Multiple Imputation for Incomplete Data With Semicontinuous Variables

Kristin N. JAVARAS and David A. VAN DYK

We consider the application of multiple imputation to data containing not only partially missing categorical and continuous variables, but also partially missing 'semicontinuous' variables (variables that take on a single discrete value with positive probability but are otherwise continuously distributed). As an imputation model for data sets of this type, we introduce an extension of the standard general location model proposed by Olkin and Tate; our extension, the blocked general location model, provides a robust and general strategy for handling partially observed semicontinuous variables. In particular, we incorporate a two-level model for the semicontinuous variables into the general location model. The first level models the probability that the semicontinuous variable takes on its point mass value, and the second level models the distribution of the variable given that it is not at its point mass. In addition, we introduce EM and data augmentation algorithms for the blocked general location model with missing data; these can be used to generate imputations under the proposed model and have been implemented in publicly available software. We illustrate our model and computational methods via a simulation study and an analysis of a survey of Massachusetts Megabucks Lottery winners.

KEY WORDS: Data augmentation; EM algorithm; General location model; Missing data; Survey data.

1. INTRODUCTION

1.1 Multiple Imputation With Semicontinuous Variables

Missing data often complicate statistical analysis, particularly in the survey setting where item nonresponse is frequently encountered. A common method for handling such missing data is to impute them before any statistical analysis is performed. This method divides data analysis into two phases. First, in the *imputation phase*, an imputation of the dataset (i.e., a complete version of the data with all missing values imputed) is generated. Second, in the *analysis phase*, the imputed dataset is analyzed as if there were no missing data. Usually, the imputation phase is conducted in consultation with the data collector, incorporates available expert knowledge, and uses sophisticated statistical tools (e.g., Monte Carlo methods for fitting highly structured models). The analysis phase can be repeated by numerous end-users who are less familiar with the data collection mechanism and have fewer statistical tools. The basic advantage of this strategy is that it separates the complications caused by the missing data from the basic statistical analysis. Thus, from a computational standpoint, imputation simplifies the analysis phase so that it requires only standard statistical methods (i.e., methods for data with no missing values). More importantly, from an inferential standpoint, this method incorporates the knowledge and information available to the data collector into the imputation. Thus, imputation methods are more than computational tools; they are a mode of inference that allows for the sequential input of multiple sources of information (see, e.g., Meng 1994).

Unfortunately, it is quite unusual for the imputation to be completely deterministic; that is, even with the expert knowledge of the data collector, there is uncertainty involved in the imputation. Valid inference must account for this uncertainty. Multiple imputation (MI) (e.g., Rubin 1987) is an oft-adopted approach for obtaining such valid inference while accounting for missing data. In the imputation phase of MI, M imputations (rather than just one imputation) of the dataset are generated. Ordinarily, a model-based method is used to generate the M imputations, which then represent repeated random draws from a model for the nonresponse, which is a conditional distribution of the *imputation model* for the complete data. In the analysis phase, each of the M imputed datasets is analyzed as if there were no missing data, and the M sets of estimates and corresponding errors are combined according to simple *combining rules*. The power of MI lies in its ability to return valid inferences—that is, ones that reflect the additional variability due to the missing values under the imputation model—while requiring only standard statistical methods in the analysis phase.

Of course, it is important that the imputation model be suitable for the particular dataset of interest. For certain types of datasets, existing models can make suitable imputation models. For instance, the *general location model* (GLoM) used by Olkin and Tate (1961) and Krzanowski (1980, 1982) is a standard imputation model for incomplete datasets containing continuous and categorical variables; it is appropriate as long as its assumption of normality and variance constraints are suitable for the continuous variables in the dataset (or for some transformations of them). Imputations of the incomplete categorical and continuous dataset can be generated using a GLoM for the imputation model by employing either the expectation-maximization (EM) algorithm (Dempster, Laird, and Rubin 1977) developed by Little and Schluchter (1985) for the GLoM or the data augmentation algorithm (Tanner and Wong 1987) developed by Schafer (1997) for the GLoM.

But what if the incomplete dataset contains not only continuous and categorical variables, but also *semicontinuous variables*, as is common in practice? A semicontinuous variable is defined as a variable that takes on a specific value (say, zero, without loss of generality) with a positive probability, but otherwise takes on values that can be modeled by a continuous distribution. Some common examples of semicontinuous variables

Kristin N. Javaras is a doctoral candidate, Department of Statistics, University of Oxford, Oxford OX1 3TG, U.K. (E-mail: *javaras@stats.ox.ac.uk*). David A. van Dyk is Associate Professor, Department of Statistics, University of California at Irvine, Irvine, CA 92697 (E-mail: *dvd@ics.uci.edu*). This work was partially supported by National Science Foundation grant DMS-01-04129, the Rhodes Trust, and the U.S. Census Bureau. The authors thank John Barnard, Xiao-Li Meng, and Donald Rubin for insights and advice proffered during numerous discussions and Brian Ripley for his expert computer and programming assistance and recommendations. In addition, they thank all of the aforementioned people, as well as the JASA editorial staff and the anonymous referees, for their valuable comments on various drafts of this article.

Journal of the American Statistical Association
September 2003, Vol. 98, No. 463, Theory and Methods
DOI 10.1198/016214503000000611

(with point masses at zero) are number of years of higher education, dollar amount of unemployment benefits, and number of years of exposure to a particular carcinogen. As is the case for these examples, the continuous values of a particular semicontinuous variable are often restricted to a certain part of the real line (e.g., only nonnegative values are allowed in the foregoing examples). Also, it is not uncommon for the continuous values of a given semicontinuous variable to have a skewed distribution; however, if a semicontinuous variable is treated as continuous, then normalizing transformations such as the log transformation are problematic because of the point mass at zero. In this article we introduce a new model, the *blocked general location model* (BGLoM), that is an extension of the standard GLoM and that can be used as an imputation model for a dataset containing semicontinuous variables. The BGLoM is an alternative to existing imputation methods for semicontinuous variables proposed by Herzog and Rubin (1983) and Heeringa, Little, and Raghunathan (2002), both of which are discussed in Section 1.3. Throughout this article, we assume that the missing-data mechanism is *ignorable*, which requires that the missing data be missing at random, as defined by Rubin (1976), and that the parameters of the imputation model and missing-data mechanism be distinct (Little and Rubin 1987; Rubin 1987). Under this assumption, the BGLoM is an appropriate imputation model as long as its distributional assumptions are suitable for the dataset at hand (possibly after some transformations of the variables).

Even with an appropriate imputation model, it is important that the analysis model be *congenial* with the imputation model (Meng 1994). If the imputation and analysis models differ, then the resulting inference may not be valid. For example, if a particular interaction is assumed to be zero in the imputation model, then the imputed values will not show the interaction; as a result, any interaction existing in the observed data will be attenuated in the final analysis. Examples such as this one led Meng to suggest the following:

> To accommodate a wide variety of subject-motivated analyses that will be performed on the imputed datasets, the imputation model should be as objective and general as the imputer's resources allow. This implies that general and saturated models are preferred to models with special structures (e.g., models that assume certain interactions are zero), and imputation models should include predictors that are likely to be part of potential analyses even if these predictors are known to have limited predictive power for the existing incomplete observations.

(For further discussion of the issue of imputation model generality, see also Rubin 1980, 1996; Clogg, Rubin, Schenker, Schultz, and Weidman 1991; Schenker, Treiman, and Weidman 1988, 1993; Liu and Rubin 1998; Barnard, McCulloch, and Meng 2000.) This concern about imputation model generality is clearly of fundamental importance and is a driving force in our model specification; we allow more flexibility in the variance structure than is available with the standard GLoM.

After presenting the BGLoM, we describe several methods that can be used to generate imputations of the dataset under a BGLoM imputation model. We also introduce an EM algorithm and a data augmentation algorithm for the BGLoM with missing data, both of which can be used to implement these methods of generating imputations. Once the imputations have been generated, a method appropriate for complete data with semicontinuous variables can be used by the end-user to analyze the imputed datasets. Although there is some shortage of appropriate methods, a variety of useful tools do exist. Many of these tools model the semicontinuous variables in two parts, as does our BGLoM (see Olsen and Schafer 2001 for a review of these two-part models).

1.2 Motivating Example

This article is motivated by the Massachusetts Megabucks Lottery Winners Survey (MMLWS) dataset, which contains background information and various measures of economic behavior for individuals who won the Massachusetts Megabucks Lottery in the mid-1980s. This dataset is interesting from an economic standpoint because the lottery effectively produced a randomized experiment that assigned large sums of cash to certain individuals, thereby allowing investigation into the effects of unearned income on economic behavior (Imbens, Rubin, and Sacerdote 2001). In addition to continuous and categorical variables, the MMLWS dataset includes semicontinuous variables, such as annual income due to employment in 1995, which has a sizable point mass at zero corresponding to individuals who were unemployed in 1995. Further, in the MMLWS dataset, almost all of the variables are *partially missing*, having at least one missing value.

Previous analyses of the MMLWS dataset by Imbens, Rubin, and Sacerdote (2001) utilized the *complete-case* approach to handle the missing data: Units with missing values for one or more of the variables of interest were discarded before any data analysis was begun. As the authors acknowledge, the complete-case approach results at the very least in a loss of information and potentially in serious biases; thus, it was decided to reanalyze the MMLWS dataset using MI. However, as Schafer (1997, sec. 10.2.4) notes, there is a dearth of suitable model-based imputation methods for incomplete datasets with semicontinuous variables; not surprisingly, then, none of the existing models and accompanying imputation generating methods considered were deemed appropriate means of implementing the MI approach for the MMLWS dataset.

Throughout the article, we refer to a subset of the MMLWS dataset for illustrative purposes. This subset includes the binary categorical variable gender ($0 =$ female, $1 =$ male), the continuous variable winnings (dollar amount won in the lottery), and the semicontinuous variables 1992 earnings (annual income from employment in 1992) and 1995 earnings (annual income from employment in 1995). Both semicontinuous variables have point masses located at zero that comprise those individuals who were not employed in 1992 and 1995, respectively. More generally, the *dataset of interest* refers to a dataset with r_0 semicontinuous, q_0 continuous, and p_0 categorical variables, any of which can be partially missing.

1.3 Existing Imputation Models and Model-Based Methods

In this section we outline three existing imputation models and methods that might be, but should not be, used to generate imputations of the dataset of interest. The first imputation model that might be suitable is the aforementioned GLoM. Under the standard GLoM, the categorical variables have a marginal distribution that is multinomial across the cells of a contingency table produced by crossing the levels of the categorical variables. Conditional on the categorical variables, the continuous variables have a multivariate normal distribution with means that vary across the cells and a covariance matrix that is common across the cells. Because of this conditional formulation, generating imputations under a GLoM imputation model is generally relatively easy.

Although the GLoM and the accompanying EM and data augmentation algorithms could be used to generate imputations of the dataset of interest if all of its semicontinuous variables were treated as continuous variables, doing so would result in questionable imputations for the missing values of these variables. This is the case because the point mass value (here, zero) for one such variable would be imputed with probability zero under a continuous model; for instance, in our example no values of zero would be imputed for 1992 earnings, which is a somewhat dubious result given the nontrivial level of unemployment among the individuals with observed values. Additionally, as noted earlier, it may not be possible to use common normalizing transformations to make the standard GLoM's within-cell normality assumption more appropriate for a skewed semicontinuous variable. Finally, the continuous model may result in imputations outside the range of possible values (e.g., negative 1992 earnings). Although some of these difficulties can sometimes be mitigated by rounding values that are impossible or simply near the point mass to the point mass, such post hoc corrections generally require ad hoc rules and ignore the inappropriate nature of the Gaussian model for semicontinuous data.

A second existing option is the model-based method proposed by Herzog and Rubin (1983), in which a partially missing semicontinuous variable is imputed in two stages. However, this method is intended for the special case in which there is only one semicontinuous variable and all other variables are fully observed. Likewise, a third existing option, used by Heeringa, Little, and Raghunathan (2002), used a two-stage model for semicontinuous variables. In their examples, however, not only are all continuous and categorical variables completely observed, but also the indicator variables for the point mass values of the semicontinuous variables are completely observed (i.e., only the continuous values of the semicontinuous variables might be missing). Neither of these assumptions is appropriate for our more general dataset of interest.

Although none of these options is suitable for the dataset of interest, we adopt certain aspects of all three in our formulation of an appropriate imputation model. In particular, Heeringa, Little, and Raghunathan (2002) suggest that their model can be generalized to handle more general missing-data patterns and more general variance structures. It is these generalizations that we address.

1.4 Organization of the Article

In Section 2, we introduce the BGLoM, highlighting how it differs from the standard GLoM. We also describe how one might generate imputations of the dataset of interest under a BGLoM imputation model, using EM and data augmentation algorithms developed for the BGLoM with missing data. In Section 3, we describe the encouraging performance of the BGLoM in both a simulation study and an analysis of the MMLWS subset example. Concluding remarks regarding the benefits and limitations of our BGLoM formulation are provided in Section 4. In the appendixes we present the full details of the EM and data augmentation algorithms that we develop to fit the BGLoM.

2. A NEW IMPUTATION MODEL: THE BLOCKED GENERAL LOCATION MODEL

We now introduce a new model, the BGLoM, which is a suitable imputation model for datasets with partially missing variables of all three types provided that the model's distributional assumptions are appropriate and that the missing-data mechanism is ignorable. However, before describing the BGLoM, we discuss a popular and convenient way of treating semicontinuous variables used in the BGLoM. Further, we show how this treatment necessitates an extension of the GLoM, thereby leading to the formulation of the BGLoM.

2.1 Two-Part Models for Semicontinuous Variables

Similar to the way in which Herzog and Rubin (1983) treated the semicontinuous variable in their method, we replace each of the r_0 semicontinuous variables by a *constructed categorical variable* that has two levels (zero if the semicontinuous variable takes on the point mass value, and one otherwise) and a *constructed continuous variable* that is *relevant* if and only if the constructed categorical variable has a value of 1 (see also Dunn, Manning, Morris, and Newhouse 1983; Manning et al. 1981; Heeringa, Little, and Raghunathan 2002). When the constructed continuous variable is relevant, it equals the corresponding semicontinuous variable; when irrelevant, the constructed continuous variable is not defined. For instance, in the MMLWS subset example, 1992 earnings would be replaced with a positive earnings indicator variable that takes a value of one if an individual was working in 1992 and zero otherwise, and a constructed continuous variable that is relevant only when the positive earnings indicator variable takes a value of one and that represents the earnings of an individual employed in 1992. Analogous replacements would occur for 1995 earnings.

These $2r_0$ constructed variables can be combined with the dataset's p_0 *pure* (i.e., nonconstructed) categorical variables and q_0 pure continuous variables to produce $p = p_0 + r_0$ *combined* (i.e., pure and constructed) categorical variables and $q = q_0 + r_0$ combined continuous variables. In our illustrative example, the result would be three combined continuous variables and three combined categorical variables. (In our notation we use a zero in the subscript to refer to summaries of the pure variables and a prime in the superscript to refer to summaries of the constructed variables; summaries of the combination of pure and constructed variables have no such adornment.)

2.2 Extending the General Location Model

After redefining the dataset of interest in terms of combined continuous and categorical variables only, as is described in Section 2.1, it may appear at first glance that the standard GLoM can be used as a model for the redefined data. To see why this is not the case, we consider the contingency table produced by crossing the levels of the combined categorical variables, which we term the *augmented contingency table.* This p-dimensional augmented contingency table has $C = 2^{r_0} \prod_{j=1}^{p_0} c_j$

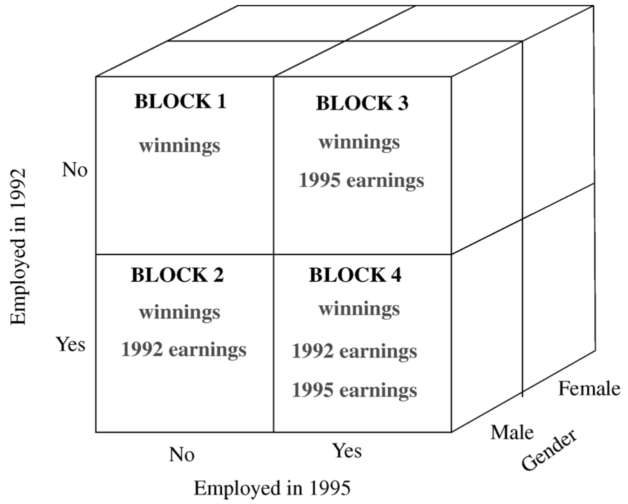

Figure 1. Augmented Contingency Table for the MMLWS Subset Example. The blocks are determined by the semicontinuous variables, via the constructed categorical variables, and have different sets of relevant constructed continuous variables. These variables, in addition to the pure continuous variable, are listed in gray type in each block.

cells, where c_j is the number of levels for pure categorical variable j. In this table, different sets of constructed continuous variables are relevant in different cells. For a concrete illustration of this phenomenon, we turn to our MMLWS subset example: the eight-cell augmented contingency table for this example is portrayed in Figure 1, which shows that the constructed continuous variable earnings in 1995 is relevant in all cells corresponding to employment in both 1992 and 1995, but not in cells corresponding to employment in 1992 and unemployment in 1995. Thus, cells of the former variety have one more constructed continuous variable than do cells of the latter variety.

Because not all cells in the augmented contingency table contain the same set of continuous variables, the covariance matrices for the continuous variables do not refer to the same set of variables (or even have the same dimensions) in every cell, unlike the covariance matrix in the standard GLoM. For this reason, it is necessary, as Schafer (1997, pp. 381–382) notes, to extend the standard GLoM to make it appropriate for the redefined dataset; below we introduce just such an extension, the BGLoM. In contrast, the strategy suggested by Heeringa, Little, and Raghunathan (2002) entails treating the variables that are not relevant in each block as missing data; see the last paragraph of Section 2.4 for more on this strategy.

2.3 Definition of Blocks

Before we present the particular extension that we propose in this article, the BGLoM, we must introduce the concept of a *block*, because the concept is integral to the formulation of the BGLoM. We seek to formalize the concept that not all cells in the augmented contingency table have the same set of relevant continuous variables. To do so, we divide the table into mutually exclusive and exhaustive blocks, where a block is defined as the set of all cells that share the same values for the r_0 constructed categorical variables. The augmented contingency table can be divided into 2^{r_0} blocks, which are formed by crossing the levels of the constructed categorical variables. Within each block is a smaller p_0-dimensional contingency table with $C_0 = \prod_{j=1}^{p_0} c_j$ cells; this smaller table is produced by crossing the levels of the pure categorical variables while holding the levels of the constructed categorical variables constant. As a result of the definition for a block, the same constructed continuous (and, of course, the same pure continuous) variables are relevant in every cell within a given block.

Figure 1 depicts the four blocks in the MMLWS subset example. The blocks, each of which consists of two cells, recede into the page, and, for each block, the relevant continuous variables are enumerated in gray type. Block 1 corresponds to unemployment in both years; block 2, to employment in 1992 and unemployment in 1995; block 3, to unemployment in 1992 and employment in 1995; and block 4, to employment in both years. Within each block are two cells that correspond to the "male" and "female" levels of the only pure categorical variable, gender. Further, the continuous variable winnings can be observed in all four blocks, but the constructed continuous variable earnings in 1995 is relevant only in blocks 3 and 4, and the constructed continuous variable earnings in 1992 is relevant only in blocks 2 and 4.

2.4 Definition of the Blocked General Location Model

Proceeding with the definition of the BGLoM, we assume a completely saturated multinomial model for the combined categorical variables across the cells of the augmented contingency table. We also assume that the joint distribution of the pure continuous variables and the relevant constructed continuous variables is multivariate normal within the cells of the augmented contingency table. In the BGLoM, the means of this joint distribution are assumed to differ across the cells, as is the case in the standard GLoM. However, the variances and covariances are assumed to be constant only within, and not across, blocks, unlike in the standard GLoM where the covariance matrix is constant across all cells. Obviously, because the same set of continuous variables is relevant in every cell within a given block, the dimensions of the covariance matrix are the same for all cells in a block. In principle, we could constrain certain variances and covariances to be equal across within-block covariance matrices, which would reduce the number of parameters to be estimated. However, in the interest of simplifying the following presentation (and because a more general imputation model is often desirable, as is discussed later in this section), we do not incorporate such constraints here.

To formalize the BGLoM, we let $W_1, \ldots, W_{p_0}$ denote the pure categorical variables, let $Z_1, \ldots, Z_{q_0}$ denote the pure continuous variables, and let $S_1, \ldots, S_{r_0}$ denote the semicontinuous variables, all recorded for n individuals. For semicontinuous variable S_j, define a constructed categorical variable W_{p_0+j} that takes the value 0 if $S_j = 0$ and 1 if $S_j \neq 0$, as well as a constructed continuous variable Z_{q_0+j} that is relevant and equal to S_j if $W_{p_0+j} = 1$ and *irrelevant* if $W_{p_0+j} = 0$. The combined sets of categorical and continuous variables are denoted by $W = (W_1, \ldots, W_p)$ and $Z = (Z_1, \ldots, Z_q)$, respectively. We use lower-case letters to refer to the rows of W and Z; thus, w_i and z_i represent the (combined) categorical and (combined) continuous variables for individual i. Finally, we denote the complete dataset by $Y = \{(w_i, z_i^{\text{rel}}),\ i = 1, \ldots, n\}$, where z_i^{rel} represents the components of z_i that are relevant. Throughout, we make some effort to follow the notation of Schafer (1997, chap. 9).

The complete-data sampling distribution can be factored into

$$p(Y|\theta) = \prod_{i=1}^{n} p(z_i^{\text{rel}}|\theta, w_i) p(w_i|\theta), \tag{1}$$

where θ is the set of unknown model parameters. We begin with the marginal distribution of the categorical variables. Each w_i can be classified into a p-way contingency table with C cells. We assume a completely saturated multinomial model for this contingency table. Specifically, we let $U = (u_{ic})$ be an $n \times C$ matrix with u_{ic} indicating whether individual i belongs to cell c: in other words, we let $u_{ic} = I\{w_i \text{ belongs to cell } c\}$ and assume

$$(u_{\cdot 1}, \ldots, u_{\cdot C}) \sim \text{multinomial}\,(n, \pi), \tag{2}$$

where $u_{\cdot c} = \sum_{i=1}^{n} u_{ic}$ for $c = 1, \ldots, C$ and $\pi = (\pi_1, \ldots, \pi_C)$ is a probability vector.

The conditional distribution of the continuous variables given the categorical variables depends on w_i through its cell and block classification. To keep track of the relationship between individual, cell, and block indexes, we introduce six possibly set-valued functions. Specifically, $\mathcal{I}(c,)$ and $\mathcal{I}(, b)$ denote the set of indexes of individuals in cell c and in block b, respectively, $\mathcal{C}(w_i,)$ and $\mathcal{C}(, b)$ denote the index of the cell containing individuals with categorical variables w_i and the indexes of the cells in block b, respectively, and $\mathcal{B}(w_i',)$ and $\mathcal{B}(, c)$ denote the index of the block containing individuals with constructed categorical variables w_i' and containing cell c, respectively. These definitions are formalized in Table 1. We can now complete the model specification with

$$z_i^{\text{rel}} \mid \theta, w_i \sim \text{N}(\mu_c, \Sigma_b), \tag{3}$$

where $c = \mathcal{C}(w_i,)$, $b = \mathcal{B}(w_i',) = \mathcal{B}(, c)$, μ_c is a vector of length q_b^{rel}, and Σ_b is a $(q_b^{\text{rel}} \times q_b^{\text{rel}})$ positive definite matrix, with q_b^{rel} the number of relevant continuous variables in block b.

Using (1)–(3), we can construct the complete-data likelihood function,

$$L(\theta|Y) \propto \left(\prod_{c=1}^{C} \pi_c^{u_{\cdot c}} \right) \prod_{b=1}^{B} |\Sigma_b|^{-u_{\cdot\cdot}^b/2} \times \exp\left[-\frac{1}{2} \left\{ \sum_{c \in \mathcal{C}(,b)} \sum_{i \in \mathcal{I}(c,)} (z_i^{\text{rel}} - \mu_c)^\top \Sigma_b^{-1} (z_i^{\text{rel}} - \mu_c) \right\} \right], \tag{4}$$

Table 1. Mappings Between Individual, Cell, and Block Indexes

$\mathcal{I}(\text{c},) = \{\text{i} : w_i = w(c)\}$	*The set of indexes of individuals in cell c*
$\mathcal{I}(, b) = \{\text{i} : w_i' = w'(\text{b})\}$	The set of indexes of individuals in block b
$\mathcal{C}(\text{w}_i,) = \{\text{c} : w(c) = w_i\}$	The index of the cell containing individuals with categorical variables w_i
$\mathcal{C}(, b) = \{\text{c} : w(c)' = w'(\text{b})\}$	The set of indexes of cells in block b
$\mathcal{B}(\text{w}_i',) = \{\text{b} : w'(\text{b}) = w_i'\}$	The index of the block containing individuals with constructed categorical variables w_i'
$\mathcal{B}(, c) = \{\text{b} : w'(\text{b}) = w(c)'\}$	The index of the block containing cell c

NOTE: Cell membership of individuals is uniquely determined by w_i and block membership is determined by w_i', the component of w_i corresponding to the constructed categorical variables. We denote the value of w_i corresponding to cell c by $w(c)$, its subvector corresponding to the constructed categorical variables by $w(c)'$, and the value of w_i' corresponding to block b by $w'(b)$. With these definitions, we define the index functions.

where $\theta = (\pi, \mu_1, \ldots, \mu_C, \Sigma_1, \ldots, \Sigma_B)$, $B = 2^{r_0}$ is the number of blocks, and $u_{\cdot\cdot}^b = \sum_{i=1}^{n} \sum_{c \in \mathcal{B}(,c)} u_{ic}$ is the total count in block b. In some cases we examine the posterior distribution $p(\theta|Y) \propto L(\theta|Y)p(\theta)$, where $p(\theta)$ is the independent reference prior distribution

$$p(\theta) \propto \left(\prod_{c=1}^{C} \pi_c^{\alpha_c - 1} \right) \left(\prod_{b=1}^{B} |\Sigma_b|^{-\frac{(q_b^{\text{rel}}+1)}{2}} \right), \tag{5}$$

where $\alpha = (\alpha_1, \ldots, \alpha_C)$ is a vector of user-specified hyperparameters. In (5) we choose a marginal Dirichlet prior distribution for π, a flat prior distribution for each μ_c, and the standard noninformative Jeffreys prior distribution for each Σ_b.

At this point, we pause to discuss why the BGLoM's assumption of block-specific covariance matrices, as opposed to a common covariance matrix for all blocks, makes it a more suitable imputation model. First, as discussed in Section 1.1, it is important that the imputation model be as objective and general as possible; unnecessary constraints should be avoided whenever possible. This certainly applies to avoiding constraints on the cell variances in the GLoM. Indeed, the desire to avoid such constraints is behind the generalizations of the GLoM proposed by Liu and Rubin (1998) and by Barnard, McCulloch, and Meng (2000). Constraining the cell variances in a GLoM to be the same can, for example, lead to dramatic overcoverage of multiple imputation confidence intervals (Barnard 1995, sec. 6.6.2). Thus, constraining variance elements to be equal across blocks should be a method of last resort to be used only when the data are not rich enough to estimate block-specific covariance matrices. If the sparsity of a particular dataset necessitates constraining some or all of the variance elements to be equal across blocks (to reduce the number of parameters to be estimated), then it is possible to incorporate these constraints into the BGLoM, as is noted in the first paragraph of this section. However, such constraints on the variances lead to computational issues, which is the second reason we choose to assume that the covariance matrix varies across blocks. More specifically, because different blocks contain different sets of relevant continuous variables, attempting to constrain variance elements to be equal across blocks can lead to messy, and often intractable calculations. One reasonably straightforward strategy for implementing the assumption that the entire covariance matrix (for all the pure and constructed continuous variables) is constant across all blocks entails treating the values of continuous variables as missing data in blocks where those variables are not relevant (e.g., values of 1995 earnings in cells corresponding to no earnings in 1995). Heeringa, Little, and Raghunathan (2002), for example, adopt this approach and impute the irrelevant values to have a common covariance matrix (see also Little and Su 1987; Little and Raghunathan 1997). However, treating the variables that are not relevant in a given block as missing data raises concerns of computational efficiency, because it can seriously degrade the rates of convergence of the EM and data augmentation algorithms.

2.5 Generating Imputations Under the Blocked General Location Model

The BGLoM can be used as an imputation model for datasets with partially missing variables of all three types provided that the model's distributional assumptions are appropriate for the particular dataset of interest and that the missing-data mechanism can be assumed to be ignorable.

To allow for missing data, we dichotomize the (combined) categorical and (combined) continuous variables into missing and observed groups for each individual. In particular, let $w_i = (w_i^{\text{obs}}, w_i^{\text{mis}})$, where w_i^{obs} and w_i^{mis} represent the observed and missing components of w_i, respectively. The situation is somewhat more complicated for the continuous variables, because we may not know which components of z_i are relevant. This is the case because some of the categorical variables that determine block membership, and thus which continuous variables are relevant, may be missing. Thus we let $z_i^{\text{rel}} = \{z_i^{\text{obs}}, z_i^{\text{mis}}(b)\}$, where $z_i^{\text{mis}}(b)$ represents the components of z_i that are missing if $\mathcal{B}(w_i',) = b$.

Because we assume that the missing-data mechanism is ignorable, the observed-data log-likelihood is

$$L(\theta|Y^{\text{obs}}) = \int \cdots \int L(\theta|Y) dz_1^{\text{mis}}\{\mathcal{B}(w_1',)\} \cdots dz_n^{\text{mis}}\{\mathcal{B}(w_n',)\} \cdot dw_1^{\text{mis}} \cdots dw_n^{\text{mis}}, \qquad (6)$$

where $Y^{\text{obs}} = \{(w_i^{\text{obs}}, z_i^{\text{obs}}), i = 1, \ldots, n\}$.

Imputations of the dataset of interest can be generated under a BGLoM imputation model using a variety of methods, two of which are described here. One potential method entails generating imputed datasets from the BGLoM with parameters set equal to their estimates, usually the (observed-data) maximum likelihood estimates (MLEs). An EM algorithm developed for the BGLoM with missing data can be used to calculate the MLEs of the model parameters, given the observed data [i.e., the value of θ that maximizes (6)]. Then imputed values for the missing data can be generated from the BGLoM at the MLEs using straightforward conditional simulation techniques made possible by the model's convenient conditional formulation.

A second method, which is Bayesian, uses the prior distribution on the BGLoM parameters given in (5) and requires simulating the joint posterior distribution of the unobserved data and the BGLoM parameters, given the observed data. We suggest using Markov chain Monte Carlo techniques or, more specifically, a data augmentation algorithm developed for the BGLoM with missing data. Imputed values for the missing data are sampled from the Markov chain (or chains) simulated by the data augmentation algorithm once the chain has achieved stationarity.

To implement these two methods of generating imputations, we develop EM and data augmentation algorithms for the BGLoM with missing data. Details of the algorithms are presented in Appendixes A and B.

3. NUMERICAL RESULTS

3.1 Available Software for Multiple Imputation Under the Blocked General Location Model

We implemented the EM and data augmentation algorithms for the BGLoM with missing data in S-PLUS and R, with certain common subfunctions coded in C; an R package containing these implementations is available on request from the first author. In this section, we apply these routines first in a simulation study designed to investigate the properties of various parameter estimates and then in an analysis of the MMLWS subset data.

3.2 A Simulation Study

We used the R implementation of the BGLoM algorithms to perform a two-stage simulation experiment. In both stages of this experiment, we generated 1,000 datasets with $n = 250$ and with three (pure) continuous variables, two semicontinuous variables (with point masses at zero), and one (pure) trichotomous categorical variable, doing so from a known BGLoM; we then made these datasets incomplete using a missing-at-random mechanism (Rubin 1976).

The first stage of the simulation experiment was designed to investigate the frequentist properties of estimates of the (known) BGLoM parameters. In this part of the experiment, we applied the EM and data augmentation algorithms for the BGLoM to each of the 1,000 incomplete datasets. When sampling from the posterior distribution with the data augmentation algorithm, we use the prior distribution in (5) with $\alpha_c = 1 + (.05)n/12$ for $c = 1, \ldots, 12$. Our choice of the hyperparameter vector was motivated by a desire to have a slightly informative prior distribution for the cell probabilities; relative to a flat prior distribution, our specification effectively divides the equivalent of 5% of the data evenly among the 12 cells. For each dataset, maximum likelihood estimates (MLEs) produced by the EM algorithm and 100 parameter draws generated by the data augmentation algorithm were retained for each BGLoM parameter. We collected the 100 draws of each parameter from a single data augmentation chain, started from the MLEs, by sampling every third draw beginning after a burn-in period of 150 iterations. We made the sampling decision by examining autocorrelation plots for each parameter from a number of preliminary runs, all of which suggested that lag-three draws were effectively uncorrelated; we determined the burn-in period by computing the R-hat statistic (Gelman and Rubin 1992) for various burn-in period lengths using three data augmentation chains begun from overdispersed starting values.

For each BGLoM parameter, we used the 1,000 MLEs and the 1,000 "means" of the retained parameter draws to assess the parameter recovery properties of the EM and data augmentation algorithms, respectively. We obtained each of the 1,000 means for a given parameter by first applying the relevant "normalizing" transformation to the 100 retained draws, then taking the mean of the transformed draws, and finally applying the inverse transformation to that mean; the transformations were $\log(\pi/(1-\pi))$, μ, $\log(\sigma^2)$, and $\log((\rho+1)/(\rho-1))$ for the cell probabilities, within-cell means, within-block variances, and within-block correlations, respectively. In addition, we used 94% equitailed empirical credible intervals for each BGLoM parameter to further investigate the performance of the data augmentation algorithm, as well as the impact of the prior distributions used in the algorithm. For a particular parameter, we computed the 1,000 94% intervals by taking the range of the middle 94 (out of the 100) draws of that parameter for each of the 1000 datasets. The reason for using an even-numbered coverage, as opposed to, say, 95%, was that an equitailed interval was then constructed. The results from the first simulation experiment, summarized in Table 2 and Figure 2, were encouraging. As can be seen in the figure, the coverage frequencies of the credible intervals for the (known) BGLoM parameters varied tightly around their selected credible levels [see Fig. 2(a)], and the priors used in the data augmentation algorithm did not

Table 2. Simulation Stage 1 Results for a Randomly Selected Subset of the BGLoM Parameters

Simulation criteria[b]	*Parameter*[a]			
	π_6	$\mu_{3,C2}$	$\sigma^2_{4(C2,C2)}$	$\rho_{2(P2,C1)}$
True parameter value	.065	12.00	1,000.00	−.018
Average MLE	.062	13.73	1,000.86	−.038
Estimated bias of MLE	−.003	1.73	.86	−.020
Average posterior mean	.061	13.82	1,249.26	−.017
Estimated bias of posterior mean	−.004	1.82	249.26	.001
Average lower bound	.028	−7.00	785.48	−.347
Average upper bound	.121	36.94	2,134.75	.343
Coverage of the true parameter	1.00	.91	.95	1.00
Coverage of the MLE	.96	1.00	.96	1.00

[a] The four selected BGLoM parameters are: (1) π_6, the cell 6 probability; (2) $\mu_{3,C2}$, the mean of constructed continuous variable 2 in cell 3; (3) $\sigma^2_{4(C2,C2)}$, the variance of constructed continuous variable 2 in block 4; and (4) $\rho_{2(P2,C1)}$, the correlation of pure continuous variable 2 and constructed continuous variable 1 in block 2.

[b] In the results for the first stage of the simulation experiment, *average* refers to the average across the 1,000 simulated datasets, the MLE is computed with the EM algorithm, the *estimated bias* of an estimate is the difference between the average estimate across the 1,000 simulated datasets and the true value of the parameter, the *posterior mean* is the (retransformed) mean of 100 transformed ("uncorrelated", post–burn-in) draws from the posterior distribution obtained using the data augmentation algorithm, the *lower bound* and *upper bound* are the bounds of the 94% empirical equitailed credible interval computed from 100 ("uncorrelated" post–burn-in) draws, and the *coverage of the true parameter* and the *coverage of the MLE* are the proportion of the 1,000 94% empirical credible intervals that contain the true parameter value and the maximum likelihood estimate, respectively.

seem to have a large impact, except in the case of very small cell probabilities (e.g., the cell probability corresponding to the outlying point in both cell probability boxplots). Additionally, as can be seen by the small size of the estimated biases of the MLEs and posterior means in rows 3 and 5 of the table, both algorithms did a good job recovering the BGLoM parameters.

The second stage of the simulation experiment was designed to investigate how the MI analysis approach with a BGLoM as the imputation model performs relative to the complete-case analysis approach and relative to the MI analysis approach with a GLoM as the imputation model. In this third and last approach, we use the standard GLoM (Olkin and Tate 1961); because this model makes no allowance for semicontinuous variables, we treat them simply as continuous variables. Regardless of which of the three approaches was used to handle the missing data, we analyzed the simulated datasets by fitting a two-part model to each of the semicontinuous variables: for each semicontinuous variable, we used a logistic regression to model the probability that the semicontinuous variable takes on its point mass value, and used a Gaussian linear regression to model the semicontinuous variable given that it does not take on its point mass value. In all four regressions (two regressions for each semicontinuous variable), we included an intercept and the same five covariates: indicator variables for two levels of the categorical variable, denoted by $I\{W_1 = 2\}$ and $I\{W_1 = 3\}$, and the three continuous variables, denoted by Z_1, Z_2, and Z_3. For each of the 1,000 simulated datasets, we

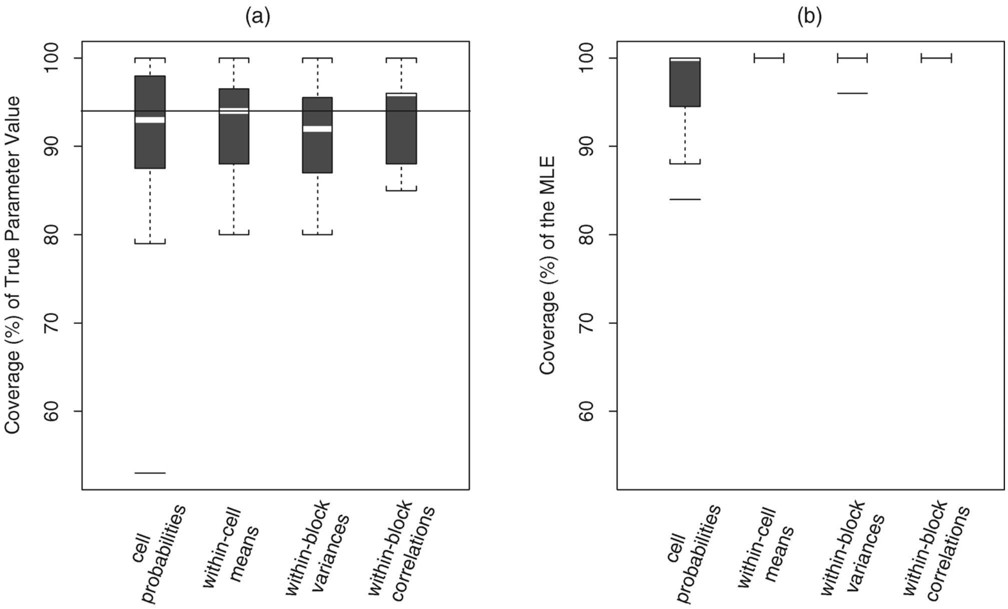

Figure 2. Coverage Frequencies for 94% Credible Intervals, by Parameter Group. These boxplots summarize the distribution of coverage of (a) the true parameter values and (b) the MLEs, obtained by simulation for the four types of parameters in the BGLoM. The horizontal line in (a) corresponds to the target coverage of 94%, and lines below the boxplots represent outlying coverage frequencies.

Table 3. Simulation Stage 2 Results

	Mean distance to gold standard estimate			*Mean length of intervals*			*% of intervals containing gold standard estimate*		
Covariates	*Case-deletion*	*GLoM*	*BGLoM*	*Case-deletion*	*GLoM*	*BGLoM*	*Case-deletion*	*GLoM*	*BGLoM*
Logistic-linear regression for first semicontinuous variable									
Intercept	.06	.14	.07	1.38	1.01	1.04	100	100	100
$I\{W_1 = 2\}$	.28	.14	.38	2.06	1.54	1.66	100	100	100
$I\{W_1 = 3\}$	.49	.16	.14	2.27	1.69	1.59	100	100	100
Z_1	.00	.00	.00	.01	.01	.01	100	100	100
Z_2	.00	.00	.00	.01	.01	.01	100	100	100
Z_3	.00	.00	.00	.01	.01	.01	100	100	100
Gaussian linear regression for first semicontinuous variable									
Intercept	20.71	16.02	2.72	49.28	42.33	38.33	100	85	100
$I\{W_1 = 2\}$	21.73	6.19	4.38	73.51	60.14	51.51	100	100	100
$I\{W_1 = 3\}$	13.26	12.42	2.89	62.76	56.83	42.61	100	100	100
Z_1	.01	.13	.09	.41	.40	.26	100	91	100
Z_2	.00	.09	.03	.34	.28	.18	100	99	100
Z_3	.12	.02	.09	.47	.38	.27	100	100	100
Logistic-linear regression for second semicontinuous variable									
Intercept	.29	.07	.11	1.37	1.01	1.05	100	100	100
$I\{W_1 = 2\}$	.33	.68	.51	2.09	1.65	1.53	100	79	100
$I\{W_1 = 3\}$	.52	.16	.26	2.11	1.66	1.47	100	100	100
Z_1	.00	.00	.00	.01	.01	.01	100	96	100
Z_2	.00	.00	.00	.01	.01	.01	100	100	100
Z_3	.00	.00	.00	.01	.01	.01	100	100	100
Gaussian linear regression for second semicontinuous variable									
Intercept	4.77	5.56	5.13	42.98	29.02	26.61	100	100	100
$I\{W_1 = 2\}$	4.60	2.57	3.02	60.71	44.11	37.90	100	100	100
$I\{W_1 = 3\}$	10.14	11.35	7.73	83.72	59.87	51.52	100	100	100
Z_1	.14	.05	.05	.51	.32	.28	100	100	100
Z_2	.17	.05	.01	.43	.27	.25	100	100	100
Z_3	.06	.04	.03	.47	.33	.27	100	100	100

performed the four regressions four different times, once on the complete version of each dataset (before missing-at-random missingness was imposed), and three times on the incomplete version of each dataset using three different approaches to handling incomplete data. First, we analyzed the complete version of each dataset by simply fitting the four regressions; we used this "gold standard" analysis later to assess the performance of the three incomplete-data analysis approaches. Second, we analyzed the incomplete version of each dataset using a complete-case analysis approach; we discarded cases (e.g., individuals) with any missing data and fit the four regression models to only the completely observed cases. Third and fourth, we analyzed the incomplete version of each dataset using the MI approach with two different imputation models, first the standard GLoM and then the BGLoM. In both of these MI analysis approaches, we imputed five complete datasets from five parallel data augmentation chains, each started from the relevant MLEs; these chains were allowed burn-in periods of 400 and 1,000 iterations for the standard GLoM and BGLoM imputation models, respectively. Any negative semicontinuous variable values generated under the standard GLoM were replaced by zero, the point mass value; no such negative values were generated under the BGLoM. For the standard GLoM imputations and also the BGLoM imputations, we obtained estimates and standard errors of the various regression parameters from each of the five imputed datasets, and then applied the MI *combining rules* to the five sets of errors and estimates to produce a grand estimate and a 95% confidence interval for each parameter.

Estimates of the 24 regression coefficients were retained for all four analyses of each dataset, and confidence intervals for all 24 coefficients were retained for the three incomplete-data analyses. We used these retained estimates and confidence intervals to compare the performance of the three incomplete-data approaches. We measured performance for each coefficient by the mean (absolute) difference between each approach's 1,000 estimates and the gold standard estimate, by the mean length of each approach's 1,000 confidence intervals, and by the percentage of each approach's 1,000 confidence intervals that covered the corresponding gold standard estimate. As shown in Table 3, the BGLoM approach performed much better than the other two incomplete-data approaches. Its confidence intervals, although usually narrower than those of the other approaches, generally achieved higher coverage frequencies, and its estimates were on average closer to the gold standard estimates.

Further results and details of the simulation experiment, including the BGLoM parameter values used to generate datasets in both stages of the simulation experiment, are available from the authors on request.

3.3 Analysis of the Massachusetts Megabucks Lottery Winners Survey Subset Data

We analyzed the MMLWS subset data using a *case-deletion* approach and also the MI approach, implemented first with a standard GLoM imputation model and then with a BGLoM imputation model. The MMLWS subset data consist of the variables gender, winnings, 1992 earnings, and 1995 earnings for the 327 surveyed individuals who won large prizes (more than 22,000 1986 U.S. dollars). These variables contain 4 (1%), 0 (0%), 120 (37%), and 161 (49%) missing values, respectively, making it necessary to use a missing-data method to analyze the data.

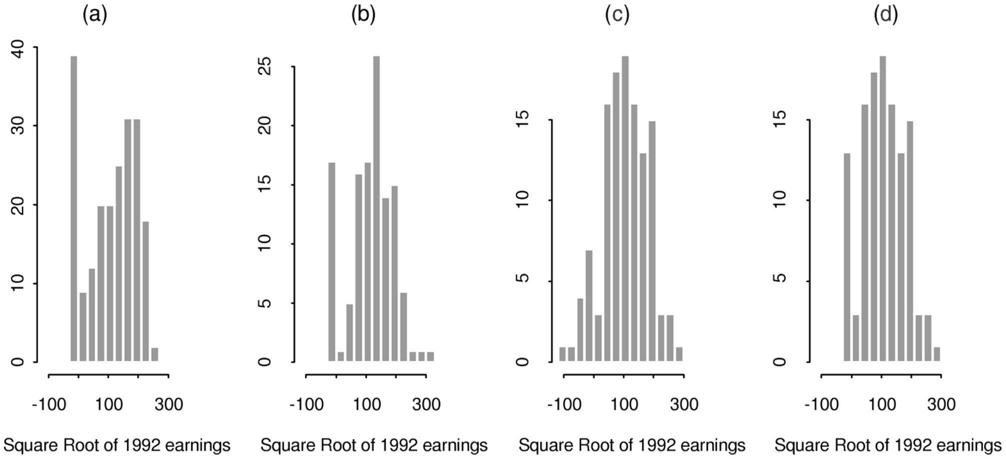

Figure 3. Histograms for (a) the 207 Observed Values of $\sqrt{1992\ \text{Earnings}}$, (b) the 120 Imputed Values of $\sqrt{1992\ \text{Earnings}}$ in One BGLoM-Imputed Dataset, (c) the 120 Imputed Values of $\sqrt{1992\ \text{Earnings}}$, Without Rounding, in One Standard GLoM-Imputed Dataset, and (d) the 120 Imputed Values of $\sqrt{1992\ \text{Earnings}}$, With Negative Imputations Rounded to the Point Mass at zero, in One Standard GLoM-Imputed Dataset. In each histogram, the 1992 earnings values are on the square root scale, and the bar located at zero corresponds to those individuals who were unemployed in 1992.

In the case-deletion approach, we removed units with missing values for any of the variables involved in a particular analysis during that analysis. Note that this approach differs from the complete-case approach of Imbens, Rubin, and Sacerdote (2001), in which units with missing values for any of the variables of interest (in any of the analyses) were removed permanently before beginning the data analysis.

We found the MI approach with either a standard GLoM or a BGLoM imputation model to be a potentially appropriate method of handling the missing data, although some concern might typically arise over the suitability of the missing-at-random assumption for the earnings variables. We generated imputations of the MMLWS subset data under the standard GLoM imputation model (with the two semicontinuous earnings variables treated as continuous) using the "mix" library provided by Joseph Schafer for S-PLUS (but ported to R). We generated imputations under the BGLoM imputation model using the aforementioned R package. For both imputation models, winnings were imputed on the log scale, and 1992 earnings and 1995 earnings were imputed on the square root scale, as suggested by graphical inspection of their observed values within the augmented contingency table cells. We generated 10 imputed datasets under both the GLoM, which took about 1 minute, and the BGLoM, which took about 15 minutes, with a 1-GHz Pentium III processor. Under both models, we generated the imputed datasets by sampling, after a burn-in period of 450 iterations, from parallel chains started at the relevant EM-produced MLEs; we based our decision to use a burn-in period of length 450 on time series plots and R-hat statistics for the respective model parameters. For both the GLoM and the BGLoM imputation models, we used the prior distribution specified in (5) when sampling from the posterior distribution; for the GLoM, both hyperparameters in α were equal to one, making the prior distribution flat, and for the BGLoM, all eight hyperparameters were equal to two. In the imputed datasets generated under the standard GLoM imputation model, all negative imputations of the (square-rooted) 1992 earnings and 1995 earnings variables were set to zero, thus making them part of the point mass for those variables. This procedure was unnecessary for the BGLoM-imputed datasets, because none of the imputed values for the two earnings variables were negative. Figure 3 compares the observed values of 1992 earnings (on the square root scale) with the imputed values from two datasets, one dataset imputed under the BGLoM and the other imputed under the GLoM. Note that the two rightmost histograms both correspond to the GLoM-imputed values; the negative imputations that appear in the first of these plots have been rounded to the point mass at zero in the second of these plots. Also note that the GLoM-imputed values (whether rounded or not) and the BGLoM-imputed values, besides differing from each other, both differ from the observed values in the leftmost plot, which suggests that the missingness mechanism for 1992 earnings is not missing completely at random.

We used the standard GLoM- and BGLoM-imputed datasets to investigate the following parameters: (1) the mean earnings of employed individuals in 1992 and in 1995; (2) the proportions of unemployed individuals in 1992 and in 1995; (3) the coefficients from a logistic regression of 1995 employment status on gender, the log of winnings, and 1992 employment status; and (4) the coefficients from a linear regression of the square root of 1995 earnings on gender, the log of winnings, and the square root of 1992 earnings (left as a semicontinuous variable), for only those individuals who were employed in 1995. For the standard GLoM imputations and also the BGLoM imputations, we obtained estimates and standard errors of these parameters from each of the 10 imputed datasets, and then applied the MI *combining rules* to the 10 sets of errors and estimates to produce a grand estimate and a 95% confidence interval for each parameter. For the 1992 and 1995 proportions, we performed the combining of estimates and errors on the logit scale to make the normal assumptions of the combining rules more appropriate, and, then retransformed the resulting estimate and confidence interval upper and lower bounds to the original scale using the inverse logit transformation.

The estimates and confidence intervals produced by the case-deletion analysis and the standard GLoM and BGLoM MI

Table 4. Results of Analyzing the MMLWS Subset Data

	Case-deletion		*GLoM*		*BGLoM*	
Parameter	*Estimate*	*95% Conf. interval*	*Estimate*	*95% Conf. interval*	*Estimate*	*95% Conf. interval*
Mean 1992 earnings for those employed	23,121	(20,643, 25,600)	20,958	(18,635, 23,281)	28,622	(22,909, 34,335)
Mean 1995 earnings for those employed	22,116	(19,385, 24,847)	20,053	(17,441, 22,664)	29,341	(25,871, 32,810)
Proportion unemployed in 1992	.19	(.14, .24)	.15	(.11, .20)	.18	(.11, .24)
Proportion unemployed in 1995	.12	(.07, .17)	.09	(.05, .14)	.18	(.09, .27)
Logistic regression: Intercept	5.66	(−2.53, 13.85)	4.01	(−5.21, 13.23)	7.29	(−.72, 15.29)
Logistic regression: Gender coefficient	.34	(−.83, 1.51)	.30	(−.87, 1.47)	.53	(−1.52, 2.59)
Logistic regression: 1992 employ. status coefficient	2.97	(1.72, 4.21)	2.97	(1.88, 4.07)	2.05	(.82, 3.27)
Logistic regression: Log(winnings) coefficient	−.40	(−1.00, .19)	−.27	(−.94, .39)	−.55	(−1.13, .03)
Linear regression: Intercept	99.23	(−25.81, 224.27)	98.45	(−31.26, 228.17)	40.84	(−58.64, 140.31)
Linear regression: Gender coefficient	3.46	(−12.76, 19.68)	2.89	(−16.24, 22.01)	−9.13	(−23.26, 5.01)
Linear regression: $\sqrt{1992 \text{ earnings}}$ coefficient	.56	(.45, .66)	.61	(.50, .72)	.65	(.53, .78)
Linear regression: Log(winnings) coefficient	−2.78	(−12.10, 6.53)	−3.43	(−13.07, 6.20)	2.56	(−4.81, 9.92)

NOTE: The table compares the estimates and 95% confidence intervals resulting from various case-deletion analyses and analyses of the MI datasets imputed under the GLoM and BGLoM. There are notable differences, particularly for mean earnings for those employed in 1992 and 1995 and for the proportion employed in 1992 and 1995.

analyses are presented in Table 4. To investigate the sensitivity of the two MI analyses to the particular imputed datasets, we replicated both the GLoM and BGLoM analyses 3 more times, using 10 different imputed datasets each time; the results from these replications were quite similar to those from the single replication summarized in Table 4. As can be seen in this table, the same regression coefficients in the linear and the logistic regression are significant regardless of which of the three analysis approaches was used, but the estimates produced under the three methods differ considerably for many coefficients. This latter fact is also true for the estimates of the marginal parameters in the first four rows of the table; in fact, for some of these parameters (e.g., the mean of 1995 earnings), the GLoM and BGLoM confidence intervals do not even overlap.

Overall, the MMLWS subset data analysis illustrates that the BGLoM's more reasonable model assumptions for semicontinuous variables (relative to those of the GLoM) can have a substantive impact on the resulting statistical analysis. The results of this MMLWS subset data analysis and of the simulation experiment suggest that under reasonable conditions, the BGLoM and the accompanying EM and data augmentation algorithms proposed here are a viable means of implementing the MI approach for datasets with partially missing semicontinuous, continuous, and categorical variables.

4. DISCUSSION

The strong relationship between the BGLoM and the standard GLoM means that many of the strengths and weaknesses of the standard GLoM apply equally to the BGLoM. Here we outline some of these similarities, but also emphasize several salient differences.

Computationally, the EM and data augmentation algorithms for fitting the BGLoM with missing data can be formulated through simple and elegant modifications to the corresponding algorithms (Little and Schluchter 1985; Schafer 1997) for the standard GLoM with missing data. More specifically, the constant within-block covariance assumption in the BGLoM means that operations performed for the dataset as a whole in the standard GLoM algorithms are merely performed within blocks in the corresponding BGLoM algorithms.

As with the GLoM, the model specification of the BGLoM is not always appropriate. For example, the assumption of a multinomial distribution for the combined categorical variables is not ideal if the dataset has categorical variables of the ordinal variety, because the information contained in the ordering of the variable levels is lost. The assumptions of multivariate normality and constant within-block covariance matrices can also be problematic. Fortunately, though, these two particular constraints can be relaxed. For instance, the BGLoM setting could be imposed on Liu and Rubin's (1998) ellipsoidally symmetric extended GLoM, which allows for different but proportional within-cell covariance matrices and within-cell multivariate t distributions for the continuous variables, where the degrees of freedom of these distributions vary across cells. Alternatively, it may be possible to transform the data in such a way that the distributional assumptions become more appropriate. For example, suppose (as is often the case) that a given semicontinuous variable's values are restricted to the positive part of the real line and have a skewed distribution, which brings the assumption of within-cell normality into question. Then, to make this assumption more appropriate for the corresponding constructed continuous variable, the log transformation could be applied to all continuous values of the semicontinuous variable. We emphasize that this solution is much easier with the BGLoM than with the standard GLoM, at least in the common case of a nonnegative variable with a point mass at zero, because the log transformation cannot be applied unless the point mass is removed.

One additional concern arises with the BGLoM because its parameter space is larger than that of the GLoM. The assumption that the covariance matrices of the constructed continuous variables are constant only within blocks, and not across all cells, can greatly increase the dimension of the free parameters, especially when there are either many blocks or many pure continuous variables. Although this added flexibility in the model is sometimes more appropriate than assuming constant variance across all cells, it requires larger datasets for acceptable statistical inference. The larger parameter space can also lead to computational difficulties. Not only is the efficiency of the EM and data augmentation algorithms reduced, but also the algorithms may not converge in cases where the dataset is too sparse (i.e., has a small number of units and/or significant amounts of missing data). Both the computational and the more important inferential difficulties can, in principle, be mitigated by placing restrictions on the BGLoM parameters. For example, the cell probabilities and within-cell means could be forced to satisfy a log-linear model and a linear model, respectively, as has been done for the standard GLoM (Krzanowski 1982; Little and Schluchter 1985). Alternatively, certain variances or correlations could be constrained to be equal across blocks, or a common prior distribution could be used on the scalar variances or correlations to borrow strength across blocks.

Despite these concerns, both the MMLWS subset data analysis and the simulation experiment described in Section 3 suggest that the BGLoM and our fitting algorithms perform well in practice, at least in situations where the dataset is sufficiently nonsparse (relative to the number of BGLoM parameters) and where the BGLoM is deemed to be an appropriate imputation model. The MMLWS application in particular illustrates the real inferential advantage of the BGLoM for statistical analysis of datasets with partially observed categorical, continuous, and semicontinuous variables.

APPENDIX A: THE EM ALGORITHM FOR THE BLOCKED GENERAL LOCATION MODEL WITH INCOMPLETE DATA

Here and in Appendix B, we continue to avoid constraints on the elements of the within-block covariance matrices in the interest of simplicity. In general, a patterned within-block covariance matrix (e.g., Σ_b diagonal) can be accommodated by an (often simple) modification to the M-step of the EM algorithm (or to the P-step of the data augmentation algorithm). However, if we wish to constrain certain elements of two or more within-block covariance matrices to be equal, it may be necessary to alter the data augmentation scheme to maintain a simple M-step (or P-step). Consider, for example, two blocks where the set of relevant continuous variables for the first block is a subset of the relevant continuous variables for the second block. We can easily accommodate the constraint that the corresponding elements of the two within-block covariance matrices are equal; this could be done by treating the variables that are relevant in the second block, but not in the first block, as missing data in the first block. The method of Heeringa, Little, and Raghunathan (2002) is an example of this strategy in which all of the corresponding elements of the within-block covariance matrices are constrained to be equal and all of the variables that are not relevant are treated as missing data. Addressing the numerous possible variations of this strategy would be notationally tedious, and thus we avoid doing so here.

A.1 Overview

The EM algorithm is a well-known iterative method for computing the modes of marginal distributions such as $p(\theta|Y^{\text{obs}}) \propto L(\theta|Y^{\text{obs}})p(\theta)$, where $L(\theta|Y^{\text{obs}})$ is the marginal distribution given in (6). Here we focus on maximum likelihood estimation by taking $p(\theta) \propto 1$. Beginning with starting value $\theta^{(0)}$, EM proceeds by computing

$$\theta^{(t+1)} = \operatorname*{arg\,max}_{\theta} \mathrm{E}\{\log L(\theta|Y)p(\theta)|Y^{\text{obs}}, \theta^{(t)}\} \tag{A.1}$$

for $t = 0, 1, \ldots$. Computing the expectation in (A.1) is known as the E-step of EM, and computing the maximization in (A.1) is known as the M-step of EM. This procedure, which often entails only straightforward computation, is guaranteed to increase the log posterior at each iteration.

Beginning with the E-step, we note that $\log L(\theta|Y)$, as a function of Y, is linear in a set of sufficient statistics; thus the expectation in (A.1) can be computed by calculating the expectation of these sufficient statistics. Here the sufficient statistics are the within-block sum of squares of the relevant continuous variables, the within-cell sum of the relevant continuous variables, and the cell counts. These can be represented formally by

$$T_{1b} = (Z^b)^\top \operatorname{diag}\{U^b(U^b)^\top\} Z^b \tag{A.2}$$

and

$$T_{2b} = (U^b)^\top Z^b, \tag{A.3}$$

for $b = 1, \ldots, B$, and

$$T_3 = U^\top U, \tag{A.4}$$

respectively, where diag(M) is a diagonal matrix with diagonal elements equal to those of M, Z^b is the $n \times q_b^{\text{rel}}$ submatrix of Z with columns corresponding to the continuous variables relevant to block b and rows corresponding to the individuals, and U^b is the $n \times C_0$ submatrix of U with columns corresponding to the cells within block b and rows again corresponding to the individuals, with $C_0 = \prod_{j=1}^{p_0} c_j$ the number of cells within each block. (Note that the columns of U^b indicate the cell within block b, if any, to which each individual belongs.) Thus the E-step consists of computing

$$\hat{T}_{1b}^{(t+1)} = \mathrm{E}\big[(Z^b)^\top \operatorname{diag}\{U^b(U^b)^\top\} Z^b | Y^{\text{obs}}, \theta^{(t)}\big]$$

and

$$\hat{T}_{2b}^{(t+1)} = \mathrm{E}\big\{(U^b)^\top Z^b | Y^{\text{obs}}, \theta^{(t)}\big\},$$

for $b = 1, \ldots, B$, and

$$\hat{T}_3^{(t+1)} = \mathrm{E}\big(U^\top U | Y^{\text{obs}}, \theta^{(t)}\big).$$

Details are given in Appendix A.2.

We now turn to the other step of the EM algorithm, the M-step, which is simple for the BGLoM because it belongs to an exponential family. More specifically, the maximization in (A.1) is accomplished by computing

$$\pi^{(t+1)} = \frac{1}{n}\hat{T}_3^{(t+1)}\mathbf{1}$$

and, for $b = 1, \ldots, B$,

$$M_b^{(t+1)} = \big(\hat{T}_{3b}^{(t+1)}\big)^{-1}\hat{T}_{2b}^{(t+1)}$$

and

$$\Sigma_b^{(t+1)} = \frac{1}{\operatorname{trace}(\hat{T}_{3b}^{(t+1)})}\big\{\hat{T}_{1b}^{(t+1)} - \big(\hat{T}_{2b}^{(t+1)}\big)^\top\big(\hat{T}_{3b}^{(t+1)}\big)^{-1}\hat{T}_{2b}^{(t+1)}\big\},$$

where $\mathbf{1}$ is a column vector of 1s, $M_b^{(t+1)}$ is a $(C_0 \times q_b^{\text{rel}})$ matrix with rows equal to the cell means of relevant variables in block b [i.e., $\mu_c^{(t+1)}$ for $c \in \mathcal{C}(, b)$], and $\hat{T}_{3b}^{(t+1)}$ is the $(C_0 \times C_0)$ submatrix of $\hat{T}_3^{(t+1)}$ with rows and columns corresponding to the cells within block b.

A.2 Computations for the E-step

Here we provide the details of the generic computation of $\hat{T}_{1b}^{(t+1)}$, $\hat{T}_{2b}^{(t+1)}$, and $\hat{T}_3^{(t+1)}$. We begin with $\hat{T}_3^{(t+1)}$, which is by construction a diagonal matrix with the expected cell counts as diagonal elements. Thus, we need compute only

$$\mathrm{E}(u_{\cdot c}|Y^{\text{obs}}, \theta) = \sum_{i=1}^{n} \mathrm{E}(u_{ic}|Y^{\text{obs}}, \theta), \tag{A.5}$$

evaluated at $\theta = \theta^{(t)}$, for each c. The set of possible cell memberships for individual i is determined by w_i^{obs} and denoted by $\mathcal{C}^{\text{obs}}(w_i^{\text{obs}},)$.

Thus, we can compute (A.5) using

$$\begin{aligned}
&\mathrm{E}(u_{ic}|Y^{\mathrm{obs}},\theta)\\
&\quad = I_{\{c\in\mathcal{C}^{\mathrm{obs}}(w_i^{\mathrm{obs}},)\}}\pi_c\big|\Sigma^{\mathrm{obs}}_{\mathcal{B}(,c)i}\big|^{-\frac{1}{2}}\\
&\qquad \times \exp\Big\{-\frac{1}{2}\big(z_i^{\mathrm{obs}}-\mu_{ci}^{\mathrm{obs}}\big)^{\top}\big(\Sigma^{\mathrm{obs}}_{\mathcal{B}(,c)i}\big)^{-1}\big(z_i^{\mathrm{obs}}-\mu_{ci}^{\mathrm{obs}}\big)\Big\}\\
&\qquad \times \Bigg(\sum_{c'\in\mathcal{C}^{\mathrm{obs}}(w_i^{\mathrm{obs}},)} \pi_{c'}\big|\Sigma^{\mathrm{obs}}_{\mathcal{B}(,c')i}\big|^{-\frac{1}{2}}\\
&\qquad \times \exp\Big\{-\frac{1}{2}\big(z_i^{\mathrm{obs}}-\mu_{c'i}^{\mathrm{obs}}\big)^{\top}\big(\Sigma^{\mathrm{obs}}_{\mathcal{B}(,c')i}\big)^{-1}\big(z_i^{\mathrm{obs}}-\mu_{c'i}^{\mathrm{obs}}\big)\Big\}\Bigg)^{-1},
\end{aligned} \tag{A.6}$$

where $\Sigma_{bi}^{\mathrm{obs}}$ is the submatrix of Σ_b with rows and columns corresponding to the components of z_i^{obs}, and μ_{ci}^{obs} is the subvector of μ_c with components corresponding to those of z_i^{obs}. If $z_i^{\mathrm{obs}}=\emptyset$, then we replace (A.6) with $\mathrm{E}(u_{ic}|Y^{\mathrm{obs}},\theta) = I_{\{c\in\mathcal{C}^{\mathrm{obs}}(w_i^{\mathrm{obs}},)\}}\pi_c/\sum_{c'\in\mathcal{C}^{\mathrm{obs}}(w_i^{\mathrm{obs}},)}\pi_{c'}$.

Next, we address the calculation of $\hat{T}_{2b}^{(t+1)}$, which is computed elementwise for each b. More specifically, we compute the conditional expectation of the sum of relevant continuous variable j over individuals in cell $c\in\mathcal{C}(,b)$, using

$$\begin{aligned}
\big(\hat{T}_{2b}^{(t+1)}\big)_{cj} &= \mathrm{E}\Bigg(\sum_{i=1}^{n} u_{ic}z_{ij}|Y^{\mathrm{obs}},\theta^{(t)}\Bigg)\\
&= \sum_{i=1}^{n}\mathrm{E}\big\{u_{ic}\mathrm{E}\big(z_{ij}|Y^{\mathrm{obs}},\theta^{(t)},u_{ic}\big)|Y^{\mathrm{obs}},\theta^{(t)}\big\}\\
&= \sum_{i=1}^{n}\mathrm{E}\big(u_{ic}|Y^{\mathrm{obs}},\theta^{(t)}\big)\mathrm{E}\big(z_{ij}|Y^{\mathrm{obs}},\theta^{(t)},u_{ic}=1\big),
\end{aligned} \tag{A.7}$$

where $c\in\mathcal{C}(,b)$, $Z_j=(z_{1j},\ldots,z_{nj})^{\top}$ is relevant in block b, and expression (A.7) follows because u_{ic} is an indicator variable. [We emphasize that the subscript j of $(\hat{T}_{2b}^{(t+1)})_{cj}$ does not necessarily correspond to the column j of $\hat{T}_{2b}^{(t+1)}$ but instead refers to continuous variable j.] The conditional expectation of u_{ic} in (A.7) is given in (A.5); thus, we need compute only $\mathrm{E}(z_{ij}|Y^{\mathrm{obs}},\theta^{(t)},u_{ic}=1)$ for each i. If z_{ij} is not observed, then this expectation is computed via a multivariate regression on z_i^{obs} that assumes $z_i^{\mathrm{rel}}\sim\mathrm{N}(\mu_c,\Sigma_{\mathcal{B}(,c)})$ because $u_{ic}=1$. (Here we use the sweep operator to describe the calculations required for this regression; readers unfamiliar with the sweep operator might wish to consult Schafer 1997, sec. 5.2.4.) Specifically, we construct

$$\xi_b^{(t)} = \begin{pmatrix} \Sigma_b^{(t)} & (M_b^{(t)})^{\top} \\ M_b^{(t)} & I_b \end{pmatrix}, \tag{A.8}$$

where I_b is a $(q_b^{\mathrm{rel}}\times q_b^{\mathrm{rel}})$ identity matrix, and then sweep $\xi_b^{(t)}$ on the positions in $\Sigma_b^{(t)}$ corresponding to the elements of z_i^{obs}. We denote the upper-left and lower-left submatrices that result from this sweep operation by $\Sigma^{\star}$ and $M^{\star}$, respectively. (Note that the dependencies on individual, block in which the individual is assumed to be, and iteration are suppressed in the notation used for these two submatrices and their elements; however, this should not obscure the fact that the sweep operation is in general repeated many times in each E-step.) Finally, we can compute

$$\mathrm{E}\big(z_{ij}|Y^{\mathrm{obs}},\theta,u_{ic}=1\big) = \begin{cases} z_{ij} & \text{if } z_{ij}\in z_i^{\mathrm{obs}} \\ m_{cj}^{\star} + \displaystyle\sum_{k:z_{ik}\in z_i^{\mathrm{obs}}} \sigma_{jk}^{\star}z_{ik} & \text{if } z_{ij}\in z_i^{\mathrm{mis}}\{\mathcal{B}(,c)\}, \end{cases} \tag{A.9}$$

where $m_{cj}^{\star}$ is the element of $M^{\star}$ corresponding to cell c and variable Z_j, and $\sigma_{jk}^{\star}$ is the element of $\Sigma^{\star}$ corresponding to variables Z_j and Z_k. We emphasize again that the subscripts on $m_{cj}^{\star}$ and $\sigma_{jk}^{\star}$ refer to the corresponding continuous variables rather than to locations in $M^{\star}$ and $\Sigma^{\star}$, which can differ from the variable number because some continuous variables may not be relevant.

We turn from $\hat{T}_{2b}^{(t+1)}$ to $\hat{T}_{1b}^{(t+1)}$, which is also computed elementwise for each b. Specifically, we compute the conditional expectation of the sum (over individuals in block b) of the product of relevant continuous variables j and k,

$$\begin{aligned}
\big(\hat{T}_{1b}^{(t+1)}\big)_{jk} &= \mathrm{E}\Bigg(\sum_{i=1}^{n}\sum_{c\in\mathcal{C}(,b)} u_{ic}z_{ij}z_{ik}\Big|Y^{\mathrm{obs}},\theta^{(t)}\Bigg)\\
&= \sum_{i=1}^{n}\sum_{c\in\mathcal{C}(,b)} \mathrm{E}\big(z_{ij}z_{ik}|Y^{\mathrm{obs}},\theta^{(t)},u_{ic}=1\big) && \text{(A.10)}\\
&\qquad \times \mathrm{E}\big(u_{ic}|Y^{\mathrm{obs}},\theta^{(t)}\big), && \text{(A.11)}
\end{aligned}$$

where (A.10)–(A.11) follows exactly as (A.7) and the j and k subscripts on $(T_{1b}^{(t+1)})_{jk}$ again refer to the corresponding relevant continuous variables, not element locations. Because the expectation in (A.11) has already been calculated in (A.6), we need compute only each $\mathrm{E}(z_{ij}z_{ik}|Y^{\mathrm{obs}},\theta^{(t)},u_{ic}=1)$. This is done using

$$\mathrm{E}(z_{ij}z_{ik}) = \begin{cases} z_{ij}z_{ik} & \text{if } z_{ij},z_{ik}\in z_i^{\mathrm{obs}} \\ \mathrm{E}(z_{ij})z_{ik} & \text{if } z_{ij}\in z_i^{\mathrm{mis}}\{\mathcal{B}(,c)\} \text{ and } z_{ik}\in z_i^{\mathrm{obs}} \\ z_{ij}\mathrm{E}(z_{ik}) & \text{if } z_{ij}\in z_i^{\mathrm{obs}} \text{ and } z_{ik}\in z_i^{\mathrm{mis}}\{\mathcal{B}(,c)\} \\ \mathrm{E}(z_{ij})\mathrm{E}(z_{ik})+\sigma_{jk}^{\star} & \text{if } z_{ij},z_{ik}\in z_i^{\mathrm{mis}}\{\mathcal{B}(,c)\}, \end{cases}$$

where all expectations are conditional on Y^{obs}, $\theta^{(t)}$, and $u_{ic}=1$; where Z_j and Z_k are relevant in block $\mathcal{B}(,c)$; and where $\sigma_{jk}^{\star}$ is from the output of the appropriate application of the sweep operator, as in (A.9).

This completes the calculations needed for the E-step.

APPENDIX B: DATA AUGMENTATION ALGORITHM FOR THE BLOCKED GENERAL LOCATION MODE WITH INCOMPLETE DATA

The joint posterior distribution, $p(\theta,Y|Y^{\mathrm{obs}})\propto L(\theta|Y)p(\theta)$, can be summarized by using Monte Carlo methods to obtain a sample from $p(\theta,Y|Y^{\mathrm{obs}})$. Here we use the data augmentation algorithm (Tanner and Wong 1987), an iterative algorithm that constructs a Markov chain that under mild regularity conditions converges to the joint posterior distribution (see Roberts 1996, for convergence results). More specifically, we start at $\theta^{(0)}$ and iterate as follows:

I-step: Draw $Y^{(t+1)}$ from $p(Y|Y^{\mathrm{obs}},\theta^{(t)})$.
P-step: Draw $\theta^{(t+1)}$ from $p(\theta|Y^{(t+1)})$.

For sufficiently large t_0, we can consider $\{(\theta^{(t)},X^{(t)}), t=t_0,\ldots,t_0+l\}$ to be a sample from the joint posterior distribution. Thus, one method of creating multiple imputed datasets is to select an effectively independent subset of $(Y^{(t)}, t=t_0,\ldots,t_0+l)$.

In the I-step, we use the factorization in (1) and sample first from $u_i^{(t)}\sim p(u_i|Y^{\mathrm{obs}},\theta^{(t)})$ and then from $(z_i^{\mathrm{mis}}(b))^{(t)}\sim p(z_i^{\mathrm{mis}}(b)|Y^{\mathrm{obs}}, \theta^{(t)},u_i^{(t)})$, where u_i is the row of U corresponding to individual i and $b=\mathcal{B}(,c)$, with c such that $u_{ic}^{(t)}=1$. The distribution $p(u_i|Y^{\mathrm{obs}},\theta)$ is multinomial with cell probabilities given in (A.6) and

$$\begin{aligned}
&p\big(z_i^{\mathrm{mis}}\{\mathcal{B}(,c)\}|Y^{\mathrm{obs}},\theta^{(t)},u_{ic}=1\big)\\
&\qquad = \mathrm{N}\big(\mathrm{E}\big[z_i^{\mathrm{mis}}\{\mathcal{B}(,c)\}|Y^{\mathrm{obs}},\theta^{(t)},u_{ic}=1\big],(\Sigma^{\star})^{\mathrm{mis}}\big),
\end{aligned} \tag{B.1}$$

where the components of the mean vector are given in (A.9), and $(\Sigma^\star)^{\mathrm{mis}}$ is the submatrix of $\Sigma^\star$ with rows and columns corresponding to the components of $z_i^{\mathrm{mis}}\{\mathcal{B}(,c)\}$, with $\Sigma^\star$ the output from the appropriate sweep operator described in Appendix A.2. The I-step is completed by tabulating the complete-data sufficient statistics, $T_{1b}^{(t+1)}$ and $T_{2b}^{(t+1)}$, for each b, and $T_3^{(t+1)}$, using (A.2)–(A.4).

In the P-step, we sample $\theta^{(t+1)}$ using

$$\pi^{(t+1)} \mid Y^{(t+1)} \sim \text{Dirichlet}\big(\alpha + T_3^{(t+1)}\mathbf{1}\big),$$

$$\Sigma_b^{(t+1)} \mid Y^{(t+1)} \sim \text{inverse Wishart}\Big[\text{trace}\big(T_{3b}^{(t+1)}\big) - C_0, \\ \big\{\hat{T}_{1b}^{(t+1)} - \big(\hat{T}_{2b}^{(t+1)}\big)^\top \big(\hat{T}_{3b}^{(t+1)}\big)^{-1} \hat{T}_{2b}^{(t+1)}\big\}^{-1}\Big],$$

and

$$\mu_c^{(t+1)} \mid Y^{(t+1)}, \Sigma_{\mathcal{B}(,c)}^{(t+1)} \\ \sim \mathrm{N}\Big\{\mathrm{E}\big(\mu_c | Y^{(t+1)}, \Sigma_{\mathcal{B}(,c)}^{(t+1)}\big), \frac{1}{(T_3^{(t+1)})_c} \Sigma_{\mathcal{B}(,c)}^{(t+1)}\Big\},$$

where $\mathrm{E}(\mu_c | Y^{(t+1)}, \Sigma_{\mathcal{B}(,c)}^{(t+1)})$ is the row of $(T_{3,\mathcal{B}(,c)}^{(t+1)})^{-1} T_{2,\mathcal{B}(,c)}^{(t+1)}$ corresponding to cell c and $(T_3^{(t+1)})_c$ is the diagonal element of $T_3^{(t+1)}$ corresponding to cell c. (Readers unfamiliar with the inverse Wishart distribution or how to draw from it might wish to consult Schafer 1997, pp. 150–151, 184.)

[Received February 2002. Revised January 2003.]

REFERENCES

Barnard, J. (1995), "Cross-Match Procedures for Multiple-Imputation Inference: Bayesian Theory and Frequentist Evaluation," unpublished doctoral thesis, University of Chicago, Dept. of Statistics.

Barnard, J., McCulloch, R., and Meng, X. L. (2000), "Modeling Covariance Matrices in Terms of Standard Deviations and Correlations, With Applications to Shrinkage," *Statistica Sinica*, 10, 1281–1311.

Clogg, C. C., Rubin, D. B., Schenker, N., Schultz, B., and Weidman, L. (1991), "Multiple Imputation of Industry and Occupation Codes in Census Public Use Samples Using Bayes Logistic Regression," *Journal of the American Statistical Association*, 86, 68–78.

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

Dunn, N., Manning, W. G., Morris, C. N., and Newhouse, J. P. (1983), "A Comparison of Alternative Models for the Demand for Medical Care," *Journal of Business & Economic Statistics*, 1, 115–126.

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.

Gelman, A., and Rubin, D. B. (1992), "Inference From Iterative Simulation Using Multiple Sequences" (with discussion), *Statistical Science*, 7, 457–472.

Heeringa, S. G., Little, R. J. A., and Raghunathan, T. E. (2002), "Multivariate Imputation of Coarsened Survey Data on Household Wealth," in *Survey Nonresponse*, eds. R. M. Groves, D. A. Dillman, J. L. Eltinge, and R. J. A. Little, New York: Wiley, pp. 357–371.

Herzog, T. N., and Rubin, D. B. (1983), "Using Multiple Imputations to Handle Nonresponse in Sample Surveys," in *Incomplete Data in Sample Surveys* (Vol. 2), eds. W. G. Madow, I. Olkin, and D. B. Rubin, New York: Academic Press, pp. 209–245.

Imbens, G. W., Rubin, D. B., and Sacerdote, B. (2001), "Estimating the Effect of Unearned Income on Labor Earnings, Savings, and Consumption: Evidence From a Survey of Lottery Players," *American Economic Review*, 91, 778–794.

Krzanowski, W. J. (1980), "Mixtures of Continuous and Categorical Variables in Discriminant Analysis," *Biometrics*, 36, 493–499.

——— (1982), "Mixtures of Continuous and Categorical Variables in Discriminant Analysis: A Hypothesis Testing Approach," *Biometrics*, 38, 198–206.

Little, R. J. A., and Raghunathan, T. E. (1997), "Should Imputation of Missing Data Condition on All Observed Variables?," in *Proceedings of the Section on Survey Research Methods Section, American Statistical Association*, pp. 617–622.

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

Little, R. J. A., and Schluchter, M. D. (1985), "Maximum Likelihood Estimation for Mixed Continuous and Categorical Data With Missing Values," *Biometrika*, 72, 492–512.

Little, R. J. A., and Su, H. L. (1987), "Missing-Data Adjustments for Partially-Scaled Variables," in *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 644–649.

Liu, C., and Rubin, D. B. (1998), "Ellipsoidally Symmetric Extensions of the General Location Model for Mixed Categorical and Continuous Data," *Biometrika*, 85, 673–688.

Manning, W. G., Morris, C. N., Newhouse, J. P., Orr, L. L., Dunn, N., Keeler, E. B., Leibowitz, A., Marquis, K. H., Marquis, M. S., and Phelps, C. E. (1981), "A Two-Part Model of the Demand for Medical Care: Preliminary Results From the Health Insurance Experiment," in *Health, Economics, and Health Economics*, eds. J. van der Gaag and M. Perlman, Amsterdam: North-Holland, pp. 103–104.

Meng, X. L. (1994), "Multiple-Imputation Inference With Uncongenial Sources of Input" (with discussion), *Statistical Science*, 9, 538–573.

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

Olsen, M. K., and Schafer, J. L. (2001), "A Two-Part Random-Effects Model for Semicontinuous Longitudinal Data," *Journal of the American Statistical Association*, 96, 730–745.

Roberts, G. O. (1996), "Markov Chain Concepts Related to Sampling Algorithms," in *Markov Chain Monte Carlo in Practice*, eds. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, London: Chapman & Hall, pp. 45–57.

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

——— (1980), *Handling Nonresponse in Sample Surveys by Multiple Imputation*, Washington, DC: U.S. Bureau of the Census.

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

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

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

Schenker, N., Treiman D. J., and Weidman, L. (1988), "Evaluation of Multiply-Imputed Public-Use Tapes," in *Proceedings of the Section on Survey Research Methods Section, American Statistical Association*, pp. 85–92.

——— (1993), "Analysis of Public Use Decennial Census Data With Multiply-Imputed Industry and Occupation Codes," *Journal of the Royal Statistical Society*, Ser. C, 42, 545–556.

Tanner, M. A., and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation" (with discussion), *Journal of the American Statistical Association*, 82, 528–550.