \ldots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \ldots & \boldsymbol{\Omega}_1 \end{bmatrix} = \boldsymbol{I}_{m_1} \otimes \boldsymbol{\Omega}_1 \tag{37}$$

where $\boldsymbol{I}_{m_1}$ is the $m_1 \times m_1$ identity matrix and $\otimes$ is the Kronecker product. That is, $\boldsymbol{\Omega}$ is a block-diagonal matrix whose diagonal is m_1 copies of $\boldsymbol{\Omega}_1$. Similarly, a relative precision factor is $\boldsymbol{\Delta} = \boldsymbol{I}_{m_1} \otimes \boldsymbol{\Delta}_1$ where $\boldsymbol{\Delta}_1$ is any $q_1 \times q_1$ matrix satisfying $\boldsymbol{\Delta}_1^\top \boldsymbol{\Delta}_1 = \boldsymbol{\Omega}_1$.

The calculation of $\boldsymbol{R}_{XX}$,

$$\boldsymbol{R}_{ZZ} = \begin{bmatrix} \boldsymbol{R}_{ZZ(1)} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{R}_{ZZ(2)} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{R}_{ZZ(m_1)} \end{bmatrix}, \quad \boldsymbol{R}_{ZX} = \begin{bmatrix} \boldsymbol{R}_{ZX(1)} \\ \boldsymbol{R}_{ZX(2)} \\ \vdots \\ \boldsymbol{R}_{ZX(m_1)} \end{bmatrix} \tag{38}$$

and $\boldsymbol{r}_{Zy} = \left[\boldsymbol{r}_{Zy(1)}, \boldsymbol{r}_{Zy(2)}, \dots, \boldsymbol{r}_{Zy(m_1)}\right]^{\mathsf{T}}$ can be performed in blocks, using a series of QR decompositions based on the corresponding blocks in $\boldsymbol{T}$

$$\begin{bmatrix} \boldsymbol{\Delta}_1 \\ \boldsymbol{T}_{ZZ(i)} \end{bmatrix} = \boldsymbol{Q}_i \begin{bmatrix} \boldsymbol{R}_{ZZ(i)} \\ \mathbf{0} \end{bmatrix} \text{ followed by } \boldsymbol{Q}_i^{\mathsf{T}} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \boldsymbol{T}_{ZX(i)} & \boldsymbol{t}_{Zy(i)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{R}_{ZX(i)} & \boldsymbol{r}_{Zy(i)} \\ \boldsymbol{R}_{XX(i)} & \boldsymbol{r}_{Xy(i)} \end{bmatrix}, \tag{39}$$

and

$$\begin{bmatrix} \boldsymbol{R}_{XX(1)} & \boldsymbol{r}_{Xy(1)} \\ \vdots & \vdots \\ \boldsymbol{R}_{XX(m_1)} & \boldsymbol{r}_{Xy(m_1)} \\ \boldsymbol{T}_{XX} & \boldsymbol{t}_{Xy} \\ \mathbf{0} & t_{yy} \end{bmatrix} = \boldsymbol{Q}_0 \begin{bmatrix} \boldsymbol{R}_{XX} & \boldsymbol{r}_{Xy} \\ \mathbf{0} & r_{yy} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}. \tag{40}$$

Noting that $|\boldsymbol{\Omega}| = |\boldsymbol{\Omega}_1|^{m_1} = |\boldsymbol{\Delta}_1|^{2m_1}$ and that $|\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z} + \boldsymbol{\Omega}| = \prod_{i=1}^{m_1} |\boldsymbol{R}_{ZZ(i)}|^2$, we can evaluate the profiled log-likelihood (10) or the profiled log-restricted-likelihood (15).

To evaluate the ECME increments we note that

$$\begin{aligned} \operatorname{tr}\left[\mathsf{D}_i(\boldsymbol{\Omega})\boldsymbol{\Omega}^{-1}\right] &= m_1 \operatorname{tr}\left[\mathsf{D}_i(\boldsymbol{\Omega}_1)\boldsymbol{\Omega}_1^{-1}\right] = m_1 \operatorname{tr}\left[\boldsymbol{\Delta}_1^{-\mathsf{T}}\mathsf{D}_i(\boldsymbol{\Omega}_1)\boldsymbol{\Delta}_1^{-1}\right] \\ \operatorname{tr}\left[\mathsf{D}_i(\boldsymbol{\Omega})\left(\frac{\hat{\boldsymbol{b}}(\boldsymbol{\theta}_0)}{\sigma_1}\frac{\hat{\boldsymbol{b}}(\boldsymbol{\theta}_0)^{\mathsf{T}}}{\sigma_1} + (\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z} + \boldsymbol{\Omega}(\boldsymbol{\theta}_0))^{-1}\right)\right] &= \operatorname{tr}\left[\boldsymbol{A}_1^{\mathsf{T}}\mathsf{D}_i(\boldsymbol{\Omega})\boldsymbol{A}_1\right] \\ \operatorname{tr}\left[\mathsf{D}_i(\boldsymbol{\Omega})\left(\frac{\hat{\boldsymbol{b}}(\boldsymbol{\theta}_0)}{\sigma_R}\frac{\hat{\boldsymbol{b}}(\boldsymbol{\theta}_0)^{\mathsf{T}}}{\sigma_R} + \boldsymbol{V}_{\boldsymbol{b}}\right)\right] &= \operatorname{tr}\left[\boldsymbol{A}_{R1}^{\mathsf{T}}\mathsf{D}_i(\boldsymbol{\Omega})\boldsymbol{A}_{R1}\right] \end{aligned} \tag{41}$$

where the matrices $\boldsymbol{A}_1$ and $\boldsymbol{A}_{R1}$ are obtained from the QR decompositions

$$\boldsymbol{U}_1\boldsymbol{A}_1 = \begin{bmatrix} \hat{\boldsymbol{b}}_1^\mathsf{T}/\sigma_1 \\ \boldsymbol{R}_{ZZ(1)}^{-\mathsf{T}} \\ \vdots \\ \hat{\boldsymbol{b}}_{m_1}^\mathsf{T}/\sigma_1 \\ \boldsymbol{R}_{ZZ(m_1)}^{-\mathsf{T}} \end{bmatrix} \quad \text{and} \quad \boldsymbol{U}_{R1}\boldsymbol{A}_{R1} = \begin{bmatrix} \hat{\boldsymbol{b}}_1^\mathsf{T}/\sigma_R \\ \boldsymbol{R}_{ZZ(1)}^{-\mathsf{T}} \\ -\boldsymbol{R}_{XX}^{-\mathsf{T}}\boldsymbol{R}_{ZX(1)}^\mathsf{T}\boldsymbol{R}_{ZZ(1)}^{-\mathsf{T}} \\ \vdots \\ \hat{\boldsymbol{b}}_{m_1}^\mathsf{T}/\sigma_R \\ \boldsymbol{R}_{ZZ(m_1)}^{-\mathsf{T}} \\ -\boldsymbol{R}_{XX}^{-\mathsf{T}}\boldsymbol{R}_{ZX(m_1)}^\mathsf{T}\boldsymbol{R}_{ZZ(m_1)}^{-\mathsf{T}} \end{bmatrix} \tag{42}$$

The matrix $\boldsymbol{\Omega}_1$ must be positive definite and symmetric. If no further structure is imposed on it then $k = q_1(q_1+1)/2$ and a suitable value of $\boldsymbol{\Delta}_1(\boldsymbol{\theta}_1)$ can be calculated as $\boldsymbol{\Delta}_1 = \sqrt{m_1}\boldsymbol{A}_1^{-\mathsf{T}}$ for ML estimation and $\boldsymbol{\Delta}_{R1} = \sqrt{m_1}\boldsymbol{A}_{R1}^{-\mathsf{T}}$ for REML. If further structure is imposed on $\boldsymbol{\Omega}_1(\boldsymbol{\theta})$, so that $k < q_1(q_1+1)/2$, then the gradient equations, (33) or (34), must be solved for the ECME update.

Evaluation of the gradients (27 or 28) of the profiled objective functions can be simplified in the same way as the ECME increment is, as can the first term in the Hessian (29 or 30).

The other terms in the Hessian can be simplified in various ways. Because $\mathsf{D}_i\boldsymbol{\Omega} = \boldsymbol{I}_{m_1} \otimes \mathsf{D}_i\boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}^{-1} = \boldsymbol{I}_{m_1} \otimes \boldsymbol{\Omega}_1^{-1}$

$$\begin{aligned} \operatorname{tr}\left(\mathsf{D}_j\boldsymbol{\Omega}\boldsymbol{\Omega}^{-1}\mathsf{D}_i\boldsymbol{\Omega}\boldsymbol{\Omega}^{-1}\right) &= m_1 \operatorname{tr}\left(\mathsf{D}_j\boldsymbol{\Omega}_1\boldsymbol{\Omega}_1^{-1}\mathsf{D}_i\boldsymbol{\Omega}_1\boldsymbol{\Omega}_1^{-1}\right) \\ &= m_1 \operatorname{tr}\left(\boldsymbol{\Delta}_1^{-\mathsf{T}}\mathsf{D}_j\boldsymbol{\Omega}_1\boldsymbol{\Delta}_1^{-1}\boldsymbol{\Delta}_1^{-\mathsf{T}}\mathsf{D}_i\boldsymbol{\Omega}_1\boldsymbol{\Delta}_1^{-1}\right) \end{aligned} \tag{43}$$

From the vectors $\boldsymbol{c}_{\ell i}$, $\boldsymbol{d}_{\ell i}$, and $\boldsymbol{g}_{\ell i}$, $\ell = 1, \ldots, m_1$; $i = 1, \ldots, k$ defined as

$$\begin{aligned} \boldsymbol{c}_{\ell i} &= \mathsf{D}_i\boldsymbol{\Omega}_1\hat{\boldsymbol{b}}_\ell \\ \boldsymbol{d}_{\ell i} &= \boldsymbol{R}_{ZZ(\ell)}^{-\mathsf{T}}\boldsymbol{c}_{\ell i} \\ \boldsymbol{g}_{\ell i} &= -\boldsymbol{R}_{XX}^{-\mathsf{T}}\boldsymbol{R}_{ZX(\ell)}^\mathsf{T}\boldsymbol{d}_{\ell i} \end{aligned} \tag{44}$$

we can evaluate

$$\begin{aligned} \left(\hat{\boldsymbol{b}}^\mathsf{T}\mathsf{D}_j\boldsymbol{\Omega}\hat{\boldsymbol{b}}\right)\left(\hat{\boldsymbol{b}}^\mathsf{T}\mathsf{D}_i\boldsymbol{\Omega}\hat{\boldsymbol{b}}\right) &= \left(\sum_{\ell=1}^{m_1}\boldsymbol{b}_\ell^\mathsf{T}\boldsymbol{c}_{\ell j}\right)\left(\sum_{\ell=1}^{m_1}\boldsymbol{b}_\ell^\mathsf{T}\boldsymbol{c}_{\ell i}\right) \\ \hat{\boldsymbol{b}}^\mathsf{T}\mathsf{D}_j\boldsymbol{\Omega}\boldsymbol{V}_{\boldsymbol{b}}\mathsf{D}_i\boldsymbol{\Omega}\hat{\boldsymbol{b}} &= \sum_{\ell=1}^{m_1}\boldsymbol{d}_{\ell j}^\mathsf{T}\boldsymbol{d}_{\ell i} + \left(\sum_{\ell=1}^{m_1}\boldsymbol{g}_{\ell j}\right)^\mathsf{T}\left(\sum_{\ell=1}^{m_1}\boldsymbol{g}_{\ell i}\right) \end{aligned} \tag{45}$$

The remaining terms in the Hessian expressions are evaluated as

$$\begin{aligned}\operatorname{tr}\left[\mathsf{D}_j\boldsymbol{\Omega}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\boldsymbol{\Omega})^{-1}\mathsf{D}_i\boldsymbol{\Omega}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}+\boldsymbol{\Omega})^{-1}\right] = \\ \sum_{\ell=1}^{m_1}\operatorname{tr}\left[\boldsymbol{R}_{ZZ(\ell)}^{-\mathsf{T}}\mathsf{D}_j\boldsymbol{\Omega}_1\boldsymbol{R}_{ZZ(\ell)}^{-1}\boldsymbol{R}_{ZZ(\ell)}^{-\mathsf{T}}\mathsf{D}_i\boldsymbol{\Omega}_1\boldsymbol{R}_{ZZ(\ell)}^{-1}\right] \quad (46)\end{aligned}$$

or

$$\begin{aligned}&\operatorname{tr}\left[\mathsf{D}_j\boldsymbol{\Omega}\boldsymbol{V_b}\mathsf{D}_i\boldsymbol{\Omega}\boldsymbol{V_b}\right] = \\ &\sum_{\ell=1}^{m_1}\operatorname{tr}\left[\boldsymbol{R}_{ZZ(\ell)}^{-\mathsf{T}}\mathsf{D}_j\boldsymbol{\Omega}_1\boldsymbol{R}_{ZZ(\ell)}^{-1}\boldsymbol{R}_{ZZ(\ell)}^{-\mathsf{T}}\mathsf{D}_i\boldsymbol{\Omega}_1\boldsymbol{R}_{ZZ(\ell)}^{-1}\right] \\ &+2\sum_{\ell=1}^{m_1}\operatorname{tr}\left[\boldsymbol{R}_{ZZ(\ell)}^{-\mathsf{T}}\mathsf{D}_j\boldsymbol{\Omega}_1\boldsymbol{R}_{ZZ(\ell)}^{-1}\boldsymbol{R}_{ZX(\ell)}\boldsymbol{R}_{XX}^{-1}\boldsymbol{R}_{XX}^{-\mathsf{T}}\boldsymbol{R}_{ZX(\ell)}^{\mathsf{T}}\boldsymbol{R}_{ZZ(\ell)}^{-\mathsf{T}}\mathsf{D}_i\boldsymbol{\Omega}_1\boldsymbol{R}_{ZZ(\ell)}\right] \\ &+\operatorname{tr}\left[\left(\sum_{\ell=1}^{m_1}\boldsymbol{R}_{XX}^{-\mathsf{T}}\boldsymbol{R}_{ZX(\ell)}^{\mathsf{T}}\boldsymbol{R}_{ZZ(\ell)}^{-\mathsf{T}}\mathsf{D}_j\boldsymbol{\Omega}_1\boldsymbol{R}_{ZZ(\ell)}^{-1}\boldsymbol{R}_{ZX(\ell)}\boldsymbol{R}_{XX}^{-1}\right)\right. \\ &\quad\left.\left(\sum_{\ell=1}^{m_1}\boldsymbol{R}_{XX}^{-\mathsf{T}}\boldsymbol{R}_{ZX(\ell)}^{\mathsf{T}}\boldsymbol{R}_{ZZ(\ell)}^{-\mathsf{T}}\mathsf{D}_i\boldsymbol{\Omega}_1\boldsymbol{R}_{ZZ(\ell)}^{-1}\boldsymbol{R}_{ZX(\ell)}\boldsymbol{R}_{XX}^{-1}\right)\right] \quad (47)\end{aligned}$$

By working with the components of $\boldsymbol{\Omega}$ instead of the whole matrix we save a considerable amount of storage and computation because q_1 is typically very small (values of one or two are common) while m_1 can be very large. We have worked with cases where m_1 is in the millions. In most cases $p \geq q_1$ but it is unusual for p to exceed, say, one hundred.

After the pre-decomposition the storage required is approximately $m_1q_1(q_1+p)$ locations and the number of floating point operations (FLOPs) per iteration for the function evaluation is on the order of $m_1q_1^3$ for the decompositions and $m_1pq_1^2$ for the multiplications in (39), followed by $m_1q_1p^2$ for (40). Evaluation of the determinants of $\boldsymbol{R}_{ZZ(i)}, i = 1, \ldots, m_1$ and $\boldsymbol{R}_{XX}$ is trivial because these matrices are triangular. The dominant term in the FLOP count is $m_1q_1(q_1 + p)^2$.

This is also the order of the computation for the ECME update, the gradient calculation and the Hessian calculation.

We can simplify some of the expressions for the Hessian by noting that, for arbitrary $q_1 \times q_1$ matrices $\boldsymbol{U}$ and $\boldsymbol{V}$,

$$\begin{aligned}\operatorname{tr}\left[\mathsf{D}_i\boldsymbol{\Omega}_1\boldsymbol{U}\mathsf{D}_j\boldsymbol{\Omega}_1\boldsymbol{V}\right] &= \sum_{k_1}\sum_{k_2}\sum_{k_3}\sum_{k_4} d_{k_1k_2}^{(i)}d_{k_3k_4}^{(j)}\operatorname{tr}\left[\boldsymbol{e}_{k_1}\boldsymbol{e}_{k_2}^\mathsf{T}\boldsymbol{U}\boldsymbol{e}_{k_3}\boldsymbol{e}_{k_4}^\mathsf{T}\boldsymbol{V}\right] \\ &= \sum_{k_1}\sum_{k_2}\sum_{k_3}\sum_{k_4} d_{k_1k_2}^{(i)}d_{k_3k_4}^{(j)}\left(\boldsymbol{e}_{k_2}^\mathsf{T}\boldsymbol{U}\boldsymbol{e}_{k_3}\boldsymbol{e}_{k_4}^\mathsf{T}\boldsymbol{V}\boldsymbol{e}_{k_1}\right) \\ &= \sum_{k_1}\sum_{k_2}\sum_{k_3}\sum_{k_4} d_{k_1k_2}^{(i)}d_{k_3k_4}^{(j)}u_{k_2k_3}v_{k_4k_1}\end{aligned}$$

where $d_{jk}^{(i)}$ is the (j,k)th element of $\mathsf{D}_i\boldsymbol{\Omega}_1$, $\boldsymbol{e}_k$ is the k-th column of the identity matrix of size $q_1 \times q_1$, and u_{ij} and v_{ij} are the (i,j)th elements of $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively.

For the one-level model, all the terms involved in the Hessian can be rewritten using this principle and so the full Hessian can be written as

$$\sum_{k_1k_2k_3k_k} d_{k_1k_2}^{(i)} d_{k_3k_4}^{(j)} \boldsymbol{H}_{k_1k_2k_3k_4}$$

where $\boldsymbol{H}$ is an array of size $q_1 \times q_1 \times q_1 \times q_1$.

6 Models with multiple grouping factors

We will consider models with multiple grouping factors where the random effects associated with each grouping factor are independent between groups and are i.i.d within groups. That is, the matrix $\boldsymbol{\Omega}$ consists of s blocks of the form $\boldsymbol{I}_{m_j} \otimes \boldsymbol{\Omega}_j$, $j = 1, \ldots, s$. We will give details for $s = 2$, in which case

$$\boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{I}_{m_1} \otimes \boldsymbol{\Omega}_1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_{m_2} \otimes \boldsymbol{\Omega}_2 \end{bmatrix}. \tag{48}$$

Extensions to more than two grouping factors follow naturally.

The block-diagonal structure in (48) will also be present in the factor, $\boldsymbol{\Delta}$, and in all the derivatives with respect to components of $\boldsymbol{\theta}$. That is, $\boldsymbol{\Delta}$ consists of s blocks of the form $\boldsymbol{I}_{m_j} \otimes \boldsymbol{\Delta}_j$, $j = 1, \ldots, s$ where $\boldsymbol{\Delta}_j$ is of size $q_j \times q_j$. We assume that each component of $\boldsymbol{\theta}$ determines only one block in $\boldsymbol{\Omega}$ and we designate the block associated with component j as $b(j)$ where $1 \leq b(j) \leq s$, $j = 1, \ldots, k$. Then $\mathsf{D}_j\boldsymbol{\Omega}$ is zero except for the $b(j)$th diagonal block which is of the form $\boldsymbol{I}_{m_{b(j)}} \otimes \mathsf{D}_j\boldsymbol{\Omega}_{b(j)}$. Notice that the second derivative, $\mathsf{D}_j\mathsf{D}_i\boldsymbol{\Omega}$, will be zero if $b(i) \neq b(j)$.

We divide the random effects vector $\boldsymbol{b}$ into $s = 2$ blocks and subdivide the jth block into m_j components of length q_j denoted by $\boldsymbol{b}_{ji}$, $j = 1, \ldots, s$; $i = 1, \ldots, m_j$.

If $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}$ is split into blocks corresponding to the blocks in $\boldsymbol{\Omega}$ then the diagonal blocks in $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}$ are themselves block diagonal but, unlike the situation with a single grouping factor, $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}$ and $\boldsymbol{T}_{ZZ}$ will have non-zero off-diagonal blocks. These off-diagonal blocks can be sparse or dense according to whether the grouping factors are nested or crossed. We distinguish three cases: completely crossed, partially crossed, and strictly nested.

Completely crossed grouping factors usually occur in designed experiments. For example, biological assays are often conducted by measuring the optical density of liquid samples in wells arranged in a grid on a plate. A common arrangement is 96 wells in a grid of 8 rows by 12 columns. If we assigned one set of random effects to the rows and another set of random effects to the columns then each row would occur with each column, resulting in completely crossed random effects.

An example of nested grouping factors would be a longitudinal study, say records of annual achievement test scores, of students in several schools. For the student factor to be strictly nested within the school factor we require that each student attend only one school during the period of the study. In most large studies this will not be the case. We expect some students will attend more than one school but we do not expect every student to attend every school. In such a case the grouping factors are neither strictly nested nor completely crossed. We describe this situation as partially crossed grouping factors.

If we divide $\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Z}_1 & \boldsymbol{Z}_2 \end{bmatrix}$ according to the two levels of random effects and correspondingly divide

$$\boldsymbol{T}_{ZZ} = \begin{bmatrix} \boldsymbol{T}_{ZZ11} & \boldsymbol{T}_{ZZ12} \\ \boldsymbol{0} & \boldsymbol{T}_{ZZ22} \end{bmatrix} \text{ and } \boldsymbol{R}_{ZZ} = \begin{bmatrix} \boldsymbol{R}_{ZZ11} & \boldsymbol{R}_{ZZ12} \\ \boldsymbol{0} & \boldsymbol{R}_{ZZ22} \end{bmatrix} \tag{49}$$

then the block diagonal structure of $\boldsymbol{Z}_1^\intercal \boldsymbol{Z}_1$ is also present in $\boldsymbol{T}_{ZZ11}$ and we can store and manipulate it accordingly. Even though $\boldsymbol{Z}_2^\intercal \boldsymbol{Z}_2$ is block diagonal there can be non-zero off-diagonals induced in $\boldsymbol{T}_{ZZ22}$ from $\boldsymbol{T}_{ZZ12}$ when taking the Cholesky decomposition. This is called "fill-in" [6].

If we consider the matrix $\boldsymbol{T}_{ZZ12}$ to consist of $m_1 \times m_2$ blocks of size $q_1 \times q_2$ then the (i,j)th such block will necessarily be zero if the ith level of the first grouping factor does not occur in combination with the jth level of the second grouping factor.

For strictly nested grouping factors there will be only one non-zero block in each of the m_1 sets of q_1 adjacent rows. Furthermore, this pattern guarantees that there will be no fill-in of $\boldsymbol{T}_{ZZ22}$ and that it can be stored as a block-diagonal matrix consisting of m_2 blocks of size $q_2 \times q_2$. The matrix $\boldsymbol{R}_{ZZ}$ has the same pattern of sparsity so $\boldsymbol{T}_{ZZ}$ and $\boldsymbol{R}_{ZZ}$ only require storage of size $(m_1 q_1 + m_2 q_2)(q_1 + q_2 + p)$ and computing $\boldsymbol{R}_{ZZ}$ is of order $(m_1 q_1 + m_2 q_2)(q_1 + q_2 + p)^2$ FLOPs. The same sparsity pattern will be present in $\boldsymbol{R}_{ZZ}^{-1}$, which we can calculate explicitly, allowing for the evaluation of all the formulas for the ECME update, the gradient, and the Hessian. These results generalize to an arbitrary number of strictly nested levels of grouping factors (and are

incorporated this way in our software).

For fully crossed grouping factors $\boldsymbol{T}_{ZZ12}$ and $\boldsymbol{T}_{ZZ22}$ are both dense and would need to be stored and manipulated as such. Because we can condense the calculation for the first group of random effects only, we choose the order of the grouping factors so that $m_1 q_1 \geq m_2 q_2$. Once the order of the grouping factors is established then the calculations involving $\boldsymbol{T}_{ZZ12}$ and $\boldsymbol{T}_{ZZ22}$ are essentially the same as those involving $\boldsymbol{T}_{ZX}$ and $\boldsymbol{T}_{XX}$ except that producing $\boldsymbol{R}_{ZZ22}$ from $\boldsymbol{T}_{ZZ22}$ also involves $\boldsymbol{\Delta}_2$.

The amount of storage required for fully crossed grouping factors is on the order of $(m_1 q_1 + m_2 q_2)(q_1 + m_2 q_2 + p)$ locations and producing $\boldsymbol{R}_{ZZ}$ requires on the order of $(m_1 q_1 + m_2 q_2)(q_1 + m_2 q_2 + p)^2$ FLOPs, which, obviously, can be considerably more than $(m_1 q_1 + m_2 q_2)(q_1 + q_2 + p)^2$. However, fully crossed grouping factors result in $n \geq m_1 m_2$ and usually come from designed experiments so we do not expect $m_2 q_2$ to be extremely large.

The most interesting case is partially crossed grouping factors where $\boldsymbol{T}_{ZZ12}$ is sparse but does not obey that pattern that there is only one non-zero $q_1 \times q_2$ block in each of the m_1 sets of q_1 adjacent rows. Having more than one non-zero block in such a set of rows (say because a student attended more than one school during the course of the study) does not generate fill-in in $\boldsymbol{T}_{ZZ12}$ but does generate fill-in in $\boldsymbol{T}_{ZZ22}$ and in $\boldsymbol{R}_{ZZ22}$.

If the extent of the crossing is moderate (i.e. an individual student may attend more than one school during the study but no student attends a large proportion of all the schools in the study, so that most of the $q_1 \times q_2$ blocks in any one of the m_1 sets of q_1 adjacent rows are zero), then it will be advantageous to use sparse matrix representations of $\boldsymbol{T}_{ZZ12}$ (and $\boldsymbol{R}_{ZZ12}$) and to generate $\boldsymbol{R}_{ZZ22}$ using methods for the Cholesky decomposition of sparse semi-definite matrices. This particular calculation has been extensively studied because it is important in the implementation of interior-point methods in mathematical programming (Wright (1997, pp. 253–254) [6]).

A critical part of algorithms for the Cholesky decomposition of sparse semi-definite matrices is choosing an ordering of the columns of $\boldsymbol{Z}_2$ so as to minimize fill-in in $\boldsymbol{R}_{ZZ22}$. This only needs to be done once and can be based on the pattern of crossing of the grouping factors.

The amount of storage and computation required to generate $\boldsymbol{R}_{ZZ}$ for partially crossed grouping factors will fall between that for strictly nested grouping factors and that for fully crossed grouping factors. It is unlikely that $\boldsymbol{R}_{ZZ}$ can be inverted in place when the grouping factors are partially crossed. We expect that it will be feasible to evaluate the objective function, the ECME update, and the gradient but that it may not be feasible to evaluate the Hessian.

7 Implementation

We have implemented the computational methods for evaluating the profiled objective function (either log-likelihood or log-restricted-likelihood), the ECME increment and the gradient in the `lme4` package for R [7] (www.r-project.org). Some comparisons of the speed and stability of this implementation versus our previous implementation in the `nlme` package for R are available in Debroy (2003) [8].

Future versions of this software will allow for crossed random effects and will provide the Hessian, at least for the completely crossed and strictly nested cases.

8 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. We thank Deepayan Sarkar for helpful discussions.

A Differentiating the penalized RSS

Using (21) and the relationships

$$\mathsf{D}\,\boldsymbol{\Phi} = \begin{bmatrix}\mathbf{0}\\ \boldsymbol{I}\end{bmatrix}\mathsf{D}\,\boldsymbol{\Delta}\begin{bmatrix}\boldsymbol{I}\ \mathbf{0}\end{bmatrix}, \quad \boldsymbol{\Phi}^{+}\tilde{\boldsymbol{y}} = \begin{bmatrix}\hat{\boldsymbol{b}}\\ \widehat{\boldsymbol{\beta}}\end{bmatrix}, \quad \text{and} \quad P^{\perp}\tilde{\boldsymbol{y}} = \begin{bmatrix}\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}-\boldsymbol{Z}\hat{\boldsymbol{b}}\\ -\boldsymbol{\Delta}\hat{\boldsymbol{b}}\end{bmatrix} \tag{A.1}$$

we obtain

$$\nabla r^2_{yy}(\boldsymbol{\theta}) = 2\hat{\boldsymbol{b}}^{\mathsf{T}}\boldsymbol{\Delta}^{\mathsf{T}}\mathsf{D}\,\boldsymbol{\Delta}\hat{\boldsymbol{b}} = \hat{\boldsymbol{b}}^{\mathsf{T}}\mathsf{D}\,\boldsymbol{\Omega}\hat{\boldsymbol{b}} \tag{A.2}$$

The (j,i)th element of the Hessian $\nabla^2 r^2_{yy}(\boldsymbol{\theta}) = \mathsf{D}_j\mathsf{D}_i r^2_{yy}$ is

$$\begin{aligned}
\tfrac{1}{2}\mathsf{D}_j\mathsf{D}_i r^2_{yy} &= -\mathsf{D}_j\left(\tilde{\boldsymbol{y}}^\mathsf{T} P^\perp \mathsf{D}_i \boldsymbol{\Phi}\boldsymbol{\Phi}^+\tilde{\boldsymbol{y}}\right) \\
&= -\tilde{\boldsymbol{y}}^\mathsf{T}\left(\mathsf{D}_j P^\perp \mathsf{D}_i\boldsymbol{\Phi}\boldsymbol{\Phi}^+ + P^\perp\mathsf{D}_j\mathsf{D}_i\boldsymbol{\Phi}\boldsymbol{\Phi}^+ + P^\perp\mathsf{D}_i\boldsymbol{\Phi}\mathsf{D}_j\boldsymbol{\Phi}^+\right)\tilde{\boldsymbol{y}} \\
&= \tilde{\boldsymbol{y}}^\mathsf{T}\left[P^\perp\mathsf{D}_j\boldsymbol{\Phi}\boldsymbol{\Phi}^+ + \left(\boldsymbol{\Phi}^+\right)^\mathsf{T}\mathsf{D}_j\boldsymbol{\Phi}^\mathsf{T}P^\perp\right]\mathsf{D}_i\boldsymbol{\Phi}\boldsymbol{\Phi}^+\tilde{\boldsymbol{y}} + \hat{\boldsymbol{b}}^\mathsf{T}\boldsymbol{\Delta}^\mathsf{T}\mathsf{D}_j\mathsf{D}_i\boldsymbol{\Delta}\hat{\boldsymbol{b}} \\
&\quad + \tilde{\boldsymbol{y}}^\mathsf{T}P^\perp\mathsf{D}_i\boldsymbol{\Phi}\left[\boldsymbol{\Phi}^+\mathsf{D}_j\boldsymbol{\Phi}\boldsymbol{\Phi}^+ - \left(\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\right)^{-1}\mathsf{D}_j\boldsymbol{\Phi}^\mathsf{T}P^\perp\right]\tilde{\boldsymbol{y}} \\
&= \tilde{\boldsymbol{y}}^\mathsf{T}\left(\boldsymbol{\Phi}^+\right)^\mathsf{T}\mathsf{D}_j\boldsymbol{\Phi}^\mathsf{T}\mathsf{D}_i\boldsymbol{\Phi}\boldsymbol{\Phi}^+\tilde{\boldsymbol{y}} + \hat{\boldsymbol{b}}^\mathsf{T}\boldsymbol{\Delta}^\mathsf{T}\mathsf{D}_j\mathsf{D}_i\boldsymbol{\Delta}\hat{\boldsymbol{b}} \\
&\quad - \tilde{\boldsymbol{y}}^\mathsf{T}\left(\boldsymbol{\Phi}^+\right)^\mathsf{T}\mathsf{D}_j\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\left(\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^\mathsf{T}\mathsf{D}_i\boldsymbol{\Phi}\boldsymbol{\Phi}^+\tilde{\boldsymbol{y}} \\
&\quad + \tilde{\boldsymbol{y}}P^\perp\mathsf{D}_j\boldsymbol{\Phi}\left(\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^\mathsf{T}\mathsf{D}_i\boldsymbol{\Phi}\boldsymbol{\Phi}^+\tilde{\boldsymbol{y}} \\
&\quad + \tilde{\boldsymbol{y}}P^\perp\mathsf{D}_i\boldsymbol{\Phi}\left(\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^\mathsf{T}\mathsf{D}_j\boldsymbol{\Phi}\boldsymbol{\Phi}^+\tilde{\boldsymbol{y}} \\
&\quad - \tilde{\boldsymbol{y}}P^\perp\mathsf{D}_i\boldsymbol{\Phi}\left(\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\right)^{-1}\mathsf{D}_j\boldsymbol{\Phi}^\mathsf{T}P^\perp\tilde{\boldsymbol{y}}
\end{aligned} \tag{A.3}$$

Writing

$$\boldsymbol{w}_i = \mathsf{D}_i\boldsymbol{\Phi}\boldsymbol{\Phi}^+\tilde{\boldsymbol{y}} = \begin{bmatrix}\mathbf{0}\\ \mathsf{D}_i\boldsymbol{\Delta}\hat{\boldsymbol{b}}\end{bmatrix}, \tag{A.4}$$

$$\boldsymbol{u}_i = \left(\boldsymbol{R}^{-1}\right)^\mathsf{T}\boldsymbol{\Phi}^\mathsf{T}\mathsf{D}_i\boldsymbol{\Phi}\boldsymbol{\Phi}^+\tilde{\boldsymbol{y}} = \left(\boldsymbol{R}^{-1}\right)^\mathsf{T}\begin{bmatrix}\boldsymbol{\Delta}^\mathsf{T}\mathsf{D}_i\boldsymbol{\Delta}\hat{\boldsymbol{b}}\\ \mathbf{0}\end{bmatrix}, \tag{A.5}$$

and

$$\boldsymbol{v}_i = -\left(\boldsymbol{R}^{-1}\right)^\mathsf{T}\mathsf{D}_i\boldsymbol{\Phi}^\mathsf{T}P^\perp\tilde{\boldsymbol{y}} = \left(\boldsymbol{R}^{-1}\right)^\mathsf{T}\begin{bmatrix}\mathsf{D}_i\boldsymbol{\Delta}^\mathsf{T}\boldsymbol{\Delta}\hat{\boldsymbol{b}}\\ \mathbf{0}\end{bmatrix}, \tag{A.6}$$

where $\boldsymbol{R}$ is the Cholesky decomposition of $\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi} = \boldsymbol{R}^\mathsf{T}\boldsymbol{R}$, expression (A.3) becomes

$$\tfrac{1}{2}\mathsf{D}_j\mathsf{D}_i r^2_{yy}(\boldsymbol{\theta}) = \hat{\boldsymbol{b}}^\mathsf{T}\boldsymbol{\Delta}^\mathsf{T}\mathsf{D}_j\mathsf{D}_i\boldsymbol{\Delta}\hat{\boldsymbol{b}} + \boldsymbol{w}_j^\mathsf{T}\boldsymbol{w}_i - (\boldsymbol{u}_j+\boldsymbol{v}_j)^\mathsf{T}(\boldsymbol{u}_j+\boldsymbol{v}_j) \tag{A.7}$$

Noting that $\boldsymbol{u}_i + \boldsymbol{v}_i = (\boldsymbol{R}^{-1})^\mathsf{T}\begin{bmatrix}\mathsf{D}_i\boldsymbol{\Omega}\hat{\boldsymbol{b}}\\ \mathbf{0}\end{bmatrix}$, we have

$$\mathsf{D}_j\mathsf{D}_i r^2_{yy}(\boldsymbol{\theta}) = \hat{\boldsymbol{b}}^\mathsf{T}\left[\mathsf{D}_j\mathsf{D}_i\boldsymbol{\Omega} - \mathsf{D}_j\boldsymbol{\Omega}\boldsymbol{V}_{\boldsymbol{b}}\mathsf{D}_i\boldsymbol{\Omega} - \mathsf{D}_i\boldsymbol{\Omega}\boldsymbol{V}_{\boldsymbol{b}}\mathsf{D}_j\boldsymbol{\Omega}\right]\hat{\boldsymbol{b}}. \tag{A.8}$$

References

[1] C. Henderson, Applications of Linear Models in Animal Breeding, University of Guelph, 1984.

[2] G. H. Golub, V. Pereyra, The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate, SIAM J. of Numerical Analysis 10 (2) (1973) 413–432.

[3] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. of the Royal Statistical Society, Series B 39 (1977) 1–22.

[4] C. Liu, D. B. Rubin, The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence, Biometrika 81 (1994) 633–648.

[5] S. W. Raudenbush, A. S. Bryk, Hierarchical Linear Models: Applications and Data Analysis Methods, 2nd Edition, Sage, 2002.

[6] S. J. Wright, Primal-Dual Interior Point Methods, SIAM, 1997.

[7] R. Ihaka, R. Gentleman, R: A language for data analysis and graphics, Journal of Computational and Graphical Statistics 5 (3) (1996) 299–314.

[8] S. DebRoy, Computational methods for mixed-effects models, Ph.D. thesis, U. of Wisconsin – Madison (2003).