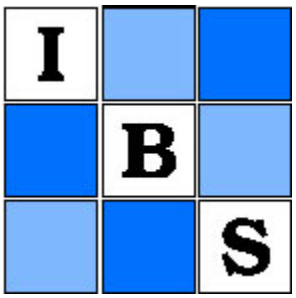

BIOMETRICS 42, 805–820
December 1986

Unbalanced Repeated-Measures Models with Structured Covariance Matrices

Robert I. Jennrich

Departments of Mathematics and Biomathematics, University of California, Los Angeles, California 90024, U.S.A.

and

Mark D. Schluchter

Department of Epidemiology and Biostatistics, University of South Carolina, Columbia, South Carolina 29208, U.S.A.

SUMMARY

The question of how to analyze unbalanced or incomplete repeated-measures data is a common problem facing analysts. We address this problem through maximum likelihood analysis using a general linear model for expected responses and arbitrary structural models for the within-subject covariances. Models that can be fit include standard univariate and multivariate models with incomplete data, random-effects models, and models with time-series and factor-analytic error structures. We describe Newton–Raphson and Fisher scoring algorithms for computing maximum likelihood estimates, and generalized EM algorithms for computing restricted and unrestricted maximum likelihood estimates. An example fitting several models to a set of growth data is included.

1. Introduction

A distinguishing characteristic of longitudinal studies and repeated-measures experiments is that each subject is observed at several different times or under different experimental conditions. Classical analytic techniques have centered around two models for the within-subject covariances: the univariate mixed model (e.g., Winer, 1971, Chap. 7) which assumes that observations from the same subject have a constant variance and a common correlation, and the general multivariate repeated-measures (Bock, 1963; Cole and Grizzle, 1966; Timm, 1980) and growth curve (Potthoff and Roy, 1964; Grizzle and Allen, 1969) models which make no assumptions about the structure of the within-subject covariance matrix.

In many cases, the classical analytic techniques cannot be used, either because some observations are missing or because the design is unbalanced for some other reason—for example, the presence of time-varying covariates. A variety of techniques for handling unbalanced repeated-measures designs are reviewed in unpublished work by Cook (paper presented at the Joint Statistical Meetings, Toronto, 1983) and Berk (1984). In addition to iterative maximum likelihood procedures, these methods include the use of unbalanced least squares (Bartlett, 1937; Rubin, 1976a; Schwertman, 1978) and generalized least squares techniques using consistent estimates of the covariance matrix obtained from incomplete data (Kleinbaum, 1973a, 1973b; LaVange and Helms, paper presented at the Joint Statistical Meetings, Toronto, 1983).

Key words: Covariance structure; EM algorithm; Growth data; Incomplete multivariate data; Newton algorithm; Repeated measures; Scoring algorithm.

A systematic approach to the analysis of incomplete and unbalanced data is to specify a model and to estimate parameters of the model using maximum likelihood or restricted maximum likelihood. The original focus of much of the work on maximum likelihood estimation with incomplete data was centered around the mixed-ANOVA model (Hartley and Rao, 1967; Hemmerle and Hartley, 1973; Jennrich and Sampson, 1976) and the multivariate normal model with unstructured covariance matrix and no within-subject design (Anderson, 1957; Trawinski and Bargman, 1964; Hartley and Hocking, 1971; Orchard and Woodbury, 1972; Beale and Little, 1975). More recent work has extended both models to allow arbitrary linear models to describe the mean structure and intermediate types of covariance structure. Maximum likelihood estimation procedures under general random-effects models for incomplete data have been studied by Laird and Ware (1982), Cook (unpublished 1983 paper), and Fairclough and Helms (unpublished paper presented at the Symposium on the Interface, Lexington, Kentucky, 1985). Ware (1985) discusses maximum likelihood estimation under a similar model with three types of covariance structures: multivariate, random-effects, and autoregressive time series; and Szatrowski (1983) and Helms (unpublished paper presented at the Joint Statistical Meetings, Philadelphia, 1984) consider models for incomplete data with linear covariance structure.

Among the major statistical software packages there are very few programs for the unbalanced model. Those we know of are restricted to models with simple between-within error structure. The GLM procedure in SAS is probably the most useful. It proceeds by blocking on subjects—that is, by treating the between component of error as an unknown constant. An ABSORB feature is used to keep the program from being overloaded with additional parameters. This makes within-subject comparisons statistically exact under appropriate assumptions but, because of confounding, destroys the possibility of making between-subject comparisons or recovering between-subject information. The BMDP program P3V can treat the between error components as random and hence can be used for both within- and between-subject comparisons and to recover between-subject information. The program was not designed for repeated-measures applications, however, and as a consequence it self-destructs when the number of subjects reaches 80 or so. To our knowledge there is nothing in SPSS, MINITAB, or GLIM that is directly applicable to the unbalanced problem. They can all block on subjects, of course, but without an ABSORB feature, this is of limited use.

In this paper, we address the problem of computing maximum likelihood estimates under a very general model where the expected values of the responses are described as arbitrary linear functions of unknown regression parameters, and the within-subject covariances are modeled as arbitrary functions of a set of unknown covariance parameters. This model is more general than the models employed by previous authors in that we allow an arbitrary parametrization of the covariance matrix. Nonstandard design features such as time-varying covariates and incomplete data are automatically incorporated into the model. The univariate mixed-model ANOVA and multivariate models are special cases of this model obtained by specifying the structural form of the within-subject covariance matrix to be compound symmetric (constant variance, constant off-diagonal) or unstructured (fully parametrized). Other interesting structures that can be fit include random-effects structures, time-series, and factor-analytic structures.

We describe three general algorithms for computing maximum likelihood estimates of the regression and covariance parameters: Newton–Raphson, Fisher scoring, and an algorithm combining scoring with generalized EM. In Section 2, we present the basic model and describe a spectrum of important covariance structures that can be fit. The three basic algorithms are described in Section 3. In Section 4, we discuss how computational efficiency can be improved through the use of specialized coding, describe how the basic algorithms can be modified in order to guarantee convergence, and also discuss problems that exist

when the maximum likelihood estimate of a covariance matrix is not positive definite. An example illustrating our methods by fitting a series of different mean and covariance structure models to a set of growth data is given in Section 5. In the last section we discuss how our algorithms are related to the work of others, and give some final remarks.

2. The Model

Let $\mathbf{y}_i$ be a $t_i \times 1$ vector containing the responses for subject i, where $i = 1, \ldots, n$. The $\mathbf{y}_i$ are assumed to follow the model

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{e}_i, \tag{1}$$

where $\mathbf{X}_i$ is a $t_i \times p$ known matrix, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown regression parameters, and the $\mathbf{e}_i$ are independently distributed as $\mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_i)$. We further assume that the elements of each matrix $\boldsymbol{\Sigma}_i$, for $i = 1, \ldots, n$, are known functions of q unknown covariance parameters contained in the vector $\boldsymbol{\theta}$. We shall write $\boldsymbol{\Sigma}_i$ as $\boldsymbol{\Sigma}_i(\boldsymbol{\theta})$ when we wish to emphasize its dependence on $\boldsymbol{\theta}$. The regression parameters $\boldsymbol{\beta}$ vary independently of the covariance parameters $\boldsymbol{\theta}$.

The ability to model the $\boldsymbol{\Sigma}_i$ allows one to examine several or many alternative structures for the $\boldsymbol{\Sigma}_i$, each structure having important subject matter interpretations. Even if interest is mainly in the regression parameters $\boldsymbol{\beta}$, efficiency of their estimates may be improved considerably by modeling the $\boldsymbol{\Sigma}_i$ parsimoniously. This is especially likely when sample sizes are small and the data are unbalanced.

The simplest form for the $\boldsymbol{\Sigma}_i$ is one that arises from independent, constant-variance observations:

$$\boldsymbol{\Sigma}_i = \sigma^2\mathbf{I}_i$$

where $\mathbf{I}_i$ is the $t_i \times t_i$ identity matrix. One may generalize this by allowing σ^2 to vary from group to group, or by allowing the variance to vary from observation to observation within subject. The latter assumes that each $\boldsymbol{\Sigma}_i$ is diagonal.

Another important special case is the *random-coefficients model*:

$$\boldsymbol{\Sigma}_i = \mathbf{Z}_i\boldsymbol{\phi}\mathbf{Z}_i' + \sigma^2\mathbf{I}_i.$$

This model arises by assuming

$$\mathbf{e}_i = \mathbf{Z}_i\mathbf{b}_i + \mathbf{u}_i,$$

where $\mathbf{Z}_i$ is a known $t_i \times k$ matrix and $\mathbf{b}_i$ and $\mathbf{u}_i$ are independent random vectors with $\mathbf{b}_i \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\phi})$ and $\mathbf{u}_i \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I}_i)$. Laird and Ware (1982) consider this model in the case where $\boldsymbol{\phi}$ is unstructured. In their unpublished 1985 paper, Fairclough and Helms consider the case where $\boldsymbol{\phi} = \boldsymbol{\phi}(\boldsymbol{\theta})$ has arbitrary linear structure. The special case where each $\mathbf{Z}_i$ is a vector of ones gives the simple between-within mixed-model ANOVA structure, called compound symmetry.

Still another important special case, which we shall call the *incomplete data model*, is obtained by assuming each $\boldsymbol{\Sigma}_i$ is actually a submatrix of a $T \times T$ matrix $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$. This model arises in situations where a fixed number T of measurements, corresponding to different times or experimental conditions, are to be collected on each of n subjects, but not all of the subjects' responses are observed.

An important extension of the incomplete data model is one that allows groups of subjects to have unequal covariance matrices. That is, for each subject i in group g, $\boldsymbol{\Sigma}_i$ is assumed to be a submatrix of a $T \times T$ matrix $\boldsymbol{\Sigma}_g(\theta)$.

In Table 1, we have listed several important structural models for $\boldsymbol{\Sigma}$ in the incomplete data model that can be fit. In addition to structures already mentioned, the table contains

Table 1
Possible covariance structures

Structure	q	Description
Independent observations	1	$\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$, $\mathbf{I} = T \times T$ identity matrix
Compound symmetry	2	$\boldsymbol{\Sigma} = \sigma_b^2\mathbf{1}\mathbf{1}' + \sigma^2\mathbf{I}$, $\mathbf{1} = T \times 1$ vector of 1's; $\mathbf{I} = T \times T$ identity matrix
Random-effects (k effects)	$\frac{1}{2}k(k+1)+1$	$\boldsymbol{\Sigma} = \mathbf{Z}\boldsymbol{\phi}\mathbf{Z}' + \sigma^2\mathbf{I}$, $\mathbf{Z} = T \times k$ known matrix; $\boldsymbol{\phi} = k \times k$ unknown dispersion matrix
Factor-analytic (k factors)	$T + kT - k(k-1)/2$[a]	$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\psi}$, $\boldsymbol{\Lambda} = T \times k$ matrix of unknown factor loadings; $\boldsymbol{\psi}$ = unknown diagonal matrix of unique variances
First-order autoregressive	2	$\sigma_{ij} = \sigma^2\rho^{\lvert i-j\rvert}$
Banded, or general autoregressive	T	$\sigma_{ij} = \theta_k$, $k = \lvert i-j\rvert + 1$
Unstructured	$\frac{1}{2}T(T+1)$	$\sigma_{11} = \theta_1$, $\sigma_{12} = \theta_2$, ..., $\sigma_{TT} = \theta_q$

[a] $k(k-1)/2$ factor loadings in Λ set equal to zero to fix the rotation.

factor-analytic and stationary time-series structures as well as the fully parametrized (unstructured) structure. Except for the first-order autoregressive and factor-analytic models, the models in Table 1 are examples of linear covariance structures. That is, $\boldsymbol{\Sigma}$ can be written as $\theta_1\mathbf{G}_1 + \cdots + \theta_q\mathbf{G}_q$ where $\mathbf{G}_1, \ldots, \mathbf{G}_q$ are known matrices and $\theta_1, \ldots, \theta_q$ are unknown parameters. Our program allows the user to specify an arbitrary linear structure by defining $\mathbf{G}_1, \ldots, \mathbf{G}_q$. Other models—for example, other ARMA models (Box and Jenkins, 1970), may be handled by providing subroutines to define the desired covariance structure and its parametric derivatives. Structures analogous to those in Table 1 exist for the $\boldsymbol{\Sigma}_i$ in the general model.

The vector $\boldsymbol{\beta}$ will range over all of Euclidean p-space. We will consider two domains for $\boldsymbol{\theta}$. In the case of Newton–Raphson and Fisher scoring algorithms below, $\boldsymbol{\theta}$ will range over all q-vectors such that all $\boldsymbol{\Sigma}_i(\boldsymbol{\theta})$ are positive definite. For the algorithms using EM in the incomplete data model, $\boldsymbol{\theta}$ will range over all q-vectors such that $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is nonnegative definite and all $\boldsymbol{\Sigma}_i(\boldsymbol{\theta})$ are positive definite. This is a more restricted set.

3. Algorithms for Maximum Likelihood

General descriptions of the algorithms we have developed are given in this section. The likelihood function and its derivatives are given in Section 3.1. The Newton–Raphson and Fisher scoring algorithms, described in Section 3.2, can be used with the general model of Section 2. The hybrid EM scoring algorithm of Section 3.3 can be used only with the incomplete data model of Section 2. A restricted maximum likelihood algorithm for this model is given in Section 3.4.

3.1 *The Likelihood Function and Its Derivatives*

The log-likelihood λ for the data $\mathbf{y}_i, \ldots, \mathbf{y}_n$ is

$$\lambda = \text{Const.} - \frac{1}{2}\sum_{i=1}^{n} \log|\boldsymbol{\Sigma}_i| - \frac{1}{2}\sum_{i=1}^{n} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'\boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}). \tag{2}$$

The score vector $\mathbf{s}$ and Hessian matrix $\mathbf{H}$ are

$$\mathbf{s} = \begin{bmatrix} \mathbf{s}_\beta \\ \mathbf{s}_\theta \end{bmatrix} = \begin{bmatrix} \partial\lambda/\partial\boldsymbol{\beta} \\ \partial\lambda/\partial\boldsymbol{\theta} \end{bmatrix} \tag{3}$$

and

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_{\beta\beta} & \mathbf{H}_{\beta\theta} \\ \mathbf{H}_{\theta\beta} & \mathbf{H}_{\theta\theta} \end{bmatrix} = \begin{bmatrix} \partial^2\lambda/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta} & \partial^2\lambda/\partial\boldsymbol{\beta}\partial\boldsymbol{\theta} \\ \partial^2\lambda/\partial\boldsymbol{\theta}\partial\boldsymbol{\beta} & \partial^2\lambda/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta} \end{bmatrix}. \tag{4}$$

Expressions for the elements of $\mathbf{s}$ and $\mathbf{H}$ are:

$$\mathbf{s}_\beta = \sum_{i=1}^{n} \mathbf{X}_i'\boldsymbol{\Sigma}_i^{-1}\mathbf{e}_i; \tag{5}$$

$$[\mathbf{s}_\theta]_r = \frac{1}{2}\sum_{i=1}^{n} \operatorname{tr} \boldsymbol{\Sigma}_i^{-1}(\mathbf{e}_i\mathbf{e}_i' - \boldsymbol{\Sigma}_i)\boldsymbol{\Sigma}_i^{-1}\dot{\boldsymbol{\Sigma}}_{ir}, \quad r = 1, \ldots, q; \tag{6}$$

$$\mathbf{H}_{\beta\beta} = -\sum_{i=1}^{n} \mathbf{X}_i'\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i; \tag{7}$$

$$[\mathbf{H}_{\beta\theta}]_{jr} = -\sum_{i=1}^{n} \mathbf{x}_{ij}'\boldsymbol{\Sigma}_i^{-1}\dot{\boldsymbol{\Sigma}}_{ir}\boldsymbol{\Sigma}_i^{-1}\mathbf{e}_i, \quad j = 1, \ldots, p, \quad r = 1, \ldots, q; \tag{8}$$

and

$$[\mathbf{H}_{\theta\theta}]_{rs} = -\frac{1}{2}\sum_{i=1}^{n} \operatorname{tr} \boldsymbol{\Sigma}_i^{-1}\dot{\boldsymbol{\Sigma}}_{ir}\boldsymbol{\Sigma}_i^{-1}(2\mathbf{e}_i\mathbf{e}_i' - \boldsymbol{\Sigma}_i)\boldsymbol{\Sigma}_i^{-1}\dot{\boldsymbol{\Sigma}}_{is} + \frac{1}{2}\sum_{i=1}^{n} \operatorname{tr} \boldsymbol{\Sigma}_i^{-1}(\mathbf{e}_i\mathbf{e}_i' - \boldsymbol{\Sigma}_i)\boldsymbol{\Sigma}_i^{-1}\ddot{\boldsymbol{\Sigma}}_{i,rs}, \quad r, s, = 1, \ldots, q; \tag{9}$$

where $\mathbf{e}_i = \mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}$, $\mathbf{x}_{ij}$ is the jth column of $\mathbf{X}_i$, $\dot{\boldsymbol{\Sigma}}_{ir} = \partial\boldsymbol{\Sigma}_i/\partial\theta_r$, and $\ddot{\boldsymbol{\Sigma}}_{i,rs} = \partial^2\boldsymbol{\Sigma}_i/\partial\theta_r\partial\theta_s$. The elements of $\dot{\boldsymbol{\Sigma}}_{ir}$ and $\ddot{\boldsymbol{\Sigma}}_{i,rs}$, respectively, are the first and second derivatives of elements of $\boldsymbol{\Sigma}_i$ with respect to $\theta_1, \ldots, \theta_q$.

3.2 *Newton–Raphson and Fisher Scoring Algorithms*

The Newton–Raphson algorithm is an iterative procedure that computes new parameter values $\tilde{\boldsymbol{\beta}}$, $\tilde{\boldsymbol{\theta}}$ from current values $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ using

$$\begin{bmatrix}\tilde{\boldsymbol{\beta}}\\ \tilde{\boldsymbol{\theta}}\end{bmatrix} = \begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{\theta}\end{bmatrix} - \begin{bmatrix}\mathbf{H}_{\beta\beta} & \mathbf{H}_{\beta\theta}\\ \mathbf{H}_{\theta\beta} & \mathbf{H}_{\theta\theta}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{s}_\beta\\ \mathbf{s}_\theta\end{bmatrix} \tag{10}$$

where the Hessian and score vector in (10) are computed using current values of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$.

The Fisher scoring algorithm replaces the Hessian matrix by its expectation. These expectations are:

$$\mathrm{E}\mathbf{H}_{\beta\beta} = -\sum_{i=1}^{n} \mathbf{X}_i'\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i; \tag{11}$$

$$\mathrm{E}\mathbf{H}_{\beta\theta} = \mathbf{0}; \tag{12}$$

$$\mathrm{E}[\mathbf{H}_{\theta\theta}]_{rs} = -\frac{1}{2}\sum_{i=1}^{n} \operatorname{tr}(\boldsymbol{\Sigma}_i^{-1}\dot{\boldsymbol{\Sigma}}_{ir}\boldsymbol{\Sigma}_i^{-1}\dot{\boldsymbol{\Sigma}}_{is}), \quad r, s = 1, \ldots, q. \tag{13}$$

Because $\mathrm{E}\mathbf{H}_{\beta\theta} = \mathbf{0}$, the updates of $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\theta}}$ are obtained by solving separate equations. The new values $\tilde{\boldsymbol{\beta}}$ are obtained by a generalized least squares step with weight matrices $\boldsymbol{\Sigma}_i^{-1}$:

$$\tilde{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \mathbf{X}_i'\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{n} \mathbf{X}_i'\boldsymbol{\Sigma}_i^{-1}\mathbf{y}_i\right). \tag{14}$$

New estimates $\tilde{\boldsymbol{\theta}}$ are then computed by a scoring step:

$$\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta} + \mathbf{I}_{\theta\theta}^{-1}\mathbf{s}_\theta, \tag{15}$$

where $\mathbf{I}_{\theta\theta}$ is the negative of the matrix $\mathrm{E}\mathbf{H}_{\theta\theta}$ defined by (13) and $\mathbf{s}_\theta$ is computed from (6) using the updated residuals $\mathbf{e}_i = \mathbf{y}_i - \mathbf{X}_i\tilde{\boldsymbol{\beta}}$. This represents a slight modification of the strict definition of the scoring algorithm in that $\mathbf{s}_\theta$ is computed using the updated value $\tilde{\boldsymbol{\beta}}$ rather than the current value $\boldsymbol{\beta}$.

3.3 *Hybrid EM Scoring Algorithm for the Incomplete Data Model*

The algorithm described in this section is restricted to use with the incomplete data model. The advantage of this algorithm over the Newton and scoring algorithms is in fitting covariance matrices with large numbers of parameters (e.g., large unstructured $\boldsymbol{\Sigma}$).

The algorithm consists of two parts. The first part [step (i) below] uses generalized least squares to update $\boldsymbol{\beta}$, which is equivalent to maximizing the likelihood with respect to $\boldsymbol{\beta}$ while holding $\boldsymbol{\theta}$ fixed. The second part [steps (ii)–(iv) below] comprise one iteration of a

generalized EM (GEM) algorithm (Dempster, Laird, and Rubin, 1977) to update $\boldsymbol{\theta}$, using $\mathbf{e}_i^* \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$ as complete data and subvectors $\mathbf{e}_i = \mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}$ as observed data, and assuming $\boldsymbol{\beta}$ is equal to its updated value. The steps of the algorithm are as follows:

(i) Compute updated estimates $\tilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ using equation (14).

(ii) Using the updated estimates of the parameters $\tilde{\boldsymbol{\beta}}$ and the current estimates of the parameters $\boldsymbol{\theta}$, compute $\hat{\boldsymbol{e}}_i^*$ and $\mathbf{R}_i$, where

$$\hat{\mathbf{e}}_i^* = \mathrm{E}(\mathbf{e}_i^* \mid \mathbf{e}_i) \tag{16}$$

and

$$\mathbf{R}_i = \mathrm{cov}(\mathbf{e}_i^* \mid \mathbf{e}_i). \tag{17}$$

For example, if

$$\mathbf{e}_i^* = \begin{bmatrix} e_i \\ e_i+ \end{bmatrix},$$

where $\mathbf{e}_i = \mathbf{y}_i - \mathbf{X}_i\tilde{\boldsymbol{\beta}}$, then

$$\mathrm{E}(\mathbf{e}_i^* \mid \mathbf{e}_i) = \mathbf{M}_i\mathbf{e}_i,$$

where

$$\mathbf{M}_i = \begin{bmatrix} \mathbf{I} \\ \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1} \end{bmatrix} \tag{18}$$

and

$$\mathrm{cov}(\mathbf{e}_i^* \mid \mathbf{e}_i) = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \end{bmatrix},$$

where $\boldsymbol{\Sigma}_{11}$, $\boldsymbol{\Sigma}_{12}$, and $\boldsymbol{\Sigma}_{22}$ are the appropriate submatrices of $\boldsymbol{\Sigma}$.

(iii) Using $\hat{\mathbf{e}}_i^*$ and $\mathbf{R}_i$, compute the matrix

$$\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\mathbf{e}}_i^*\hat{\mathbf{e}}_i^{*\prime} + \mathbf{R}_i). \tag{19}$$

(iv) Compute the update $\boldsymbol{\theta}$: When $\boldsymbol{\Sigma}$ is an unstructured dispersion matrix use the update $\tilde{\boldsymbol{\Sigma}} = \mathbf{S}$. If $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$ is structured, use a "scoring step":

$$\Delta\boldsymbol{\theta} = \mathbf{I}^{-1}\mathbf{s},$$

where

$$[\mathbf{s}]_r = \tfrac{1}{2}\,\mathrm{tr}\,\boldsymbol{\Sigma}^{-1}(\mathbf{S} - \boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{\Sigma}}_r$$

and

$$[\mathbf{I}]_{rs} = \tfrac{1}{2}\,\mathrm{tr}\,\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{\Sigma}}_r\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{\Sigma}}_s.$$

Let $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta} + \Delta\boldsymbol{\theta}$ and check to see if $h(\tilde{\boldsymbol{\theta}}) > h(\boldsymbol{\theta})$, where

$$h(\boldsymbol{\theta}) = -\log\,|\,\boldsymbol{\Sigma}(\boldsymbol{\theta})\,| - \mathrm{tr}\,\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\mathbf{S}.$$

If $h(\boldsymbol{\theta})$ is not increased, use partial stepping to increase it. We use step-halving, replacing $\Delta\boldsymbol{\theta}$ by $\Delta\boldsymbol{\theta}/2$ until $h(\tilde{\boldsymbol{\theta}}) > h(\boldsymbol{\theta})$. This condition must be met to guarantee that this part of the algorithm is in fact a step of the generalized EM algorithm, which in turn guarantees that the likelihood is increased at each step.

3.4 *Restricted Maximum Likelihood Using Generalized EM*

Let $\bar{\boldsymbol{\beta}}$ be the ordinary least squares estimates of $\boldsymbol{\beta}$, given by

$$\bar{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \mathbf{X}_i'\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{n} \mathbf{X}_i'\mathbf{y}_i\right). \tag{20}$$

The restricted maximum likelihood (REML) estimates of $\boldsymbol{\theta}$ are those values that maximize the likelihood of the least squares residuals, $\bar{\mathbf{e}}_i = \mathbf{y}_i - \mathbf{X}_i\bar{\boldsymbol{\beta}}$, $i = 1, \ldots, n$. The log-likelihood of the residuals, λ_{REML}, is written as

$$\lambda_{\text{REML}} = \text{Const.} - \frac{1}{2}\sum_{i=1}^{n} \log|\boldsymbol{\Sigma}_i| - \frac{1}{2}\log\left|\sum_{i=1}^{n} \mathbf{X}_i'\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i\right| - \frac{1}{2}\sum_{i=1}^{n} \tilde{\mathbf{e}}_i'\boldsymbol{\Sigma}_i^{-1}\tilde{\mathbf{e}}_i,$$

where $\tilde{\mathbf{e}}_i = \mathbf{y}_i - \mathbf{X}_i\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ is defined by (14). Note that λ_{REML} does not involve $\boldsymbol{\beta}$. A true GEM algorithm that uses the $\mathbf{e}_i^*$ as complete data and the $\hat{\mathbf{e}}_i^*$ as observed data proceeds as follows:

(i) Compute $\tilde{\boldsymbol{\beta}}$, $\hat{\mathbf{e}}_i^*$, $\mathbf{R}_i$, $i = 1, \ldots, n$, and $\mathbf{S}$ exactly as in steps (i)–(iii) of the algorithm of Section 3.3.

(ii) Using the current value of $\boldsymbol{\theta}$, compute

$$\mathbf{S}_2 = \mathbf{S} + \frac{1}{n}\sum_{i=1}^{n} \mathbf{M}_i\mathbf{X}_i\mathbf{V}\mathbf{X}_i'\mathbf{M}_i',$$

where $\mathbf{V} = (\sum_{i=1}^{n} \mathbf{X}_i'\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i)^{-1}$ and $\mathbf{M}_i$ [equation (18)] is the matrix of regression coefficients for the regression of $\mathbf{e}_i^*$ on $\mathbf{e}_i$.

(iii) Update $\boldsymbol{\theta}$ precisely as described in step (iv) of the hybrid algorithm of Section 3.3, using $\mathbf{S}_2$ in place of $\mathbf{S}$. For example, if $\boldsymbol{\Sigma}$ is unstructured, then use the update $\tilde{\boldsymbol{\Sigma}} = \mathbf{S}_2$ and otherwise use a "scoring step." The converged value of $\boldsymbol{\theta}$ is the REML estimate, and the converged value of $\boldsymbol{\beta}$ is the generalized least squares estimate (14) with the $\boldsymbol{\Sigma}_i$ evaluated at the REML estimate of $\boldsymbol{\theta}$. One desirable property of REML estimates is that in some balanced cases they agree with familiar ANOVA estimates.

A simple method for obtaining initial parameter values, applicable with all four algorithms and any covariance structure, is to set the initial values of $\boldsymbol{\beta}$ equal to the ordinary least squares estimates (20) and to choose the initial values of $\boldsymbol{\theta}$ so that $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = s^2\mathbf{I}$, where s^2 is the residual mean square from this regression. More elaborate methods for obtaining initial values can be devised, if desired, when fitting a particular type of covariance structure.

3.5 *Standard Errors*

Standard error estimates can be computed from the inverse of the Fisher information matrix [the negative of equations (11)–(13)] or from the inverse of the empirical information matrix [the negative of equations (7)–(9)]. In either case the matrices are evaluated at the final estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. As indicated in the discussion, the use of the empirical information matrix may be preferable when used with incomplete data generated by some missing-data mechanisms. In the algorithms using EM, the information matrix for $\boldsymbol{\theta}$ is not computed, and consequently no standard errors for the elements of $\boldsymbol{\theta}$ are produced.

4. Implementation

We have attempted to resolve a number of the problems that arise when one is trying to produce more than demonstration software. These include considerations of computational efficiency, and modifications to the raw algorithms to make them ascent algorithms,

and to ensure graceful behavior when the likelihood cannot be maximized by positive definite $\boldsymbol{\Sigma}_i$.

4.1 *Steps to Improving Computational Efficiency*

The amount of computation required by each of the three algorithms increases roughly linearly with the number of subjects, n. The calculation of $\boldsymbol{\Sigma}_i^{-1}$ in the scoring and Newton–Raphson algorithms, and analogously, the calculation of the pieces of $\mathbf{S}$ and $\mathbf{S}_2$ involving $\mathbf{R}_i$ and $\mathbf{M}_i$ in the algorithms using EM, can be made more efficient by grouping subjects with equal $\boldsymbol{\Sigma}_i$. In the case of the incomplete data model it is helpful to store $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^{-1}$ and pivot or reverse pivot to obtain each $\boldsymbol{\Sigma}_i^{-1}$ using the matrix that requires the smallest number of pivots.

All four algorithms compute $\Sigma \mathbf{X}_i' \boldsymbol{\Sigma}_i^{-1} \mathbf{X}_i$, which can be expensive if either T ($T = \max t_i$ or order of $\boldsymbol{\Sigma}$) or p is large, requiring on the order of $np^2T/2 + npT^2$ operations per iteration. At times, considerable savings in computation can be realized by utilizing the fact that some design variables are constant across time, so that the corresponding column in $\mathbf{X}_i$ is a multiple of the vector $\mathbf{1} = (1, \ldots, 1)'$. The cost of computing the portion of $\mathbf{X}_i' \boldsymbol{\Sigma}_i^{-1} \mathbf{X}_i$ corresponding to constant covariates can then be reduced essentially to the cost of computing $\mathbf{1}' \boldsymbol{\Sigma}_i^{-1} \mathbf{1}$ for that subject. Dummy variables representing main effects and interactions of between-subject grouping factors are included in the category of design variables that do not change across time.

The scoring step for $\boldsymbol{\theta}$ in either of the algorithms using generalized EM is usually not a major part of the total computation; however, the computation of $\mathbf{H}_{\theta\theta}$ or $\mathbf{I}_{\theta\theta}$ in the Newton and scoring algorithms becomes extensive when either q or T is large. In particular, if $\boldsymbol{\Sigma}$ is unstructured, then $q = T(T + 1)/2$ and the direct computation of $\mathbf{I}_{\theta\theta}$ requires on the order of nT^6 operations. When $\boldsymbol{\Sigma}$ is unstructured, however, the $\dot{\boldsymbol{\Sigma}}_{ir}$, $r = 1, \ldots, q$, have a very simple form that can be exploited with special code; for example, if $\boldsymbol{\Sigma}^{-1} = [\sigma^{ij}]$ and $\dot{\boldsymbol{\Sigma}}_{rs} = \partial \boldsymbol{\Sigma} / \partial \sigma_{rs}$,

$$\operatorname{tr} \boldsymbol{\Sigma}^{-1} \dot{\boldsymbol{\Sigma}}_{rs} \boldsymbol{\Sigma}^{-1} \dot{\boldsymbol{\Sigma}}_{tu} = \begin{cases} 2(\sigma^{st}\sigma^{ru} + \sigma^{rt}\sigma^{su}) & r \neq s,\ t \neq u \\ 2\sigma^{rt}\sigma^{st} & r \neq s,\ t = u \\ (\sigma^{rt})^2 & r = s,\ t = u \end{cases}.$$

With specialized code for unstructured $\boldsymbol{\Sigma}$, the number of operations required to compute $\mathbf{I}_{\theta\theta}$ is reduced to the order nT^4. Specialized code can be written to speed computations for other special covariance structures as well. For example, when $\boldsymbol{\Sigma}$ has the compound symmetry structure, special code can be written so that the number of operations to compute $\mathbf{I}_{\theta\theta}$ is linear in n and T. The computation of $\mathbf{X}_i' \boldsymbol{\Sigma}_i^{-1} \mathbf{X}_i$ can be greatly simplified also. At present, we have written special code to calculate $\mathbf{I}_{\theta\theta}$ for the compound symmetry structure and for unstructured $\boldsymbol{\Sigma}$ in the scoring algorithm.

When the second derivatives $\ddot{\boldsymbol{\Sigma}}_{i,rs}$, $r, s = 1, \ldots, q$, are nonzero, the Newton algorithm may require substantially more computation than scoring because of the need to calculate the second term in (9). In our implementation we have ignored the second term in (9) when computing $\mathbf{H}_{\theta\theta}$. If the elements of $\boldsymbol{\Sigma}$ are linear functions of $\boldsymbol{\theta}$, as is true for all but the factor-analytic and first-order autoregressive structures in Table 1, all $\ddot{\boldsymbol{\Sigma}}_{i,rs} = 0$ and the Newton algorithm is unchanged. When the $\boldsymbol{\Sigma}_i(\boldsymbol{\theta})$ are approximately linear in $\boldsymbol{\theta}$, the modified Newton algorithm will generally converge faster than the scoring algorithm. An additional advantage of this modification is that the user does not need to specify the second derivatives of the $\boldsymbol{\Sigma}_i$ when programming a special covariance structure.

4.2 *Modifications to Improve Convergence*

Several fairly simple modifications of the three maximum likelihood algorithms can be made in order to make them ascent algorithms and ensure convergence.

These modifications ensure that the following two conditions are met: (i) For the Newton and scoring algorithms, each $\boldsymbol{\Sigma}_i$, for $i = 1, \ldots, n$, must be positive definite at each step. For the EM scoring algorithm, $\boldsymbol{\Sigma}$ must be positive definite at each step. (ii) For the Newton–Raphson and scoring algorithms, the log-likelihood λ must increase at each step. For the EM scoring algorithm, $h(\boldsymbol{\theta})$ must increase.

Each of the three algorithms involves computing an increment vector $\Delta\boldsymbol{\gamma}$ and $\mathbf{Q}^{-1}\mathbf{g}$, where $\mathbf{g}$ is the gradient of a function of $\boldsymbol{\gamma}$ whose increase guarantees the increase of the log-likelihood λ or the increase of $h(\boldsymbol{\theta})$. In the Newton algorithm $\boldsymbol{\gamma} = (\boldsymbol{\beta}, \boldsymbol{\theta})$ and in the scoring and EM scoring algorithms $\boldsymbol{\gamma} = \boldsymbol{\theta}$. During inversion, the diagonal of $\mathbf{Q}$ is enhanced, if necessary, to ensure that the resulting inverse is positive definite. With this done a small enough step in the direction of $\Delta\boldsymbol{\gamma}$ will satisfy (i) and (ii). For partial stepping, if required, we use step-halving.

A simple way to enhance the diagonal of $\mathbf{Q}$ (which is what we have implemented) is to monitor the encountered tolerance of $\mathbf{Q}$ (ratio of current pivot element p_{ii} to original diagonal q_{ii}) during its inversion by Gauss–Jordan pivoting. If the encountered tolerance drops below a threshold of 10^{-6}, we replace the current pivot element by $10^{-6}q_{ii}$ before making the pivot. Similarly, when checking condition (i), we use the operational definition that $\boldsymbol{\Sigma}$ is positive definite if the smallest tolerance encountered in its inversion exceeds 10^{-6}. All computations are in double precision.

When using the Newton algorithm, we start with one or more scoring steps and switch to Newton–Raphson when the change in the log-likelihood is less than 1. If the Hessian has low tolerance in any subsequent step, we switch to scoring for that step.

4.3 *Nonpositive Definite Estimates for* $\boldsymbol{\Sigma}$ *and* $\boldsymbol{\Sigma}_i$

A situation that can arise in the incomplete data model when there are no complete vectors $\mathbf{y}_i$ is that there are well-defined positive definite estimates for the $\boldsymbol{\Sigma}_i$, but the corresponding value of $\boldsymbol{\Sigma}$ is not positive definite. When this occurs, the Newton and scoring algorithms converge to a well-defined maximum likelihood estimate $\hat{\boldsymbol{\Sigma}}$ having one or more nonpositive eigenvalues. However, iterates of $\boldsymbol{\Sigma}$ in the EM scoring algorithm must be positive definite, so that the limiting value of $\boldsymbol{\Sigma}$ must be nonnegative definite. In theory, one should be able to detect this situation by monitoring the encountered tolerance of $\boldsymbol{\Sigma}$, which should tend to zero. However, our experiences are that the algorithm converges extremely slowly in this situation, with the encountered tolerance of $\boldsymbol{\Sigma}$ still far from our threshold of 10^{-6} after a large number of iterations. In their unpublished 1985 paper, Fairclough and Helms also report slow convergence of their EM algorithm when the estimated covariance matrix is nearly singular. Our best indicator of this situation at present is a measure of the convergence rate $r_k = (\lambda_k - \lambda_{k-1})/(\lambda_{k-1} - \lambda_{k-2})$ where λ_k is the value of λ at the kth iteration. This approaches 1 as $\boldsymbol{\Sigma}$ becomes singular.

A less common problem occurs when λ has no finite maximum. As the model was defined in Section 2, the maximum likelihood estimate does not exist when this occurs. In this situation the scoring and Newton algorithms converge toward values of $\boldsymbol{\Sigma}_i$, at least one of which is operationally singular, giving a reasonable indication of the nature of the problem. The step-halving discussed in Section 4.3 prevents each $\boldsymbol{\Sigma}_i$ from actually becoming singular. One might attempt to show that λ is approaching infinity, but this fails because the matrix inversions required to compute λ break down long before λ gets large. In this situation the limits of the $\boldsymbol{\Sigma}_i$ are not unique and may change if different initial values are used. This reflects the fact that in general in this situation many different singular $\boldsymbol{\Sigma}_i$ will make the likelihood infinite. In general, in this situation, the EM scoring algorithm will not drive λ to infinity even in theory. Its slow convergence often fails to give a clear indication that $\boldsymbol{\Sigma}$ is becoming singular. The r_k, however, again tend to approach 1, indicating at least that some problem exists.

The problems discussed in this subsection arise only under rather special circumstances. They should not be considered typical. However, they do exist, and have arisen on real data sets.

4.4 *Constraints on Covariance Parameters* $\boldsymbol{\theta}$

We have not restricted $\boldsymbol{\theta}$ beyond what is implied naturally by the algorithms employed. This can lead to negative estimates for variance components even with the two algorithms using EM. For example, in the compound symmetry model

$$\boldsymbol{\Sigma} = \sigma_b^2 \mathbf{1}\mathbf{1}' + \sigma^2 \mathbf{I},$$

the estimate of σ_b^2 will be negative with probability roughly .5 when the population value of $\sigma_b^2 = 0$. One could introduce quadratic programming code to produce restricted estimates for σ_b^2 with the natural restriction being $\hat{\sigma}_b^2 \geq 0$. While the restricted estimates have the better mean squared error properties, there are some advantages to using the unrestricted estimates. The asymptotic distribution of the unrestricted estimate is normal even when $\sigma_b^2 = 0$ and the null distribution of the likelihood ratio test of the hypothesis $\sigma_b^2 = 0$ is asymptotically χ^2 with 1 degree of freedom. Neither of these statements holds for the restricted estimates.

5. Example

Potthoff and Roy (1964) present a set of growth data for 11 girls and 16 boys. For each subject, the distance (mm) from the center of the pituitary to the pterygomaxillary fissure was recorded at the ages of 8, 10, 12, and 14. None of the data are missing. The results of fitting eight different models to these data are summarized in Table 2. This example is intended to illustrate the use of mean and covariance structures and does not represent a complete analysis of these data.

Let y_{gst} denote the response (distance) for the sth subject in sex group $g = 1, 2$, at time $t = 1, 2, 3, 4$. The first model uses a separate mean for each (g, t) pair and a common unstructured covariance $\boldsymbol{\Sigma}$ for the responses from each subject:

Model 1 $\quad y_{gst} = \beta_{gt} + e_{gst}$

where independently for each (g, s),

$$\mathbf{e}_{gs} = (e_{gs1}, \ldots, e_{gs4})' \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}).$$

The ML estimates for the β_{gt} are the cell means $\bar{y}_{g\cdot t}$ and the ML estimate of $\boldsymbol{\Sigma}$ is $(1/n)\mathbf{W}$, where $\mathbf{W}$ is the pooled within-group matrix of corrected sums of squares and cross-products and $n = 27$. The ML estimates are shown in Table 3. Plots of $\bar{y}_{g\cdot t}$ on age x_t suggest that the relation between mean distance and age is linear with possibly separate lines for girls and boys. This suggests the next model.

Model 2 $\quad y_{gst} = \alpha_g + \beta_g x_t + e_{gst}$

with, as in Model 1, an unstructured covariance matrix $\boldsymbol{\Sigma}$ for the $\mathbf{e}_{gs}$. The likelihood ratio test comparing Model 2 against Model 1 is not significant, which favors the null hypothesis that the relationship between mean distance and age is linear for each sex. To test for common slope we fit Model 3.

Model 3 $\quad y_{gst} = \alpha_g + \beta x_t + e_{gst}$

with, as before, an unstructured covariance matrix $\boldsymbol{\Sigma}$ for the $\mathbf{e}_{gs}$. The likelihood ratio test for Model 3 versus Model 1 is significant, indicating that the slopes differ.

Models 4–8 use the same model for the means as Model 2, but differ in what is assumed about the structure of $\boldsymbol{\Sigma}$.

Table 2
Summary of models fit

Model	Model for means	Model for Σ	Number of parameters $(p+q)$	-2λ	Goodness-of-fit test: Comparison model	χ^2	df
1	$E(y_{gsl}) = \beta_{gl}$	Unstructured	18	416.509			
2	$E(y_{gsl}) = \alpha_g + \beta_g X_l$	Unstructured	14	419.477	1	2.968	4
3	$E(y_{gsl}) = \alpha_g + \beta X_l$	Unstructured	13	426.153	2	6.676	1**
4	$E(y_{gsl}) = \alpha_g + \beta_g X_l$	Banded	8	424.643	2	5.166	6
5	$E(y_{gsl}) = \alpha_g + \beta_g X_l$	AR(1)	6	440.681	2	21.204	8**
6	$E(y_{gsl}) = \alpha_g + \beta_g X_l$	Random coefficients	8	427.806	2	8.329	6
7	$E(y_{gsl}) = \alpha_g + \beta_g X_l$	Compound symmetry	6	428.639	6	0.833	2
8	$E(y_{gsl}) = \alpha_g + \beta_g X_l$	$\sigma^2\mathbf{I}$	5	478.242	7	50.833	1***

** $P < .01$.
*** $P < .001$.

Table 3
Maximum likelihood estimates
(standard errors obtained from the inverse Fisher information matrix)

		Estimated age/sex-specific means				Estimates of variances and correlations				
		Boys		Girls						
Model	Age	Estimate	SE	Estimate	SE	Variance	Correlation matrix			
1	8	22.875	.560	21.182	.675	5.0143	1.0000			
	10	23.813	.492	22.227	.594	3.8748	.5707	1.0000		
	12	25.719	.611	23.091	.737	5.9775	.6613	.5632	1.0000	
	14	27.469	.537	24.091	.648	4.6164	.5216	.7262	.7281	1.0000
2	8	22.457	.489	21.236	.590	5.1192	1.0000			
	10	24.110	.453	22.189	.546	3.9280	.5443	1.0000		
	12	25.764	.471	23.142	.568	5.9798	.6526	.5607	1.0000	
	14	27.418	.536	24.095	.647	4.6180	.5188	.7190	.7276	1.0000
3	8	22.815	.480	20.770	.569	5.0857	1.0000			
	10	24.164	.453	22.119	.546	3.9528	.5623	1.0000		
	12	25.513	.466	23.468	.557	6.0605	.6421	.5416	1.0000	
	14	26.863	.517	24.818	.600	5.0492	.4754	.6521	.7249	1.0000
4	8	22.638	.530	21.217	.639	4.9438	1.0000			
	10	24.232	.475	22.168	.573	9.9438	.6171	1.0000		
	12	25.827	.475	23.120	.573	4.9438	.6888	.6171	1.0000	
	14	27.421	.530	24.072	.639	4.9438	.4737	.6888	.6171	1.0000
5	8	22.749	.543	21.192	.655	4.8908	1.0000			
	10	24.288	.435	22.159	.525	4.8908	.6071	1.0000		
	12	25.827	.435	23.127	.525	4.8908	.3686	.6071	1.0000	
	14	27.366	.543	24.094	.655	4.8908	.2238	.3686	.6071	1.0000
6	8	22.616	.507	21.209	.611	4.6216	1.0000			
	10	24.184	.467	22.168	.563	4.6839	.6210	1.0000		
	12	25.753	.483	23.127	.583	4.9363	.6014	.6336	1.0000	
	14	27.322	.551	24.086	.665	5.3788	.5729	.6226	.6586	1.0000
7	8	22.616	.521	21.209	.628	4.9052	1.0000			
	10	24.184	.474	22.168	.572	4.9052	.6178	1.0000		
	12	25.753	.474	23.127	.572	4.9052	.6178	.6178	1.0000	
	14	27.322	.521	24.086	.628	4.9052	.6178	.6178	.6178	1.0000
8	8	22.616	.463	21.209	.559	4.9052	1.0000			
	10	24.184	.303	22.168	.366	4.9052	.0000	1.0000		
	12	25.753	.303	23.127	.366	4.9052	.0000	.0000	1.0000	
	14	27.322	.463	24.086	.559	4.9052	.0000	.0000	.0000	1.0000

Model 4 $\boldsymbol{\Sigma}$ is banded.

A likelihood ratio test indicates no significant difference between this and Model 2.

Model 5 $\boldsymbol{\Sigma}$ is AR(1),

that is, $\boldsymbol{\Sigma}$ has a first-order autoregressive structure. Model 5 exhibits a significant lack of fit when compared to Model 2, which is consistent with the observation that the correlations estimated using Model 2 do not exhibit the ρ, ρ^2, ρ^3 pattern required under the AR(1) structure.

A natural structure for $\boldsymbol{\Sigma}$ is provided by the random coefficients model:

Model 6 $y_{gst} = a_{gs} + b_{gs}x_t + w_{gst}$

where independently for each (g, s),

$$(a_{gs}, b_{gs}) \sim \mathrm{N}((\alpha_g, \beta_g), \boldsymbol{\phi})$$

where $\boldsymbol{\phi}$ is an arbitrary 2×2 covariance matrix and the w_{gst} are independent $\mathrm{N}(0, \sigma^2)$. Model 6 fits the data, as judged by the likelihood ratio test comparing it with Model 2. The only estimate $\hat{\phi}_{rs}$ that appears significantly different from zero when compared with its standard error estimate is $\hat{\phi}_{11}$. This suggests the mixed model:

Model 7 $y_{gst} = a_{gs} + \beta_g x_t + w_{gst}$

where the a_{gs} are independent $\mathrm{N}(\alpha_g, \sigma_a^2)$. For this model $\boldsymbol{\Sigma}$ has compound symmetry structure. Model 7 also appears to fit the data as judged by the likelihood ratio test comparing it with Model 6. Finally, the hypothesis $\sigma_a^2 = 0$ suggests fitting the independence model:

Model 8 $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$.

This model does not appear to fit when judged by the likelihood ratio test comparing it to Model 7.

At least as judged by our model-building exercise, Models 1, 2, 4, 6, and 7 all appear to fit the data. Of these, Model 7 is the most parsimonious. It is also the standard univariate repeated-measures model with a between-subjects component a_{gs} and within-subjects component w_{gst}.

It should be clear that all eight models have the structure set forth in Section 2. In Model 6, for example, $\boldsymbol{\Sigma} = \mathbf{Z}\boldsymbol{\phi}\mathbf{Z}' + \sigma^2\mathbf{I}$, where

$$\mathbf{Z} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \end{bmatrix}.$$

Because the data are complete and balanced within the sex groups in this example, there exist noniterative solutions to Models 1, 6, 7, and 8 and the scoring algorithm converges in one iteration. The maximum likelihood estimates (MLEs) of the parameters of Models 1 and 8 are well known; Model 1 is a standard multivariate model with no structure on the mean vectors or covariance matrix, and Model 8 is a standard univariate regression model. Explicit formulas for the MLEs of mean and covariance parameters for balanced random-effects models, which include Model 6, are given by unpublished work by Laird, Lange, and Stram (paper presented at the Joint Statistical Meetings, Las Vegas, 1985). Model 2 is a multivariate growth curve model of the type considered by Potthoff and Roy (1964). Although closed-form expressions for the MLEs under Model 2 are known (Khatri, 1966; Grizzle and Allen, 1969), the scoring algorithm does not converge in one iteration for this model. We are not aware of any closed-form expressions for the MLEs under Models 4 and 5. Of course, our approach can be used to fit these same models iteratively when the

data are not balanced, as would be the case if some subjects had fewer than four measurements present.

6. Discussion

Direct comparisons of the Newton, scoring, and EM scoring algorithms in terms of required computation are difficult, because this depends to a large degree on how efficiently the algorithms are coded. The Newton algorithm, with a quadratic convergence rate, generally converges in a small number of iterations, with a higher cost per iteration. At the other extreme, EM scoring has the lowest cost per iteration, but at times requires a large number of iterations. The scoring algorithm is intermediate in terms of cost per iteration and required number of iterations. However, the cost per iteration of scoring is often not much less than that of the Newton algorithm, whereas the scoring algorithm sometimes requires a considerably higher number of iterations than does the Newton algorithm. Because the scoring algorithm is often more robust to poor starting values than is the Newton algorithm, a good compromise is to start with scoring, and to switch to the Newton algorithm after the first several steps, as we have done.

When the number of covariance parameters q is not large, any of the algorithms will generally be satisfactory, but we prefer the Newton algorithm because it is not restricted to the incomplete data model and because of its generally clean and fast convergence. With large q, as when fitting a large unstructured covariance matrix ($T > 10$, say), only the EM scoring algorithm is feasible because the cost per iteration of the Newton and scoring algorithms becomes excessive.

Another consideration is that the Newton and scoring algorithms produce standard errors based on the empirical and Fisher information matrices as a by-product. The EM scoring algorithm gives no standard errors for $\hat{\boldsymbol{\theta}}$, but these can be obtained by taking a single Newton or scoring step after convergence.

The incomplete data model discussed in Section 2 may result from a missing-data mechanism operating on a set of complete data. Our methods, which do not explicitly incorporate such a mechanism, apply as long as the mechanism can be ignored. The most commonly occurring ignorable mechanism involves observations that are randomly missing as though by throwing darts at a complete data matrix. One may then use the marginal distributions for the observed data assuming these have the form set forth in Section 2.

There are missing-data mechanisms that lead to the same likelihood for $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ as that derived from the model in Section 2, even though, because of the mechanism, the observed data itself do not follow the model assumed there. Rubin (1976b) identifies such mechanisms. In this situation, one may proceed with likelihood-based inferences. This justifies, for example, the use of the empirical information matrix for supplying standard errors, but, as Rod Little has observed (personal communication), it does not justify the use of the Fisher (expected) information matrix computed without regard to the missing-data mechanism. This is a motivation for providing both forms of standard errors.

For the special case where $\boldsymbol{\Sigma}$ is unstructured, Ware (1985) outlined a "modified Gauss–Seidel" algorithm that is similar to our EM scoring algorithm. Ware's algorithm differed in that it iterated on steps (ii)–(iv), which Ware noted was undesirable because it required two levels of iteration. Our algorithm avoids two levels of iteration by taking a single step of generalized EM to update $\boldsymbol{\theta}$ given $\boldsymbol{\beta}$.

If no data are incomplete, then our EM scoring and scoring algorithms are equivalent. In general, these two algorithms will produce similar sequences of iterates when only a small amount of data is missing.

The scoring and Newton algorithms are designed to solve the general problem set forth in Section 3. The algorithms using EM solve subproblems and seem to work because they

exploit the restrictions that lead to these subproblems. The use of $\mathbf{e}_i^*$ as complete data, for example, exploits the restrictions of the incomplete data model.

Acknowledgements

This work was supported by NIMH Grant 37188, while Mark Schluchter was with the Department of Biomathematics, University of California, Los Angeles. Requests for reprints should be sent to him at his current address. Robert Jennrich was supported by NIMH Grant 37188 and by NSF Grant MCS-8301587.

Résumé

La question de savoir comment analyser des données incomplètes ou désequilibrées correspondants à des mesures repetées est un problème que rencontrent frequement les analystes. Nous proposons d'etudier ce problème en utilisant le maximum de vraisemblance en posant un modèle linéaire general pour l'esperance et avec une structure arbitraire pour les covariances intra-individu. Les modèles que l'on peut traiter incluent: les modèles univariables ou multivariables standarts avec données manquantes, les modèles à effets aléatoires, ainsi que les séries chronologiques et les erreurs à structure factorielle. On decrit les algorithmes de Newton–Raphson et "Fisher scoring" pour calculer les estimateurs du maximum de vraisemblance restreint ou non restreint. Un exemple de l'ajustement de plusieurs modèles à des données de croissance est proposé.

References

Anderson, T. W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some data are missing. *Journal of the American Statistical Association* **52,** 200–203.

Bartlett, M. S. (1937). Some examples of statistical methods of research in agriculture and applied biology. *Journal of the Royal Statistical Society, Supplement* 4, 137–170.

Beale, E. M. L. and Little, R. J. A. (1975). Missing values in multivariate analysis. *Journal of the Royal Statistical Society, Series B* **37,** 129–146.

Berk, K. N. (1984). Computing for unbalanced repeated measures experiments. In *Proceedings of the Eleventh Annual SUGI Conference*, 754–760. Cary, North Carolina: SAS Institute.

Bock, R. D. (1963). Multivariate analysis of repeated measurements. In *Problems in Measuring Change*, C. W. Harris (ed.), 85–103. Madison, Wisconsin: University of Wisconsin Press.

Box, G. E. P. and Jenkins, G. M. (1970). *Time Series Analysis: Forecasting and Control*, revised edition. San Francisco: Holden-Day.

Cole, J. W. L. and Grizzle, J. E. (1966). Applications of multivariate analysis of variance to repeated measurements experiments. *Biometrics* **22,** 810–828.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with Discussion). *Journal of the Royal Statistical Society, Series B* **39,** 1–38.

Grizzle, J. E. and Allen, D. M. (1969). Analysis of growth and dose–response curves. *Biometrics* **25,** 357–381.

Hartley, H. O. and Hocking, R. R. (1971). The analysis of incomplete data. *Biometrics* **27,** 783–808.

Hartley, H. O. and Rao, J. N. K. (1967). Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika* **54,** 93–108.

Hemmerle, W. J. and Hartley, H. O. (1973). Computing maximum likelihood estimates for the mixed AOV model using the *W* transform. *Technometrics* **15,** 819–831.

Jennrich, R. I. and Sampson, P. F. (1976). Newton–Raphson and related algorithms for maximum likelihood variance component estimation. *Technometrics* **18,** 11–17.

Khatri, C. G. (1966). A note on the MANOVA model applied to problems in growth curve. *Annals of the Institute of Statistical Mathematics* **18,** 75–86.

Kleinbaum, D. G. (1973a). Testing linear hypotheses in generalized multivariate linear models. *Communications in Statistics* **1,** 433–457.

Kleinbaum, D. G. (1973b). A generalization of the growth curve model which allows missing data. *Journal of Multivariate Analysis* **3,** 117–124.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38,** 963–974.

Orchard, T. and Woodbury, M. A. (1972). A missing information principle: Theory and applications. In *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability* **1,** 697–715. Berkeley, California: University of California Press.

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.

Rubin, D. B. (1976a). Noniterative least squares estimates, standard errors and *F*-tests for analysis of variance with missing data. *Journal of the Royal Statistical Society, Series B* **38,** 270–274.

Rubin, D. B. (1976b). Inference and missing data. *Biometrika* **63,** 581–592.

Schwertman, N. C. (1978). A note on the Greenhouse–Geisser correction for incomplete data split-plot analysis. *Journal of the American Statistical Association* **73,** 393–396.

Szatrowski, T. H. (1983). Missing data in the one-population multivariate normal patterned mean and covariance matrix testing and estimation problem. *Annals of Statistics* **11,** 947–958.

Timm, N. H. (1980). Multivariate analysis of variance of repeated measures. In *Handbook of Statistics, Analysis of Variance*, Vol. 1, P. R. Krishnaiah (ed.), 41–87. New York: North-Holland.

Trawinski, I. M. and Bargman, R. E. (1964). Maximum likelihood estimation with incomplete multivariate data. *Annals of Mathematical Statistics* **35,** 647–657.

Ware, J. H. (1985). Linear models for the analysis of longitudinal studies. *The American Statistician* **39,** 95–101.

Winer, B. J. (1971). *Statistical Principles in Experimental Design*, 2nd edition. New York: McGraw-Hill.

Received August 1985; *revised July* 1986.