Journal of Educational and Behavioral Statistics
Spring 1997, Vol. 22, No. 1, pp. 77–108

Hierarchical Linear Models for Multivariate Outcomes

Yeow Meng Thum
Consortium on Chicago School Research
University of Chicago

Keywords: *empirical Bayes, hierarchical linear models, individual differences, maximum likelihood, missing outcomes, multilevel data, multivariate repeated measures, multivariate two-stage models, random effects, sensitivity,* t *prior, quasi-Newton*

*In this article, we develop a class of two-stage models to accommodate three common characteristics of behavioral data. First, behavior is invariably multivariate in its conceptualization and communication. Separate univariate analyses of related outcome variables are fraught with potential interpretive blind spots for the researcher. This practice also suffers, from an inferential standpoint, because it fails to take advantage of any redundant information in the outcomes. Second, studies of behavior, especially in experimental research, employ smaller samples. This situation raises issues of robustness of inference with respect to outlying individuals. Third, the outcome variable may have observations missing because of accidents or by design. The model permits the estimation of the full spectrum of plausible measurement error structures while using all the available information. Maximum likelihood estimates are obtained for various members of a multivariate hierarchical linear model (MHLM), and, in the context of several illustrative examples, these estimates match closely the results from a Bayesian approach to the normal-normal MHLM and to the normal-*t *MHLM.*

It is fair to say that in recent years univariate multilevel models, estimated under large sample assumptions, have attained near-paradigmatic proportions in quantitative educational research (Bock, 1989; de Leeuw & Kreft, 1986; Goldstein, 1987; Longford, 1987; Raudenbush, 1988; Raudenbush & Bryk, 1986; Rogosa & Willett, 1985). But because behavior is most often multivariate, both in its conceptualization and in the way it is reported, multilevel

---

Major portions of this article originally appeared as a contribution to the "New Developments in Multilevel Modeling" session at the 1994 Annual Meeting of the American Educational Research Association, New Orleans, April 4–8. Research reported in this article is based on dissertation work done while the author was with the Department of Psychology at the University of Chicago. In addition, the associate editor, two anonymous referees, R. Darrell Bock, Anthony S. Bryk, Ken Frank, David W. Kerbow, Stephen W. Raudenbush, Michael H. Seltzer, and Benjamin D. Wright have all contributed much appreciated insights and comments to earlier versions of this article. The author is solely responsible for all remaining errors.

analyses employing these models do not take proper advantage of the richly textured information borne by multiple measures of behavior. For example, a teacher's attitudes towards school reform proposals and his efforts in this connection are obviously related. Analyzing the attitudinal set and the behavioral reactions toward reform separately may conceal informative patterns in the data from the researcher's view. Even when a univariate approach is adopted with confidence, it is frequently inevitable that the researcher and his audience want to integrate into a more coherent picture the conclusions from his parallel investigations. But without a model for exploring the interrelationships jointly, information which is helpful from an inferential standpoint is discarded because the correlations among the outcomes or the random parameters of related univariate models are assumed to be zero. When any of these components are correlated, whether within a univariate model (e.g., measurement errors in a repeated measures design) or between the random effects of two univariate models, the informational redundancy will generally alter the magnitudes of the variance components. Corresponding changes in the standard error estimates that result with a multivariate approach may, in turn, modify the conclusions drawn from the separate univariate analyses.

This article presents a *multivariate hierarchical linear model* (MHLM) for multiple continuous outcomes. Our two-stage model is developed as a methodological complement to a natural-science, population concept of behavioral processes (Thum, 1994). The Stage 1 model in the MHLM is a multivariate analysis of covariance model for exploring the controlling factors of individual performance. At Stage 2, individual differences in performance are described by a multivariate normal or a multivariate t prior. Thus, in addition to dealing with multiple outcomes, the MHLM with a multivariate t prior accommodates another common characteristic of behavioral data, which is that the number of subjects may be small.

The MHLM adds to several recent multivariate formulations of the normal-normal multilevel model that have appeared in Goldstein and McDonald (1988), Muthén (1989), and Schmidt (1969). More recently, McDonald and Goldstein (1989) and Longford and Muthén (1992) restate the general LISREL model (Jöreskog & Sörbom, 1984) and Schmidt's (1969) early results for multivariate random effects models for the unbalanced case. Another approach to multivariate multilevel data can be formulated in terms of a univariate three-stage model, since a three-stage model with no residual variance at any one level is formally equivalent to a two-stage model, typically the lowest (see Goldstein, 1987, or Longford, 1993). As is, the normal-normal formulation of the MHLM is a straightforward simplification of any three-stage hierarchical linear model, such as that of Raudenbush and Bryk (1986). Note, however, that while a reduced three-stage model is two-stage multivariate by virtue of restricting one level of variation, the MHLM is multivariate in the usual sense that its Stage 1 model is simply the familiar general linear model with a less restrictive error structure.

In a preliminary section we develop the MHLM, placing particular stress on the components of the model that are meaningful to behavioral research. For lack of space, we give only a brief outline of model estimation. We estimate both the normal-normal model and the normal-multivariate t model by maximum likelihood, using a reliable variable-metric (or quasi-Newton) algorithm which supports linear equality constraints on point estimators. Estimation of the normal-multivariate t model involves numerical integration by a one-dimensional Gauss-Laguerre quadrature. In all cases, standard errors are computed from analytic expressions of the Hessian evaluated at the minima (again, by an approximation for models with a multivariate t prior). Full details of model estimation can be found in Thum (1994). We next discuss briefly three worked applications of the model, placing particular stress on estimation results. First, data from a survey of teacher attitudes towards school reform proposals and their subsequent efforts in that regard serve to illustrate the overall approach. In a second example, the results of a normal-t variant of the MHLM for studying the vocabulary growth of infants, in which the number of subjects is relatively small, is compared with a Bayesian solution via the Gibbs sampler (Seltzer, 1993). Third, results for alternative models with independently and identically distributed (iid) normal errors or with serially correlated errors for the growth data presented in Pothoff and Roy (1964) are used to illustrate model estimation when some outcomes are missing at random.

A Multivariate Hierarchical Linear Model

Modeling Individual Performance

With repeated measurements, a data set in the behavioral sciences typically displays a two-level nested structure in that the jth of N independent data sets contains n_j observations on the same set of variables for each subject, j, in the sample. The data from each individual subject may consist of either univariate or multivariate repeated measurements of outcomes and conditions. We therefore consider the fairly general situation in which a set of k related outcome measures, $\mathbf{y}'_{ij} = [y_{ij}^{(1)}\ y_{ij}^{(2)} \cdots y_{ij}^{(k)}]$, is observed for each of n_j replications for individual j ($i = 1, 2, \ldots, n_j$ and $j = 1, 2, \ldots, N$).

Linear Measurement Effects

Models for measurement effects require no special introduction. In psychology, they range from models of growth in short time-series experiments to models of method effects in test validation studies. From the jth subject, we obtain n_j observations on k outcome measures, which results in an $n_j \times k$ multivariate response matrix, $\mathbf{Y}_j$. A linear model for measurement effects employs the $k \times m$ model matrix $\mathbf{M}_j$,

$$\mathscr{E}[\mathbf{Y}_j|\boldsymbol{\mu}_j] = \mathbf{1}_j\boldsymbol{\mu}'_j\mathbf{M}'_j, \tag{1}$$

where $\mathbf{1}_j$ is the $n_j \times 1$ unit vector, $\boldsymbol{\mu}_j$ is the $m \times 1$ vector of unknown measurement effects, and $m \leq k$. In most situations, the measurement effects model is fixed over j (see example in next section). In some important cases to be considered elsewhere, a more general form of (1) is required to accommodate instances where the number of outcomes varies from one observation to another, so that $[\mathbf{M}'_{1j}\, \mathbf{M}'_{2j} \cdots \mathbf{M}'_{n_j j}]'$ replaces $\mathbf{M}_j$. We will explore only a second important case in which $\mathbf{M}_j$ is a row-subset of $\mathbf{M}$. This is useful when the patterns of available responses differ from one individual to another, sometimes by design but oftentimes because the observations are missing. To simplify our presentation, we will assume a common measurement effects model, $\mathbf{M}$, until situations requiring its change with i or with j are considered.

Experimental Treatment and Covariate Adjustment

Suppose that, as in preference research, a subject j is asked to rate n_j multiattributed objects of dimension p on k criterion variables. Then, aside from the response matrix $\mathbf{Y}_j$, we also obtain an $n_j \times p$ matrix of known object attributes, $\mathbf{A}_j$. Alternatively, subject j could be exposed to n_j treatment conditions of dimension p. In this case, $\mathbf{A}_j$ represents the within-subject treatment design for subject j. In the MHLM, adjusting the basic measurement effects model (1) for treatment effects is a simple application of the usual adjustment strategy employed in multivariate analysis of covariance. Recall the generic multivariate analysis of covariance model

$$\mathscr{E}[\mathbf{Y}_j|\boldsymbol{\mu}_j, \boldsymbol{\Xi}_j] = \mathbf{A}_j\boldsymbol{\Xi}_j + \mathbf{1}_j\boldsymbol{\mu}_j'\mathbf{M}_j', \tag{2}$$

where $\boldsymbol{\Xi}_j$ is a $p \times k$ matrix of unknown random regression parameters. Results of this general formulation are well understood (see Pothoff & Roy, 1964; Bock, 1975).

From this point on, the usual formulation of the normal-normal linear model follows. Our notation needs only minor adjustments in the face of the added complexities (see Table 1 for notation). For the n_j k-variate observation vectors for individual j, we propose the following mixed-effects linear *response model* for each individual:

$$\mathscr{E}[\text{vec}(\mathbf{Y}'_j)|\boldsymbol{\beta}_j] = E[\mathbf{y}_j|\boldsymbol{\beta}_j] = \mathbf{A}_j^*\boldsymbol{\alpha} + \mathbf{B}_j^*\boldsymbol{\beta}_j. \tag{3}$$

As in Equation 1, $\mathbf{Y}_j$ is the $n_j \times k$ matrix of responses. Thus, $\mathbf{y}_j$ is $n_j k \times 1$. Using

$$\mathbf{A}_j^*\boldsymbol{\alpha} = \text{vec}([\mathbf{A}_j\boldsymbol{\Xi} + \mathbf{1}_j\boldsymbol{\mu}'\mathbf{M}_j']'),$$

we have represented, in effect, the coefficients $\boldsymbol{\alpha}$ as $[\text{vec}(\boldsymbol{\Xi}')\ \boldsymbol{\mu}']'$. In a simplified model with a common measurement effects model, the predictor

TABLE 1
Multivariate multilevel data

		Covariates		
		Stage 1		
Group	Outcome	Fixed	Random	Stage 2
n_1	$\mathbf{y}_1$	$\mathbf{A}_1^*$	$\mathbf{B}_1^*$	$\mathbf{x}_1'$
n_2	$\mathbf{y}_2$	$\mathbf{A}_2^*$	$\mathbf{B}_2^*$	$\mathbf{x}_2'$
n_3	$\mathbf{y}_3$	$\mathbf{A}_3^*$	$\mathbf{B}_3^*$	$\mathbf{x}_3'$
⋮	⋮	⋮	⋮	⋮
n_j	$\mathbf{y}_j$	$\mathbf{A}_j^*$	$\mathbf{B}_j^*$	$\mathbf{x}_j'$
⋮	⋮	⋮	⋮	⋮
n_N	$\mathbf{y}_N$	$\mathbf{A}_N^*$	$\mathbf{B}_N^*$	$\mathbf{x}_N'$

matrix with fixed coefficients $\mathbf{A}_j^*$ comprises $[\mathbf{A}_j \otimes \mathbf{I}_k \; \mathbf{1}_j \otimes \mathbf{M}_A]$, and the predictor matrix with random coefficients $\mathbf{B}_j^*$ consists of $[\mathbf{B}_j \otimes \mathbf{I}_k \; \mathbf{1}_j \otimes \mathbf{M}_B]$.[1] $[\mathbf{A}_j \otimes \mathbf{I}_k]$ is the $n_j k \times p_a$ matrix of individual level covariates with fixed coefficients, and $[\mathbf{B}_j \otimes \mathbf{I}_k]$ is the $n_j k \times p_b$ matrix of covariates with random coefficients. The matrices $\mathbf{M}_A$ and $\mathbf{M}_B$, of column orders m_a and m_b, respectively, generally refer to partitions of the $k \times (m_a + m_b)$ measurement effects model matrix $\mathbf{M}$ according to whether coefficients of component columns are considered, respectively, fixed or random. Any combinations of these component predictor matrices may be specified to fit the desired model. In the usual mixed-model interpretation of (3), $\boldsymbol{\alpha}$ is an unknown $(p_a k + m_a)$ vector of fixed regression effects, and $\boldsymbol{\beta}_j$ is an unknown $(p_b k + m_b)$ vector of random regression coefficients. From the researcher's point of view, one useful interpretation of the distinction between $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}_j$ is that $\boldsymbol{\alpha}$ represents comparable interindividual performance on the first set of individual characteristics, whereas $\boldsymbol{\beta}_j$ represents varying interindividual degrees of regression on others. If covariates with random coefficients are absent, (3) reduces to the normal unilevel fixed-effects regression model.

Correlated Errors of Measurement

To complete the Stage 1 model of individual performance, we assume that, conditional on $\boldsymbol{\beta}_j$, $\mathbf{y}_j$ is multinormally distributed as

$$\pi(\mathbf{y}_j|\boldsymbol{\beta}_j; \boldsymbol{\alpha}, \boldsymbol{\Theta}) = N(\mathbf{A}_j^*\boldsymbol{\alpha} + \mathbf{B}_j^*\boldsymbol{\beta}_j, \boldsymbol{\Theta}_j), \tag{4}$$

where $\boldsymbol{\Theta}_j = [\mathbf{I}_j \otimes \boldsymbol{\Theta}]$. $\boldsymbol{\Theta}$ denotes the $k \times k$ variance-covariance matrix of measurement error. Although the Stage 1 measurement errors are independently and identically distributed, the form of $\boldsymbol{\Theta}$ remains deliberately general because the errors of measurement associated with multivariate outcomes are neither likely to be equal nor independent. Important forms of $\boldsymbol{\Theta}$ include:

(i) $\theta^2\mathbf{I}_k$
(ii) $\text{diag}(\theta_1^2, \theta_2^2, \ldots, \theta_k^2)$
(iii) $\boldsymbol{\Theta}$
(iv) $\theta^2\{\rho^{|l-1|}\}, \quad l = 1, 2, \ldots, k$
(v) arbitary equality constraints

The $\boldsymbol{\Theta}$ error model (i) states that all outcomes measure one underlying construct and that they furthermore err to a similar degree, implying that the outcomes are, for all intents and purposes, iid. As a result, $\boldsymbol{\Theta}$ model (i) is not distinguishable from the usual error model for univariate outcomes. Model (ii) relaxes this assumption by positing that the errors of measurement may well be heteroscedastic. A general symmetric $\boldsymbol{\Theta}$ in model (iii) suggests further that the measurement errors may be correlated as well. Model (iv) provides for a first-order correlation, ρ, among errors of adjacent measurements in short time-series data. For model fitting purposes, error models (v) serve further theoretical and practical needs by allowing individual elements of $\boldsymbol{\Theta}$ to be arbitarily constrained. Later we show that the procedures for estimating patterned matrices and arbitary linear constraints may be effected easily with gradient algorithms, such as the quasi-Newton in particular.

Incomplete Multivariate Responses

A common problem with multivariate response data is the frequent occurrence of incomplete observations in the response vector. Responses may be incomplete in some experiments because the unrecorded observations have been missing more or less by accident. In other multivariate experiments, responses are incomplete because a different subset of the variables under study is recorded in each subgroup of experimental units. Following the work of Trawinski and Bargmann (1964) for the multivariate general linear model, we let

$$\mathbf{y}_j = (\mathbf{I}_j \otimes \mathbf{S}_j^{*\prime})\mathbf{z}_j. \tag{5}$$

$\mathbf{z}_j$ is the $(n_j k \times 1)$ complete data vector, and $\mathbf{S}_j^*$ is a $k \times k_j$ incidence matrix which selects the columns of the complete data matrix to be analyzed $\mathbf{y}_j$. For example,

$$\mathbf{S}_{24}^* = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}$$

if only two (columns 1 and 3) out of three possible columns of the 24th subject are available for analysis. Obviously, $\mathbf{S}_j^* = \mathbf{I}_k$ if data are complete.

$\mathbf{y}_j$ is now an $(n_j k_j \times 1)$ vector, and, consequently, the rank of $\boldsymbol{\Theta}$ is dependent on how large k_j gets over j. Model (5) is particularly useful in mixed longitudinal designs or when outcomes in a repeated measures design are missing in a $\boldsymbol{\Theta}$ model with first-order correlations (see Example 3). In the extreme, a model for univariate outcomes in which the error variances change with j may also be defined if each $\mathbf{S}_j$ is unique to each individual (see also Schlucter, 1988).

Modeling Individual Differences

The Stage 1 linear model in Equation 4 represents the summary we are giving individual j based on a set of information, $[\mathbf{A}_j^* \ \mathbf{B}_j^*]$, that we think is relevant to his observed behavior, $\mathbf{y}_j$. The essential character of each individual's performance is contained in the $(p_a k + m_a + p_b k + m_b)$-fold regression vector, $[\boldsymbol{\alpha}' \ \boldsymbol{\beta}_j']$. This summary will range in effectiveness from one individual to another because, as we have argued previously, behavior will often vary systematically from one individual to another, but it is hoped that overall our theory of how people behave is valid enough so that the model describes most individuals sufficiently well. Then, the task of a theory of individual differences is to express, in reproducible terms, the degree of variability of observed characteristics and to explain the performance of individuals in the target population in terms of theoretically interesting characteristics of individual subjects and the range of conditions under which subjects function.

Unlike conventional analyses which fail to distinguish behavioral fluctuations within an individual from variation among individuals in the population, we pose at Stage 2 a multivariate linear model for the varying indexes of individual functioning, $\boldsymbol{\beta}_j$, from Equation 4. The corresponding *individual difference model* is

$$\pi(\boldsymbol{\beta}_j; \boldsymbol{\gamma}, \boldsymbol{\Psi}) = N(\mathbf{X}_j\boldsymbol{\gamma}, \boldsymbol{\Psi}) \tag{6}$$

for the large-sample normal case at Stage 2 or, when N is small, as is typical in many experimental studies, a robust multivariate t prior with λ degrees of freedom:

$$\pi(\boldsymbol{\beta}_j; \boldsymbol{\gamma}, \boldsymbol{\Psi}) = t(\lambda, \mathbf{X}_j\boldsymbol{\gamma}_T, \boldsymbol{\Psi}_T). \tag{7}$$

Here, $\mathbf{X}_j$ is a block-diagonal matrix consisting of $(p_b k + m_b)$ blocks of the row vector of known unit-level predictors $\mathbf{x}_j$, or a subvector of it. $\boldsymbol{\gamma}$ is a column vector of fixed effects of the appropriate length, and $\boldsymbol{\Psi}$ is the $((p_b k + m_b) \times (p_b k + m_b))$ residual parameter variance-covariance matrix. If $\mathbf{X}_j$ is the identity matrix, we have specified a *random regression model* or a *population model* and $\boldsymbol{\gamma}$ is a stochastic average of subject performance in the population. This Stage 2 model ((6) or (7)) is a formal device for aggregating over regressions (Bryk & Raudenbush, 1985; Rubin, 1978). We see that

$\boldsymbol{\Psi}$ indicates the extent of subject variation in performance in the population. For any other subject-level correlates of psychological performance, $\mathbf{X}_j$, a *conditional model* for $\boldsymbol{\beta}_j$ is estimated and $\boldsymbol{\gamma}$ is a representation of modulated population behavior. Like treatment effects, effects due to measurement method are permitted to vary among subjects, as well.

Correlated Performance Components

In summary, Stage 2 of the MHLM provides a means for examining the differences in individual performance. Generally, the regression coefficients, $\boldsymbol{\beta}_j$, may be selectively fixed. When all of them are fixed, $\boldsymbol{\Psi}$ is null (model (i)) and we return to a fixed effects unilevel regression model. For some applications, the coefficients may be assumed a priori uncorrelated (model (iii)), or they may further be expected to be comparable (model (ii)). Linear equality constraints are as relevant here as they are for the elements of $\boldsymbol{\Theta}$. The most frequent patterns of residual parameter variance are:

(i) $\mathbf{0}$
(ii) $\psi^2\mathbf{I}_{(b+m_b)}$
(iii) $\text{diag}(\psi_1^2, \psi_2^2, \ldots, \psi_{(p_b\text{k}+m_b)}^2)$
(iv) $\boldsymbol{\Psi}$
(v) $\psi^2\{\omega^{|l-1|}\}, \qquad l = 1, 2, \ldots, (p_b\text{k} + m_b)$
(vi) arbitary equality constraints

Estimating the MHLM

For the class of models subsumed by the MHLM, we develop a quasi-Newton algorithm for its estimation. As we have argued in Thum (1994), this approach has several advantages over other, currently popular *complete-data methods* for maximum likelihood such as Fisher scoring (Longford, 1987), Newton-Raphson (Jennrich & Schlucter, 1986; Lindstrom & Bates, 1988), and iterative generalized least squares (Goldstein, 1987), and over *missing-data methods* such as the EM algorithm (Bryk & Raudenbush, 1992; Dempster, Laird, & Rubin, 1977), data augmentation (Seltzer, 1991), and the Gibbs sampler (Casella & George, 1992; Gelfand, Hills, Racine-Poon, & Smith, 1990; Seltzer, 1993). Only the first derivatives of the log-likelihood are required for its solution. The inverse Hessian is updated by a simple recursive formula during each cycle; thus, the algorithm dispenses with the time-consuming task of evaluating and inverting the Hessian in every iteration. The savings are considerable even in cases, such as the normal-normal hierarchical linear model, in which the Hessian does not depend on the data, $\mathbf{y}_j$. More important, perhaps, is that with reasonable starting values, convergence is stable and quadratic. Standard errors may be computed from well known expressions of the expected information matrix on convergence.

Variable Metric Methods

To motivate the computational approach taken in this article, we note that both the Newton-Raphson method and Fisher scoring are examples of a class of gradient algorithms

$$\mathbf{v}_{i+1} = \mathbf{v}_i - \lambda \mathbf{H}_i \mathbf{d}_i \tag{8}$$

for minimizing $f(\mathbf{v})$. The subscript i denotes the current iteration and $\mathbf{v}_i$ the current estimate of the parameters $\mathbf{v}$. $\mathbf{d}$ contains the vector of first derivatives, while $\mathbf{H}$ is a symmetric, positive-definite matrix. λ is an appropriately chosen scalar, to be obtained by a linear search or a step-length method. The use of λ usually produces relatively superior performance. If $\mathbf{H}_i$ is the identity matrix, (8) is the robust albeit slow *method of steepest descent*. As previously noted, (8) is the Newton-Raphson algorithm when $\mathbf{H}_i$ is the inverse Hessian. In statistical applications, Fisher scoring uses the inverse of the expected information matrix.

Consideration of other forms of $\mathbf{H}_i$ in (8) leads to so-called *quasi-Newton* or *variable metric* methods (Davidon, 1959; Fletcher & Powell, 1963). These procedures seek simple recursive formulas for approximating $\mathbf{H}_i$ in order to produce algorithms based only on the first derivatives (thus dispensing with the need to evaluate, and to invert, the matrix of second derivatives) that will match Newton's method in convergence rate and with increased reliability. In addition, $\mathbf{H}_i$ converges to the inverse Hessian as i approaches infinity. Several such formulas for recursively approximating $\mathbf{H}_i$ have been developed (see Burley, 1974; Huang, 1970). Following Fletcher (1970), Shanno and Phua (1974) suggested the Broyden-Fletcher-Shanno (BFS) matrix update formula

$$\mathbf{H}_{i+1} = (\mathbf{I} - \mathbf{p}_i\mathbf{g}_i'/\mathbf{p}_i'\mathbf{g}_i)\mathbf{H}_i(\mathbf{I} - \mathbf{p}_i\mathbf{g}_i'/\mathbf{p}_i'\mathbf{g}_i) + \mathbf{p}_i\mathbf{p}_i'/\mathbf{p}_i'\mathbf{g}_i, \tag{9}$$

where $\mathbf{p}_i = \mathbf{v}_{i+1} - \mathbf{v}_i$ and $\mathbf{g}_i = \mathbf{d}_{i+1} - \mathbf{d}_i$. The computational steps for the BFS algorithm are:

(1) In iteration i, $(f_i, \mathbf{v}_i, \mathbf{d}_i, \mathbf{H}_i)$ are known.
(2) Choose λ to minimize $F(\lambda) = f(\mathbf{v}_i + \lambda \mathbf{H}_i \mathbf{d}_i)$.
(3) Compute $\mathbf{v}_{i+1} = \mathbf{v}_i - \lambda_{\min} \mathbf{H}_i \mathbf{d}_i$.
(4) Update $\mathbf{H}_i$ to $\mathbf{H}_{i+1}$ by (9).
(5) Go to Step 1 if convergence is not attained.

See Shanno and Phua (1974) for complete details.

Linear Equality Constraints

When using gradient methods, we may also selectively fix certain coefficients by an appropriate incidence matrix $\mathbf{S}_v$ such that the parameters consid-

ered as variables in $f(\cdot)$ are $\mathbf{v}_v = \mathbf{S}_v\mathbf{v}$. To fix parameters 2 and 3 of a 5×1 parameter vector $\mathbf{v}$, for example, we have

$$\begin{bmatrix} v_1 \\ v_4 \\ v_5 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \end{bmatrix}.$$

Linear equality constraints on $\mathbf{v}_v$ may be imposed by noting that the unique parameter vector $\mathbf{u}$ is simply $\mathbf{v}_v = \mathbf{S}_u\mathbf{u}$. For example, constraining $v_4 = v_5$, we have

$$\begin{bmatrix} v_1 \\ v_4 \\ v_5 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}.$$

The relation of $\mathbf{u}$ to $\mathbf{v}$ is therefore $\mathbf{u} = \mathbf{S}_c\mathbf{v}$, where the general matrix of constraints on $\mathbf{v}$ is

$$\mathbf{S}_c = [\mathbf{S}_u'\mathbf{S}_u]^{-1}\mathbf{S}_u'\mathbf{S}_v. \tag{10}$$

In our example, $u_1 = v_1$ and $u_2 = (v_4 + v_5)/2$ as expected. Newton-Raphson or Fisher scoring solutions in terms of $\mathbf{u}$ may now be effected by transforming the matrix of first derivatives by

$$\left[\frac{\partial f}{\partial \mathbf{u}}\right] = \mathbf{S}_c\left[\frac{\partial f}{\partial \mathbf{v}}\right] \tag{11}$$

and, if the Hessian is to be evaluated, by further noting that

$$\left[\frac{\partial^2 f}{\partial \mathbf{u}\partial \mathbf{u}'}\right] = \mathbf{S}_c\left[\frac{\partial^2 f}{\partial \mathbf{v}\partial \mathbf{v}'}\right]\mathbf{S}_c'. \tag{12}$$

Multivariate Normal Prior

First, we consider the normal-normal MHLM, suitable when N is moderate to large. It is well known from the specifications of the Stage 1 multivariate normal model (4) and the Stage 2 multivariate normal prior (6) of the MHLM above that $\mathbf{y}_j$ is marginally distributed normal

$$\mathbf{y}_j \sim N(\mathbf{C}_j\boldsymbol{\eta}, \boldsymbol{\Sigma}_j), \tag{13}$$

where

$$\mathbf{C}_j = \mathbf{C}_j^*\mathbf{X}_j^* = [\mathbf{A}_j^* \; \mathbf{B}_j^*]\mathbf{X}_j^*,$$

$$\mathbf{X}_j^* = \begin{bmatrix} \mathbf{I}_{(p_a+m_a)} & 0 \\ 0 & \mathbf{X}_{j\mathrm{B}} \end{bmatrix},$$

$$\boldsymbol{\eta} = \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\gamma} \end{bmatrix},$$

$$\boldsymbol{\Sigma}_j = \mathbf{B}_j^*\boldsymbol{\Psi}\mathbf{B}_j^{*\prime} + \boldsymbol{\Theta}_j, \quad \text{and}$$

$$\boldsymbol{\Theta}_j = \mathbf{I}_j \otimes \boldsymbol{\Theta}.$$

Maximum Likelihood

Assuming that responses, $\mathbf{y} = (y_1, y_2, \ldots, \mathbf{y}_N)$, of subjects are independent, the natural logarithm of the marginal data likelihood is

$$l_{\mathrm{F}}(\boldsymbol{\eta}, \boldsymbol{\Theta}, \boldsymbol{\Psi}|\mathbf{y}) = -\frac{1}{2}\sum_{j=1}^{N}\{n_j k \ln(2\pi) + \ln|\boldsymbol{\Sigma}_j| + \mathrm{tr}(\boldsymbol{\Sigma}_j^{-1}\mathbf{T}_j)\}, \tag{14}$$

where

$$\mathbf{T}_j = (\mathbf{y}_j - \mathbf{C}_j\boldsymbol{\eta})(\mathbf{y}_j - \mathbf{C}_j\boldsymbol{\eta})'.$$

The exact computational formulas for maximizing (14), and the form of the basic statistics necessary for efficient calculations, will depend on the specific structure of $\mathbf{B}_j^*$, and the specific structure of $\mathbf{A}_j^*$ if nonvarying coefficients are present in the Stage 1 model. To complete our presentation, we will summarize the basic results for the response model (3), where

$$\mathbf{A}_j^* = [\mathbf{A}_j \otimes \mathbf{I}_k \quad \mathbf{1}_j \otimes \mathbf{M}_A]$$

and

$$\mathbf{B}_j^* = [\mathbf{B}_j \otimes \mathbf{I}_k \quad \mathbf{1}_j \otimes \mathbf{M}_B],$$

in the remainder of this treatment. It will be seen that the basic statistics for this simple variant of the MHLM are the sums of squares and of cross-products for every j.

The free parameters, $(\boldsymbol{\eta}, \boldsymbol{\Theta}, \boldsymbol{\Psi})$, in (14) have explicit solutions when

subgroup numbers are equal, that is, $n_j = n$. A number of iterative solutions have been used in the more general case where subgroup numbers vary. If the variance components ($\boldsymbol{\Theta}$, $\boldsymbol{\Psi}$) are known, the fixed effects $\boldsymbol{\eta}$ may be obtained by the generalized least squares estimator

$$\tilde{\boldsymbol{\eta}} = \left[\sum_{j=1}^{N} \mathbf{C}_j' \boldsymbol{\Sigma}_j^{-1} \mathbf{C}_j\right]^{-1} \sum_{j=1}^{N} \mathbf{C}_j' \boldsymbol{\Sigma}_j^{-1} \mathbf{y}_j; \tag{15}$$

otherwise, the OLS residuals from (4) provide good starting values for estimating $\boldsymbol{\Theta}$, and the raw residuals from (6) for a preliminary estimate of $\boldsymbol{\Psi}$ (see Lindstrom & Bates, 1988).

First derivatives under ML. Much of the following derivations involves results that are well documented elsewhere (Dwyer, 1967; MacRae, 1974; Magnus & Neudecker, 1979; McCulloch, 1982). So, proofs will be left deliberately brief. Thum (1994, chapter 4) also detailed further simplified computing formulas. Writing $\text{vec}(\mathbf{K}_j') = \mathbf{C}_j\boldsymbol{\eta}$ and $\mathbf{D}_j = \mathbf{Y}_j - \mathbf{K}_j$, we have for $\boldsymbol{\eta}$,

$$\frac{\partial l_F}{\partial \boldsymbol{\eta}} = -\sum_{j=1}^{N} \mathbf{C}_j' \boldsymbol{\Sigma}_j^{-1} \text{vec}(\mathbf{D}_j') \; ; \tag{16}$$

for $\boldsymbol{\Psi}$,

$$\frac{\partial l_F}{\partial \boldsymbol{\Psi}} = \frac{1}{2} \sum_{j=1}^{N} \mathbf{B}_j^{*\prime} \boldsymbol{\Sigma}_j^{-1} (\boldsymbol{\Sigma}_j - \mathbf{T}_j) \boldsymbol{\Sigma}_j^{-1} \mathbf{B}_j^* \; ; \tag{17}$$

and for $\boldsymbol{\Theta}$,

$$\frac{\partial l_F}{\partial \boldsymbol{\Theta}_j} = \frac{1}{2} \boldsymbol{\Sigma}_j^{-1} (\boldsymbol{\Sigma}_j - \mathbf{T}_j) \boldsymbol{\Sigma}_j^{-1}. \tag{18}$$

Let $\mathbf{L}_i = [\mathbf{e}_i \otimes \mathbf{I}_k]$, where $\mathbf{e}_i$ is an $n_j \times 1$ null vector except for a "1" in its ith location. We then have the computational formula, in terms of $\boldsymbol{\Theta}$,

$$\frac{\partial l_F}{\partial \boldsymbol{\Theta}} = \frac{1}{2} \sum_{j=1}^{N} \sum_{i=1}^{n_j} \{\mathbf{L}_i' \boldsymbol{\Sigma}_j^{-1} \mathbf{L}_i - \mathbf{L}_i' \boldsymbol{\Sigma}_j^{-1} \mathbf{T}_j \boldsymbol{\Sigma}_j^{-1} \mathbf{L}_i\}.$$

Expected information matrix under ML. Bock (1989) shows that the expected information matrix under ML is

$$\mathbf{I}_F(\boldsymbol{\eta}, \text{vech}(\boldsymbol{\Psi}), \text{vech}(\boldsymbol{\Theta})) = \begin{bmatrix} \mathbf{I}_F(\boldsymbol{\eta}) & & \text{(symmetric)} \\ \mathbf{0} & \mathbf{I}_F(\text{vech}(\boldsymbol{\Psi})) & \\ \mathbf{0} & \mathbf{I}_F(\text{vech}(\boldsymbol{\Psi}), \text{vech}(\boldsymbol{\Theta})) & \mathbf{I}_F(\text{vech}(\boldsymbol{\Theta})) \end{bmatrix}. \tag{19}$$

The covariance partitions for $(\boldsymbol{\eta}, \text{vech}(\boldsymbol{\Psi}))$ and $(\boldsymbol{\eta}, \text{vech}(\boldsymbol{\Theta}))$ are null because they involve the expectations of third-order moments of residuals.

For $\boldsymbol{\eta}$, note that the residuals $\text{vec}(\mathbf{D}_j') \sim N(\mathbf{0}, \boldsymbol{\Sigma}_j)$, so that

$$\begin{aligned}\mathbf{I}_F(\boldsymbol{\eta}) &= \mathscr{E}\left[\frac{\partial l_F}{\partial \boldsymbol{\eta}} \frac{\partial l_F}{\partial \boldsymbol{\eta}'}\right] \\ &= \sum_{j=1}^{N} \mathbf{C}_j' \boldsymbol{\Sigma}_j^{-1} \mathscr{E}[\text{vec}(\mathbf{D}_j')\text{vec}(\mathbf{D}_j')'] \boldsymbol{\Sigma}_j^{-1} \mathbf{C}_j \\ &= \sum_{j=1}^{N} \mathbf{C}_j' \boldsymbol{\Sigma}_j^{-1} \mathbf{C}_j. \end{aligned} \tag{20}$$

For $\boldsymbol{\Psi}$, again let $p = (p_b k + m_b)$ be the order of $\boldsymbol{\Psi}$. Now note that

$$\begin{aligned}\frac{\partial l_F}{\partial \text{vech}(\boldsymbol{\Psi})} &= \frac{1}{2} \mathbf{G}_p' \sum_{j=1}^{N} \text{vec}\left(\frac{\partial l_F}{\partial \boldsymbol{\Psi}}\right) \\ &= -\frac{1}{2} \mathbf{G}_p' \sum_{j=1}^{N} (\mathbf{B}_j^{*\prime} \boldsymbol{\Sigma}_j^{-1} \otimes \mathbf{B}_j^{*\prime} \boldsymbol{\Sigma}_j^{-1}) \text{vec}(\mathbf{T}_j - \boldsymbol{\Sigma}_j)\end{aligned}$$

by using the well known relation between the vec(·) operator and the Kronecker product

$$\text{vec}(\boldsymbol{ABC}) = (\boldsymbol{C}' \otimes \boldsymbol{A})\text{vec}(\boldsymbol{B}),$$

for matrices $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ which are comformable to multiplication in that order, and McCulloch's (1982, p. 680) Theorem 1. In the latter, $\mathbf{G}_a$ is a unique $a^2 \times a(a+1)/2$ matrix of "0" and "1" such that

$$\text{vec}(\boldsymbol{A}) = \mathbf{G}_a \text{vech}(\boldsymbol{A})$$

for any symmetric matrix $\boldsymbol{A}$ of order a. The information matrix for $\boldsymbol{\Psi}$ is therefore

$$\begin{aligned}\mathbf{I}_F(\text{vech}(\boldsymbol{\Psi})) &= \mathscr{E}\left[\frac{\partial l_F}{\partial \text{vech}(\boldsymbol{\Psi})} \frac{\partial l_F}{\partial \text{vech}(\boldsymbol{\Psi})'}\right] \\ &= \frac{1}{2} \mathbf{G}_p' \sum_{j=1}^{N} (\mathbf{B}_j^{*\prime} \boldsymbol{\Sigma}_j^{-1} \mathbf{B}_j^{*} \otimes \mathbf{B}_j^{*\prime} \boldsymbol{\Sigma}_j^{-1} \mathbf{B}_j^{*}) \mathbf{G}_p. \end{aligned} \tag{21}$$

Thum

For $\boldsymbol{\Theta}$, using the vec(·) notation again, the information of the unique elements of $\boldsymbol{\Theta}$, is

$$\mathbf{I}_{\mathrm{F}}(\mathrm{vech}(\boldsymbol{\Theta})) = \mathscr{E}\left[\frac{\partial l_{\mathrm{F}}}{\partial \mathrm{vech}(\boldsymbol{\Theta})}\frac{\partial l_{\mathrm{F}}}{\partial \mathrm{vech}(\boldsymbol{\Theta})'}\right]$$

$$= \frac{1}{2}\mathbf{G}_k' \sum_{j=1}^{N}\sum_{i,h}^{n_j} (\mathbf{L}_i'\boldsymbol{\Sigma}_j^{-1}\mathbf{L}_h \otimes \mathbf{L}_i'\boldsymbol{\Sigma}_j^{-1}\mathbf{L}_h)\mathbf{G}_k. \tag{22}$$

Similarly, for $(\boldsymbol{\Psi}, \boldsymbol{\Theta})$,

$$\mathbf{I}_{\mathrm{F}}(\mathrm{vech}(\boldsymbol{\Psi}), \mathrm{vech}(\boldsymbol{\Theta})) = \mathscr{E}\left[\frac{\partial l_{\mathrm{F}}}{\partial \mathrm{vech}(\boldsymbol{\Psi})}\frac{\partial l_{\mathrm{F}}}{\partial \mathrm{vech}(\boldsymbol{\Theta})'}\right]$$

$$= \frac{1}{2}\mathbf{G}_p' \sum_{j=1}^{N}\sum_{i=1}^{n_j} (\mathbf{B}_j^{*\prime}\boldsymbol{\Sigma}_j^{-1}\mathbf{L}_i \otimes \mathbf{B}_j^{*\prime}\boldsymbol{\Sigma}_j^{-1}\mathbf{L}_i)\mathbf{G}_k. \tag{23}$$

Restricted Maximum Likelihood

Using (14), the maximum likelihood estimates (MLF estimates) of $(\boldsymbol{\Theta}, \boldsymbol{\Psi})$ are, however, biased downward because they fail to adjust for the degrees of freedom used in estimating $\boldsymbol{\eta}$. The precision of the maximum likelihood estimates is thus inflated, and the bias may be appreciable in small to moderate sample sizes. Harville (1977) suggested using the residual, or *restricted*, likelihood

$$l_{\mathrm{R}}(\boldsymbol{\eta}, \boldsymbol{\Theta}, \boldsymbol{\Psi}|\mathbf{y}) = l_{\mathrm{F}}(\boldsymbol{\eta}, \boldsymbol{\Theta}, \boldsymbol{\Psi}|\mathbf{y}) + \frac{1}{2}\ln\left|\sum_l \mathbf{C}_l'\boldsymbol{\Sigma}_l^{-1}\mathbf{C}_l\right| \tag{24}$$

for estimating $(\boldsymbol{\Theta}, \boldsymbol{\Psi})$ instead. The resulting REML estimates agree well with results for balanced designs for which closed-form solutions are available.

From (24), it is easy to see that the derivatives for a REML solution require an additional component in the MLF formulas for $\boldsymbol{\Psi}$ and for $\boldsymbol{\Theta}$. Writing

$$\mathbf{F} = \sum_l \mathbf{C}_l'\boldsymbol{\Sigma}_l^{-1}\mathbf{C}_l,$$

we have

$$\frac{\partial \frac{1}{2}\ln|\mathbf{F}|}{\partial \boldsymbol{\Psi}} = -\frac{1}{2}\sum_{j=1}^{N} \mathbf{B}_j^{*\prime}\boldsymbol{\Sigma}_j^{-1}\mathbf{C}_j\mathbf{F}^{-1}\mathbf{C}_j'\boldsymbol{\Sigma}_j^{-1}\mathbf{B}_j^{*}. \tag{25}$$

As for $\boldsymbol{\Theta}$, similar derivation predictably gives

$$\frac{\partial \frac{1}{2}\ln|\mathbf{F}|}{\partial \boldsymbol{\Theta}} = -\frac{1}{2}\sum_{j=1}^{N}\sum_{i=1}^{n_j}\mathbf{L}_i'\boldsymbol{\Sigma}_j^{-1}\mathbf{C}_j\mathbf{F}^{-1}\mathbf{C}_j'\boldsymbol{\Sigma}_j^{-1}\mathbf{L}_i. \tag{26}$$

These expressions may then be employed, as in the previous section, to give the expected information matrix under REML.

Patterned Matrices

In the preceding sections, we develop the first and second derivatives—(11) and (12), respectively—for a scalar function of linear, arbitarily constrained parameter vectors. Here we briefly consider the first and second derivatives of patterned matrices, such as when the variance-covariance matrix may be presumed to be univariate iid, to be diagonal, or to evidence first-order correlations.

When $\boldsymbol{\Theta} = \theta^2\mathbf{I}_k$, we have the well known result

$$\frac{\partial l_F}{\partial \theta^2} = \operatorname{tr}\left(\frac{\partial l_F}{\partial \boldsymbol{\Theta}}\right),$$

since $\partial\boldsymbol{\Theta}/\partial\theta^2 = \mathbf{I}_k$. It is also easily verified that

$$\operatorname{tr}\left(\frac{\partial l_F}{\partial \boldsymbol{\Theta}}\right) = \operatorname{vech}(\mathbf{I}_k)'\,\frac{\partial l_F}{\partial \operatorname{vech}(\boldsymbol{\Theta})}.$$

Thus, writing in terms of previous results for the information matrix under ML,

$$\mathbf{I}_F(\theta^2) = \operatorname{vech}(\mathbf{I}_k)'\mathbf{I}_F(\operatorname{vech}(\boldsymbol{\Theta}))\operatorname{vech}(\mathbf{I}_k).$$

For $\boldsymbol{\Theta} = \operatorname{diag}(\theta_1^2, \theta_2^2, \ldots, \theta_k^2)$, simply extract the rows of $\partial l_F/\partial\boldsymbol{\Theta}$ which correspond to the diagonal elements of $\boldsymbol{\Theta}$ for the first derivatives. Similarly, the rows and columns of $\mathbf{I}_F(\operatorname{vech}(\boldsymbol{\Theta}))$ which correspond to the diagonal elements of $\boldsymbol{\Theta}$ provide the correct information matrix.

We may write $\boldsymbol{\Theta} = \theta^2\{\rho^{|l-j|}\}$, $l, j = 1, 2, \ldots, k$ and $-1 < \rho < 1$, when errors display first-order correlation, that is,

$$\boldsymbol{\Theta} = \theta^2\begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{k-1} \\ \rho & 1 & \rho & \cdots & \rho^{k-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{k-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{k-1} & \rho^{k-2} & \rho^{k-3} & \cdots & 1 \end{bmatrix} = \theta^2\mathbf{K}_\theta.$$

Arguing as before, we see that

$$\frac{\partial l_F}{\partial \theta^2} = \text{tr}\left(\frac{\partial l_F}{\partial \boldsymbol{\Theta}} \frac{\partial \boldsymbol{\Theta}}{\partial \theta^2}\right) = \text{vech}(\mathbf{K}_\theta)' \frac{l_F}{\partial \text{vech}(\boldsymbol{\Theta})}$$

and

$$\mathbf{I}_F(\theta^2) = \text{vech}(\mathbf{K}_\theta)'\mathbf{I}_F(\text{vech}(\boldsymbol{\Theta}))\text{vech}(\mathbf{K}_\theta).$$

Writing $\partial\boldsymbol{\Theta}/\partial\rho = \theta^2\partial\mathbf{K}_\theta/\partial\rho = \mathbf{H}_\theta$, we similarly have

$$\frac{\partial l_F}{\partial \rho} = \text{vech}(\mathbf{H}_\theta)' \frac{\partial l_F}{\partial \text{vech}(\boldsymbol{\Theta})}$$

and

$$\mathbf{I}_F(\rho) = \text{vech}(\mathbf{H}_\theta)'\mathbf{I}_F(\text{vech}(\boldsymbol{\Theta}))\text{vech}(\mathbf{H}_\theta).$$

Furthermore, the covariance term is clearly

$$\mathbf{I}_F(\theta^2, \rho) = \text{vech}(\mathbf{K}_\theta)'\mathbf{I}_F(\text{vech}(\boldsymbol{\Theta}))\text{vech}(\mathbf{H}_\theta).$$

A good illustration of the use of patterned error covariance structure may be found in the variant of the model for first-order correlation attempted by Hopper and Mathews (1982) which factors $\boldsymbol{\Theta}$ as

$$\theta_1\{\theta_2\mathbf{I}_k + (1 - \theta_2)\theta_3\mathbf{K}_\theta\},$$

where $\theta_1 \geq 0$ represents the total dispersion, θ_2 is interpreted as a heritability parameter, and $-1 \leq \theta_3 \leq 1$ as some decay constant. Finally, we conclude this section by noting that a similar strategy underlies derivations for other patterned matrices, such as the band matrix, although the use of arbitary constraints as suggested earlier may become increasingly flexible and efficient from a programming standpoint as the number of unique parameters in $\boldsymbol{\Theta}$ grows. All the above results pertain to derivatives under a restricted likelihood, as well.

Multivariate t *Prior*

We assume that $\boldsymbol{\beta}_j$ is drawn from the p-variate multivariate t distribution with λ degrees of freedom

$$\pi(\boldsymbol{\beta}_j; \lambda, \boldsymbol{\gamma}, \boldsymbol{\Psi}) = t(\lambda, \mathbf{X}_j\boldsymbol{\gamma}, \boldsymbol{\Psi}), \tag{27}$$

where λ is positive and may be a noninteger (Cornish, 1954; Dunnett & Sobel, 1954).

Several properties of the multivariate t are well known. Small λ values represent heavy-tailed densities, and (27) approaches the multivariate normal distribution with covariance matrix $\boldsymbol{\Psi}$ as λ approaches infinity. If $\lambda > 2$,

$$\left[\frac{\lambda}{\lambda - 2}\right]\boldsymbol{\Psi}$$

may be viewed as the *subjective covariance matrix* of $\boldsymbol{\beta}_j$ (Berger, 1985, p. 245). In the case of univariate β_j and $\lambda = 1$, (27) is Cauchy $C(\lambda, \psi^{1/2})$. The principal advantage of this model is that it has all the familiar interpretational components of (6). At the same time, the use of heavy-tailed priors, such as the t-density, robustifies inferences against distributional assumptions in the sense that inferences are relatively insensitive to moderate misspecification of the prior distribution (see also Berger, 1985, p. 228). This property of a multivariate t prior is, of course, analogous to the effectiveness of the multivariate t in reducing the influence of outlying observations on the likelihood (Lange, Little, & Taylor, 1989). Furthermore, the wide range in distributional forms of the multivariate t density provides the researcher with an invaluable tool with which to assess the sensitivity of his inference to prior distributional assumptions.

Maximum Likelihood

It turns out that the desired solution requires only simple modifications of the algorithm for ML estimation of the multivariate normal-normal model. We begin by employing a convenient hierarchical representation of the multivariate t density:

$$t(\lambda, \mathbf{X}_j\boldsymbol{\gamma}, \boldsymbol{\Psi}) = \int_0^\infty N\left(\mathbf{X}_j\boldsymbol{\gamma}, \frac{1}{v}\boldsymbol{\Psi}\right) G\left(\frac{\lambda}{2}, \frac{2}{\lambda}\right) dv$$

for $\lambda \geq 2$ and $v > 0$. ($G(\lambda_1, \lambda_2)$ is the gamma density with parameters $\lambda_1 > 0$ and $\lambda_2 > 0$.) It is easy to see that the resulting marginal density for subject j is

$$h_j = \int_0^\infty h_j^*(v)\, \pi(v)\, dv,$$

where

$$h_j^*(v) = N(\mathbf{C}_j\boldsymbol{\eta}, \boldsymbol{\Sigma}_j(v)),$$

$$\pi(v) = G\left(\frac{\lambda}{2}, \frac{2}{\lambda}\right),$$

and

$$\boldsymbol{\Sigma}_j(v) = \left[\frac{1}{v}\mathbf{B}_j^*\boldsymbol{\Psi}\mathbf{B}_j^{*\prime} + \boldsymbol{\Theta}_j\right],$$

which leads to the marginal likelihood

$$\begin{aligned} l_t(\boldsymbol{\eta}, \boldsymbol{\Theta}, \boldsymbol{\Psi}|\mathbf{y}) &= \sum_{j=1}^{N} \ln h_j \\ &= \sum_{j=1}^{N} \ln\left(\int_0^\infty h_j^*(v)\, \pi(v)\, dv\right). \end{aligned} \tag{28}$$

Here, v denotes the variable of integration, even as writing $h_j^*(v)$ could have implied that the density depends on v. $h_j^*(v)$ is, of course, a function of ($\boldsymbol{\eta}$, $\boldsymbol{\Theta}$, $\boldsymbol{\Psi}$), assuming additionally that λ is known. (Freeing λ is the basis of *adaptive* robust procedures.) Evaluation of the likelihood function will involve a one-dimensional integral for $v > 0$, which may be accomplished adequately by, for example, the Gauss-Laguerre quadrature for integrals that can be expressed in the form

$$\int_0^\infty f(x)\, e^{-x}\, dx$$

(see, e.g., Weeg & Reed, 1966, p. 159).

Derivatives. Generally, the derivative of l_t with respect to some vector $\boldsymbol{\nu}$ is

$$\begin{aligned} \frac{\partial l_t}{\partial \boldsymbol{\nu}} &= \sum_{j=1}^{N} \frac{1}{h_j}\frac{\partial h_j}{\partial \boldsymbol{\nu}} \\ &= \sum_{j=1}^{N} \frac{1}{h_j}\int_0^\infty \left(\frac{\partial h_j^*(v)}{\partial \boldsymbol{\nu}}\right) \pi(v)\, dv. \end{aligned} \tag{29}$$

Since

$$\frac{\partial h_j^*(v)}{\partial \boldsymbol{\nu}} = h_j^* \frac{\partial \ln h_j^*(v)}{\partial \boldsymbol{\nu}},$$

the first derivatives are simply

$$\frac{\partial l_t}{\partial \boldsymbol{\nu}} = \sum_{j=1}^{N} \frac{1}{h_j} \int_0^{\infty} \left(\frac{\partial \ln h_j^*(v)}{\partial \boldsymbol{\nu}} \right) h_j^*(v)\, \pi(v)\, dv. \tag{30}$$

For each subject j, (30) is the corresponding derivative for the normal-normal model weighted by the t marginal. To obtain standard errors, note that although the matrix of second derivatives is rather complex under (28), the expected information matrix at convergence is simply

$$\begin{aligned} \mathbf{I}_t(\boldsymbol{\nu}) &= \sum_{j=1}^{N} \frac{1}{h_j} \int_0^{\infty} \mathscr{E}\left[\left(\frac{\partial \ln h_j^*(v)}{\partial \boldsymbol{\nu}} \right) \left(\frac{\partial \ln h_j^*(v)}{\partial \boldsymbol{\nu}'} \right) \right] h_j^*(v)\, \pi(v)\, dv \\ &= \sum_{j=1}^{N} \frac{1}{h_j} \int_0^{\infty} \mathbf{I}_{F_j}(v, \boldsymbol{\nu})\, h_j^*(v)\, \pi(v)\, dv. \end{aligned} \tag{31}$$

Thus, all the basic programming structures, already in place for the normal-normal model, remain valid for the normal-t case.

Examples

Teacher Engagement in School Reform

This analysis is based on ongoing research at the Consortium on Chicago School Research at the University of Chicago.[2] Data are drawn from a survey conducted to evaluate the impact of the reform agenda of Chicago's public schools on the teacher's attitudes and behavior, and vice versa. A total of 11,253 primary school teachers from 401 schools provided information on the variables listed in Table 2. We examine school-level correlates of two outcome variables measuring teacher's level of engagement in reform, *TIME* and *PLAN*.

We begin with two separate unconditional models, one for each indicator of support since reform:

$$\begin{aligned} y_{ij}^{(1)} &= \beta_j^{(1)} + e_{ij}^{(1)}, \\ \beta_j^{(1)} &= \gamma^{(1)} + r_j^{(1)}; \end{aligned}$$

and

$$\begin{aligned} y_{ij}^{(2)} &= \beta_j^{(2)} + e_{ij}^{(2)}, \\ \beta_j^{(2)} &= \gamma^{(2)} + r_j^{(2)}, \end{aligned}$$

where the superscript "1" denotes the outcome variable *TIME*, and "2" denotes *PLAN*. Results are given in columns 2 and 3 of Table 3. For both outcome

TABLE 2
Study of teachers' engagement in school reform: Definitions of variables

Variable	Interpretation
Outcomes[a]	
PLAN	Spends more time participating in school governance
TIME	Spends more time on specific teaching activities
Teacher-level covariates	
CLASSRM	1 if teacher assigned to classroom, 0 otherwise
EFFICACY	Teacher feels efficacious
GRPK_3	1 if K–3 teacher, 0 otherwise
MINORITY	1 if minority teacher, 0 otherwise
YRSTGHT	Number of years taught
School-level correlates	
APSLOIN	Arcsine of % low income
COMACH89	Overall re-reform achievement
LSENR	Log of total school enrollment
PROFCOM	Level of professional community support

[a]Outcomes are Rasch measures in logits.

variables, there seems to be greater variability in the measured levels of engagement among teachers than across schools. A third model which considered both outcomes jointly,

$$[y_{ij}^{(1)}\ y_{ij}^{(2)}] = [\beta_j^{(1)}\ \beta_j^{(2)}] + [e_{ij}^{(1)}\ e_{ij}^{(2)}] \tag{32}$$

TABLE 3
Teachers' level of engagement in school reform

	MHLM model		
Parameter	*TIME* Est. (*SE*)	*PLAN* Est. (*SE*)	*TIME, PLAN* Est. (*SE*)
Intercept			
TIME	0.0064 (0.0307)		0.0058 (0.0306)
PLAN		−0.5998 (0.0213)	−0.5999 (0.0212)
Residual parameter variance			
TIME	0.2238 (0.0263)		0.2235 (0.0263)
TIME/PLAN			0.0212 (0.0131)
PLAN		0.1401 (0.0127)	0.1395 (0.0127)
Error variance			
TIME	3.8663 (0.0525)		3.8664 (0.0525)
TIME/PLAN			0.5267 (0.0195)
PLAN		0.9985 (0.0136)	0.9986 (0.0136)

and

$$[\beta_j^{(1)}\ \beta_j^{(2)}] = [\gamma^{(1)}\ \gamma^{(2)}] + [r_j^{(1)}\ r_j^{(2)}], \tag{33}$$

provides additional information. Column 4 in Table 3 suggests that measurement errors are not uncorrelated as assumed by the independent analyses (correlation is .27). We also find that the mean levels of engagement, as measured separately by $\beta^{(1)}$ and $\beta^{(2)}$, are positively correlated (.16). The redundancy in information in the measures does not have as substantial an impact on the standard errors (from .307 to .306 for *TIME* and from .0213 to .0212 for *PLAN*) as expected, but this is probably due to the counteracting effect of a large sample size. Nevertheless, the likelihood ratio chi-square for adding the two covariance terms in the joint multivariate model when compared with the independent analyses is 47,526.124 + 32,535.640 − 79,242.127 = 819.637 on 2 degrees of freedom.

Results for a final set of conditional models are summarized in Table 4. Again, similar analyses are performed for each outcome variable (columns 2 and 3). We first adjust the level of engagement for the teacher-level covariates in Table 2. The set of school-level variables was then used to model the adjusted mean level of engagement for each school. Compared with the unconditional model above, the adjusted level of engagement as measured by *TIME* seemed to improve since reform where there is a higher level of support from other members of the school as a professional community (*PROFCOM*). Higher achievement at the school level prior to reform (*COMACH89*) also seemed to be associated with a lower adjusted level of commitment of additional time to teaching, however. These factors were sufficiently powerful to reduce parameter variance from .2238 (in the unconditional model) to .1340. At the same time, decreases in the level of teacher participation in school governance since reform (*PLAN*) were found in larger schools (*LSENR*), schools with more low-income students (*APSLOIN*), and, surprisingly, schools where there was evidence of higher support from the school professional community (*PROFCOM*). These results are, however, somewhat weak. The reduction in the variance of the adjusted level of participation is modest at best, from 0.1401 to 0.1372.

Column 4 in Table 4 gives the results of the corresponding multivariate model which, in extenso, is

$$[y_{ij}^{(1)}\quad y_{ij}^{(2)}] = \left[\sum_p a_{ijp}\alpha_p^{(1)} \quad \sum_p a_{ijp}\alpha_p^{(2)}\right] + \left[\sum_q x_{jq}\gamma_q^{(1)} \quad \sum_q x_{jq}\gamma_q^{(2)}\right] + [r_j^{(1)}\quad r_j^{(2)}] + [e_{ij}^{(1)}\quad e_{ij}^{(2)}], \tag{34}$$

for $p = 1, 2, \ldots, 5$ and $q = 1, 2, \ldots, 5$. a_{ij1} is teacher-level covariate $CLASSRM_{ij}$, a_{ij2} is $GRPK\_3_{ij}$, and so on. Similarly, x_{j1} is the intercept term,

TABLE 4
Teacher-level and school-level correlates of teachers' level of engagement in reform

Parameters	MHLM model		
	TIME Est. (*SE*)	*PLAN* Est. (*SE*)	*TIME, PLAN* Est. (*SE*)
Teacher covariates of *TIME*			
CLASSRM	−0.3470 (0.0434)		−0.3473 (0.0434)
GRPK_3	0.0149 (0.0480)		0.0147 (0.0479)
MINORITY	0.2101 (0.0433)		0.2089 (0.0433)
YRSTGHT	−0.0249 (0.0184)		−0.0244 (0.0184)
EFFICACY	0.4827 (0.0187)		0.4835 (0.0187)
Teacher covariates of *PLAN*			
CLASSRM		−0.0857 (0.0226)	−0.0850 (0.0226)
GRPK_3		−0.0923 (0.0250)	−0.0922 (0.0250)
MINORITY		−0.0009 (0.0231)	0.0002 (0.0231)
YRSTGHT		0.1597 (0.0096)	0.1593 (0.0096)
EFFICACY		0.1001 (0.0098)	0.1004 (0.0098)
School correlates of *TIME*			
INTERCEPT	0.0081 (0.0266)		0.0082 (0.0266)
APSLOIN	0.0575 (0.0382)		0.0577 (0.0382)
LSENR	−0.0344 (0.0295)		−0.0348 (0.0295)
COMACH89	−0.1377 (0.0401)		−0.1378 (0.0401)
PROFCOM	0.1551 (0.0300)		0.1556 (0.0300)
School correlates of *PLAN*			
INTERCEPT		−0.5960 (0.0211)	−0.5959 (0.0211)
APSLOIN		−0.0640 (0.0301)	−0.0640 (0.0301)
LSENR		−0.0681 (0.0230)	−0.0684 (0.0229)
COMACH89		−0.0002 (0.0312)	−0.0004 (0.0311)
PROFCOM		−0.0572 (0.0236)	−0.0575 (0.0236)
Residual parameter variance			
TIME	0.1340 (0.0190)		0.1346 (0.0190)
TIME/PLAN			0.0382 (0.0112)
PLAN		0.1372 (0.0124)	0.1368 (0.0124)
Error variance			
TIME	3.6065 (0.0489)		3.6063 (0.0489)
TIME/PLAN			0.4770 (0.0184)
PLAN		0.9605 (0.0130)	0.9606 (0.0130)

x_{j2} is the school-level correlate $APSLOIN_j$, and so on. As before, the pattern of results for the fixed effects did not appear to contradict conclusions drawn from the separate univariate analyses. Note that the standard error for the school-size (*LSENR*) effect decreased. While the measurement errors are correlated as before, the residual parameter variance now shows a modest correlation of .28. The improvement in fit with the multivariate model is still evident. The likelihood ratio chi-square is now 44,647.156 + 32,104.884 − 77,986.933 = 765.107 on 2 degrees of freedom.

Infant Vocabulary Growth

In this example, we consider robust inference for fixed effects under maximum likelihood. In particular, we compare several models for data from a developmental study of vocabulary growth in infants (Huttenlocher, Haight, Bryk, Seltzer, & Lyons, 1991). Two groups of mother-infant pairs, with 6 boys and 5 girls in each, provide from 5 to 7 observations per infant. Of interest is how gender differences and differential amounts of maternal speech output occurring in normal mother-child interactions within a specified observation period relate to the cumulative growth of an infant's vocabulary size. The time dimension is $(AGE_{ij} - 12)$ months, assuming that at 12 months a child's vocabulary size is zero.

Huttenlocher et al. (1991) found the following quadratic model suitable for describing individual growth in vocabulary:

$$y_{ij} = \beta_j(AGE_{ij} - 12)^2 + e_{ij}, \tag{35}$$

where β_j is the acceleration of vocabulary acquisition for child j. The effects of *GROUP* ("0" for Group 1 or "1" for Group 2), *GIRL* ("1" for girls), and *MOMSPEAK* (natural log of the number of words spoken to child j in a 3-hour observation period at age 16 months) are introduced in an individual difference model for the acceleration measure β_j:

$$\beta_j = \gamma_0 + \gamma_1\, GROUP_j + \gamma_2\, MOMSPEAK_j + \gamma_3\, GIRL_j + r_j. \tag{36}$$

We begin by examining the overall summaries of fit for the above model using several alternative likelihoods, each with 6 parameters, given in Table 5. Fit for the REML solution is shown along with results for three models under ML. Note that fits under REML and MLF are not always directly

TABLE 5
*Alternative ML quadratic growth models (2*log(likelihood))*

REML	MLF	t_{11}	t_4
1,288.28	1,278.44	1,277.26	1,275.73

comparable. We see that for these data a t_4 residual parameter error model gives the best maximum likelihood solution under MLF.

We next compare estimates from a *joint* maximum likelihood solution with results from a Bayesian approach which computes *marginal* estimates as previously given by Seltzer (1993) using the Gibbs sampler under alternative priors for the acceleration parameter β_j. It is argued that a Bayesian approach, by incorporating uncertainty about the variance components, produces more realistic (larger) estimates of estimation errors for the fixed effects. This is evident from a rough comparison of the 95% credibility interval of the Bayesian estimate with the standard error for the ML estimate for each fixed effect under the same prior for β_j in Table 6. One notable result of this comparison is that the Bayesian model calls into question the presence of a gender effect (γ_3).

For each fixed effect, Table 6 also provides a comparison of the effect of employing alternative priors under a Bayesian approach with maximum likelihood. Changes in the estimates under different priors for each effect show a similar pattern within each approach. Of particular interest are the measures of estimation errors. Because a heavy-tailed distribution, such as a t with 4 or 11 degrees of freedom, downweights outlying observations by an inverse function of the Mahalanobis distance adjusted for the t degrees

TABLE 6
Fixed effects estimates under alternative priors

		Model			
		Bayesian[a]		Max. lik.	
Effect	Prior	Mean	95% C.I.	Est.	*SE*
Base	Normal	−4.89	(−11.25, 1.50)	−4.92	(2.58)
(γ_0)	t_{11}	−4.71	(−10.62, 1.03)	−4.64	(2.17)
	t_4	−4.52	(−9.79, 0.68)	−4.39	(1.64)
Group	Normal	−1.11	(−1.94, −0.30)	−1.11	(0.33)
(γ_1)	t_{11}	−1.15	(−1.91, −0.36)	−1.16	(0.29)
	t_4	−1.18	(−1.89, −0.45)	−1.23	(0.23)
Maternal	Normal	0.88	(0.07, 1.70)	0.89	(0.33)
speech	t_{11}	0.86	(0.14, 1.61)	0.85	(0.28)
(γ_2)	t_4	0.84	(0.18, 1.50)	0.82	(0.21)
GIRL	Normal	0.80	(−0.07, 1.66)	0.80	(0.35)
(γ_3)	t_{11}	0.90	(0.08, 1.70)	0.93	(0.30)
	t_4	0.98	(0.21, 1.71)	1.04	(0.24)
Residual	Normal			0.57	(0.18)
variance	t_{11}			0.42	(0.13)
(ψ)	t_4			0.24	(0.08)

[a]See Seltzer, 1993.

of freedom, it generally produces smaller variance estimates. In our example, this translates first into a more modest estimate of the residual parameter variance in the growth model and then into smaller standard errors for point estimates. The overall agreement of both methods is remarkable for this data, keeping in mind that the Bayesian approach produces marginal mean estimates as compared with the joint modal estimates of the likelihood function under maximum likelihood.

Growth Curve With Data Missing

It is all too common in behavioral data for some outcome measures to be missing, either by accident or by design. Using data presented in Potthoff and Roy (1964) and further reanalyzed in Jennrich and Schlucter (1986) and Little and Rubin (1987), we present maximum likelihood solutions to a growth curve with a random intercept and a random slope. This model is not the best possible model for the data, and the results below are presented for the sole purpose of illustrating the feasibility of MHLMs with missing outcome data.

The outcome vector, $\mathbf{y}_{ij}$, comprises distances from the center of the pituitary to the maxillary fissure obtained at ages 8, 10, 12, and 14 for each of 16 boys and 11 girls in the study. Let

$$\mathbf{M} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}.$$

Furthermore, we set

$$\mathbf{X}_j = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & x_j \end{bmatrix},$$

with x_j equal to 1 for a boy or 0 for a girl.

Results are summarized in Table 7. With complete data, an autoregressive error model with lag 1 for Θ fits the data better than a model with the usual assumption that measurement errors are independently and identically distributed normal (see columns 2 and 3). The autocorrelation is a large $-.4746$, so we may conclude that the error variance is reduced with a compensating increase in residual parameter variance (now significant). Columns 4 and 5 contain the results for models with an $AR(1)$ error structure for a subset of the original data (about 8% of the outcomes were ignored), as selected by Little and Rubin (1987, p. 159), under a normal prior and a

TABLE 7
Alternative MHLMs for Potthoff-Roy growth data when some data are missing

	Complete data		Missing data	
	Normal		Normal	t_4
Parameter	Θ iid normal Est. (*SE*)	Θ *AR*(1) Est. (*SE*)	Θ *AR*(1) Est. (*SE*)	Θ *AR*(1) Est. (*SE*)
Constant	20.7222	20.6675	20.5841	20.6732
	(0.4619)	(0.4461)	(0.4633)	(0.3819)
Linear.INT	0.8774	0.8829	0.9001	0.8887
	(0.1760)	(0.1803)	(0.1859)	(0.1573)
Linear.BOY	0.7474	0.7601	0.7741	0.6519
	(0.2038)	(0.2117)	(0.2148)	(0.1836)
Residual parameter variance				
Const/Const	3.1865	4.3302	4.3704	2.2889
	(1.6443)	(1.5168)	(1.6253)	(0.9320)
Const/Linear	−0.1379	−0.5622	−0.6208	−0.2144
	(0.3981)	(0.3754)	(0.4132)	(0.2255)
Linear/Linear	0.0996	0.3059	0.3310	0.1593
	(0.1374)	(0.1349)	(0.1485)	(0.0834)
Error variance				
θ	1.7162	1.1920	1.2590	1.2003
	(0.3303)	(0.2319)	(0.2678)	(0.2477)
ρ		−0.4746	−0.5120	−0.4789
		(0.1728)	(0.1806)	(0.1726)
−2 Log Likelihood	428.5342	424.6055	392.5972	390.2039

t prior respectively. The latter model fitted the data marginally better, with an all-around reduction in standard error estimates.

Summary and Discussion

During the past decade, there has been clear and growing recognition among educational researchers who are interested in evaluating program effects, in the study of individual change, and in describing a multitude of contextual influences on observed patterns of effects or change that the standard general linear model falls short in meeting the vexing methodological problems posed by unbalanced, nested, longitudinal data. These problems demand a more flexible class of models that are capable of accommodating a wider range of structural, as well as distributional, assumptions. Although the steady accumulation of innovative responses which has followed this

recognition began as several independent streams of methodological research (see Bryk & Raudenbush, 1992, p. 3, for a brief ontogeny of the two-stage hierarchical linear model), the class of multilevel models that has taken shape in educational research boasts several useful extensions for educational and behavioral data. Broadly, they include models for data with two as well as three levels of nesting, for discrete outcomes, for data designs with crossed random effects, and for multivariate and latent variables. Collectively, the new texts by Bryk and Raudenbush (1992), Goldstein (1987, 1995), and Longford (1993) provide the authoritative coverage of these important methodological developments in recent educational research. This article elaborated on the *standard* univariate two-stage normal-normal model to accommodate three common characteristics of behavioral data.

First, we note that behavior is invariably multivariate in its conceptualization and communication. Separate univariate analyses of related outcome variables are fraught with potential interpretive blind spots for the researcher. This practice also suffers, from an inferential standpoint, because it fails to take advantage of any redundant information in the outcomes. Multivariate outcomes include, in their most general form, *doubly* multivariate response vectors (Bock, 1975). In this article, we discussed specifications of a measurement model matrix and the use of covariates following the general linear model at Level 1.

Other alternatives to the Stage 1 multivariate model are based on the three-level model. These approaches to multivariate, normal-normal models are identical in the following way: Level 1 of the three-level model takes the form of a measurement model for the unrepeated response vector of each lowest-level unit. Generally, a measurement design matrix of 1s and 0s (the identity matrix, or its subset, and even a matrix of contrasts, estimated or defined a priori, could be used) and, possibly, a matrix of covariates could be specified. In one approach, Raudenbush, Rowan, and Kang (1991) used the univariate, Level 1 model as a *true score* model. But presupposing that the Level 1 error term is iid normal is a simplification that is not easily justified unless we are analyzing a repeated measures problem. If the observed scores are derived measurements which are further accompanied by their standard errors, we can rescale the Level 1 model into an *iid unit-normal model*, hence fixing the Level 1 error variance at unity, but this still presumes that the errors of measurement are uncorrelated. In yet another approach, adopted by both Goldstein (1995, p. 70) and Longford (1993, p. 117), the authors suggest fixing the error variance at Level 1 at zero, thus collapsing the lowest two levels of the three-level model into a single multivariate lower-level model—and the univariate, three-level model into a multivariate, two-level model. For some three-level programs, this last strategy required setting the measurement error component to a small but arbitrary value. Alternatively, the likelihood can be rewritten to dispense with estimating the measurement

error component entirely—and this is where reduction of the three-level approach is equivalent to the two-level MHLM.

Second, studies of behavior, especially in experimental research, employ smaller samples. This situation raises issues of robustness of inference with respect to outlying individuals. The approach using a multivariate t is shown to be useful and relatively easy to implement. Maximum likelihood estimates are then obtained for various members of a multivariate hierarchical linear model in the context of several illustrative examples. Results match closely with estimates from a Bayesian approach both for the normal-normal MHLM and for the normal-t MHLM.

But while the t-prior offers some protection in this regard, problems with small samples do not end here. We recognize that parameter estimation and inference for these models are based on large sample assumptions, and obviously this would be problematic in many applications involving only small to moderate numbers of nesting units, N (e.g., Seltzer, 1993). In particular, when inferences concerning fixed effects are drawn (e.g., parameters capturing how differences in school characteristics relate to differences in SES-achievement slopes), maximum likelihood estimates of the variance components are treated as known, true values. (Such an approach is termed *empirical Bayes*.) But when N is small, MLEs for the variance components may be poorly determined, and confidence intervals based on them will, as a consequence, be highly misleading (Seltzer, 1994).

One alternative is to assert that we know enough about the parameters, even if only by assumption at the start, to augment the basic model with selected priors. In a marked departure from using maximum likelihood optimization of the empirical Bayes model, Seltzer (1993) illustrates a Bayesian hierarchical linear approach that involves placing priors on the variance components in the hierarchical linear model. The Bayesian approach takes into account the uncertainty in variance components by treating them as unknowns, each given a prior probability distribution. Computational procedures such as the Gibbs sampler are then used to integrate over all the unknowns in the model to obtain the marginal posterior distributions of the parameters of interest. It should be clear that, independently of the method of estimation or the specific forms of inference, differences in the two analyses can be predicted generally. For example, we expect that increasing the number of units will reduce any major discrepancies as the information in the data begins to overwhelm the intended constraints enforced by the chosen prior.

A series of more detailed comparisons of the results from the Bayesian model with the results from an empirical Bayes analysis could well be the topic of future studies. Any insights from such comparisons serve not only to remove potentially conflicting conclusions sometimes offered by the two approaches (as regards the gender effect example) but also to serve as a useful guide considering the relatively heavy computational burden of Bayesian

computations via tools such as the Gibbs sampler. Using recent results on efficient integration (e.g., Leonard, Hsu, & Tsui, 1989, and others), we may explore various approximations to the marginal densities of fixed effects and variance components from their joint posterior density under the Bayesian model, to be obtained via a quasi-Newton algorithm such as the one employed in this article, and compare the inference based on these approximations with results based on marginal posterior densities from a Bayesian analysis via the Gibbs sampler. These comparisons could be made within the context of a small simulation study that varies, principally, the number of Level 2 units and the number of random effects in a two-stage hierarchical linear model.

A third and often troublesome feature of multivariate behavioral data is that the outcome vector may have observations missing due to accidents, or by design. The quasi-Newton algorithm employed in this study permits the estimation of the full spectrum of plausible measurement error structures while using all the available information. Internally, it turns out that missing measures are handled in exactly the same way in multivariate two-stage models based on collapsing Level 1 into Level 2 of a three-level model (Longford, 1993).

Notes

[1]Throughout this study, $\otimes$ denotes the left Kronecker product.

[2]See *Charting Reform: The Teacher's Turn* by the Consortium on Chicago School Research, 1991, Chicago: Author. We thank Eric Camburn for his assistance with the data.

References

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis.* New York: Springer-Verlag.

Bock, R. D. (1975). *Multivariate statistical methods in behavioral research.* New York: McGraw-Hill.

Bock, R. D. (1989). Addendum—Measurement of human variation: A two-stage model. In R. D. Bock (Ed.), *Multilevel analysis of educational data* (pp. 319–342). New York: Academic Press.

Bryk, A. S., & Raudenbush, S. W. (1985). Empirical Bayes meta-analysis. *Journal of Educational Statistics, 10,* 75–98.

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

Burley, D. M. (1974). *Studies in optimization.* New York: Wiley.

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

Cornish, E. A. (1954). The multivariate *t*-distribution associated with a set of normal standard deviates. *Australian Journal of Physics, 7,* 531–542.

Davidon, W. C. (1959). *Variable metric method for minimization* (Rep. ANL-5990 (Rev.)). Argonne, IL: Argonne National Laboratory.

de Leeuw, J., & Kreft, I. (1986). Random coefficient models for multilevel analysis. *Journal of Educational Statistics, 11*, 57–85.

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

Dunnett, C. W., & Sobel, M. (1954). A bivariate generalization of Student's *t*-distribution with tables for certain special cases. *Biometrika, 41*, 153–169.

Dwyer, P. S. (1967). Some applications of matrix derivatives in multivariate analysis. *Journal of the American Statistical Association, 62*, 607–625.

Fletcher, R. (1970). A new approach to variable metric algorithms. *Computer Journal, 13*, 317–322.

Fletcher, R., & Powell, M. J. D. (1963). A rapidly convergent descent method for minimization. *Computer Journal, 6*, 163–168.

Gelfand, A. E., Hills, S., Racine-Poon, A., & Smith, A. F. M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association, 85*, 398–409.

Goldstein, H. (1987). *Multilevel models in education and social research*. London: Oxford University Press.

Goldstein, H. (1995). *Multilevel statistical models*. New York: Halstead Press.

Goldstein, H., & McDonald, R. P. (1988). A general model for the analysis of multilevel data. *Psychometrika, 53*, 455–467.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association, 72*, 320–340.

Hopper, J. L., & Mathews, J. D. (1982). Extensions of the multivariate normal models for pedigree analysis. *Annals of Human Genetics, 46*, 373–383.

Huang, H. Y. (1970). Unified approach to quadratically convergent algorithms for function minimization. *Journal of Optimization Theory and Applications, 5*, 405–423.

Huttenlocher, J. E., Haight, W., Bryk, A. S., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology, 27*, 236–248.

Jennrich, R. J., & Schlucter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics, 42*, 805–820.

Jöreskog, K. G., & Sörbom, D. (1984). *LISREL VI: Analysis of linear structural relationships by maximum likelihood, instrumental variables and least squares methods*. Mooresville, IN: Scientific Software.

Lange, K. L., Little, R. J. A., & Taylor, J. M. G. (1989). Robust statistical modeling using the *t* distribution. *Journal of the American Statistical Association, 84*, 881–896.

Leonard, T., Hsu, J. S. J., & Tsui, K. W. (1989). Bayesian marginal inference. *Journal of the American Statistical Association, 84*, 1051–1058.

Lindstrom, M. J., & Bates, D. M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models of repeated-measures data. *Journal of the American Statistical Association, 83*, 1014–1022.

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

Longford, N. T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested effects. *Biometrika, 74*, 817–827.

Longford, N. T. (1993). *Random coefficient models*. Oxford, England: Clarendon Press.

Longford, N. T., & Muthén, B. O. (1992). Factor analysis for clustered observations. *Psychometrika, 57*, 581–597.

MacRae, E. C. (1974). Matrix derivatives with an application to an adaptive linear decision problem. *Annals of Statistics, 2*, 337–346.

Magnus, J. R., & Neudecker, H. (1979). The commutation matrix: Some properties and applications. *Annals of Statistics, 7*, 381–394.

McCulloch, C. E. (1982). Symmetric matrix derivatives with applications. *Journal of the American Statistical Association, 77*, 679–682.

McDonald, R. P., & Goldstein, H. (1989). Balanced versus unbalanced designs for linear structural relations in two-level data. *British Journal of Mathematical and Statistical Psychology, 42*, 215–232.

Muthén, B. (1989). Latent variable modelling in heterogeneous populations. *Psychometrika, 54*, 557–585.

Potthoff, R. F., & Roy, S. N. (1964). A generalized multivariate of variance model useful especially for growth curve problems. *Biometrika, 69*, 657–660.

Raudenbush, S. W. (1988). Educational applications of hierarchical models: A review. *Journal of Educational Statistics, 13*, 85–116.

Raudenbush, S. W., & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of Education, 59*, 1–17.

Raudenbush, S. W., Rowan, B., & Kang, S. J. (1991). A multilevel, multivariate model for studying school climate with estimation via the EM algorithm and application to U.S. high-school data. *Journal of Educational Statistics, 16*, 295–330.

Rogosa, D. R., & Willett, J. B. (1985). Understanding correlates of change by modelling individual differences in growth. *Psychometrika, 50*, 203–228.

Rubin, D. B. (1978). A note on Bayesian, likelihood, and sampling distribution inferences. *Journal of Educational Statistics, 3*, 189–201.

Schlucter, M. D. (1988). Analysis of incomplete multivariate data using linear models with structured covariance matrices. *Statistics in Medicine, 7*, 317–324.

Schmidt, W. H. (1969). *Covariance structure analysis of the multivariate random effects model*. Unpublished doctoral dissertation, University of Chicago.

Seltzer, M. H. (1991). *The use of data augmentation in fitting hierarchical linear models to educational data*. Unpublished doctoral dissertation, University of Chicago.

Seltzer, M. H. (1993). Sensitivity analysis for fixed effects in hierarchical models: A Gibbs sampling approach. *Journal of Educational Statistics, 18*, 207–235.

Seltzer, M. H. (1994, April). *Inference for variance components in hierarchical models: Problems and solutions*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Shanno, D. F., & Phua, K. H. (1974). Algorithm 500: Minimization of unconstrained multivariate functions. *ACM Transactions on Mathematical Software, 2*, 87–94.

Trawinski, I. M., & Bargmann, R. W. (1964). Maximum likelihood with incomplete multivariate data. *Annals of Mathematical Statistics, 35*, 647–657.

Thum, Y. M. (1994). *Analysis of individual variation: A multivariate hierarchical linear model for behavioral data.* Unpublished doctoral dissertation, University of Chicago.

Weeg, G. P., & Reed, G. A. (1966). *Introduction to numerical analysis.* Waltham, MA: Blaisdell.

Author

YEOW MENG THUM is Analysis Director, Consortium on Chicago School Research, University of Chicago, Chicago, IL 60637. He specializes in models of educational and psychological change and growth.

Received February 8, 1995
Revision received September 5, 1995
Accepted October 17, 1995

References

Empirical Bayes Meta-Analysis
Stephen W. Raudenbush; Anthony S. Bryk
Journal of Educational Statistics, Vol. 10, No. 2. (Summer, 1985), pp. 75-98.
Stable URL:
http://links.jstor.org/sici?sici=0362-9791%28198522%2910%3A2%3C75%3AEBM%3E2.0.CO%3B2-Z

Explaining the Gibbs Sampler
George Casella; Edward I. George
The American Statistician, Vol. 46, No. 3. (Aug., 1992), pp. 167-174.
Stable URL:
http://links.jstor.org/sici?sici=0003-1305%28199208%2946%3A3%3C167%3AETGS%3E2.0.CO%3B2-R

Random Coefficient Models for Multilevel Analysis
Jan de Leeuw; Ita Kreft
Journal of Educational Statistics, Vol. 11, No. 1. (Spring, 1986), pp. 57-85.
Stable URL:
http://links.jstor.org/sici?sici=0362-9791%28198621%2911%3A1%3C57%3ARCMFMA%3E2.0.CO%3B2-Z

Maximum Likelihood from Incomplete Data via the EM Algorithm
A. P. Dempster; N. M. Laird; D. B. Rubin
Journal of the Royal Statistical Society. Series B (Methodological), Vol. 39, No. 1. (1977), pp. 1-38.
Stable URL:
http://links.jstor.org/sici?sici=0035-9246%281977%2939%3A1%3C1%3AMLFIDV%3E2.0.CO%3B2-Z

LINKED CITATIONS

A Bivariate Generalization of Student's t-Distribution, with Tables for Certain Special Cases
Charles W. Dunnett; Milton Sobel
Biometrika, Vol. 41, No. 1/2. (Jun., 1954), pp. 153-169.
Stable URL:
http://links.jstor.org/sici?sici=0006-3444%28195406%2941%3A1%2F2%3C153%3AABGOSW%3E2.0.CO%3B2-2

Some Applications of Matrix Derivatives in Multivariate Analysis
Paul S. Dwyer
Journal of the American Statistical Association, Vol. 62, No. 318. (Jun., 1967), pp. 607-625.
Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28196706%2962%3A318%3C607%3ASAOMDI%3E2.0.CO%3B2-I

Sampling-Based Approaches to Calculating Marginal Densities
Alan E. Gelfand; Adrian F. M. Smith
Journal of the American Statistical Association, Vol. 85, No. 410. (Jun., 1990), pp. 398-409.
Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28199006%2985%3A410%3C398%3ASATCMD%3E2.0.CO%3B2-3

Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems
David A. Harville
Journal of the American Statistical Association, Vol. 72, No. 358. (Jun., 1977), pp. 320-338.
Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28197706%2972%3A358%3C320%3AMLATVC%3E2.0.CO%3B2-9

Unbalanced Repeated-Measures Models with Structured Covariance Matrices
Robert I. Jennrich; Mark D. Schluchter
Biometrics, Vol. 42, No. 4. (Dec., 1986), pp. 805-820.
Stable URL:
http://links.jstor.org/sici?sici=0006-341X%28198612%2942%3A4%3C805%3AURMWSC%3E2.0.CO%3B2-L

Robust Statistical Modeling Using the t Distribution
Kenneth L. Lange; Roderick J. A. Little; Jeremy M. G. Taylor
Journal of the American Statistical Association, Vol. 84, No. 408. (Dec., 1989), pp. 881-896.
Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28198912%2984%3A408%3C881%3ARSMUTD%3E2.0.CO%3B2-U

LINKED CITATIONS

Bayesian Marginal Inference

Tom Leonard; John S. J. Hsu; Kam-Wah Tsui

Journal of the American Statistical Association, Vol. 84, No. 408. (Dec., 1989), pp. 1051-1058.

Stable URL:

http://links.jstor.org/sici?sici=0162-1459%28198912%2984%3A408%3C1051%3ABMI%3E2.0.CO%3B2-C

Newton-Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data

Mary J. Lindstrom; Douglas M. Bates

Journal of the American Statistical Association, Vol. 83, No. 404. (Dec., 1988), pp. 1014-1022.

Stable URL:

http://links.jstor.org/sici?sici=0162-1459%28198812%2983%3A404%3C1014%3ANAEAFL%3E2.0.CO%3B2-J

A Fast Scoring Algorithm for Maximum Likelihood Estimation in Unbalanced Mixed Models with Nested Random Effects

Nicholas T. Longford

Biometrika, Vol. 74, No. 4. (Dec., 1987), pp. 817-827.

Stable URL:

http://links.jstor.org/sici?sici=0006-3444%28198712%2974%3A4%3C817%3AAFSAFM%3E2.0.CO%3B2-A

Matrix Derivatives with an Application to an Adaptive Linear Decision Problem

Elizabeth Chase MacRae

The Annals of Statistics, Vol. 2, No. 2. (Mar., 1974), pp. 337-346.

Stable URL:

http://links.jstor.org/sici?sici=0090-5364%28197403%292%3A2%3C337%3AMDWAAT%3E2.0.CO%3B2-M

The Commutation Matrix: Some Properties and Applications

Jan R. Magnus; H. Neudecker

The Annals of Statistics, Vol. 7, No. 2. (Mar., 1979), pp. 381-394.

Stable URL:

http://links.jstor.org/sici?sici=0090-5364%28197903%297%3A2%3C381%3ATCMSPA%3E2.0.CO%3B2-Y

Symmetric Matrix Derivatives with Applications

Charles E. McCulloch

Journal of the American Statistical Association, Vol. 77, No. 379. (Sep., 1982), pp. 679-682.

Stable URL:

http://links.jstor.org/sici?sici=0162-1459%28198209%2977%3A379%3C679%3ASMDWA%3E2.0.CO%3B2-Q

LINKED CITATIONS

A Generalized Multivariate Analysis of Variance Model Useful Especially for Growth Curve Problems

Richard F. Potthoff; S. N. Roy

Biometrika, Vol. 51, No. 3/4. (Dec., 1964), pp. 313-326.

Stable URL:

http://links.jstor.org/sici?sici=0006-3444%28196412%2951%3A3%2F4%3C313%3AAGMAOV%3E2.0.CO%3B2-6

Educational Applications of Hierarchical Linear Models: A Review

Stephen W. Raudenbush

Journal of Educational Statistics, Vol. 13, No. 2. (Summer, 1988), pp. 85-116.

Stable URL:

http://links.jstor.org/sici?sici=0362-9791%28198822%2913%3A2%3C85%3AEAOHLM%3E2.0.CO%3B2-T

A Hierarchical Model for Studying School Effects

Stephen Raudenbush; Anthony S. Bryk

Sociology of Education, Vol. 59, No. 1. (Jan., 1986), pp. 1-17.

Stable URL:

http://links.jstor.org/sici?sici=0038-0407%28198601%2959%3A1%3C1%3AAHMFSS%3E2.0.CO%3B2-1

A Multilevel, Multivariate Model for Studying School Climate with Estimation Via the EM Algorithm and Application to U. S. High-School Data

Stephen W. Raudenbush; Brian Rowan; Sang Jin Kang

Journal of Educational Statistics, Vol. 16, No. 4. (Winter, 1991), pp. 295-330.

Stable URL:

http://links.jstor.org/sici?sici=0362-9791%28199124%2916%3A4%3C295%3AAMMMFS%3E2.0.CO%3B2-Q

A Note on Bayesian, Likelihood, and Sampling Distribution Inferences

Donald B. Rubin

Journal of Educational Statistics, Vol. 3, No. 2. (Summer, 1978), pp. 189-201.

Stable URL:

http://links.jstor.org/sici?sici=0362-9791%28197822%293%3A2%3C189%3AANOBLA%3E2.0.CO%3B2-O

Sensitivity Analysis for Fixed Effects in the Hierarchical Model: A Gibbs Sampling Approach

Michael H. Seltzer

Journal of Educational Statistics, Vol. 18, No. 3. (Autumn, 1993), pp. 207-235.

Stable URL:

http://links.jstor.org/sici?sici=0362-9791%28199323%2918%3A3%3C207%3ASAFFEI%3E2.0.CO%3B2-D