Maximum Likelihood Computations with Repeated Measures: Application of the EM Algorithm

Author(s): Nan Laird, Nicholas Lange and Daniel Stram

Source: *Journal of the American Statistical Association*, Mar., 1987, Vol. 82, No. 397 (Mar., 1987), pp. 97-105

Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association

Stable URL: https://www.jstor.org/stable/2289129

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at https://about.jstor.org/terms

Taylor & Francis, Ltd. and *American Statistical Association* are collaborating with JSTOR to digitize, preserve and extend access to *Journal of the American Statistical Association*

This content downloaded from 100.15.107.84 on Tue, 15 Dec 2020 03:11:24 UTC
All use subject to https://about.jstor.org/terms

Maximum Likelihood Computations With Repeated Measures: Application of the EM Algorithm

NAN LAIRD, NICHOLAS LANGE, and DANIEL STRAM*

The purpose of this article is to consider the use of the EM algorithm (Dempster, Laird, and Rubin 1977) for both maximum likelihood (ML) and restricted maximum likelihood (REML) estimation in a general repeated measures setting using a multivariate normal data model with linear mean and covariance structure (Anderson 1973). Several models and methods of analysis have been proposed in recent years for repeated measures data; Ware (1985) presented an overview. Because the EM algorithm is a general-purpose, iterative method for computing ML estimates with incomplete data, it has often been used in this particular setting (Dempster et al. 1977; Andrade and Helms 1984; Jennrich and Schluchter 1985).

There are two apparently different approaches to using the EM algorithm in this setting. In one application, each experimental unit is observed under a standard protocol specifying measurements at each of n occasions (or under n conditions), and incompleteness implies that the number of measurements actually collected on each unit is less than the requisite n for at least some units. In this circumstance, incompleteness may be modeled if one regards the measurements actually collected as the observed data, the conceptual set of n measurements on each individual as the complete data, and the unobserved data as the missing measurements on those units with fewer than n observations. Application of the EM algorithm in this setting [referred to as "missing data" in Dempster et al. (1977) and "incomplete data" in Jennrich and Schluchter (1985)] was discussed by Orchard and Woodbury (1972), Beale and Little (1975), and Jennrich and Schluchter (1985).

One drawback of this approach in the longitudinal data setting is that the multivariate model with linear mean and covariance structure does not, in general, possess closed-form solutions even with complete data (Anderson 1973; Szatrowski 1980). Thus implementing the EM algorithm requires either an iterative M step within each EM iteration or the use of a generalized EM (GEM) algorithm that requires only that the complete data likelihood be increased rather than maximized at each M step. A second drawback is that this approach requires specification of the covariates for both the observed and the missing observations. If the covariates are unknown for the missing observations, arbitrary values must be specified, which may affect the rate but not the final point of convergence (Jennrich and Schluchter 1985).

The second application of the EM algorithm arises naturally when we use mixed models to analyze serial measurements. In this setting, the incomplete data are modeled quite differently. The observed data are as before, that is, the measurements actually collected on each unit. The complete data, however, consist of the observed data plus the unobservable random parameters and error terms specified in the mixed model. Thus the missing data (the random parameters and error terms) would not be viewed as data in the traditional statistical sense. Laird and Ware (1982) and Andrade and Helms (1984) took this approach.

This article shows that the latter approach is more general and encompasses the missing-data approach as a special case. This result has several important applications. First, it means that EM algorithms encoded for models with random effects can also be used for multivariate normal models with arbitrary covariance structure and missing data. Second, this approach avoids specification of covariates for missing observations. Finally, use of the general formulation means that closed-form solutions for the complete data maximization will exist for a much broader class of models, enabling one to avoid use of GEM or iterations within each M step.

For a certain class of multivariate growth curve models with random effects structure (Reinsel 1982), closed-form solutions exist for both ML and REML estimates of the mean and covariance parameters. Formulas for these closed-form solutions are given that are applicable whenever the solution is not on the boundary.

The choice of starting values for the EM iterations is important, since the EM algorithm will not, in general, converge from arbitrary starting values to the closed-form solution (if it exists) in one iteration. Several possibilities for starting values are given.

The rate of convergence of the EM algorithm is generally linear. The actual speed of convergence in two data examples is shown to depend heavily on both the actual data set and the assumed structure for the covariance matrix. We discuss two methods for accelerating convergence, which we find are most useful when the covariance matrix is assumed to have a random effects structure. When the covariance matrix is assumed to be arbitrary, the EM iterations reduce to familiar iteratively reweighted least squares (IRLS) computations. The EM algorithm has the unusual property in this setting that when all of the data are complete (no missing observations), the iterations are still IRLS, but the rate of convergence changes from linear to quadratic.

KEY WORDS: Mixed models; Restricted maximum likelihood; Growth curves with random parameters; Aitken acceleration; Linear patterned covariance matrices.

1. INTRODUCTION

We begin by defining the multivariate normal data model and by demonstrating its generality for modeling both the mean vector and the correlation structure of the observations. Next, we give the iterative equations that define the EM algorithm. Finally, we discuss the existence of closed-form solutions in balanced data cases, computation of starting values for the iterations, and methods for speeding convergence. Two data examples are used to illustrate features of convergence.

2. THE GENERAL LINEAR MIXED MODEL FOR REPEATED MEASURES, GROWTH CURVE, OR SERIAL MEASUREMENT DATA

2.1 General Formulation and Relation to ANOVA Models

The general model that we use to characterize the common structure of repeated measures, growth curve, or serial measurements data is that described by Laird and Ware (1982). Specifically, let $\mathbf{y}_i$ denote an $n_i \times 1$ vector of n_i measurements observed on the ith experimental unit. We assume the model

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\alpha} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i, \tag{2.1}$$

* Nan Laird is Professor, Department of Biostatistics, School of Public Health, Harvard University, Boston, MA 02115. Nicholas Lange is Instructor, Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139. Daniel Stram is Visiting Scientist, Radiation Effects Research Foundation, Hiroshima, Japan. Support for this research was provided by National Institutes of Health Grant GM29745. The authors appreciate comments from Ken Byrk on an earlier draft of this article. The article is one of several papers organized with the editorial assistance of J. Michael Steele, Colin Goodall, and Douglas M. Bates.

© 1987 American Statistical Association
Journal of the American Statistical Association
March 1987, Vol. 82, No. 397, Theory and Methods

This content downloaded from
100.15.107.84 on Tue, 15 Dec 2020 03:11:24 UTC
All use subject to https://about.jstor.org/terms

where $\mathbf{X}_i$ and $\mathbf{Z}_i$ are known $n_i \times p$ and $n_i \times q$ design matrices, $\boldsymbol{\alpha}$ is a vector of fixed effects to be estimated, and $\mathbf{b}_i$ and $\mathbf{e}_i$ are independent random vectors distributed as $N(\mathbf{0}, \mathbf{D})$ and $N(\mathbf{0}, \sigma^2\mathbf{I}_i)$, respectively. Here $\mathbf{D}$ is a positive semidefinite $q \times q$ matrix of unknown parameters to be estimated, and $\mathbf{I}_i$ is the $n_i \times n_i$ identity matrix. It thus follows that

$$E(\mathbf{y}_i) = \mathbf{X}_i\boldsymbol{\alpha},$$

$$\text{var}(\mathbf{y}_i) = \boldsymbol{\Sigma}_i = \sigma^2\mathbf{I}_i + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T,$$

and

$$\text{cov}(\mathbf{y}_i, \mathbf{y}_j) = \mathbf{0}, \qquad i \neq j,$$

for $i = 1, \ldots, m$ units.

The form of $\boldsymbol{\Sigma}_i = \text{var}(\mathbf{y}_i)$ is the familiar variance components structure. It offers an approach to modeling $\boldsymbol{\Sigma}_i$ that is especially useful when individuals have varying patterns or times of observations, or large numbers of observations, as in the data examples in Hui and Berger (1983) and Dempster, Rubin, and Tsutakawa (1981). In some settings, each unit may be observed under a standard protocol, and, provided the number of observations on each unit is not large, one may prefer to assume that $\text{var}(\mathbf{y}_i)$ is unrestricted. This may be achieved in the variance components model by taking $\sigma^2 = 0$, and $\mathbf{Z}_i = \mathbf{I}_{n\times n}$ if a unit is measured at all n occasions. A unit missing observations at some occasions ($n_i < n$) would have $\mathbf{Z}_i$ as the rows of the identity matrix corresponding to occasions when measurements were made. Then

$$\text{var}(\mathbf{y}_i) = \mathbf{D} \quad \text{when } n_i = n$$

or

$$\text{var}(\mathbf{y}_i) = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T \equiv \mathbf{D}_i \quad \text{when } n_i < n,$$

where $\mathbf{D}_i$ denotes the $n_i \times n_i$ submatrix formed by deleting the rows and columns of $\mathbf{D}$ corresponding to the missing observations. When one is using the EM algorithm, setting $\sigma^2 = 0$ is easily accomplished by choice of starting values.

For any model, we let $\boldsymbol{\theta}$ denote the vector consisting of σ^2 and the unique elements of $\mathbf{D}$ to be estimated. In general, no restrictions are put on $\mathbf{D}$; in Section 2.3 we discuss the utility of assuming special structure on $\mathbf{D}$. Jennrich and Schluchter (1985) considered both first-order and general autoregressive structures for $\mathbf{D}$, as well as $\mathbf{D}$ arbitrary.

If we write $\mathbf{y}^T = (\mathbf{y}_1^T, \ldots, \mathbf{y}_m^T)$, then the linear mixed model for $\mathbf{y}$ has exactly the same form as the model discussed by Harville (1977), except that the implied covariance structure for $\mathbf{y}$ is block diagonal, with the m $\boldsymbol{\Sigma}_i$'s making up the elements on the diagonal. Thus our model is a special case of Harville's in which the only random factors are individual units and the interactions of unit with occasion or period variables. If $\mathbf{D}$ is diagonal and each $\mathbf{Z}_i$ consists of only zeros and ones, our model reduces to a special case of the general analysis of variance (ANOVA) mixed model, with one random factor corresponding to experimental units and any other random factor being an interaction of unit with another categorical variable. The compound symmetry model is a common simplification of the repeated measures model, where each $\mathbf{Z}_i$ is an $n_i \times 1$ vector of ones. Our model is thus in some ways more general and in some ways more restrictive than the general ANOVA mixed model.

2.2 A Growth Curve Formulation

The mixed model that we use can also be seen to have origins in the growth curve literature, where a slightly different approach to modeling has been traditional. Here it is more natural to specify the model characterizing the growth curve for each individual unit and then to model the parameters of the individual growth curves as linear functions of individual characteristics. Specifically, for the growth curve formulation, we assume that

$$\mathbf{y}_i = \mathbf{Z}_i\boldsymbol{\beta}_i + \mathbf{e}_i, \qquad i = 1, \ldots, m,$$

where $\mathbf{y}_i$, $\mathbf{e}_i$, and $\mathbf{Z}_i$ are as previously defined. The component $\mathbf{Z}_i\boldsymbol{\beta}_i$ defines the ith individual's growth curve. The $\boldsymbol{\beta}_i$'s are random parameters unique to each individual; they are assumed to be independently distributed as $N(\mathbf{A}_i\boldsymbol{\alpha}, \mathbf{D})$, where $\boldsymbol{\alpha}$ and $\mathbf{D}$ are as previously defined, and $\mathbf{A}_i$ is a $q \times p$ design matrix. Thus

$$E(\mathbf{y}_i) = \mathbf{Z}_i\mathbf{A}_i\boldsymbol{\alpha} = \mathbf{X}_i\boldsymbol{\alpha}$$

and

$$\text{var}(\mathbf{y}_i) = \sigma^2\mathbf{I}_i + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T,$$

where $\mathbf{X}_i = \mathbf{Z}_i\mathbf{A}_i$. It follows that $\boldsymbol{\beta}_i = \mathbf{A}_i\boldsymbol{\alpha} + \mathbf{b}_i$, and thus $\mathbf{b}_i$ can be viewed as a generalized residual vector referring to a deviation of the parameters of an individual's growth curve from the parameters of the population growth curve. This is in contrast to the ordinary residual defined as

$$\mathbf{r}_i = \mathbf{y}_i - \mathbf{X}_i\boldsymbol{\alpha}. \tag{2.2}$$

It is often more natural to think in the context of the growth curve model, and many investigators find it easier to conceptualize the model using this approach. Using the growth curve formulation appears to imply certain modeling limitations. Specifically, the dimension of $\mathbf{Z}_i$ determines not only the structure of the covariance matrix $\boldsymbol{\Sigma}_i$ but also the model for the mean vector, $\mathbf{Z}_i\mathbf{A}_i\boldsymbol{\alpha}$. To model changes in level adequately, one may be forced to make the covariance structure overly complex. A second limitation is that forcing $\mathbf{X}_i = \mathbf{Z}_i\mathbf{A}_i$ implies that the only covariates that are allowed to change over time for an individual are those that can be expressed as interactions of individual covariates with the time design variables. In most settings, only one element in each column of $\mathbf{A}_i$ will be nonzero, thus any column of $\mathbf{X}_i$ must take the form $\mathbf{Z}_i^{(j)}a_i^{(j)}$, where $\mathbf{Z}_i^{(j)}$ refers to the jth column of $\mathbf{Z}_i$, and $a_i^{(j)}$ is a suitable scalar. With the traditional growth curve formulation, individual characteristics such as smoking status or treatment group cannot vary over time. A referee has noted, however, that any model of the form (2.1) can be written as a generalized growth curve where we specify certain elements of both $\boldsymbol{\alpha}$ and $\mathbf{D}$ to be zero. Setting the $\mathbf{Z}_i^*$ of the

This content downloaded from
100.15.107.84 on Tue, 15 Dec 2020 03:11:24 UTC
All use subject to https://about.jstor.org/terms

growth curve equal to $\mathbf{Z}_i^* = (\mathbf{X}_i, \mathbf{Z}_i)$, $\boldsymbol{\alpha}^{*T} = (\boldsymbol{\alpha}^T, \mathbf{0})$, $\mathbf{A}_i = \mathbf{I}_i$, and

$$\mathbf{D}^* = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{bmatrix}$$

gives a model equivalent to (2.1).

A special case arises with complete and balanced data, where $\mathbf{Z}_i = \mathbf{Z}$ and $n_i = n$ for all i and $\mathbf{A}_i = \mathbf{I} \otimes \mathbf{a}_i^T$ for an $r \times 1$ vector $\mathbf{a}_i$, with $qr = p$ and $\otimes$ the direct (Kronecker) product. If we now let $\mathbf{Y}$ denote an $n \times m$ matrix with ith column $\mathbf{y}_i$, we can write

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\Psi}\mathbf{A} + \mathbf{R}, \tag{2.3}$$

where $\mathbf{R}$ is an $n \times m$ matrix with ith column $\mathbf{r}_i$ defined in (2.2), $\boldsymbol{\Psi}$ is a $q \times r$ matrix obtained by suitably reshaping $\boldsymbol{\alpha}$, and $\mathbf{A}$ is an $r \times m$ matrix with ith column $\mathbf{a}_i$. Thus in this case our general model is the same as the growth curve model discussed by Potthoff and Roy (1964), Khatri (1966), and Grizzle and Allen (1969), except those papers generally assume that $\mathbf{r}_i$ is $\mathbf{N}(\mathbf{0}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma}$ an arbitrary $n \times n$ covariance matrix, whereas our general model assumes that $\boldsymbol{\Sigma} = \sigma^2\mathbf{I} + \mathbf{Z}\mathbf{D}\mathbf{Z}^T$. This connection will be important in Section 4.1 on existence of closed-form solutions.

2.3 Special Covariance Structures

When $\mathbf{D}$ is assumed to be arbitrary, the implied covariance structure of $\mathbf{y}_i$, $\sigma^2\mathbf{I} + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T$, may be too restrictive. If we let $\mathbf{Z}_i$ contain only the design on time, and not the individual characteristics, the covariance matrix is seen to be a function only of the pattern of observations and not of individual characteristics. In some settings, the covariance matrix may differ for subpopulations.

Such dependence can be handled, using our general model, by allowing $\mathbf{Z}_i$ to contain not only the design on time but also interactions of the time variables with individual characteristics, and specifying that $\mathbf{D}$ be block diagonal. This is best illustrated by an example. Suppose we have a simple setting involving two treatment groups, with individuals belonging uniquely to either one group or the other. Assume that repeated observations on an individual over time may be modeled as a linear growth curve with intercept and slope depending on treatment group. Let $\mathbf{a}_i^T = (0, 1)$ or $(1, 0)$ depending on whether the ith individual is in the first or second treatment group, let $\mathbf{Z}_i^{(1)}$ be a vector of ones, and let $\mathbf{Z}_i^{(2)}$ be a vector of times of measurement for the ith individual. Writing

$$\mathbf{y}_i = \mathbf{Z}_i \otimes \mathbf{a}_i^T\boldsymbol{\alpha} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i$$

would yield the desired model for the mean, with the covariance structure not depending on treatment ($\mathbf{a}_i$). Now suppose that we write

$$\mathbf{y}_i = \mathbf{Z}_i \otimes \mathbf{a}_i^T\boldsymbol{\alpha} + \mathbf{Z}_i \otimes \mathbf{a}_i^T\mathbf{b}_i^* + \mathbf{e}_i,$$

where $\mathbf{b}_i^*$ is $N(\mathbf{0}, \mathbf{D}^*)$, with

$$\mathbf{D}^* = \begin{bmatrix} \mathbf{D}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_2 \end{bmatrix}$$

and $\mathbf{D}_1$ and $\mathbf{D}_2$ being 2×2 positive definite matrices. We now see that if $\mathbf{a}_i^T = (1, 0)$ (treatment group 1),

$$\mathbf{y}_i = \mathbf{Z}_i \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} + \mathbf{Z}_i \begin{pmatrix} b_{i1}^* \\ b_{i2}^* \end{pmatrix} + \mathbf{e}_i,$$

and for $\mathbf{a}_i^T = (0, 1)$,

$$\mathbf{y}_i = \mathbf{Z}_i \begin{pmatrix} \alpha_3 \\ \alpha_4 \end{pmatrix} + \mathbf{Z}_i \begin{pmatrix} b_{i3}^* \\ b_{i4}^* \end{pmatrix} + \mathbf{e}_i.$$

It follows that for treatment group 1,

$$\text{var}(\mathbf{y}_i) = \sigma^2\mathbf{I} + \mathbf{Z}_i\mathbf{D}_1\mathbf{Z}_i^T,$$

whereas for treatment group 2,

$$\text{var}(\mathbf{y}_i) = \sigma^2\mathbf{I} + \mathbf{Z}_i\mathbf{D}_2\mathbf{Z}_i^T.$$

The four-dimensional vector $\mathbf{b}_i^*$ is an artificial construct, since half of its components never actually enter the model. Defining $\mathbf{b}_i^*$ as such shows that only minor modifications are needed to introduce this level of complexity into the general model. The restriction that $\mathbf{D}^*$ be block diagonal is easily implemented in this model when one is using the EM algorithm by simply ensuring that the starting values satisfy this restriction. Allowing σ^2 to depend on a covariate would require modification of the basic algorithm presented in the next section.

3. COMPUTING FORMULAS FOR ML AND REML ESTIMATION

In this section we describe the computing formulas for implementing the EM algorithm to produce maximum likelihood (ML) or restricted maximum likelihood (REML) estimates of the parameters. The equations are given without proof or derivation because they appear in numerous other papers, including Laird and Ware (1982). These equations have been encoded in a FORTRAN program, originally written by N. Cook (1982a) and modified by Daniel Stram.

Let ω ($\omega = 0, 1, \ldots, \infty$) index the iterations, where $\omega = 0$ refers to the starting values (described in the next section) and $\omega = \infty$ refers to convergence. For $\boldsymbol{\alpha}$ we have

$$\boldsymbol{\alpha}^{(\omega)} = \left(\sum_{i=1}^{m} \mathbf{X}_i^T\mathbf{W}_i^{(\omega)}\mathbf{X}_i \right)^{-1} \sum_{i=1}^{m} \mathbf{X}_i^T\mathbf{W}_i^{(\omega)}\mathbf{y}_i, \tag{3.1}$$

where

$$\mathbf{W}_i^{(\omega)} = [\boldsymbol{\Sigma}_i^{(\omega)}]^{-1} \tag{3.2}$$

and

$$\boldsymbol{\Sigma}_i^{(\omega)} = \sigma^{(\omega)^2}\mathbf{I} + \mathbf{Z}_i\mathbf{D}^{(\omega)}\mathbf{Z}_i^T. \tag{3.3}$$

Also define

$$\mathbf{b}_i^{(\omega)} = \mathbf{D}^{(\omega)}\mathbf{Z}_i^T\mathbf{W}_i^{(\omega)}\mathbf{r}_i^{(\omega)}, \tag{3.4}$$

where

$$\mathbf{r}_i^{(\omega)} = \mathbf{y}_i - \mathbf{X}_i\boldsymbol{\alpha}^{(\omega)}. \tag{3.5}$$

Let $\boldsymbol{\theta}$ denote the vector that contains σ^2 and the non-

This content downloaded from 100.15.107.84 on Tue, 15 Dec 2020 03:11:24 UTC
All use subject to https://about.jstor.org/terms

redundant, unknown components of $\mathbf{D}$. Equations (3.1)–(3.5) hold for both ML and REML estimation of $\boldsymbol{\theta}$.

For ML estimation of $\boldsymbol{\theta}$ we have, in addition,

$$\sigma^{(\omega+1)^2} = \left\{ \sum_{i=1}^{m} [(\mathbf{r}_i^{(\omega)} - \mathbf{Z}_i\mathbf{b}_i^{(\omega)})^T(\mathbf{r}_i^{(\omega)} - \mathbf{Z}_i\mathbf{b}_i^{(\omega)}) + \sigma^{(\omega)^2} \operatorname{tr}(\mathbf{I} - \sigma^{(\omega)^2}\mathbf{W}_i^{(\omega)})] \right\} \Big/ N \tag{3.6}$$

and

$$\mathbf{D}^{(\omega+1)} = \sum_{i=1}^{m} [\mathbf{b}_i^{(\omega)}\mathbf{b}_i^{(\omega)T} + \mathbf{D}^{(\omega)} \times (\mathbf{I} - \mathbf{Z}_i^T\mathbf{W}_i^{(\omega)}\mathbf{Z}_i\mathbf{D}^{(\omega)})]/m, \tag{3.7}$$

where $N = \sum_{i=1}^{m} n_i$ is the total number of observations.

For REML estimation we have

$$\sigma^{(\omega+1)^2} = \left\{ \sum_{i=1}^{m} [(\mathbf{r}_i - \mathbf{Z}_i\mathbf{b}_i^{(\omega)})^T(\mathbf{r}_i - \mathbf{Z}_i\mathbf{b}_i^{(\omega)}) + \sigma^{(\omega)^2} \operatorname{tr}(\mathbf{I} - \sigma^{(\omega)^2}\mathbf{P}_i^{(\omega)})] \right\} \Big/ N \tag{3.8}$$

and

$$\mathbf{D}^{(\omega+1)} = \sum_{i=1}^{m} [\mathbf{b}_i^{(\omega)}\mathbf{b}_i^{(\omega)T} + \mathbf{D}^{(\omega)} \times (\mathbf{I} - \mathbf{Z}_i^T\mathbf{P}_i^{(\omega)}\mathbf{Z}_i\mathbf{D}^{(\omega)})]/m, \tag{3.9}$$

where

$$\mathbf{P}_i^{(\omega)} = \mathbf{W}_i^{(\omega)} \left[\mathbf{I} - \mathbf{X}_i \left(\sum_{i=1}^{m} \mathbf{X}_i^T\mathbf{W}_i^{(\omega)}\mathbf{X}_i \right)^{-1} \mathbf{X}_i^T\mathbf{W}_i^{(\omega)} \right]. \tag{3.10}$$

At convergence, we have in addition

$$\hat{\operatorname{var}}(\hat{\boldsymbol{\alpha}}) = \left(\sum_{i=1}^{m} \mathbf{X}_i^T\mathbf{W}_i^{\infty}\mathbf{X}_i \right)^{-1}$$

and

$$\hat{\operatorname{var}}(\hat{\mathbf{b}}_i - \mathbf{b}_i) = \mathbf{D}^{\infty}[\mathbf{I} - \mathbf{Z}_i^T\mathbf{P}_i^{\infty}\mathbf{Z}_i\mathbf{D}^{\infty}].$$

Note that application of these formulas requires repeated inversion of each $\boldsymbol{\Sigma}_i$ to obtain $\mathbf{W}_i$, even though $\sum_{i=1}^{m} \mathbf{X}_i^T\mathbf{W}_i\mathbf{X}_i$ needs to be inverted only once per iteration. But by using the matrix identity

$$\mathbf{W}_i = [\mathbf{I} - \mathbf{Z}_i(\sigma^2\mathbf{D}^{-1} + \mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T]/\sigma^2, \tag{3.11}$$

we see that to compute $\mathbf{W}_i$ we need only invert $\mathbf{D}$ once and then invert the matrices $\sigma^2\mathbf{D}^{-1} + \mathbf{Z}_i^T\mathbf{Z}_i$, which ordinarily have smaller dimension than the $\boldsymbol{\Sigma}_i$'s. A drawback of this approach to inverting $\boldsymbol{\Sigma}_i$ is that it requires $\mathbf{D}$ to be nonsingular, whereas $\mathbf{W}_i$ is well defined for arbitrary positive semidefinite $\mathbf{D}$ provided that $\sigma^2 > 0$. Other simplifications occur in Equations (3.7)–(3.10) if $\mathbf{D}^{-1}$ exists. A referee has pointed out that in some cases (q large relative to n_i) it may be computationally more efficient to use Cholesky decompositions than to invert $\boldsymbol{\Sigma}_i$ (Kennedy and Gentle 1980, sec. 7.4; Dongarra, Bunch, Moler, and Stewart 1979).

Certain data and model configurations may give rise to an ML or REML estimate of $\mathbf{D}$ on the boundary of the parameter space ($\hat{\mathbf{D}}$ not positive semidefinite). In this instance, convergence of the algorithm is not guaranteed. From (3.7) and (3.9) it would appear that $\mathbf{D}^{(\omega)}$ will always be positive semidefinite provided $\mathbf{D}^{(0)}$ is positive semidefinite and $\sigma^{(0)} > 0$, although $\mathbf{D}^{(\omega)}$ may approach a singular matrix in the limit. We have too little practical experience with this problem to suggest what the actual behavior of the algorithm will be in such cases.

It is instructive to see what form these equations take when $\sigma^2 = 0$ and $\mathbf{Z}_i = \mathbf{I}_{n\times n}$ or an appropriate subset of its rows. Assume that $\sigma^{(\omega)} = 0$ for any $\omega = 0, 1, \ldots$. Then from (3.4), $\mathbf{Z}_i\mathbf{b}_i^{(\omega)} = \mathbf{r}_i^{(\omega)}$, and thus from (3.6) or (3.8), $\sigma^{(\omega+1)^2} = 0$. For ML estimation of $\mathbf{D}$, we have

$$\mathbf{D}^{(\omega+1)} = \sum_{i=1}^{m} (\mathbf{b}_i^{(\omega)}\mathbf{b}_i^{(\omega)T} + \mathbf{R}_i^{(\omega)})/m, \tag{3.12}$$

where

$$\mathbf{b}_i^{(\omega)} = \mathbf{r}_i^{(\omega)} \quad \text{when } n_i = n$$
$$= \mathbf{D}^{(\omega)}\mathbf{Z}_i^T\mathbf{D}_i^{(\omega)-1}\mathbf{r}_i^{(\omega)} \quad \text{otherwise}$$

and

$$\mathbf{R}_i^{(\omega)} = \mathbf{0} \quad \text{when } n_i = n$$
$$= \mathbf{D}^{(\omega)}(\mathbf{I} - \mathbf{Z}_i^T\mathbf{D}_i^{(\omega)^{-1}}\mathbf{Z}_i\mathbf{D}^{(\omega)}) \quad \text{otherwise.}$$

If $n_i = n$ for all i (no missing data), the algorithm reduces to IRLS:

$$\mathbf{D}^{(\omega+1)} = \sum_{i=1}^{m} (\mathbf{r}_i^{(\omega)}\mathbf{r}_i^{(\omega)T})/m.$$

For REML estimation of $\mathbf{D}$, we have

$$\mathbf{D}^{(\omega+1)} = \sum_{i=1}^{m} (\mathbf{b}_i^{(\omega)}\mathbf{b}_i^{(\omega)T} + \mathbf{R}_i^{(\omega)} + \mathbf{T}_i^{(\omega)})/m, \tag{3.13}$$

where $\mathbf{R}_i^{(\omega)}$ is as before, and

$$\mathbf{T}_i^{(\omega)} = \mathbf{D}^{(\omega)}\mathbf{Z}_i^T\mathbf{D}_i^{(\omega)-1}\mathbf{X}_i \times \left(\sum_{i=1}^{m} \mathbf{X}_i^T\mathbf{D}_i^{(\omega)-1}\mathbf{X}_i \right)^{-1} \mathbf{X}_i^T\mathbf{D}_i^{(\omega)-1}\mathbf{Z}_i\mathbf{D}^{(\omega)}.$$

Note that if $\mathbf{Z}_i = \mathbf{I}$, $\mathbf{D}_i^{(\omega)} = \mathbf{D}^{(\omega)}$, and $\mathbf{R}_i^{(\omega)} = \mathbf{0}$ as before, then

$$\mathbf{T}_i^{(\omega)} = \mathbf{X}_i \left(\sum_{i=1}^{m} \mathbf{X}_i^T\mathbf{D}^{(\omega)-1}\mathbf{X}_i \right)^{-1} \mathbf{X}_i^T,$$

so the REML version of IRLS (all $n_i = n$) becomes

$$\mathbf{D}^{(\omega+1)} = \sum_{i=1}^{m} \left[\mathbf{r}_i^{(\omega)}\mathbf{r}_i^{(\omega)T} + \mathbf{X}_i \left(\sum_{i=1}^{m} \mathbf{X}_i^T\mathbf{D}^{(\omega)-1}\mathbf{X}_i \right)^{-1} \mathbf{X}_i^T \right] \Big/ m.$$

Formulas similar to (3.12) and (3.13) were given by Jennrich and Schluchter (1985), which are applicable when $\mathbf{D}$ is unstructured. If $\mathbf{D}$ has a patterned structure, the

This content downloaded from 100.15.107.84 on Tue, 15 Dec 2020 03:11:24 UTC
All use subject to https://about.jstor.org/terms

general form of the M step must be modified and (3.12) and (3.13) no longer hold. For those cases in which the M step has no closed form, Jennrich and Schluchter suggested replacing the M step with one "scoring step," which yields a generalized EM (GEM) algorithm.

For checking convergence of the algorithm, it is also useful to have expressions for the log-likelihoods. For ML we have

$$L_{\mathrm{ML}}(\boldsymbol{\alpha}, \boldsymbol{\theta}) = \sum_{i=1}^{m} (\ln|\mathbf{W}_i| - \mathbf{r}_i^T\mathbf{W}_i\mathbf{r}_i)/2,$$

and for REML we have

$$L_{\mathrm{REML}}(\boldsymbol{\theta}) = \left[\sum_{i=1}^{m} [\ln|\mathbf{W}_i| - \mathbf{r}_i^{*T}\mathbf{W}_i\mathbf{r}_i^*] - \ln\left|\sum_{i=1}^{m} \mathbf{X}_i^T\mathbf{W}_i\mathbf{X}_i\right|\right]\Big/2,$$

where

$$\mathbf{r}_i^* = \mathbf{y}_i - \mathbf{X}_i\left(\sum_{i=1}^{m} \mathbf{X}_i^T\mathbf{W}_i\mathbf{X}_i\right)^{-1}\sum_{i=1}^{m} \mathbf{X}_i^T\mathbf{W}_i\mathbf{y}_i.$$

4. EXISTENCE OF CLOSED-FORM SOLUTIONS FOR BALANCED DATA SETTINGS AND COMPUTATION OF STARTING VALUES

For complete and balanced data, that is, when $n_i = n$ and $\mathbf{Z}_i = \mathbf{Z}$ for all i, a sufficient condition for the existence of closed-form ML and REML estimates for all parameters $(\boldsymbol{\alpha}, \boldsymbol{\theta})$ is that the general model (2.1) take the more restrictive growth-curve form (2.3). Szatrowski and Miller (1980) gave examples of certain mixed ANOVA models for which this condition is not necessary; however, their examples cannot be specified in the model (2.1) form, because they include random effects that are not indexed by individual units or by levels of metameter variables, or interactions between units and metameter variables.

In this section, we give noniterative expressions for both ML and REML estimators of $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ in the growth-curve formulation. We also discuss two simple generalizations of the growth-curve model that also yield closed-form estimates. Finally, we give expressions for starting values of the covariance component estimates to be used in iterative computations for unbalanced data settings, where closed-form estimates do not exist.

4.1 Closed-Form Estimates for $\boldsymbol{\alpha}$

For the growth-curve model (2.3) of Section 2.2, in which $\mathbf{X}_i = \mathbf{Z} \otimes \mathbf{a}_i^T$ for all i, the generalized least squares estimator of location (3.1) reduces to the ordinary least squares (OLS) estimator, namely,

$$\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\alpha}}_{\mathrm{OLS}} = \left[\sum_{i=1}^{m} \mathbf{X}_i^T\mathbf{X}_i\right]^{-1}\sum_{i=1}^{m} \mathbf{X}_i^T\mathbf{y}_i, \tag{4.1}$$

regardless of the forms of the $\mathbf{Z}$ and $\mathbf{D}$ matrices. To see this, reshape $\boldsymbol{\alpha}_{p\times 1}$ of model (2.1) into the $\boldsymbol{\Psi}_{q\times r}$ of the growth-curve model (2.3) so that

$$\hat{\boldsymbol{\alpha}} = \mathrm{vec}(\hat{\boldsymbol{\Psi}}^T), \tag{4.2}$$

where $\mathrm{vec}(\hat{\boldsymbol{\Psi}}^T)$ stacks the r columns of $\hat{\boldsymbol{\Psi}}^T$ beneath each other to form a $p \times 1$ vector. The ML estimate of $\boldsymbol{\Psi}$ under arbitrary covariance structure (Khatri 1966; Grizzle and Allen 1969) is

$$\hat{\boldsymbol{\Psi}}^* = (\mathbf{Z}^T\mathbf{S}^{-1}\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{S}^{-1}\mathbf{Y}\mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1},$$

where

$$\mathbf{S} = \mathbf{Y}(\mathbf{I} - \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A})\mathbf{Y}^T.$$

One can apply Grizzle and Allen's (1969) approach directly to the case in which $\boldsymbol{\Sigma} = \sigma^2\mathbf{I} + \mathbf{Z}\mathbf{D}\mathbf{Z}^T$, to show that

$$\hat{\boldsymbol{\Psi}} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{Y}\mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}. \tag{4.3}$$

Straightforward matrix algebra is used to show that (4.3) is equivalent to (4.1) and (4.2) in the balanced growth-curve model. The preceding results appear in the statistical literature in various forms and are established in a variety of ways (Anderson 1971, theorem 10.2.1; Szatrowski 1980, theorem 1; Reinsel 1982, 1984; Azzalini 1985).

4.2 Closed-Form Expressions for the Covariance Components

For the growth-curve model (2.3), noniterative ML estimates of σ^2 and $\mathbf{D}$ are

$$\hat{\sigma}^2_{\mathrm{ML}} = \mathrm{tr}(\mathbf{Y}^T\mathbf{M}_Z\mathbf{Y})/(m(n - q)) \tag{4.4}$$

and

$$\hat{\mathbf{D}}_{\mathrm{ML}} = m^{-1}\mathbf{C}_Z\mathbf{Y}\mathbf{M}_A\mathbf{Y}^T\mathbf{C}_Z^T - \hat{\sigma}^2_{\mathrm{ML}}(\mathbf{Z}^T\mathbf{Z})^{-1}, \tag{4.5}$$

where

$$\mathbf{M}_Z = \mathbf{I}_n - \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T,$$

$$\mathbf{M}_A = \mathbf{I}_m - \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A},$$

and

$$\mathbf{C}_Z = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T.$$

Expressions (4.4) and (4.5) are perhaps most simply verified by referring to Azzalini (1985), who studied a growth-curve model more general than (2.3). For REML estimation, one can extend Azzalini's approach to show that $\hat{\sigma}^2_{\mathrm{REML}} = \hat{\sigma}^2_{\mathrm{ML}}$ and that $\hat{\mathbf{D}}_{\mathrm{REML}}$ is $\hat{\mathbf{D}}_{\mathrm{ML}}$ with m^{-1} replaced by $(m - r)^{-1}$ as one might anticipate. Reinsel (1982, 1985) gave expressions for unbiased estimates of σ^2 and $\mathbf{D}$ that are identical to the REML estimates given here. Equations (4.4) and (4.5) only yield ML (or REML) estimates if $\hat{\mathbf{D}}$ is positive semidefinite.

4.3 Two Simple Generalizations of the Growth-Curve Model

The requirement that all individuals share a common $\mathbf{Z}$ matrix in the growth-curve formulation may be overly restrictive in certain data-analytic settings. We can generalize (2.3) by defining different intraindividual design matrices $\mathbf{Z}^*$ and $\mathbf{Z}$ for the mean and covariance structures, re-

This content downloaded from 100.15.107.84 on Tue, 15 Dec 2020 03:11:24 UTC
All use subject to https://about.jstor.org/terms

spectively, where

$$\underset{n\times(q+s)}{\mathbf{Z}^*} = [\underset{n\times q}{\mathbf{Z}} : \underset{n\times s}{\mathbf{Z}_S}],$$

such that $\mathbf{Z}^T\mathbf{Z}_s = \mathbf{0}$. Another generalization of the growth-curve model accommodates special covariance structures, such as those described in Section 2.3. Under such generalizations, closed-form ML and REML estimates of all parameters exist and are similar to those for the standard growth-curve model with random effects given previously. For more detail on growth curve modeling with patterned covariance matrices, see Lange and Laird (1986).

4.4 Starting Values for Iterative Computations

An iterative scheme must be employed when closed-form estimators for the covariance components do not exist, and such a procedure must begin with initial estimates of σ^2 and $\mathbf{D}$. Criteria for good starting values are (a) initial estimates can be obtained under all configurations of data and models, and (b) if closed-form expressions for $\hat{\sigma}^2$ and $\hat{\mathbf{D}}$ exist, the method of obtaining starting values should find them. In case (b), our empirical evidence has shown that the EM method of iteration does not converge in one cycle to the MLE's of σ^2 and $\mathbf{D}$ from any allowable starting values, yet such a property has not been formally demonstrated. The lack of such a feature in the EM approach makes the choice of initial estimates all the more important.

The choice of starting values is determined by the form of the model and the characteristics of the data. Following Cook (1982b), starting values for σ^2 and $\mathbf{D}$ under the general model (2.1) can be computed from OLS estimates of $\boldsymbol{\alpha}$ and $\mathbf{b}_i$ as

$$\hat{\sigma}_0^2 = \left(\sum_{i=1}^m \mathbf{y}_i^T\mathbf{y}_i - \hat{\boldsymbol{\alpha}}_0^T\sum_{i=1}^m \mathbf{X}_i^T\mathbf{y}_i - \sum_{i=1}^m \hat{\mathbf{b}}_i^T\mathbf{Z}_i^T(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\alpha}}_0)\right)\Big/(N - (m-1)q - p)$$

and

$$\hat{\mathbf{D}}_0 = \sum_{i=1}^m \hat{\mathbf{b}}_i\hat{\mathbf{b}}_i^T/m - \hat{\sigma}_0^2\sum_{i=1}^m (\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}/m,$$

where $\hat{\boldsymbol{\alpha}}_0$ is from (4.1) and

$$\hat{\mathbf{b}}_i = (\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\alpha}}_0).$$

For the growth-curve model of Section 2.2, where each $\mathbf{Z}_i$ is of full rank, these values can be improved. From Cook (1982b), we may let

$$\hat{\mathbf{b}}_i = (\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T\mathbf{y}_i$$

and begin iterations with

$$\hat{\sigma}_0^2 = \left(\sum_{i=1}^m \mathbf{y}_i^T\mathbf{y}_i - \sum_{i=1}^m \hat{\mathbf{b}}_i^T\mathbf{Z}_i^T\mathbf{y}_i\right)\Big/(N - mq)$$

and

$$\hat{\mathbf{D}}_0 = \left[\sum_{i=1}^m \hat{\mathbf{b}}_i\hat{\mathbf{b}}_i^T - \left(\sum_{i=1}^m \hat{\mathbf{b}}_i\right)\left(\sum_{i=1}^m \hat{\mathbf{b}}_i\right)^T\Big/m\right]\Big/(m-1) - \hat{\sigma}_0^2\sum_{i=1}^m (\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}/m. \quad (4.6)$$

Notice that all of the preceding starting values require $\mathbf{Z}_i^T\mathbf{Z}_i$ of full rank. Whenever $\mathbf{Z}_i^T\mathbf{Z}_i$ is singular, as it would be when $n_i < q$ for some i, we may exclude such individuals from the computation of $\hat{\sigma}_0^2$ and $\hat{\mathbf{D}}_0$. Reinsel (1985) gave similar unbiased estimates for the growth-curve model, which could also be used. In any case, one should check for $\hat{\mathbf{D}}_0$ positive semidefinite before beginning the iterations.

When fitting models with special, group-dependent covariance structures such as those described in Section 2.3, we may specify starting values for the block diagonal $\mathbf{D}^*$ by using (4.6) applied to each group. From (3.7) or (3.9) and the construction of the $\mathbf{b}_i^*$ in Section 2.3, it is clear that elements of $\mathbf{D}^*$ not in its diagonal blocks will remain set to 0 in subsequent iterations.

5. SPEEDING CONVERGENCE OF THE ALGORITHM

5.1 Characterizing Convergence

A common criticism of the use of the EM algorithm in many settings, not just in variance component estimation, is that it can be extremely slow to converge, even when other methods such as Newton–Raphson or Fisher's scoring converge rapidly. The reason for this is that the EM algorithm is a first-order successive substitution method and will exhibit linear convergence at the end of the iterations. For either ML or REML, we may write

$$\boldsymbol{\theta}^{(\omega)} = \mathbf{g}(\boldsymbol{\theta}^{(\omega-1)}) \quad (5.1)$$

for the appropriate mapping $\mathbf{g}$. Using the first term of a Taylor series expansion of $\mathbf{g}$ we have

$$\boldsymbol{\theta}^{(\omega+1)} - \boldsymbol{\theta}^{(\omega)} = \mathbf{g}(\boldsymbol{\theta}^{(\omega)}) - \mathbf{g}(\boldsymbol{\theta}^{(\omega-1)}) \doteq \mathbf{J}^{(\omega-1)}(\boldsymbol{\theta}^{(\omega)} - \boldsymbol{\theta}^{(\omega-1)}),$$

where $\mathbf{J}^{(\omega-1)}$ is the matrix of partial derivatives $\mathbf{J} = \partial\mathbf{g}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ evaluated at $\boldsymbol{\theta}^{(\omega-1)}$. For ω large enough we will have $\mathbf{J}^{(\omega)} \doteq \mathbf{J}^\infty$ and thus

$$\boldsymbol{\theta}^{(\omega+1)} - \boldsymbol{\theta}^{(\omega)} \doteq \mathbf{J}^\infty(\boldsymbol{\theta}^{(\omega)} - \boldsymbol{\theta}^{(\omega-1)}),$$

where $\mathbf{J}^\infty$ is $\mathbf{J}$ evaluated at the limiting $\boldsymbol{\theta}^\infty$. Thus $\mathbf{J}^\infty$ determines the rate of convergence of the algorithm near a limit point.

In the exponential families setting,

$$\mathbf{J}^\infty = [\mathrm{var}(\mathbf{t} \mid \boldsymbol{\theta}^\infty)]^{-1}\mathrm{var}(\mathbf{t} \mid \boldsymbol{\theta}^\infty, \mathbf{y}),$$

where $\mathbf{t}$ is the complete data vector of sufficient statistics and $\mathbf{y}$ is the observed data (Dempster, Laird, and Rubin 1977). In our setting $\mathbf{t}$ is composed of quadratic forms in

This content downloaded from 100.15.107.84 on Tue, 15 Dec 2020 03:11:24 UTC
All use subject to https://about.jstor.org/terms

the residual vectors $\mathbf{b}_i$ and $\mathbf{e}_i$ ($i = 1, \ldots, m$). It follows that the eigenvalues of $\mathbf{J}^\infty$ are all between 0 and 1. These eigenvalues may be interpreted as fractions of missing information, since var($\mathbf{t} \mid \boldsymbol{\theta}^\infty$) is the information about $\boldsymbol{\theta}$ in the complete data vector (assuming that $\boldsymbol{\theta}$ is the natural parameter), and var($\mathbf{t} \mid \boldsymbol{\theta}^\infty, \mathbf{y}$) may be interpreted as the information in the unobserved or missing data vector. Then $\mathbf{J}^\infty = \mathbf{0}$ would imply no missing information and a supralinear convergence rate. This will happen in the general multivariate setting when all individuals have complete data ($\sigma^2 = 0$ and $\mathbf{Z}_i = \mathbf{I}_{n\times n}$).

Further iterations produce differences in the parameter estimates iteratively as

$$\boldsymbol{\theta}^{(k+\omega+1)} - \boldsymbol{\theta}^{(k+\omega)} \doteq (\mathbf{J}^\infty)^{k+1}(\boldsymbol{\theta}^{(\omega)} - \boldsymbol{\theta}^{(\omega-1)}). \quad (5.2)$$

But this implies (e.g., see Gerald 1970, p. 182) that the left side of (5.2) will approach the eigenvector associated with λ, the largest eigenvalue of $\mathbf{J}^\infty$ (so long as λ is distinct). In the limit then, λ will dominate the convergence. A λ near 1 implies very slow convergence; a λ near 0 implies nearly quadratic convergence.

5.2 Speeding Convergence

Any linearly convergent successive substitution algorithm can be accelerated by using multivariate forms of the Aitken acceleration method (Gerald 1970). The basic idea is to employ an estimate either of $\mathbf{J}^\infty$ or of λ to change the convergence behavior of the EM algorithm from linear to quadratic.

It is useful to monitor the convergence of the EM algorithm by estimating λ in the course of the iterations. One reasonable estimate of λ is

$$\hat{\lambda} = \sum_{i=1}^{s} (\theta_i^{(\omega)} - \theta_i^{(\omega-1)})/s(\theta_i^{(\omega-1)} - \theta_i^{(\omega-2)}), \quad (5.3)$$

where s is the number of components of $\boldsymbol{\theta}$. This is the mean of the ratios of the differences of the individual parameter estimates obtained in the two most recent iterations. If all of the parameter changes are approximately proportional, that is, if

$$(\boldsymbol{\theta}^{(\omega)} - \boldsymbol{\theta}^{(\omega-1)}) \doteq \hat{\lambda}(\boldsymbol{\theta}^{(\omega-1)} - \boldsymbol{\theta}^{(\omega-2)})$$

for $i = 1, \ldots, s$, and if $\hat{\lambda}$ is between 0 and 1, then it is appropriate to use $\hat{\lambda}$ to speed convergence. From (5.2) we can write

$$\boldsymbol{\theta}^\infty - \boldsymbol{\theta}^{(\omega-1)} = \boldsymbol{\theta}^{(\omega)} - \boldsymbol{\theta}^{(\omega-1)} + \boldsymbol{\theta}^{(\omega+1)} - \boldsymbol{\theta}^{(\omega)} + \cdots$$

$$\doteq \sum_{k=0}^{\infty} \lambda^k(\boldsymbol{\theta}^{(\omega)} - \boldsymbol{\theta}^{(\omega-1)}) = 1/(1-\lambda)(\boldsymbol{\theta}^{(\omega)} - \boldsymbol{\theta}^{(\omega-1)}).$$

Thus we can approximate $\boldsymbol{\theta}^\infty$ by

$$\hat{\boldsymbol{\theta}}^\infty = \boldsymbol{\theta}^{(\omega-1)} + 1/(1-\hat{\lambda})(\boldsymbol{\theta}^{(\omega)} - \boldsymbol{\theta}^{(\omega-1)}). \quad (5.4)$$

The expression for $\hat{\boldsymbol{\theta}}^\infty$ (5.4) could then be used instead of $\boldsymbol{\theta}^{(\omega+1)}$ in further iterations. Of course, it would be advisable to check that $\hat{\boldsymbol{\theta}}^\infty$ actually increases the likelihood over $\boldsymbol{\theta}^{(\omega)}$. This is similar to applying a univariate Aitken acceleration to each of the parameters being estimated.

Another approach to speeding up the algorithm is to estimate $\mathbf{J}^\infty$ rather than λ and use a multivariate generalization of the Aitken acceleration procedure. Following the same logic, we have

$$\boldsymbol{\theta}^\infty \doteq \boldsymbol{\theta}^{(\omega-1)} + \left\{\sum_{k=0}^{\infty} (\mathbf{J}^\infty)^k\right\}(\boldsymbol{\theta}^{(\omega)} - \boldsymbol{\theta}^{(\omega-1)}).$$

Since $\mathbf{J}^\infty$ has all of its eigenvalues between 0 and 1, the power series converges to $(\mathbf{I} - \mathbf{J}^\infty)^{-1}$. Thus to speed convergence we may try

$$\hat{\boldsymbol{\theta}}^\infty = \boldsymbol{\theta}^{(\omega-1)} + (\mathbf{I} - \hat{\mathbf{J}}^\infty)^{-1}(\boldsymbol{\theta}^{(\omega)} - \boldsymbol{\theta}^{(\omega-1)}). \quad (5.5)$$

After checking that $\hat{\boldsymbol{\theta}}^\infty$ increases the likelihood over $\boldsymbol{\theta}^{(\omega)}$, we replace $\boldsymbol{\theta}^{(\omega+1)}$ with $\hat{\boldsymbol{\theta}}^{(\infty)}$.

How then does one estimate $\mathbf{J}^\infty$? One could of course estimate $\mathbf{J}^\infty$ by $\mathbf{J}^{(\omega)}$. For ML estimation, it is not too hard to give explicit formulas for $\mathbf{J}^{(\omega)}$, either by directly differentiating the updating formulas presented in Equations (3.6)–(3.9) or the formulas in Dempster et al. (1977), or by using methods discussed by Louis (1982). These calculations, however, would seem to get unbearably messy for REML estimation. It is, nevertheless, not generally necessary to know the form of $\mathbf{J}^{(\omega)}$ to attempt the speedup. We can instead approximate $\mathbf{J}^\infty$ from the past history of the iterations themselves. Thus for $\omega > s$ we can approximate $\mathbf{J}^{(\omega)}$ as

$$\mathbf{J} = \boldsymbol{\theta}_s^\omega[\boldsymbol{\theta}_s^{\omega-1}]^{-1}, \quad (5.6)$$

where $\boldsymbol{\theta}_s^\omega$ is an $s \times s$ matrix of form

$$[\boldsymbol{\theta}^{(\omega)} - \boldsymbol{\theta}^{(\omega-1)} \mid \boldsymbol{\theta}^{(\omega-1)} - \boldsymbol{\theta}^{(\omega-2)} \mid \cdots \mid \boldsymbol{\theta}^{(\omega-s+1)} - \boldsymbol{\theta}^{(\omega-s)}].$$

As ω approaches ∞ this procedure becomes numerically unstable because

$$(\boldsymbol{\theta}^{(\omega-1)} - \boldsymbol{\theta}^{(\omega-2)}) \doteq \lambda(\boldsymbol{\theta}^{(\omega-2)} - \boldsymbol{\theta}^{(\omega-3)}),$$

and the inverse of $\boldsymbol{\theta}_s^{(\omega-1)}$ no longer exists. Of course, when this occurs we can simply switch to the λ method to accomplish the same thing.

5.3 Examples

We illustrate these acceleration techniques on two data sets. The first set comes from a longitudinal experiment in energy conservation (Stram, Laird, and Ware 1985) and the second comes from a longitudinal study of low lead exposures in infants (Waternaux, Laird, and Ware 1985). In the first example, $m = 138$, $p = 8$, $q = 4$, and n_i varies from 18 to 24. When individual growth curves were fitted to each experimental unit the R^2's were uniformly high, between .8 and .98, suggesting relatively small residual errors $\mathbf{e}_i$.

Figure 1 shows plots of several of the variance component estimates, against iteration number, calculated for the energy conservation data. For illustrative purposes, extremely poor initial values for $\mathbf{D}$ and σ^2 were used here. After six iterations, we calculated λ and the RMSE of the summands of (5.3) as equal to .2204 and .0265, respectively. Since λ was relatively close to 0 with a small het-

This content downloaded from
100.15.107.84 on Tue, 15 Dec 2020 03:11:24 UTC
All use subject to https://about.jstor.org/terms

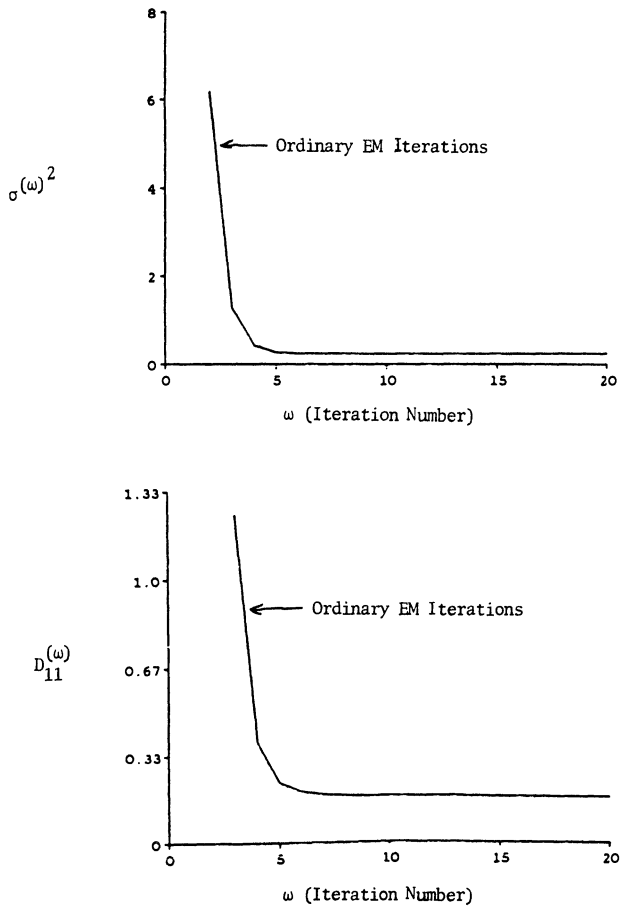

Figure 1. Plots of Variance Component Estimates Against Iteration Number for the Energy Conservation Example. The y axis is the value of the current estimate, and the x axis is the iteration number.

erogeneity over the s summands of (5.3), we expect at this point in the iterations that the EM will converge readily, as seen in Figure 1. We could apply (5.4) at this point with $\omega = 6$, although it is probably unnecessary because λ is so small.

The second example has $m = 214$ and $n_i = 3$ for most units, with $n_i = 2$ or 1 for about 15% of the units. To illustrate the point that convergence depends on parameterization, we use two models for the variance structure. First, we assume that $\sigma^2 > 0$ and $\mathbf{Z}_i$ is an $n_i \times 2$ matrix whose first column is a column of ones and the second is a column of ages at observation (there are three possible ages). Thus $q = 2$, $s = 4$, and $p = 5$ for this parameterization.

In contrast to the first example, when individual linear growth curves are fitted for units with $n_i = 3$, the R^2's (adjusted for degrees of freedom) are generally low, between .05 and .10. The convergence is very slow for these data. Figure 2 gives plots of the variance components versus iteration number. We notice that the first few iterations, starting from fairly poor initial values, produced large step sizes. In the later iterations the algorithm was very slow to approach its final values. Even after more than 100 iterations the variance component estimates continued to change in the third decimal place. After six iterations of the EM on these data we estimated $\mathbf{J}$ using (5.6) as

$$\hat{\mathbf{J}} = \begin{bmatrix} .7607 & 2.7226 & 1.3997 & -4.1169 \\ -.0178 & 1.7790 & .3019 & -1.3840 \\ -.4552 & -6.0342 & 2.5391 & 8.0242 \\ -.1122 & -.5491 & -.5458 & 1.4118 \end{bmatrix}.$$

The largest eigenvalue of this matrix equals .899, which corresponds well with the slow convergence of the estimates observed in Figure 2. The use of the λ method at iteration 6 seemed inappropriate, however, because the summands in (5.3), namely

$$(\theta_i^{(6)} - \theta_i^{(5)})/(\theta_i^{(5)} - \theta_i^{(4)}), \qquad i = 1, \ldots, 4,$$

varied from 2.75 to .09, indicating that $(\boldsymbol{\theta}^{(5)} - \boldsymbol{\theta}^{(4)})$ was nowhere near an eigenvector of $\mathbf{J}^{\infty}$. Nevertheless, good results were obtained for these data by the use of the multivariate Aitken acceleration method (5.5) when this procedure was applied at the 6th, 12th, and 18th iterations. The results are shown in Figure 2 as the line on the plots beginning at iteration 7.

For our second parameterization of this problem, the mean vector $\mathbf{X}_i\boldsymbol{\alpha}$ remains the same but we assume that $\sigma^2 = 0$, $\mathbf{Z}_i = \mathbf{I}$ if $n_i = 3$, and a subset of the rows of $\mathbf{I}$ if $n_i < 3$. Now $q = 3$ and $s = 6$, and we estimate two additional

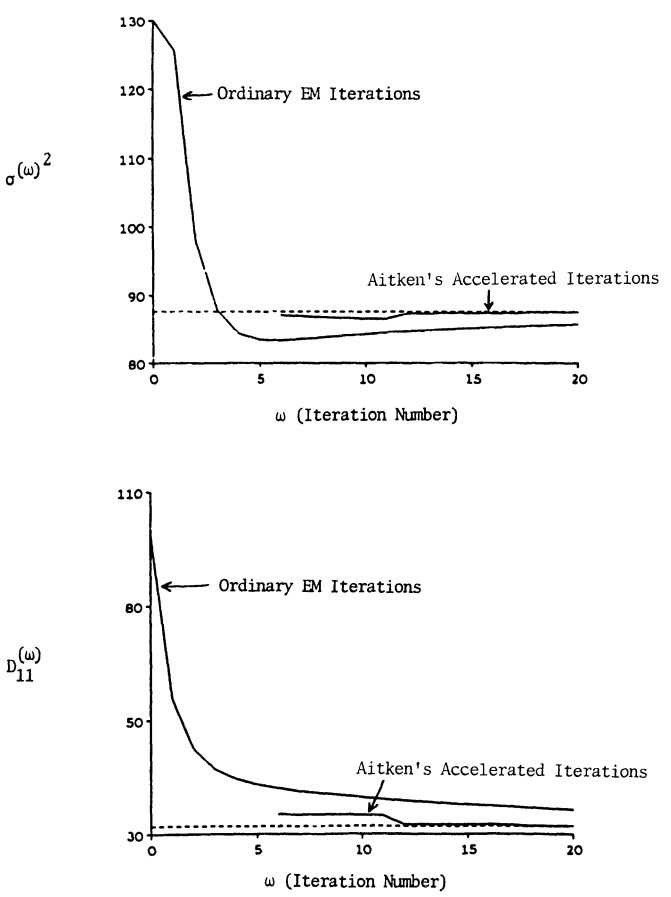

Figure 2. Plots of Variance Component Estimates Against Iteration Number for the Second Example, Using the Growth Curve Structure for the Covariance Matrix. Also shown are the results of Aitken's acceleration procedure as the shorter solid line. The dotted line denotes the final point of convergence. The y axis is the value of the current estimate, and the x axis is the iteration number.

This content downloaded from
100.15.107.84 on Tue, 15 Dec 2020 03:11:24 UTC
All use subject to https://about.jstor.org/terms

variance components. In this case, the algorithm converges in only nine iterations. After the fourth iteration, $\hat{\lambda} = .162$, and the summands in (5.13) range from .12 to .19, suggesting that $\boldsymbol{\theta}^{(4)} - \boldsymbol{\theta}^{(3)}$ is close to an eigenvector of $\mathbf{J}^{\infty}$. The proximity of $\hat{\lambda}$ to 0 indicates a relatively low fraction of missing data, and its value is very close to the fraction of individuals with $n_i < 3$. [We note that if we fail to initialize $\sigma^2 = 0$, the algorithm still converges to the same value of $\hat{\boldsymbol{\Sigma}}_i = \hat{\sigma}^2\mathbf{I}_i + \mathbf{Z}_i\hat{\mathbf{D}}\mathbf{Z}_i^T$ for some arbitrary $\hat{\sigma}^2 > 0$, but convergence is much slower, with λ depending on initial σ^2. In fact, because (3.11) requires $\sigma^2 > 0$, we set $\sigma^2 = .0001$ in the calculations. Inverting $\boldsymbol{\Sigma}_i$ directly would solve this problem.]

Finally, we refit this model to the data after deleting any individual with $n_i < 3$. Since we now have no missing data, we should have $\lambda = 0$ (IRLS). Now the algorithm converges in four iterations; at the end $\hat{\lambda} = .024$ and the summands in (5.3) ranged from .021 to .028. It is clear, then, that if the repeat observations correspond to a small set of distinct ages, the use of an arbitrary structure for $\boldsymbol{\Sigma}$ rather than a variance component structure may give large gains in terms of faster convergence, especially if the fraction of missing data is small.

Our recommendation for exploiting these extremely simple procedures for accelerating convergence is to attempt to use Aitken's acceleration method (5.5) first, but, if $\boldsymbol{\theta}_s^{(\omega-1)}$ is severely ill-conditioned, to switch to the λ method (5.4), where the largest eigenvalue λ is estimated by (5.3). In passing we note that the computational burden of these techniques is far less than that of performing an EM step and thus should always be considered as a convergence accelerator for any linearly convergent iterative algorithm.

[*Received May 1985. Revised July 1986.*]

REFERENCES

Anderson, T. W. (1971), *The Statistical Analysis of Time Series*, New York: John Wiley.

——— (1973), "Asymptotically Efficient Estimation of Covariance Matrices With Linear Structure," *The Annals of Statistics*, 1, 135–141.

Andrade, D. F., and Helms, R. W. (1984), "Maximum Likelihood Estimates in the Multivariate Normal With Patterned Mean and Covariance Via the EM Algorithm," *Communications in Statistics—Theory and Methods*, 13, 2239–2251.

Azzalini, A. (1985), "Growth Curves Analysis for Patterned Covariance Matrices," technical report, University of Padua, Italy, Dept. of Statistical Sciences.

Beale, E. M. L., and Little, R. J. A. (1975), "Missing Values in Multivariate Analysis," *Journal of the Royal Statistical Society*, Ser. B, 37, 129–145.

Cook, N. (1982a), "A General Linear Model Approach to Longitudinal Data Analysis," unpublished D.Sc. thesis, Harvard School of Public Health, Dept. of Biostatistics.

——— (1982b), "A FORTRAN Program for Random-effects Models," technical report, Harvard School of Public Health, Dept. of Biostatistics.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data Via the EM Algorithm," *Journal of the Royal Statistical Society*, Ser. B, 39, 1–38.

Dempster, A. P., Rubin, D. B., and Tsutakawa, R. K. (1981), "Estimation in Covariance Component Models," *Journal of the American Statistical Association*, 76, 341–353.

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

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

Grizzle, J. E., and Allen, D. M. (1969), "Analysis of Growth and Dose Response Curves," *Biometrics*, 25, 357–382.

Harville, D. A. (1977), "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems," *Journal of the American Statistical Association*, 72, 320–340.

Hui, S. L., and Berger, J. O. (1983), "Empirical Bayes Estimation of Rates in Longitudinal Studies," *Journal of the American Statistical Association*, 78, 753–761.

Jennrich, R. I., and Schluchter, M. D. (1985), "Unbalanced Repeated Measures Models With Structured Covariance Matrices," technical report, University of California, Los Angeles, Dept. of Biomathematics.

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

Khatri, C. G. (1966), "A Note on a MANOVA Model Applied to Problems in Growth Curve," *Annals of the Institute of Statistical Mathematics*, 18, 75–86.

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

Lange, N., and Laird, N. M. (1986), "Random-Effects and Growth-Curve Modeling for Balanced and Complete Longitudinal Data," technical report, Harvard School of Public Health, Dept. of Biostatistics.

Louis, T. A. (1982), "Finding the Observed Information Matrix When Using the EM Algorithm," *Journal of the Royal Statistical Society*, Ser. B, 44, 226–233.

Orchard, T., and Woodbury, M. A. (1972), "A Missing Information Principle: Theory and Applications," in *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1), eds. L. M. LeCam, J. Neyman, and E. L. Scott, Berkeley: University of California Press, pp. 697–715.

Potthoff, R. F., and Roy, S. N. (1964), "A Generalized Multivariate Analysis of Variance Model Useful Especially for Growth Curve Problems," *Biometrika*, 51, 313–326.

Reinsel, G. (1982), "Multivariate Repeated-Measurement or Growth Curve Models With Multivariate Random-Effects Covariance Structure," *Journal of the American Statistical Association*, 77, 190–195.

——— (1984), "Estimation and Prediction in a Multivariate Random-Effects Generalized Linear Model," *Journal of the American Statistical Association*, 79, 406–414.

——— (1985), "Mean Squared Error Properties of Empirical Bayes Estimators in a Multivariate Random Effects General Linear Model," *Journal of the American Statistical Association*, 80, 642–650.

Stram, D. O., Laird, N. M., and Ware, J. H. (1985), "An Algorithmic Approach for the Fitting of a General Mixed ANOVA Model Appropriate in Longitudinal Studies," unpublished paper presented at the Seventeenth Symposium on the Interface of Computer Science and Statistics in Lexington, Kentucky.

Szatrowski, T. H. (1980), "Necessary and Sufficient Conditions for Explicit Solutions in the Multivariate Normal Estimation Problem for Patterned Means and Covariances," *The Annals of Statistics*, 8, 802–810.

Szatrowski, T. H., and Miller, J. J. (1980), "Explicit Maximum Likelihood Estimates From Balanced Data in the Mixed-Model of the Analysis of Variance," *The Annals of Statistics*, 8, 811–819.

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

Waternaux, C., Laird, N. M., and Ware, J. H. (1985), "Methods for Analysis of Longitudinal Data: Blood Lead Concentrations and Cognitive Development," technical report, Harvard School of Public Health, Dept. of Biostatistics.

This content downloaded from
100.15.107.84 on Tue, 15 Dec 2020 03:11:24 UTC
All use subject to https://about.jstor.org/terms