Computing Gaussian Likelihoods and Their Derivatives for General Linear Mixed Models

Article *in* SIAM Journal on Scientific Computing · November 1994

DOI: 10.1137/0915079

CITATIONS
97

READS
352

3 authors, including:

Russ Wolfinger
SAS Institute
262 PUBLICATIONS **21,779** CITATIONS
SEE PROFILE

John Sall
SAS Institute
8 PUBLICATIONS **263** CITATIONS
SEE PROFILE

COMPUTING GAUSSIAN LIKELIHOODS AND THEIR DERIVATIVES FOR GENERAL LINEAR MIXED MODELS

RUSS WOLFINGER, RANDY TOBIAS, AND JOHN SALL *

Abstract. Algorithms are described for computing the Gaussian likelihood or restricted likelihood corresponding to a general linear mixed model. Included are arbitrary covariance structures for both the random effects and errors. Formulas are also given for the first and second derivatives of the likelihoods, thus enabling a Newton-Raphson implementation. The algorithms make heavy use of the Cholesky decomposition, the sweep operator, and the W-transformation. Also described are the modifications needed for variance profiling, Fisher scoring, and MIVQUE(0), as well as the computational order of the procedures.

Key words. Cholesky decomposition, Fisher scoring, MIVQUE(0), Newton-Raphson, profiling, sweep, W-transformation

AMS subject classifications. 62J10, 62J99

1. Introduction.

1.1. The Model. The most common statistical model is the general linear model, which has the following signal-plus-noise form:

$$y = X\beta + \epsilon$$

Here y represents a known data vector of length n, β is a vector of p unknown parameters with known design matrix X, and ϵ is an unknown error vector. For inference purposes, ϵ is usually assumed to have a Gaussian (normal) distribution with mean 0 and variance matrix $\sigma^2 I$, where I is the $n \times n$ identity matrix. The model is "general" in the sense that the columns of X may consist of either known explanatory variables, as in regression, or dummy variables with 0s and 1s indicating the presence or absence of an effect, as in analysis of variance (ANOVA). The parameters in β are the objects of primary interest, and they are usually estimated using the method of least squares. Classical statistical inference via t- and F-tests are then possible under the above Gaussian assumption.

Probably the most attractive feature of the general linear model is its considerable flexibility in modeling the signal (in the form of the mean) of the data. However, the assumption that the variance matrix of the noise vector equals $\sigma^2 I$ is often too limiting, as under normality this implies that each observation is statistically independent.

The general linear mixed model lifts this restriction by taking the form

$$y = X\beta + Zv + \epsilon$$

where y, X, β, and ϵ, are as above, and v is a vector of g unknown parameters with known design matrix Z. The unknown parameters in β and v correspond to *fixed-effects* and *random-effects*, respectively, and the inclusion of both defines a *mixed model*. The parameters in β are still often the objects of primary interest, with their estimation being an important analysis goal. The parameters in v are considered to be random variables, and thus account for variability in the data beyond that modeled

* SAS Institute Inc., SAS Campus Drive, Cary, North Carolina, 27513 (sasrdw@unx.sas.com, sasrdt@unx.sas.com, sall@sas.com).

in ϵ. To be more precise about this variance modeling, assume v and ϵ have Gaussian distributions with 0 expectations and

$$Var\begin{bmatrix} v \\ \epsilon \end{bmatrix} = \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix}$$

where both G and R are nonsingular. As such, the variance of y is

$$V = ZGZ' + R$$

The above general definition of the mixed model subsumes that of the classical linear mixed model, in which $R = \sigma^2 I$ and G is a diagonal matrix of variance components [12], [13]. The simplest example of the classical case is the randomized block ANOVA, in which the experimental units occur in clusters, or *blocks*, and each unit is subjected to a level of some experimental factor, the *treatment*. Treatment is the fixed effect, and so X consists of t columns of dummy variables, where t is the number of treatment levels, each column indicating the treatment level of the observations. Similarly, block is the random effect, and Z consists of b dummy columns, where b is the number of blocks, each column indicating block membership. Finally, G equals σ_b^2 times the $b \times b$ identity matrix, σ_b^2 being the variance component for blocks.

Other common statistical analyses falling within this unifying paradigm include the covariance structure approach to repeated measures [11], split-plot and incomplete-block designs [20], MANOVA, seemingly unrelated regressions [23], random coefficients [22], shrinkage estimators, and best linear unbiased predictors [15].

This article presents the details of fitting mixed models with arbitrarily parameterized covariance structures for both G and R. The difficulty is in estimating the parameters of G and R, and many possible estimation methods exist. Several traditional approaches use the method of moments, which solves a system of equations relating expected mean squares to observed ones. However, these methods are often only applicable to classical mixed models for balanced data, that is, data with equal numbers of observations for each random effect. Instead, we make use of the Gaussian assumption and employ the method of maximum likelihood as well as one of its variants, restricted/residual maximum likelihood (REML) [6].

Unfortunately, except in special cases, the maximum likelihood or REML estimates of the parameters in G and R must be computed by numerical methods. The resulting problems can be very computationally intensive, and only in the past decade have sufficient computer resources been routinely available to conquer them. This article shows how to efficiently compute the likelihoods and their first two derivatives, thus enabling a Newton-Raphson optimization scheme. The techniques are implemented in the SAS/STAT MIXED procedure [16].

1.2. Outline. §2 presents the details of computing the likelihood and the restricted likelihood of the general linear mixed model. §3 contains formulas for the first and second derivatives of these likelihoods, including those needed for variance profiling and Fisher scoring. MIVQUE(0) initial estimates are the topic of Section 4 and computational order that of §5.

2. Likelihoods. Given particular covariance structures for G and R, let θ denote the vector of unknown parameters in V; assume it is of length q. we adopt a normal-theory maximum-likelihood approach to estimation [6], [2], [10], [19]. This approach is preferred to the traditional ANOVA method [18].

2.1. General Form of Likelihoods. The negative of twice the Gaussian log likelihood for the general linear mixed model is

$$-2l(\beta,\theta|y) = \log|V(\theta)| + (y - X\beta)'V^{-1}(\theta)(y - X\beta) + n\log 2\pi$$

Minimizing this expression analytically for β yields

$$b(\theta) = [X'V^{-1}(\theta)X]^{-1}X'V^{-1}(\theta)y$$

and substitution into the original equation produces the negative of twice the profile/concentrated log likelihood for θ:

$$-2l(\theta|y) = \log|V(\theta)| + [y - Xb(\theta)]'V^{-1}(\theta)[y - Xb(\theta)] + n\log 2\pi$$

Profiling means reducing the dimension of an objective function by analytic substitution. We perform it so that the numerical optimization can be carried out only over the parameters in θ. Assuming $\hat{\theta}$ is the resulting optimum, β is estimated by $b(\hat{\theta})$. The larger the dimension of β, the greater the savings in time profiling offers over the full numerical optimization on (β,θ).

The formula for the negative of twice the restricted/residual log likelihood by definition does not involve β, and is given by

$$\begin{aligned} -2l_R(\theta|y) = {} & \log|V(\theta)| + [y - Xb(\theta)]'V^{-1}(\theta)[y - Xb(\theta)] + \\ & \log|X'V^{-1}(\theta)X| + (n-p)\log 2\pi \end{aligned}$$

The final two expressions above are the objective functions for maximum likelihood and REML, respectively. Ignoring the constant terms, they can be written as

$$\begin{aligned} -2l(\theta|y) &= l_1(\theta) + l_2(\theta) \\ -2l_R(\theta|y) &= l_1(\theta) + l_2(\theta) + l_3(\theta) \end{aligned}$$

where

$$\begin{aligned} l_1(\theta) &= \log|V(\theta)| \\ l_2(\theta) &= r(\theta)'V^{-1}(\theta)r(\theta) \\ l_3(\theta) &= \log|X'V^{-1}(\theta)X| \end{aligned}$$

and

$$r(\theta) = y - Xb(\theta)$$

2.2. Factoring and Profiling the Residual Variance. It is possible to proceed one step further analytically by factoring out one parameter from V; this factorization may or may not be natural and desirable, depending on the structure of R. Assuming that it is desirable, denote the one parameter by σ^2, and let θ^* be the new remaining parameter vector, with $q-1$ elements. A common example is a diagonal G with variance components as diagonal elements and $R = \sigma^2 I$. In this case θ^* contains each unknown variance component divided by σ^2. The likelihoods become

$$\begin{aligned} -2l(\theta^*,\sigma^2|y) &= n\log\sigma^2 + l_1(\theta^*) + l_2(\theta^*)/\sigma^2 + n\log 2\pi \\ -2l_R(\theta^*,\sigma^2|y) &= (n-p)\log\sigma^2 + l_1(\theta^*) + l_2(\theta^*)/\sigma^2 + l_3(\theta^*) + (n-p)\log 2\pi \end{aligned}$$

Minimizing analytically for σ^2 yields

$$\begin{aligned}\hat{\sigma}^2(\theta^*) &= l_2(\theta^*)/n \\ \hat{\sigma}^2_R(\theta^*) &= l_2(\theta^*)/(n-p)\end{aligned}$$

Back substitution produces the profile likelihoods

$$\begin{aligned}-2l^*(\theta^*|y) &= l_1(\theta^*) + n\log l_2(\theta^*) + n + n\log(2\pi/n) \\ -2l^*_R(\theta^*|y) &= l_1(\theta^*) + (n-p)\log l_2(\theta^*) + l_3(\theta^*) + \\ &\quad (n-p) + (n-p)\log[2\pi/(n-p)]\end{aligned}$$

The only substantial difference between these and the previous expressions is that l_2 has been replaced by a multiple of $\log l_2$. The minimization thus proceeds along the same lines as when not profiling the variance.

2.3. Likelihood Calculations. Direct construction of V^{-1} in the expressions for the likelihoods can be computationally prohibitive when the number of data points n is large. As a method for reducing the order of the calculations, we employ an extension of the W-transformation [7], [5]. Begin by constructing the following cross-products matrix:

$$W_0 = \begin{bmatrix} X'R^{-1}X & X'R^{-1}Z & X'R^{-1}y \\ Z'R^{-1}X & Z'R^{-1}Z & Z'R^{-1}y \\ y'R^{-1}X & y'R^{-1}Z & y'R^{-1}y \end{bmatrix}$$

Assuming R is block-diagonal, this construction is best carried out by first premultiplying the corresponding blocks of X, Z, and y by the inverse of the Cholesky root of R, which we denote by $R^{-1/2}$. The Cholesky root is taken blockwise, thereby keeping the dimensions small. Regarding singularity tolerances for the Cholesky root, we recommend using the original diagonal elements of R times $10^4\zeta$, where ζ is a machine-dependent constant defined as the largest floating-point number for which $\zeta + 1 = \zeta$; typically $\zeta = 10^{-16}$. Twice the sum of the logs of the positive diagonal elements of $R^{-1/2}$ equals $-\log|R|$, and should be computed for use below.

The next step involves the sweep operator, which is closely related to Gauss-Jordan elimination and the Forward Doolittle procedure [4]. Sweeping the first partition of a symmetric, nonnegative definite matrix of the form

$$\begin{bmatrix} A & B \\ B' & C \end{bmatrix}$$

produces

$$\begin{bmatrix} A^- & A^-B \\ -B'A^- & C - B'A^-B \end{bmatrix}$$

where $^-$ denotes a g_2-generalized inverse. If A is nonsingular $A^- = A^{-1}$.

To compute likelihoods, sweep the first partition of the following augmentation of W_0:

$$\begin{bmatrix} I + L'Z'R^{-1}ZL & L'W_0(Z,\cdot) \\ W_0(\cdot,Z)L & W_0 \end{bmatrix}$$

Here L is the lower-triangular Cholesky root of G so that $G = LL'$. This use of L accomodates the case when G is singular [8]. The augmented matrix does not actually have to be stored in memory when G is diagonal [5]. In this case the sweep can be performed using only a vector of pivots and the elements in W_0. When G is non-diagonal, the lower triangle of the augmented matrix needs to be stored in memory. As with the Cholesky decomposition, we recommend using the original diagonal elements times $10^4\zeta$ as the singularity tolerance for the sweep.

Summing the logs of the positive pivots during the sweep produces

$$\begin{aligned}\log|I + L'Z'R^{-1}ZL| &= \log|I + ZLL'Z'R^{-1}| \\ &= \log|R + ZGZ'| + \log|R^{-1}| \\ &= \log|V| - \log|R|.\end{aligned}$$

Adding this expression to $\log|R|$ computed above yields l_1.

The sweep results in the following matrix:

$$\begin{bmatrix} X'V^{-1}X & X'V^{-1}Z & X'V^{-1}y \\ Z'V^{-1}X & Z'V^{-1}Z & Z'V^{-1}y \\ y'V^{-1}X & y'V^{-1}Z & y'V^{-1}y \end{bmatrix}$$

Then sweeping the submatrix

$$\begin{bmatrix} X'V^{-1}X & X'V^{-1}y \\ y'V^{-1}X & y'V^{-1}y \end{bmatrix}$$

on all but its last row produces

$$\begin{bmatrix} (X'V^{-1}X)^- & b \\ b' & l_2 \end{bmatrix}$$

The sum of the log of the pivots from this sweep is l_3, and $(X'V^{-1}X)^-$ is the approximate variance matrix of b.

One further reduction in dimensionality can be accomplished when ZGZ' and R are both block-diagonal, and the blocking form of R is the same or nested within that of ZGZ'. We call such blocks of ZGZ' *subjects*. In this case the data from different subjects are statistically independent, and the entire procedure above can be performed subjectwise on the data, successively accumulating l_1, l_2, and l_3.

3. Derivatives. The maximum likelihood or REML objective functions can be optimized numerically using a Newton-Raphson algorithm, which is preferred to the Expectation-Maximization (EM) algorithm because of its convergence properties and information matrices [13]. One proviso for Newton-Raphson is that the covariance structures of G and R be twice-differentiable. Possible structures for G and R in this class include diagonal, compound symmetry, unstructured, autoregressive, Toeplitz, and several spatial types [11], [24]. A Kalman filter approach can possibly be used to obtain second derivatives of more complicated time-series structures such as autoregressive-moving average and autoregressive conditionally heteroskedastic. Other heteroskedastic structures involving β as a part of the variance function can be difficult to differentiate, although pseudo-likelihoods can be constructed in this case [1].

Assuming the likelihoods can be differentiated, define the gradient vector and Hessian matrix as follows:

$$\begin{aligned} g_k &= \frac{\partial}{\partial\theta} l_k(\theta) \\ H_k &= \frac{\partial^2}{\partial\theta\partial\theta'} l_k(\theta) \end{aligned}$$

for $k = 1, 2, 3$. We subsequently drop the functional reference to θ, and use subscripting outside brackets to denote elements and dots to denote differentiation.

3.1. General Form of Derivatives. The following expressions are similar to those given in [11], [13]:

$$\begin{aligned} [g_1]_r &= tr\left(V^{-1}\dot{V}_r\right) \\ [g_2]_r &= -r'V^{-1}\dot{V}_rV^{-1}r \\ [g_3]_r &= -tr\left(X^{*\prime}V^{-1}\dot{V}_rV^{-1}X^*\right) \end{aligned}$$

for $r = 1, \ldots, q$, and recall q is the number of elements in θ. Here $X^* = XC$ for a matrix C satisfying $CC' = (X'V^{-1}X)^-$.

$$\begin{aligned} [H_1]_{rs} &= -tr\left(V^{-1}\dot{V}_rV^{-1}\dot{V}_s\right) + tr\left(V^{-1}\ddot{V}_{rs}\right) \\ [H_2]_{rs} &= 2r'V^{-1}\dot{V}_rV^{-1}\dot{V}_sV^{-1}r \\ &\quad -2r'V^{-1}\dot{V}_rV^{-1}X^*X^{*\prime}V^{-1}\dot{V}_sV^{-1}r \\ &\quad -r'V^{-1}\ddot{V}_{rs}V^{-1}r \\ [H_3]_{rs} &= 2tr\left(X^{*\prime}V^{-1}\dot{V}_rV^{-1}\dot{V}_sV^{-1}X^*\right) \\ &\quad -tr\left(X^{*\prime}V^{-1}\dot{V}_rV^{-1}X^*X^{*\prime}V^{-1}\dot{V}_sV^{-1}X^*\right) \\ &\quad -tr\left(X^{*\prime}V^{-1}\ddot{V}_{rs}V^{-1}X^*\right) \end{aligned}$$

for $r, s = 1, \ldots, q$. The second term in $[H_2]_{rs}$ is not found in the above references; it results from profiling on β. Note that the final term in each of the H_k expressions vanishes if $\ddot{V}_{rs} = 0$, which is usually the case, except for most time-series and spatial structures.

The components g_3, H_2, and H_3 can be written more concisely if we define

$$\begin{aligned} H_2^r &= X^{*\prime}V^{-1}\dot{V}_rV^{-1}r \\ H_2^{r,s} &= 2r'V^{-1}\dot{V}_rV^{-1}\dot{V}_sV^{-1}r - r'V^{-1}\ddot{V}_{rs}V^{-1}r \\ H_3^r &= X^{*\prime}V^{-1}\dot{V}_rV^{-1}X^* \\ H_3^{r,s} &= 2X^{*\prime}V^{-1}\dot{V}_rV^{-1}\dot{V}_sV^{-1}X^* - X^{*\prime}V^{-1}\ddot{V}_{rs}V^{-1}X^* \end{aligned}$$

With this notation,

$$\begin{aligned} [g_3]_r &= -tr\left(H_3^r\right) \\ [H_2]_{rs} &= H_2^{r,s} - 2H_2^{r\prime}H_2^s \\ [H_3]_{rs} &= tr\left(H_3^{r,s} - H_3^rH_3^s\right) \end{aligned}$$

3.2. G **Derivatives.** In this section we consider an efficient approach to computing derivatives with respect to those elements of θ in G. As with the likelihood calculations, the method is an extension of the W-transformation [7], [4], and avoids direct construction of V^{-1}. All of the computations are performed using the partitions of the following matrix:

$$\begin{aligned} W &= \begin{bmatrix} W(X,X) & W(X,Z) & W(X,r) \\ W(Z,X) & W(Z,Z) & W(Z,r) \\ W(r,X) & W(r,Z) & W(r,r) \end{bmatrix} \\ &= \begin{bmatrix} X^{*\prime}V^{-1}X^* & X^{*\prime}V^{-1}Z & X^{*\prime}V^{-1}r \\ Z'V^{-1}X^* & Z'V^{-1}Z & Z'V^{-1}r \\ r'V^{-1}X^* & r'V^{-1}Z & r'V^{-1}r \end{bmatrix} \end{aligned}$$

W can be computed by changing y to r and X to X^* in the matrix obtained in the likelihood calculations.

Note that if r corresponds to an element of G, then $\dot{V}_r = Z\dot{G}_r Z'$. Substituting this into the general formulas and performing cyclic permutations inside of the trace operator yield convenient forms for the derivatives. Intuitively, this algorithm turns the inversion problem "inside out." The following are the results:

$$\begin{aligned} [g_1]_r &= tr\Big(W(Z,Z)\dot{G}_r\Big) \\ [g_2]_r &= -W(Z,r)'\dot{G}_r W(Z,r) \\ [g_3]_r &= -tr\Big(W(X,Z)\dot{G}_r W(Z,X)\Big) \end{aligned}$$

$$\begin{aligned} [H_1]_{rs} &= -tr\Big(W(Z,Z)\dot{G}_r W(Z,Z)\dot{G}_s\Big) \\ &\quad +tr\Big(W(Z,Z)\ddot{G}_{rs}\Big) \\ [H_2]_{rs} &= 2W(r,Z)\dot{G}_r W(Z,Z)\dot{G}_s W(Z,r) \\ &\quad -2W(r,Z)\dot{G}_r W(Z,X)W(X,Z)\dot{G}_s W(Z,r) \\ &\quad -W(r,Z)\ddot{G}_{rs}W(Z,r) \\ [H_3]_{rs} &= 2tr\Big(W(X,Z)\dot{G}_r W(Z,Z)\dot{G}_s W(Z,X)\Big) \\ &\quad -tr\Big(W(X,Z)\dot{G}_r W(Z,X)W(X,Z)\dot{G}_s W(Z,X)\Big) \\ &\quad -tr\Big(W(X,Z)\ddot{G}_{rs}W(Z,X)\Big) \end{aligned}$$

The simplifying components are

$$\begin{aligned} H_2^r &= W(X,Z)\dot{G}_r W(Z,r) \\ H_2^{r,s} &= 2W(r,Z)\dot{G}_r W(Z,Z)\dot{G}_s W(Z,r) - W(r,Z)\ddot{G}_{rs}W(Z,r) \\ H_3^r &= W(X,Z)\dot{G}_r W(Z,X) \\ H_3^{r,s} &= 2W(X,Z)\dot{G}_r W(Z,Z)\dot{G}_s W(Z,X) - W(X,Z)\ddot{G}_{rs}W(Z,X) \end{aligned}$$

As in the likelihood calculations, the dimensionality of the above calculations can be reduced when ZGZ' is block-diagonal and its blocks contain those of R. Then the entire procedure above can be performed blockwise, successively accumulating the derivatives on a subject-by-subject basis.

3.3. R **Derivatives.** For derivatives with respect to those elements of θ in R, we make heavy use of the identity

$$V^{-1} = R^{-1} - R^{-1}ZMZ'R^{-1}$$

where

$$M = (G^{-1} + Z'R^{-1}Z)^{-1}$$

For a derivation, see [9] or [17] §10.8.

When R is block-diagonal, this expression shows that operations involving V^{-1} can be computed without inverting large matrices. Our strategy is to substitute this expression for V^{-1} into the general formulas and then to cyclically permute matrices inside the trace operator. We also use the fact that if r corresponds to an element of R, then $\dot{V}_r = \dot{R}_r$.

The resulting expansions are lengthy, but many of the terms equal 0 for some important special cases. For example, if $M = 0$, as is the case when there are no random effects or when computing MIVQUE(0) estimates, then only one term in each of the expansions is needed. Also, several of the terms drop out if $\ddot{R}_{rs} = 0$. Finally, none of the equations are needed if R is equal to $\sigma^2 I$. The following are the results:

$$\begin{aligned}
[g_1]_r &= tr\left(R^{-1}\dot{R}_r\right) \\
&\quad -tr\left(MZ'R^{-1}\dot{R}_rR^{-1}Z\right) \\
[g_2]_r &= -r'R^{-1}\dot{R}_rR^{-1}r \\
&\quad +2r'R^{-1}ZMZ'R^{-1}\dot{R}_rR^{-1}r \\
&\quad -r'R^{-1}ZMZ'R^{-1}\dot{R}_rR^{-1}ZMZ'R^{-1}r \\
[g_3]_r &= -tr(H_3^r)
\end{aligned}$$

$$\begin{aligned}
[H_1]_{rs} &= -tr\left(R^{-1}\dot{R}_rR^{-1}\dot{R}_s\right) \\
&\quad +tr\left(MZ'R^{-1}\dot{R}_rR^{-1}\dot{R}_sR^{-1}Z\right) \\
&\quad +tr\left(MZ'R^{-1}\dot{R}_sR^{-1}\dot{R}_rR^{-1}Z\right) \\
&\quad -tr\left(MZ'R^{-1}\dot{R}_rR^{-1}ZMZ'R^{-1}\dot{R}_sR^{-1}Z\right) \\
&\quad +tr\left(R^{-1}\ddot{R}_{rs}\right) \\
&\quad -tr\left(MZ'R^{-1}\ddot{R}_{rs}R^{-1}Z\right) \\
[H_2]_{rs} &= H_2^{r,s} - 2H_2^{r'}H_2^s \\
[H_3]_{rs} &= tr(H_3^{r,s} - H_3^rH_3^s)
\end{aligned}$$

where

$$
\begin{aligned}
H_2^r \;=\;& X^{*\prime}R^{-1}\dot{R}_rR^{-1}r \\
&-X^{*\prime}R^{-1}ZMZ'R^{-1}\dot{R}_rR^{-1}r \\
&-X^{*\prime}R^{-1}\dot{R}_rR^{-1}ZMZ'R^{-1}r \\
&+X^{*\prime}R^{-1}ZMZ'R^{-1}\dot{R}_rR^{-1}ZMZ'R^{-1}r
\end{aligned}
$$

$$
\begin{aligned}
H_2^{r,s} \;=\;& 2r'R^{-1}\dot{R}_rR^{-1}\dot{R}_sR^{-1}r \\
&-2r'R^{-1}ZMZ'R^{-1}\dot{R}_rR^{-1}\dot{R}_sR^{-1}r \\
&-2r'R^{-1}\dot{R}_rR^{-1}ZMZ'R^{-1}\dot{R}_sR^{-1}r \\
&+2r'R^{-1}ZMZ'R^{-1}\dot{R}_rR^{-1}ZMZ'R^{-1}\dot{R}_sR^{-1}r \\
&-2r'R^{-1}\dot{R}_rR^{-1}\dot{R}_sR^{-1}ZMZ'R^{-1}r \\
&+2r'R^{-1}ZMZ'R^{-1}\dot{R}_rR^{-1}\dot{R}_sR^{-1}ZMZ'R^{-1}r \\
&+2r'R^{-1}\dot{R}_rR^{-1}ZMZ'R^{-1}\dot{R}_sR^{-1}ZMZ'R^{-1}r \\
&-2r'R^{-1}ZMZ'R^{-1}\dot{R}_rR^{-1}ZMZ'R^{-1}\dot{R}_sR^{-1}ZMZ'R^{-1}r \\
&-r'R^{-1}\ddot{R}_{rs}R^{-1}r \\
&+2r'R^{-1}\ddot{R}_{rs}R^{-1}ZMZ'R^{-1}r \\
&-r'R^{-1}ZMZ'R^{-1}\ddot{R}_{rs}R^{-1}ZMZ'R^{-1}r
\end{aligned}
$$

$$
\begin{aligned}
H_3^r \;=\;& X^{*\prime}R^{-1}\dot{R}_rR^{-1}X^* \\
&-X^{*\prime}R^{-1}ZMZ'R^{-1}\dot{R}_rR^{-1}X^* \\
&-X^{*\prime}R^{-1}\dot{R}_rR^{-1}ZMZ'R^{-1}X^* \\
&+X^{*\prime}R^{-1}ZMZ'R^{-1}\dot{R}_rR^{-1}ZMZ'R^{-1}X^*
\end{aligned}
$$

$$
\begin{aligned}
H_3^{r,s} \;=\;& 2X^{*\prime}R^{-1}\dot{R}_rR^{-1}\dot{R}_sR^{-1}X^* \\
&-2X^{*\prime}R^{-1}ZMZ'R^{-1}\dot{R}_rR^{-1}\dot{R}_sR^{-1}X^* \\
&-2X^{*\prime}R^{-1}\dot{R}_rR^{-1}ZMZ'R^{-1}\dot{R}_sR^{-1}X^* \\
&+2X^{*\prime}R^{-1}ZMZ'R^{-1}\dot{R}_rR^{-1}ZMZ'R^{-1}\dot{R}_sR^{-1}X^* \\
&-2X^{*\prime}R^{-1}\dot{R}_rR^{-1}\dot{R}_sR^{-1}ZMZ'R^{-1}X^* \\
&+2X^{*\prime}R^{-1}ZMZ'R^{-1}\dot{R}_rR^{-1}\dot{R}_sR^{-1}ZMZ'R^{-1}X^* \\
&+2X^{*\prime}R^{-1}\dot{R}_rR^{-1}ZMZ'R^{-1}\dot{R}_sR^{-1}ZMZ'R^{-1}X^* \\
&-2X^{*\prime}R^{-1}ZMZ'R^{-1}\dot{R}_rR^{-1}ZMZ'R^{-1}\dot{R}_sR^{-1}ZMZ'R^{-1}X^* \\
&-X^{*\prime}R^{-1}\ddot{R}_{rs}R^{-1}X^* \\
&+2X^{*\prime}R^{-1}\ddot{R}_{rs}R^{-1}ZMZ'R^{-1}X^* \\
&-X^{*\prime}R^{-1}ZMZ'R^{-1}\ddot{R}_{rs}R^{-1}ZMZ'R^{-1}X^*
\end{aligned}
$$

Each of the terms above can be computed piecewise if R is block-diagonal. This is carried out by expanding the $\dot{R}_r$ factor into a sum of the blocks. The individual blocks can be computed by looping through the random effects, while the other factors in the term need to be computed previously. The use of $R^{-1/2}$ throughout is recommended.

3.4. Cross Derivatives. Here we derive expressions for $[H_k]_{rs}$ when r corresponds to a parameter from R and s corresponds to a parameter from G. This effectively means repeating some of the formulas from the previous section with $\dot{R}_s$ replaced by $Z\dot{G}_s Z'$. We assume that G and R share none of the same parameters, so all of the 2-dot terms are zero. Also note that H_2^r, H_2^s, H_3^r, and H_3^s will already be computed by one of the previous methods.

$$\begin{aligned}
[H_1]_{rs} &= -tr\left(Z'R^{-1}\dot{R}_r R^{-1}Z\dot{G}_s\right) \\
&\quad +2tr\left(Z'R^{-1}\dot{R}_r R^{-1}Z\dot{G}_s Z'R^{-1}ZM\right) \\
&\quad -tr\left(Z'R^{-1}\dot{R}_r R^{-1}ZMZ'R^{-1}Z\dot{G}_s Z'R^{-1}ZM\right) \\
[H_2]_{rs} &= H_2^{r,s} - 2H_2^{r\prime}H_2^s \\
[H_3]_{rs} &= tr\left(H_3^{r,s} - H_3^r H_3^s\right)
\end{aligned}$$

where

$$\begin{aligned}
H_2^{r,s} &= 2A_1^r B\dot{G}_s B'Z'R^{-1}r \\
H_3^{r,s} &= 2A_2^r B\dot{G}_s B'Z'R^{-1}X^*
\end{aligned}$$

and

$$\begin{aligned}
A_1^r &= r'R^{-1}\dot{R}_r R^{-1}Z \\
&\quad -r'R^{-1}ZMZ'R^{-1}\dot{R}_r R^{-1}Z \\
A_2^r &= X^{*\prime}R^{-1}\dot{R}_r R^{-1}Z \\
&\quad -X^{*\prime}R^{-1}ZMZ'R^{-1}\dot{R}_r R^{-1}Z \\
B &= MZ'R^{-1}Z - I.
\end{aligned}$$

The matrix $W_0(Z,\cdot)$ obtained in the likelihood calculations can be used in computing the above expressions.

3.5. Derivatives when Factoring or Profiling the Residual Variance. We now describe derivatives appropriate for the objective functions described in Section 2.2. Two different forms are presented: the first uses the factored objective functions and considers σ^2 as a parameter to be differentiated; the second uses the profiled objective functions and eliminates σ^2 from the optimization problem. The second method is more efficient for optimization purposes because it has one less parameter than the first. For the first case, the formulas are as follows:

$$\begin{aligned}
\frac{\partial}{\partial\theta^*}[-2l(\theta^*,\sigma^2)] &= \begin{bmatrix} g_1 + g_2/\sigma^2 \\ n/\sigma^2 - l_2/\sigma^4 \end{bmatrix} \\
\frac{\partial}{\partial\theta^*}[-2l_R(\theta^*,\sigma^2)] &= \begin{bmatrix} g_1 + g_2/\sigma^2 + g_3 \\ (n-p)/\sigma^2 - l_2/\sigma^4 \end{bmatrix} \\
\frac{\partial^2}{\partial(\theta^*,\sigma^2)\partial(\theta^*,\sigma^2)}[-2l(\theta^*,\sigma^2)] &= \begin{bmatrix} H_1 + H_2/\sigma^2 & -g_2/\sigma^4 \\ -g_2'/\sigma^4 & -n/\sigma^4 + 2l_2/\sigma^6 \end{bmatrix} \\
\frac{\partial^2}{\partial(\theta^*,\sigma^2)\partial(\theta^*,\sigma^2)}[-2l_R(\theta^*,\sigma^2)] &= \begin{bmatrix} H_1 + H_2/\sigma^2 + H_3 & -g_2/\sigma^4 \\ -g_2'/\sigma^4 & -(n-p)/\sigma^4 + 2l_2/\sigma^6 \end{bmatrix}
\end{aligned}$$

TABLE 1
Overall derivatives for maximum likelihood estimation.

Technique	Gradient	Hessian
Newton	$g_1 + g_2$	$H_1 + H_2$
Scoring	$g_1 + g_2$	$-H_1$
Factoring-Newton	$\begin{bmatrix} g_1 + g_2/\sigma^2 \\ n/\sigma^2 - l_2/\sigma^4 \end{bmatrix}$	$\begin{bmatrix} H_1 + H_2/\sigma^2 & -g_2/\sigma^4 \\ -g_2'/\sigma^4 & -n/\sigma^4 + 2l_2/\sigma^6 \end{bmatrix}$
Factoring-Scoring	$\begin{bmatrix} g_1 + g_2/\sigma^2 \\ n/\sigma^2 - l_2/\sigma^4 \end{bmatrix}$	$\begin{bmatrix} -H_1 & (g_1 + g_3)/\sigma^2 \\ (g_1 + g_3)'/\sigma^2 & n/\sigma^4 \end{bmatrix}$
Profiling-Newton	$g_1 + g_2^*$	$H_1 + H_2^* - g_2^* g_2^{*\prime}/n$
Profiling-Scoring	$g_1 + g_2^*$	$-H_1 - g_2^* g_2^{*\prime}/n$

For the second case, let g_2^* and H_2^* denote g_2 and H_2 divided by the appropriate $\hat{\sigma}^2$. The profiling formulas are then as follows:

$$\begin{aligned}
\frac{\partial}{\partial \theta^*}[-2l^*(\theta^*)] &= g_1 + g_2^* \\
\frac{\partial}{\partial \theta^*}[-2l_R^*(\theta^*)] &= g_1 + g_2^* + g_3 \\
\frac{\partial^2}{\partial(\theta^*, \sigma^2)\partial(\theta^*, \sigma^2)}[-2l^*(\theta^*, \sigma^2)] &= H_1 + H_2^* - g_2^* g_2^{*\prime}/n \\
\frac{\partial^2}{\partial(\theta^*, \sigma^2)\partial(\theta^*, \sigma^2)}[-2l_R^*(\theta^*, \sigma^2)] &= H_1 + H_2^* - g_2^* g_2^{*\prime}/(n-p) + H_3
\end{aligned}$$

These profiling formulas result from

$$\begin{aligned}
\frac{\partial}{\partial \theta} \log l_2 &= \frac{\frac{\partial}{\partial \theta} l_2}{l_2} \\
\frac{\partial^2}{\partial \theta \partial \theta'} \log l_2 &= \frac{\frac{\partial^2}{\partial \theta \partial \theta'} l_2}{l_2} - \left[\frac{\partial}{\partial \theta} \log l_2\right]\left[\frac{\partial}{\partial \theta} \log l_2\right]'
\end{aligned}$$

3.6. Overall and Scoring Derivatives. So far we have seen how to calculate $g_k = \frac{\partial}{\partial \theta} l_k$ and $H_k = \frac{\partial^2}{\partial \theta \partial \theta'} l_k$, $k = 1, 2, 3$. In this section we show how to put these together to form overall first and second derivatives. Once formed, the Newton-Raphson step is obtained by computing the inverse of the second derivative matrix, the Hessian, times the first derivative vector, the gradient.

Tables 1 and 2 present the overall derivatives for maximum likelihood and REML estimation, respectively. The maximum likelihood formulas are profiled with respect to the fixed effects. The rows labeled "Factoring" and "Profiling" correspond to the results from the previous section concerning the residual variance parameter, σ^2. For the factoring and profiling formulas, all of the derivatives are evaluated at θ^*. Also, g_2^* and H_2^* denote g_2 and H_2 divided by the appropriate $\hat{\sigma}^2$.

Also in Tables 1 and 2 are formulas for a modification of the Newton-Raphson algorithm known as Fisher scoring. A scoring algorithm works by replacing the second derivative matrix with its expectation at the true values of the parameters. This entails setting $b(\theta)$ equal to β before taking expectations, thereby changing r to $Zv + e$. The only random components are those with a subscript of 2, and the resulting

TABLE 2
Overall derivatives for REML estimation.

Technique	Gradient	Hessian
Newton	$g_1 + g_2 + g_3$	$H_1 + H_2 + H_3$
Scoring	$g_1 + g_2 + g_3$	$-H_1 + H_3$
Factoring-Newton	$\begin{bmatrix} g_1 + g_2/\sigma^2 + g_3 \\ (n-p)/\sigma^2 - l_2/\sigma^4 \end{bmatrix}$	$\begin{bmatrix} H_1 + H_2/\sigma^2 + H_3 & -g_2/\sigma^4 \\ -g_2'/\sigma^4 & -(n-p)/\sigma^4 + 2l_2/\sigma^6 \end{bmatrix}$
Factoring-Scoring	$\begin{bmatrix} g_1 + g_2/\sigma^2 \\ (n-p)/\sigma^2 - l_2/\sigma^4 \end{bmatrix}$	$\begin{bmatrix} -H_1 + H_3 & (g_1 + g_3)/\sigma^2 \\ (g_1 + g_3)'/\sigma^2 & (n+p)/\sigma^4 \end{bmatrix}$
Profiling-Newton	$g_1 + g_2^* + g_3$	$H_1 + H_2^* + H_3 - g_2^* g_2^{*\prime}/(n-p)$
Profiling-Scoring	$g_1 + g_2^* + g_3$	$-H_1 + H_3 - g_2^* g_2^{*\prime}/(n-p)$

expressions are as follows:

$$\begin{aligned} E_{b=\beta}[l_2(\theta)] &= n \\ E_{b=\beta}[l_2(\theta^*)] &= \sigma^2 n \\ E_{b=\beta}[g_2(\theta)] &= -g_1(\theta) \\ E_{b=\beta}[g_2(\theta^*)] &= -\sigma^2 g_1(\theta^*) \\ E_{b=\beta}[H_2(\theta)] &= -2H_1(\theta) + tr\left(V^{-1}\ddot{V}\right) \\ E_{b=\beta}[H_2(\theta^*)] &= -2\sigma^2 H_1(\theta^*) + \sigma^2 tr\left(V^{-1}\ddot{V}\right) \end{aligned}$$

The scoring formulas in Tables 1 and 2 omit the 2-dot terms above, and are less computationally demanding than the full Newton-Raphson approach. Also, the profiling-scoring formulas are only a partial implementation, as it is cumbersome to take the expectation of $g_2^* g_2^{*\prime}$. Another partial scoring implementation involves using

$$E[rr'] = V - X^* X^{*\prime}$$

directly, although this technique has nearly the same computational order as full Newton-Raphson.

One may wish to transform the Hessian for the (θ^*, σ^2) parameterization into that for the θ parameterization. This is accomplished by

$$H(\theta) = BH(\theta^*, \sigma^2)B' + A$$

For example, if $\theta^* = (\theta_1^*, \theta_2^*)$

$$A = \begin{bmatrix} 0 & 0 & -g(\theta_1^*)/\sigma^4 \\ 0 & 0 & -g(\theta_2^*)/\sigma^4 \\ -g(\theta_1^*)/\sigma^4 & -g(\theta_2^*)/\sigma^4 & 2/\sigma^4[\theta_1^* g(\theta_1^*) + \theta_2^* g(\theta_2^*)] \end{bmatrix}$$

$$B = \begin{bmatrix} 1/\sigma^2 & 0 & 0 \\ 0 & 1/\sigma^2 & 0 \\ -\theta_1^*/\sigma^2 & -\theta_2^*/\sigma^2 & 1 \end{bmatrix}$$

Note that A vanishes both at the optimum and when scoring.

4. Initial Estimates. We recommend the MIVQUE(0) method [14] for computing initial estimates. One advantage of this approach is that in certain balanced cases it produces the REML estimates, thus eliminating the need for Newton-Raphson iterations. Disadvantages include the fact that the resulting V may not be positive definite

and that MIVQUE(0) may not be appropriate for many time-series and spatial structures. In these cases we suggest setting V to some easily obtained and sensible value, or using user-supplied initial values. A brief explanation of MIVQUE(0) estimates now follows, and then their computing formulas.

As before, let y be data with variance matrix V, with V containing q unknown parameters denoted by the vector θ. An assumption for the MIVQUE(0) approach is that one can write

$$V = \sum_{r=1}^{q} \theta_r \dot{V}_r$$

This essentially requires V to have a general linear structure with $\ddot{V}_{rs} = 0$, an assumption not satisfied for many time-series and spatial structures. Define

$$P = V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}$$

The MIVQUE(0) estimates of θ solve the equations

$$\left[tr\left(P\dot{V}_r P\dot{V}_s\right)\right]_{r,s} \theta = \left[y'P\dot{V}_r Py\right]_r$$

where P is computed at $V = I$. It is a special case of the MIVQUE(V_0) method which consists of setting the quadratic forms $y'P\dot{V}_r Py$ equal to their expected values given $V = V_0$ [18]. By examining §3.1, these equations are equivalent to

$$-(H_1 + H_3)\theta = -g_2$$

We solve this system for θ when using REML.

For maximum likelihood, we solve the modified MIVQUE(0) equations

$$\left[tr\left(V^{-1}\dot{V}_r V^{-1}\dot{V}_s\right)\right]_{r,s} \theta = \left[y'P\dot{V}_r Py\right]_r$$

which consist of setting the quadratic forms $y'P\dot{V}_r Py$ equal to the expected values of $r'V^{-1}\dot{V}_r V^{-1}r$ given $V = I$. These equations reduce to

$$-H_1\theta = -g_2$$

When profiling σ^2 out of V, the following formulas are used. For REML,

$$\begin{bmatrix} \left[tr\left(P\dot{V}_r P\dot{V}_s\right)\right]_{r,s=1}^{q-1} & \left[tr\left(P\dot{V}_r P\right)\right]_{r=1}^{q-1} \\ \left(\left[tr\left(P\dot{V}_s P\right)\right]_{s=1}^{q-1}\right)' & tr\left(P^2\right) \end{bmatrix} \begin{bmatrix} \sigma^2\theta^* \\ \sigma^2 \end{bmatrix} = \begin{bmatrix} \left[y'P\dot{V}_r Py\right]_{r=1}^{q-1} \\ y'P^2y \end{bmatrix}$$

which under the idempotency of P becomes

$$\begin{bmatrix} -(H_1 + H_3) & g_1 + g_3 \\ g_1' + g_3' & n - p \end{bmatrix} \begin{bmatrix} \sigma^2\theta^* \\ \sigma^2 \end{bmatrix} = \begin{bmatrix} -g_2 \\ l_2 \end{bmatrix}$$

For maximum likelihood, we have

$$\begin{bmatrix} \left[tr\left(V^{-1}\dot{V}_r V^{-1}\dot{V}_s\right)\right]_{r,s=1}^{q-1} & \left[tr\left(V^{-1}\dot{V}_r V^{-1}\right)\right]_{r=1}^{q-1} \\ \left(\left[tr\left(V^{-1}\dot{V}_s V^{-1}\right)\right]_{s=1}^{q-1}\right)' & tr\left(V^{-2}\right) \end{bmatrix} \begin{bmatrix} \sigma^2\theta^* \\ \sigma^2 \end{bmatrix} = \begin{bmatrix} \left[y'P\dot{V}_r Py\right]_{r=1}^{q-1} \\ y'Py \end{bmatrix}$$

TABLE 3
Notation for order expressions.

Symbol	Number
p	columns of X
g	columns of Z
n	observations
q	covariance parameters
t	maximum observations per subject
s	subjects

which when $V = I$ becomes

$$\begin{bmatrix} -H_1 & g_1 \\ g_1' & n \end{bmatrix} \begin{bmatrix} \sigma^2\theta^* \\ \sigma^2 \end{bmatrix} = \begin{bmatrix} -g_2 \\ l_2 \end{bmatrix}$$

5. Computational Order. Using the notation from Table 3, the following are estimates of the computational speed of the algorithms described above. For likelihood calculations, the cross-products matrix construction is of order $n(p+g)^2$ and the sweep operations are of order $(p+g)^3$. The first derivative calculations for parameters in H are of order qg^3 for maximum likelihood and $q(g^3 + pg^2 + p^2g)$ for REML. If ZGZ' is block-diagonal with blocks corresponding to subjects, then replace g by g/s and q by qs in these calculations. The first derivative calculations for parameters in R are of order $qs(t^3 + gt^2 + g^2t)$ for maximum likelihood and $qs(t^3 + (p+g)t^2 + (p^2+g^2)t)$ for REML. For the second derivatives, replace q by $q(q+1)/2$ in the first derivative expressions. When there are both G and R parameters, then additional calculations are required of order equal to the sum of the orders for G and R; this is the most computationally intensive case. The approximate memory requirement in bytes is $40(p^2+g^2) + 32(p+g)^2$.

Note that the second derivatives cost approximately $q/2$ times as much as the first derivatives, which in turn cost approximately q times as much as likelihood evaluations. This leads one to suspect that first-derivative (quasi-Newton) or derivative-free methods may outperform standard Newton-Raphson for large problems. Sparse matrix techniques are also a possibility, and progress is being made primarily by animal breeders, who often have millions of records [21].

To provide a brief illustration of the efficiency of the algorithms described in this paper, Tables 4 and 5 describe several small examples and their run times on different computing systems. All calculations are performed with Release 6.08.01 of the SAS/STAT MIXED procedure [16], and Table 5 displays time to completion of PROC MIXED with its default options. Approximately 5-10 percent of the time is spent in initial data set-up and post-convergence calculations, while the remainder is consumed by a ridge-stabilized Newton-Raphson algorithm with REML MIVQUE(0) starting values (§4). Convergence is assumed when the relative orthogonality criterion [13] is less than 10^{-8}.

Table 4 lists the dimensions of each example using the notation from Table 3. In addition, the number of covariance parameters, q, is divided into those corresponding to G (q_G) and R (q_R), indicating the appropriate formulas from §3. Table 4 also includes the number of likelihood evaluations and Newton-Raphson iterations required for convergence.

Examples 1, 3, 6, and 7 are all simple random effects models with diagonal G and R. Examples 6 and 7 employ the same data and model; however, Example 6 breaks

TABLE 4
Dimensions of run-time examples, using notation from Table 3.

Example	p	g	n	q	q_G	q_R	t	s	Number of Evaluations	Number of Iterations
1	3	33	308	7	6	1	31	11	4	2
2	5	0	2002	10	0	10	4	930	4	2
3	108	285	240	8	7	1	240	1	9	3
4	84	21	108	6	2	4	36	3	11	6
5	4	0	306	2	0	2	306	1	4	2
6	231	630	3008	3	2	1	21	30	6	3
7	231	630	3008	3	2	1	3008	1	6	3

TABLE 5
Elapsed times in minutes:seconds for the examples in Table 4 running on various systems.

Example	486 PC Windows 3.1 16 mb	486 PC OS/2 2.1 16 mb	HP 9000/720 HP-UX 9.01 32 mb	IBM 3090 MVS 32 mb	IBM 3090 MVS-vector 32 mb	Convex 3800 ConvexOS 512 mb
1	00:05	00:03	00:02	00:01	00:01	00:01
2	04:40	01:12	00:42	00:28	00:25	00:33
3	03:23	02:20	00:54	00:48	00:35	00:10
4	09:53	01:51	01:05	00:46	00:44	00:44
5	52:57	07:16	04:07	02:47	02:36	02:41
6	53:49	14:42	06:05	09:26	09:02	01:56
7	52:57	21:11	08:43	05:25	03:42	02:08

Z into 30 subjects during the computations as described at the end of Section 2. In contrast, Example 7 operates on the entire 630 columns of Z at once. Examples 2 and 5 have no G matrix; Example 2 has a large, unstructured-block-diagonal R matrix, and Example 5 has a 306×306 dense R matrix. Finally, Example 4 has a diagonal G matrix and a Toeplitz-block-diagonal R matrix.

Table 5 compares the computing speed of these examples for several common hardware configurations. The times increase roughly as one one moves down and left on the table, and all of them are less than an hour. As expected, the PCs ran the slowest, with OS/2 outperforming Windows by a factor ranging from 2 to 6. The HP Unix workstation ran approximately twice as fast as OS/2 and 1.5 times slower than a standard IBM mainframe system. The addition of a vector facility on the IBM mainframe produces savings ranging from 5-30 percent over the standard configuration. The Convex system was the fastest of the six.

One anomaly in Table 5 is the slowness of the mainframes for Example 6 as compared with Example 7. Because of its block-diagonal nature, one would expect Example 6 to run faster than Example 7, as it did on most of the other systems. Further investigation showed that the elapsed times of the mainframes for Example 6 were more than 3 times the CPU times, whereas in the other 6 examples they were less than 1.5 times the CPU times. The extra time resulted from I/O inefficiency, and a re-run under virtual I/0 produced an Example 6 time of 02:18 for the 3090 with the vector facility.

For a rough comparison with previously published results, Lindstrom and Bates [13] report a run time of 6.32 seconds for Example 1 on a Vax 11/750 running 4.3 BSC Unix. They utilize Newton-Raphson formulas based on the QR decomposition. The QR method has higher numerical stability than the sweep-based methods described in this paper because it operates on X and not $X'X$. However, the QR method is

approximately two times slower than the sweep [3] §5.3.9, and Example 1 provides a confirmation.

In summary, the algorithms appear to be able to handle moderately sized problems quite well. More rigorous investigation is called for to precisely determine their practicality.

Acknowledgements. We thank several referees for helpful comments, Tim Gregoire at VPI for pointing out [9], and Leigh Ihnen and Connie Dunbar at SAS for aid with the run-time examples.

REFERENCES

[1] R. J. CARROLL AND D. RUPPERT, *Transformation and Weighting in Regression.* Chapman and Hall, London, 1988.

[2] A. P. DEMPSTER, M. R. SELWYN, C. M. PATEL, AND A. J. ROTH, *Statistical and computational aspects of mixed model analysis*, Appl. Statist., 33 (1984), pp. 203-214.

[3] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations,* Second ed.,. The Johns Hopkins University Press, Baltimore, MD, 1989.

[4] J. H. GOODNIGHT, *A tutorial on the sweep operator*, Amer. Statist., 33 (1979), pp. 149-158.

[5] J. H. GOODNIGHT AND W. J. HEMMERLE, *A simplified algorithm for the W-transformation in variance component estimation*, Technom., 21 (1979), pp. 265-268.

[6] D. A. HARVILLE, *Maximum likelihood approaches to variance component estimation and to related problems*, J. of the Amer. Statist. Assoc., 72 (1977), pp. 320-338.

[7] W. J. HEMMERLE AND H. O. HARTLEY, *Computing maximum likelihood estimates for the mixed AOV model using the W-transformation*, Technom., 15 (1973), pp. 819-831.

[8] C. R. HENDERSON, *Applications of Linear Models in Animal Breeding.* University of Guelph, Canada, 1984.

[9] C. R. HENDERSON, O. KEMPTHORNE, S. R. SEARLE, AND C. M. VON KROSIGK, *The estimation of environmental and genetic traits from records subject to culling*, Biometrics, 15 (1959), pp. 192-218.

[10] R. R. HOCKING, *The Analysis of Linear Models.* Brooks/Cole, Monterey, CA, 1985.

[11] R. I. JENNRICH AND M. D. SCHLUCHTER, *Unbalanced repeated-measures models with structured covariance matrices*, Biometrics, 42 (1986), pp. 805-820.

[12] N. M. LAIRD AND J. H. WARE, *Random-effects models for longitudinal data*, Biometrics, 38 (1982), pp. 963-974.

[13] M. J. LINDSTROM AND D. M. BATES, *Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data*, J. of the Amer. Statist. Assoc., 83 (1988), pp. 1014-1022.

[14] C. R. RAO, *Estimation of variance and covariance components in linear models*, J. of the Amer. Statist. Assoc., 67 (1972), pp. 112-115.

[15] G. K. ROBINSON, *That BLUP is a good thing: the estimation of random effects* (with discussion), Statist. Science, 6 (1991), pp. 15-51.

[16] SAS INSTITUTE INC., *Chapter 16: The MIXED Procedure*, in SAS Technical Report P-229, SAS/STAT Software: Changes and Enhancements, Release 6.07, SAS Institute Inc., Cary, NC, 1992.

[17] S. R. SEARLE, *Linear Models*, John Wiley and Sons, New York, 1971.

[18] ———, *Mixed models and unbalanced data: wherefrom, whereat, and whereto?* Comm. in Statist. - Theory and Meth., 17 (1988), pp. 935-968.

[19] S. R. SEARLE, G. CASELLA, AND C. E. MCCULLOCH, *Variance Components*, John Wiley and Sons, New York, 1992.

[20] W. W. STROUP, *Predictable functions and prediction space in the mixed model procedure*, in Applications of Mixed Models in Agriculture and Related Disciplines, Southern Cooperative Series Bulletin No. 343, Louisiana Agricultural Experiment Station, Baton Rouge, 1989, pp. 39-48.

[21] L. D. VAN VLECK, *The revolution in statistical computing: from least squares to DFREML*, in Proceedings 41st National Breeders Roundtable, Poultry Breeders of America and Southeastern Poultry and Egg Assoc., May 7-8., St. Louis, MO, 1992.

[22] E. F. VONESH AND R. L. CARTER, *Efficient inference for random-coefficient growth curve models with unbalanced data*, Biometrics, 43 (1987), pp. 617-628.

[23] A. ZELLNER, *An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias*, J. of the Amer. Statist. Assoc., 57 (1962), pp. 348-368.

[24] D. L. ZIMMERMAN AND D. A. HARVILLE, *A random field approach to the analysis of field-plot experiments and other spatial experiments*, Biometrics, 47 (1991), pp. 223-239.