Multiple imputation in multiple classification and multiple-membership structures

Recai M. Yucel[1], Hong Ding[2], Ali Kerem Uludag[2] and Donald Tomaskovic-Devey[3]
Department of Epidemiology and Biostatistics, University at Albany, SUNY[1]
Division of Biostatistics and Epidemiology, University of Massachusetts, Amherst[2]
Department of Sociology, University of Massachusetts, Amherst[3]

Abstract

In data systems with complexities due to nested/non-nested clustering and multiple-membership, missing values present an added analytic challenge to the statistical analyses. We develop model-based multiple imputation (MI) inference which has been a popular method in the analyses of missing data. Adaptations of multivariate generalizations of the mixed-effects models are used as imputation model. These models are modified to handle multivariate responses and observational units with possibly overlapping membership of clusters that are not necessarily hierarchical. Markov Chain Monte Carlo techniques are used to simulate and draw imputations from underlying joint posterior predictive distributions. Brief discussion on handling mixture of variable types and calibration techniques for post-imputation checks will be provided. Relevant concepts on both multiple-membership and non-nested clustering are demonstrated longitudinal administrative data with panel missingness as well as arbitrary item nonresponse.

KEY WORDS: Multiple imputation, Bayesian inference, missing data, multiple membership, mixed-effects

1. Introduction

Principled missing-data techniques especially those using the multiple-imputation (MI) paradigm (Rubin, 1976) have developed significantly since 1980s. Most of these techniques rely on relatively straightforward model assumptions such as independent and identically distributed units or clustered data. These methods are available to practitioners in software packages such as SAS PROC MI (SAS Institute 2001) (for cross-sectional data) and R package pan (Schafer and Yucel 2002), MlwiN mimacro (Carpenter and Kenward 2008) (for multilevel data). Building on these well-established methods, we develop model-based MI techniques for analyzing clustered incomplete data with multiple membership and non-nestedness. Our strategy jointly models variables subject to missing values in such settings leading to multivariate extension of a multiple membership and multiple classification model as first suggested by Browne, Goldstein, and Rasbash (2001). Below we describe the example that motivated this research and we believe it is useful to illustrate multiple membership as well as multiple classification problem.

1.1 Motivating Example

Since 1966 the U.S. Equal Employment Opportunity Commission (EEOC) has been collecting yearly workplace surveys describing outcomes on equal employment opportunity (EEO) . Private sector firms with more than 50 employees (25 if federal contractors), are required to submit yearly reports on the race/ethnic and sex composition of their work force in each establishment with 25 or more employees, about 696691 across US. These reports contain establishment employment counts of sex by five race/ethnic groups (White, Black, Hispanic, Asian/ Pacific Islander, American Indian/ Alaskan Native) distributed across nine occupational categories (officials and managers, professionals, technicians, sales workers, office and clerical workers, craft workers, operatives, laborers, and service workers). These reports also include information on the establishments parent company, industry, and geographic location. Each record states whether or not the parent company is a federal contractor.

Unit of analysis in the substantive analyses is defined to be an establishment. Each establishment has repeated observations over time. At any one point in time establishments are nested within firms. Firms that are federal contractors are required to practice affirmative action. We observe federal contractor status as a firm characteristic. Establishments are also nested within industries. Industries provide normative models of appropriate workplace organization. Industries with more diversity in group representation may encourage managerial integration at the workplace level. We observe the proportion of status group representation in total antd managerial industry employment. Establishments are also nested within spatial contexts. The local labor market from which labor is drawn influences the ability to hire from various status groups. For each outcome variable we observe that groups proportional representation in the local labor market. A second spatial context is the state an establishment is found within. States represent a political context that may influence workplace behavior. Prior research suggests that as the percent minority in states increase discrimination in various institutions (education, law, voting, as well as employment) increase as well. Other research suggests that unions were strong supporters of civil rights law. We observe percent black, Hispanic, and unionized at the state level to model their influence on state as political context. Figure 1 depicts this complicated structure of nesting.

Establishments can also shift industries and firms over

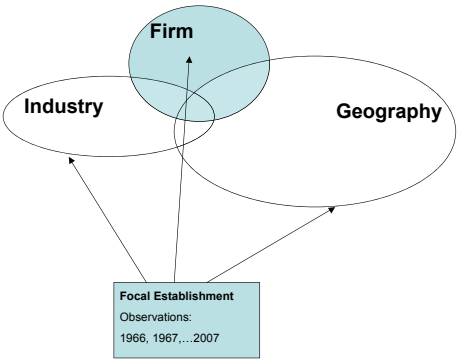

Figure 1: Non-nested depiction of EEO data

time. Note that shift is permanent, making this multiple membership problem different from the ones seen in educational or genetic studies. For example, Good Foods Establishment in Chicago is under "Good Foods" firm several years within grocery and retails industry, then it gets bought out by "Better Foods" which now is under retail industry. The multiple membership problem is also depicted in Figure 2.

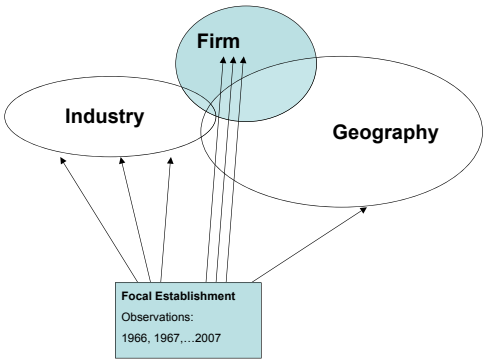

Figure 2: Depiction of multiple membership

Substantively we explore confidential private sector workplace panel data collected annually by the U.S. Equal Employment Opportunity Commission (EEOC) since 1966. Some of the substantive goals of this study have been handicapped by incompleteness of the database, due to either lost records from early years (1967–1970, 1973 and 1975) of data collection or simple item nonresponse. This produced two methodological problem. First problem results from missing panels representing the period of most rapid change in sex and race employment relations, analyses that ignore them would potentially bias the analyses. In addition, these surveys produce data on the multiple membership context of responding establishments (owning firm, industry, state) as well as the typical panel contexts of time and repeated observations. These two problems motivated our basic goal of developing joint imputation models allowing for multiple membership and multiple classification (3MC) models.

1.2 Example

Consider a model for organizational change in the sex and race composition of managerial occupations as a function of their firm, industrial, and geographic contexts, with an emphasis both on estimating the variance components associated with multiple membership and the influence of observed variables within each membership on shifts in managerial composition. For the sake of discussion, let's focus on a simplified version of this model that only reflects the structure of the data. A random-intercept-only model on the response variable y_i is given as follows (Browne, Goldstein, Rasbash, 2002),

$$y_i = \sum_{l=1}^{p_f} x_{il}\beta_l + b^{(2)}_{Est(i)} + b^{(3)}_{Geog(i)} + \sum_{c\in Ind(i)} w^{(4)}_{i,c} b^{(4)}_c + \sum_{d\in Firm(i)} w^{(5)}_{i,d} b^{(5)}_d + \epsilon_i, \quad (1)$$

where y_i denotes outcome of interest for the observation level (establishment/year) i; Est(i), Geog(i), Ind(i) and Firm(i) are i^{th} observation's establishment (year), geography, firm and industry, respectively. β represents effects common to all establishments and establishment-specific effects due to nesting/clustering factors are denoted by $b^{(2)}, b^{(3)}, b^{(4)}, b^{(5)}$, which are random intercepts for classification 2 (establishment), classification 3 (geography), classification 4 (Firm) and classification 5 (Industry). Weights due to multiple membership are contained in $w^{(4)}_{i,c}$, indicating the weight given to Classification 4 for establishment/year i. Finally, distributional specifications on error terms and random effects are: $\epsilon_i \sim N_{n_i}(0, \sigma^2 V_i)$, $b^{(2)}_{Est(i)} \sim N(0, \psi^{(2)}), b^{(3)}_{Geog(i)} \sim N(0, \psi^{(3)}), b^{(4)}_{Ind(i)} \sim N(0, \psi^{(4)}), b^{(5)}_{Firm(i)} \sim N(0, \psi^{(5)})$.

How to best handle missing values in the response or covariates remains to be a challenging issue for most researchers. As most statistical analyses and estimation procedures are not designed to handle missing values, especially at this complexity level, it may be tempting to choose unprincipled methods such as case deletion or ad-hoc methods of single imputation. Biased estimates, understated variances or lower coverage rates are some of the major problems that are generally associated with these "unprincipled" methods. Comprehensive discussion on such adverse effects is given by Rubin (1987), Little and Rubin (2002) and a review of modern missing-data techniques is given by Schafer and Graham (2002).

1.3 Previous research

Extensive literature exists on model-fitting techniques and inferences for multilevel data sets, some of the sources include Diggle, Liang, and Zeger (1994), Vonesh and Chinchilli (1997), Pinheiro and Bates (2000), Verbeke and Molenberghs (2000), McCulloch and Searle (2001), Demidenko (2004), and Fitzmaurice, Laird, and Ware (2004). Together, these provide a clear and comprehensive discussion of state-of-the-art methods for estimation, testing, and prediction in the context of linear, generalized linear, and nonlinear mixed-effects modeling. Browne, Goldstein, and Rasbash (2001) provides Bayesian model-fitting techniques using MCMC simulation techniques for applications with multiple membership and cross-classification.

Since the landmark paper on MI and EM (Rubin 1978, Dempster, Laird, and Rubin 1977), literature on missing data methods have become quite extensive. Applications of these well-known missing-data techniques in multilevel settings with arbitrary patterns of missing values, however, have not been equally well-developed. Several researchers showed that, when the missingness is only on the response variable, under certain conditions (such as missing at random as defined below), the inferences under mixed-effects models are valid. Some of the limited work on MI in multilevel applications include Liu, Taylor, and Belin 2000, Schafer and Yucel 2002, (Carpenter and Kenward 2008). This work primarily deals multilevel applications where the observations may appear in multiple clustering factors and some of the clustering factors are not necessarily in a "hiearchical" nature. In a modelling sense, we modify the model suggested by Browne, Goldstein, and Rasbash (2001) to allow multiple responses and use similar Bayesian arguments within the MCMC simulation framework.

2. Models

Any missing data method assumes certain structures for the mechanisms generating either missing-data (missingness mechanism) or data intended to be collected. In the following parts, we briefly summarize the commonly assumed structures on missingness mechanism, and our model used to base multiple imputations.

2.1 Models for missingness mechanism

Explicit or implicit assumptions are made about the missing-data mechanisms in all missing-data methods. To set the notation for the discussion of missingness mechanisms, let R denote the set of missing-value indicators; note that R has the same dimension as Y, and it is always observed. Each element of R takes the value of 0 or 1 depending on whether the corresponding element in Y is missing or observed, respectively. Similar to Y, R can be seen as a random variable; and the conditional distribution of R given Y depends on a set of parameters, say γ.

Most tools available to the practitioners of missing data methods (e.g. SAS PROC MIXED (Littell, Miliken, Stroup, and Russell 1996), R packages norm (Schafer 2000) or Splus library missing and pan (Schafer and Yucel 2002)) assume *missing at random* (MAR) as the missingness mechanism. This assumption implies that the missingness probability may depend on the observed data but not on the missing data over the conditional distribution of R given Y_{obs}. More formally, $P(R \mid Y_{obs}, Y_{mis}, \gamma) = P(R \mid Y_{obs}, \gamma)$. Some misconceptions among practitioners exist due to the name MAR. The missing values do not occur at random under MAR; when they do, the mechanism is, in fact, *missing completely at random* (MCAR). Under MCAR, the missingness probabilities are independent from both Y_{obs} and Y_{mis}: $P(R \mid Y_{obs}, Y_{mis}, \gamma) = P(R \mid \gamma)$. When the missingness probabilities depend on Y_{mis}, the missingness mechanism is called *missing not at random* (MNAR). Under MNAR, one must posit a model for the complete data as well as for R. These models are usually very sensitive to the model assumptions (see detailed discussion by Schafer and Graham (2002) for more information). Another important concept is the "ignorability" of the missingness mechanism, and it is often seen as an implied condition once MAR is assumed. Specifically, ignorability occurs when the mechanism is MAR and the parameters γ and θ are distinct: $f(Y_{obs}, R \mid \theta, \gamma) = f(Y_{obs} \mid \theta) f(R \mid \gamma)$ (see Little and Rubin (2002) and Schafer (1997) for more details).

2.2 Imputation models

The theory of MI does not require any particular assumption on missingness, it can be performed under any type of missingness-mechanism. However, the structure from which multiple imputations are drawn needs to be specified so that missing values are replaced by draws from the posterior predictive distribution of missing data (i.e. the conditional distribution of the missing data, given the observed data and the unknown parameters). This typically involves positing a parametric model for the data and using it to derive this conditional distribution. In multilevel data applications, multivariate extensions of the mixed-effects models based on normality have often been perceived as a natural assumption because (1) it reflects the design features; and (2) the conditional distribution of the missing data given the observed data is easily tractable (Schafer and Yucel 2002). Several studies demonstrated that under moderate missingness, most parametric assumptions do not matter on the validity of the inferences and more important emphasis should be put on the imputation model reflecting the important data features such as clustering (**?**). For this reason, our models are specifically designed to consider non-nested classifications and multiple-meberships of observation unit establishments. Once the multiple imputations are created under these model, say m times (in most problems $m < 10$), an ana-

lyst's model is fitted with these imputed data, resulting a set of m set of coefficients and associated standard errors. These results are then combined using rules by (Rubin 1987). Other combining rules that operate on other inferential quantities (e.g. p-values) are also available, see for example Li, Meng, Raghunathan, and Rubin (1991) and Rubin (1987).

2.3 Multivariate Multiple Membership and Multiple Classification (4MC) Models

For the sake of clarity, we will use a scalar representation to easily express multiple membership as well as multiple classification. Let $Y_{i1}, \ldots, Y_{ir}$ denote a set of r incompletely-observed variables for i^{th} establishment and year combination. $Est(i), Geog(i), Ind(i)$ and $Frm(i)$ refer to the appropriate classifications at the establishment, geography, industry and firm level. Our modeling strategy regards repeated observations over years as nested within establishments, i.e. establishment itself is a classification factor. A random-intercept-only model for all of the corresponding clustering factors is given by

$$\begin{aligned} y_{ij} &= \sum_{l=1}^{p_f} x_{ijl}\beta_{lj} + b^{(2)}_{Est(i),j} + b^{(3)}_{Geog(i),j} + \\ &\quad \sum_{c\in Ind(i)} w^{(4)}_{i,c} b^{(4)}_{c,j} + \sum_{d\in Firm(i)} w^{(5)}_{i,d} b^{(5)}_{d,j} + \epsilon_{ij}, \quad (2) \end{aligned}$$

where x_{ij} denotes a set of *completely-observed* p_f covariates for the i^{th} establishment/year observation and j^{th} response variable. β_{lj} is the set of p_f coefficients corresponding to covariates in x_{ij}. The random intercepts for classification levels at $Est(i), Geog(i), Ind(i)$ and $Frm(i)$ are

$$\begin{aligned} vec(b^{(2)}_{Est(i)}) &\sim N(0, \Psi^{(2)}) \\ vec(b^{(3)}_{Geog(i)}) &\sim N(0, \Psi^{(3)}) \\ vec(b^{(4)}_{Ind(i)}) &\sim N(0, \Psi^{(4)}) \\ vec(b^{(5)}_{Firm(i)}) &\sim N(0, \Psi^{(5)}). \end{aligned}$$

Note that the influence of multiple industries and firms on an establishment is adjusted to reflect the appropriate weights by $w^{(4)}_{i,c}$ and $w^{(5)}_{i,d}$. The model specification is concluded by the assumption on the error term: $vec(\epsilon_i) \sim N_r(0, \Sigma)$ and $\Psi^{(c)}$ can be assumed as block-diagonal or unstructured, where c denotes the underlying classification for establishment, geography, industry or firm. The multiple membership weights are pre-assigned and typically a frequency of appearance of the unit is used.

Finally, we assume standard prior distributions on the variance parameters:

$$\begin{aligned} \beta &\sim \text{uniform on } \mathcal{R}^{pr} \text{ (improper)}, \\ (\Psi^{(c)})^{-1} &\sim Wishart(\nu^{(c)}_1, \Lambda^{(c)}_1), \nu_1 \geq r, \\ \Sigma^{-1} &\sim Wishart(\nu_2, \Lambda_2), \nu_2 \geq r, \end{aligned}$$

where

$$\begin{aligned} (\nu^{(c)}_1)^{-1}(\Lambda^{(c)}_1)^{-1} &: \text{prior guess for } \Psi^{(c)},\ c = 2,3,4,5, \text{ and} \\ \nu_2^{-1}\Lambda_2^{-1} &: \text{prior guess for } \Sigma. \end{aligned}$$

3. Inferential algorithms

Our ultimate interest is to generate M independent draws of missing data, $Y^{(1)}_{mis}, \ldots, Y^{(M)}_{mis}$ from the posterior predictive distribution for the missing data derived under the model given by (2):

$$P(Y_{mis} \,|\, Y_{obs}) = \int P(Y_{mis} \,|\, Y_{obs}, \theta)\, P(\theta \,|\, Y_{obs})\, d\theta, \quad (3)$$

where $P(\theta \,|\, Y_{obs})$ is the observed-data posterior density, which is proportional to the product of the prior densities given in Section 2.3 and the observed-data likelihood function

$$L(\theta \mid Y_{obs}) = \int L(\theta \mid Y_{obs}) dY_{mis}.$$

After imputation, the resulting M versions of the complete data are analyzed separately by complete-data methods, and the results are combined using simple arithmetic to obtain inferences that effectively incorporate uncertainty due to missing data. As shown by Rubin (1987), quality inferences can often be obtained with a very small number (e.g., M = 5) of imputations. Methods for combining the results of the complete-data analyses are given by Rubin (1987, 1996) and reviewed by Schafer (1997, chap. 4).

Except in trivial special cases, the posterior predictive distribution (3) for our model cannot be simulated directly. We create random draws of Y_{mis} from $P(Y_{mis} \mid Y_{obs})$ by techniques of Markov chain Monte Carlo (MCMC) called a Gibbs Sample. In Gibbs sampler, one generates a sequence of dependent random variates whose distribution converges to the desired target distributions $P(Y_{mis} \mid Y_{obs})$ and $P(\theta \mid Y_{obs})$. Specifically, it updates the current version of the unknown parameters $\theta^{(t)} = ((\Psi^{(c)})^{(t)}, \Sigma^{(t)}, \beta^{(t)})$ and missing data $Y^{(t)}_{mis}$ are updated in successive steps. Below we describe in detail these steps for one classification for multiple membership and extend it to multiple classification in the following section.

3.1 Algorithm for one classification with multiple membership

Suppose that observational units are grouped under a single clustering factor (e.g. establishment/year within firm) and allowed to appear in more than one cluster. The following is the proposed model for a set of incompletely observed variables:

$$y_{ij} = \sum_{l=1}^{p_f} x_{il}\beta_{lj} + \sum_{c\in Firm(i)} w^{(2)}_{i,c} \sum_{l=1}^{p_2} z^{(2)}_{il} b^{(2)}_{c,lj} + \epsilon_{ij}, \quad (4)$$

where $vec(b_i^{(2)}) \sim N_{rp_2}(0, \Psi^{(2)})$ and $\epsilon_i \sim N(0, \Sigma)$. For clarity of the algorithmic details, we let b_i denote $b_i^{(2)}$. Similarly, let $\theta^{(t)} = (\beta^{(t)}, \Sigma^{(t)}, (\Psi^{(2)})^{(t)})$, missing data $Y_{mis}^{(t)}$, $B^{(t)} = (b_1^{(t)}, b_2^{(t)}, \ldots, b_{N_{firm}}^{(t)})$ denote the current state of unknowns. Conditionals of a Gibbs sampler are given by

$$\begin{aligned} b_i^{(t+1)} &\sim P(b_i \mid Y_{obs}, Y_{mis}^{(t)}, \theta^{(t)}), && (5)\\ \theta^{(t+1)} &\sim P(\theta \mid Y_{obs}, Y_{mis}^{(t)}, B^{(t+1)}), && (6)\\ y_{i(mis)}^{(t+1)} &\sim P(y_{i(mis)} \mid Y_{obs}, B^{(t+1)}, \theta^{(t+1)}), && (7) \end{aligned}$$

where $i = 1, \ldots, N_{firm}$. Given $\theta^{(0)}$ and ${Y_{mis}}^{(0)}$, (1)–(3) define a cycle of MCMC called Gibbs sampler, and $\{\theta^{(t)}\} \to^d P(\theta \mid Y_{obs})$, and $\{Y_{mis}^{(t)}\} \to^d P(Y_{mis} \mid Y_{obs})$ as $t \to \infty$. The specific forms of the conditionals follow from straightforward application of Bayes' theorem and some algebra. Below these conditionals are provided, details are available upon request.

3.1.1 Conditionals of the Gibbs: $P(b_i \mid Y_{obs}, Y_{mis}, \theta)$

The pairs (y_i, b_i) assumed to be independent for $i = 1, 2 \ldots, N_{firm}$, with

$$\begin{aligned} vec(y_i) \mid b_i, \theta &\sim N(\mu_i, V_i)\\ vec(b_i) \mid \theta &\sim N(0, \Psi^{(2)}), \end{aligned}$$

where $\mu_i = E(vec(y_i) \mid b_i, \theta) = vec(X_i\beta + \sum_{c=1}^{C} w_{i,c} Z_i b_i)$ and $V_i = V(vec(y_i) \mid b_i, \theta) = \Sigma \otimes I_{n_{Frm(i)}})$ as the rows of y_i for a given establishment are assumed to be independent. Bayes' theorem implies

$$vec(b_i) \mid y_i, \theta \sim N(vec(\tilde{b}_i), U_i),$$

where

$$\begin{aligned} vec(\tilde{b}_i) &= U_i(\Sigma \otimes Z_i^T) vec(y_i - X_i\beta)\\ U_i &= ((\Psi^{(2)})^{-1} + (\Sigma^{-1} \otimes Z_i^T Z_i))^{-1} \end{aligned}$$

3.1.2 Conditionals of the Gibbs: $P(\theta \mid Y_{obs}, Y_{mis}, B)$

Simulation of θ in (6) proceeds as follows: First draw $(\Psi^{(2)})^{(-1)}$ from a Wishart distribution with degrees of freedom $\nu + N_{Firm}$ and scale $(\Lambda^{(-1)} + \sum_i^{N_{firm}} b_i b_i^T)$. This result follows from simple application of Bayes theorem on the joint density of the random-effects $(b_i \mid \Psi^{(2)} \sim N(0, \Psi^{(2)})$, independently for $i = 1, 2, \ldots, N_{firm}$) and conjugate prior $(\Psi^{(2)})^{(-1)} \sim W(\nu, \Lambda)$. Next calculate the ordinary least-square coefficients

$$\hat{\beta} = \left(\sum_i X_i^T X_i\right)^{(-1)} \sum_i X_i^T (y_i - \sum_{c=1}^{C} w_{i,c} Z_i b_i)$$

and residuals $\hat{\epsilon}_i = y_i - X_i\hat{\beta} - \sum_{c=1}^{C} w_{i,c} Z_i b_i$, and draw $\Sigma^{(-1)}$:

$$\Sigma^{-1} \mid Y_{obs}, Y_{mis}, \{b_i\}_{i=1}^{N_{firm}}, \Psi^{(2)}, \beta \sim W(\nu_1 - p_f + \sum_i^{N_{firm}} n_i, (\Lambda_1^{-1} + \sum_{i=1}^{N_{firm}} \hat{\epsilon}_i^T \hat{\epsilon}_i)^{-1}).$$

Finally, draw β from a multivariate normal distribution centered at $\hat{\beta}$ with covariance matrix $V(\hat{\beta})$ where

$$\begin{aligned} \hat{\beta} &= \left(\sum_i X_i^T X_i\right)^{(-1)} \sum_i X_i^T (y_i - \sum_{c=1}^{C} w_{i,c} Z_i b_i)\\ V(\hat{\beta}) &= \Sigma^{(-1)} \otimes \left(\sum_i X_i^T X_i\right)^{(-1)}. \end{aligned}$$

3.1.3 Drawing missing data: $P(y_{i(mis)} \mid Y_{obs}, B, \theta)$

Our goal in (7) is to draw from the following conditional using the most recent state of unknowns:

$$y_{i(mis)} \sim P(y_{i(mis)} \mid Y_{obs}, B, \theta),\ i = 1, \ldots, N_{firm}.$$

This task can easily be accomplished by noticing that the rows of $\epsilon_i = y_i - X_i\beta - Z_i b_i$ are independet and normally distributed with mean 0 and covariance matrix Σ, or

$$\epsilon_i \mid Y_{obs}\beta, b_i, \Sigma, \Psi \sim N(0, \Sigma \otimes I_{n_i}).$$

This implies that, in any row of ϵ_i, the missing elements have an intercept-free multivariate normal regression on the observed elements; the slopes and residual covariances for this regression can be quickly calculated by inverting the square submatrix of Σ corresponding to the observed variables. That is $\epsilon_{i,M(s)} \sim N(E(\epsilon_{i,M(s)} \mid \epsilon_{i,M(s)}), V(\epsilon_{i,M(s)} \mid \epsilon_{i,M(s)}))$, where

$$E(\epsilon_{i,M(s)} \mid \epsilon_{i,O(s)}) = \Sigma_{M(s),O(s)} \Sigma_{O(s),O(s)}^{(-1)} \epsilon_{i,O(s)}$$

$$V(\epsilon_{i,M(s)} \mid \epsilon_{i,O(s)}) = \Sigma_{M(s),O(s)} \Sigma_{O(s),O(s)}^{(-1)} \Sigma_{O(s),M(s)}$$

$M(s), O(s)$ denote the missing and observed values in missingness patters s, respectively. Because computations are performed over distinct missingness patterns, our algorithm's computational cost per iteration is relatively low.

3.2 Incorporating additional classifications

In EEOC data, observational establishment units are classified by numerous factors as depicted in Figures 1 and 2. Incorporating these additional classification factors is a matter of simulating a conditional distribution of the related parameters in the Gibbs sampler defined by (6)–(7). Consider, for example, adding an establishment

level random-effects that will account for correlated measurements per establishment (recall that each establishment has repeated measurements over years). A natural extension of of model (4) can be written as

$$y_{ij} = \sum_{l=1}^{p_f} x_{il}\beta_{lj} + b^{(2)}_{Est(i),j} + \sum_{c\in Firm(i)} w^{(3)}_{i,c} \sum_{l=1}^{p_3} z^{(3)}_{il} b^{(3)}_{c,lj} + \epsilon_{ij} \quad (8)$$

In this new formulation, "establishment" specific random-effects $b^{(2)}_{Est(i),j}$ are among the unknowns along with θ, $B^{(3)}$ (which was $B^{(2)}$ in Section 3.1) and Y_{mis}. The new Gibbs sampler now adds the conditional of $b^{(2)}_{Est(i),j}$. Note that for fixed values of $b^{(2)}_{Frm(i)}$, the model

$$y^{\star}_{ijk} = \sum_{l=1}^{p_f} x_{ijl}\beta_{kl} + \sum_{c\in Firm(i)} w^{(3)}_{i,c} \sum_{l=1}^{p_3} z^{(3)}_{il} b^{(3)}_{c,lj} + \epsilon_{ijk},$$

is same as the previous model of Section 3.1 with $y^{\star}_{ijk} = y_{ijk} - b^{(2)}_{Est(i),j}$. And for fixed values of $b^{(2)}_{Est(i),j}$, the model (8) reduces to

$$y^{\star}_{ijk} = \sum_{l=1}^{p_f} x_{ijl}\beta_{kl} + b^{(2)}_{Est(i),j} + \epsilon_{ijk},$$

where $y^{\star}_{ijk} = y_{ijk} - \sum_{c\in Firm(i)} w^{(3)}_{i,c} \sum_{l=1}^{p_3} z^{(3)}_{il} b^{(3)}_{c,lj}$. Note that both of these implied conditional models are same as the one given in Section 3.1.

4. A simulation study

Our limited simulation study evaluates the frequentist characteristics of the 4MC-based MI in a repetitive sampling setting. It consists of data generation (clustered data with multiple membership), imposing missing values under MAR, MI inference under a hypothetical analyst's model, parameter estimation, and evaluation of this estimation:

Data generation:
A bivariate intercept-only linear mixed-effects model allowing for multiple membership was used to generate complete data:

$$y_{ij} = \beta_{0j} + \sum_{c\in Firm(i)} w^{(2)}_{i,c} b^{(2)}_{c,j} + \epsilon_{ij},$$

where $j = 1, 2$, $i = 1, 2, \ldots, N = 1000$ and firm of establishment i is one of the 100 firms: $Firm(i) \in (1, 2, \ldots, 100)$. We assumed 80% of the establishments belong to only one firm and 20% appear in two firms with weights 0.5 ($w^{(2)}_{i,c} = 0.5$). The coefficients of the data model above were set to $\beta_{0,1} = 0.5, \beta_{0,2} = 1$, $Var(\epsilon_1) = 1.2, Var(\epsilon_2) = 0.7, Cov(\epsilon_1, \epsilon_2) = -0.2$, finally, $Var(b_1) = 0.8, Var(b_2) = 1.5, Cov(b_1, b_2) = 0.3$.

Imposing missing values and analyst's model
We imposed missing values under the missingness mechanism defined by

$$\log \frac{P(r_{Y_2} = 1 \mid Y_1)}{1 - P(r_{Y_2} = 1 \mid Y_1)} = \beta^m_0 + \beta^m_1 y_1,$$

where β_0 and β_1 are set to values to produce around 30% missingness.

Evaluating MI under 4MC
Our goal was to mimic the practice of multiple imputation. To make this as real as possible, we assumed the model given in (3) underlies the substantive goal of the analysis, i.e. the analyst's model. Hence, we performed MI inference on the estimation parameters of this model. We then evaluated the performance of MI under 4MC by comparing its the coverage rates to available-cases only analyses and MI under a model ignoring multiple membership (Schafer and Yucel 2002, referred as PAN below). We define (CR) as the percentage of times that the true parameter value is covered in the 95% confidence interval. Here the true parameter value is the average parameter estimate across the simulations before the missing values are imposed. If a procedure is working well, the actual coverage should be close to the nominal rate of 95% in our study. If the procedure results in CRs that are close to 100% or below 85%, extra caution should be taken when using that procedure.

Table 1 summarizes the results of our limited simulation study. Results show that both MI procedure (4MC and PAN) capture the true unknown parameter values, however the procedure ignoring multiple membership (PAN) leads to confidence intervals with much lower coverage rate than a nominal rate of 95%. MI under 4MC achieve excellent coverage rates. Similar behavior is seen in capturing the random-effect variance, MI under PAN underestimates the true value. The performance of the MI under either model is far more superior than the unprincipled method of simple case-deletion, which leads to significant biases as well as dismal coverage rates, in some cases it is as dramatic as less than half of the nominal rate.

Table 1: Summary of the results under MI under 4MC. PI stands for parameter of interest

PI	Before deletion	After deletion	MI under 4MC	MI under PAN
β_0	1.25	1.15 (0.803)	1.25 (0.945)	1.25 (0.955)
β_1	0.47	0.34 (0.691)	0.46 (0.939)	0.48 (0.929)
$\psi^{(2)}$	1.087	0.849 (0.454)	1.10 (0.959)	1.01 (0.89)

5. Discussion

In this manuscript, we describe an MI-based strategy to estimate models from data structures complicated by

non-nested clusterings, multiple membership of units to these clusters and missing data. This research was motivated by interest in characterizing the change in sex and race composition of managerial occupations using an administrative longitudinal data. The proposed approach, however, has a broader relevance to many other fields where such complexities coupled with missing data are seen. It is common to see such non-nested clusters with multiple membership in health services research (e.g. patients treated both in multiple hospitals and/or multiple doctors) or education (e.g. longitudinal studies on students in multiple schools).

In this paper, "multiple membership" weights were assumed fixed and unknown. In problems where the cluster identifiers are not directly observed but rather some of their determinants are observed (e.g. ambiguous geneotype assignments in genetic association studies), these weights are often modeled and estimated (Foulkes, Yucel, and Li In press). Introduction of this additional modeling step is a meter of adding a Gibbs sampler step. However, It should be noted that one of the implication of the increased rate of missingness would be somewhat decreased performance measures due to increased fraction of missing information (Little and Rubin 2002).

The computational algorithms for simulating MIs under the proposed model could be made to mix faster. This generally implies reducing or de-conditioning on simulated values of some of the unknowns such as random-effects. De-conditioning may greatly increase the computational cost per iteration, however, and some limited experience suggests that the additional effort required to do so is not worthwhile. With modern computers, iterations of the Gibbs are performed quickly even with the large datasets provided that sufficient physical memory is available to store Y_{obs},$Y_{mis}^{(t)}$ and covariate matrices.

Our current work includes a comprehensive simulation study (both Monte Carlo and repetitive sampling from the administrative dataset) to evaluate our method and applying our methods to EEOC data for substantive results. Our model allows only continuous data, the future extensions will include models for categorical or mixtures of continuous and categorical data in similar settings. When the number of categorical items is large, estimation-related computational difficulties can occur. In such settings we will incorporate conditional variable-by-variable approach (Raghunathan, Lepkowski, and VanHoewyk 2001; Yucel and Raghunathan 2006). Finally, given the comprehensive and rich nature of the EEOC data, alternative ad-hoc imputation methods (e.g. predictive mean matching) can be quite successful. These methods usually make minimal assumptions and may lead to underestimation of the standard-errors due to their single-imputation nature. However, with improvements to adjust the standard errors, they can be valuable source of missing-data technique in rich data systems such as EEOC.

References

Browne, J., Goldstein, H., and Rasbash, J. (2001), "Multiple membership multiple classification (MMMC) models," *Statistical Modelling*, 1, 103–124.

Carpenter, J. and Kenward, M. (2008), *Instructions for MLwiN multiple imputation macros.*, Bristol, UK: Centre for Multilevel Modelling.

Demidenko, E. (2004), *Mixed Models: Theory and Applications*, New York: John Wiley and Sons.

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.

Diggle, P., Liang, K., and Zeger, S. (1994), *Analysis of longitudinal data*, Oxford: Oxford University Press.

Fitzmaurice, G., Laird, N., and Ware, J. (2004), *Applied Longitudinal Analysis*, New York: John Wiley and Sons.

Foulkes, A. S., Yucel, R. M., and Li, X. (In press), "Mixed modelling with ambiguous clusters: A likelihood-based approach," *Biostatistics*.

Li, K., Meng, X., Raghunathan, T., and Rubin, D. (1991), "Signicance levels from repeated p-values with multiply-imputed data," *Statistica Sinica*, 1, 65–92.

Littell, R., Miliken, G., Stroup, W., and Russell, D. (1996), *SAS System for Mixed Models*, Carey, NC: SAS Publishing.

Little, R. J. A. and Rubin, D. B. (2002), *Statistical Analysis with Missing Data, Second Edition*, New York: J. Wiley & Sons, New York.

Liu, M., Taylor, J., and Belin, T. (2000), "Multiple Imputation and Posterior Simulation for Multivariate Missing Data in Longitudinal Studies," *Biometrics*, 56, 1157–1163.

McCulloch, C. and Searle, S. (2001), *Generalized, Linear and Mixed Models*, New York: John Wiley and Sons.

Pinheiro, J. and Bates, D. (2000), *Mixed-Effects Models in S and S-PLUS*, New York: Springer-Verlag Inc.

Raghunathan, T. E., Lepkowski, J. M., and VanHoewyk, J. (2001), "A multivariate technique for multiply imputing missing values using a sequence of regression models," *Survey Methodology*, 27, 1–20.

Rubin, D. B. (1978), "Multiple Imputations in Sample Surveys: A Phenomenological Bayesian Approach to Nonresponse," in *ASA Proceedings of the Section on Survey Research Methods*.

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

SAS Institute (2001), *SAS/Stat User's Guide, Version 8.2*, Carey, NC: SAS Publishing.

Schafer, J. (2000), *Multiple imputation of incomplete multivariate normal data*, The Pennsylvania State University, PA, USA.

Schafer, J. and Yucel, R. (2002), "Computational strategies for multivariate linear mixed-effects models with missing values," *Journal of Computational and Graphical Statistics*, 11, 421–442.

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

Schafer, J. L. and Graham, J. W. (2002), "Missing Data: Our View of the State of the Art," *Psychological Methods*, 7, 147–177.

Verbeke, G. and Molenberghs, G. (2000), *Linear Mixed Models for Longitudinal Data*, New York: Springer-Verlag Inc.

Vonesh, E. F. and Chinchilli, V. M. (1997), *Linear and nonlinear models for the analysis of repeated measurements*, Marcel Dekker Inc.

Yucel, R. M. and Raghunathan, T. (2006), "Sequential Hierarchical Regression Imputation (SHRIMP)," in *ASA Proceedings of the Joint Statistical Meetings*, American Statistical Association.