Computational Strategies for Multivariate Linear Mixed-Effects Models With Missing Values

Joseph L. SCHAFER and Recai M. YUCEL

This article presents new computational techniques for multivariate longitudinal or clustered data with missing values. Current methodology for linear mixed-effects models can accommodate imbalance or missing data in a single response variable, but it cannot handle missing values in multiple responses or additional covariates. Applying a multivariate extension of a popular linear mixed-effects model, we create multiple imputations of missing values for subsequent analyses by a straightforward and effective Markov chain Monte Carlo procedure. We also derive and implement a new EM algorithm for parameter estimation which converges more rapidly than traditional EM algorithms because it does not treat the random effects as "missing data," but integrates them out of the likelihood function analytically. These techniques are illustrated on models for adolescent alcohol use in a large school-based prevention trial.

Key Words: EM algorithm; Longitudinal data; Markov chain Monte Carlo; Multiple imputation.

1. INTRODUCTION

1.1 THE MODEL

Multivariate longitudinal or clustered data are characterized by multiple responses measured (a) at multiple occasions for each subject or (b) for subjects nested within naturally occurring groups. Examples include multiple exam or test scores recorded for students across time, and multiple items at a single occasion for students in more than one school. Sensible methods for analyzing such data will appreciate both the relationships among the

Joseph L. Schafer is Associate Professor, Department of Statistics and The Methodology Center, The Pennsylvania State University, University Park, PA 16802 (E-mail: jls@stat.psu.edu). Recai M. Yucel is Statistician and Instructor in Medicine (Health Policy), Institute for Health Policy and Harvard Medical School, Boston, MA 02115 (E-mail: yucel@gem.mgh.harvard.edu). Authors' names are given in alphabetical order.

Journal of Computational and Graphical Statistics, Volume 11, Number 2, Pages 437–457

response variables and potential correlations among observations from the same individual or cluster. This article discusses a multivariate version of a popular linear mixed-effects model for longitudinal or clustered data and applies this model to datasets with missing values.

Let y_i denote an $n_i \times r$ matrix of multivariate responses for sample unit i, $i = 1, 2, \ldots, m$, where each row of y_i is a joint realization of variables $Y_1, Y_2, \ldots, Y_r$. We consider situations where portions of $y_1, \ldots, y_m$ are ignorably missing in the sense described by Rubin (1976) and Little and Rubin (1987). Our model for the complete data is

$$y_i = X_i\beta + Z_i b_i + \epsilon_i, \tag{1.1}$$

where X_i $(n_i \times p)$ and Z_i $(n_i \times q)$ are known covariate matrices, β $(p \times r)$ is a matrix of regression coefficients common to all units, and b_i $(q \times r)$ is a matrix of coefficients specific to unit i. In popular terminology, β and b_i are called "fixed effects" and "random effects," respectively. We assume that the n_i rows of ϵ_i are independently distributed as $N(0, \Sigma)$, and that the random effects are distributed as $\text{vec}(b_i) \sim N(0, \Psi)$ independently for $i = 1, \ldots, m$ (the "vec" operator vectorizes a matrix by stacking its columns). Without conditioning on $b_1, \ldots, b_m$, the implied model for $\text{vec}(y_i)$ is normal with mean $\text{vec}(X_i\beta)$ and covariance matrix

$$W_i^{-1} = (I_r \otimes Z_i)\Psi(I_r \otimes Z_i)^T + (\Sigma \otimes I_{n_i}). \tag{1.2}$$

In longitudinal applications, times of measurement may be incorporated into X_i and Z_i, allowing relevant aspects of the growth curves (e.g., intercepts and slopes) to vary by subject.

1.2 PREVIOUS WORK

The univariate ($r = 1$) version of our model,

$$y_i \sim N(X_i\beta,\, Z_i\psi Z_i^T + \sigma^2 I_{n_i}), \tag{1.3}$$

and more general univariate models have been extensively treated by Laird and Ware (1982); Jennrich and Schluchter (1986); Laird, Lange, and Stram (1987); Lindstrom and Bates (1988); and others. A variety of software is available for fitting these linear mixed-effects models. Commercial packages include HLM (Bryk, Raudenbush, and Congdon 1996) and MLn (Multilevel Models Project 1996). Similar procedures are now found in SAS (Littell, Milliken, Stroup, and Wolfinger 1996), S-Plus (Mathsoft, Inc. 1997), and STATA (Stata Corporation 1997). These programs can handle unbalanced longitudinal data, with measurements taken at an arbitrary set of time points for each subject. Responses that are missing, either unintentionally or by design, are ignored in the computations along with the corresponding rows of X_i and Z_i. An important limitation of these methods is that missing values must be confined to the single response variable; missing values on predictors are not allowed.

Despite the popularity of single-response models, multivariate versions have received scant treatment in the literature. A model similar to (1.1) was considered by Reinsel (1984) who derived closed-form estimates with completely observed y_i and balanced designs. More recently, Shah, Laird, and Schoenfeld (1997) extended the EM-type algorithm of Laird and Ware (1982) to a bivariate ($r = 2$) setting. In common econometric terminology, their model is analogous to "seemingly unrelated regression" (Zellner 1962) whereas ours corresponds to "standard multivariate regression." The added generality of the seemingly unrelated model comes at a high cost, making the resulting algorithms impractical for more than a few response variables. In certain situations, it may be possible to recast the multivariate model as a univariate one by stacking the columns of y_i and applying existing software (e.g., SAS Proc Mixed) with a user-specified covariance structure. In most applications, however, this approach quickly becomes impractical. Examples for only $r = 2$ response variables with complete data (Shah, Laird, and Schoenfeld 1997) and incomplete data (Verbeke and Molenberghs 2000) require complicated SAS macros. As the number of variables and number of individuals or time-points per cluster grow, the dimension of the response increases rapidly, and usage of SAS Proc Mixed becomes practically impossible.

Perhaps one reason why little attention has been paid to the multivariate models is that it is often natural to regard one of the variables as a response and the others as potential predictors. When the predictors have missing values, however, joint modeling of the multiple responses becomes helpful or even necessary; some type of modeling assumptions must be applied to $Y_1, \ldots, Y_r$ to achieve an efficient solution, even if the parameters of interest pertain only to the conditional model for one variable given the others.

In panel studies where individuals are assessed at a common set of occasions, models equivalent to ours may be formulated as latent growth curves (McArdle 1988; Meredith and Tisak 1990) and fit with structural-equations software. Two programs for structural equations, Mx (Neale 1994) and Amos (Arbuckle 1995), perform ML estimation from datasets with missing values. In principle, missing values can also be accommodated in other structural-equations software using a multiple groups approach (Allison 1987; Muthén, Kaplan, and Hollis 1987) but the implementation can be tedious. A disadvantage of the latent growth-curve formulation is that the measurements must be taken at a small number of common time points for all subjects. The method does not apply to clustered situations where the rows of y_i represent subjects nested within a group.

Schafer (1997) derived likelihood-based and Bayesian methods for independent multivariate observations with arbitrary patterns of missing values. In certain cases, this methodology can be applied to longitudinal data by treating the same outcome at different time points as distinct variables. Because this approach does not take into account the longitudinal structure, it may introduce more parameters than can be well estimated from the observed data.

1.3 SCOPE OF THIS ARTICLE

In the following sections, we develop computational techniques for applying the multivariate linear mixed model (1.1) to datasets with missing values. Two approaches

are discussed. The first one, described in Section 2, is to generate multiple imputations for the missing values using Markov chain Monte Carlo (MCMC). We extend the methodology of Schafer (1997) to groups of correlated multivariate observations, making it applicable to a variety of cluster samples and panel studies. In one sense, the material in Section 2 could be regarded as straightforward application of existing MCMC methods described elsewhere (e.g., Gilks, Richardson, and Spiegelhalter 1996). However, many of the the details of our implementation—especially where missing data are involved—might not be obvious even to readers familiar with MCMC. With careful attention to these computational details, the method is very effective and may be applied to datasets that are quite large.

Section 3 describes a second set of techniques which produce maximum-likelihood estimates or posterior modes. These methods may be used to estimate the parameters of model (1.1) directly from the incomplete data. They may also be used in conjunction with the MCMC methods of Section 2, helping the user to obtain good quality starting values and to select prior distributions for unknown variance components. Mode-finding algorithms are also helpful for testing model fit. The major innovation of Section 3 is a newly formulated EM algorithm which performs substantially better than previous methods.

Section 4 illustrates our methods by applying them to data from the Adolescent Alcohol Prevention Trial, a longitudinal study of substance-use attitudes and behaviors. Finally, Section 5 discusses the limitations of our model and future extensions. Procedures discussed here will be made available in a stand-alone program called PAN (Schafer and Yucel 2001) which operates in the Windows environment. PAN can be downloaded free of charge from http://www.stat.psu.edu/~jls/misoftwa.html.

2. METHODS FOR MULTIPLE IMPUTATION

2.1 MULTIPLE IMPUTATION BY MCMC

Suppose that portions of $Y = (y_1, y_2, \ldots, y_m)$ are ignorably missing. Let $y_{i(\text{obs})}$ and $y_{i(\text{mis})}$ denote the observed and missing parts of y_i, respectively, and let $Y_{\text{obs}} = (y_{1(\text{obs})}, y_{2(\text{obs})}, \ldots, y_{m(\text{obs})})$ and $Y_{\text{mis}} = (y_{1(\text{mis})}, y_{2(\text{mis})}, \ldots, y_{m(\text{mis})})$ denote all observed and missing responses. Unknown parameters are denoted by $\theta = (\beta, \Sigma, \Psi)$. For the fixed effects and residual covariances, we assume that $\beta \in \mathcal{R}^{pr}$ and $\Sigma > 0$. Depending on the application, we may allow Ψ to be either (a) unstructured or (b) block diagonal with r nonzero blocks of size $q \times q$ corresponding to the individual columns of b_i.

Multiple imputation, developed by Rubin (1987, 1996), is an increasingly popular method for handling missing values. For multiple imputation, we generate M independent draws $Y_{\text{mis}}^{(1)}, \ldots, Y_{\text{mis}}^{(M)}$ from a posterior predictive distribution for the missing data,

$$P(Y_{\text{mis}} \,|\, Y_{\text{obs}}) = \int P(Y_{\text{mis}} \,|\, Y_{\text{obs}}, \theta)\, P(\theta \,|\, Y_{\text{obs}})\, d\theta, \tag{2.1}$$

where $P(\theta \,|\, Y_{\text{obs}})$ is the observed-data posterior density, which is proportional to the product

of a prior density $\pi(\theta)$ and the observed-data likelihood function

$$L(\theta \,|\, Y_{\text{obs}}) \;=\; \int L(\theta \,|\, Y)\, dY_{\text{mis}}.$$

After imputation, the resulting M versions of the complete data are analyzed separately by complete-data methods, and the results are combined using simple arithmetic to obtain inferences that effectively incorporate uncertainty due to missing data. As shown by Rubin (1987), quality inferences can often be obtained with a very small number (e.g., $M = 5$) of imputations. Methods for combining the results of the complete-data analyses are given by Rubin (1987, 1996) and reviewed by Schafer (1997, chap. 4).

When a model is used as a device for imputation, the meaning or interpretation of its parameters is not crucial; the utility of the model lies in its ability to predict and simulate missing observations. A sensible imputation method for multivariate longitudinal or clustered data should preserve basic relationships among variables and correlations among observations from the same subject or cluster. The model (1.1) is capable of preserving these effects. In many cases, post-imputation analyses will be based on models less elaborate; for example, a model for one response variable given the others. In other cases, effective analyses may be carried out under a model somewhat different from that used to impute missing values. The performance of multiple imputation when the imputer's and analyst's models differ was addressed by Meng (1994) and Rubin (1996). In practice, inference by multiple imputation is fairly robust to departures from the imputation model because that model effectively applies not to the entire dataset but only to its missing parts. We have used (1.1) as the basis for imputing binary and ordinal responses, rounding off the continuous imputed values to the nearest category. Simulations have shown that the biases incurred by such rounding procedures may be minor (Schafer 1997). At best this is only an approximate solution; a more principled but complicated approach may involve introducing random effects into the general location model for multivariate data with continuous and categorical variables (Olkin and Tate 1961; Schafer 1997).

Except in trivial special cases, the posterior predictive distribution (2.1) for our model cannot be simulated directly. We create random draws of Y_{mis} from $P(Y_{\text{mis}} \mid Y_{\text{obs}})$ by techniques of Markov chain Monte Carlo (MCMC). In MCMC, one generates a sequence of dependent random variates whose distribution converges to the desired target. Overviews of MCMC were given by Gelfand et al. (1990); Smith and Roberts (1993); Tanner (1993); and in the chapters of Gilks, Richardson, and Spiegelhalter (1996). Schafer (1997) described MCMC for multivariate continuous and categorical missing data problems, but did not consider mixed models with random effects. Applications of MCMC to univariate linear mixed models have been made by a number of authors, including Gelfand, Hills, Racine-Poon, and Smith (1990); Zeger and Karim (1991); Liu and Rubin (1995); and Carlin (1996). These MCMC methods rely on simplifications that result when the random effects are assumed known. If $B = (b_1, b_2, \ldots, b_m)$ were known, then inferences about θ would separate into two simpler problems: (a) a normal-theory inference about Ψ based on B, and (b) a normal-theory inference about (β, Σ) based on $(y_i - Z_i b_i)$, $i = 1, \ldots, m$. This simplification is also an underlying feature of conventional EM algorithms for random-

effects model as well, to be discussed in Section 3. Unlike EM, however, MCMC allows us to circumvent manipulations on large matrices by alternately conditioning on simulated values of the random effects and the missing data.

2.2 A GIBBS SAMPLER

In a slight abuse of notation, let $A^* \sim P(A)$ denote simulation of a random variate A^* from a distribution or density function $P(A)$. Consider an iterative simulation algorithm in which current versions of the unknown parameters $\theta^{(t)} = (\beta^{(t)}, \Sigma^{(t)}, \Psi^{(t)})$ and missing data $Y_{\text{mis}}^{(t)}$ are updated in three steps: first,

$$b_i^{(t+1)} \sim P\left(b_i \mid Y_{\text{obs}}, Y_{\text{mis}}^{(t)}, \theta^{(t)}\right) \tag{2.2}$$

independently for $i = 1, \ldots, m$; next,

$$\theta^{(t+1)} \sim P\left(\theta \mid Y_{\text{obs}}, Y_{\text{mis}}^{(t)}, B^{(t+1)}\right); \tag{2.3}$$

and finally,

$$y_{i(\text{mis})}^{(t+1)} \sim P\left(y_{i(\text{mis})} \mid Y_{\text{obs}}, B^{(t+1)}, \theta^{(t+1)}\right) \tag{2.4}$$

for $i = 1, \ldots, m$. Given starting values $\theta^{(0)}$ and $Y_{\text{mis}}^{(0)}$, these steps define one cycle of an MCMC procedure called a Gibbs sampler. Executing the cycle repeatedly creates sequences $\{\theta^{(1)}, \theta^{(2)}, \ldots\}$ and $\{Y_{\text{mis}}^{(1)}, Y_{\text{mis}}^{(2)}, \ldots\}$ whose limiting distributions are $P(\theta \mid Y_{\text{obs}})$ and $P(Y_{\text{mis}} \mid Y_{\text{obs}})$, respectively.

Implementing (2.3) requires a prior distribution for θ. It is known that in mixed-effects models, improper prior distributions for the covariance components may lead to Gibbs samplers that do not converge to proper posteriors, even though each step of the cycle is well-defined. For this reason, proper prior distributions for the covariance matrices are highly recommended. For simplicity, we apply independent inverted Wishart priors $\Sigma^{-1} \sim W(\nu_1, \Lambda_1)$ and $\Psi^{-1} \sim W(\nu_2, \Lambda_2)$, where $W(\nu, \Lambda)$ denotes a Wishart variate with $\nu > 0$ degrees of freedom and mean $\nu\Lambda > 0$. This prior is appropriate for a model with unstructured Ψ; versions for block-diagonal Ψ will be discussed later. These priors exist provided that $\Lambda_1 > 0$, $\Lambda_2 > 0$, $\nu_1 \geq r$ and $\nu_2 \geq qr$. In choosing values for the hyperparameters, it is helpful to regard $\nu_1^{-1}\Lambda_1^{-1}$ and $\nu_2^{-1}\Lambda_2^{-1}$ as prior guesses for Σ and Ψ with confidence equivalent to ν_1 and ν_2 degrees of freedom, respectively. Small values for ν_1 and ν_2 make the prior densities relatively diffuse, reducing their impact on the final inferences. For β, we use an improper uniform "density" over $\mathcal{R}^{pr}$.

Under these priors, each of the steps (2.2)–(2.4) is derived by straightforward application of Bayes' theorem. In our model, the pairs (y_i, b_i) are distributed as

$$\begin{aligned} \text{vec}(y_i) \mid b_i, \theta &\sim N(\,\text{vec}(X_i\beta + Z_i b_i),\, (\Sigma \otimes I_{n_i})\,), \\ \text{vec}(b_i) \mid \theta &\sim N(0, \Psi) \end{aligned}$$

independently for $i = 1, \ldots, m$. It follows that

$$\text{vec}(b_i) \,|\, y_i, \theta \;\sim\; N(\text{vec}(\tilde{b_i}), U_i),$$

where

$$\begin{aligned} \text{vec}(\tilde{b_i}) &= U_i \,(\Sigma^{-1} \otimes Z_i^T)\, \text{vec}(y_i - X_i\beta), && (2.5) \\ U_i &= (\,\Psi^{-1} + (\Sigma^{-1} \otimes Z_i^T Z_i)\,)^{-1}. && (2.6) \end{aligned}$$

Simulation of θ in (2.3) proceeds as follows: First, draw Ψ^{-1} from a Wishart distribution with degrees of freedom $\nu_2' = \nu_2 + m$ and scale $\Lambda_2' = (\Lambda_2^{-1} + B^T B)^{-1}$. Next, calculate the ordinary least-squares coefficients

$$\hat{\beta} \;=\; \left(\sum_{i=1}^{m} X_i^T X_i\right)^{-1} \left(\sum_{i=1}^{m} X_i^T (y_i - Z_i b_i)\right)$$

and residuals $\hat{\varepsilon}_i = y_i - X_i\hat{\beta} - Z_i b_i$, and draw Σ^{-1} from a Wishart distribution with degrees of freedom $\nu_1' = \nu_1 - p + \sum_{i=1}^m n_i$ and scale $\Lambda_1' = \left(\Lambda_1^{-1} + \sum_{i=1}^m \hat{\varepsilon}_i^T \hat{\varepsilon}_i\right)^{-1}$. Finally, draw β from a multivariate normal distribution centered at $\hat{\beta}$ with covariance matrix $\Sigma \otimes V$, where $V = \left(\sum_{i=1}^m X_i^T X_i\right)^{-1}$. For simulating β, it is helpful to note that if G and H are upper-triangular square roots of Σ and V, respectively ($G^T G = \Sigma$ and $H^T H = V$), then $G \otimes H$ is an upper-triangular square root of $\Sigma \otimes V$.

To carry out the final step (2.4) of the Gibbs sampler, notice that the rows of $\varepsilon_i = y_i - X_i\beta - Z_i b_i$ are independent and normally distributed with mean zero and covariance matrix Σ. Therefore, in any row of ε_i, the missing elements have an intercept-free multivariate normal regression on the observed elements; the slopes and residual covariances for this regression can be quickly calculated by inverting the square submatrix of Σ corresponding to the observed variables. Drawing the missing elements in ε_i from these regressions and adding them to the corresponding elements of $X_i\beta + Z_i b_i$ completes the simulation of $y_{i(\text{mis})}$.

2.3 IMPLEMENTATION ISSUES

The Gibbs sampler defined by (2.2)–(2.4) is not the only one that could be implemented for this problem; as noted by Liu and Rubin (1995) in the univariate case, a wide variety of alternative MCMC algorithms are possible. If any of the steps (2.2)–(2.4) could be carried out without conditioning on simulated values of Y_{mis} or B, then the algorithm could be made to converge in fewer iterations. De-conditioning may greatly increase the computational cost per iteration, however, and some limited experience suggests that the additional effort required to do so is not worthwhile. With modern computers, iterations of (2.2)–(2.4) can be performed quickly even with the large datasets provided that sufficient physical memory is available to store Y_{obs}, $Y_{\text{mis}}^{(t)}$, and the covariate matrices X_i and Z_i.

The convergence behavior of this algorithm is governed by two factors: the amount of information about θ carried in Y_{mis} relative to Y_{obs}; and the degree to which the random

effects b_i can be estimated from y_i. If the missing portions of y_i exert high leverage over components of θ, or if the b_i are poorly estimated (i.e., if the within-unit precision matrices $\Sigma^{-1} \otimes Z_i^T Z_i$ tend to be small relative to ψ^{-1}), then convergence can be slow. Convergence may also be slow when the number of subjects m is large, because for large m the posterior distribution for Ψ given $b_1, \ldots, b_m$ becomes very tight, causing the drawn value for Ψ to be close to its previous value. When producing multiple imputations, slow convergence is not disastrous because in most cases only a few independent draws of Y_{mis} are needed. If the algorithm is believed to achieve approximate stationarity by T cycles, then M imputations of Y_{mis} can be generated in MT cycles. Convergence can be informally assessed by examining time-series plots, autocorrelations, and so on. for individual elements or functions of θ. In particular, one should pay close attention to the elements of Ψ because these parameters tend to exhibit high autocorrelations. Formal and informal convergence diagnostics for MCMC were discussed by Gilks, Richardson, and Spiegelhalter (1996) and Schafer (1997, chap. 4).

Notice that any row of y_i that is completely missing may be omitted from consideration, along with the corresponding rows of X_i and Z_i, without changing the form of the complete-data model (1.1). Ignoring these rows will eliminate unnecessary computation at each cycle and reduce the rate of missing information, speeding the overall convergence. These rows of data may be restored at the final imputation step (2.4) to produce a fully completed dataset.

2.4 PRIOR GUESSES AND ALTERNATIVE COVARIANCE STRUCTURES

When specifying values for the hyperparameters, our usual practice is to set $\nu_1 = r$ and $\nu_2 = qr$ to make the priors as dispersed as possible and minimize their subjective influence. We typically set $\Lambda_1^{-1} = \nu_1 \hat{\Sigma}$ and $\Lambda_2^{-1} = \nu_2 \hat{\Psi}$, where $\hat{\Sigma}$ and $\hat{\Psi}$ are reasonable prior guesses for Σ and Ψ. If no prior guesses are available, the data themselves may be used to obtain them; the EM algorithms of Section 3 are excellent tools for pursuing these guesses.

Excellent prior guesses for Σ and Ψ may also be obtained by temporarily supposing that Σ is diagonal and Ψ is block-diagonal. Under these conditions, the multivariate model separates into independent univariate models for each of the r columns of y_i, and ML or RML estimates of the variance components may be quickly calculated using existing software for linear mixed-effects models. When data are sparse and some aspects of Σ or Ψ are not well estimated, diagonal and block-diagonal prior guesses for Σ and Ψ, respectively, tend to stabilize the computational procedures in much the same way that ridge regression stabilizes estimated coefficients when collinearity is present. The use of ridge-like priors with incomplete and sparse multivariate data was described by Schafer (1997).

When modeling a large number of response variables at once, it may be advantageous to restrict Ψ to a block-diagonal structure—not only for the purpose of obtaining prior guesses, but also when running the Gibbs sampler itself. If Ψ is block-diagonal, then independent inverted Wishart prior distributions may be applied to the $q \times q$ nonzero blocks, $\Psi_j^{-1} \sim W(\nu_j, \Lambda_j)$ for $j = 1, 2, \ldots, r$. Weak priors are obtained by setting $\nu_j = q$ and $\Lambda_j^{-1} = \nu_j \hat{\Psi}_j$, where Ψ_j is an estimate or prior guess for Ψ_j. The distributions for these blocks in step

(2.3) become $\Psi_j^{-1} \sim W(\nu_j', \Lambda_j')$, where $\nu_j' = \nu_j + m$, $\Lambda_j'^{-1} = \Lambda_j^{-1} + \sum_{i=1}^m b_{ij} b_{ij}^T$, and b_{ij} is the jth column of b_i.

The choice between an unstructured or block-diagonal Ψ will depend on both theoretical and practical considerations. A block diagonal structure indicates no a priori associations between the random effects for any two response variables Y_j and $Y_{j'}$. In a multivariate cluster sample with many variables, many units per cluster, but relatively few clusters, it may simply not be possible to estimate covariances among the random effects for all response variables. It is important to note that even if Ψ is block-diagonal, the columns of b_i are not independent in an a posteriori sense because (2.6) is not block-diagonal. A formal likelihood ratio test to choose between the unstructured and block-diagonal forms for ψ is possible with the EM procedures in Section 3.

3. ALGORITHMS FOR MODE-FINDING

3.1 IMPORTANCE OF MODE-FINDING PROCEDURES

The Gibbs sampler of Section 2 is an effective method for imputing missing values in the y_i matrices under the multivariate model (1.1). In principle it may also be used to simulate Bayesian estimates for θ, but in many cases estimates are more easily found with EM. Deterministic parameter estimation or mode-finding algorithms are a desirable accompaniment to MCMC simulation procedures (Gelman, Carlin, Stern, and Rubin 1995; Carlin 1996; Schafer 1997). MCMC requires starting values for the unknown model parameters; ML estimates can provide excellent starting values. As described earlier, ML estimates may provide guidance for specifying prior distributions required by MCMC. Finally, an algorithm for ML estimation can help to reveal pathological situations where the likelihood function is unusually shaped, with multiple modes or suprema on the boundary.

The first method is a Fisher scoring procedure which applies when $y_1, \ldots, y_m$ are fully observed. The second method, discussed in Section 3.3, is a new EM algorithm which incorporates Fisher scoring into the M-step; this procedure may be used when the response matrices y_i are partially missing. This new EM algorithm tends to converge more quickly than conventional EM algorithms for mixed-effects models because the random effects are not included in EM's formulation of "missing data." Implementation of the new algorithm is somewhat more complicated, but the per-iteration execution time compares favorably to that of conventional EM in many examples. In a few cases, this new algorithm is less stable than conventional EM. A hybrid procedure that combines stability with rapid convergence is described in Section 3.4.

3.2 FISHER SCORING

After the general presentation of EM by Dempster, Laird, and Rubin (1977), EM and its extensions have been extensively applied to the univariate model (1.3). EM is designed

for ML estimation with incomplete data and in situations that can be formulated as missing-data problems. Conventional applications of EM to mixed-effects models treat the random coefficients as missing data, capitalizing on a factorization of the augmented-data likelihood,

$$L(\theta\,|\,Y,B) \;=\; L(\Psi\,|\,B)\,L(\beta,\sigma^2\,|\,Y,B). \tag{3.1}$$

The overall maximum of (3.1) with respect to θ can be found by maximizing each of the two factors separately, neither of which requires iteration. Each cycle of EM maximizes the expected logarithm of (3.1), where the expectation is taken with respect to the conditional distribution of B given Y with the parameters fixed at their current estimates. With some effort, these EM conventional algorithms for the univariate model can be extended to the multivariate case. Shah, Laird, and Schoenfeld (1997) extended the EM-type algorithm of Laird and Ware (1982) to a bivariate ($r = 2$) response, both for complete y_i and for incomplete y_i.

Conventional EM algorithms which operate on (3.1) may suffer from very slow convergence. We have found that when there are no missing values in y_i—or, more generally, when entire rows in y_i are missing—the likelihood can be maximized more quickly by Fisher scoring.

The likelihood function arising from the marginal normal distribution for y_i is

$$L(\theta) \;\propto\; \prod_{i=1}^{m} |W_i|^{1/2} \exp\left\{-\frac{1}{2}\delta_i^T W_i \delta_i\right\},$$

where $\delta_i = \text{vec}(y_i - X_i\beta)$ and W_i is defined by (1.2). Using the relationship $|W_i| = |\Sigma \otimes I_{n_i}|^{-1}|\Psi|^{-1}|U_i|$ and ignoring constants of proportionality, the logarithm of L becomes

$$\ell(\theta) = -\frac{N}{2}\log|\Sigma| - \frac{m}{2}\log|\Psi| + \frac{1}{2}\sum_{i=1}^{m}\log|U_i| - \frac{1}{2}\sum_{i=1}^{m}\delta_i^T\, W_i\delta_i. \tag{3.2}$$

Fisher scoring updates the current estimate $\theta^{(t)}$ by solving the linear system $C\theta^{(t+1)} = d$, where $C = -E\ell''(\theta^{(t)})$ and $d = C\theta^{(t)} + \ell'(\theta^{(t)})$. Upon convergence, the final value of C^{-1} provides an estimated covariance matrix for $\hat{\theta}$.

For convenience, we take derivatives with respect to β and the nonredundant elements of Ψ^{-1} and Σ^{-1}. These matrices can be expressed as

$$\begin{aligned}\Psi^{-1} &= \sum_{j=1}^{g}\omega_j G_j,\\ \Sigma^{-1} &= \sum_{j=1}^{h}\sigma_j F_j,\end{aligned}$$

where $G_1, G_2, \ldots, G_g$ and $F_1, F_2, \ldots, F_h$ are known symmetric matrices of dimensions $rq \times rq$ and $r \times r$, respectively. The number of free parameters in Ψ is $g = rq(rq+1)/2$ when Ψ is unstructured and $g = rq(q+1)/2$ when it is block-diagonal. The first derivatives of $\ell(\theta)$ are $\partial\ell/\partial\text{vec}(\beta) = -\Gamma^{-1}\text{vec}(\beta - \tilde{\beta})$,

$$\frac{\partial\ell}{\partial\omega_j} \;=\; \frac{1}{2}\sum_{i=1}^{m}\text{tr}\left(\Psi - U_i - \text{vec}(\tilde{b}_i)\text{vec}(\tilde{b}_i)^T\right)G_j,$$

and

$$\frac{\partial \ell}{\partial \sigma_l} = \frac{1}{2}\sum_{i=1}^{m} \text{tr}\left(n_i \Sigma F_l - (F_l \otimes Z_i^T Z_i)U_i - \text{vec}(\tilde{\epsilon}_i) F_l \text{vec}(\tilde{\epsilon}_i)^T\right),$$

where $\text{vec}(\tilde{\epsilon}_i) = \text{vec}(y_i - X_i\beta - Z_i\tilde{b}_i)$, and $\tilde{\beta}$ is obtained by generalized least squares (GLS),

$$\begin{aligned} \text{vec}(\tilde{\beta}) &= \Gamma \sum_{i=1}^{m} (I_r \otimes X_i)^T W_i \,\text{vec}(y_i), \\ \Gamma^{-1} &= \sum_{i=1}^{m} (I_r \otimes X_i)^T W_i (I_r \otimes X_i). \end{aligned}$$

Taking expectations over the distribution of y_i for fixed θ, one can show that $E(\tilde{\beta}) = \beta$, $E(\text{vec}(\hat{b}_i)) = 0$, and $E(\text{vec}(\hat{b}_i)(\text{vec}(\hat{b}_i))^T)$=$\Psi - U_i$. Using these facts and algebraic manipulation, it follows that

$$E\left(\frac{\partial^2 \ell}{\partial \text{vec}(\beta) \partial (\text{vec}(\beta))^T}\right) = -\Gamma$$

and

$$E\left(\frac{\partial^2 \ell}{\partial \omega_j \partial (\text{vec}(\beta))^T}\right) = E\left(\frac{\partial^2 \ell}{\partial \sigma_j \partial (\text{vec}(\beta))^T}\right) = 0.$$

Moreover,

$$\begin{aligned} E\left(\frac{\partial^2 \ell}{\partial \omega_j \partial \omega_k}\right) &= -\frac{1}{2}\sum_{i=1}^{m} \text{tr}(\Psi - U_i) G_j (\Psi - U_i) G_k, \\ E\left(\frac{\partial^2 \ell}{\partial \omega_j \partial \sigma_k}\right) &= -\frac{1}{2}\sum_{i=1}^{m} \text{tr} U_i (F_k \otimes Z_i^T Z_i) U_i G_j, \\ E\left(\frac{\partial^2 \ell}{\partial \sigma_j \partial \sigma_k}\right) &= -\frac{1}{2}\sum_{i=1}^{m} \text{tr}\left(n_i \Sigma F_j \Sigma F_k \right. \\ &\quad - (F_k \otimes Z_i^T Z_i) U_i (F_k \otimes Z_i^T Z_i) \\ &\quad \left. - 2(F_j \Sigma F_k \otimes Z_i^T Z_i) U_i\right). \end{aligned}$$

Because the cross-derivatives of β with the covariance parameters have zero expectation, the scoring step for θ separates into independent linear updates for β and (Ψ, Σ). The updated estimate for β is the GLS estimate $\tilde{\beta}$ under the current estimated covariance parameters. Collecting the free covariance parameters into vectors, $\omega = (\omega_1, \omega_2, \ldots, \omega_g)^T$, $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_h)^T$, and $\eta = (\omega^T, \sigma^T)^T$, the updated covariance estimates are found by solving $C\eta^{(t+1)} = d$ with

$$C = -\begin{bmatrix} E\left(\frac{\partial^2 \ell}{\partial \omega \partial \omega^T}\right) & E\left(\frac{\partial^2 \ell}{\partial \omega \partial \sigma^T}\right) \\ E\left(\frac{\partial^2 \ell}{\partial \sigma \partial \omega^T}\right) & E\left(\frac{\partial^2 \ell}{\partial \sigma \partial \sigma^T}\right) \end{bmatrix}$$

and $d = C\eta^{(t)} + \ell'(\eta)$. Updated estimates for Ψ and Σ are obtained by inversion of $\sum_j \omega_j G_j$ and $\sum_j \sigma_j F_j$. In typical situations, the algorithm converges by 10–15 cycles. Note that scoring-updated estimates for Ψ and Σ are not guaranteed to be positive definite; if the estimates stray outside the parameter space, a step-halving procedure is used to bring them back in.

3.3 EM ALGORITHM

We now discuss a procedure that can be used when arbitrary portions of the response matrices $Y = (y_1, y_2, \ldots, y_m)$ are ignorably missing. We embed our scoring procedure within an EM algorithm which augments the observed data with missing portions of y_i but not random effects. The performance of this algorithm is best when the proportion of partially observed rows in y_i is small, and degrades if the observed data become very sparse; however, it does not tend to slow down merely when the random effects are poorly estimated. The E-step calculates the expectation of the complete-data log-likelihood function (3.2) with respect to the conditional distribution of Y_{mis} given Y_{obs} under a current estimate of θ. The M-step updates the estimate of θ, maximizing this expected log-likelihood by scoring. Details are provided below.

For the E-step, note that (3.2) is a linear function of the sufficient statistics $\text{vec}(y_i)$ and $\text{vec}(y_i)\text{vec}(y_i)^T$. It follows from (1.1) that $\text{vec}(y_i)$ and $\text{vec}(b_i)$ are jointly normal with covariance matrix

$$\begin{bmatrix} (I_r \otimes Z_i)\Psi(I_r \otimes Z_i)^T & (I_r \otimes Z_i)\Psi \\ \Psi(I_r \otimes Z_i)^T & \Psi \end{bmatrix}. \tag{3.3}$$

One way to find the necessary expectations is to begin with (3.3), whose dimension is $(rq + rn_i) \times (rq + rn_i)$, and apply an orthogonalization method (e.g. sweep) for $i = 1, 2, \ldots, m$. This strategy may work in small examples but becomes prohibitively expensive as n_i or r grows. Instead, we capitalize on the fact that the rows of y_i are conditionally independent given b_i with constant covariance.

Consider the expectation of the first complete-data sufficient statistic,

$$E(y_i \mid y_{i(\text{obs})}) = E\left(E(y_i \mid y_{i(\text{obs})}, b_i) \mid y_{i(\text{obs})}\right).$$

This calculation requires access to the distributions of $y_{i(\text{mis})}$ given $(y_{i(\text{obs})}, b_i)$ and b_i given $y_{i(\text{obs})}$. The former is simple because, given b_i, the rows of $y_i^* = y_i - X_i\beta - Z_i b_i$ are independent and identically distributed as $N(0, \Sigma)$. Therefore, the missing elements in any row of y_i^* have, given the observed elements and b_i, an intercept-free regression on the observed elements; the parameters of this regression can be obtained by inverting the square submatrix of Σ corresponding to the observed elements. Letting $y^*_{ij(\text{mis})}$ and $y^*_{ij(\text{obs})}$ denote the missing and observed portions of the jth row of y_i^*, we have

$$E(y^*_{ij(\text{mis})} \mid y_{i(\text{obs})}, b_i) = \Sigma_{21}\Sigma_{11}^{-1} y^*_{ij(\text{obs})},$$

where Σ_{11} is the square submatrix of Σ corresponding to the observed elements and Σ_{21} is the rectangular submatrix of covariances between the missing and observed elements.

Finally, because y_i^* is a linear function of b_i, the expectation of y_i without conditioning on b_i is obtained by direct substitution of $E(b_i \mid y_{i(\text{obs})})$ for b_i. Notice that the value of $\Sigma_{21}\Sigma_{11}$ varies by missingness pattern but not by observational units $i = 1, 2, \ldots, m$; computations can be reduced by grouping rows with identical missingness patterns across units. The parameters of the distribution of b_i given $y_{i(\text{obs})}$ are obtained by applying a reverse-sweep procedure to $\hat{b}_i$ and U_i, as defined in Section 2.2, to de-condition upon $y_{i(\text{mis})}$.

For the second sufficient statistic $\text{vec}(y_i)\text{vec}(y_i)^T$, one can apply a similar argument, first calculating the conditional expectation given b_i and $y_{i(\text{obs})}$, then averaging over the distribution of b_i given $y_{i(\text{obs})}$. Let y_{ijk} denote the kth element of the jth row of y_i. The formula for the expectation of $y_{ijk}y_{ij'k'}$ depends on whether y_{ijk} and $y_{ij'k'}$ are observed or missing, and whether they are in the same ($j = j'$) or different ($j \neq j'$) rows. It is easy to see that the expectation of $y_{ijk}y_{ij'k'}$ given $y_{i(\text{obs})}$ is given by: $y_{ijk}y_{ij'k'}$ if both are observed; $y_{ijk}E(y_{ij'k'} \mid y_{i(\text{obs})})$ if y_{ijk} is observed and $y_{ij'k'}$ is missing; and

$$E(y_{ijk} \mid y_{i(\text{obs})})E(y_{ij'k'} \mid y_{i(\text{obs})}) + \text{cov}(y_{ijk}, y_{ij'k'} \mid y_{i(\text{obs})})$$

if both are missing. The covariance between y_{ijk} and $y_{ij'k'}$ given $y_{i(\text{obs})}$ is equal to

$$\text{cov}(A_{ijk}, A_{ij'k'} \mid y_{i(\text{obs})}) + [\Sigma_{22\cdot 1}]_{kk'}$$

if they are in the same row, and

$$\text{cov}(A_{ijk}, A_{ij'k'} \mid y_{i(\text{obs})})$$

if they are in different rows, where

$$A_{ijk} \;=\; E(y_{ijk} \mid b_i, y_{i(\text{obs})})$$

comes from the regression predictions for the missing elements in the jth row of y_i given the observed elements. The covariance $\text{cov}(A_{ijk}, A_{ij'k'} \mid y_{i(\text{obs})})$ is obtained by noting that it is a linear function of the elements of the covariance matrix for b_i given $y_{i(\text{obs})}$.

The M-step requires us to maximize the expected log-likelihood computed in the E-step. This expected log-likelihood has nearly the same form as (3.2) and can be maximized by a slight modification of the Fisher scoring procedure. Minor changes must be made to the function ℓ and its first derivatives, but the expected second derivatives remain the same. The first derivatives of $\ell_e = E(\ell \mid Y_{\text{mis}})$ with respect to the elements of θ are

$$\begin{aligned}
\frac{\partial \ell_e}{\partial \text{vec}(\beta)} &= -\left(\sum_{i=1}^{m}(I_r \otimes X_i)^T W_i (I_r \otimes X_i)\right)\text{vec}(\beta - \tilde{\beta}), \\
\frac{\partial \ell_e}{\partial \omega_j} &= \frac{1}{2}\sum_{i=1}^{m}\text{tr}\left(\Psi - U_i - (\Sigma^{-1} \otimes Z_i^T Z_i)\right. \\
&\qquad \left. U_i T_i U_i (\Sigma^{-1} \otimes Z_i^T Z_i)\right) G_j, \\
\frac{\partial \ell_e}{\partial \sigma_l} &= \frac{1}{2}\sum_{i=1}^{m}\text{tr}\left(n_i \Sigma F_l - (F_l \otimes Z_i^T Z_i)U_i\right. \\
&\qquad \left. -W_i(\Sigma F_j \Sigma \otimes I_{n_i})W_i T_i\right),
\end{aligned}$$

where

$$\begin{aligned} \text{vec}(\tilde{\beta}) &= \Gamma \sum_{i=1}^{m} (I_r \otimes X_i)^T W_i E(\text{vec}(y_i) \mid \theta, y_{i(\text{obs})}), \\ T_i &= E\left\{ \text{vec}(y_i - X_i\beta)\text{vec}(y_i - X_i\beta)^T \,|\, y_{i(\text{obs})}, \theta \right\}. \end{aligned}$$

After calculating these derivatives, we update the parameters in the same fashion as in Section 3.2.

In practice, it is not necessary to iterate until the scoring procedure converges within each M-step; one step of scoring is usually sufficient, provided that ℓ_e has increased. The resulting procedure becomes a generalized EM (GEM) algorithm rather than EM, in the terminology of Dempster, Laird, and Rubin (1977), and is usually well-behaved. Slightly faster convergence can often be achieved by a simple reparameterization, taking logarithms of the diagonal elements of Ψ^{-1} and Σ^{-1} for scoring, which seems to help when the maximum lies near the boundary of the parameter space. Derivatives with respect to these parameters are found by the expressions above and a chain rule.

3.4 FURTHER POINTS

Mode-finding algorithms, especially scoring, may require good starting values. We obtain starting values as follows: For each response variable Y_j, we fit univariate linear mixed model (1.3) using the cases for which Y_j is observed. Fast and stable algorithms described in a technical report (Schafer 1998) provide ML estimates for the portions of Σ, Ψ and β pertaining to Y_j. Off-diagonal elements of Σ and blocks of Ψ are initially set to zero.

Although our algorithm converges more quickly than conventional EM methods for mixed-effects models, it may be less stable when the log-likelihood is oddly shaped. To improve stability, we combine our method with a conventional EM procedure based on the augmented-data likelihood (3.1), substituting one step of conventional EM if a single step of scoring fails to increase the log-likelihood.

If random effects are eliminated ($\Psi = 0$), the model reduces to a standard multivariate regression $y_i = X_i\beta + \epsilon_i$ where the rows of ϵ are independently distributed as $N(0, \Sigma)$. In this situation, ML estimates of (β, Σ) may be found by a straightforward extension of EM algorithms for incomplete multivariate normal data (Schafer 1997, chap. 5). Note that a hypothesis test for $\Psi = 0$ should not be performed by standard likelihood-ratio methods because the null model places rq parameters on on the boundary of the parameter space, making the limiting distribution under null hypothesis rather complicated (Stram and Lee 1995). The standard chi-square limiting distribution does apply when testing the null hypothesis that Ψ is block-diagonal versus the unstructured alternative.

As an alternative to Fisher scoring, one might consider optimizing the expected log-likelihood by a sequence of constrained maximizations. For example, one could maximize with respect to β holding (Ψ, Σ) constant; then with respect to Ψ holding (β, Σ) constant; and then with respect to Σ holding (β, Ψ) constant. This would produce an ECM algorithm,

a useful generalization of EM described by Meng and Rubin (1993). In this example, however, two of the three constrained maximizations would require an iterative method such as Newton–Raphson, leading to no substantial simplification.

As with any EM algorithm, the procedure of Section 3.3 does not automatically produce correct standard errors for parameter estimates. If necessary, standard errors could be found by the supplemented EM (SEM) method of Meng and Rubin (1991). In most cases, however, multiple imputation as described in Section 2 will produce standard errors in a more straightforward and less cumbersome fashion.

Finally, consider the related problem of restricted maximum likelihood (RML) estimation, which maximizes the indefinite integral of the likelihood with respect to β. This function is

$$L_1(\theta) \quad \propto \quad |\Gamma|^{1/2} \prod_{i=1}^{m} |W_i|^{1/2} \exp\left\{ -\frac{1}{2} \text{vec}(y_i - X_i\tilde{\beta})^T \, W_i \text{vec}(y_i - X_i\tilde{\beta}) \right\},$$

where Γ and $\tilde{\beta}$ are as defined in Section 3.2. Our algorithms for ML estimates may be modified to compute RML estimates. One may approximate the expected second derivatives of $\ell_1(\theta) = \log L_1(\theta)$ by the expected second derivatives of $\ell(\theta)$, but first derivatives are more complicated because $\tilde{\beta}$ is a function of the unknown covariance parameters. These changes affect both the scoring procedure for complete y_i and the M-step for incomplete y_i.

4. EXAMPLE

4.1 ADOLESCENT ALCOHOL PREVENTION TRIAL

Data for this example were taken from the Adolescent Alcohol Prevention Trial (AAPT), a longitudinal school-based intervention study of substance use in the Los Angeles, CA, area (Hansen and Graham 1991). A sample of 3,574 school children received questionnaires yearly in grades 5–10 to measure substance-use attitudes and behaviors. We examined three important variables derived from the AAPT questionnaire: Y_1 = DRINKING, a composite measure of self-reported alcohol use; Y_2 = POSCON, a measure of the perceived positive consequences of use; and Y_3 = NEGCON, a measure of the perceived negative consequences of use. Many values of these variables were missing due to absenteeism and attrition, which we will assume to be ignorable (Little and Rubin 1987; Rubin 1976). The ignorability assumption has been considered in detail by Graham, Hofer, and Piccinin (1994) and is thought to be somewhat plausible; the primary reasons for attrition were ordinary moving and migration of students among schools and districts. Moreover, a large portion of truly ignorable missing data were missing by design; in some years, Y_2 and Y_3 were omitted at random from one-third of the questionnaires, and in other years these measures were not collected at all. Missingness rates for the three variables are shown in Table 1, and means and standard deviations by year are shown in Table 2.

Table 1. Missingness Rates (%) by Grade

	Grade					
	5	*6*	*7*	*8*	*9*	*10*
DRINKING	2	24	24	33	35	44
POSCON	47	55	62	100	66	63
NEGCON	48	56	62	100	100	100

For one analysis, researchers wanted to fit linear growth curves to predict Y_1 from Y_2, Y_3, and other important covariates including gender. This analysis was not a straightforward application of a linear mixed-effects model because of the high rates of missing values on the covariates Y_2 and Y_3. We multiply imputed values for Y_1, Y_2, and Y_3 under our multivariate model, allowing the growth modeling to proceed with standard software. Our imputation model specified linear trends over time with random slopes and intercepts for each of the $r = 3$ variables, a fixed effect for gender, and a gender by time interaction. Each X_i matrix had $p = 4$ columns corresponding to an intercept, grade, gender, and gender $\times$ grade; and each Z_i had $q = 2$ columns corresponding to intercept and grade. Notice from Table 2 that both the average level of DRINKING and its variation increase dramatically over time. To make the assumption of a constant residual covariance matrix Σ more plausible, reported alcohol use was re-expressed as the logarithm of (DRINKING+5).

Because NEGCON is entirely missing for the last three years of the study, the likely values of this variable for grades 8–10 are being inferred from two sources: extrapolation from grades 5–7 based on the assumption of linear growth, and the residual covariances among the three response variables which are assumed to be constant across time. Neither of these assumptions can be effectively tested from the data at hand, so inferences pertaining to NEGCON are heavily model-based.

4.2 MODE FINDING AND IMPUTATION

Prior to imputation, we examined alternative covariance structures using the estimation procedures of Section 3.3. Despite the high rates of missingness, our EM algorithm converged to a maximum relative parameter change of 0.0001 in only 104 iterations for the unstructured-Ψ model and 95 for the block-diagonal version. Without random effects

Table 2. Means (standard deviations) of Observed Variables by Grade

	Grade					
	5	*6*	*7*	*8*	*9*	*10*
DRINKING	−1.43	−1.12	−0.57	0.09	1.29	1.97
	(1.33)	(1.96)	(2.73)	(3.47)	(4.40)	(4.78)
POSCON	1.30	1.34	1.48	—	1.84	1.96
	(0.61)	(0.62)	(0.74)	—	(0.89)	(0.91)
NEGCON	2.94	3.05	3.07	—	—	—
	(0.76)	(0.75)	(0.77)	—	—	—

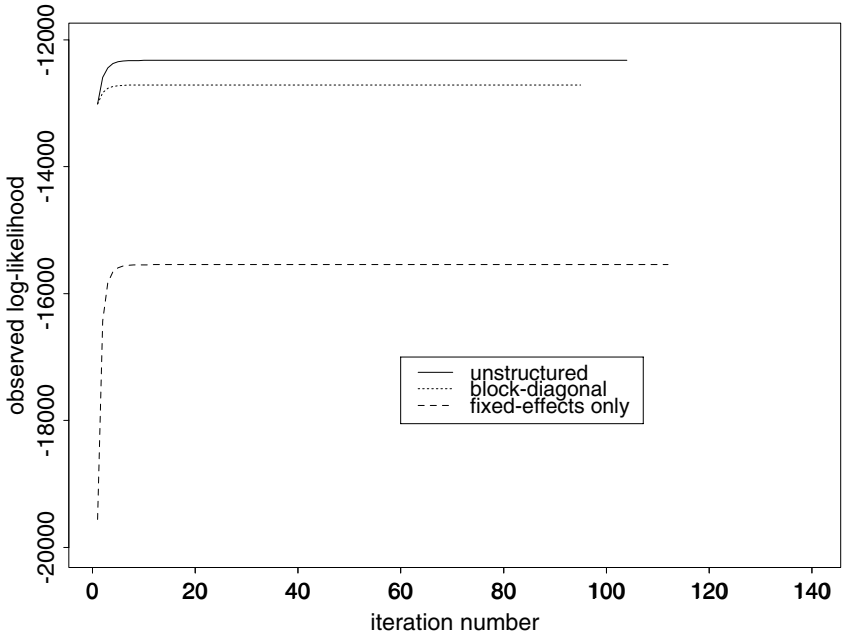

Figure 1. Convergence behaviors under different covariance structures.

($\Psi = 0$) EM again converged in approximately 100 steps. Values of the log-likelihood for all iterations are plotted in Figure 1. The likelihood-ratio statistic for testing the block-diagonal model against the unstructured alternative is 776.86; comparing this value to χ^2_{12} yields a p value of essentially zero.

In contrast to these EM algorithms, we anticipated that the Gibbs sampler of Section 2 would converge rather slowly, because that procedure augments the observed data by simulated random effects at each cycle. With only six occasions, the individual random slopes and intercepts for Y_1, Y_2, and Y_3 are not well estimated; moreover, the large sample size causes the augmented-data posterior distribution for Ψ to become very tight, inducing a high degree of correlation from one cycle to the next. To assess convergence, we ran our Gibbs sampler for an initial 2,000 cycles using an unstructured Ψ and mild prior distributions; we set $\nu_1 = 3$, $\Lambda_1^{-1} = 3\hat{\Sigma}$, $\nu_2 = 6$, and $\Lambda_2^{-1} = 6\hat{\Psi}$, where $\hat{\Sigma}$ and $\hat{\Psi}$ were obtained from EM. Time-series plots and sample autocorrelations for the elements of Ψ suggested that several hundred cycles were needed for the dependence to die out. Based on this information, we continued the Gibbs sampler for a total of 11,000 cycles, taking the simulated values of Y_{mis} stored at cycles 2,000, 3,000, ..., 11,000 as multiple imputations. Re-estimating the autocorrelations from cycles 1,001–11,000, we verified that the dependence in the elements of θ had indeed died down by lag 200, so the ten stored imputations could reasonably be regarded as independent draws from $P(Y_{\text{mis}} \mid Y_{\text{obs}})$. Each 1,000 cycles required approximately 17 minutes on a 400 MhZ Pentium II workstation.

Table 3. Estimated Coefficients, Standard Errors, Degrees of Freedom, and Percent Missing Information From Multiply-Imputed Growth-Curve Analysis

	est.	*SE*	*df*	*% missing*
intercept	−2.572	0.084	19	71
grade (1=5th, . . ., 6=10th)	0.386	0.011	35	53
sex (0=female, 1=male)	0.370	0.046	324	17
sex × grade	−0.105	0.013	88	33
POSCON	0.549	0.023	17	76
NEGCON	−0.090	0.023	15	80

4.3 POST-IMPUTATION ANALYSIS

After imputation, we analyzed the data by a conventional mixed-effects model for the logarithm of (DRINKING+5). The model was a version of (1.3) with fixed effects for gender, grade, gender×grade, POSCON and NEGCON, plus random intercepts and slopes for grade. ML estimates were computed from each imputed data set and combined using Rubin's (1987) rules for multiple-imputation inference for scalar estimands. Results of this procedure are summarized in Table 3. The point estimates are simply the averages of the ML estimates across the ten imputations. The standard errors incorporate uncertainty due to missing data as well as ordinary sampling variability. The degrees of freedom shown are the estimated degrees of freedom appropriate for hypothesis tests and interval estimates based on a Student's t-approximation. All coefficients are highly statistically significant.

Table 3 also displays the estimated percent rate of missing information for each estimand as derived by Rubin (1987). The high rates of missing information indicate that inferences for all coefficients (except sex) may be highly dependent upon the form of the imputation model and the assumption of ignorable nonresponse. The latter assumption is not particularly troubling for these data, because the majority of missing values are missing by design. Certain assumptions of the imputation model, however—in particular, the assumed linear growth for NEGCON and constancy of the residual covariances across time—are not really testable from the observed data, so results from this analysis should be interpreted with caution.

Despite these strong caveats, the estimates in Table 3 provide some intriguing and plausible interpretations about the behavior of this cohort. The positive coefficient for sex indicates that boys reported higher average rates of alcohol use than girls in the initial years of the study. The negative effect for sex×grade, however, shows that girls exhibit higher rates of increase than boys, so that the girls' average overtakes the boys' by grade 8. The large positive effect of POSCON indicates that increasing perceptions about the positive consequences of alcohol use are highly associated with increasing levels of reported use. The negative coefficient for NEGCON suggests that increasing beliefs about negative consequences do tend to reduce levels of use, but the effect is much smaller than that of POSCON. These results are consistent with those of previous studies (MacKinnon et al. 1991) which demonstrated that perceived positive consequences may be influential

determinants of substance-use behavior, but beliefs about negative consequences have little discernible effect.

5. DISCUSSION

The algorithms developed here represent an important step in helping researchers to analyze multivariate longitudinal or clustered data with missing values. If the dataset contains only a few large clusters, the MCMC procedure described in Section 2 will converge rapidly. With many small clusters the algorithm works very reliably but convergence may be slow. The EM methods of Section 3 were designed specifically for many small clusters and perform best in that setting.

It is straightforward to show that the multivariate mixed-effects model (1.1) implies a conditional univariate model of the form (1.3) for each response variable given the others, where the others are incorporated into the columns of X_i. Thus, the imputation procedures in Section 2 are appropriate for longitudinal analyses with partially missing covariates, when those covariates are later going to be incorporated into an analytic model as linear fixed effects. In many studies, however, one would like to preserve and detect certain nonlinear associations and interactions. For example, in the first analysis of Section 4, it would have been interesting to see whether the association between POSCON and DRINKING may have been increasing or decreasing over time; the imputation model, however, imputed the missing values under an assumption of a constant POSCON$\times$DRINKING association. Extensions of the multivariate model to allow more elaborate fixed associations such as POSCON $\times$ DRINKING $\times$ grade, or random associations such as POSCON $\times$ DRINKING $\times$ subject, are an important topic of ongoing research.

Another limitation of our methods is that they currently allow only two levels of nesting. Many studies involve multivariate longitudinal data that are clustered further into larger units (e.g., repeated multivariate measurements on students within schools). Extending the Gibbs sampler of Section 2 to accommodate additional levels of random effects is a simple matter, but extending the scoring and EM procedures of Section 3 is not.

Another important limitation pertains to missing covariates at the subject or cluster level, for example, non-time-varying covariates. If these covariates have no missing values, they can be handled under the current model by simply moving them to the matrix X_i. When missing values are present, however, they should be explicitly modeled and imputed. More specifically, let $V_i = (v_{i1}, v_{i2}, \ldots, v_{ik})^T$ denote a set of variables describing unit i that appear in some form in the columns of X_i. If one is willing to impose a simple parametric distribution on V_i such as multivariate normal, then Gibbs sampler given by (2.2)–(2.4) can easily be extended in the following fashion. Given V_i, the conditional distribution of y_i is be given by (1.1), and marginally the distribution of V_i is a multivariate normal distribution. Conditionally upon the random effects b_i, the joint distribution for V_i and y_i is still a multivariate normal with $(y_i - Z_i b_i)$ appended to the variables in V_i.

Our model assumes that the rows of y_i are conditionally independent given b_i with common covariance matrix Σ. In the univariate case, this assumption is commonly relaxed by allowing a residual covariance matrix of the form $\sigma^2 V_i$, where V_i has a simple (e.g.,

autoregressive or banded) pattern with a small number of unknown parameters. Sensible multivariate extensions of these patterned covariance structures produces models and algorithms that are complicated even apart from missing data. For example, the obvious extension of $\text{vec}(\epsilon_i) \sim N(0, (\Sigma \otimes I_{n_i}))$ to $\text{vec}(\epsilon_i) \sim N(0, (\Sigma \otimes V_i))$ seems too restrictive for many longitudinal datasets, because the response variables $Y_1, \ldots, Y_r$ would be required to have an identical autocorrelations. Accounting for autocorrelated residuals in a plausible manner may prove be a daunting task in the multivariate case. In many cases, apparent nonzero correlations among the rows of ϵ_i may arise because of a misspecified model for the mean structure over time. The problem may sometimes be reduced or eliminated by including additional (e.g., higher-order polynomial) terms for time in the covariate matrices X_i or Z_i.

ACKNOWLEDGMENTS

This work was funded by NIH grants 2R44CA65147 and 1-P50-DA10075. Thanks to John Graham for providing the data used in Section 4.

[Received December 1999. Revised January 2001.]

REFERENCES

Allison, P. D. (1987), "Estimation of Linear Models With Incomplete Data," in *Sociological Methodology*, ed. C. Clogg, Washington, DC: American Sociological Association, pp. 71–103.

Arbuckle, J. L. (1995), *Amos Users' Guide*, Chicago, IL: Small Waters.

Bryk, A. S., Raudenbush, S. W., and Congdon, R. T. (1996), *Hierarchical Linear and Nonlinear Modeling with HLM/2L and HLM/3L Programs*, Chicago, IL: Scientific Software International.

Carlin, B. P. (1996) "Hierarchical Longitudinal Modelling," in *Markov Chain Monte Carlo in Practice*, eds, W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, London, U.K.: Chapman & Hall, pp. 303–319.

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

Gelfand, A. E., Hills, S. E., Racine-Poon, A., and Smith, A. F. M. (1990), "Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling," *Journal of the American Statistical Association*, 85, 972–985.

Gelman, G., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995), *Bayesian Data Analysis*, London, U.K.: Chapman & Hall.

Gilks, W. R., Richardson, S., Spiegelhalter, D. J. (1996), *Markov Chain Monte Carlo in Practice*, London, U.K.: Chapman and Hall.

Graham, J. W., Hofer, S. M., and Piccinin, A. M. (1994) "Analysis With Missing Data in Drug Prevention Research," in *Advances in Data Analysis for Prevention Intervention Research*, eds. L.M. Collins and L.A. Seitz, Bethesda, MD: National Institute on Drug Abuse, pp. 13–63.

Hansen, W. B., and Graham, J. W. (1991), "Preventing Alcohol, Marijuana, and Cigarette use Among Adolescent: Peer Pressure Resistance Training Versus Establishing Conservative Norms," *Preventive Medicine*, 20, 414–430.

Jennrich, R. I., and Schluchter, M. D. (1986), "Unbalanced Repeated-Measures Models With Structured Covariance Matrices," *Biometrics*, 38, 967–974.

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

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

Lindstrom, M. J., and Bates, D. M. (1988), "Newton-Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data," *Journal of the American Statistical Association*, 83, 1014–1022.

Littell, R. C., Milliken, G. A., Stroup, W. W., and Wolfinger, R. D. (1996), *SAS System for Mixed Models*, Cary, NC: SAS Institute, Inc.

Little, R. J. A., and Rubin, D. B. (1987), *Statistical Analysis with Missing Data*, New York: Wiley.

Liu, C., and Rubin, D. B. (1995), "Application of the ECME Algorithm and the Gibbs Sampler to General Linear Mixed Models," *Proceedings of the 17th International Biometric Conference*, 1, 97–107.

MacKinnon, D. P., Johnson, C. A., Pentz, M. A., Dwyer, J. H., Hansen, W. B., Flay, B. R., and Wang, E. Y. (1991), "Mediating Mechanisms in a School-Based Drug Prevention Program: First-Year Effects of the Midwestern Prevention Project," *Health Psychology*, 10, 164–172.

MathSoft, Inc. (1997), *S-PLUS User's Guide*, Data Analysis Product Division, Seattle, WA: MathSoft.

McArdle, J. (1988), "Dynamic but Structural Modeling of Repeated Measures Data," in *Handbook of Multivariate Experimental Psychology*, eds. J. R. Nesselroade and R. B. Catell, New York: Plenum.

Meng, X. L. (1994), "Multiple-Imputation Inferences with Uncongenial Sources of Input" (with discussion), *Statistical Science*, 10, 538–573.

Meng, X. L., and Rubin, D. B. (1991), "Using EM to Obtain Asymptotic Variance-Covariance Matrices: The SEM Algorithm," *Journal of the American Statistical Association*, 86, 899–909.

——— (1993), "Maximum Likelihood Estimation via the ECM Algorithm: A General Framework," *Biometrika*, 80, 267–278.

Meredith, W., and Tisak, J. (1990), "Latent Curve Analysis," *Psychometrika*, 55, 105–122.

Multilevel Models Project (1996), *Multilevel Modeling Applications—A Guide for Users of MLn*, ed. Geoff Woodhouse, London: Institute of Education, University of London.

Muthén, B., Kaplan, D., and Hollis, M. (1987), "On Structural Equation Modeling With Data that are not Missing Completely at Random," *Psychometrika*, 55, 107–122.

Neale, M. C. (1994), *Mx: Statistical Modeling* (2nd ed.), Box 710 MCV, Richmond, VA 23298: Dept. of Psychiatry.

Olkin, I., and Tate, R. F. (1961), "Multivariate Correlation Models With Mixed Discrete and Continuous Variables," *The Annals of Mathematical Statistics*, 32, 448–465.

Reinsel, G. (1984) "Multivariate Repeated-Measurement or Growth Curve Models With Multivariate Random-Effects Covariance Structure," *Journal of the American Statistical Association*, 77, 190–195.

Rubin, D. B. (1976), "Inference and Missing Data," *Biometrika*, 63, 581–592.

——— (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

——— (1996), "Multiple Imputation After 18+ Years," *Journal of the American Statistical Association*, 91, 473–489.

Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data*, London, U.K.: Chapman & Hall.

——— (1998), "Some Improved Procedures for Linear Mixed Models," Technical Report, Department of Statistics, The Pennsylvania State University.

Schafer, J. L., and Yucel, R. M. (2001), PAN: *Multiple imputation for multivariate panel data,* software for Windows 95/98/NT. Available at http://www.stat.psu.edu/~jls/misoftwa.html.

Shah, A., Laird, N., Schoenfeld, D. (1997), "A Random-Effects Model for Multiple Characteristics With Possibly Missing Data," *Journal of the American Statistical Association*, 92, 775–779.

Smith, A. F. M., and Roberts, G. O. (1993), "Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods," *Journal of the Royal Statistical Society*, Ser. B, 55, 3–24.

Stata Corporation (1997), *Stata Reference Manual*, College Station, TX: Stata Press.

Stram, D. O., and Lee, J. W. (1995), Correction to "Variance Components Testing in the Longitudinal Mixed Effects Model," *Biometrics*, 51, 1196.

Tanner, M. A. (1993), *Tools for Statistical Inference, Methods for the Exploration of Posterior Distributions and Likelihood Functions*, (Second Edition), New York: Springer-Verlag.

Verbeke, G., and Molenberghs, G. (2000), *Linear Mixed Models for Longitudinal Data*, New York: Springer-Verlag.

Zellner, A. (1962), "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias," *Journal of the American Statistical Association*, 57, 348–368.