Title
Pattern-Mixture Models for Multivariate Incomplete Data

Author
Roderick J. A. Little

Publication Date
2011-10-24

Pattern-Mixture Models for Multivariate Incomplete Data

RODERICK J. A. LITTLE*

Consider a random sample on variables $X_1, \ldots, X_V$ with some values of X_V missing. *Selection models* specify the distribution of $X_1, \ldots, X_V$ over respondents and nonrespondents to X_V, and the conditional distribution that X_V is missing given $X_1, \ldots, X_V$. In contrast, *pattern-mixture models* specify the conditional distribution of $X_1, \ldots, X_V$ given that X_V is observed or missing respectively and the marginal distribution of the binary indicator for whether or not X_V is missing. For multivariate data with a general pattern of missing values, the literature has tended to adopt the selection-modeling approach (see for example Little and Rubin); here, pattern-mixture models are proposed for this more general problem. Pattern-mixture models are chronically underidentified; in particular for the case of univariate nonresponse mentioned above, there are no data on the distribution of X_V given $X_1, \ldots, X_{V-1}$ in the stratum with X_V missing. Thus the models require restrictions or prior information to identify the parameters. *Complete-case restrictions* tie unidentified parameters to their (identified) analogs in the stratum of complete cases. Alternative types of restriction tie unidentified parameters to parameters in other missing-value patterns or sets of such patterns. This large set of possible identifying restrictions yields a rich class of missing-data models. Unlike ignorable selection models, which generally requires iterative methods except for special missing-data patterns, some pattern-mixture models yield explicit ML estimates for general patterns. Such models are readily amenable to Bayesian methods and form a convenient basis for multiple imputation. Some previously considered noniterative estimation methods are shown to be maximum likelihood (ML) under a pattern-mixture model. For example, Buck's method for continuous data, corrected as in Beale and Little (1975), and Brown's estimators for nonrandomly missing data are ML for pattern-mixture models with particular complete-case restrictions. Available-case analyses, where the mean and variance of X_j are computed using all cases with X_j observed and the correlation (or covariance) of X_j and X_k is computed using all cases with X_j and X_k observed, are also close to ML for another pattern-mixture model. Asymptotic theory for this class of estimators is outlined.

KEY WORDS: EM algorithm; Imputation; Maximum likelihood; Missing values; Monotone missing data; Multiple imputation; Nonresponse.

1. INTRODUCTION

1.1 The Problem

I consider the analysis of an $(n \times V)$ matrix $\mathbf{X} = \{x_{ij}\}$ from a random sample of n observations on V variables $X_1, \ldots, X_V$, where x_{ij} is the value of variable X_j for case i. A standard statistical analysis is contemplated, such as summarizing the joint distribution of $X_1, \ldots, X_V$ by the sample mean and covariance matrix, computing the regression of one variable on the others, or fitting a log-linear model to the contingency table formed by categorical X_j's. The objective here is to develop a corresponding analysis when some components of X are missing.

Most statistical packages discard incomplete cases and carry out a *complete-case* (CC) *analysis* using only the cases with $X_1, \ldots, X_V$ all observed. This approach is simple but clearly inefficient and generally requires the strong assumption that the complete cases are a random subsample of all cases. The strategy followed in this article is to formulate a statistical model for the joint distribution of X and the missing data mechanism and then to estimate parameters of the model by ML or Bayesian methods. Rubin (1976) formalized models for the missing-data mechanism by introducing the stochastic $(n \times V)$ *missing-data indicator matrix* $\mathbf{M}$, with entries $m_{ij} = 0$ if x_{ij} is observed and $m_{ij} = 1$ if x_{ij} is missing. A full parametric model then specifies the joint distribution of X and M indexed by a set of unknown parameters. (Fully observed covariates can be treated as fixed and conditioned in the specification of the model distributions; such conditioning is suppressed here to keep notation as simple as possible.)

1.2 Two Classes of Models

Two ways of specifying the joint distribution of X and M can be contrasted (Little and Rubin 1987, chap. 10). *Selection models* specify

$$p(X, M|\theta, \psi) = p(X|\theta)p(M|X, \psi), \tag{1}$$

where $p(X|\theta)$ represents the complete-data model for X, $p(M|X, \psi)$ represents the model for the missing-data mechanism, and (θ, ψ) are unknown parameters. *Pattern-mixture models* specify

$$p(X, M|\varphi, \pi) = p(X|M, \varphi)p(M|\pi), \tag{2}$$

where the distribution of X is conditioned on the missing data pattern M. The term "pattern-mixture" reflects the fact that the resulting marginal distribution of X is a mixture of distributions. Glynn, Laird, and Rubin (1986) and Rubin (1987) used the term "mixture" for models of this kind; "pattern" is added to make the nature of the mixing more explicit.

Equations (1) and (2) are simply two different ways of factoring the joint distribution of X and M. When M is independent of X, Rubin calls the data *missing completely at random* (MCAR), and the two specifications are equivalent when $\theta = \varphi$ and $\psi = \pi$. When the missing data are not MCAR and distributional assumptions are added, (1) and

* Roderick J. A. Little is Professor, Department of Biomathematics, UCLA School of Medicine, Los Angeles, CA 90024. This research was supported by National Institute of Mental Health Grant MH37188. The author gratefully acknowledges many useful comments from Michael Pearlman, Don Rubin, Yong-Xiao Wang, Robert Weiss, and an anonymous associate editor and referee.

(2) can yield different models, as shown by the following basic example.

Example 1. Monotone Normal Data with Two Variables. Suppose in a panel study with two time points, $V = 2$, X_j is a variable measured at time j ($j = 1, 2$), X_1 is fully observed, and X_2 is sometimes missing due to attrition. There are then two missing-data patterns, complete cases, for which the ith row of M is $(m_{i1}, m_{i2}) = (0, 0)$, and cases with X_2 missing, for which $(m_{i1}, m_{i2}) = (0, 1)$. A well-known selection model for these data assumes that (a) (x_{i1}, x_{i2}) are bivariate normal with mean and covariance matrix $\theta = (\mu, \Sigma)$, and (b) m_{i2} given (x_{i1}, x_{i2}) is Bernoulli with probability

$$p(m_{i2} = 1 \mid x_{i1}, x_{i2}, \psi) = g(\psi_0 + \psi_1 x_{i1} + \psi_2 x_{i2}), \quad (3)$$

for some function g taking values in the range (0, 1). On the other hand, the normal pattern-mixture model assumes that (a) given $m_{i2} = r$, (x_{i1}, x_{i2}) is bivariate normal with mean and covariance matrix $\varphi^{(r)} = \{\mu^{(r)}, \Sigma^{(r)}\}$, and $r = 0, 1$; and (b) m_{i2} is marginally Bernoulli with $pr(m_{i2} = 1) = \pi$. This model implies that marginally (X_1, X_2) is a mixture of two normal distributions, rather than normal as in the selection model. The two models are equivalent when the data are MCAR, because then missingness of X_2 is independent of X_1 and X_2, $\psi_1 = \psi_2 = 0$, $g(\psi_0) = \pi$ and $\mu^{(r)} = \mu$, $\Sigma^{(r)} = \Sigma$ for $r = 0, 1$.

Many ML missing-data methods in the literature are based on *ignorable* selection models of the form (1), where (a) θ and ψ are distinct parameters and (b) the data are *missing at random* (MAR); that is,

$$p(M \mid X, \psi) = p(M \mid X_{obs}, \psi), \quad (4)$$

where X_{obs} denotes the set of observed components of $\{x_{ij}\}$ in the data matrix $\mathbf{X}$. Thus the conditional distribution of M given X does not depend on unobserved (missing) components of X. Rubin (1976) showed that ML inference for θ under such models does not depend on $p(M \mid X, \psi)$ and can be based on the likelihood obtained by integrating missing values in X out of the density $p(X \mid \theta)$ in (1). The selection model in Example 1 is ignorable if θ and ψ are distinct and $\psi_2 = 0$, because then missingness of X_2 depends only on the values of X_1, which are always observed. Note that the MAR condition (4) is less restrictive than MCAR.

Pattern-mixture models are often underidentified. In particular, for the pattern-mixture model in Example 1, the data supply no information about the parameters of the distribution of X_2 given X_1 for the pattern with X_2 missing. Rubin (1977) addressed this problem by introducing Bayesian prior distributions relating these parameters to corresponding parameters in the CC stratum. A degenerate form of prior imposes *identifying restrictions* on the stratum parameters. In Example 1, this approach might lead us to identify the parameters of the distribution of X_2 and X_1 in P_0 and P_1, yielding standard analyses. But for more general missing-data patterns, the class of potential identifying restrictions is much richer, yielding a variety of interesting estimation methods.

1.3 Choice of Target Parameters

If the data are not MCAR, then the parameters in selection and pattern-mixture models have a different interpretation; in particular, θ concerns the marginal distribution of X, whereas φ concerns the conditional distribution of X given M. In certain contexts the parameters of the conditional distribution can be of substantive interest. For example, in a panel survey it may be of interest to compare demographic characteristics of movers ($m_{ij} = 0$) and stayers ($m_{ij} = 1$). In surveys where a missing value arises because a question was inapplicable, it may make sense to treat nonresponse as defining a category of the population, in which case parameters that are conditional on the lack of a response (Little and Rubin 1987, chap. 1) are of obvious interest. Pattern-mixture models clearly are a natural way of modeling such situations.

Even so, it is more frequently the case that parameters of the marginal distribution of X are the main focus of interest. Given this emphasis, selection modeling may seem more natural, but pattern-mixture models are not precluded. When a pattern-mixture model is adopted, the parameters of the marginal distribution of X can be expressed as functions of φ and π, and inferences can be obtained for these parametric functions. In particular, ML estimates are derived by substituting ML estimates for φ and π in these functions.

Continuous distributions are summarized here by the mean and covariance matrix of X. A referee has noted that these are not necessarily appropriate summaries for the mixtures of normal distributions implied by normal pattern-mixture models, such as that in Example 1. This is true if the patterns define well-separated normal distributions, a situation that could be spotted to some degree through inspection of the empirical marginal distributions of the variables. If bimodality of the mixture distribution is absent or minor, the marginal mean and covariance matrix are useful summaries.

1.4 Pattern-Set Mixture Models

The ideas of selection modeling and pattern-mixture modeling can be combined. The following extension is useful in Section 2.4.3. Divide the set of patterns into two subsets, labelled $s = 1$ and $s = 2$, and let X_s, M_s denote the components of X, M for subset s ($s = 1, 2$). Then a selection model,

$$p(X_1, M_1 \mid s = 1, \theta, \psi) = p(X_1 \mid s = 1, \theta) p(M_1 \mid X_1, s = 1, \psi), \quad (5)$$

can be specified for subset 1, and a pattern-mixture model,

$$p(X_2, M_2 \mid s = 2, \varphi, \pi) = p(X_2 \mid M_2, s = 2, \varphi) p(M_2 \mid s = 2, \pi), \quad (6)$$

can be specified for subset 2. Equations (5) and (6), combined with the marginal distribution of s, define a pattern-set mixture model, because the mixture involves a set of patterns with $s = 1$. More complex pattern-set mixture models can be envisaged, but will not be considered here.

1.5 What This Article is About

Previous discussions of pattern-mixture models have been largely confined to the case of univariate nonresponse. Rubin (1978) mentioned the possibility of using them in more general situations, but gave no examples. This article explores

pattern-mixture models and pattern-set mixture models for a general pattern of missing data. Section 2 considers the simple but important special case of incomplete bivariate data. Section 3 outlines extensions to a general pattern of missing data, and Section 4 compares pattern-mixture models with ignorable selection models and makes some concluding remarks.

2. INCOMPLETE BIVARIATE DATA

2.1 Pattern-Mixture Models for Bivariate Data

I first consider bivariate data; that is, $V = 2$. Let $m_i = (m_{i1}, m_{i2})$ denote the missing-data pattern for case i, with possible values (0, 0), (0, 1), (1, 0), and (1, 1). It simplifies the notation to use integer labels for these patterns; hence define $r_i = r(m_i)$ so that $r\{(0, 0)\} = 0$, $r\{(1, 0)\} = 1$, $r\{(0, 1)\} = 2$, and $r\{(1, 1)\} = 3$. Here and more generally, 0 indexes the complete-case pattern. For pattern r, $0 \le r \le 3$, let P_r denote the set of sample cases (see Fig. 1), let n_r denote the number of cases in the sample, and let $n = \Sigma_r n_r$. The *saturated pattern-mixture model* for bivariate data is defined as follows:

(a) $r_i = r(m_i)$ is iid multinomial with index 1 and probabilities

$$p(r_i = r) = \pi_r;\ \pi_0 + \pi_1 + \pi_2 + \pi_3 = 1.$$

(b) $x_i = (x_{i1}, x_{i2})$ given $r_i = r$ are independent with density $p(x_{i1}, x_{i2} | r_i = r, \varphi^{(r)})$.

(c) The parameters $\pi = \{\pi_r\}$ and $\varphi = \{\varphi^{(r)}\}$ are distinct.

This model is underidentified. Specifically, for pattern r let $\varphi_u^{(r)}$ and $\varphi_{v\cdot u}^{(r)}$ denote the parameters of the marginal distribution of X_u and the conditional distribution of X_v given X_u and $u = 1$, $v = 2$ or $u = 2$, $v = 1$. The likelihood has the form

$$L(\varphi, \pi | M, X_{obs}) = \prod_{r=0}^{3} \pi_r^{n_r} \prod_{i \in P_0} p(x_{i1}, x_{i2} | r_i = 0, \varphi^{(0)}) \times \prod_{i \in P_1} p(x_{i1} | r_i = 1, \varphi_1^{(1)}) \prod_{i \in P_2} p(x_{i2} | r_i = 2, \varphi_2^{(2)}). \quad (7)$$

Restrictions or prior information are needed to identify $\varphi_{2\cdot 1}^{(1)}$, $\varphi_{1\cdot 2}^{(2)}$, and $\varphi^{(3)}$, because these parameters do not appear in the likelihood. Consider, for example, the following extension of Example 1.

Example 2: Bivariate Normal Pattern-Mixture Model. Suppose that X_1 and X_2 are bivariate normal for each pattern and $\varphi^{(r)} = (\mu_1^{(r)}, \mu_2^{(r)}, \sigma_{11}^{(r)}, \sigma_{22}^{(r)}, \sigma_{12}^{(r)})$ denotes the means, variances, and covariance of X_1 and X_2 for pattern r. The dimension of φ is thus $5 \times 4 = 20$. Eleven of these parameters are not identified, namely the three parameters of the regression of X_2 on X_1 for P_1, the three parameters of the regression of X_1 on X_2 for P_2, and the five parameters of the joint distribution of X_1 and X_2 for P_3.

From the standpoint of ML estimation of parameters of the marginal distribution of X, it is only necessary to identify parameters of patterns r that are observed ($n_r > 0$). In particular, for Example 2, let $\theta = (\mu, \Sigma)$, $\Sigma = \{\sigma_{jk}\}$ denote the mean and covariance matrix of the marginal distribution of X. Then $\mu_j = \Sigma_r \pi_r \mu_j^{(r)}$, $\sigma_{jk} = \Sigma_r \pi_r \{\sigma_{jk}^{(r)} + (\mu_j^{(r)} - \mu_j)(\mu_k^{(r)} - \mu_k)\}$ $(1 \le j, k \le 2)$. The ML estimates are thus $\hat{\mu}_j = \Sigma_r \hat{\pi}_r \hat{\mu}_j^{(r)} = \Sigma_{r:n_r>0} \hat{\pi}_r \hat{\mu}_j^{(r)}$, $\hat{\sigma}_{jk} = \Sigma_{r:n_r>0} \hat{\pi}_r \{\hat{\sigma}_{jk}^{(r)} + (\hat{\mu}_j^{(r)} - \hat{\mu}_j)(\hat{\mu}_k^{(r)} - \hat{\mu}_k)\}$. Because $\hat{\pi}_r = n_r/n = 0$ when $n_r = 0$, patterns not in the sample receive no weight and can be ignored.

The choice of identifying restrictions for observed patterns should reflect contextual knowledge of the nature of the missing-data mechanisms. Even so, restrictions that yield explicit ML estimates of φ are of practical interest. Sufficient (but not necessary) conditions for this to happen are that (a) the estimable parameters in (7) are distinct for each pattern; (b) ML estimation for the estimable parameters in (7), for each pattern taken separately, is noniterative; (c) the restrictions just identify (that is, do not overidentify) the model; and (d) resulting estimates of φ and π, and functions of these estimates, lie within the parameter space. Under these conditions, the three products in the likelihood (7) can be maximized separately as three complete-data problems: $\hat{\varphi}^{(0)}$ from the data in P_0, $\hat{\varphi}_1^{(1)}$ from the data in P_1, and $\hat{\varphi}_2^{(2)}$ from the data in P_2. ML estimates of other parameters follow by using the restrictions to express them as functions of $\varphi^{(0)}$, $\varphi_1^{(1)}$, and $\varphi_2^{(2)}$ and π, and then substituting ML estimates of these parameters.

The MCAR assumption that X and M are independent yields a strong set of identifying restrictions, namely

$$\varphi^{(0)} = \varphi^{(1)} = \varphi^{(2)} = \varphi^{(3)} = \theta; \quad (8)$$

that is, the φ's all equal the parameter of the marginal distribution of X. The pattern-mixture model with restrictions (8) reduces to an ignorable selection model. ML for this

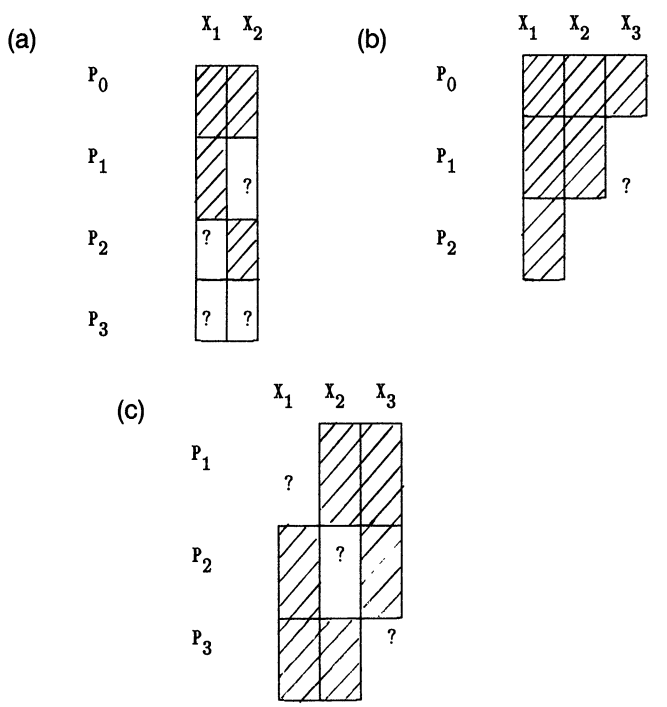

Figure 1. Selected Missing-Data Patterns For Two and Three Variables: (a) bivariate general pattern; (b) trivariate monotone pattern; (c) trivariate, bivariate response.

model is in general iterative, because the restrictions (8) overidentify the pattern-mixture model, hence violating condition (c). My interest here is with weaker (and hence potentially more plausible) alternative assumptions to (8).

Definition 1: Restrictions on The Missing-Variable Distribution. The joint distribution of X for each pattern is the product of the marginal distribution of the observed variables and the conditional distribution of the missing variables given the observed variables; I call the latter the *missing-variable (MV) distribution* for that pattern; identifying restrictions on the parameters of MV distributions are called *missing-variable (MV) restrictions.*

2.2 Complete-Case Missing-Variable Restrictions

The MV distribution for a pattern P_r is said to be *equated* to a pattern P_s if it is assumed to equal the corresponding distribution defined for pattern P_s. *Complete-case missing-variable (CCMV) restrictions* equate all MV distributions to P_0: hence for bivariate data,

$$\varphi_{2\cdot 1}^{(1)} = \varphi_{2\cdot 1}^{(0)}, \qquad \varphi_{1\cdot 2}^{(2)} = \varphi_{1\cdot 2}^{(0)}, \qquad \varphi^{(3)} = \varphi^{(0)}. \quad (9)$$

This set of restrictions is weaker than (8); we now apply them to Example 2.

Example 3 (Example 2 continued): Normal Pattern-Mixture Model with CCMV Restrictions. For the model of Example 2, $\hat{\varphi}^{(0)}$ is the sample mean and covariance matrix of X_1 and X_2 for cases in P_0, $\hat{\varphi}_1^{(1)}$ is the sample mean and variance of X_1 for cases in P_1, and $\hat{\varphi}_2^{(2)}$ is the sample mean and variance of X_2 for cases in P_2. The mean of X_1 is

$$\mu_1 = \sum_{r=0}^{3} \pi_r \mu_1^{(r)},$$

as noted previously. Applying (9) to unidentified parameters in this sum yields

$$\mu_1^{(3)} = \mu_1^{(0)}; \qquad \mu_1^{(2)} = \beta_{10\cdot 2}^{(2)} + \beta_{12\cdot 2}^{(2)}\mu_2^{(2)} = \beta_{10\cdot 2}^{(0)} + \beta_{12\cdot 2}^{(0)}\mu_2^{(2)},$$

where $\beta_{10\cdot 2}^{(r)}$ and $\beta_{12\cdot 2}^{(r)}$ are the intercept and coefficient of X_2 in the regression of X_1 on X_2 for pattern r. Substituting ML estimates yields

$$\hat{\mu}_1 = (\hat{\pi}_0 + \hat{\pi}_3)\hat{\mu}_1^{(0)} + \hat{\pi}_1\hat{\mu}_1^{(1)} + \hat{\pi}_2(\hat{\beta}_{10\cdot 2}^{(0)} + \hat{\beta}_{12\cdot 2}^{(0)}\hat{\mu}_2^{(2)}), \quad (10)$$

where $\hat{\pi}_r$ is the proportion of the sample with pattern r, $\hat{\mu}_j^{(r)}$ denotes the sample mean of X_j for cases in P_r, and $\hat{\beta}_{10\cdot 2}^{(0)}$ and $\hat{\beta}_{12\cdot 2}^{(0)}$ are the intercept and slope from the least squares regression of X_1 on X_2, computed using cases in P_0.

For pattern r, let $\sigma_{uu}^{(r)}$ denote the variance of X_u and let $\sigma_{vv\cdot u}^{(r)}$ denote the residual variance of X_v given X_u. Then the ML estimate of the variance of X_1 is

$$\hat{\sigma}_{11} = \sum_{r=0}^{3} \hat{\pi}_r\hat{\sigma}_{11}^{(r)} + \sum_{r=0}^{3} \hat{\pi}_r(\hat{\mu}_1^{(r)} - \hat{\mu}_1)^2, \quad (11)$$

where $\hat{\sigma}_{11}^{(0)}$ and $\hat{\sigma}_{11}^{(1)}$ are the sample variances of X_1 for cases in P_0 and P_1, $\hat{\sigma}_{11}^{(2)} = \hat{\sigma}_{11\cdot 2}^{(0)} + [\hat{\beta}_{12\cdot 2}^{(0)}]^2\hat{\sigma}_{22}^{(2)}$, and $\hat{\sigma}_{11}^{(3)} = \hat{\sigma}_{11}^{(0)}$. The ML estimate of the covariance of X_1 and X_2 is

$$\hat{\sigma}_{12} = \sum_{r=0}^{3} \hat{\pi}_r\hat{\sigma}_{12}^{(r)}, \quad (12)$$

where $\hat{\sigma}_{12}^{(0)}$ is the sample covariance for cases in P_0, $\hat{\sigma}_{12}^{(1)} = \hat{\beta}_{21\cdot 1}^{(0)}\hat{\sigma}_{11}^{(1)}$, $\hat{\sigma}_{12}^{(2)} = \hat{\beta}_{12\cdot 2}^{(0)}\hat{\sigma}_{22}^{(2)}$, and $\hat{\sigma}_{12}^{(3)} = \hat{\sigma}_{12}^{(0)}$. Corresponding expressions for $\hat{\mu}_2$ and $\hat{\sigma}_{22}$ are obtained by symmetry.

Because the pattern-mixture model with restrictions (8) is a submodel of the model with restrictions (9), estimates from the latter are not fully efficient if the data are in fact MCAR. The loss of efficiency is examined in the next example.

Example 4 (Example 3 Continued): Normal Pattern-Mixture Model with CCMV Restrictions: Efficiency of Estimates Under MCAR. Consider the asymptotic variances of estimates of μ_1 for the model of Example 2 with restrictions (8) or (9), when $n_3 = 0$ and the data are in fact MCAR. The asymptotic variance of the estimate (10) can be shown to be $\sigma_{11}/\hat{n}_1$, where

$$\hat{n}_1 = \frac{n\pi_0}{\pi_0 + \pi_2(\pi_0 + \pi_2)(1 - \rho^2)}$$

is the effective sample size and $\rho = \mathrm{Corr}(X_1, X_2)$. The asymptotic variance of the ML estimate of μ_1 under the MCAR model is obtained by calculating and inverting the information matrix, and equals $\sigma_{11}/\tilde{n}_1$, where

$$\tilde{n}_1 = n\left(1 - \pi_2 + \frac{\pi_0\pi_2\rho^2}{\pi_0 + \pi_2(1 - \rho^2)}\right).$$

Hence the proportional loss in effective sample size from using (10) when the data are MCAR is

$$(\tilde{n}_1 - \hat{n}_1)/\tilde{n}_1 = \frac{\pi_2^2\pi_1(1 - \pi_1)(1 - \rho^2)^2}{\pi_0^2 + \pi_2\pi_0(1 - \rho^2) + \pi_2^2\pi_1(1 - \pi_1)(1 - \rho^2)^2}, \quad (13)$$

after some algebra. This loss is 0 when π_1 or π_2 is 0, and it also tends to 0 as ρ^2 tends to 1 and the value of using data in P_2 increases. When $\pi_0 = \pi_1 = \pi_2 = \frac{1}{3}$, the loss is 18.2% when $\rho^2 = 0$, 3.6% when $\rho^2 = .5$ and .7% when $\rho^2 = .8$. These losses can be compared to the loss in effective sample size from estimating μ_1 from its sample mean over cases in P_0 and P_1, which is the ML estimate under the independence restriction $\rho = 0$. This loss is 0 when $\rho^2 = 0$, 14.3% when $\rho^2 = .5$, 25% when $\rho^2 = .8$, and 33.3% when $\rho^2 = 1$ (and the dropped cases in P_2 have asymptotically the same information about μ_1 as do the cases in P_0 and P_1).

A referee notes that when data are in fact MCAR, ML for the pattern-mixture model with CCMV restrictions can be seen as avoiding iteration by fitting a larger model; efficiency in such settings can be increased by applying the adjustment method of Cox and Wermuth (1990). This is an interesting idea that merits consideration. When data are not MCAR, the method is ML under (9), and its value rests on the plausibility of that identifying restriction, a question considered further in Section 4.

The fact that the loss of information (13) is 0 when both π_3 and either π_1 or π_2 equal 0 is no accident, because in those cases the ML estimates of θ under (8) or (9) are identical; (10)–(12) can be shown to be the ML estimates under (8) by factoring the likelihood (Anderson 1957; Little and Rubin 1987, chap. 6; Rubin 1974). When $n_3 = 0$ and n_1 and n_2 are both non-zero, ML estimation under (8) requires it-

eration, and the EM algorithm (Dempster, Laird, and Rubin 1977) can be applied. The estimates (10)–(12) are then the output from the first iteration of EM, with starting values computed from complete cases.

The next example provides a pattern-mixture model for bivariate categorical data.

Example 5. A Multinomial Pattern-Mixture Model for Contingency Tables with Supplemental Margins. For categorical X_1 and X_2 with J and K levels, data in P_0 in Figure 1a can be arranged as a $(J \times K)$ contingency table, and the data in patterns P_1 and P_2 form supplemental $(J \times 1)$ and $(1 \times K)$ margins. Table 1 displays two such data sets with $J = K = 2$, analyzed in Little (1982, 1985). A basic pattern-mixture model assumes that for pattern r, X_1 and X_2 are multinomial with probabilities $\varphi^{(r)} = \{\varphi_{jk}^{(r)}\}$, where

$$\varphi_{jk}^{(r)} = pr(X_1 = j, X_2 = k \mid R = r).$$

Let + denote summation over a subscript. Taking into account the constraints $\varphi_{++}^{(r)} = 1$ for all r, this model has $4(JK - 1)$ distinct parameters.

Let $\theta = \{\theta_{jk}\}$, $\theta_{jk} = pr(X_1 = j, X_2 = k)$ denote the probabilities of the marginal distribution of X_1 and X_2, summed over patterns. Then $\theta_{jk} = \Sigma_r \pi_r \varphi_{jk}^{(r)}$. ML estimates of θ can be found by finding the ML estimates of π and φ and substituting them in these expressions.

With CCMV restrictions (9), $\hat{\varphi}_{jk}^{(0)}$ are simply the sample proportions for the complete cases, $\hat{\varphi}_{j+}^{(1)}$ are the sample proportions for the supplemental P_1 margin, and $\hat{\varphi}_{+k}^{(2)}$ are the sample proportions for the supplemental P_2 margin. ML estimates of θ are obtained by allocating the supplemental counts into the two-way table using conditional probabilities determined from $\{\hat{\varphi}_{jk}^{(0)}\}$. Table 2 illustrates the procedure for the data in Table 1a.

As in the case of the normal model, this procedure is the first iteration of the EM algorithm for ML estimation under MCAR restrictions (8) (Chen and Fienberg 1974). The E step allocates the supplemental margins into the table using conditional probabilities based on the current estimates of θ, and the M step re-estimates θ by the sample proportions from the filled-in data. The E step and M step are then repeated until convergence. ML estimates under (8) and under (9) are displayed in Table 3 for the two data sets in Table 1; they can be compared with the estimates from the complete cases only (row 1). For Table 1a with a moderate amount of missing data, the two sets of ML estimates are close, reflecting the fact that both make use of all the data. For Table 1b the fraction of incomplete cases is very large and the supplemental margin on X_2 differs greatly from the corresponding CC margin. The three sets of estimates are quite different, reflecting sensitivity to the choice of model.

Table 1. Two (2 x 2) Tables with Supplemental Margins on Both Variables

a)

P_0: X_1 and X_2 observed

X_1 \ X_2	1	2	All
1	100	50	150
2	75	75	150
All	175	125	300

P_1: X_1 only observed

X_1	
1	30
2	60
All	90

P_2: X_2 only observed

1	2	All
28	60	88

b)

P_0: X_1 and X_2 observed

X_1 \ X_2	1	2	All
1	35	10	45
2	5	50	55
All	40	60	100

P_1: X_1 only observed

X_1	
1	225
2	220
All	445

P_2: X_2 only observed

1	2	All
48	407	455

Table 2. ML Estimation for Multinomial Pattern-Mixture Model with Complete-Case Restrictions, Applied to Data in Table 1a

a) Estimates $\{\hat{\varphi}_{jk}^{(0)}\}$ of Cell Probabilities for P_0:

X_1 \ X_2	1	2	All
1	100/300 = .3333	50/300 = .1667	.5
2	75/300 = .25	75/300 = .25	.5
All	.5833	.4167	

b) Estimates $\{\hat{\varphi}_{jk}^{(1)}\}$ of Cell Probabilities for P_1, from Allocating P_1-margin Into Table Using Conditional Probabilities Based on a):

X_1 \ X_2	1	2	All
1	$\frac{(30)(.3333)}{.5} = 20$	$\frac{(30)(.1667)}{.5} = 10$	30
2	$\frac{(60)(.25)}{.5} = 30$	$\frac{(60)(.25)}{.5} = 30$	60
All			90

c) Estimates $\{\hat{\varphi}_{jk}^{(2)}\}$ of Cell Probabilities for P_2, from Allocating P_2-margin Into Table Using Conditional Probabilities Based on a):

X_1 \ X_2	1	2	All
1	$\frac{(28)(.3333)}{.5833} = 16$	$\frac{(60)(.1667)}{.4167} = 24$	
2	$\frac{(28)(.25)}{.5833} = 12$	$\frac{(60)(.25)}{.4167} = 36$	
All	28	60	88

d) Estimates $\{\hat{\theta}_{jk}\}$ of Marginal Probabilities, Aggregating Over Patterns:

$\hat{\theta}_{11} = (100 + 20 + 16)/478 = .285$; $\hat{\theta}_{12} = (50 + 10 + 24)/478 = .176$;

$\hat{\theta}_{21} = (75 + 30 + 12)/478 = .245$; $\hat{\theta}_{22} = (75 + 30 + 36)/478 = .295$;

Table 3. Estimates of Cell Probabilities (×1,000) for Tables 1a and 1b, for Three Methods

Method	Cell probabilities θ_{11}	θ_{12}	θ_{21}	θ_{22}
Table 1a				
(a)				
CC Analysis	333	167	250	250
Pattern-Mixture, CC Restrictions	285	176	245	295
Ignorable Selection Model	279	174	239	308
Table 1b				
(b)				
CC Analysis	350	100	50	500
Pattern-Mixture, CC Restrictions	252	128	31	589
Ignorable Selection Model	150	319	17	514

2.3 Pattern-Mixture Models With Other CC Restrictions

It seems natural to place identifying restrictions on the parameters of MV distributions, but restrictions on other parameters also are possible, as in the following example.

Example 6: Monotone Data with Two Variables (Continued): Brown's Estimators. Consider bivariate normal data as in Example 1, so the pattern is Figure 1a with $n_2 = n_3 = 0$. Suppose that

$$pr(X_2 \text{ missing} | X_1, X_2) = g(X_2); \tag{14}$$

that is, missingness of X_2 depends on the value of X_2. It is easily seen that (14) implies that the distribution of X_1 given X_2 is the same for P_0 and P_1. Hence this mechanism suggests a normal pattern-mixture model with restrictions

$$\varphi_{1\cdot 2}^{(1)} = \varphi_{1\cdot 2}^{(0)},$$

rather than the MVCC restrictions $\varphi_{2\cdot 1}^{(1)} = \varphi_{2\cdot 1}^{(0)}$. ML estimates under this model can be shown to be equivalent to the "protective" estimators proposed by Brown (1990, eq. 4.3). Detailed properties of this interesting model and variants will be developed in a future article.

2.4 Alternatives to Complete-Case Restrictions

2.4.1 Limitations of Complete-Case Restrictions. CC restrictions (definition 2) equate all distributions to P_0. As a practical matter, this is unappealing if the number of complete cases is small—for the normal model, for example, at least three complete cases are needed to yield consistent estimates of μ and Σ. On a modeling level, it may be more reasonable to equate the MV distribution for an incomplete pattern to patterns other than P_0.

In particular, for bivariate data the MV distribution for the pattern P_3 with both X_1 and X_2 missing is simply the joint distribution of X_1 and X_2. CC restrictions equate this distribution to the joint distribution of X_1 and X_2 in P_0. But we might expect cases in P_3 to be more like cases in P_1 or P_2 than like cases in P_0. For example, suppose that X_j represents a set of variables from a panel survey at time j, and the sample frame consists of a list of names and addresses. If a primary reason for missing data is the sample person moving from the listed address to an undetermined or inaccessible location, then cases in P_1, P_2, and P_3 are "movers" and cases in P_0 are "stayers". If mover/stayer status is related to the survey variables, then it makes more sense to equate P_3 to P_1 and P_2 rather than to P_0. Of course, P_1 and P_2 are incomplete, and hence their MV distributions need to be equated to P_0; however, equating P_3 to P_1 and P_2 uses the information in P_0 in a more indirect way.

The usual treatment of P_3 is simply to drop these cases from the analysis. This does *not* correspond to CC restriction; rather, it implies that respondents ($r = 0$, 1, or 2) and nonrespondents ($r = 3$) have the same distribution of X_1 and X_2; that is,

$$p(x_1, x_2|\theta) = p(x_1, x_2|\theta, r = 3) = p(x_1, x_2|\theta, r \neq 3).$$

Hence dropping P_3 in effect equates the MV distribution of X_1 and X_2 in P_3 to the set of patterns $\{P_0, P_1, P_2\}$.

These examples indicate the need for alternatives to CC restrictions that identify parameters by equating to sets of more than one pattern. Two approaches are proposed: *equating of parameters* and *equating via pattern-set mixture models.*

2.4.2 Equating Parameters to a Set of Patterns. Let $\gamma^{(r)}$ be a parameter indexing the distribution of x_i for pattern r in the saturated pattern-mixture model. Then $\gamma^{(r)}$ is equated to the set of patterns $\mathcal{S}$ if

$$\gamma^{(r)} = \Sigma_{s\in\mathcal{S}}\pi_s\gamma^{(s)}/\Sigma_{s\in\mathcal{S}}\pi_s. \tag{15}$$

In the next example, parameters of MV distributions are equated using (15).

Example 7 (Example 5 Continued): 2 × 2 Table with Supplemental Margins. Consider Table 1a supplemented by additional cases with both X_1 and X_2 missing (pattern P_3), and suppose that as in Example 5, parameters of the MV distributions of P_1 and P_2 are equated to P_0 and parameters of the MV distribution for P_3 are equated to $\{P_0, P_1, P_2\}$. ML estimation distributes cases in P_1 and P_2 as before, and cases in P_3 according to the probabilities in row 2 of Table 3; overall estimates of θ are unchanged, namely

$$\hat\theta_{11} = .285, \quad \hat\theta_{12} = .176, \quad \hat\theta_{21} = .245, \quad \text{and } \hat\theta_{22} = .295.$$

As noted earlier, an argument can be made that P_3 parameters should be equated to the other incomplete patterns, $\{P_1, P_2\}$. Then cases in $\{P_3\}$ are distributed using the probabilities obtained by merging the allocated counts for these two patterns; that is:

$$\hat\varphi_{11}^{(3)} = .202, \quad \hat\varphi_{12}^{(3)} = .191, \quad \hat\varphi_{21}^{(3)} = .236, \quad \text{and } \hat\varphi_{22}^{(3)} = .371,$$

which differ from the estimates in row 2 of Table 3. (In this case equating of parameters is equivalent to equating the MV distribution.) For example, with $n_3 = 100$, final estimates are

$$\hat\theta_{11} = .270, \quad \hat\theta_{12} = .178, \quad \hat\theta_{21} = .243, \quad \text{and } \hat\theta_{22} = .308.$$

As with CC restrictions, restrictions based on (15) need not be applied to the parameters of MV distributions. Consider bivariate normal data with the pattern of Figure 1a. If one regards P_1 and P_2 as similar strata, one might equate parameters of the marginal distributions of X_1 and X_2 in these patterns and equate the correlations or covariances to P_0. Equating correlations leads to the following set of six identifying restrictions:

$$\mu_1^{(2)} = \mu_1^{(1)}, \quad \sigma_{11}^{(2)} = \sigma_{11}^{(1)}, \quad \mu_2^{(1)} = \mu_2^{(2)},$$
$$\sigma_{22}^{(1)} = \sigma_{22}^{(2)}, \quad \rho^{(2)} = \rho^{(1)} = \rho^{(0)}, \tag{16}$$

where $\rho^{(j)}$ denotes the correlation for pattern P_j. The resulting ML estimate of μ_1 is a weighted average of the values of X_1, with cases in P_1 given $(n_1 + n_2)/n_1$ times the weight of cases in P_0; cases in P_1 are effectively representing the cases in P_1 and P_2 in this model. If instead of (16), $\mu_1^{(2)}$ and $\sigma_{11}^{(2)}$ are equated to P_1 *and* P_0 and $\mu_2^{(1)}$ and $\sigma_{22}^{(1)}$ are equated to P_2 *and* P_0, then the resulting ML estimates of the means and variances can be shown to be available-case estimates, as discussed by Little and Rubin (1987, chap. 3); the ML estimate

of the correlation differs slightly from the available-case estimate.

One drawback of placing restrictions on parameters not indexing the MV distribution is that resulting estimates of parameters of the marginal distribution of (X_1, X_2) may not lie in the parameter space. For example, if covariances rather than correlations are identified across patterns in (16), then the implied estimate of the correlation of X_1 and X_2 may not lie between -1 and 1.

2.4.3 Equating Via Pattern-Set Mixture Models. Equatings of parameters are not invariant to transformation. For example, equating correlations in equation (16) yields different solutions to equating covariances. Another approach is to define a selection model (5) for a subset of patterns and equate parameters to that subset. The following example, suggested by a referee, motivates this extension.

Example 8: Tying a Nonmonotone Pattern to a Monotone Set of Patterns. Consider data as in Figure 1a, with $n_3 = 0$. Suppose that the bulk of the data belong to P_0 and P_1, but a smaller but nonnegligible set belong to P_2. Let $s = 1$ index cases in P_0 or P_1, and let $s = 2$ index cases in P_2. Suppose that data with $s = 1$ are assumed to follow a normal ignorable selection model. Cases with $s = 2$ are modeled as normal, and the MV distribution of $X_2 | X_1$ is equated to the (normal) distribution of $X_2 | X_1$ given $s = 1$:

$$p(x_2 | x_1, s = 2, \varphi^{(2)}) = p(x_2 | x_1, s = 1, \theta_{2\cdot 1}^{(01)}),$$

where $\theta_{2\cdot 1}^{(01)}$ are the parameters of the normal linear regression of x_2 on x_1 for cases with $s = 1$. Note that if a normal pattern-mixture model was defined for P_0 and P_1, then the conditional distribution of X_2 given X_1 and $s = 1$ would be a mixture of normals and could not be equated with the (normal) distribution of X_2 given X_1 and $s = 2$. Replacing the pattern-mixture model for P_0 and P_1 by the selection model avoids this difficulty.

ML calculations for this model are illustrated by deriving the ML estimate of μ_1:

$$\hat{\mu}_1 = (\hat{\pi}_0 + \hat{\pi}_1)\hat{\mu}_1^{(01)} + \hat{\pi}_2\hat{\mu}_1^{(2)}, \tag{17}$$

where (01) denotes pooling over patterns 0 and 1. Applying the identifying restrictions for P_2 yields

$$\hat{\mu}_1^{(2)} = \hat{\beta}_{10\cdot 2}^{(2)} + \hat{\beta}_{12\cdot 2}^{(2)}\hat{\mu}_2^{(2)} = \hat{\beta}_{10\cdot 2}^{(01)} + \hat{\beta}_{12\cdot 2}^{(01)}\hat{\mu}_2^{(2)}, \tag{18}$$

where (01) denotes ML estimate for patterns 0 and 1, pooled. Reexpressing the (01) parameters in terms of parameters of the factorization X_1 and $X_2 | X_1$ yields

$$\hat{\beta}_{10\cdot 2}^{(01)} = (\hat{\sigma}_{22\cdot 1}^{(0)}\hat{\mu}_1^{(01)} - \hat{\sigma}_{11}^{(01)}\hat{\beta}_{20\cdot 1}^{(0)}\hat{\beta}_{21\cdot 1}^{(0)})/\hat{\sigma}_{22}^{(2)}$$

$$\text{and } \hat{\beta}_{12\cdot 2}^{(01)} = \hat{\beta}_{21\cdot 1}^{(0)}\hat{\sigma}_{11}^{(01)}/\hat{\sigma}_{22}^{(2)},$$

where $\hat{\sigma}_{22}^{(2)} = \hat{\sigma}_{22\cdot 1}^{(0)} + \hat{\beta}_{21\cdot 1}^{(0)2}\hat{\sigma}_{11}^{(01)}$. Substituting these expressions in (17) and (18) yields

$$\hat{\mu}_1 = \hat{\mu}_1^{(01)} + \hat{\beta}_{21\cdot 1}^{(0)}\hat{\sigma}_{11}^{(01)}\{\hat{\beta}_{21\cdot 1}^{(0)}(\hat{\mu}_1^{(0)} - \hat{\mu}_1^{(01)}) + (\hat{\mu}_2^{(2)} - \hat{\mu}_2^{(0)})\}/\hat{\sigma}_{22}.$$

Expressions for other parameter estimates can be derived in a similar fashion.

3. GENERAL PATTERNS

3.1 Introduction

I now outline extensions of the ideas of Section 2 for data with a general pattern of missing values. Again, let P_0 denote the set of complete cases, and suppose that there are T incomplete data patterns $P_1, \ldots, P_T$ in the population and $t \le T$ patterns $P_0, \ldots, P_t$ appear in the sample; let n_r denote the number of cases with pattern r, with $\sum_{r=0}^{t} n_r = n$. For case i in P_r, let $x_{obs,i}^{(r)}$ represent the set of observed variables, and let $x_{mis,i}^{(r)}$ represent the set of missing variables.

To define the saturated pattern-mixture model, let r_i take the value r for cases in P_r and suppose that r_i is multinomial with probabilities $p(r_i = r) = \pi_r$, $r = 0, 1, \ldots, T$. Also let $p(x_i | r_i = r, \varphi^{(r)})$ be the density for x_i given that $r_i = r$ and factorize it into the marginal distribution of the observed part of x_i and the conditional distribution of the missing part of x_i given the observed part:

$$p(x_i | r_i = r, \varphi^{(r)}) = p(x_{obs,i}^{(r)} | r_i = r, \varphi_{obs,r}^{(r)})p(x_{mis,i}^{(r)} | r_i = r, x_{obs,i}, \varphi_{mis,r}^{(r)}),$$

where $\varphi_{obs,r}^{(r)}$ and $\varphi_{mis,r}^{(r)}$ are functions of $\varphi^{(r)}$ and are assumed distinct for all r. (Here $\varphi_{mis,r}^{(r)}$ is short for the more descriptive but cumbersome $\varphi_{mis,r\cdot obs,r}^{(r)}$). The likelihood is

$$L(\pi, \varphi | X_{obs}, M) = \prod_{r=0}^{t}\left\{\pi_r^{n_r}\prod_{i\in P_r} p(x_{obs,i}^{(r)} | r_i = r, \varphi_{obs,r}^{(r)})\right\}. \tag{19}$$

Restrictions are needed to identify the parameters $\{\varphi_{mis,r}^{(r)}: r = 1, \ldots, t\}$, which do not appear in (19). As before, restrictions can be omitted for patterns that do not appear in the sample. Under MCAR restrictions, $\varphi^{(r)} = \theta$ for all r, (19) reduces to the likelihood for the corresponding ignorable selection model.

3.2 CCMV Restrictions for General Patterns

CCMV restrictions take the form

$$\varphi_{mis,r}^{(r)} = \varphi_{mis,r}^{(0)}, \quad 1 \le r \le t. \tag{20}$$

Example 9: Multivariate Normal Mixture Model. Suppose that for $r = 0, \ldots, T$, $p(x_i | r_i = r, \varphi^{(r)})$ denotes the V-variate normal distribution with parameter $\varphi^{(r)} = (\mu^{(r)}, \Sigma^{(r)})$, where $\mu^{(r)}$ is a $(V \times 1)$ mean vector and $\Sigma^{(r)}$ is a $(V \times V)$ covariance matrix. Then $\varphi_{obs,r}^{(r)}$ contains the mean and covariance matrix of the observed components of x_i for pattern r, and the sample mean and covariance matrices for each pattern maximize the likelihood (19). ML estimates of the unconditional mean and covariance matrix of x_i are obtained by regressing the missing components of x_i for each pattern on the observed components, using the complete cases to estimate the regression coefficients and residual covariance matrix. For consistent estimates, at least $V + 1$ complete cases are needed so that the sample covariance matrix from cases in P_0 is positive definite with probability 1. Then the mean of X_j has ML estimate

$$\hat{\mu}_j = \sum_{r=0}^{t} \hat{\pi}_r\hat{\mu}_j^{(r)},$$

where $\hat{\mu}_j^{(r)}$ is the sample mean of X_j for patterns r with X_j observed, and the regression estimate of $\mu_j^{(r)}$ from the regression on observed X_k's, for patterns with X_j missing. ML estimates of the elements of the covariance matrix of x_i are multivariate analogs of expressions given previously for the bivariate case.

Beale and Little (1975) described this method as corrected Buck, because it is essentially Buck's (1960) method for estimating μ and Σ, with a correction to the covariance estimates for consistency. The estimates are also the result of the first iteration of Orchard and Woodbury's (1972) EM algorithm for ML estimation under the normal ignorable selection model, with starting values based on complete cases. The link between ML for the pattern-mixture model with CCMV restrictions and the first iteration of EM for the ignorable selection model also applies to the multivariate generalization of the bivariate multinomial model of Example 5.

Table 4. *Wave Nonresponse Patterns for 20,938 Individuals in the ISDP*

Type	Pattern	Count	Percent	Count	Percent
a) Complete	11111			15,458	73.8
b) Attritors	1111?	419	2.0		
	111??	631	3.0		
	11???	896	4.3		
	1????	1254	6.0		
Subtotal		3200	15.3	3,200	15.3
c) Late entrants	?1111	333	1.6		
	??111	102	0.5		
	???11	89	0.4		
	????1	78	0.4		
Subtotal		602	2.9	602	2.9
d) Re-entrants,	1?111	380	1.8		
One wave missed	11?11	419	2.0		
	111?1	198	0.9		
Subtotal		997	4.8	997	4.8
e) Other patterns (20)				681	3.3
TOTAL				20,938	100.0

3.3 Other Restrictions

Because CCMV restrictions equate parameters to P_0, the method requires a reasonable number of complete cases. Under MCAR, the method can be inefficient if the fraction of complete cases is small. Alternatives to CCMV restrictions discussed in Sections 2.3 and 2.4 extend in obvious ways to more general patterns of data—for the general case, the possibilities seem almost endless. Discussion here is confined to three examples, and details are omitted.

Example 10: Trivariate Monotone Data. In an extension of Example 1, let X_1, X_2, and X_3 be outcomes in a panel survey with three time points. Suppose that the data have the monotone pattern of Figure 1b, where P_0 are complete cases, P_1 contains cases that drop out at time 3, and P_2 contains cases that drop out at time 2. The ignorable selection model equates the distribution of $X_2 | X_1$ in P_2 to (P_0, P_1); the pattern-mixture model with CCMV restrictions equates this distribution to P_0. It is plausible that because cases in P_1 and P_2 drop out of the sample, cases in P_2 are more like cases in P_1 than like cases in P_0, and hence an alternative equates the distribution of $X_2 | X_1$ in P_2 to P_1. This yields the restrictions

$$\varphi_{3\cdot 12}^{(2)} = \varphi_{3\cdot 12}^{(1)} = \varphi_{3\cdot 12}^{(0)} \quad \text{and} \quad \varphi_{2\cdot 1}^{(2)} = \varphi_{2\cdot 1}^{(1)}.$$

Note that because n_0 will usually be considerably larger than n_1, $\varphi_{2\cdot 1}^{(1)}$ will be less well estimated than will $\varphi_{2\cdot 1}^{(0)}$, so fitting this model entails some loss in precision. Hence a preliminary test is suggested to check whether P_0 and P_1 really differ with respect to the regression of X_2 on X_1.

Example 11: Tying Nonmonotone Patterns to a Prevalent Monotone Pattern. Table 4, taken from Little and David (1983), summarizes the missing-data pattern for 20,938 individuals surveyed in the Income Survey Development Program (ISDP), a panel survey of income that preceded the Survey of Income and Program Participation currently conducted by the U.S. Census Bureau. The survey had six waves; each column in Table 4 represents a wave. Note that most cases fall into the monotone pattern of attrition (types a and b), but some cases fall outside this pattern (types c, d, and e); ML estimation under ignorable selection models involves iterative methods. A way of avoiding iteration is to base analysis on a pattern-set mixture model, where complete and attriting cases (a and b) are modeled using an ignorable selection model and MV distributions for the other patterns are then equated to corresponding distributions for a and b defined by the selection model. This strategy allows for the simplicity of analysis of the monotone pattern, while retaining data that fall outside the pattern. It may be preferable to other ways of avoiding iteration, such as discarding data outside the monotone pattern (Marini, Olsen, and Rubin 1980) or imposing conditional independence assumptions on the variables (Andersson and Perlman 1989).

Example 12: Restrictions for Trivariate Normal Data with No Complete Cases. Consider pattern-mixture models for normal data with the pattern of Figure 1c. The model has $(3 \times 9) = 27$ φ parameters, 15 of which can be estimated from the bivariate observations. Complete-case restrictions are impossible, because there are no complete cases. Suppose instead that the mean and variance of X_1 in P_1 are equated to (P_2, P_3), the mean and variance of X_2 in P_2 are equated to (P_1, P_3), the mean and variance of X_3 in P_3 are equated to (P_1, P_2), and the correlations are equated across all three patterns. The resulting estimates of the mean and covariance matrix are similar to estimates from available-case analysis (see, for example, Little and Rubin 1987, chap. 3).

3.4 Theory for the General Model

3.4.1 Consistency. Despite the potentially large numbers of parameters in pattern-mixture models, the number of parameters is bounded as the sample size increases, because the number of patterns is bounded by T. Hence the usual consistency property of ML under the assumed model applies to identified parameters. This is not the case for some approaches; for example, methods that treat missing data as parameters have poor consistency properties, because the number of parameters increase with the sample size unless the fraction of missing data tends to 0 (Little and Rubin 1983).

Other consistency results can be obtained using the following result. Let the parameter $\delta = \delta(\pi, \varphi)$ be a continuously differentiable function of the pattern-mixture parameters, and suppose that (a) standard regularity assumptions hold for the distributions of observed data in each pattern; (b) for all r, $\hat{\varphi}^{(r)}_{obs,r} \to \varphi^{(r)}_{obs,r}$ and $\hat{\pi}_r \to \pi_r$ as $n \to \infty$; and (c) the identifying restrictions of the model are correct. Then $\hat{\varphi} \to \varphi$, and hence $\delta(\hat{\pi}, \hat{\varphi}) \to \delta$ as $n \to \infty$.

The identifying restrictions considered in this article are weaker than the MCAR assumption, so (c) will be satisfied under MCAR. If data are not MCAR, then validity of the restrictions becomes important. Condition (b) does not necessarily require the full distributional assumptions of the model. In particular, consider the normal pattern-mixture model just identified, with $\{\varphi^{(r)}_{obs,r}\}$ distinct for each pattern r. Then $\hat{\varphi}^{(r)}_{obs,r}$ consists of the sample mean and covariance matrix of the observed variables in pattern r, which are consistent for their population analogs under finite second-moment assumptions. Hence normality of this normal pattern-mixture model is not a requirement for ML estimates to be consistent. This result appears to extend to overidentified models, but no proof is offered. Interestingly, the result is stronger than corresponding consistency results for the normal ignorable selection model. Under that model, ML estimates are known to be consistent under MCAR, but consistency results that relax the normality assumption are not currently available under MAR.

3.4.2 Large Sample Covariance Matrix. Variance formulas derived from inverting information matrices do not require MCAR, but remain valid under the weaker assumptions about the missing-data mechanism implied by the identifying restrictions. Results are offered for the just-identified models where the parameters $\{\varphi^{(r)}_{obs,r}\}$ are distinct across patterns. The observed information matrix then has the block diagonal form

$$I(\hat{\pi}, \hat{\varphi}) = \text{diag}\{I(\hat{\pi}); I(\hat{\varphi}^{(0)}); I(\hat{\varphi}^{(1)}_{obs,1}); \ldots; I(\hat{\varphi}^{(t)}_{obs,t})\},$$

where $I(\hat{\pi})$ is a $(t-1) \times (t-1)$ multinomial information matrix, with (r, s) element $\hat{\pi}_r(1 - \hat{\pi}_r)/n$ for $r = s$ and $-\hat{\pi}_r\hat{\pi}_s/n$ for $r \neq s$, and $I(\hat{\varphi}^{(r)}_{obs,r})$ is the information matrix based on cases in P_r.

Let $\delta = \Sigma_r \pi_r \delta_r(\varphi)$ be a parameter of interest with ML estimate $\hat{\delta} = \Sigma_r \hat{\pi}_r \delta_r(\hat{\varphi})$, let $\hat{\delta}_r = \delta_r(\hat{\varphi})$, and let $\hat{\delta}_{r,u}$ denote the vector of partial derivatives of δ_r with respect to $\varphi^{(u)}_{obs,u}$, evaluated at $\hat{\varphi}$. Then, asymptotically, $\hat{\delta} - \delta$ is normal with mean 0 and asymptotic variance

$$\hat{\text{var}}(\hat{\delta} - \delta) = n^{-1} \sum_{r=0}^{t} \hat{\pi}_r(\hat{\delta}_r - \hat{\delta})^2 + \sum_{r=0}^{t} \sum_{s=0}^{t} \hat{\pi}_r\hat{\pi}_s c_{rs}, \quad (21)$$

where

$$c_{rs} = \sum_{u=0}^{t} \hat{\delta}^T_{r,u} I^{-1}(\hat{\varphi}^{(u)}_{obs,u}) \hat{\delta}_{s,u}. \quad (22)$$

These formulas require enough cases in each pattern for the information matrices $I(\varphi^{(r)}_{obs,r})$ to be invertible. For the normal pattern-mixture model, this requires

$$n_r \geq V_{obs,r} + 1$$

for all patterns r, where $V_{obs,r}$ is the number of observed variables for pattern r. This is of course a minimal requirement, particularly for patterns used to identify other parameters. For example, complete-case restrictions require at least $V + 1$ complete cases, but in practice more than that are needed to compute the required regressions with reasonable precision. The variance (21) of parameter estimates that aggregate over patterns may be small, even though the variance of estimates of parameters in rare patterns is high.

The information matrices $I(\varphi^{(u)}_{obs,u})$ in (22) are defined for each pattern separately and correspond to complete-data problems; the main work is in computing the partial derivatives and substituting them in (22). For the examples in Section 2, the computing was simplified by applying identifying restrictions to the ML estimate and computing the asymptotic variance of the resulting function directly. With complex identifying restrictions these formulas are tedious to apply, and simulation-based methods such as bootstrapping may have considerable appeal.

4. DISCUSSION

For problems involving complex patterns of missing data, selection models have been proposed and yield useful methods (Little and Rubin 1987). Why then consider pattern-mixture models? Some reasons are offered.

1. Pattern-mixture models provide a flexible class of models for data that are not MCAR. The observed data can provide evidence of departures from MCAR. In particular, for Example 1 MCAR implies that the distribution of X_1 is the same for respondents and nonrespondents to X_2—an assumption that is readily tested; for example, equality of the means can be assessed via a simple two-group t test. This idea is generalized to multivariate missing data in the BMDP8D program in Dixon (1989); a global MCAR test based on a pattern-mixture model with pooled within-pattern covariance matrix was presented in Little (1988b).

2. Pattern-mixture models are close to the way in which survey samplers view the nonresponse problem (see, for example, Cochran 1963, chap. 13); some have argued that by modeling the distribution of $X_1, \ldots, X_V$ separately for each missing-data pattern, problems of identifiability are made explicit that are obscured in the selection modeling approach (Glynn, Laird, and Rubin 1986; Rubin 1977). Nonignorable selection models like (3) are often identified by untestable distributional assumptions. For particular choices of g in (3), we obtain simple cases of the probit selection model of Heckman (1976) and the logit selection model of Greenlees, Reece, and Zieschang (1982). The parameters $\theta = (\mu, \Sigma)$ and ψ are identified, but estimates rely heavily on normal assumptions (Glynn, Laird, and Rubin 1986; Little 1982; Little and Rubin 1987, chap. 11).

3. Lack of robustness of nonignorable selection models to untestable assumptions has led some (including this author) to favor *ignorable* selection models. But the MAR assumption (4) that underpins such models is not necessarily appropriate or even natural in multivariate settings. Consider, for example, the bivariate pattern of Figure 1a. The MAR assumption can be shown to imply that for each case i,

$$m_{i1} \perp\!\!\!\perp x_{i1} \mid x_{i2} \quad \text{and} \quad m_{i2} \perp\!\!\!\perp x_{i2} \mid x_{i1}, \quad (23)$$

where m_{ij} is the missing-value indicator for x_{ij} and $\perp\!\!\!\perp$ denotes independence. The pattern-mixture model with complete-case restrictions (9) implies that

$$m_{i1} \perp\!\!\!\perp x_{i1} | (x_{i2}, m_{i2} = 0)$$
$$\text{and} \quad m_{i2} \perp\!\!\!\perp x_{i2} | (x_{i1}, m_{i1} = 0). \tag{24}$$

The choice between models hinges on whether (23) or (24) is a better assumption in practice; one might argue that (24) is preferable, because independence of m_{ij} and x_{ij} is conditioned not only on the value of the other variable but also on the missing-data indicator for that variable. Plausible alternatives to (23) and (24) can be constructed; for example, a number of authors (Baker and Laird 1988; Brown 1990; Fay 1986; Little 1985) have considered inference assuming

$$m_{i1} \perp\!\!\!\perp x_{i2} | x_{i1} \quad \text{and} \quad m_{i2} \perp\!\!\!\perp x_{i1} | x_{i2}, \tag{25}$$

where missingness of a variable depends only on the value of that variable. Such mechanisms can also be modeled using the pattern-mixture approach, as shown in Example 6.

We would like the data to discriminate between these alternative assumptions and the assumptions implied by the other identifying restrictions in Section 2; however, tests would be driven by distributional assumptions such as normality. In the multinomial setting all just-identified models yield perfect fits to the observed data (Little and Rubin 1987, sec. 11.6), so choices need to be based on knowledge of the missing-data mechanism. If knowledge is limited, then it may be more honest to present a range of estimates under alternative assumptions about the missing-data mechanism. This is a difficult area where more work seems needed.

4. We have seen that for some problems, ML for selection models requires numerical methods, whereas just-identified pattern-mixture models lead to ML estimates with explicit forms. Such models are readily amenable to Bayesian small-sample analyses, as in Little (1988a). They also provide a basis for multiple imputation of general missing-data patterns, for which Rubin (1987, sec. 5.6) has noted that more tools are needed. Other methods for simulating the posterior distribution, such as data augmentation (Tanner 1990), might be usefully applied to these models.

A useful development of the models proposed here is to attempt to assess sensitivity to departures from the identifying restrictions. Rubin (1987) suggested that differences between the distribution of $x_{mis,i}$ given $x_{obs,i}$ among respondents and nonrespondents can be modeled simply by adding simple bias and scale quantities to the draws from the predictive distribution. Such quantities are easily incorporated into pattern-mixture models without complicating the subsequent analysis, although deciding appropriate values of these quantities may be no easy matter.

[*Received November 1990. Revised April 1992.*]

REFERENCES

Anderson, T. W. (1957), "Maximum Likelihood Estimation for the Multivariate Normal Distribution when Some Observations are Missing," *Journal of the American Statistical Association,* 52, 200–203.

Andersson, S. A., and Perlman, M. D. (1989), Lattice-Ordered Conditional Independence Models for Missing Data," Technical Report No. 179, Department of Statistics, University of Washington.

Baker, S. G., and Laird, N. M. (1988), "Regression Analysis for Categorical Variables with Outcome Subject to Nonignorable Nonresponse," *Journal of the American Statistical Association,* 83, 62–69.

Beale, E. M. L., and Little, R. J. A. (1975), "Missing Values in Multivariate Analysis," *Journal of the Royal Statistical Society,* Ser. B, 37, 129–145.

Buck, S. F. (1960), "A Method of Estimation of Missing Values in Multivariate Data Suitable for Use with an Electronic Computer," *Journal of the Royal Statistical Society,* Ser. B, 22, 302–306.

Brown, C. H. (1990), "Protecting Against Nonrandomly Missing Data in Longitudinal Studies," *Biometrics,* 46, 143–157.

Chen, T., and Fienberg, S. E. (1974), "Two-Dimensional Contingency Tables with Both Completely and Partially Classified Data," *Biometrics,* 30, 629–642.

Cox, D. R., and Wermuth, N. (1990), "An Approximation to Maximum Likelihood Estimates in Reduced Models," *Biometrika,* 77, 747–762.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society,* Ser. B, 39, 1–38.

Fay, R. E. (1986), "Causal Models for Nonresponse," *Journal of the American Statistical Association,* 81, 354–365.

Glynn, R. J., Laird, N. M., and Rubin, D. B. (1986), "Selection Modeling Versus Mixture Modeling with Nonignorable Nonresponse," in *Drawing Inferences from Self-Selected Samples,* ed. H. Wainer, New York: Springer-Verlag, pp. 115–142.

Greenlees, W. S., Reece, J. S., and Zieschang, K. D. (1982), "Imputation of Missing Values When the Probability of Response Depends on the Value Being Imputed," *Journal of the American Statistical Association,* 77, 251–261.

Heckman, J. (1976), "The Common Structure of Statistical Models of Truncation, Sample Selection, Limited Dependent Variables, and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement,* 5, 475–492.

Little, R. J. A. (1982), "Models for Nonresponse in Sample Surveys," *Journal of the American Statistical Association,* 77, 237–250.

——— (1985), "Nonresponse Adjustments in Longitudinal Surveys: Models for Longitudinal Data," *Bulletin of the International Statistical Institute, Invited Papers,* Section 15.1, 1–18.

——— (1988a), "Small-Sample Inference About Means from Bivariate Normal Data with Missing Values," *Computational Statistics and Data Analysis,* 7, 161–178.

——— (1988b), "A Test of Missing Completely at Random for Multivariate Data with Missing Values," *Journal of the American Statistical Association,* 83, 1198–1202.

Little, R. J. A., and David, M. (1983), "Weighting Adjustments for Nonresponse in Panel Surveys," working paper, U.S. Bureau of the Census, Washington, DC.

Little, R. J. A., and Rubin, D. B. (1983), "On Jointly Estimating Parameters and Missing Data by Maximizing the Complete-Data Loglikelihood," *The American Statistician,* 37, 218–220.

——— (1987), *Statistical Analysis with Missing Data,* New York: John Wiley.

——— (1989), "Missing Data in Social Science Data Sets," *Sociological Methods and Research,* 18, 292–326; reprinted in *Modern Methods of Data Analysis,* eds. J. S. Long and J. Fox, Newbury Park, CA: Sage Press, pp. 374–409.

Marini, M. M., Olsen, A. R., and Rubin, D. B. (1980), "Maximum Likelihood Estimation in Panel Studies with Missing Data," in *Sociological Methodology,* 11, 314–357.

Orchard, T., and Woodbury, M. A. (1972), "A Missing Information Principle: Theory and Applications," *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability,* 1, 697–715.

Rubin, D. B. (1974), "Characterizing the Estimation of Parameters in Incomplete Data Problems," *Journal of the American Statistical Association,* 69, 467–474.

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

——— (1977), "Formalizing Subjective Notions About the Efffect of Nonrespondents in Sample Surveys," *Journal of the American Statistical Association,* 72, 538–543.

——— (1978), "Multiple Imputations in Sample Surveys—A Phenomenological Bayesian Approach," *Proceedings of the Survey Research Methods Section, American Statistical Association,* 20–34.

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

Tanner, M. (1990), *Tools for Statistical Inference: Observed Data and Data Augmentation Methods,* New York: Springer-Verlag.