# 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, $\beta$ $(p \times r)$ 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 $\varepsilon_i$ 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 $\Sigma$ 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 $\Sigma$. In particular, they assumed a diagonal form for $\Sigma$ which is often unrealistic and undesirable. Correlations among the columns of $\epsilon_i$ 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 $\beta$ can be removed from this likelihood function in one of two ways: profiling, in which $\beta$ is replaced by its conditional maximum given $(\Sigma, \Psi)$; and marginalizing, in which the likelihood is replaced by its indefinite integral with respect to $\beta$. Both the profile and marginal likelihoods can be written in closed form as functions of the generalized least-squares estimate for $\beta$ given $(\Sigma, \Psi)$. 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 $\Psi$ has a special patterned structure, $\Psi = \Sigma \otimes \Upsilon$ for some $q \times q$ matrix $\Upsilon$. 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 $\Psi$ and $(\beta, \Sigma)$,

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

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 $\theta$. In the multivariate model, the pairs $(y_i, b_i)$ are distributed according to

$$
\begin{aligned}
{y_i}^V \mid b_i, \theta &\sim N(\,(X_i\beta + Z_i b_i)^V,\, (\Sigma \otimes I_{n_i})\,), && (5)\\
{b_i}^V \mid \theta &\sim N(0, \Psi), && (6)
\end{aligned}
$$

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

$$
\begin{aligned}
\tilde{b}_i{}^V &= \Gamma_i\,(\Sigma^{-1} \otimes Z_i^T)\,(y_i - X_i\beta)^V, && (7)\\
\Gamma_i &= (\,\Psi^{-1} + (\Sigma^{-1} \otimes Z_i^T Z_i)\,)^{-1}. && (8)
\end{aligned}
$$

Calculating $\Gamma_i$ 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

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

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 $\theta$ is not of high priority. Rather than attempting direct likelihood-based inferences about $\theta$, 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 $\theta$. 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 $\Sigma$ and $\Psi$ with confidence based on $\nu_1$ and $\nu_2$ degrees of freedom, respectively. Small values for $\nu_1$ and $\nu_2$ make the prior densities relatively diffuse, reducing their impact on the final inferences. For $\beta$, 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 $\theta$ in (12) proceeds as follows: First, draw $\Psi^{-1}$ 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 $\Sigma^{-1}$ 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 $\beta$ 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 $\beta$, it is helpful to note that if $G$ and $H$ are upper-triangular square roots of $\Sigma$ 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 $\Sigma$. Therefore, in any row of $\varepsilon_i$, 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 $\Sigma$ corresponding to the observed variables. Drawing the missing elements in $\varepsilon_i$ 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 $\theta$ 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 $\theta$, 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 $\psi^{-1}$), 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 $\theta$ 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 $\Sigma$ 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 $\theta$ were then examined. As anticipated, the elements of $\Psi$ 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. 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 $\theta$ 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 $\sigma^2$ denotes the univariate version of $\Sigma$). 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 $\theta$; (b) a constrained maximization of the expected loglikelihood for $\theta_2$ given the previous estimate of $\theta_1$, 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 $\theta_1$ given the updated estimate of $\theta_2$. 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 $\Psi$ for the random effects may be limiting in situations where some aspects of $\Psi$ 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 $\Psi$ 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 $\Sigma$. 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 $\epsilon_i$ 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.