Newton–Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data

MARY J. LINDSTROM and DOUGLAS M. BATES*

We develop an efficient and effective implementation of the Newton–Raphson (NR) algorithm for estimating the parameters in mixed-effects models for repeated-measures data. We formulate the derivatives for both maximum likelihood and restricted maximum likelihood estimation and propose improvements to the algorithm discussed by Jennrich and Schluchter (1986) to speed convergence and ensure a positive-definite covariance matrix for the random effects at each iteration. We use matrix decompositions to develop efficient and computationally stable implementations of both the NR algorithm and an EM algorithm (Laird and Ware 1982) for this model. We compare the two methods (EM vs. NR) in terms of computational order and performance on two sample data sets and conclude that in most situations a well-implemented NR algorithm is preferable to the EM algorithm or EM algorithm with Aitken's acceleration. The term *repeated measures* refers to experimental designs where there are several individuals and several measurements taken on each individual. In the mixed-effects model each individual's vector of responses is modeled as a parametric function, where some of the parameters or "effects" are random variables with a multivariate normal distribution. This model has been successful because it can handle unbalanced data (different designs for different individuals), missing data (observations on all individuals are taken at the same design points, but some individuals have missing data), and jointly dependent random effects. The price for this flexibility is that the parameter estimates may be difficult to compute. We propose some new methods for implementing the EM and NR algorithms and draw conclusions about their performance. We also discuss extensions of the mixed-effects model to incoporate nonindependent conditional error structure and nested-type designs.

KEY WORDS: Growth curve; Longitudinal data; Random effects.

1. INTRODUCTION

We use a modification of the mixed-effects model described by Laird and Ware (1982) for repeated-measures data to express the observation vector, $\mathbf{y}_i$, of length n_i for individual i as

$$\mathbf{y}_i \mid \mathbf{b}_i \sim N(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i, \sigma^2\boldsymbol{\Lambda}_i), \qquad i = 1, \ldots, M,$$

where $\mathbf{y}_i$ is independent of $\mathbf{y}_j$ for all $i \neq j$, $\boldsymbol{\Lambda}_i$ does not depend on i except that its size must be $n_i \times n_i$ (it is often assumed that $\boldsymbol{\Lambda}_i = \mathbf{I}$), $\boldsymbol{\beta}$ is a vector of p fixed population parameters, and $\mathbf{b}_i$ is a vector of q random effects associated with individual i. Furthermore, the distribution of these random effects is assumed to be $N(\mathbf{0}, \sigma^2\mathbf{D})$. It follows that, marginally, the $\mathbf{y}_i$ are independent multivariate normal vectors with mean $\mathbf{X}_i\boldsymbol{\beta}$ and covariance matrix $\boldsymbol{\Sigma}_i = \sigma^2(\boldsymbol{\Lambda}_i + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T)$. We can write the combined model for all of the data in a matrix form by letting

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_M \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_M \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_M \end{bmatrix},$$

$$\boldsymbol{\Sigma} = \operatorname{diag}(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \ldots, \boldsymbol{\Sigma}_M),$$

$\tilde{\mathbf{D}} = \operatorname{diag}(\mathbf{D}, \mathbf{D}, \ldots, \mathbf{D})$, $\mathbf{Z} = \operatorname{diag}(\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_M)$, and $\boldsymbol{\Lambda} = \operatorname{diag}(\boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2, \ldots, \boldsymbol{\Lambda}_M)$. Thus the model for the entire observation vector is $\mathbf{y} \mid \mathbf{b} \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}, \sigma^2\boldsymbol{\Lambda})$, where $\mathbf{b} \sim N(\mathbf{0}, \sigma^2\tilde{\mathbf{D}})$ and the marginal distribution of $\mathbf{y}$ is

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}), \qquad \boldsymbol{\Sigma} = \sigma^2(\boldsymbol{\Lambda} + \mathbf{Z}\tilde{\mathbf{D}}\mathbf{Z}^T). \quad (1.1)$$

When $\boldsymbol{\Lambda}$ and $\mathbf{D}$ are known, the standard estimators for $\boldsymbol{\beta}$ and $\mathbf{b}$ are the generalized least squares estimator

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} \quad (1.2)$$

and the posterior mean $\hat{\mathbf{b}}(\boldsymbol{\theta}) = \tilde{\mathbf{D}}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}))$, where $\mathbf{V} = \boldsymbol{\Lambda} + \mathbf{Z}\mathbf{D}\mathbf{Z}^T$. The vector $\boldsymbol{\theta}$ contains the unique elements of $\mathbf{D}$ and the parameters in $\boldsymbol{\Lambda}$.

We will consider both maximum likelihood (ML) and restricted maximum likelihood (RML) estimators for the variance components σ and $\boldsymbol{\theta}$. The ML estimators for σ and $\boldsymbol{\theta}$ are obtained by maximizing the log-likelihood corresponding to the marginal density of $\mathbf{y}$ [Eq. (1.1)] for $\boldsymbol{\beta}$, σ, and $\boldsymbol{\theta}$. We denote the nonconstant part of this "full" log-likelihood as l_F, where

$$l_F(\boldsymbol{\beta}, \sigma, \boldsymbol{\theta} \mid \mathbf{y}) = -\tfrac{1}{2}\log|\sigma^2\mathbf{V}| - \tfrac{1}{2}\sigma^{-2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \quad (1.3)$$

A criticism of the ML estimators for the variance components is that they are biased downward because they do not take into account the loss in degrees of freedom from the estimation of $\boldsymbol{\beta}$. The RML method corrects for this by defining estimates of the variance components as the maximizers of the log-likelihood based on $N - p$ linearly independent error contrasts, where N is the total number of observations from all individuals ($N = \sum_{i=1}^{M} n_i$). This log-likelihood, derived by Harville (1974), is

$$l_R(\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \sigma, \boldsymbol{\theta} \mid \mathbf{y}) = -\tfrac{1}{2}\log|\sigma^{-2}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}| + l_F(\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \sigma, \boldsymbol{\theta} \mid \mathbf{y}), \quad (1.4)$$

where $\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ is defined in Eq. (1.2).

In this article we discuss computational procedures for

* Mary J. Lindstrom is Statistician, Biostatistics Center, and Douglas M. Bates is Associate Professor, Department of Statistics, both at the University of Wisconsin, Madison, WI 53706. This research was supported by National Institutes of Health Grant CA 18332-13 and the Wisconsin Alumni Research Foundation. The authors are grateful for the helpful comments of a referee.

Journal of the American Statistical Association
December 1988, Vol. 83, No. 404, Computational Statistics

obtaining values for the estimators described previously. No restrictions are placed on the design matrices $\mathbf{X}$ and $\mathbf{Z}$. In particular, unbalanced designs and missing data are handled automatically. We concentrate on the case in which the conditional errors are assumed to be iid ($\mathbf{\Lambda} = \mathbf{I}$) and $\sigma^2\mathbf{D}$ is a general covariance matrix. In Section 7 we briefly discuss models for serially correlated conditional errors and models where some of the random effects are assumed to be independent of others (e.g., grouped data) and/or have values that vary with an experimental unit other than individual (e.g., nested designs).

In Section 2 we give an overview of our computational methods, and in Section 3 we present the derivatives necessary for the Newton–Raphson (NR) algorithm. We describe an implementation of the NR algorithm using matrix decompositions in Section 4, and in Section 5 we use these same decompositions to implement the EM algorithm in an efficient and computationally simple way. In Section 6 we present two examples comparing the NR and EM algorithms, and in Section 7 we discuss our conclusions and some extensions.

2. COMPUTATIONAL METHODS

In all but the simplest cases, iterative methods must be used to find estimates for the parameters in mixed-effects models for repeated-measures data (Ware 1985). An EM algorithm for both ML and RML estimation of the variance components was described by Laird and Ware (1982) and implemented by Laird, Lange, and Stram (1987), who used Aitken's acceleration (Gerald 1970) to improve the speed of convergence. Jennrich and Schluchter (1986) discussed the NR algorithm and the EM algorithm for ML estimation for a more general model that includes the mixed model as a special case.

In this article we present the necessary derivative formulas for implementing the NR algorithm for both ML and RML estimation. We include four improvements to the algorithm proposed by Jennrich and Schluchter (1986). First, we simplify the algorithm by removing σ from the computations. This reduces the number of iterations required and improves the overall convergence behavior. Second, we take advantage of the conditional linearity of $\boldsymbol{\beta}$ by replacing the value of $\boldsymbol{\beta}$ at the ωth iteration, $\boldsymbol{\beta}^{(\omega)}$, with the generalized least squares estimate, $\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}^{(\omega)})$ [Eq. (1.2)]. This will speed the convergence slightly (Bates and Lindstrom 1986) and facilitate the computation of the derivatives for the next iteration. Third, we make use of the matrix decompositions described in Section 4 to provide stable and efficient formulas for calculating the necessary derivatives. In general, such decompositions are recommended (Chambers 1977; Dongarra, Bunch, Moler, and Stewart 1979; Kennedy and Gentle 1980; Stewart 1973) because they are much less susceptible to round-off errors, they use storage efficiently, and they speed execution by reducing the order of the computations. [We implement the EM algorithm of Laird et al. (1987), using these matrix decompositions in Sec. 5.] Fourth, we optimize the log-likelihood as a function of the nonzero entries of $\mathbf{L}$, the Cholesky factor of $\mathbf{D}$ (i.e., $\mathbf{L}^T\mathbf{L} = \mathbf{D}$ and $\mathbf{L}$ is upper triangular) rather than $\mathbf{D}$. This transforms the constrained optimization problem to an unconstrained problem and ensures positive-definite estimates for $\mathbf{D}$. This will dramatically improve the convergence properties of the algorithm.

We use the standard method of inflating the diagonal elements to handle a nonpositive-definite Hessian during the iteration process. We also use step halving to ensure an increase in the log-likelihood at each iteration (Kennedy and Gentle 1980).

The formulas given in this article for the NR and EM algorithms were implemented in FORTRAN 77 on a Vax 11/750. EM test results were verified by comparison with a wholly independent FORTRAN implementation of the EM algorithm for RML estimation (provided by D. O. Stram, Dept. of Biostatistics, Harvard University). NR results were compared with a MATLAB (Moler 1981) implementation of all formulas for a specific example. Starting values were calculated using the formulas given by Laird et al. (1987) and implemented using the matrix decompositions described in Section 4. Timing results and other comparisons of the algorithms are given in Sections 6 and 7.

3. DERIVATIVES OF THE LOG–LIKELIHOOD FOR THE NEWTON–RAPHSON ALGORITHM

If $\boldsymbol{\beta}$, σ, and $\boldsymbol{\theta}$ maximize l_F, then $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}(\boldsymbol{\theta})$. Also, the RML estimates of σ and $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\beta}}$ evaluated at the RML estimate of $\boldsymbol{\theta}$ may be calculated by redefining l_R as a function of $\boldsymbol{\beta}$ rather than $\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ and maximizing it for $\boldsymbol{\beta}$, σ, and $\boldsymbol{\theta}$. We rewrite the expressions for l_F and l_R (incorporating this change in l_R) using a modification of the general formulation of Jennrich and Schluchter (1986), where $\mathbf{V}$ is an arbitrary function of the parameter vector $\boldsymbol{\theta}$. That is,

$$l_F(\boldsymbol{\beta}, \sigma, \boldsymbol{\theta} \mid \mathbf{y}) = -\frac{1}{2}\sum_{i=1}^{M} \log|\sigma^2\mathbf{V}_i(\boldsymbol{\theta})| - \frac{1}{2}\sigma^2 \sum_{i=1}^{M} \mathbf{r}_i^T\mathbf{V}_i^{-1}(\boldsymbol{\theta})\mathbf{r}_i \quad (3.1)$$

and

$$l_R(\boldsymbol{\beta}, \sigma, \boldsymbol{\theta} \mid \mathbf{y}) = -\frac{1}{2}\log|\sigma^2 \sum_{i=1}^{M} \mathbf{X}_i^T\mathbf{V}_i^{-1}(\boldsymbol{\theta})\mathbf{X}_i| + l_F(\boldsymbol{\beta}, \sigma, \boldsymbol{\theta} \mid \mathbf{y}), \quad (3.2)$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are disjoint and $\mathbf{r}_i = \mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}$. In the mixed-effects model of Section 1 $\mathbf{V}(\boldsymbol{\theta}) = (\mathbf{\Lambda} + \mathbf{ZDZ}^T)$, where $\boldsymbol{\theta}$ contains the unique elements of $\mathbf{D}$ and the parameters in $\mathbf{\Lambda}$. We use the notation $\mathbf{V}(\boldsymbol{\theta})$ to emphasize the dependence of $\mathbf{V}$ on $\boldsymbol{\theta}$. To save space we will often drop the argument and simply write $\mathbf{V}$.

To simplify the computations we solve for the ML estimate (and RML estimate) of σ^2 as a function of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. That is,

$$\hat{\sigma}^2_{\mathrm{ML}}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{M} \mathbf{r}_i^T\mathbf{V}_i(\boldsymbol{\theta})^{-1}\mathbf{r}_i$$

and

$$\hat{\sigma}^2_{\mathrm{RML}}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \frac{1}{N-p}\sum_{i=1}^{M} \mathbf{r}_i^T \mathbf{V}_i^{-1}(\boldsymbol{\theta})\mathbf{r}_i.$$

We substitute these expressions into Equations (3.1) and (3.2) to obtain the profile log-likelihoods of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. These are

$$p_F(\boldsymbol{\beta}, \boldsymbol{\theta} \mid \mathbf{y}) = -\frac{1}{2}\sum_{i=1}^{M} \log|\mathbf{V}_i(\boldsymbol{\theta})| - \frac{N}{2}\log\left[\sum_{i=1}^{M} \mathbf{r}_i^T \mathbf{V}_i^{-1}(\boldsymbol{\theta})\mathbf{r}_i\right] \quad (3.3)$$

and

$$p_R(\boldsymbol{\beta}, \boldsymbol{\theta} \mid \mathbf{y}) = -\frac{1}{2}\log\left|\sum_{i=1}^{M} \mathbf{X}_i^T \mathbf{V}_i^{-1}(\boldsymbol{\theta})\mathbf{X}_i\right| - \frac{1}{2}\sum_{i=1}^{M} \log|\mathbf{V}_i(\boldsymbol{\theta})| - \frac{1}{2}(N-p)\log\left[\sum_{i=1}^{M} \mathbf{r}_i^T \mathbf{V}_i^{-1}(\boldsymbol{\theta})\mathbf{r}_i\right]. \quad (3.4)$$

The NR optimization may be done using either the original log-likelihoods [Eqs. (3.1) and (3.2)] or these profile log-likelihoods. We recommend optimizing the profile log-likelihood, since it will usually require fewer iterations, the derivatives are somewhat simpler, and the convergence is more consistent. We have also encountered examples where the NR algorithm failed to converge when optimizing the likelihood including σ but was able to optimize the profile likelihood with ease. The inverse Hessian of the profile log-likelihood at convergence is an estimate for the marginal variance–covariance matrix for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ and can be used to find approximate confidence intervals for $\boldsymbol{\beta}$. If the Hessian of the original log-likelihood is desired, then it can be calculated after convergence is obtained for the profile log-likelihood. The additional derivatives required are straightforward.

Following the style of Jennrich and Schluchter (1986) we first give formulas for the derivatives of the profile log-likelihoods for the general formulation [Eqs. (3.3) and (3.4)] with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. We then calculate the derivatives for the mixed-effects model with independent conditional errors, that is, when $\boldsymbol{\Sigma}_i = \sigma^2\mathbf{V}_i(\boldsymbol{\theta}) = \sigma^2(\mathbf{I} + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T)$ and $\boldsymbol{\theta} = \mathrm{vec}(\mathbf{D})$ ($\mathbf{D}$ is not assumed to be symmetric at this point). Finally, we use the chain rule to find the derivatives for the mixed-effects model with respect to the transformed $\boldsymbol{\theta}$ (the upper triangular elements of L). This last step is essential. The transformation to an unconstrained problem improves the performance of the original algorithm dramatically by eliminating the possibility of optima with nonpositive-definite $\mathbf{D}$.

3.1 Derivatives With Respect to β and θ

The derivatives of p_F and p_R with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are calculated without assuming that $\mathbf{V}_i$ is symmetric, to avoid calculating the derivative of a symmetric matrix with respect to one of its elements. After differentiation we simplify the expressions by using the fact that $\mathbf{V}_i = \mathbf{V}_i^T$, but we maintain the distinction between $\partial\mathbf{V}_i/\partial\boldsymbol{\theta}$ and $\partial\mathbf{V}_i^T/\partial\boldsymbol{\theta}$ to allow the later calculation of the derivatives with respect to the entries of $\mathbf{L}$, the Cholesky factor of $\mathbf{D}$. [The derivatives of l_F with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, where $\boldsymbol{\Sigma}_i$ is assumed to be symmetric, were described by Jennrich and Schluchter (1986)].

The convention for vector derivatives we will follow is a slight variation of that of Bard (1974). That is, if x is a scalar, $\mathbf{y}$ is a $p \times 1$ vector, $\mathbf{z}$ is a $q \times 1$ vector, and $\mathbf{A}$ is a $p \times q$ matrix, then $\partial x/\partial\mathbf{y}$ and $[\partial x/\partial\mathbf{y}^T]^T$ are $p \times 1$ vectors and $\partial\mathbf{y}/\partial\mathbf{z}^T$, $\partial\mathbf{A}/\partial x$, $\partial^2 x/\partial\mathbf{z}^T\partial\mathbf{y}$, and $\partial x/\partial\mathbf{A}$ are all $p \times q$ matrices.

To conserve space we will list only the derivatives of $\mathbf{r}_i^T\mathbf{V}_i^{-1}\mathbf{r}_i$, $\log|\mathbf{V}_i|$, and $\log|\Sigma_{i=1}^M \mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i|$. The derivatives of p_F and p_R can be calculated easily from these quantities:

$$\frac{\partial \mathbf{r}_i^T\mathbf{V}_i^{-1}\mathbf{r}_i}{\partial\boldsymbol{\beta}} = -2\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{r}_i, \qquad \frac{\partial \mathbf{r}_i^T\mathbf{V}_i^{-1}\mathbf{r}_i}{\partial\boldsymbol{\theta}_j} = -\mathbf{r}_i^T\mathbf{A}_{ij}\mathbf{r}_i,$$

$$\frac{\partial^2 \mathbf{r}_i^T\mathbf{V}_i^{-1}\mathbf{r}_i}{\partial\boldsymbol{\beta}^T\,\partial\boldsymbol{\beta}} = 2\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i,$$

$$\frac{\partial^2 \mathbf{r}_i^T\mathbf{V}_i^{-1}\mathbf{r}_i}{\partial\boldsymbol{\theta}_j\,\partial\boldsymbol{\beta}} = \mathbf{X}_i^T(\mathbf{A}_{ij} + \mathbf{A}_{ij}^T)\mathbf{r}_i,$$

$$\frac{\partial^2 \mathbf{r}_i^T\mathbf{V}_i^{-1}\mathbf{r}_i}{\partial\boldsymbol{\theta}_k\,\partial\boldsymbol{\theta}_j} = -\mathbf{r}_i^T\frac{\partial\mathbf{A}_{ij}}{\partial\boldsymbol{\theta}_k}\mathbf{r}_i, \qquad \frac{\partial \log|\mathbf{V}_i|}{\partial\boldsymbol{\theta}_j} = \mathrm{tr}\left[\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\boldsymbol{\theta}_j}\right],$$

$$\frac{\partial^2 \log|\mathbf{V}_i|}{\partial\boldsymbol{\theta}_k\,\partial\boldsymbol{\theta}_j} = \mathrm{tr}\left[-\mathbf{A}_{ik}^T\frac{\partial\mathbf{V}_i}{\partial\boldsymbol{\theta}_j} + \mathbf{V}_i^{-1}\frac{\partial^2\mathbf{V}_i}{\partial\boldsymbol{\theta}_k\,\partial\boldsymbol{\theta}_j}\right],$$

$$\frac{\partial \log|\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}|}{\partial\boldsymbol{\theta}_j} = -\sum_{i=1}^{M}\mathrm{tr}[\mathbf{H}^{-1}\mathbf{X}_i^T\mathbf{A}_{ij}\mathbf{X}_i],$$

and

$$\frac{\partial^2 \log|\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}|}{\partial\boldsymbol{\theta}_k\,\partial\boldsymbol{\theta}_j} = -\mathrm{tr}\left[\mathbf{H}^{-1}\sum_{i=1}^{M}(\mathbf{X}_i^T\mathbf{A}_{ik}\mathbf{X}_i) \times \mathbf{H}^{-1}\sum_{i=1}^{M}(\mathbf{X}_i^T\mathbf{A}_{ij}\mathbf{X}_i)\right] - \mathrm{tr}\left[\mathbf{H}^{-1}\sum_{i=1}^{M}\left(\mathbf{X}_i^T\frac{\partial\mathbf{A}_{ij}}{\partial\boldsymbol{\theta}_k}\mathbf{X}_i\right)\right],$$

where

$$\mathbf{A}_{ij} = \mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\boldsymbol{\theta}_j}\mathbf{V}_i^{-1},$$

$$\frac{\partial\mathbf{A}_{ij}}{\partial\boldsymbol{\theta}_k} = -\mathbf{V}_i^{-1}\left(\frac{\partial\mathbf{V}_i}{\partial\boldsymbol{\theta}_k}\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\boldsymbol{\theta}_j} - \frac{\partial^2\mathbf{V}_i}{\partial\boldsymbol{\theta}_k\,\partial\boldsymbol{\theta}_j} + \frac{\partial\mathbf{V}_i}{\partial\boldsymbol{\theta}_j}\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\boldsymbol{\theta}_k}\right)\mathbf{V}_i^{-1},$$

and

$$\mathbf{H} = \sum_{i=1}^{M}\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i.$$

3.2 Derivatives With Respect to β and vec(D)

If $\mathbf{D}$ is not assumed to be symmetric, then $\partial\mathbf{V}_i/\partial\mathbf{D}_{jk} = \mathbf{Z}_i^{(j)}\mathbf{Z}_i^{(k)T}$, where $\mathbf{Z}_i^{(i)}$ denotes the ith column of $\mathbf{Z}_i$. From the equations in Section 3.1, with $\boldsymbol{\theta} = \mathrm{vec}(\mathbf{D})$ and $\boldsymbol{\beta} =$

$\hat{\boldsymbol{\beta}}(\mathbf{D})$, it follows that

$$\frac{\partial \mathbf{r}_i^T \mathbf{V}_i^{-1} \mathbf{r}_i}{\partial \boldsymbol{\beta}} = \mathbf{0}, \qquad \frac{\partial^2 \mathbf{r}_i^T \mathbf{V}_i^{-1} \mathbf{r}_i}{\partial \boldsymbol{\beta}^T \partial \boldsymbol{\beta}} = 2\mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i,$$

$$\frac{\partial \mathbf{r}_i^T \mathbf{V}_i^{-1} \mathbf{r}_i}{\partial \operatorname{vec}(\mathbf{D})} = -\operatorname{vec}(\mathbf{v}_i \mathbf{v}_i^T),$$

$$\frac{\partial^2 \mathbf{r}_i^T \mathbf{V}_i^{-1} \mathbf{r}_i}{\partial \operatorname{vec}(\mathbf{D})^T \partial \boldsymbol{\beta}} = \mathbf{v}_i^T \otimes \mathbf{C}_i^T + \mathbf{C}_i^T \otimes \mathbf{v}_i^T,$$

$$\frac{\partial^2 \mathbf{r}_i^T \mathbf{V}_i^{-1} \mathbf{r}_i}{\partial \operatorname{vec}(\mathbf{D}^T)^T \partial \operatorname{vec}(\mathbf{D})} = \mathbf{B}_i + \mathbf{B}_i^T,$$

$$\frac{\partial \log|\mathbf{V}_i|}{\partial \operatorname{vec}(\mathbf{D})} = \operatorname{vec}(\mathbf{Z}_i^T \mathbf{V}_i^{-1} \mathbf{Z}_i),$$

$$\frac{\partial^2 \log|\mathbf{V}_i|}{\partial \operatorname{vec}(\mathbf{D}^T)^T \partial \operatorname{vec}(\mathbf{D})} = -\mathbf{Z}_i^T \mathbf{V}_i^{-1} \mathbf{Z}_i \otimes \mathbf{Z}_i^T \mathbf{V}_i^{-1} \mathbf{Z}_i,$$

$$\frac{\partial \log|\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X}|}{\partial \operatorname{vec}(\mathbf{D}^T)} = -\operatorname{vec}\left(\sum_{i=1}^{M} \mathbf{C}_i \mathbf{H}^{-1} \mathbf{C}_i^T\right),$$

and

$$\begin{aligned}
&\frac{\partial^2 \log|\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X}|}{\partial \operatorname{vec}(\mathbf{D}^T)^T \partial \operatorname{vec}(\mathbf{D})} \\
&\quad = -\left[\sum_{i=1}^{M} (\mathbf{C}_i \mathbf{H}^{(-1/2)T} \otimes \mathbf{C}_i \mathbf{H}^{(-1/2)T})\right] \\
&\qquad \times \left[\sum_{i=1}^{M} (\mathbf{H}^{-1/2} \mathbf{C}_i^T) \otimes (\mathbf{H}^{-1/2} \mathbf{C}_i^T)\right] \\
&\qquad + \sum_{i=1}^{M} [\mathbf{C}_i \mathbf{H}^{-1} \mathbf{C}_i^T \otimes \mathbf{Z}_i^T \mathbf{V}_i^{-1} \mathbf{Z}_i \\
&\qquad + \mathbf{Z}_i^T \mathbf{V}_i^{-1} \mathbf{Z}_i \otimes \mathbf{C}_i \mathbf{H}^{-1} \mathbf{C}_i^T],
\end{aligned}$$

where $\mathbf{v}_i = \mathbf{Z}_i^T \mathbf{V}_i^{-1} \mathbf{r}_i$, $\mathbf{B}_i = \mathbf{Z}_i^T \mathbf{V}_i^{-1} \mathbf{Z}_i \otimes \mathbf{v}_i \mathbf{v}_i^T$, $\mathbf{C}_i = \mathbf{Z}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i$, $\mathbf{H} = \mathbf{X}^T \mathbf{V}^{-1} \mathbf{X}$, and $\otimes$ denotes the Kronecker product. To find derivatives with respect to $\operatorname{vec}(\mathbf{D})^T$ $\operatorname{vec}(\mathbf{D})$, permute the columns of the derivatives with respect to $\operatorname{vec}(\mathbf{D}^T)^T$ $\operatorname{vec}(\mathbf{D})$ appropriately.

3.3 Derivatives With Respect to vec(L)

To find derivatives with respect to $\mathbf{L}$ we calculate the derivatives of $\mathbf{D}$ with respect to $\mathbf{L}$ and then apply the chain rule. If $g(\mathbf{D})$ is a scalar- or vector-valued function of $\mathbf{D}$, then

$$\frac{\partial g(\mathbf{D})}{\partial \operatorname{vec}(\mathbf{L})} = \tilde{\mathbf{L}}\left(\frac{\partial g(\mathbf{D})}{\partial \operatorname{vec}(\mathbf{D}^T)} + \frac{\partial g(\mathbf{D})}{\partial \operatorname{vec}(\mathbf{D})}\right),$$

where $\tilde{\mathbf{L}}$ is defined to be the $q^2 \times q^2$ matrix $\operatorname{diag}(\mathbf{L}, \mathbf{L}, \ldots, \mathbf{L})$. Also, if $f(\mathbf{D})$ is a scalar-valued function of $\mathbf{D}$, then

$$\begin{aligned}
\frac{\partial^2 f(\mathbf{D})}{\partial (\mathbf{L}^{(k)})^T \partial \mathbf{L}^{(j)}} &= \left(\frac{\partial f(\mathbf{D})}{\partial \mathbf{D}_{jk}} + \frac{\partial f(\mathbf{D})}{\partial \mathbf{D}_{kj}}\right) \mathbf{I} \\
&\quad + 2\mathbf{L}\left(\frac{\partial^2 f(\mathbf{D})}{\partial \mathbf{D}^{[k]} \partial \mathbf{D}^{(j)}} + \frac{\partial^2 f(\mathbf{D})}{\partial (\mathbf{D}^{(k)})^T \partial \mathbf{D}^{(j)}}\right) \mathbf{L}^T,
\end{aligned}$$

where $\mathbf{D}^{[k]}$ is the kth row of $\mathbf{D}$.

The gradient and Hessian are created by first calculating the gradient and Hessian for the entire parameter vector $(\boldsymbol{\beta}^T, \operatorname{vec}(\mathbf{L})^T)^T$ and then deleting the entries that correspond to the entries in $\mathbf{L}$ that are 0 by definition.

4. MATRIX DECOMPOSITIONS FOR THE NEWTON–RAPHSON ALGORITHM

Under the assumption that the variance–covariance of $\mathbf{y}_i$ given $\mathbf{b}_i$ is $\sigma^2 \mathbf{I}$, we can substantially reduce the complexity of calculations and the amount of data that must be stored by first reducing the matrices $\mathbf{X}_i$ and $\mathbf{Z}_i$ to an orthogonal-triangular (or QR) form. We form

$$[\mathbf{Z}_i, \mathbf{X}_i] = \mathbf{Q}_i \mathbf{R}_i = [\mathbf{Q}_{i(1)}, \mathbf{Q}_{i(2)}, \mathbf{Q}_{i(3)}] \begin{bmatrix} \mathbf{R}_{i(11)} & \mathbf{R}_{i(12)} \\ \mathbf{0} & \mathbf{R}_{i(22)} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \tag{4.1}$$

where $\mathbf{Q}$ is orthogonal and $\mathbf{R}_{i(11)}$ and $\mathbf{R}_{i(22)}$ are $q \times q$ and $p \times p$ upper triangular matrices. The matrix $\mathbf{R}_{i(12)}$ is $q \times p$.

Applying $\mathbf{Q}_i^T$ to the data vector $\mathbf{y}_i$ produces $\mathbf{c}_i = \mathbf{Q}_i^T \mathbf{y}_i = (\mathbf{y}_i^T \mathbf{Q}_{i(1)}, \mathbf{y}_i^T \mathbf{Q}_{i(2)}, \mathbf{y}_i^T \mathbf{Q}_{i(3)})^T = (\mathbf{c}_{i(1)}^T, \mathbf{c}_{i(2)}^T, \mathbf{c}_{i(3)}^T)^T$, with $\mathbf{c}_{i(1)} \mid \mathbf{b}_i \sim N(\mathbf{R}_{i(12)}\boldsymbol{\beta} + \mathbf{R}_{i(11)}\mathbf{b}_i, \sigma^2 \mathbf{I})$, $\mathbf{c}_{i(2)} \sim N(\mathbf{R}_{i(22)}\boldsymbol{\beta}, \sigma^2 \mathbf{I})$, and $\mathbf{c}_{i(3)} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$ for $i = 1, \ldots, M$. Since $\mathbf{c}_{i(2)}$ and $\mathbf{c}_{i(3)}$ do not depend on $\mathbf{b}_i$, the dimensions of the vectors in the log-likelihood expressions can be reduced. We can further reduce dimensions by decomposing

$$\mathbf{R}_{(22)} = \begin{bmatrix} \mathbf{R}_{1(22)} \\ \mathbf{R}_{2(22)} \\ \cdot \\ \cdot \\ \mathbf{R}_{M(22)} \end{bmatrix}_{[Mp,p]} = \tilde{\mathbf{Q}}\tilde{\mathbf{R}} = \tilde{\mathbf{Q}}_{(1)}\tilde{\mathbf{R}}_{(1)}$$

and forming

$$\mathbf{w} = \tilde{\mathbf{Q}}^T \mathbf{c}_{(2)} = \begin{bmatrix} \tilde{\mathbf{Q}}_{(1)}^T \mathbf{c}_{(2)} \\ \tilde{\mathbf{Q}}_{(2)}^T \mathbf{c}_{(2)} \end{bmatrix} = \begin{bmatrix} \mathbf{w}_{(1)} \\ \mathbf{w}_{(2)} \end{bmatrix},$$

where $\mathbf{c}_{(2)} = [\mathbf{c}_{1(2)}^T, \mathbf{c}_{2(2)}^T, \ldots, \mathbf{c}_{M(2)}^T]^T$, $\mathbf{w}_1 \sim N(\tilde{\mathbf{R}}_{(1)}\boldsymbol{\beta}, \sigma^2 \mathbf{I})$, and $\mathbf{w}_2 \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$. To aid readers we indicate the dimensions of some of the matrices in these formulas by a pair of subscripts enclosed in brackets (or a single subscript if the matrix is square).

Because the distributions of $\mathbf{w}_2$ and $\mathbf{c}_{i(3)}$ do not depend on any parameters except σ^2, we can express the likelihood using only $\mathbf{R}_{i(11)}$, $\mathbf{R}_{i(12)}$, $\tilde{\mathbf{R}}_{(1)}$, $\mathbf{c}_{i(1)}$, $\mathbf{w}_1$, and $k = \|\mathbf{w}_2\|^2 + \sum_{i=1}^{M} \|\mathbf{c}_{i(3)}\|^2$.

First we form $\tilde{\mathbf{V}}_i = \mathbf{I} + \mathbf{R}_{i(11)} \mathbf{D} \mathbf{R}_{i(11)}^T$ and

$$\mathbf{c} = \begin{bmatrix} \mathbf{c}_{1(1)} \\ \mathbf{c}_{2(1)} \\ \vdots \\ \mathbf{c}_{M(1)} \\ \mathbf{w}_1 \end{bmatrix}, \qquad \mathbf{R} = \begin{bmatrix} \mathbf{R}_{1(12)} \\ \mathbf{R}_{2(12)} \\ \vdots \\ \mathbf{R}_{M(12)} \\ \tilde{\mathbf{R}}_{(1)} \end{bmatrix},$$

$$\tilde{\mathbf{V}} = \operatorname{diag}(\tilde{\mathbf{V}}_1, \tilde{\mathbf{V}}_2, \ldots, \tilde{\mathbf{V}}_M, \mathbf{I}_{[p]})$$

to produce

$$\begin{aligned}
l_F(\boldsymbol{\beta}, \sigma, \mathbf{D} \mid \mathbf{y}) &= -(N/2)\log(\sigma^2) - \tfrac{1}{2}\log|\tilde{\mathbf{V}}| \\
&\quad - \tfrac{1}{2}\sigma^2[(\mathbf{c} - \mathbf{R}\boldsymbol{\beta})^T \tilde{\mathbf{V}}^{-1}(\mathbf{c} - \mathbf{R}\boldsymbol{\beta}) + k]
\end{aligned} \tag{4.2}$$

and

$$l_R(\hat{\boldsymbol{\beta}}(\mathbf{D}), \sigma, \mathbf{D} \mid \mathbf{y}) = -\tfrac{1}{2}(N - p)\log(\sigma^2) - \tfrac{1}{2} \times \log|\mathbf{R}^T\tilde{\mathbf{V}}^{-1}\mathbf{R}| - \tfrac{1}{2}\log|\tilde{\mathbf{V}}| - \tfrac{1}{2}\sigma^{-2}[(\mathbf{c} - \mathbf{R}\hat{\boldsymbol{\beta}}(\mathbf{D}))^T\tilde{\mathbf{V}}^{-1}(\mathbf{c} - \mathbf{R}\hat{\boldsymbol{\beta}}(\mathbf{D})) + k]. \quad (4.3)$$

These transformed log-likelihoods have the same form for the estimation of $\boldsymbol{\beta}$ and $\mathbf{b}$ as the original log-likelihoods [Eqs. (1.3) and (1.4)] with the substitutions $\mathbf{X} = \mathbf{R}$, $\mathbf{Z}_i = \mathbf{R}_{i(11)}$, $\mathbf{V} = \tilde{\mathbf{V}}$, and $\mathbf{y} = \mathbf{c}$. Thus the conditional parameter estimates can be written as

$$\hat{\boldsymbol{\beta}}(\mathbf{D}) = (\mathbf{R}^T\tilde{\mathbf{V}}^{-1}\mathbf{R})^{-1}\mathbf{R}\tilde{\mathbf{V}}^{-1}\mathbf{c}$$
$$\hat{\mathbf{b}}_i(\mathbf{D}) = \mathbf{D}\mathbf{R}_{i(11)}^T\tilde{\mathbf{V}}_i^{-1}(\mathbf{c}_{i(1)} - \mathbf{R}_{i(12)}\hat{\boldsymbol{\beta}}(\mathbf{D})). \quad (4.4)$$

4.1 Calculation of V, β, and b

First we derive expressions for $\mathbf{V}$, $\hat{\mathbf{b}}(\mathbf{D})$, and $\hat{\boldsymbol{\beta}}(\mathbf{D})$, then we use them to obtain efficient implementations of the EM and NR algorithms.

Assume that $\mathbf{D}$ is available from the last iteration. Then from Equation (4.1),

$$\mathbf{V}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T + \mathbf{I} = \mathbf{Q}_i \operatorname{diag}(\tilde{\mathbf{V}}_i, \mathbf{I}_{[n_i-q]})\mathbf{Q}_i^T = \mathbf{Q}_i \operatorname{diag}(\mathbf{K}_i^T\mathbf{K}_i, \mathbf{I}_{[n_i-q]})\mathbf{Q}_i^T,$$

where $\mathbf{K}_i^T = [\mathbf{R}_{i(11)}\mathbf{L}^T, \mathbf{I}_{[q]}]$ and $\mathbf{L}$ is the Cholesky factor of $\mathbf{D}$. To simplify this expression, decompose $\mathbf{K}_i$ into $\mathbf{S}_i\mathbf{T}_i = \mathbf{S}_{i(1)}\mathbf{T}_{i(1)}$ with a QR decomposition and then $\tilde{\mathbf{V}}_i = \mathbf{T}_{i(1)}^T\mathbf{T}_{i(1)}$. If we define $\mathbf{T} = \operatorname{diag}(\mathbf{T}_{1(1)}, \mathbf{T}_{2(1)}, \ldots, \mathbf{T}_{M(1)}, \mathbf{I}_{[p]})$, then $\tilde{\mathbf{V}} = \mathbf{T}^T\mathbf{T}$.

To calculate $\hat{\boldsymbol{\beta}}(\mathbf{D})$, define $\mathbf{g} = \mathbf{T}^{-T}\mathbf{c}$ and decompose $\mathbf{T}^{-T}\mathbf{R} = \mathbf{F}\mathbf{G} = \mathbf{F}_{(1)}\mathbf{G}_{(1)}$. Then $\hat{\boldsymbol{\beta}}(\mathbf{D}) = \mathbf{G}_{(1)}^{-1}\mathbf{F}_{(1)}^T\mathbf{g}$ and the residual vector, $\mathbf{u} = \mathbf{T}^{-T}(\mathbf{c} - \mathbf{R}\hat{\boldsymbol{\beta}}(\mathbf{D}))$. Using the fact that $\mathbf{R}_{i(11)}^T\mathbf{T}_{i(1)}^{-1} = \mathbf{L}^{-1}\mathbf{S}_{i(11)}$, where $\mathbf{S}_{i(11)}$ is the first q rows of $\mathbf{S}_{i(1)}$, we have

$$\hat{\mathbf{b}}_i(\mathbf{D}) = \mathbf{D}\mathbf{R}_{i(11)}^T\tilde{\mathbf{V}}_i^{-1}(\mathbf{c}_{i(1)} - \mathbf{R}_{i(12)}\hat{\boldsymbol{\beta}}(\mathbf{D})) = \mathbf{D}\mathbf{R}_{i(11)}^T\mathbf{T}_{i(1)}^{-1}\mathbf{u}_{i(1)} = \mathbf{L}^T\mathbf{S}_{i(11)}\mathbf{u}_{i(1)}, \quad (4.5)$$

where $\mathbf{u}_{i(1)}$ is the ith q rows of $\mathbf{u}_{(1)}$.

We can also express the profile log-likelihoods in terms of the previous decompositions:

$$p_F(\hat{\boldsymbol{\beta}}(\mathbf{D}), \mathbf{D} \mid \mathbf{y}) = -\sum_{i=1}^{M}\log(abs[\operatorname{tr}\mathbf{T}_{i(1)}]) - \frac{N}{2}\log(\mathbf{u}^T\mathbf{u} + k)$$

and

$$p_R(\hat{\boldsymbol{\beta}}(\mathbf{D}), \mathbf{D} \mid \mathbf{y}) = -\log(abs[\operatorname{tr}\mathbf{G}_{(1)}]) - \sum_{i=1}^{M}\log(abs[\operatorname{tr}\mathbf{T}_{i(1)}]) - \tfrac{1}{2}(N - p)\log(\mathbf{u}^T\mathbf{u} + k).$$

4.2 Computational Equations for the Newton–Raphson Derivatives

We now present efficient formulas that can be used to calculate the derivatives required for the NR algorithm. The equations are derived by applying the expressions obtained in Section 4.1 to the corresponding factors in the transformed log-likelihoods. Note that these equations depend on the fact that the estimate of $\boldsymbol{\beta}$ at the ωth iteration, $\boldsymbol{\beta}^{(\omega)}$, has been updated to the generalized least squares estimate, $\hat{\boldsymbol{\beta}}(\mathbf{D}^{(\omega)})$.

We define a division of $\mathbf{F}_{(1)}$ and $\mathbf{F}_{(2)}$ into submatrices with Mq and p rows, respectively, as

$$\mathbf{F} = \begin{bmatrix}\mathbf{F}_{(11)} & \mathbf{F}_{(21)}\\ \mathbf{F}_{(12)} & \mathbf{F}_{(22)}\end{bmatrix}$$

and let the $q \times p$ matrix $\mathbf{F}_{i(11)}$ be the ith q rows of $\mathbf{F}_{(11)}$. Then the expressions required to calculate the derivatives in Section 3.2 are

$$\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{r}_i = \mathbf{0}, \qquad \mathbf{X}^T\mathbf{V}^{-1}\mathbf{X} = \mathbf{G}_{(1)}^T\mathbf{G}_{(1)},$$
$$\mathbf{Z}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i = \mathbf{L}^{-1}\mathbf{S}_{i(11)}\mathbf{F}_{i(11)}\mathbf{G}_{(1)},$$
$$\mathbf{Z}_i^T\mathbf{V}_i^{-1}\mathbf{r}_i = \mathbf{L}^{-1}\mathbf{S}_{i(11)}\mathbf{u}_{i(1)},$$
$$\mathbf{Z}_i^T\mathbf{V}_i^{-1}\mathbf{Z}_i = \mathbf{L}^{-1}\mathbf{S}_{i(11)}\mathbf{S}_{i(11)}^T\mathbf{L}^{-T},$$

and

$$\sum_{i=1}^{M}\mathbf{r}_i^T\mathbf{V}_i^{-1}\mathbf{r}_i = \mathbf{u}^T\mathbf{u} + k.$$

5. ESTIMATION OF σ AND $\mathbf{D}$ VIA THE EM ALGORITHM

Laird and Ware (1982) presented the EM algorithm as a method of calculating both ML and RML estimates for the variance components σ and $\mathbf{D}$ using the projections $\mathbf{P}_{i\mathrm{ML}} = \mathbf{V}_i^{-1}$ and $\mathbf{P}_{i\mathrm{RML}} = \mathbf{V}_i^{-1} - \mathbf{V}_i^{-1}\mathbf{X}_i(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}_i^T\mathbf{V}_i^{-1}$.

Given the data and estimates from the last iteration, $\sigma^{(\omega)}$ and $\mathbf{D}^{(\omega)}$, the new estimates, $\sigma_{\mathrm{ML}}^{(\omega+1)}$ and $\mathbf{D}_{\mathrm{ML}}^{(\omega+1)}$ or $\sigma_{\mathrm{RML}}^{(\omega+1)}$ and $\mathbf{D}_{\mathrm{RML}}^{(\omega+1)}$, are calculated as

$$\mathbf{D}_{(\mathrm{R})\mathrm{ML}}^{(\omega+1)} = \mathbf{D}^{(\omega)} + \frac{1}{M}\sum_{i=1}^{M}[(\sigma^{(\omega)})^{-2}\hat{\mathbf{b}}_i\hat{\mathbf{b}}_i^T - \mathbf{D}^{(\omega)}\mathbf{Z}_i^T\mathbf{P}_{i(\mathrm{R})\mathrm{ML}}\mathbf{Z}_i\mathbf{D}^{(\omega)}]$$

and

$$(\sigma_{(\mathrm{R})\mathrm{ML}}^{(\omega+1)})^2 = (\sigma^{(\omega)})^2 + \frac{1}{N}\sum_{i=1}^{M}[(\mathbf{r}_i - \mathbf{Z}_i\hat{\mathbf{b}}_i)^T(\mathbf{r}_i - \mathbf{Z}_i\hat{\mathbf{b}}_i) - (\sigma^{(\omega)})^2\operatorname{tr}\mathbf{P}_{i(\mathrm{R})\mathrm{ML}}],$$

where $\hat{\mathbf{b}}_i$, $\mathbf{r}_i$, and $\mathbf{V}_i$ are evaluated at $\mathbf{D}^{(\omega)}$.

Laird et al. (1987) and Stram, Laird, and Ware (1986) discussed the computational details of implementing the EM algorithm for mixed-effects models in the longitudinal data setting. These papers pointed out that inverting the $n_i \times n_i$ matrices $\mathbf{V}_i$ $(i = 1, \ldots, N)$ can be avoided by noticing that $\mathbf{V}_i^{-1} = \mathbf{I} - \mathbf{Z}_i(\mathbf{Z}_i^T\mathbf{Z}_i + \mathbf{D}^{-1})^{-1}\mathbf{Z}_i^T$. The iterative calculations, however, still involve the evaluation of the M residual vectors $\mathbf{r}_i - \mathbf{Z}_i\hat{\mathbf{b}}_i$, a computation of order N. We propose an implementation of the EM algorithm in this setting, which uses matrix decompositions and reduces the order of each iteration to $M(q^3 + q^2p)$.

We first rewrite the EM iterative equations in terms of

the transformed log-likelihoods [Eqs. (4.2) and (4.3)]:

$$\mathbf{D}_{(\mathrm{R})\mathrm{ML}}^{(\omega+1)} = \mathbf{D}^{(\omega)} + \frac{1}{M}\sum_{i=1}^{M}[(\sigma^{(\omega)})^{-2}\hat{\mathbf{b}}_i\hat{\mathbf{b}}_i^T - \mathbf{D}^{(\omega)}\mathbf{R}_{i(11)}^T\tilde{\mathbf{P}}_{i(\mathrm{R})\mathrm{ML}}\mathbf{R}_{i(11)}\mathbf{D}^{(\omega)}] \quad (5.1)$$

and

$$(\sigma_{(\mathrm{R})\mathrm{ML}}^{(\omega+1)})^2 = (\sigma^{(\omega)})^2 + \frac{1}{N}\sum_{i=1}^{M}(\tilde{\mathbf{r}}_i - \mathbf{R}_{i(11)}\hat{\mathbf{b}}_i)^T(\tilde{\mathbf{r}}_i - \mathbf{R}_{i(11)}\hat{\mathbf{b}}_i) + \frac{1}{N}[(\mathbf{w}_1 - \tilde{\mathbf{R}}_{(1)}\hat{\boldsymbol{\beta}})^T(\mathbf{w}_1 - \tilde{\mathbf{R}}_{(1)}\hat{\boldsymbol{\beta}}) + k - (\sigma^{(\omega)})^2(\mathrm{tr}\,\tilde{\mathbf{P}}_{(\mathrm{R})\mathrm{ML}} + N - Mq)], \quad (5.2)$$

where $\tilde{\mathbf{r}}_i = \mathbf{c}_{i(1)} - \mathbf{R}_{i(12)}\hat{\boldsymbol{\beta}}$, $\tilde{\mathbf{P}}_{i\mathrm{ML}} = \tilde{\mathbf{V}}_i^{-1}$, $\tilde{\mathbf{P}}_{i\mathrm{RML}} = \tilde{\mathbf{V}}_i^{-1} - \tilde{\mathbf{V}}_i^{-1}\mathbf{R}_{i(12)}(\mathbf{R}^T\tilde{\mathbf{V}}^{-1}\mathbf{R})^{-1}\mathbf{R}_{i(12)}^T\tilde{\mathbf{V}}_i^{-1}$, $\tilde{\mathbf{P}}_{\mathrm{ML}} = \tilde{\mathbf{V}}^{-1}$, and $\tilde{\mathbf{P}}_{\mathrm{RML}} = \tilde{\mathbf{V}}^{-1} - \tilde{\mathbf{V}}^{-1}\mathbf{R}(\mathbf{R}^T\tilde{\mathbf{V}}^{-1}\mathbf{R})^{-1}\mathbf{R}^T\tilde{\mathbf{V}}^{-1}$. The quantities $\tilde{\mathbf{r}}_i$, $\hat{\mathbf{b}}_i$, $\hat{\boldsymbol{\beta}}$, and $\tilde{\mathbf{V}}_i$ are evaluated at $\mathbf{D}^{(\omega)}$. We can use the matrix decompositions of Section 4 to implement the EM iterative equations for these transformed log-likelihoods. To calculate the value of σ^2 for the next iteration, note that from the expression for $\hat{\mathbf{b}}(\mathbf{D})$ in Equation (4.4) we have $\tilde{\mathbf{r}}_i - \mathbf{R}_{i(11)}\hat{\mathbf{b}}_i(\mathbf{D}) = \tilde{\mathbf{r}}_i - (\tilde{\mathbf{V}}_i - \mathbf{I})\tilde{\mathbf{V}}_i^{-1}\tilde{\mathbf{r}}_i = \tilde{\mathbf{V}}_i^{-1}\tilde{\mathbf{r}}_i$ and $\tilde{\mathbf{r}}_i^T\tilde{\mathbf{V}}_i^{-1}\tilde{\mathbf{r}}_i = (\mathbf{u}_{i(1)}^T\mathbf{T}_{i(1)}^{-T}\mathbf{T}_{i(1)}^{-1}\mathbf{u}_{i(1)})$. In addition, $(\mathbf{w}_1 - \tilde{\mathbf{R}}_{(1)}\hat{\boldsymbol{\beta}})^T(\mathbf{w}_1 - \tilde{\mathbf{R}}_{(1)}\hat{\boldsymbol{\beta}}) = \mathbf{u}_{(2)}^T\mathbf{u}_{(2)}$. The definitions of $\tilde{\mathbf{P}}_{i\mathrm{RML}}$ and $\tilde{\mathbf{P}}_{i\mathrm{ML}}$ and the fact that $\mathbf{T}_{i(1)}\mathbf{R}_{i(12)} = \mathbf{F}_{i(11)}\mathbf{G}_{(1)}$ give us $\mathrm{tr}\,\tilde{\mathbf{P}}_{\mathrm{ML}} = \sum_{i=1}^{M}\mathrm{tr}[\mathbf{T}_i^{-1}\mathbf{T}_i^{-T}]$ and

$$\mathrm{tr}\,\tilde{\mathbf{P}}_{\mathrm{RML}} = \sum_{i=1}^{M}(\mathrm{tr}\,\mathbf{T}_{i(1)}^{-1}\mathbf{T}_{i(1)}^{-\mathrm{T}} - \mathrm{tr}\,\mathbf{T}_{i(1)}^{-1}\mathbf{F}_{i(11)}\mathbf{F}_{i(11)}^T\mathbf{T}_{i(1)}^{-T}) - \mathrm{tr}[\mathbf{F}_{(12)}\mathbf{F}_{(12)}^T].$$

These expressions, along with Equation (5.2), allow us to calculate $\sigma^{(\omega+1)}$.

To calculate $\mathbf{D}$ for the next iteration, first note that $\mathbf{L}\mathbf{R}_{i(11)}^T\tilde{\mathbf{P}}_{i\mathrm{ML}}\mathbf{R}_{i(11)}\mathbf{L}^T = \mathbf{S}_{i(11)}\mathbf{S}_{i(11)}^T$ and $\mathbf{L}\mathbf{R}_{i(11)}^T\tilde{\mathbf{P}}_{i\mathrm{RML}}\mathbf{R}_{i(11)}\mathbf{L}^T = \mathbf{S}_{i(11)}\mathbf{S}_{i(11)}^T - \mathbf{S}_{i(11)}\mathbf{F}_{i(11)}\mathbf{F}_{i(11)}^T\mathbf{S}_{i(11)}^T$. Thus, from Equations (5.1) and (4.5), we have

$$\mathbf{D}_{\mathrm{ML}}^{(\omega+1)} = \mathbf{L}^T\left[\mathbf{I}_{[q]} + \frac{1}{M}\sum_{i=1}^{M}((\sigma^{(\omega)})^{-2} \times \mathbf{S}_{i(11)}\mathbf{u}_{i(1)}\mathbf{u}_{i(1)}^T\mathbf{S}_{i(11)}^T - \mathbf{S}_{i(11)}\mathbf{S}_{i(11)}^T)\right]\mathbf{L}$$

and

$$\mathbf{D}_{\mathrm{RML}}^{(\omega+1)} = \mathbf{D}_{\mathrm{ML}}^{(\omega+1)} + \mathbf{L}^T\left[\frac{1}{M}\sum_{i=1}^{M}(\mathbf{S}_{i(11)}\mathbf{F}_{i(11)}\mathbf{F}_{i(11)}^T\mathbf{S}_{i(11)}^T)\right]\mathbf{L}.$$

6. EXAMPLES

To compare the three algorithms (NR, EM, and EM with Aitken's acceleration), we use each of them to find RML estimates in two examples. The first example is part of a study presented in Pierson and Ginther (1987). The data consist of the number of ovarian follicles greater than 10 mm in diameter recorded daily for each of 11 mares from three days before ovulation until three days after the next ovulation. The measurement times were scaled so that the ovulations for each mare occur at times 0 and 1. Since the ovulation cycles vary in length, the number of measurements and the (scaled) times at which they occurred vary among the mares. No mechanistic model was postulated by the researcher for the number of large follicles as a function of time, but the shape of the individual data plots suggested that a sinusoid model of the form

$$y_{ij} = (\boldsymbol{\beta}_1 + \mathbf{b}_{i1}) + (\boldsymbol{\beta}_2 + \mathbf{b}_{i2})\sin(2\pi x_{ij}) + (\boldsymbol{\beta}_3 + \mathbf{b}_{i3})\cos(2\pi x_{ij}) + e_{ij} \quad (6.1)$$

might be appropriate for these data. For purposes of this comparison we assume that $\mathbf{b}_i \sim N(\mathbf{0}, \sigma^2\mathbf{D})$ and $\mathbf{e}_{ij} \sim N(\mathbf{0}, \sigma^2)$, where no restrictions are placed on the form of $\sigma^2\mathbf{D}$ except that it be a covariance matrix. In Section 7 we discuss the implications of assuming independent conditional errors for longitudinal data.

The second example is data from a study of the effects of a calcium supplement on bone density (Smith, Sempos, Smith, and Gilligan 1988). We consider data on 74 postmenopausal women equally divided into those that received calcium and those that received a placebo. The days on which the density measurements were taken vary from woman to woman but are at intervals of approximately 3 months for the first 12 months of the study (including a baseline measurement) and then every 6 months for an additional 3 years. The model postulated for these data is a linear growth curve with different slopes and intercepts for the two groups. That is, for individuals in group 1

$$y_{ij} = (\boldsymbol{\beta}_1 + \mathbf{b}_{i1}) + (\boldsymbol{\beta}_2 + \mathbf{b}_{i2})x_{ij} + e_{ij} \quad (6.2)$$

and for individuals in group 2

$$y_{ij} = (\boldsymbol{\beta}_3 + \mathbf{b}_{i1}) + (\boldsymbol{\beta}_4 + \mathbf{b}_{i2})x_{ij} + e_{ij},$$

where once again we assume that $\mathbf{b}_i \sim N(\mathbf{0}, \sigma^2\mathbf{D})$ and $\mathbf{e}_{ij} \sim N(\mathbf{0}, \sigma^2)$ where $\sigma^2\mathbf{D}$ is a general covariance matrix.

Details on the behavior of the algorithms for these examples are shown in Table 1. For consistency, convergence of all of the algorithms was assessed by the orthogonality convergence criterion (Bates and Watts 1981). This criterion is a dimensionless quantity that measures the size of the numerical variability relative to the statistical variability. Convergence is declared when the criterion is less than .001, that is, when the ratio of the numerical uncertainty of the estimate is negligible compared with the statistical variability of the parameters. It is important to note that the orthogonality criterion requires the calculation of an NR step, so it is not usually available during EM iterations. The timings for the EM algorithm do not include the calculation of the convergence criterion. Convergence of the EM algorithm was tested every iteration for the first 20 iterations and every tenth iteration thereafter. The Aitken acceleration was calculated every 9 iterations for the examples with $q = 3$ and every 6 iterations for the example with $q = 2$.

Pairwise plots of the iteration paths for some of the parameters in the horse-follicle example are presented in Figure 1. These plots graphically demonstrate the slow

Table 1. Comparison of the NR and EM Algorithms

Example [model]	M	N	Average (n_i)	p	q	Method	Iterations required	Average second/iteration*
Follicles [Eq. (6.1)]	11	308	28	3	3	NR	4	1.58
						EM	50	.50
						EM with Aitken's	9	.50
Bone density [Eq. (6.2)]	74	804	11	4	2	NR	3	3.53
						EM	40	1.18
						EM with Aitken's	6	1.18
Bone density [Eq. (6.3)]	74	804	11	6	3	NR	6	9.21
						EM	>200	2.45
						EM with Aitken's	>200	2.45

* Timings were done on a Vax 11/750 running 4.3 BSD Unix.

convergence of the EM algorithm. The steps it takes toward the optimum usually undershoot the optimum and the directions of the increments do not change substantially from iteration to iteration. In contrast, the NR iterations quickly bracket and then converge to the optimum.

For the first two models in Table 1 the EM algorithm is quite slow compared with the NR algorithm. On the other hand, the EM algorithm with Aitken's acceleration seems to be quite competitive with the NR. In an attempt to distinguish these two algorithms, quadratic terms were added to the model for the bone-density data; that is, for individuals in group 1

$$y_{ij} = (\boldsymbol{\beta}_1 + \mathbf{b}_{i1}) + (\boldsymbol{\beta}_2 + \mathbf{b}_{i2})x_{ij} + (\boldsymbol{\beta}_3 + \mathbf{b}_{i3})x_{ij}^2 + e_{ij} \tag{6.3}$$

and for individuals in group 2

$$y_{ij} = (\boldsymbol{\beta}_4 + \mathbf{b}_{i1}) + (\boldsymbol{\beta}_5 + \mathbf{b}_{i2})x_{ij} + (\boldsymbol{\beta}_6 + \mathbf{b}_{i3})x_{ij}^2 + e_{ij}.$$

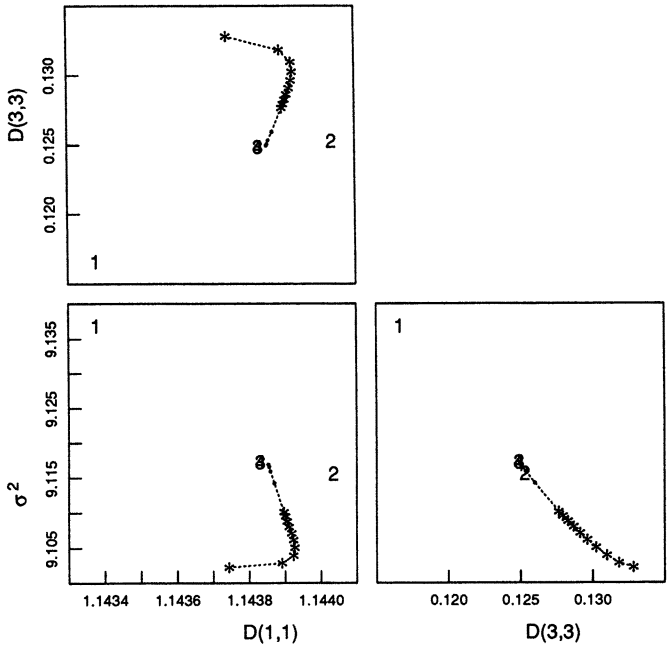

Figure 1. Detail of Iterative Paths for D_{11}, D_{33}, and σ^2 in the Ovarian Follicle Example. The numbers represent the NR iterations beginning with iteration 1 (the starting value is not shown). Iterations 3 and 4 are superimposed on the optimum. The asterisks and small dots connected by a dotted line represent the EM iterations without Aitken's acceleration. The asterisks represent iterations 1–10, and the small dots represent iterations 20, 30, 40, and 50.

It was expected that the likelihood surface for the parameters in this model would be quite flat and difficult to maximize, since the quadratic term is not a significant improvement to the model. In fact, although the NR algorithm converged quickly, the EM algorithm both with and without Aitken's acceleration failed to converge after 200 iterations. We believe that this is a fair test of the algorithms, because it is often necessary to overfit a data set before settling on a reduced model. The EM algorithm with Aitken's acceleration attained estimates that were close to the optimal values quite quickly but was then unable to converge within a reasonable number of iterations.

7. DISCUSSION AND EXTENSIONS

7.1 Discussion

Based on the work presented in this article we can compare the EM, EM with Aitken's acceleration, and NR algorithms as methods for obtaining estimates for mixed-effects models. The qualities of a good optimization algorithm include (in order of importance) iterative computations that are not prohibitively time consuming, quick and consistent convergence, and a good method for assessing convergence.

First, no matter how effective an algorithm is in other ways, it is worthless if each iteration takes too much time to be practical. When the EM algorithm is implemented as described in this article the iterations are of order dominated by $M(q^3 + q^2p)$ rather than N. This is a substantial improvement over existing methods when the sum of the lengths of the individuals' data vectors is large. Except for the initial, one-time, matrix calculations of order $(p + q)^2(N - M)$, there is no computational penalty on gathering a large amount of data for each individual. In the past the EM algorithm has been preferred over the NR algorithm because each iteration could be computed more quickly. If q is relatively small and the NR algorithm is implemented as described, however, then there is no significant computational penalty in using the NR algorithm (the iterations are of order Mq^4). As the timings in the previous section show, for moderate size M and small p and q, the iterative calculations for both the EM and NR algorithms are on the order of a few seconds on a minicomputer.

The second criterion mentioned is quick and consistent

convergence. Quick convergence is simply a small number of iterations. The number of iterations required for the NR algorithm is usually quite small compared with the number for the EM algorithm. As illustrated in Section 6, Aitken's acceleration can improve the behavior of the EM algorithm substantially but not in all situations.

Consistent convergence (i.e., convergence in a very high percentage of data examples) is also essential. The EM algorithm will always converge to a local maximum of the likelihood surface but may require a prohibitively high number of iterations. Unlike the EM algorithm, the NR algorithm is not guaranteed to converge. The NR algorithm, however, implemented as we suggest (optimizing the profile likelihood with respect to $\boldsymbol{\beta}$ and the entries of the Cholesky factor of $\mathbf{D}$), will converge to a local maximum of the likelihood surface very consistently. Reparameterization is key to ensuring consistent convergence of the NR algorithm.

The third important feature of a good optimization routine is the existence of an objective convergence criterion. The orthogonality criterion recommended for the NR algorithm (Bates and Watts 1981) is such a criterion and can be easily calculated from the information available at each iteration. This criterion cannot be used for the EM algorithm without increasing the order of the EM iterations to that of the NR iterations. The criterion usually used for the EM algorithm is the size of the change in the likelihood or parameter estimates from one iteration to the next. This is a measure of lack of progress but not of actual convergence. We view this as a major drawback of the EM algorithm.

In addition to the advantages of the NR algorithm listed previously, the Hessian matrix for the parameter vector $[\boldsymbol{\beta}^T, \boldsymbol{\theta}^T]$ is available at the end of the NR iterations. In addition, unlike the EM algorithm, the NR algorithm can be adapted to handle most of the common extensions of the mixed-effects models. In summary, we feel that the advantages of the NR algorithm outweigh the increase in the computing time per iteration except when the number of random effects is very large.

On some computing systems, it is also valuable to compare algorithms on the extent to which they can be parallelized or vectorized. Examination of the formulas for the derivatives in the NR algorithm (Sec. 4.2) or the update for the EM algorithm (Sec. 5) shows that most of the computation in both algorithms can be done in parallel for each individual, with only a small part of the computation requiring combined information.

7.2 Extensions

One source of the interest in mixed-effects models for repeated-measures data is the frequent occurrence of longitudinal repeated-measures data for which, because of serial correlation among the observations within an individual, it is usually not reasonable to assume a spherical marginal correlation structure (constant correlation off the main diagonal). For balanced data, an alternative is the full multivariate model that assumes a general marginal covariance matrix. Both the spherical and the general marginal covariance matrix can be viewed as special cases of the mixed-effects model, where $\mathbf{Z}_i = \mathbf{1}$ (a column vector of 1s) or $\mathbf{Z}_i = \mathbf{I}$, respectively. A compromise between the full multivariate model and spherical correlation structure is to choose a $\mathbf{Z}_i$ that has fewer than n_i columns but enough to model $\boldsymbol{\Sigma}_i$ adequately. Choosing the columns of $\mathbf{Z}_i$ in this way, however, may be at odds with the original motivation for including the random effects in the model, that is, to model variability in conditional expectation among individuals. In practice, the columns of $\mathbf{Z}_i$ are often chosen as a subset of the columns of $\mathbf{X}_i$, and it is hoped that this will provide a structure rich enough to model the marginal covariance effectively.

An alternative approach is to model the serial correlation directly by specifying the Λ_i matrices appropriately. For instance, an (autoregressive) AR(1) type of conditional covariance matrix was suggested (see, e.g., Jennrich and Schluchter 1986). In this model $\boldsymbol{\theta}$, the vector of variance parameters, would include the correlation parameter ρ as well as the nonzero elements of $\mathbf{L}$. A more complicated time-series type of covariance structure could be specified if desired. Unbalanced data can be handled by defining the jkth element of Λ_i as $\rho^{|d(j,k)|}$, where $d(j, k)$ is the distance between the jth and kth observations. Although the time-series type of covariance structures cannot be written in the form $\mathbf{ZDZ}^T$, there may be practical problems with the estimability of the variance parameters when both random effects and a time-series type of $\boldsymbol{\Lambda}_i$ are specified. A further extension of this idea is to allow Λ_i to depend on the fixed and/or random effects. We could then redefine $\boldsymbol{\theta}$ to be the parameters that enter only in the variance components. The NR algorithm could be implemented for these modifications, with an increase in the computational order depending on the number of additional covariance parameters specified. The EM algorithm cannot be easily generalized to handle these models.

Another generalization of mixed-effects models are models where some of the random effects are independent of others (as might be the case if the data are grouped) or have levels that vary with an experimental unit other than individual (i.e., nested or split-plot type of models). Laird et al. (1987) described an example where the random effects from two groups may have different covariance matrices. They model this by letting $\mathbf{b}_i$ include the random effects for both groups and defining the $\mathbf{Z}_i$ matrices to include a block of zeros corresponding to the random effects for the group that does not include individual i. In this model the $\mathbf{D}$ matrix is block-diagonal and the EM algorithm automatically adapts for estimating the parameters. This method works well for grouped data, but the extension to nested models would not be as straightforward.

A second method for defining models with groups that have independent random effects is to include a term in the model for each set of random effects that come from a distribution with a general (unpatterned) covariance matrix. For example, in a model with two groups where the random effects from the groups may have different covariance matrices, we would define $\mathbf{b}_{1i}$ to be the random effects for individuals in group 1. Likewise, $\mathbf{b}_{2i}$ would be the random effects for individuals in group 2. This model

can be expressed in matrix notation as

$$\mathbf{y}_i \mid \mathbf{b}_i \sim N(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_{i1}\mathbf{b}_{i1} + \mathbf{Z}_{i2}\mathbf{b}_{i2}, \sigma^2\mathbf{I}), \quad b_{ij} \sim N(\mathbf{0}, \sigma^2\mathbf{D}_j),$$

where $\mathbf{Z}_{i2} = \mathbf{0}$ if individual i is in group 1 and $\mathbf{Z}_{i1} = \mathbf{0}$ for individuals in group 2.

This method extends easily to model nested data. This is seen by noticing that the model described in Section 1 is a nested model when $\Lambda = \mathbf{I}$, since its error structure involves two levels of experimental units (observations within individual and individuals). Recall that in this model the observation vector for individual i is modeled as $\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{I}\mathbf{e}_i$, where $\mathbf{b}_i \sim N(\mathbf{0}, \sigma^2\mathbf{D})$ and $\mathbf{e}_i \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. If we rename $\mathbf{b}_i$, $\mathbf{b}_{i1}$; $\mathbf{D}$, $\mathbf{D}_1$; $\mathbf{e}$, $\mathbf{b}_{i2}$; and let $\mathbf{D}_2 = \mathbf{I}$, then we have $\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_{i1}\mathbf{b}_{i1} + \mathbf{Z}_{i2}\mathbf{b}_{i2}$, where $\mathbf{Z}_{i2} = \mathbf{I}$ and $\mathbf{b}_{ij} \sim N(\mathbf{0}, \sigma^2\mathbf{D}_j)$ for $j = 1$ and 2. Additional levels of nesting are easily handled by adding more $\mathbf{Z}$ matrices. This model in general form is $\mathbf{y} \mid \mathbf{b} \sim N(\mathbf{X}\boldsymbol{\beta} + \sum_{j=1}^{J} \mathbf{Z}_j\mathbf{b}_j, \sigma^2\boldsymbol{\Lambda})$, where $\mathbf{b}_j \sim N(\mathbf{0}, \sigma^2\tilde{\mathbf{D}}_j)$, and the marginal distribution of $\mathbf{y}$ is $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \sigma^2(\boldsymbol{\Lambda} + \sum_{j=1}^{J} \mathbf{Z}_j\tilde{\mathbf{D}}_j\mathbf{Z}_j^T)$. We denote the random effect associated with the smallest experimental unit by $\mathbf{e}$ and its covariance matrix by $\sigma^2\boldsymbol{\Lambda}$ to conform with our previous notation. In general, the $\mathbf{Z}$ and $\mathbf{D}$ matrices will be block-diagonal, where the size and number of blocks is determined by the experimental unit that is relevant for the random effect. Both the NR and EM algorithms can be easily extended to handle this model.

[*Received September 1987. Revised May 1988.*]

REFERENCES

Bard, Y. (1974), *Nonlinear Parameter Estimation,* New York: Academic Press.

Bates, D. M., and Lindstrom, M. J. (1986), "Nonlinear Least Squares With Conditionally Linear Parameters," in *Proceedings of the Statistical Computing Section, American Statistical Association,* pp. 152–157.

Bates, D. M., and Watts, D. G. (1981), "A Relative Offset Orthogonality Convergence Criterion for Nonlinear Least Squares," *Technometrics,* 23, 179–183.

Chambers, J. M. (1977), *Computational Methods for Data Analysis,* New York: John Wiley.

Dongarra, J. J., Bunch, J. R., Moler, C. B., and Stewart, G. W. (1979), *Linpack Users' Guide,* Philadelphia: Society for Industrial and Applied Mathematics.

Gerald, C. F. (1970), *Applied Numerical Analysis,* Reading, MA: Addison-Wesley.

Harville, D. A. (1974), "Bayesian Inference for Variance Components Using Only Error Contrasts," *Biometrika,* 61, 383–385.

Jennrich, R. I., and Schluchter, M. D. (1986), "Unbalanced Repeated Measures Models With Structural Covariance Matrices," *Biometrics,* 42, 805–820.

Kennedy, W. J., and Gentle, J. E. (1980), *Statistical Computing,* New York: Marcel Dekker.

Laird, N., Lange, N., and Stram, D. (1987), "Maximum Likelihood Computations With Repeated Measures: Application of the EM Algorithm," *Journal of the American Statistical Association,* 82, 97–105.

Laird, N. M., and Ware, J. H. (1982), "Random-Effects Models for Longitudinal Data," *Biometrics,* 38, 963–974.

Moler, C. (1981) "Matlab Users' Guide," technical report, University of New Mexico, Dept. of Computer Science.

Pierson, R. A., and Ginther, O. J. (1987), "Follicular Population Dynamics During the Estrous Cycle of the Mare," *Animal Reproduction Science,* 14, 219–231.

Smith, E. L., Sempos, C. T., Smith, P. E., and Gilligan, C. (1988), "Calcium Supplementation and Bone Loss in Middle-Aged Women," unpublished manuscript submitted to the *American Journal of Clinical Nutrition.*

Stewart, G. W. (1973), *Introduction to Matrix Computations,* New York: Academic Press.

Stram, D. O., Laird, N. M., and Ware, J. H. (1986), "An Algorithmic Approach for the Fitting of a General Mixed ANOVA Model Appropriate in Longitudinal Settings," in *Computer Science and Statistics: Proceedings of the Seventeenth Symposium on the Interface,* Amsterdam: North-Holland, pp. 149–158.

Ware, J. H. (1985), "Linear Models for the Analysis of Longitudinal Studies," *The American Statistician,* 39, 95–101.