Maintainer's Note

The 1997 technical report as noted at `https://science.psu.edu/stat/research/technical-reports`., together with the subsequent publications, are all included in this vignette for convenience.

1. Schafer, J. L. (1997). *Imputation of Missing Covariates under a Multivariate Linear Mixed Model.* Technical Report 97-04, Department of Statistics, The Pennsylvania State University.

2. Schafer, J. L. (2001). Multiple imputation with PAN. In L. M. Collins and A. G. Sayer (Eds.), *New Methods for the Analysis of Change* (pp. 357–377). American Psychological Association. `https://doi.org/10.1037/10409-012`

3. Schafer, J. L., and Yücel, R. M. (2002). Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational and Graphical Statistics*, **11**, 437–457.

The marijuana data set included with the package is reproduced below. Example code can be viewed using `help(ecme, package = "pan")`.

Table 1: Change in heart rate (beats per minute above baseline) measured 15 and 90 minutes after marijuana use.

	15 minutes			90 minutes		
Subject	Placebo	Low	High	Placebo	Low	High
1	16	20	16	20	-6	-4
2	12	24	12	-6	4	-8
3	8	8	26	-4	4	8
4	20	8	—	—	20	-4
5	8	4	-8	—	22	-8
6	10	20	28	-20	-4	-4
7	4	28	24	12	8	18
8	-8	20	24	-3	8	-24
9	—	20	24	8	12	—

Imputation of missing covariates under a multivariate linear mixed model

Joseph L. Schafer *

February 13, 1997

Linear mixed-effects models have been widely used in the analysis of longitudinal and clustered data. Standard fitting procedures for these models allow for imbalance due to missing responses, but little has been done for problems of missing covariates. This article presents a method for creating multiple imputations (Rubin, 1987) of missing covariates, allowing the imputed data to be analyzed by current complete-data methods. The imputation procedure relies on a multivariate extension of a popular linear mixed-effects model (Laird and Ware, 1982). The multivariate model is consistent with a conditional linear mixed model for each covariate, with fixed effects for all other covariates. The technique is illustrated on a longitudinal study of adolescent substance use with large amounts of data missing by design.

Key Words: Gibbs sampling, linear mixed-effects model, longitudinal data, random effects, repeated measures

*Assistant Professor, Department of Statistics, Pennsylvania State University, University Park, PA 16802-6202. This research was supported by grant 2R44CA65147-02 from National Institutes of Health, and by grant 1-P50-DA10075-01 from the National Institute on Drug Abuse. Thanks to John Graham for providing data from the Adolescent Alcohol Prevention Trial and input on their analysis.

1 Introduction

Let y_i denote an $n_i \times r$ matrix of multivariate data for sample unit i, $i = 1, \ldots, m$, where each row of y_i is a joint realization of variables $Y_1, \ldots, Y_r$. Let us assume that y_i follows a multivariate linear mixed model of the form

$$y_i \;=\; X_i\beta + Z_i b_i + \varepsilon_i, \tag{1}$$

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 (the "fixed effects"), and b_i $(q \times r)$ is a matrix of coefficients specific to unit i (the "random effects"). We will assume that the n_i rows of ε_i are independently distributed as $N(0, \Sigma)$, and that the random effects are distributed as $b_i^V \;\sim\; N(0, \Psi)$ independently for $i = 1, \ldots, m$. The superscript "V" indicates vectorization of a matrix by stacking its columns. No further structure will be imposed on the covariance matrices or fixed effects; we will assume only that $\beta \in \mathcal{R}^{pr}$, $\Sigma > 0$, and $\Psi > 0$. Without conditioning on $b_1, \ldots, b_m$, the model becomes

$$y_i^V \;\sim\; N(\,(X_i\beta)^V,\, (I_r \otimes Z_i)\Psi(I_r \otimes Z_i)^T + (\Sigma \otimes I_{n_i})\,). \tag{2}$$

The univariate $(r = 1)$ version,

$$y_i \;\sim\; N(\,X_i\beta,\, Z_i\Psi Z_i^T + \sigma^2 I_{n_i}\,), \tag{3}$$

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. Estimation procedures—both ordinary and restricted maximum-likelihood—for the univariate versions are available in major statistical packages. The present article discusses inference for the multivariate version when arbitrary portions of the y_i may be ignorably missing or missing at random, in the sense described by Rubin (1976) and Little and Rubin (1987).

Natural applications for model (2) include (a) analyses of multivariate longitudinal data in which a set of r variables is measured for subject i at n_i occasions; and (b) analyses of clustered multivariate cross-sectional data in which subjects are nested within groups $i = 1, \ldots, m$ of varying sizes n_i. In (a), the measurements times will typically be incorporated in some fashion into X_i and Z_i; because these matrices are not assumed to have any particular form, the model allows time-varying covariates and measurement times that vary by subject. In (b), X_i and Z_i may contain descriptors of both the subjects and the groups to which they belong, allowing simultaneous estimation of effects due to characteristics at the subject and group levels.

In many analyses, it is natural to regard one of the variables (say Y_r) as a response and the remaining variables $(Y_1, \ldots, Y_{r-1})$ as potential predictors; interest is focused on the conditional distribution of Y_r given $Y_1, \ldots, Y_{r-1}$, and the parameters governing the joint distribution of $Y_1, \ldots, Y_{r-1}$ are of little interest. Given that, multivariate models for $Y_1, \ldots, Y_r$ are still worth considering in many situations. One such situation is longitudinal modeling with missing covariates. Notice that the multivariate model (2) for $Y_1, \ldots, Y_r$ implies a conditional univariate model of the form (3) for Y_r, where the covariate matrix X_i has been augmented to include columns for $Y_1, \ldots, Y_{r-1}$. When missing values occur on $Y_1, \ldots, Y_{r-1}$, a full parametric model for $Y_1, \ldots, Y_r$ provides a vehicle for inference in the conditional univariate submodel.

More generally, a full multivariate model for $Y_1, \ldots, Y_r$ can be quite useful when imputing for nonresponse in multivariate panel data. Imputation, especially multiple imputation (Rubin, 1987), has many important advantages over other methods for handling nonresponse. If values for the missing responses can be imputed in a statistically sound manner, the imputed dataset may be used for a variety of subsequent analyses. Many multivariate incomplete-data problems that were formerly troublesome can now be handled quite routinely through model-based multiple imputation (Schafer, 1996). In a multivariate panel

study, an imputation model should simultaneously preserve the relationships among variables measured for a subject at a single point in time, and among measurements of the same variable for a subject at different points in time. Multivariate mixed-effects models such as (2) are a natural choice, because they can effectively pool information within and across panels without a massive proliferation of parameters. The assumptions of a stable residual covariance matrix Σ and errors that are conditionally (given b_i) independent across time seems especially helpful; more general structures may be computationally troublesome or difficult to estimate (see Section 5). When this model is used for imputation, only the variables to be imputed need be included among $Y_1, \ldots, Y_r$; additional covariates that are completely observed may be incorporated into X_i or Z_i without distributional assumptions.

A motivating example, to be discussed in Section 4, comes from a study of adolescent substance use. For a period of six years, school children received questionnaires designed to measure attitudes and behaviors regarding the use of controlled substances. Researchers wanted to examine interrelationships among three time-varying covariates: a composite measure of self-reported alcohol use (Y_1), and measures of the perceived positive (Y_2) and negative (Y_3) consequences of alcohol use. Large amounts of data were missing by design, because Y_2 and Y_3 were measured for at most a subsample of students in each year. Using the techniques described below, values for the missing items were multiply imputed, allowing us to subsequently fit a conventional linear growth-curve model for alcohol use given the perceived consequences of use.

A recent paper by Liu, Taylor and Belin (1995) discussed the use of a multivariate model similar to (1) for imputation of missing covariates in longitudinal studies. Their model was less general, however, because it imposed special structure upon X_i, Z_i, and Σ. In particular, they assumed a diagonal form for Σ which is often unrealistic and undesirable. Correlations among the columns of ϵ_i can be a crucial aspect of an imputation procedure, because individual-level deviations from a norm in one variable may be highly predictive of

deviations on another variable. Imputing under a multivariate model that does not allow residual correlations among $Y_1, \ldots, Y_r$ may be essentially no different from imputing each variable Y_j separately under a univariate model. In the adolescent substance-use example of Section 4, the nonzero correlations among the three time-varying covariates are crucial for predicting a child's missing value for Y_1 when Y_2 and/or Y_3 are observed, and vice-versa.

Without missing data, techniques for fitting the multivariate model (1) would be relatively straightforward extensions of existing methods for the univariate case. When missing values occur within $y_1, \ldots, y_m$ in arbitrary patterns, however, direct likelihood-based inferences about the unknown parameters $\theta = (\beta, \Sigma, \Psi)$ may be difficult to obtain. Section 2 discusses general computational strategies for fitting the multivariate linear mixed model. Section 3 presents a Gibbs sampler that may be used to create model-based multiple imputations of the missing data for subsequent analyses. The technique is applied to substance-use data in Section 4, and Section 5 presents further discussion on the use of this model and many possible extensions.

2 Strategies for model fitting

Let $Y = (y_1, \ldots, y_m)$ denote the complete data without missing values. If Y were seen, inferences about the parameters $\theta = (\beta, \Sigma, \Psi)$ could be based on a likelihood function proportional to the product $(i = 1, \ldots, m)$ of the normal density functions implied by (2). The fixed effects β can be removed from this likelihood function in one of two ways: profiling, in which β is replaced by its conditional maximum given (Σ, Ψ); and marginalizing, in which the likelihood is replaced by its indefinite integral with respect to β. Both the profile and marginal likelihoods can be written in closed form as functions of the generalized least-squares estimate for β given (Σ, Ψ). Maximizing the former produces ordinary maximum-likelihood (ML) estimates, whereas maximizing the latter leads to restricted maximum-likelihood (RML) estimates.

For the univariate ($r = 1$) version of this model, Lindstrom and Bates (1988) present Newton-Raphson algorithms for ML and RML estimation. Newton-Raphson has excellent local convergence behavior but requires careful implementation. The calculations required to obtain derivatives of the loglikelihood at each iteration are complex and can be quite expensive. The algorithms of Lindstrom and Bates (1988) are finely tuned for the univariate model, but they do not generalize easily to the multivariate case unless we assume that Ψ has a special patterned structure, $\Psi = \Sigma \otimes \Upsilon$ for some $q \times q$ matrix Υ. This structure, which forces the correlation matrices for the r columns of b_i to be identical, seems quite unrealistic in many situations. Consider, for example, a linear growth model in which the slopes and intercepts for each variable $Y_1, \ldots, Y_r$ vary by subject. The correlation between the slope and intercept of any variable Y_j expresses the degree to which individuals with high initial values of Y_j tend to also have high rates of growth for Y_j; there may be no a priori reason to believe that these tendencies should be identical, especially when the variables $Y_1, \ldots, Y_r$ are very different in nature.

Simpler methods for ML and RML estimation are based on variants of the EM algorithm. EM relies on the fact that if the random effects $B = (b_1^V, \ldots, b_m^V)^T$ were seen, the likelihood function would factor into distinct likelihoods for Ψ and (β, Σ),

$$L(\theta \mid Y, B) \;=\; L(\Psi \mid B)\, L(\beta, \Sigma \mid Y, B), \tag{4}$$

each of which can be maximized quickly without iteration. EM algorithms tend to be quite stable but may converge very slowly; in many problems, hundreds or even thousands of iterations are required. EM-type algorithms for ML and RML estimation in the univariate case were given by Laird and Ware (1982) and Laird, Lange, and Stram (1987). As pointed out by Jennrich and Schluchter (1986) and Liu and Rubin (1995), many variants of EM are possible in the univariate case; not all of these generalize easily to the multivariate case.

The key feature of EM is that at each iteration, the sufficient statistics in (4) pertaining to B must be replaced by their conditional expectations given Y and the current estimate

of θ. In the multivariate model, the pairs (y_i, b_i) are distributed according to

$$y_i{}^V \mid b_i, \theta \;\sim\; N(\,(X_i\beta + Z_i b_i)^V,\, (\Sigma \otimes I_{n_i})\,), \tag{5}$$

$$b_i{}^V \mid \theta \;\sim\; N(0, \Psi), \tag{6}$$

independently for $i = 1, \ldots, m$. It follows from Bayes's Theorem that $b_i{}^V \mid y_i, \theta \sim N(\tilde{b}_i{}^V, \Gamma_i)$, where

$$\tilde{b}_i{}^V \;=\; \Gamma_i\,(\Sigma^{-1} \otimes Z_i^T)\,(y_i - X_i\beta)^V, \tag{7}$$

$$\Gamma_i \;=\; (\,\Psi^{-1} + (\Sigma^{-1} \otimes Z_i^T Z_i)\,)^{-1}. \tag{8}$$

Calculating Γ_i by (8) requires inversion of $rq \times rq$ matrices and is the preferred method in most cases where $q < n_i$. The sufficient statistics for B required by EM are linear in the elements of B and B^TB, whose expectations are $\tilde{B} = (\tilde{b}_1{}^V, \ldots, \tilde{b}_m{}^V)^T$ and $\sum_{i=1}^m (\Gamma_i + \tilde{b}_i{}^V (\tilde{b}_i{}^V)^T)$, respectively.

Now consider what happens when portions of $Y = (y_1, \ldots, y_m)$ are ignorably missing. Let $y_{i(obs)}$ and $y_{i(mis)}$ denote the observed and missing parts of y_i, respectively, and let $Y_{obs} = \{y_{i(obs)}\}$ and $Y_{mis} = \{y_{i(mis)}\}$. The simplest EM-type algorithms for ML and RML estimation still rely on the factorization (4). At each iteration, however, one must now find the conditional expectation given Y_{obs} of sufficient statistics that are linear and quadratic functions of b_i and $y_{i(mis)}$. From (5)–(6) we see that $y_i{}^V$ and b_i^V are jointly normal with covariance matrix

$$\begin{bmatrix} (I_r \otimes Z_i)\Psi(I_r \otimes Z_i)^T + (\Sigma \otimes I_{n_i}) & (I_r \otimes Z_i)\Psi \\ \Psi(I_r \otimes Z_i)^T & \Psi \end{bmatrix}. \tag{9}$$

To find the expectations necessary for EM, one would have to repeatedly apply a sweep operator or similar orthogonalization method to these matrices of dimension $(rq + rn_i) \times (rq + rn_i)$ for $i = 1, \ldots, m$. Without imposing further structure (e.g. equality of the Z_i) on the model, the computations for even the simplest variants of EM can thus be exceedingly expensive.

3 Inference by multiple imputation

In typical applications, many of the parameters in this multivariate model are a nuisance, and obtaining quality estimates of every component of θ is not of high priority. Rather than attempting direct likelihood-based inferences about θ, let us consider inference by multiple imputation. In multiple imputation, one must generate k independent draws $Y_{mis}^{(1)}, \ldots, Y^{(k)}$ from a posterior predictive distribution of the missing data,

$$P(Y_{mis} \mid Y_{obs}) = \int P(Y_{mis} \mid Y_{obs}, \theta)\, P(\theta \mid Y_{obs})\, d\theta, \tag{10}$$

where $P(\theta \mid Y_{obs})$ is proportional to the product of the observed-data likelihood function

$$P(\theta \mid Y_{obs}) = \int L(\theta \mid Y)\, dY_{mis}$$

and a prior density function $\pi(\theta)$. After imputation, the resulting k versions of the complete data are separately analyzed using complete-data methods, and the results are combined 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. $k = 5$) of imputations. Methods for combining the results of the complete-data analyses are reviewed by Schafer (1996).

Except in trivial situations, the posterior predictive distribution (10) cannot be simulated directly. It is possible, however, to create random draws of Y_{mis} from $P(Y_{mis} \mid Y_{obs})$ using 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 methods are given by Gelfand and Smith (1990); Smith and Roberts (1993); Tanner (1993); and in the chapters of Gilks, Richardson, and Spiegelhalter (1996). Applications of MCMC to univariate linear mixed models have been made by a number of authors, including Gelfand *et al.* (1990); Zeger and Karim (1991); Liu and Rubin (1995); and Carlin (1996). Like EM, these MCMC methods rely simplifications to the likelihood

that result when the random effects are assumed known. Unlike EM, however, MCMC allows us to circumvent manipulations on the large matrices (9) by alternately conditioning on simulated values of the random effects and the missing data.

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 the current version of the unknown parameter $\theta^{(t)} = (\beta^{(t)}, \Sigma^{(t)}, \Psi^{(t)})$ and the missing data $Y_{mis}^{(t)}$ are updated in three steps:

$$b_i^{(t+1)} \sim P(b_i \mid Y_{obs}, Y_{mis}^{(t)}, \theta^{(t)}), \quad i = 1, \ldots, m; \tag{11}$$

$$\theta^{(t+1)} \sim P(\theta \mid Y_{obs}, Y_{mis}^{(t)}, B^{(t+1)}); \tag{12}$$

$$y_{i(mis)}^{(t+1)} \sim P(y_{i(mis)} \mid Y_{obs}, B^{(t+1)}, \theta^{(t+1)}) \quad i = 1, \ldots, m. \tag{13}$$

Given starting values $\theta^{(0)}$ and $Y_{mis}^{(0)}$, these three steps define a Gibbs sampler in which the sequences $\{\theta^{(t)}\}$ and $\{Y_{mis}^{(t)}\}$ converge in distribution to $P(\theta \mid Y_{obs})$ and $P(Y_{mis} \mid Y_{obs})$, respectively.

This is not the only Gibbs sampler 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 (11)–(13) could be carried out without conditioning on simulated values of Y_{mis} or B then the algorithm could be made to converge more quickly. 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 usually not worthwhile. The three-step algorithm (11)–(13) is actually among the slowest to converge in terms of number of iterations required, but iterations can be executed on a computer quickly provided that sufficient physical memory is available to store Y_{obs}, $Y_{mis}^{(t)}$, and the covariate matrices X_i and Z_i. If the algorithm is believed to have converged to stationarity by T cycles, then k imputations of Y_{mis} can be generated in kT cycles. Convergence can be informally assessed by examining the time-series plots, autocorrelations, etc. for functions of $\theta^{(t)}$. Formal and informal convergence diagnostics for MCMC are discussed

by Schafer (1996) and in the chapters of Gilks, Richardson, and Spiegelhalter (1996).

Implementation of (11)–(13) requires us to specify 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, let us apply independent inverse-Wishart distributions $\Sigma^{-1} \sim W(\nu_1, \Lambda_1)$ and $\Psi^{-1} \sim W(\nu_2, \Lambda_2)$, where $W(\nu, \Lambda)$ denotes a Wishart with $\nu > 0$ degrees of freedom and mean $\nu\Lambda > 0$. These priors are proper provided that $\nu_1 \geq r$ 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 based on 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, deriving each of the distributions in (11)–(13) becomes a straightforward application of classical Bayesian methods. The random effects b_i in (11) are drawn from multivariate normal distributions with means and covariances calculated as in (7)–(8). Simulation of θ in (12) proceeds as follows: First, draw Ψ^{-1} from a Wishart distribution with parameters $\nu_2' = \nu_2 + m$ and $\Lambda_2' = (\Lambda_2^{-1} + B^T B)^{-1}$, respectively. 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 $\hat{\varepsilon} = y_i - X_i\hat{\beta} - Z_i b_i$, 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 matrix $\Lambda_1' = \left(\Lambda_1^{-1} + \sum_{i=1}^{m} \hat{\varepsilon}_i^T \hat{\varepsilon}_i\right)^{-1}$. 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 (13) 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 $y_{i(mis)}$.

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 the y_i. If the missing portions of Y 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. 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). 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 (13) to produce a fully completed dataset.

This Gibbs sampler has been implemented by the author in Fortran-77 as a function within the statistical languages S and Splus (Becker, Chambers, and Wilks, 1988). A sequence of $T \geq 1$ Gibbs cycles is performed with a single Fortran call; the function returns the final imputed dataset $(Y_{obs}, Y_{mis}^{(T)})$ and the history $\theta^{(1)}, \ldots, \theta^{(T)}$ of parameter iterates. Starting values for θ and Y_{mis} may be supplied, or the function may be allowed to choose its own starting value. Source code and documentation for this function will soon be available at the S archive in Statlib, the statistical software distribution service located at Carnegie Mellon University (`http://lib.stat.cmu.edu/S/`). The package will be called `ipan`, for imputation of multivariate panel data.

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

4 Application: Adolescent Alcohol Prevention Trial

Data for this example were drawn from the Adolescent Alcohol Prevention Trial, a longitudinal school-based intervention study of substance use in the Los Angeles area (Hansen and Graham, 1991). Attitudes and behaviors pertaining to the use of alcohol, tobacco, and marijuana were measured by self-report questionnaires administered yearly in grades 5–10. The data exhibit typical rates of uncontrolled nonresponse due to absenteeism, attrition, etc. which we will assume to be ignorable; this assumption has been given careful consideration and is not entirely implausible (Graham, Hofer, and Piccinin, 1994). In addition, large amounts of truly ignorably missing data arose by design, because each student received only a subset of the attitudinal items in any year; in some years, certain attitudinal questions were omitted entirely. For the present analysis, we examined a cohort of $m = 3,574$ children and focused attention on three variables: `DRINKING`, a composite measure of self-reported alcohol use; `POSCON`, the perceived positive consequences of alcohol use; and `NEGCON`, the perceived negative consequences of use. `DRINKING` appeared on the questionnaire every year, whereas `POSCON` was omitted in grade 8 and `NEGCON` was omitted in grades 8–10. Missingness rates for the three variables by grade are shown in Table 1; observed means and standard deviations appear in Table 2.

An analysis was performed to assess the possible influences of `POSCON` and `NEGCON` on `DRINKING`. In this analysis, missing responses were imputed under a multivariate linear growth model with random slopes and intercepts for each of the $r = 3$ variables, plus fixed effects for gender on both the slope and intercept. Each X_i matrix had $p = 4$ columns

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)	—	—	—

corresponding to an intercept, grade, gender, and gender × 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, 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.

Due to the high rates of missing information, it was anticipated that the Gibbs sampler would converge slowly. To assess convergence, the algorithm was run for an initial 2,000 cycles under a very mild prior with $\nu_1 = 3$, $\Lambda_1^{-1} = 3I$, $\nu_2 = 6$, $\Lambda_2^{-1} = 6I$. Time-series plots and sample autocorrelations for the components of θ were then examined. As anticipated, the elements of Ψ pertaining to the slopes and intercepts of NEGCON were among the slowest to converge because of the extreme sensitivity of these parameters to missing data. Based on this exploratory run, it appeared that several hundred cycles might be sufficient to achieve approximate stationarity. The Gibbs sampler was then run for an additional 9,000 cycles,

with the simulated value of Y_{mis} stored at cycles 2,000, 3,000, ..., 11,000. Autocorrelations estimated from cycles 1,001–11,000 verified that the dependence in all components of θ had indeed died down by lag 200, so the ten stored imputations could be reasonably regarded as independent draws from $P(Y_{mis} \mid Y_{obs})$. Each 1,000 cycles required approximately 17 minutes on a Sun UltraSPARC-1 workstation, approximately one cycle per second.

After imputation, the data were analyzed by a conventional linear growth-curve model for the logarithm of (`DRINKING` + 5). The model was a version of (3) with fixed effects for gender, grade, gender × grade, `POSCON` and `NEGCON`, plus random intercepts and slopes for grade. ML estimates were computed for each imputed dataset using an ECME algorithm, an extension of EM described by Liu and Rubin (1994). In this version of ECME, the parameters were partitioned as $\theta = (\theta_1, \theta_2)$ where $\theta_1 = (\beta, \sigma^2)$ and $\theta_2 = \Psi/\sigma^2$ (here σ^2 denotes the univariate version of Σ). Each cycle of ECME consisted of (a) an E-step, in which the conditional expectations of $B = (b_1, \ldots, b_m)^T$ and $B^T B$ given Y were calculated under the current value of θ; (b) a constrained maximization of the expected loglikelihood for θ_2 given the previous estimate of θ_1, in which $B = (b_1, \ldots, b_m)^T$ and $B^T B$ are replaced by their expectations; and (c) a constrained maximization of the actual loglikelihood for θ_1 given the updated estimate of θ_2. The updating formulas are

$$
\begin{aligned}
V_i^{(t)} &= \left({\theta_2^{(t)}}^{-1} + Z_i^T Z_i \right)^{-1}, \\
\tilde{b}_i^{(t)} &= V_i^{(t)} Z_i^T (y_i - X_i \beta^{(t)}), \\
W_i^{(t)} &= I_{n_i} - Z_i V_i^{(t)} Z_i^T, \\
\theta_2^{(t+1)} &= \frac{1}{m \sigma^{2(t)}} \sum_{i=1}^{m} \left(\tilde{b}_i^{(t)} \tilde{b}_i^{(t)T} + V_i^{(t)} \right), \\
\beta^{(t+1)} &= \left(\sum_{i=1}^{m} X_i^T W_i^{(t)} X_i \right)^{-1} \left(\sum_{i=1}^{m} X_i^T W_i^{(t)} y_i \right), \\
\sigma^{2(t+1)} &= N^{-1} \sum_{i=1}^{m} (y_i - X_i \beta^{(t+1)})^T W_i^{(t)} (y_i - X_i \beta^{(t+1)}),
\end{aligned}
$$

where $N = \sum_{i=1}^{m} n_i$. This simple algorithm, which does not seem to have appeared before in the literature, ran slightly faster than any of the three ECME algorithms described by

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

Liu and Rubin (1995) on this dataset and several others. Another virtue of this algorithm is that the value of the actual loglikelihood function at each iteration is available essentially no cost. Except for additive constants, the loglikelihood can be shown to be

$$l(\theta^{(t)} \mid Y) \;=\; -\frac{N}{2}\log\sigma^{2(t)} \;-\; \frac{m}{2}\log|\theta_2^{(t)}| \;+\frac{1}{2}\sum_{i=1}^{m}\log|V_i^{(t)}|\,, \tag{14}$$

and the determinants in (14) can be obtained as byproducts of the inversions required for $V_i^{(t)}$.

Using this algorithm, ML estimates were quickly obtained from the ten imputed datasets; convergence of the parameters to four significant figures required an average of just 36 iterations. Standard errors for the fixed effects were obtained from the final value of $\sigma^2(\sum_{i=1}^m X_i^T W_i X_i)^{-1}$. The ten sets of fixed-effects estimates and their standard errors were then combined using Rubin's (1987) rules for multiple-imputation inference for scalar estimands; these and other rules for combining multiply-imputed analyses are reviewed by Schafer (1996). 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 shows the estimated percentage of missing information for each estimand as

derived by Rubin (1987). The high rates of missing information indicate that the 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 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 $\times$ 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 or no discernible effect.

5 Discussion and extensions

The multivariate mixed model (1) is a natural extension of the simple univariate model (3) which has been quite popular in the analysis of longitudinal data. The imputation procedures described in Section 3 are appropriate for longitudinal analyses with partially missing covariates, when those covariates are going to be incorporated into an analytic model as

fixed effects. These methods are also appropriate for multivariate cross-sectional studies where units are nested within naturally occurring groups (e.g. children within schools). The algorithm and software described in this article provide a principled solution to missing-data problems for this somewhat limited but important class of analyses.

The imputation model and Gibbs sampler can be extended in a number of important ways. The use of an unstructured covariance matrix Ψ for the random effects may be limiting in situations where some aspects of Ψ may be poorly estimated—for example, in multivariate cluster samples with many variables, many units per cluster, but relatively few clusters. A more parsimonious block-diagonal structure, which assumes that the random effects pertaining to the r response variables are independent, can be handled easily. Under a block-diagonal structure, the likelihood function in (4) pertaining to Ψ factors into r distinct likelihoods for the diagonal blocks, so a Gibbs sampler can draw these blocks independently. Another extension which can be easily implemented pertains to linear models with additional random effects due to higher levels of clustering; this would arise, for example, in multivariate studies where individuals are grouped into larger units and multiple observations on individuals are taken over time. Both of these features will be incorporated into future versions of the software.

We are currently investigating a number of additional extensions the model. The first extension pertains to columns of y_i that are necessarily constant across the rows $1, \ldots, n_i$. In longitudinal studies, these columns would represent covariates that do not vary over time; in clustered applications, they would represent characteristics of the clusters rather than the units nested with them. 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 must be explicitly modeled for purposes of imputation. If we are willing to impose a simple parametric distribution on these covariates (e.g. multivariate normal), then it will be straightforward to extend the Gibbs sampling procedure to impute

these as well.

Another useful extension involves interactions among the columns of y_i. The multivariate normal model allows only simple linear associations among the variables $Y_1, \ldots, Y_r$, but in many studies one would like to preserve and detect certain nonlinear associations and interactions. In the data example of Section 4, for example, it may have been useful to see whether the strong effect of `POSCON` on `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 for future research.

Finally, it will be important to extend the imputation procedures to include time-varying responses that are categorical. Under the current procedure, ordinal responses can be handled in an ad hoc fashion, imputing under a normal model and rounding off the results to the nearest category. Some evidence suggests that ad hoc rounding procedures often work well in practice (Schafer, 1996). In other situations, however, a normal model will be clearly unacceptable—for example, with nominal (unordered) responses or binary variables that are heavily skewed. Imputation methods for multivariate datasets with continuous and/or categorical variables (Schafer, 1996) should be extended to include random effects that arise from longitudinal or clustered structure.

In the current model the rows of each response matrix y_i are assumed to be conditionally independent given b_i with common covariance matrix Σ. This assumption has been relaxed by Jennrich and Schluchter (1986), Lindstrom and Bates (1988), and others in the univariate case to allow a residual covariance matrix of the form $\sigma^2 V_i$, where V_i has a simple (e.g. autoregressive or banded) pattern dependent upon one or more unknown parameters. Sensible multivariate extensions of these patterned covariance structures to a

tends to produce models and algorithms that are complex even apart from missing data. For example, the obvious extension of $\epsilon_i^V \sim N(0, (\Sigma \otimes I_{n_i}))$ to $\epsilon_i^V \sim N(0, (\Sigma \otimes V_i))$ seems too restrictive for many longitudinal datasets, because the response variables $Y_1, \ldots, Y_r$ are then required to have identical autocorrelations. Accounting for autocorrelated residuals in a sensible manner may prove be a daunting task in the multivariate case. In practice, 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.

6 References

Becker, R.A., Chambers, J.M. and Wilks, A.R. (1988) *The New S Language: A programming environment for data analysis and graphics*. Wadsworth and Brooks/Cole Advanced Books and Software, Pacific Grove, CA.

Carlin, B.P. (1996) Hierarchical longitudinal modelling. *Markov Chain Monte Carlo in Practice* (eds. W.R. Gilks, S. Richardson and D.J. Spiegelhalter), 303–319, Chapman & Hall, London.

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.

Gelfand, A.E. and Smith, A.F.M. (1990) Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398–409.

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

Graham, J.W., Hofer, S.M., and Piccinin, A.M. (1994) Analysis with missing data in drug prevention research. *Advances in Data Analysis for Prevention Intervention Research* (eds. L.M. Collins and L.A. Seitz), 13–63, National Institute on Drug Abuse.

Hansen, W.B. and Graham, J.W. (1991), "Preventing alcohol, marijuana, and cigarette use among adolescents: 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.

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

Liu, C. and Rubin, D.B. (1994) The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence. *Biometrika*, **81**, 633–648.

Liu, C. and Rubin, D.B. (1995) Application of the ECME algorithm and the Gibbs sampler to general linear mixed models. *Proceedings of the* 17*th International Biometric Conference*, 97–107.

Liu, M., Taylor, M.G. and Belin, T.R. (1995) Multiple imputation and posterior simulation for multivariate missing data in longitudinal studies. *Computing Science and Statistics: Proceedings of the 27th Symposium on the Interface*, 521–529.

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.

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

Rubin, D.B. (1987) *Multiple Imputation for Nonresponse in Surveys.* J. Wiley & Sons, New York.

Schafer, J.L. (1996) *Analysis of Incomplete Multivariate Data.* Chapman & Hall, London, in press.

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 Series B*, **55**, 3–23.

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

Zeger, S.L. and Karim, M.R. (1991) Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association*, **86**, 79–86.

CHAPTER 12

Multiple Imputation With PAN

Joseph L. Schafer

Missing values are a nuisance in many research efforts but especially so in the collection and analysis of longitudinal data. Multiple occasions bring greater opportunities for missed measurements. Fortunately, missing data is one area where statisticians have made substantial progress in recent years. In this chapter, I present a strategy for analyzing incomplete longitudinal data by multiple imputation (Rubin, 1987; Schafer, 1997a).

Missing data pose a difficulty because the overwhelming majority of paradigms and software for statistical analysis assume that the input data are complete. For this reason, the quickest and most convenient method for handling incomplete observations is case deletion, that is, ignoring participants with missing information. Case deletion suffers from a number of serious drawbacks, which have been well documented (e.g., Little & Rubin, 1987). For multivariate analyses involving a large number of items case deletion can be very inefficient, discarding an unacceptably high proportion of participants; even if the per-item rates of missingness are low, few participants may have complete data for all items. Moreover, case deletion leads to valid inferences in general only when missing data are missing completely at random (MCAR), in the sense that the discarded cases are like a random subsample of all cases. If the discarded cases differ systematically from the rest, then the resulting estimates may have potentially serious bias.

A natural alternative to case deletion is *imputation*, the practice of replacing missing data with plausible values. Various forms of imputation have been applied in federal surveys and censuses for decades (Madow, Nisselson, & Olkin, 1983). Imputation has been the survey statistician's method of choice for handling *item nonresponse*, situations in which a participant provides some infor-

---

This research was supported by Grant 1-P50-DA10075 from the National Institute on Drug Abuse and by Grant 2R44CA65147-02 from the National Cancer Institute. I extend special thanks to John Graham for providing data from the Adolescent Alcohol Prevention Trial and advice on their analysis.

mation but fails to respond to one or more individual items on a questionnaire. Imputation is attractive because it apparently solves the missing-data problem at the outset; once the missing values have been imputed, the data set can be summarized and analyzed by familiar complete-data methods. Another attractive feature of imputation is its efficiency: Unlike case deletion, imputation allows one to make full use of the data at hand.

Methods of imputation range from simple procedures, such as mean substitution—replacing each missing value with the observed mean for that variable—to elaborate hot-deck algorithms that jointly replace missing items with data obtained from donor cases chosen to match the original on selected items (e.g., Bailey, Chapman, & Kasprzyk, 1985). In longitudinal data sets with substantial participant-to-participant variation, analysts have sometimes filled in missed measurements by linear interpolation, extrapolation, or "last value carried forward." Unless great care is taken, these ad hoc imputation procedures may seriously distort important aspects of the distribution of a variable or its relationships with other variables. In general, it is desirable for the distribution of imputed values to resemble the distribution of the observed values, particularly with respect to intervariable relationships.

Even if an imputation method successfully preserves important aspects of the data distributions, a potentially serious problem remains: Imputation adds fictitious information to a data set. If imputed values are treated the same way as observed values in subsequent analyses, then the resulting inferences will be artificially precise, because the imputed values are imperfect proxies for the data they represent. With single imputation, there is no simple way to reflect uncertainty in the imputed values. In response, Rubin (1987, 1996) proposed the method of multiple imputation, by which each missing value is represented by a set of $m > 1$ simulated values. Let $Y = (Y_{obs}, Y_{mis})$ denote a generic data set, in which Y_{obs} is the observed part and Y_{mis} is the missing part. Multiple imputation replaces Y_{mis} with a set of simulated draws $Y_{mis}^{(1)}, Y_{mis}^{(2)}, \ldots, Y_{mis}^{(m)}$ from a predictive probability distribution $P(Y_{mis} | Y_{obs})$ arising from a model. After multiple imputation, one has m simulated complete data sets, $Y^{(j)} = (Y_{obs}, Y_{mis}^{(j)})$, $j = 1, 2, \ldots, m$, which are analyzed with standard complete-data methods. The results are then combined, using simple arithmetic rules, to produce overall estimates and standard errors that account for missing-data uncertainty. I reviewed these rules (Schafer, 1997a) and demonstrate them in the example near the end of this chapter.

The key idea of multiple imputation is that it treats missing data as an explicit source of random variability over which to be averaged. The process of creating imputations, analyzing the imputed data sets, and combining the results is a Monte Carlo version of averaging the statistical results over the predictive distribution $P(Y_{mis} | Y_{obs})$. In practice, a large number of multiple impu-

tations are not required; sufficiently accurate results can often be obtained with $m \leq 10$.

Carrying out multiple imputation requires two sets of assumptions. First, one must propose a model for the distribution of Y. This data model should be plausible and should bear some relation to the type of analysis to be performed. For example, one could assume that the variables in the data set are jointly normally distributed. In the case of longitudinal analyses the model should be capable of preserving the correlation structure and time trends within individuals. The second set of assumptions pertains to the manner in which the missing values became missing. It is most common to assume that the missing data are missing at random (MAR) in the technical sense defined by Rubin (1976), which means that the probabilities of missingness may depend on the observed values Y_{obs} but not on the missing data Y_{mis}. The MAR assumption is primarily a mathematical convenience that allows one to perform imputation without explicitly modeling the missing-data mechanism. In practice, MAR is essentially untestable; it cannot be verified or contradicted by examination of the observed data. If the assumption seems *prima facie* implausible, then alternative procedures can be developed by modeling the probabilities of missingness. General techniques and software for creating multiple imputations under non-MAR models have not yet been developed; this is an important area for future research. Further discussion on the plausibility and ramifications of MAR was given by Little and Rubin (1987); Graham, Hofer, and Piccinin (1994); and Schafer (1997a).

Multiple imputation is not the only principled method for handling missing data. For parametric models, a main competitor is the technique of direct maximum likelihood, sometimes called *raw* or *full-information* maximum likelihood, which maximizes a likelihood function on the basis of the observed data Y_{obs} alone. This likelihood function may be written as

$$L(\theta|Y_{obs}) = \int L(\theta|Y_{obs}, Y_{mis})\, dY_{mis}, \qquad \textbf{(12.1)}$$

where θ represents the unknown parameters of the data model, and $L(\theta|Y_{obs}, Y_{mis})$ denotes the likelihood function that one would use if no data were missing. The integration in Equation 12.1 eliminates the dependence on Y_{mis}, broadening the likelihood function to reflect the additional uncertainty due to the fact that Y_{mis} is unknown. In effect, this integration is nearly the same as the averaging over $P(Y_{mis}|Y_{obs})$ that takes place in multiple imputation. Except in very simple problems, the likelihood function Equation 12.1 tends to be complicated, often requiring complicated numerical techniques or approximations. When carried out properly, direct maximum likelihood can be statistically more efficient than multiple imputation because it is a deterministic procedure; no simulation is

involved, so no extra variability is introduced into summary statistics. (In most cases, this extra randomness introduced by multiple imputation is quite minor.) In large samples, estimates and standard errors obtained by direct maximum likelihood and by multiple imputation tend to be very similar.

Applications of direct maximum likelihood are now common in longitudinal analyses. Modern algorithms for growth modeling as implemented in hierarchical linear modeling (HLM; Bryk, Raudenbush, & Congdon, 1996), Proc Mixed in SAS (Littell, Milliken, Stroup, & Wolfinger, 1996), and similar packages are designed for unbalanced data, where measurements on each participant may be taken at a different set of time points. Responses that are missing, either unintentionally or by design, are removed from the likelihood by integration as in Equation 12.1. An important limitation of these packages is that the missing values must be confined to the response variable; missing values on predictors are not allowed. If the individuals in the study have been assessed at a common set of occasions, models equivalent to those fit by HLM and Proc Mixed can be formulated using latent growth curves (McArdle, 1988; Meredith & Tisak, 1990; Willett & Sayer, 1994) and structural equations software. Two recent programs for structural equations, Mx (Neale, 1994) and Amos (Arbuckle, 1995), perform direct maximum likelihood from a raw data set with missing values. Missing data can be accommodated in other structural equations software by using the technique of multiple groups (Allison, 1987; Duncan & Duncan, 1994; Muthén, Kaplan, & Hollis, 1987). An advantage of the latent growth curve approach is that missing values may occur on predictors as well as the response; however, the measurements must be taken at a relatively small number of common time points.

When a direct maximum-likelihood procedure is available for a particular analysis, it may indeed be the most convenient and attractive method. Despite the increasing popularity of direct maximum likelihood, however, multiple imputation still offers some unique advantages for data analysts. First, it allows them to use their favorite models and software; an imputed data set may be analyzed by virtually any method that would be appropriate if the data were complete. As computing environments and statistical models grow increasingly complex, the value of using familiar methods and software should not be underestimated. Second, there are still many classes of problems for which no direct maximum-likelihood procedure is available. For example, in longitudinal analyses there is no direct maximum-likelihood method for incomplete covariates when occasions of measurement vary by individual.

A third reason why multiple imputation can be more attractive than direct maximum likelihood is that the separation of the imputation phase from the analysis phase lends a greater flexibility to the entire process. With multiple imputation the imputer is free to use additional variables that may be helpful for imputation but that are not of direct interest for the analysis. For example,

consider a covariate that helps to explain reasons for nonresponse. Using this variable in the imputation procedure tends to reduce bias in subsequent analyses, even in analyses that do not involve that variable.

Finally, an important advantage of multiple imputation over direct maximum likelihood is that it singles out missing data as a source of random variation distinct from ordinary sampling variability. The likelihood function Equation 12.1 lumps these two types of variability together; summary statistics (e.g., standard errors) derived from direct maximum likelihood do not reveal two sources. With multiple imputation, however, the overall uncertainty is formally partitioned into sampling variability and missing-data uncertainty. This partition immediately yields an estimated rate of missing information, which can be quite helpful for assessing the impact of missing data on inferences for any parameter of interest.

The purpose of this chapter is not to criticize direct maximum likelihood in favor of multiple imputation; rather, it is my hope that more analysts will recognize the important advantages offered by both of these modern missing-data methods and begin to use them instead of case deletion or other ad hoc procedures. In most real-life applications, missing data are not the main focus of scientific inquiry but an unpleasant nuisance. Missing data should be handled quickly and effectively but without compromising the integrity of the analytic results. Multiple imputation might not be the optimal choice for every analysis, but it is a handy statistical tool and a valuable addition to a researcher's methodological toolkit.

In the remainder of this chapter, I describe a method for creating multiple imputations in longitudinal databases. Previous algorithms and software for multiple imputation, as described in Schafer (1997a), have focused on missing data in general multivariate settings. In response to the specific need for longitudinal analyses, a library of algorithms called *PAN* has been developed for imputing multivariate panel data, where a group of variables is measured for individuals at multiple time points. Alternatively, PAN may be applied to clustered data where variables are measured at a single point for participants nested within some larger unit (e.g., students within classrooms). Future versions of the software will be able to handle repeated measures and clustering simultaneously.

PAN is at present available as a library of functions for the statistical programming language S-PLUS (MathSoft, Inc., 1997).[1] Current efforts are focused on developing a version of PAN that operates as a stand-alone program in the Windows 95/98/NT environment.

---

[1]This can be downloaded free of charge from http://www.stat.psu.edu/~jls/misoftwa.html

Suppose that a group of time-varying continuous variables $Y_1, Y_2, \ldots, Y_r$ is measured for individuals $i = 1, 2, \ldots, N$ at multiple occasions. The responses for participant i may be arranged as a matrix with one column for each variable and one row for each occasion,

$$y_i = \begin{bmatrix} y_{i11} & y_{i12} & \cdots & y_{i1r} \\ y_{i21} & y_{i22} & \cdots & y_{i2r} \\ \vdots & \vdots & \ddots & \vdots \\ y_{in_i1} & y_{in_i2} & \cdots & y_{in_ir} \end{bmatrix}, \tag{12.2}$$

where y_{ijk} denotes the value of variable Y_k at occasion j. The number of occasions n_i and their temporal spacing may vary by participant. I assume that missing values occur throughout the matrices $y_1, y_2, \ldots, y_m$ and that these missing values are MAR. The immediate goal is to multiply impute the missing values so that the data can be analyzed in a straightforward manner. Ultimately, the analyst may choose to regard one column of Equation 12.2 as a response and the other columns as potential predictors in a conventional growth model. For the moment, however, I regard all r columns of y_i as random responses and model them jointly for the purpose of imputation. I construct a multivariate growth model to describe the joint distribution of the variables $Y_1, Y_2, \ldots, Y_r$, possibly given other time-varying or static covariates that are fully observed and require no imputation.

The model used by PAN was designed to preserve the following relationships: (a) relationships among the variables $Y_1, Y_2, \ldots, Y_r$ within an individual at each time point. These are reflected by the covariances among the elements of any row of y_i. (b) Growth or change in any variable Y_j within an individual across time points. This growth is reflected by trends within the columns of y_i. (c) Relationships between the response variables $Y_1, Y_2, \ldots, Y_r$ and any additional participant-level (non-time-varying) covariates included in the model. The participant-level covariates may be continuous or categorical, but they must be fully observed; missing values on these non-time-varying variables are allowed in the current version. Missing values in time-varying covariates are allowed and will be imputed, provided that they are included among $Y_1, Y_2, \ldots, Y_r$.

PAN relies on a multivariate extension of a linear mixed-effects model that has been popular for nearly 20 years. The model is

$$y_i = X_i\beta + Z_ib_i + \varepsilon_i, \tag{12.3}$$

where $X_i(n_i \times p)$ and $Z_i(n_i \times q)$ are known covariate matrices, β contains regression coefficients common to all units, and b_i contains coefficients specific to unit i. Note that Equation 12.3 is a multivariate regression; β and b_i are

matrices with r columns, one column for predicting each of the variables Y_1, Y_2, . . . , Y_r, and $\boldsymbol{\varepsilon}_i$ is also a matrix with the same dimensions as $y_i(\eta_i \times r)$. The univariate ($r = 1$) version, which was proposed by Hartley and Rao (1967) and later popularized by Laird and Ware (1982), Jennrich and Schluchter (1986), Bryk and Raudenbush (1992), and others, is the basis for many of the linear growth models in use today. The coefficients $\boldsymbol{\beta}$ and b_i are often called "fixed effects" and "random effects," respectively.

With univariate versions of this model, it is common to assume that the random effects and residuals are independently drawn from normal populations, $b_i \sim N(0, \psi)$ and $\boldsymbol{\varepsilon}_i \sim N(0, \sigma^2 I)$, $i = 1, 2, \ldots, N$, where ψ is a $q \times q$ covariance matrix and I is the identity matrix ($n_i \times n_i$). For the multivariate case, one generalizes these assumptions to

$$\text{vec}(b_i) \sim N(0, \boldsymbol{\Psi}) \tag{12.4}$$

$$\text{vec}(\boldsymbol{\varepsilon}_i) \sim N[0, (\boldsymbol{\Sigma} \otimes I)], \tag{12.5}$$

where *vec* denotes the vectorization of a matrix by stacking its columns. The covariance matrix $\boldsymbol{\Psi}$ in Equation 12.4 has dimension $qr \times qr$, and the Kronecker product notation in Equation 12.5 indicates that the rows of $\boldsymbol{\varepsilon}_i$ are independently distributed as $N(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is $r \times r$.

In typical applications, the times of measurement are incorporated into X_i, and perhaps Z_i, as linear, quadratic, or higher order polynomials, and Z_i is a subset of the columns of X_i. For example, suppose that the first two columns of X_i are $(1, 1, \ldots, 1)^T$ and $(t_1, t_2, \ldots, t_{n_i})^T$, respectively, where $t_1, t_2, \ldots, t_{n_i}$ are the times of measurement for participant i; beyond these, X_i may have additional columns containing static or time-varying covariates for participant i. Setting Z_i equal to the first column of X_i produces a model of linear growth with intercepts randomly varying by individuals; setting Z_i equal to the first two columns of X_i produces random intercepts and slopes. Centering the distribution of b_i at zero causes $\boldsymbol{\beta}$ to become the population-averaged regression coefficients and the random effects $b_1, \ldots, b_m$ become perturbations due to interparticipant variation.

Note that in this multivariate model all of the covariates in X_i and Z_i appear as predictors for each of the columns of y_i. As a result, the same group of predictors and the same type of trend over time (e.g., linear mean growth with varying slopes and intercepts) are used to describe each of the response variables $Y_1, Y_2, \ldots, Y_r$. The actual coefficients for the response variables, as contained in the r columns of $\boldsymbol{\beta}$ and b_i, vary, but the same group of predictors is applied to each response. At first glance, this may appear to be a serious limitation of the model; in many scientific contexts there is no reason to believe that $Y_1, Y_2, \ldots, Y_r$ should depend on precisely the same set of covariates. One must remember, however, that the purpose of PAN is not to construct a theoretically

meaningful model but to impute missing responses in such a way that important relations are preserved. If a covariate appears in subsequent analyses as a potential predictor of one or more of the response variables $Y_1, Y_2, \ldots, Y_r$, then that covariate should be included in the imputation model, even though its effects on some of the responses may be irrelevant or null. No biases incur by using an imputation model that is larger or more general than necessary for any given analysis. For more discussion on the purpose of imputation modeling and the interplay between the imputer's and analyst's assumptions, see Meng (1994), Rubin (1996), and Schafer (1997a, chapter 4).

The current version of PAN allows two types of assumptions about Ψ, the covariance matrix for the participant-level random effects $b_1, b_2, \ldots, b_N$. One allows the Ψ matrix to be either (a) an unstructured or arbitrary covariance matrix or (b) a block diagonal covariance matrix of the form

$$\Psi = \begin{bmatrix} \Psi_1 & 0 & \cdots & 0 \\ 0 & \Psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Psi_r \end{bmatrix}, \tag{12.6}$$

where the nonzero blocks Ψ_j, $j = 1, \ldots, r$ are covariance matrices of size $q \times q$. The unstructured Ψ allows the random effects for any two responses Y_j and Y_k to be correlated, whereas the block-diagonal form assumes that the random effects for each response are independent of those for any other response.

The choice between these two depends on both theoretical and practical considerations. Suppose that $Y_1, Y_2 \ldots, Y_r$ represent achievement scores (mathematics, reading comprehension, etc.) recorded for schoolchildren over time, and one applies a model of linear growth with intercepts and slopes that vary by individual. If there is reason to believe that growth patterns for the various achievement scores are related—for example, that participants with high rates of increase for mathematics may also tend to have high rates of increase for reading comprehension—then it would be wise to use an unstructured Ψ. As the number of response variables grows, however, it often becomes impractical to estimate covariances among all of their random effects unless the number of participants is very large; to obtain a stable estimate for Ψ one may need to specify a block-diagonal structure. Unless the correlations among the random effects for some pairs of responses are unusually strong, the potential biases incurred by using a block-diagonal Ψ rather than an unstructured Ψ tend to be minor.

The basic strategy for specifying a PAN model can be summarized as follows. First, any time-varying covariates with missing values should be placed in the columns of y_i, regardless of whether they are treated as "responses" or "predictors" in later analyses. If a variable is to be imputed, then it must be

should be included in the columns of X_i and, possibly, Z_i. These include (a) variables that may be related to $Y_1, Y_2, \ldots, Y_r$ and (b) variables that may explain missingness on $Y_1, Y_2, \ldots, Y_r$. Placing a covariate in X_i allows it to influence the distribution of any or all of the variables $Y_1, Y_2, \ldots, Y_r$ in the population. Placing a time-varying covariate in both X_i and Z_i allows its degree of influence on $Y_1, Y_2, \ldots, Y_r$ to vary across individuals. Note that static or non-time-varying covariates (e.g., gender or pretest measures) should not be included in Z_i because it is impossible to estimate participant-specific effects for such variables. Finally, polynomial terms such as *1*, *time*, *time*^{2}, and so on, may be appended to X_i and Z_i as desired, to allow the mean levels of $Y_1, Y_2, \ldots, Y_r$ and the trends in these variables over time to vary across individuals. The choice of which terms to include will depend on what types of effects are believed to exist and what effects will be investigated in subsequent analyses.

Computational Algorithms

The computational engine of PAN is a Markov chain Monte Carlo (MCMC) algorithm called a *Gibbs sampler*. MCMC is a relatively new class of simulation techniques that are especially useful in Bayesian statistical analyses. A review of MCMC is beyond the scope of this chapter, but a gentle introduction is given by Casella and George (1992) and Schafer (1997a, chapters 3–4); more comprehensive references are the volume edited by Gilks, Richardson, and Spiegelhalter (1996) and the article by Gelfand and Smith (1990). Specific details and formulas for the computations used in PAN have been provided by me (Schafer, 1997b; Yucel & Schafer, 1998).

The MCMC algorithm in PAN is based on the observation that the model specified by Equations 12.3–12.5 has the following unknown components: the missing values in $y_1, y_2, \ldots, y_N$, the random effects $b_1, b_2, \ldots, b_N$, the fixed effects β, and the covariance matrices Σ and Ψ. For the purpose of imputation, I am interested only in simulating the missing data in $y_1, y_2, \ldots, y_N$; the other unknown quantities are merely a nuisance. To simulate the missing data properly, however, one must take into account the uncertainty in these other quantities and how it contributes to missing-data uncertainty. Expressing this uncertainty through mathematical formulas is difficult, so one accounts for the interdependence among the unknown quantities through a process of iterative simulation.

PAN simulates the unknown quantities in a three-step cycle.

1. Draw random values of $b_1, b_2, \ldots, b_N$ on the basis of some plausible assumed values for the missing data and the parameters β, Σ, and Ψ.
2. Draw new random values of the unknown parameters β, Σ, and

Ψ on the basis of the assumed values for the missing data and the values of $b_1, b_2, \ldots, b_N$ obtained in Step 1.

3. Draw new random values for the missing data given the values of $b_1, b_2, \ldots, b_N$ obtained in Step 1 and the parameters obtained in Step 2.

At the end of this cycle the parameters and missing data from Steps 2 and 3 become the values assumed in Step 1 at the start of the next cycle. Repeating Steps 1, 2, and 3 in turn defines a Markov chain, a sequence in which the distribution of the unknown quantities at any cycle depends on their simulated values at the previous cycle. The state of the process at Cycle 2 may be strongly correlated with its state at Cycle 1, but at subsequent Cycles 3, 4, 5, and so on, the relationship to the original state weakens. When a sufficient number of cycles has been taken to make the resulting state essentially independent of the original state, then the process is said to have *converged* or *achieved stationarity*. On convergence, the final simulated values for the missing data have in fact come from the distribution from which multiple imputations should be drawn.

This algorithm may be used to create m multiple imputations in the following way. Starting with some plausible initial values, run the Gibbs sampler for k cycles where k is large enough to ensure convergence, and take the final simulated version of the missing data as the first imputation; then return to the original starting values, run the Gibbs sampler for another k cycles, and take the final simulated version of the missing data as the second imputation; and so on. This method requires m runs of length k cycles each. Another and perhaps more convenient way is to perform one long run of mk cycles, saving the simulated values of the missing data after cycle k, $2k$, . . . , mk as the m imputations. The latter method differs from the former only in that the final values from each subchain of length k become the starting values for the next subchain of length k.

It is important to note that convergence of an MCMC procedure means convergence to a probability distribution rather than convergence to a set of fixed values. To say that the algorithm has converged by k cycles actually means that the random state of the process at cycle $t + k$ is statistically independent of its state at cycle t for $t = 1, 2, \ldots$. After running the Gibbs sampler, one can examine the output stream over many cycles to see how many are needed to achieve this independence. Suppose that one collects and stores the simulated values for one parameter θ (a particular element of β, Ψ, or Σ) over a large number C of consecutive cycles. These values $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(C)}$ can be regarded as a time series. The lag-k autocorrelation, which is the correlation between pairs $\theta^{(t)}$ and $\theta^{(t+k)}$ ($t = 1, 2, \ldots, C - k$), can be calculated for various values of k to determine how large k must be for the correlations to die down. In principle, one should examine autocorrelations for each parameter in the model

and identify a value of k large enough to guarantee that the lag-k autocorrelations for all parameters are effectively zero. In my experiences with real data, however, I have found that the greatest levels of serial dependence are almost always seen in variance and covariance parameters, and in particular within the elements of Ψ. It is usually sufficient to monitor the behavior of the elements of Ψ because it is with respect to these parameters that the algorithm tends to converge the most slowly. For more discussion on monitoring the convergence of MCMC algorithms, see Schafer (1997a, chapter 4).

The rate of convergence of this Gibbs sampler is influenced by a combination of factors pertaining to the data and the model. First, it is affected by the amounts and patterns of missing data in the matrices $y_1, y_2, \ldots, y_N$; greater rates of missing information lead to slower convergence. It is also affected by one's ability to estimate the individual random effects $b_1, b_2, \ldots, b_N$; if estimates of random effects are highly variable, then convergence is slowed. Finally, convergence behavior is also influenced by the number of participants (N). As the sample size grows, the distribution of the random Ψ matrix at each cycle becomes more tightly concentrated around the sample covariance matrix of b_1, $b_2, \ldots, b_N$ from the previous cycle. As this distribution becomes tighter, the elements of Ψ are less free to wander away from their values at the previous cycle, producing higher correlations from one cycle to the next. It is somewhat ironic that the algorithm converges more slowly as one's ability to estimate the parameters increases. With a large number of participants and a small number of occasions per participant, it is not uncommon for the Gibbs sampler to require several hundred or even 1,000 cycles to converge. Slow convergence is not necessarily a problem, however, because in most cases only a few imputations are necessary. If $k = 1{,}000$ cycles are needed to achieve stationarity, then five imputations can be produced in 5,000 cycles, which even for a large data set requires no more than a few hours on a personal computer.

In addition to deciding how many cycles are needed, the user must also specify Bayesian prior distributions for the covariance matrices Ψ and Σ. Bayesian procedures, which are becoming increasingly popular in many areas of statistical analyses, treat unknown parameters as random variables and assign prior probability distributions to them to reflect one's knowledge of or belief about the parameters before the data are seen. An excellent introduction to the Bayesian statistical paradigm was given by Novick and Jackson (1974); for a modern overview of Bayesian modeling and computation, see Gelman, Rubin, Carlin, and Stern (1995). Some statisticians tend to prefer Bayesian procedures on principle, whereas others avoid them on principle. I hold a pragmatic view, accepting the prior distribution simply as a mathematical device that allows one to generate the imputations in a principled fashion. In applications, I like to use prior distributions that are weak or highly dispersed, reflecting a state of relative ignorance about model parameters. Weak priors tend to minimize

the subjective influence of the prior, allowing the observed data to speak for themselves.

The prior distribution most commonly applied to a covariance matrix is the inverted Wishart distribution. The Wishart, a natural generalization of the chi-square to random matrices, is discussed in standard texts on multivariate analysis (e.g., Anderson, 1984; Johnson & Wichern, 1992). The prior distribution for $\boldsymbol{\Sigma}$ is

$$\boldsymbol{\Sigma}^{-1} \sim W(a, B), \tag{12.7}$$

where $W(a, B)$ denotes a Wishart with a degrees of freedom and scale B. The scale is a symmetric, positive definite matrix with the same dimensions ($r \times r$) as $\boldsymbol{\Sigma}$. The degrees of freedom, which should be greater than or equal to r, govern the spread or variability; lower values of a make the distribution more dispersed. The user of PAN must provide numeric values for a and B^{-1}. Our usual practice is to set $a = r$ to make the prior as dispersed as possible and then to set $B^{-1} = a\hat{\boldsymbol{\Sigma}}$, where $\hat{\boldsymbol{\Sigma}}$ is a reasonable prior guess or estimate of $\boldsymbol{\Sigma}$. If a guess for $\boldsymbol{\Sigma}$ is unavailable, the data themselves may be used to obtain one. Yucel and Schafer (1998) recently developed a new expectation–maximization algorithm for calculating maximum-likelihood estimates of the parameters $\boldsymbol{\beta}$, $\boldsymbol{\Psi}$, and $\boldsymbol{\Sigma}$ from the incomplete data. Running this EM algorithm before the Gibbs sampler is an excellent way to obtain a reasonable guess for $\boldsymbol{\Sigma}$.

In a similar fashion, I also use inverted Wishart prior distributions for the between-subjects covariance matrix $\boldsymbol{\Psi}$. If $\boldsymbol{\Psi}$ is unstructured, one assumes $\boldsymbol{\Psi}^{-1} \sim W(c, D)$ where D is a $qr \times qr$ matrix and $c > qr$. My usual practice is to set $c = qr$ and $D^{-1} = c\hat{\boldsymbol{\Psi}}$, where $\hat{\boldsymbol{\Psi}}$ is a prior guess or estimate of $\boldsymbol{\Psi}$. If $\boldsymbol{\Psi}$ is taken to be block diagonal as in Equation 12.6, then independent inverted Wishart prior distributions are applied to the nonzero blocks, $\boldsymbol{\Psi}_j^{-1} \sim W(c_j, D_j)$, $j = 1, \ldots, r$, where $c_j \geq q$. To make the priors weak, one sets $c_j = q$ and $D_j^{-1} = c_j\hat{\boldsymbol{\Psi}}_j$ where $\hat{\boldsymbol{\Psi}}_j$ is an estimate or guess for $\boldsymbol{\psi}_j$. The EM algorithm described by Yucel and Schafer (1998) provides a maximum-likelihood estimate for an unstructured $\boldsymbol{\psi}$ or estimates of the submatrices $\boldsymbol{\Psi}_1, \ldots, \boldsymbol{\Psi}_r$ when $\boldsymbol{\Psi}$ is block diagonal.

An Example: Expectancies and Alcohol Use in the Adolescent Alcohol Prevention Trial

The Adolescent Alcohol Prevention Trial (AAPT) was a longitudinal school-based intervention study of substance use carried out in the Los Angeles area (Hansen & Graham, 1991). In one panel of AAPT, attitudes and behaviors pertaining to the use of alcohol, tobacco, and marijuana were measured by self-report questionnaires administered yearly in Grades 5–10. The data exhibit

typical rates of uncontrolled nonresponse due to absenteeism, attrition, and so on, which I assume to be MAR. This assumption has been given careful consideration by the researchers and appears to be plausible; for example, much of the attrition is due to students moving to other schools or districts, which is at most only weakly associated with substance use patterns (Graham et al., 1994).

In addition to this uncontrolled nonresponse, large amounts of truly MAR missing data (MCAR, in fact) arose by design. The AAPT study made use of an innovative three-form design in which each student received only a subset of the items in any year, as described in chapter 11 of this volume, by Graham, Taylor, and Cumsille. In some years, certain items were omitted entirely. For the present analysis, I examine a cohort of $m = 3,574$ children and focus attention on three variables: "drinking," a composite measure of self-reported alcohol use; POSCON, a measure of the degree to which the student perceives that alcohol use has positive consequences; and NEGCON, a measure of the perceived negative consequences of use. Drinking appeared on the questionnaire every year, where POSCON was omitted in Grade 8 and NEGCON was omitted in Grades 8–10. Missingness rates for the three variables by grade are shown in Table 12.1; observed means and standard deviations appear in Table 12.2.

My analysis will focus on the possible influences of POSCON and NEGCON on drinking. Without missing data, it would be straightforward to build a growth model for drinking that includes the expectancy measures POSCON and NEGCON as time-varying covariates. Current software for multilevel models cannot accommodate missing values on covariates, however, so I first use PAN to jointly impute the missing values for drinking, POSCON, and NEGCON.

Notice in Table 12.2 that both the average level of drinking and its variation increase dramatically over time. This is somewhat problematic, because standard growth models—and the multivariate model used by PAN—assume constant variance in a response over time. To make the assumption of constant

TABLE 12.1

Missingness Rates (%) for Three Variables by Grade

	GRADE					
VARIABLE	**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

Note. POSCON = positive consequences; NEGCON = negative consequences.

TABLE 12.2

Means and Standard Deviations of Observed Variables by Grade

	GRADE											
	5		6		7		8		9		10	
VARIABLE	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Drinking	−1.43	1.33	−1.12	1.96	−0.57	2.73	0.09	3.47	1.29	4.40	1.97	4.7
POSCON	1.30	0.61	1.34	0.62	1.48	0.74			1.84	0.89	1.96	0.9
NEGCON	2.94	0.76	3.05	0.75	3.07	0.77						

Note. POSCON = positive consequences; NEGCON = negative consequences.

variance more plausible, I transformed drinking by taking its logarithm (after adding a small constant to ensure that all values were positive). After this transformation, the increase in variation became much less noticeable. The log-transformed version of drinking was used both in the imputation procedure and in subsequent analysis described below, because the transformed version more closely fit the assumptions of both the imputation procedure and the analysis. With multiple imputation, however, it is not necessary for variables to be imputed and analyzed on the same scale. Applying transformations at the imputation phase can be a highly effective tool for preserving important distributional features of nonnormal variables, regardless of how the variables are later analyzed (Schafer & Olsen, 1998).

To set up the data for PAN, one first arranges the responses for each individual in the form of a matrix y_i of dimension 6×3, with the rows corresponding to occasions (Grades 5, . . . , 10) and columns for drinking, POSCON, and NEGCON. In devising the imputation model the primary concern is to preserve growth in the variable drinking and its potential relationships to the expectancy measures. With only six time points, the model for growth must be rather simple, so let us posit a linear model with intercepts and slopes randomly varying across individuals. That is, we create a model in which drinking, POSCON, and NEGCON are each described by a linear trend with a random intercept and a random slope, for a total of six random effects in each b_i. Random intercepts and slopes are specified by placing $(1, 1, 1, 1, 1, 1)^T$ and $(1, 2, 3, 4, 5, 6)^T$ into the columns of X_i and Z_i. Finally, to incorporate potential gender differences, I allow the population average slopes and intercepts for boys and girls to vary by adding two additional columns to each X_i matrix: $\text{sex}_i \times (1, 1, 1, 1, 1, 1)^T$ and $\text{sex}_i \times (1, 2, 3, 4, 5, 6)^T$, where sex_i is a dummy indicator for participant i's gender (0 for girl, 1 for boy).

In defining a PAN model, there is no particular importance attached to the specific coding scheme used to create the design matrices X_i and Z_i. For example, the linear effect of time could have been expressed as $(-5, -3, -1, 1, 3, 5)^T$ or any other set of equally spaced scores, and the gender effect sex_i could have been coded as any two values (e.g., -1 and $+1$) rather than as 0 and 1. The particulars of the coding scheme affect the precise meaning of the parameters in $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\Psi}$, but these parameters are not of inherent interest—the goal at this stage is not to interpret parameters but to impute the missing values in y_i. Changing the coding scheme in X_i and Z_i does not change the distribution of imputed values, provided that the linear space spanned by the columns of these design matrices does not change.

Table 12.1 indicates that NEGCON is entirely missing for the last 3 years of the study. It may seem unusual to impute a variable that is entirely missing. Under this model the likely values of NEGCON for Grades 8–10 are being inferred from two sources: extrapolation from Grades 5–7 on the basis of the

assumption of linear growth, and the residual covariances among the three response variables in $\boldsymbol{\Sigma}$, which are assumed to be constant across time. Neither of these assumptions can be effectively tested with the data at hand, so inferences pertaining to NEGCON are heavily model based. In retrospect, it would have been very helpful to collect NEGCON in the final year (Grade 10) to provide more stable estimates of this variable's growth.

Before running the Gibbs sampler, I first obtained initial estimates of the unknown parameters $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\Psi}$ by running the EM algorithm. This EM procedure, which assumed an unstructured form for $\boldsymbol{\Psi}$, converged in 134 iterations and took less than 1 h on a 400 MHz Pentium II computer. The resulting maximum-likelihood estimates for $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$ were then used to formulate weak prior distributions as described in the Computational Algorithms section.

Because of the high rates of missing information, I anticipated that the Gibbs sampler would converge slowly. To assess convergence, I ran it for an initial 2,000 cycles and examined time series plots and sample autocorrelations for a variety of parameters. As anticipated, the elements of $\boldsymbol{\Psi}$ pertaining to the slopes and intercepts of NEGCON were among the slowest to converge because of the extreme sensitivity of these parameters to missing data. On the basis of this exploratory run, it appeared that several hundred cycles might be sufficient to achieve approximate stationarity. The Gibbs sampler was then run for an additional 9,000 cycles, with the simulated value of Y_{mis} stored at cycles 2,000, 3,000, . . . , 11,000. Autocorrelations estimated from cycles 1,001–11,000 verified that the dependence in all components of $\boldsymbol{\theta}$ had indeed died down by lag 200, so the 10 stored imputations could be reasonably regarded as independent draws from $P(Y_{mis}|Y_{obs})$. The entire imputation procedure took less than 2 hr with a 400 MHz Pentium II.

After imputation, the data were analyzed by a conventional linear growth-curve model for the logarithmically transformed drinking. The model was similar to the one used for imputation, except that POSCON and NEGCON now appear as time-varying covariates rather than responses. The model included an intercept and fixed effects for gender, grade, gender × grade, POSCON, and NEGCON, plus random intercepts and slopes for grade. Time was coded as $(1, 2, 3, 4, 5, 6)^T$, and gender was expressed as a dummy indicator (0 for girls, 1 for boys). Parameter estimates were computed for each imputed data set using a procedure equivalent to that used by standard packages such as HLM.

Finally, the 10 sets of fixed-effects estimates and their standard errors were then combined using Rubin's (1987) rules for multiple-imputation inference for scalar estimands. These rules are summarized as follows. Let Q denote the quantity to be estimated, in this case a regression coefficient. Let $\hat{Q}^{(j)}$ denote the estimate of Q from the jth imputed data set, and U_j its squared standard error ($j = 1, 2, \ldots, m$). The overall estimate of Q is simply the average

$$\bar{Q} = m^{-1}\Sigma\hat{Q}^{(l)}. \tag{12.8}$$

To obtain a standard error for $\bar{Q}$, one calculates the between-imputation variance $B = (m - 1)^{-1}\Sigma(\hat{Q}^{(l)} - \bar{Q})^2$ and the within-imputation variance $\bar{U} = m^{-1}\Sigma U^{(l)}$. The estimated total variance is

$$T = (1 + m^{-1})B + \bar{U}, \tag{12.9}$$

and tests and confidence intervals are based on a Student's t approximation

$$(\bar{Q} - Q)/\sqrt{T} \sim t_\nu, \tag{12.10}$$

with degrees of freedom

$$\nu = (m - 1)\left[1 + \frac{\bar{U}}{(1 + m^{-1})B}\right]^2.$$

The ratio $r = (1 + m^{-1})B/\bar{U}$ measures the relative increase in variance due to missing data, and the rate of missing information in the system is approximately $\lambda = r/(1 + r)$. A more refined estimate of this rate is

$$\lambda = \frac{r + 2/(\nu + 3)}{1 + r}. \tag{12.11}$$

The results of this procedure are summarized in Table 12.3, which shows the overall estimates, standard errors, degrees of freedom for the t approximation, and estimated percentage rates of missing information. All coefficients are highly statistically significant. The high rates of missing information indicate that the inferences for all coefficients (except sex) may be highly dependent on the form of the imputation model and the MAR assumption. The latter assumption is not particularly troubling for these data because the majority of

TABLE 12.3

Estimated Coefficients (Est.), Standard Errors, Degrees of Freedom, and Percentage Missing Information From Multiply Imputed Growth-Curve Analysis

VARIABLE	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

Note: POSCON = positive consequences; NEGCON = negative consequences.

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 caveats, the estimates in Table 12.3 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 of 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 level of use, but the effect is much smaller than that of POSCON. These results are consistent with those of previous studies (e.g., MacKinnon et al., 1991) that demonstrate that perceived positive consequences may be influential determinants of substance use behavior, but beliefs about negative consequences have little or no discernible effect.

Discussion

The multivariate mixed model (Equation 12.3) used by PAN is a natural extension of univariate growth models, which are popular in the analysis of longitudinal data. The imputation procedures described here are appropriate for longitudinal analyses with partially missing covariates. These methods are also appropriate for multivariate cross-sectional studies in which units are nested within naturally occurring groups (e.g., children within schools). The algorithm and software described in this chapter provide a principled solution to missing-data problems for this important and frequently occurring class of analyses.

The imputation model and Gibbs sampler can be extended in a number of important ways. One extension pertains to models with additional random effects due to higher levels of clustering; this would arise, for example, in multivariate studies in which individuals are grouped into larger units and multiple observations on individuals are taken over time. Another useful extension pertains to columns of y_i that are necessarily constant across the rows $1, \ldots, n_i$. In longitudinal studies, these columns would represent covariates that do not vary over time; in clustered applications, they would represent characteristics of the clusters rather than the units nested with them. If these covariates have no missing values, they can be handled under the current model by simply moving them to the matrix X. When missing values are present, however, they

must be explicitly modeled for purposes of imputation. If one imposes a simple parametric distribution on these covariates (e.g., multivariate normal), then it is straightforward to extend the Gibbs sampling procedure to impute these as well.

Another useful extension involves interactions among the columns of y_i. The multivariate normal model allows only simple linear associations among the variables $Y_1, \ldots, Y_r$, but in many studies one would like to preserve and detect certain nonlinear associations and interactions. In the AAPT example, it may have been useful to see whether the strong effect of POSCON on 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 $\times$ participant, are an important topic for future research.

In the current PAN model, the rows of y_i are assumed to be conditionally independent given b_i with common covariance matrix Σ. This assumption has been relaxed by Jennrich and Schluchter (1986), Lindstrom and Bates (1988), and others in the univariate case to allow a residual covariance matrix of the form $\sigma^2 V_i$, where V_i has a simple (e.g., autoregressive or banded) pattern dependent on one or more unknown parameters. Extensions of these patterned covariance structures to a multivariate setting tend to produce models and algorithms that are complex even apart from missing data. For example, the obvious extension of $\text{vec}(\varepsilon_i) \sim N[0, (\Sigma \otimes I)]$ to $\text{vec}(\varepsilon_i) \sim N[0, (\Sigma \otimes V_i)]$ seems too restrictive for many longitudinal data sets, because the response variables $Y_1, \ldots, Y_r$ are then required to have identical autocorrelations. Accounting for autocorrelated residuals in a sensible manner may prove to be a daunting task in the multivariate case. In practice, 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.

References

Allison, P. D. (1987). Estimation of linear models with incomplete data. *Sociological Methodology, 17*, 71–103.

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis* (2nd ed.). New York: Wiley.

Arbuckle, J. L. (1995). *Amos users' guide*. Chicago: Small Waters.

Bailey, L., Chapman, D., & Kasprzyk, D. (1985). Nonresponse adjustment procedures

at the Census Bureau: A review. In *Proceedings of the annual research conference* (pp. 421–444). Washington, DC: U.S. Bureau of the Census.

Bryk, A., & Raudenbush, S. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

Bryk, A. S., Raudenbush, S. W., & Congdon, R. T. (1996). *Hierarchical linear and nonlinear modeling with the HLM/2L and HLM/3L programs*. Chicago: Scientific Software International.

Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *American Statistician, 46*, 167–174.

Duncan, S. C., & Duncan, T. E. (1994). Modeling incomplete longitudinal substance use data using latent variable growth curve methodology. *Multivariate Behavioral Research, 29*, 313–338.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association, 85*, 398–409.

Gelman, A., Rubin, D. B., Carlin, J., & Stern, H. (1995). *Bayesian data analysis*. London: Chapman & Hall.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1996). *Markov chain Monte Carlo in practice*. London: Chapman & Hall.

Graham, J. W., Hofer, S. M., & Piccinin, A. M. (1994). Analysis with missing data in drug prevention research. In L. Collins & L. Seitz (Eds.), *National Institute on Drug Abuse research monograph series* (Vol. 142, pp. 13–62). Washington, DC: National Institute on Drug Abuse.

Hansen, W. B., & Graham, J. W. (1991). Preventing alcohol, marijuana, and cigarette use among adolescents: Peer pressure resistance training versus establishing consecutive norms. *Preventive Medicine, 20*, 414–430.

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

Jennrich, R. I., & Schluchter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics, 38*, 967–974.

Johnson, R. A., & Wichern, D. W. (1992). *Applied multivariate statistical analysis* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

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

Lindstrom, M. J., & 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., & Wolfinger, R. D. (1996). *SAS system for mixed models*. Cary, NC: SAS Institute.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.

MacKinnon, D. P., Johnson, C. A., Pentz, M. A., Dwyer, J. H., Hansen, W. B., Flay, B. R., & 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.

Madow, W. G., Nisselson, H., & Olkin, I. (1983). *Incomplete data in sample surveys, Vol. 1: Report and case studies*. New York: Academic Press.

MathSoft, Inc. (1997). *S-PLUS user's guide*. Seattle, WA: Author.

McArdle, J. (1988). Dynamic but structural modeling of repeated measures data. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (pp. 561–614). New York: Plenum Press.

Meng, X. L. (1994). Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statistical Science, 10*, 538–573.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika, 55*, 107–122.

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika, 52*, 431–462.

Neale, M. C. (1994). *Mx: Statistical modeling* (2nd ed.). Richmond: Medical College of Virginia, Department of Psychiatry.

Novick, M. R., & Jackson, P. H. (1974). *Statistical methods for educational and psychological research*. New York: McGraw-Hill.

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

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Rubin, D. B. (1996). Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association, 91*, 473–489.

Schafer, J. L. (1997a). *Analysis of incomplete multivariate data*. London: Chapman & Hall.

Schafer, J. L. (1997b). *Imputation of missing covariates under a multivariate linear mixed model* (Tech. Rep. 97-10). University Park: Pennsylvania State University, The Methodology Center.

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research, 33*, 545–571.

Willett, J. B., & Sayer, A. G. (1994). Using covariance structure analysis to detect correlates and predictors of change. *Psychological Bulletin, 116*, 363–381.

Yucel, R., & Schafer, J. L. (1998). Fitting multivariate linear mixed models with incomplete data. In *Proceedings of the Statistical Computing Section of the American Statistical Association* (pp. 177–182). Alexandria, VA: American Statistical Association.

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)	—	—	—

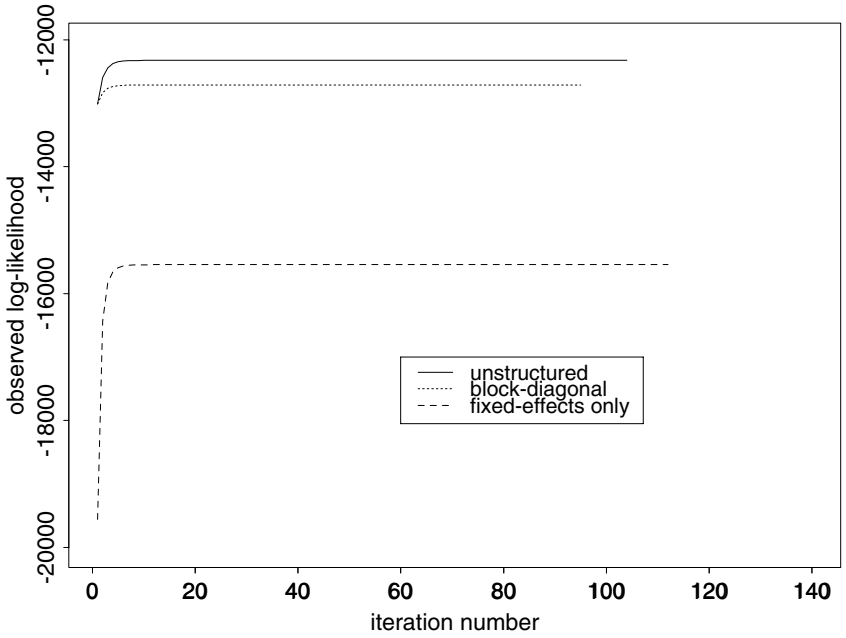

Figure 1. Convergence behaviors under different covariance structures.

($\Psi = 0$) 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.