REVSTAT – Statistical Journal
Volume 16, Number 2, April 2018, 213–230

IMPROVING BAYESIAN MIXTURE MODELS FOR MULTIPLE IMPUTATION OF MISSING DATA USING FOCUSED CLUSTERING

Authors: Lan Wei
– In4mation Insights,
Needham, MA, USA

Jerome P. Reiter
– Department of Statistical Science, Duke University,
Box 90251, Durham, NC, 27708, USA
jerry@stat.duke.edu

Received: February 2017 Revised: September 2017 Accepted: October 2017

Abstract:

- We present a joint modeling approach for multiple imputation of missing continuous and categorical variables using Bayesian mixture models. The approach extends the idea of focused clustering, in which one separates variables into two sets before estimating the mixture model. Focus variables include variables with high rates of missingness and possibly other variables that could help improve the quality of the imputations. Non-focus variables include the remainder. In this way, one can use a rich sub-model for the focus set and a simpler model for the non-focus set, thereby concentrating fitting power on the variables with the highest rates of missingness. We present a procedure for specifying which variables with low rates of missingness to include in the focus set. We examine the performance of the imputation procedure using simulation studies based on artificial data and on data from the American Community Survey.

Key-Words:

- *incomplete; nonparametric; nonresponse; survey; tensor.*

1. INTRODUCTION

Nonparametric Bayesian (NB) mixture models are useful tools for analyzing complicated data ([13], [5], [14], [3], [2]). They are especially useful as engines for multiple imputation (MI, [16], [11], [18], [9], [12], [10], [7]). NB mixture models are flexible enough to capture complex relationships among the variables, which is advantageous in MI contexts where one seeks to create completed datasets for use in multiple analyses.

In many contexts, only a few variables have high rates of missingness, and other variables are nearly or completely observed. This can create estimation difficulties when using mixture models as MI engines. In particular, with modest sample sizes and many variables, mixture models have the potential to fit the distribution of some variables well at the expense of others ([6], [19], [4]). The mixture model easily could expend its fitting power on the marginal distribution of the (nearly) completely observed variables at the expense of the distribution of the variables with high rates of missingness ([4],[20]), which could lead to poor quality imputations.

To get around this, [4] suggest using mixture models with focused clustering. Using the nomenclature in [4], the variables with high rates of missing data are called focus variables, and the others are called remainder variables. In focused clustering, the mixture model includes one set of cluster indicators for focus variables and a second set for remainder variables. The two sets are connected using a tensor factorization prior ([15]). In this way, one can use a rich sub-model for the focus set and a simpler model for the remainder set, thereby concentrating fitting power on the variables with the highest rates of missingness.

In this article, we enhance the focused clustering approach for MI to facilitate higher quality imputations. In particular, we expand the definition of focus variables to include variables with high fractions of missing data and (nearly) completely observed variables that could improve the quality of the imputations for the variables with high rates of missingness; we label the resulting set with $\mathcal{F}$. We define the non-focus variables to include those not in $\mathcal{F}$; we label these as $\mathcal{NF}$. We specify the variables to include in $\mathcal{F}$ as follows. First, we automatically put all variables with high fractions of missing values in $\mathcal{F}$. For each variable not automatically in $\mathcal{F}$, we compute its mutual information with the variables automatically in $\mathcal{F}$. We move variables with high mutual information values into $\mathcal{F}$; the remaining variables we put in $\mathcal{NF}$. We make these decisions in one step, including all variables with high mutual information values in $\mathcal{F}$. We refer to this strategy as *M*ove. We use *Stay* to refer to the strategy of putting only variables with high fractions of missingness in $\mathcal{F}$. Because *Move* allows local dependence among the variables with high amounts of missing values and (nearly) completely observed variables that can be used to predict the missing values, it can improve accuracy and, in some cases, computational efficiency.

The remainder of this article is organized as follows. In Section 2, we present the focused clustering model, which we abbreviate as HCMM-FNF for hierarchically coupled mixture model with focus/non-focus variables, and motivate the potential benefits of *Move*. In Section 3, we illustrate when *Move* engenders benefits using four simple simulation scenarios. In Section 4, we apply the strategies to data sampled from the American Community Survey. In Section 5, we conclude with a brief summary of findings.

2. SPECIFICATION OF HCMM-FNF

We indicate continuous variables with Y and categorical variables with X. We use a superscript F to denote focus variables and the superscript NF to denote non-focus variables. Thus, $Y^{(F)}$, $X^{(F)}$, $Y^{(NF)}$ and $X^{(NF)}$ are the focus continuous, focus categorical, non-focus continuous, and non-focus categorical variables, respectively. For purposes of explaining HCMM-FNF, here we assume that $\mathcal{F}$ and $\mathcal{NF}$ have been pre-specified.

For each observation $i = 1, ..., n$, we have $Y_i^{(F)} = \big(Y_{i1}^{(F)}, ..., Y_{iq^{(F)}}^{(F)}\big)^T$, $X_i^{(F)} = \big(X_{i1}^{(F)}, ..., X_{ip^{(F)}}^{(F)}\big)^T$, $Y_i^{(NF)} = \big(Y_{i1}^{(NF)}, ..., Y_{iq^{(NF)}}^{(NF)}\big)^T$, and $X_i^{(NF)} = \big(X_{i1}^{(NF)}$, ..., $X_{ip^{(NF)}}^{(NF)}\big)^T$. Let D_i be a regression design matrix containing the main effects of $X_i^{(F)}$, $Y_i^{(NF)}$, and $X_i^{(NF)}$. A similar regression approach is proposed by [15]. HCMM-FNF can be described as follows.

$$\big(Y_i^{(F)} | D_i, H_i^{(FY)} = a, -\big) \sim \mathcal{N}\big(y_i^{(F)} | D_i B_a^{(F)}, \Sigma_a^{(F)}\big), \tag{2.1}$$

$$Pr\big(X_i^{(F)} = x_i^{(F)} \,|\, H_i^{(FX)} = b, -\big) = \prod_{j=1}^{p^{(F)}} \psi_{b, x_{ij}^{(F)}}^{(F)(j)}, \tag{2.2}$$

$$\big(Y_i^{(NF)} | H_i^{(NF)} = h, -\big) \sim \mathcal{N}\big(y_i^{(NF)} | B_h^{(NF)}, \Sigma_h^{(NF)}\big), \tag{2.3}$$

$$Pr\big(X_i^{(NF)} = c_i^{(NF)} \,|\, H_i^{(NF)} = h, -\big) \sim \prod_{j=1}^{p^{(NF)}} \psi_{h, x_{ij}^{(NF)}}^{(NF)(j)}, \tag{2.4}$$

$$Pr\big(H_i^{(FY)} = a, H_i^{(FX)} = b \,|\, Z_i = z\big) = \phi_{z,a}^{(FY)} \phi_{z,b}^{(FX)}, \tag{2.5}$$

$$Pr\big(H_i^{(NF)} = h \,|\, Z_i = z\big) = \phi_{z,h}^{(NF)}, \tag{2.6}$$

$$Pr\big(Z_i = z\big) = \lambda_z. \tag{2.7}$$

$H_i^{(FY)} \in \{1, ..., k^{(FY)}\}$ is the mixture component index of $Y_i^{(F)}$. $H^{(FX)} \in \{1, ..., k^{(FX)}\}$ is the mixture component index of $X_i^{(F)}$. $H_i^{(NF)} \in \{1, ..., k^{(NF)}\}$

is the mixture component index of $Y_i^{(NF)}$ and $X_i^{(NF)}$. $Z_i \in \{1, ..., k^{(Z)}\}$ is the mixture component index of $H_i^{(F)}$ and $H_i^{(NF)}$. $B_a^{(F)}$ and $\Sigma_a^{(F)}$ are the matrix of regression coefficients and the covariance matrix in $H_i^{(FY)} = a$. $\psi^{(F)(j)}_{b,x_{ij}^{(F)}}$ is the probability of $X_{ij}^{(F)} = x_{ij}^{(F)}$ in $H_i^{(FX)} = b$. $B_h^{(NF)}$ and $\Sigma_h^{(NF)}$ are the mean vector and the covariance matrix in $H_i^{(NF)} = h$. Here, $\Sigma_h^{(NF)}$ is a diagonal matrix with non-zero entries $\left(\eta_{h,1}^{(NF)}, ..., \eta_{h,q^{(NF)}}^{(NF)}\right)$. Thus, the variables in $Y_i^{(NF)}$ are conditionally independent. Finally, $\psi^{(NF)(j)}_{h,x_{ij}^{(NF)}}$ is the probability of $X_{ij}^{(NF)} = x_{ij}^{(NF)}$ in $H_i^{(FX)} = h$.

To allow closed-form expressions for the posteriors, we take conjugacy into consideration when specifying the prior distributions. For the multinomial variables, we have

$$\psi_b^{(F)(j)} \overset{i.i.d.}{\sim} Dir\left(\gamma_{b,1}^{(j)}, ..., \gamma_{b,d_j^{(F)}}^{(j)}\right), \tag{2.8}$$

$$\psi_h^{(NF)(j)} \overset{i.i.d.}{\sim} Dir\left(\gamma_{h,1}^{(j)}, ..., \gamma_{h,d_j^{(NF)}}^{(j)}\right) \tag{2.9}$$

$$\left(\gamma_{b,1}^{(j)}, ..., \gamma_{b,d_j^{(F)}}^{(j)}\right)^T = \left(1/d_j^{(F)}, ..., 1/d_j^{(F)}\right)^T, \tag{2.10}$$

$$\left(\gamma_{h,1}^{(j)}, ..., \gamma_{h,d_j^{(NF)}}^{(j)}\right)^T = \left(1/d_j^{(NF)}, ..., 1/d_j^{(NF)}\right)^T, \tag{2.11}$$

For the multivariate normal variables, we have

$$Pr\left(B_a^{(F)}, \Sigma_a^{(F)}\right) = \mathcal{N}\left(B_0^{(F)}, I, T_B^{(F)}\right) \times \mathcal{IW}\left(\nu^{(F)}, \Sigma^{(F)}\right), \tag{2.12}$$

$$Pr\left(B_h^{(NF)}\right) = \mathcal{N}\left(B_0^{(NF)}, T_B^{(NF)}\right), \tag{2.13}$$

$$Pr\left(\eta_{h,j}^{(NF)}\right) = \mathcal{IG}\left(\nu^{(NF)}, \eta_j^{(NF)}\right), \tag{2.14}$$

where $T_B^{(F)} = Diag\left(\tau_1^{(F)}, ..., \tau_{q^{(F)}}^{(F)}\right)$ and $T_B^{(NF)} = Diag\left(\tau_1^{(NF)}, ..., \tau_{q^{(NF)}}^{(NF)}\right)$, and

$$\tau_j^{(F)} \overset{i.i.d.}{\sim} \mathcal{G}\left(\alpha_{\tau^{(F)}}, \beta_{\tau^{(F)}}\right), \tag{2.15}$$

$$\tau_j^{(NF)} \overset{i.i.d.}{\sim} \mathcal{G}\left(\alpha_{\tau^{(NF)}}, \beta_{\tau^{(NF)}}\right). \tag{2.16}$$

For the hyper-prior distributions, we have

$$\left(B_0^{(F)}, \Sigma^{(F)}\right) \sim \mathcal{N}\left(0, I, \sigma_0^{(F)^2} I\right) \times \mathcal{W}\left(\omega^{(F)}, \Sigma_0^{(F)}\right), \tag{2.17}$$

$$\left(B_0^{(NF)}\right) \sim \mathcal{N}\left(0, \sigma_0^{(NF)^2} I\right), \tag{2.18}$$

$$\left(\eta_j^{(NF)}\right) \sim \mathcal{IG}\left(\nu^{(NF)}, \eta_0^{(NF)}\right). \tag{2.19}$$

We let $\nu^{(F)} = q^{(F)} + 2$, $\nu^{(NF)} = 2$, $\omega^{(F)} = q^{(F)} + 1$, $\omega^{(NF)} = 1$, $\Sigma_0^{(F)} = I/(q^{(F)} + 1)$, and $\eta_0^{(NF)} = 1$.

The hierarchical priors for the latent variables follow a truncated version of the stick-breaking construction of the Dirichlet process ([17], [8]). We have

$$\phi_{z,a}^{(FY)} = V_{z,a}^{(FY)} \prod_{l<a} (1 - V_{z,l}^{(FY)}), \quad V_{z,a}^{(FY)} \overset{i.i.d.}{\sim} \mathcal{B}(1, \beta^{(FY)}), \quad V_{z,k^{(FY)}}^{(FY)} = 1, \tag{2.20}$$

$$\phi_{z,b}^{(FX)} = V_{z,b}^{(FX)} \prod_{l<b} (1 - V_{z,l}^{(FX)}), \quad V_{z,b}^{(FX)} \overset{i.i.d.}{\sim} \mathcal{B}(1, \beta^{(FX)}), \quad V_{z,k^{(FX)}}^{(FX)} = 1, \tag{2.21}$$

$$\phi_{z,h}^{(NF)} = V_{z,h}^{(NF)} \prod_{l<h} (1 - V_{z,l}^{(NF)}), \quad V_{z,h}^{(NF)} \overset{i.i.d.}{\sim} \mathcal{B}(1, \beta^{(NF)}), \quad V_{z,k^{(NF)}}^{(NF)} = 1, \tag{2.22}$$

$$\lambda_z = W_z \prod_{l<z} (1 - W_l), \quad W_z \overset{i.i.d.}{\sim} \mathcal{B}(1, \alpha), \quad W_{k^{(Z)}} = 1. \tag{2.23}$$

Details about the method of fitting the model can be found in Chapter 4 of [20].

Figure 1 is a graphical representation of HCMM-FNF. It is apparent that dependence between $X^{(F)}$ and all variables in $\mathcal{NF}$ is captured only by the lowest level of mixture components, which could make accurate estimation of these associations difficult. Dependence between $Y^{(F)}$ and all variables in $\mathcal{NF}$ is captured via the component regressions and the lowest level of mixture components.

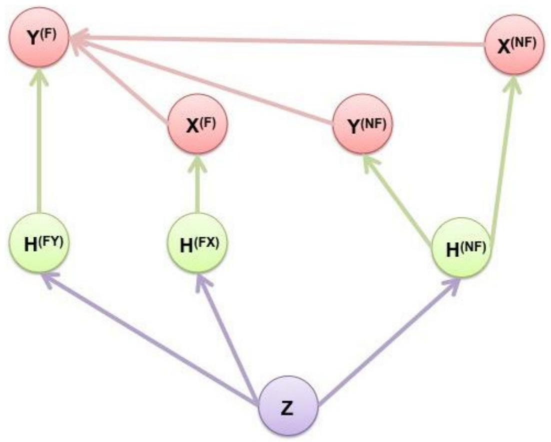


Figure 1: Graphical model representation of HCMM-FNF. $X^{(F)}$, $Y^{(F)}$, $X^{(NF)}$, and $Y^{(NF)}$ are the observed categorical and continuous variables. $H^{(F)}$ and $H^{(NF)}$ are the mixture components of $\mathcal{F}$ and $\mathcal{NF}$ variables, respectively. Z is the mixture component for $H^{(F)}$ and $H^{(NF)}$.

While this encodes dependence between $Y^{(F)}$ and all variables in $\mathcal{NF}$, we expect HCMM-FNF to do a better job capturing the joint distribution among variables within $\mathcal{F}$ than the relationships of $Y^{(F)}$ with variables in $\mathcal{NF}$, as the variables within $\mathcal{F}$ share mixture components directly. This suggests that when the associations between some variables in $Y^{(F)}$ and $Y^{(NF)}$ are strong or nonlinear, it may be advantageous to put all those variables in $\mathcal{F}$. Similarly, when $Y^{(F)}$ and

$X^{(NF)}$ are highly associated, moving $X^{(NF)}$ to $\mathcal{F}$ may improve the estimation of the associations between $Y^{(F)}$ and $X^{(NF)}$. Similarly, when some variables in $Y^{(NF)}$ are highly associated with $X^{(F)}$, or when some variables in $X^{(NF)}$ are highly associated with $X^{(F)}$, moving them to $\mathcal{F}$ could help the model estimate the associations.

These observations motivate why *Move* could lead to improved estimation over *Stay*. We now explore that possibility using simulation studies.

3. SIMULATION STUDIES

We investigate the potential of *Move* to improve the quality of imputations using four simple scenarios. To describe each scenario, let (F_0) index the focus variables automatically included in $\mathcal{F}$, i.e., those with high rates of missing values, and (NF_0) index the other variables. The sets of variables defined by (F_0) and (NF_0), which we call $\mathcal{F}_0$ and $\mathcal{NF}_0$, respectively, are those used in *Stay*. In *Move*, we put some variables in $\mathcal{NF}_0$ in $\mathcal{F}$.

3.1. Simulation scenarios and evaluation metrics

In Scenario 1, we make variables in $X^{(NF_0)}$ highly associated with some variables in $X^{(F_0)}$. We generate six binary $X^{(NF_0)}$ variables from an arbitrarily chosen joint distribution, constructed from a mixture of products of multinomial distributions. To create the dependencies between the categorical variables in $\mathcal{F}_0$ and $\mathcal{NF}_0$, we generate four $X^{(F_0)}$ variables according to Bernoulli distributions with $Pr\big(X_j^{(F_0)} = x \,|\, X_j^{(NF_0)} = x\big) = 0.9$, with $x \in \{1, 2\}$ for $j = 1, ..., 4$. Under *Move*, we put $\big(X_1^{(NF_0)}, ..., X_4^{(NF_0)}\big)$ in $\mathcal{F}$.

In Scenario 2, we make some variables in $Y^{(NF_0)}$ highly associated with variables in $X^{(F_0)}$. We generate six $Y^{(NF_0)}$ variables from an arbitrary mixture of normal distributions. We create four binary $X^{(F_0)}$ variables from Bernoulli distributions with

$$log\left(\frac{Pr\big(X_j^{(F_0)} = 2 \,|\, Y_j^{(NF_0)} = y_j^{(NF_0)}\big)}{Pr\big(X_j^{(F_0)} = 1 \,|\, Y_j^{(NF_0)} = y_j^{(NF_0)}\big)}\right) = y_j^{(NF_0)}, \tag{3.1}$$

for $j = 1, ..., 4$. Under *Move*, we put $\big(Y_1^{(NF_0)}, ..., Y_4^{(NF_0)}\big)$ in $\mathcal{F}$.

In Scenario 3, we make some variables in $X^{(NF_0)}$ highly associated with $Y^{(F_0)}$. We generate six binary $X^{(NF_0)}$ variables from an arbitrarily chosen mixture of products of multinomial distributions. We generate four $Y^{(F_0)}$ according to

$\big(Y_j^{(F_0)}|X_j^{(NF_0)}= x_j^{(NF_0)}\big) \sim \mathcal{N}\big(y_j^{(F_0)}|x_j^{(NF_0)}, 0.005\big)$, with $j = 1, ..., 4$. Under *Move*, we put $\big(X_1^{(NF_0)}, ..., X_4^{(NF_0)}\big)$ in $\mathcal{F}$.

In Scenario 4, we make some variables in $Y^{(NF_0)}$ highly associated with $Y^{(F_0)}$. We generate six $Y^{(NF_0)}$ variables from an arbitrarily chosen mixture of normal distributions. We generate four $Y^{(F_0)}$ according to $\big(Y_j^{(F_0)}|Y_j^{(NF_0)}= y_j^{(NF_0)}\big) \sim \mathcal{N}\big(0.9y_j^{(NF_0)}, 0.005\big)$, for $j = 1, ..., 4$. Under *Move*, we put $\big(Y_1^{(NF_0)}, ..., Y_4^{(NF_0)}\big)$ in $\mathcal{F}$.

We use two evaluation metrics in the simulations. Let $q_{k,j,l}^{(s)}$ be the k^{th} quantity of interest in the j^{th} repeated sample for the l^{th} imputation. The superscript (s) indicates that the estimate is from *Stay*. Similarly, we define $q_{k,j,l}^{(m)}$ for the estimate obtained from *Move*. Notations without any superscripts and subscript l, such as $q_{k,j}$, stand for the quantities from the truth, defined as the complete data without any missing values.

Metric I: We define the absolute differences as $d_{k,j,l}^{(s)} = |q_{k,j,l}^{(s)} - q_{k,j}|$ for *Stay* and $d_{k,j,l}^{(m)} = |q_{k,j,l}^{(m)} - q_{k,j}|$ for *Move*. We compute $d_{k,j}^{(s)} = (1/L)\sum_{l=1}^{L} d_{k,j,l}^{(s)}$ and $d_{k,j}^{(m)} = (1/L)\sum_{l=1}^{L} d_{k,j,l}^{(m)}$. For each quantity, we conduct a paired t-test of the hypothesis $H_0 : \mu_k^{(s)} = \mu_k^{(m)}$, where $\mu_k^{(s)}$ is the population mean of $d_{k,j}^{(s)}$ and $\mu_k^{(m)}$ is the population mean of $d_{k,j}^{(m)}$. When the p-value is below 0.01, we consider the difference between *Stay* and *Move* statistically significant.

Metric II: We define the percentage changes as $\Delta d_{k,j,l}^{(s)} = \frac{q_{k,j,l}^{(s)} - q_{k,j}}{q_{k,j}} \times 100\%$ for *Stay* and $\Delta d_{k,j,l}^{(m)} = \frac{q_{k,j,l}^{(m)} - q_{k,j}}{q_{k,j}} \times 100\%$ for *Move*. This metric is useful when the quantities of interest are not in the same units. For each quantity k, we let $\Delta d_k^{(s)} = (1/JL)\sum_{j=1}^{J}\sum_{l=1}^{L} \Delta d_{k,j,l}^{(s)}$ and $\Delta d_k^{(m)} = (1/JL)\sum_{j=1}^{J}\sum_{l=1}^{L} \Delta d_{k,j,l}^{(m)}$. We then draw box plots for all $\{\Delta d_k^{(s)}\}$ and $\{\Delta d_k^{(m)}\}$ of the same type. For example, we draw box plots of $\{\Delta d_k^{(s)}\}$ and $\{\Delta d_k^{(m)}\}$ for all possible correlations between $Y^{(F)}$ and $Y^{(NF)}$.

3.2. Results

For each scenario, we generate 100 independent datasets comprising $n = 1,000$ observations. For some variables, we make 50% of values missing completely at random (MCAR) and automatically put them in $\mathcal{F}_0$; for the remainder, we make only 1% MCAR and put them in $\mathcal{NF}_0$. In each incomplete dataset, we fit HCMM-FNF with *Move* and *Stay*, using 25,000 iterations as burn-in, which is sufficient based on standard diagnosis of MCMC convergence. After burnin, we run the chains for 1,000 iterations, and from these keep $L = 10$ imputations spaced 100 iterations apart.

Figure 2 displays results from Scenario 1 for bivariate probabilities between the categorical variables in $\mathcal{F}_0$ and $\mathcal{NF}_0$. Generally, the cell probabilities are estimated more accurately under *Move* than *Stay*. The improvements are most noticeable in the probabilities involving $\big(X_j^{(NF_0)}, X_j^{(F_0)}\big)$ where $j = 1, ..., 4$. Detailed investigation of the box plots for small values of Metric II indicates that the percentage changes under *Move* are generally smaller than those under *Stay*.

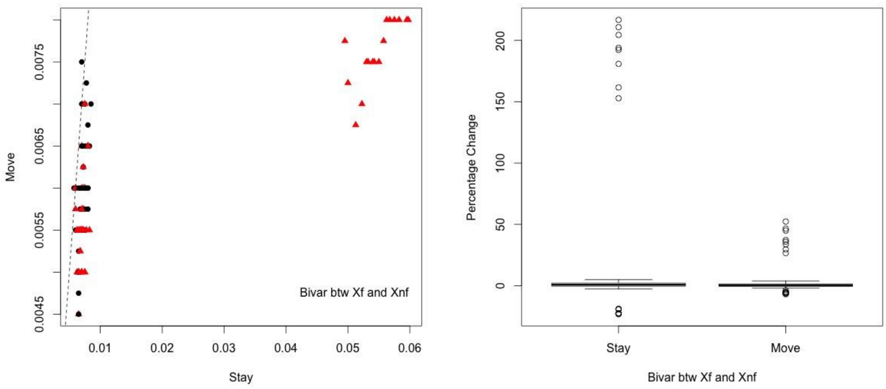


Figure 2: Bivariate cell probabilities for *Stay* and *Move* in Scenario 1. The left plot shows Metric I, where triangles correspond to p-values below 0.01 when testing for average differences in the two strategies. The right plot shows Metric II. The median of the relative differences is 0.0 for both *Stay* and *Move*.

In Scenario 2, we examine the coefficients of the logistic regressions of each $X^{(F_0)}$ variable on each $Y^{(NF_0)}$ variable. As evident in Figure 3, these coefficients are estimated more accurately in *Move* than in *Stay*. The accuracy gains are largest for the coefficients involving $\big(X_j^{(F_0)}, Y_j^{(NF_0)}\big)$ where $j = 1, ..., 4$.

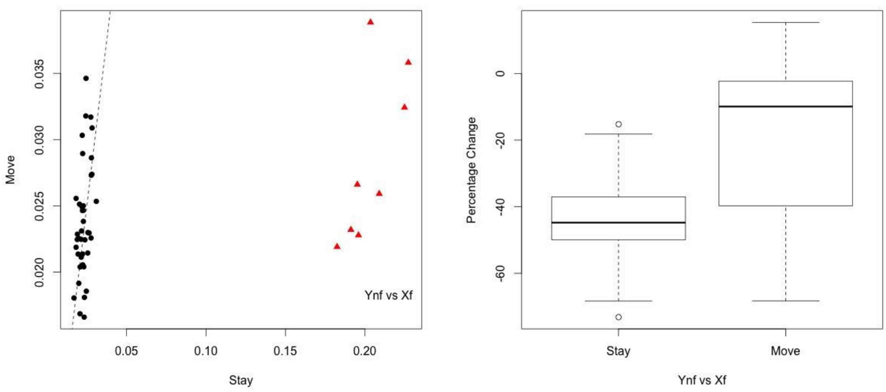


Figure 3: Coefficients in logistic regressions for *Stay* and *Move* in Scenario 2. The left plot shows Metric I, where triangles correspond to p-values below 0.01 when testing for average differences in the two strategies. The right plot shows Metric II. The median of the relative differences is -44.8 for *Stay* and -9.9 for *Move*.

In Scenario 3, we are interested in the associations between the variables in $Y^{(F_0)}$ and $X^{(NF_0)}$. We measure these associations using logistic regressions of $X_j^{(NF_0)}$ on $Y_k^{(F_0)}$ for $j \in \{1, ..., 4\}$ and $k \in \{1, ..., 6\}$. As evident in Figure 4, there are no significant differences between *Move* and *Stay* on Metric I. The box plots for Metric II show that the two medians are close, although the spread of values for *Move* is smaller than that for *Stay*.

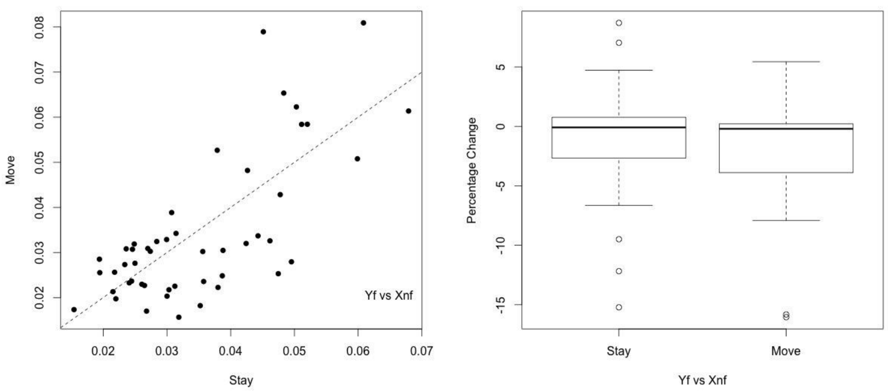


Figure 4: Coefficients in logistic regressions for *Stay* and *Move* in Scenario 3. The left plot shows Metric I, and the right plot shows Metric II. The median of the relative differences is −0.09 for *Stay* and −0.10 for *Move*.

For Scenario 4, Figure 5 displays results for the pairwise correlations of variables in $Y^{(F_0)}$ and $Y^{(NF_0)}$. There are no significant differences between *Move* and *Stay* for Metric I or Metric II.

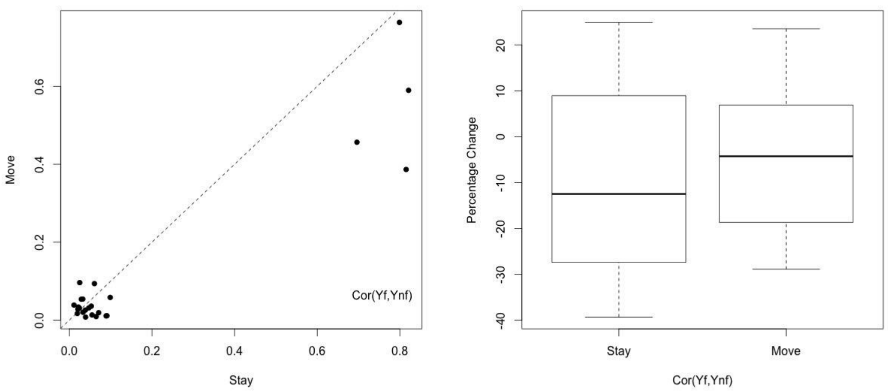


Figure 5: Pairwise correlations for *Move* and *Stay* in Scenario 4. The left plot shows Metric I, and the right plot shows Metric II. The median of the relative differences is −12.4 for *Stay* and −4.3 for *Move*.

3.3. Summary of results

When using *Stay*, associations between $X^{(F_0)}$ and $X^{(NF_0)}$ are estimated only through the tensor factorization. Apparently, in Scenario 1 this is not sufficient to capture the dependence. In contrast, by using common mixture components for all the categorical variables in $\mathcal{F}$, *Move* captures the dependence structure in Scenario 1 more effectively than *Stay*. We reach similar findings for Scenario 2, in which the local dependence enabled by *Move* captures associations involving $X^{(F_0)}$ and $Y^{(NF_0)}$ more effectively than relying only on the tensor factorization to capture the dependence. These results are in accord with the motivation we gave at the end of Section 2 for moving some (nearly) completely observed variables to $\mathcal{F}$.

For the associations between $Y^{(F_0)}$ and $\mathcal{NF}_0$, *Move* does not offer significant benefits over *Stay* in Scenarios 3 and 4. Apparently, *Stay* adequately incorporates the dependence between $Y^{(F_0)}$ and $\left(X^{(F_0)}, X^{(NF_0)}, Y^{(NF_0)}\right)$ through the mixture component regressions, so that moving variables to $\mathcal{F}$ does not noticeably improve the imputation quality. We also tried four modifications of these scenarios that use nonlinear associations between $Y^{(F_0)}$ and variables in $\mathcal{NF}_0$; see [20] for details of the designs. The performances of *Move* and *Stay* were qualitatively similar. Apparently, by using mixture distributions for the focus variables, we potentially can capture nonlinear relationships among the continuous focus variables.

4. EMPIRICAL STUDY

The findings in Section 3.3 are based on stylized simulation scenarios designed to clarify when *Move* can be advantageous. Further, in the studies we moved the nearly completely observed variables known to have strong associations with the variables in $\mathcal{F}_0$; in genuine settings we need empirical measures to identify these variables. In this section we present such measures and investigate whether or not similar behavior holds for genuine data.

4.1. Illustrative Data: The American Community Survey

The American Community Survey (ACS), an ongoing survey conducted by the U.S. Census Bureau, collects demographic, housing, social, and economic data from sampled households along with information on the people who live in these households. It is a rich and dynamic resource for public policy decision making and analysis. Researchers can access public use files from the Integrated Public Use Microdata Series (IPUMS, usa.ipums.org). Relationships among variables in the ACS can be complex and difficult to capture with standard imputation models ([15]). Thus, we can benefit from using HCMM-FNF for imputation modeling.

We subset the ACS data to include only household heads who own their living units, were employed during the year of 2010 in the state of North Carolina, and have complete data; this subset has 19,492 cases. We systematically sample 1,026 household heads as our working dataset. To facilitate reasonable computation time, we choose the 16 variables in Table 1. Since IPUMS processes the raw data, the percentage of missing values for each variable in the IPUMS file is less than 2%. We therefore introduce additional missing values for purposes of the empirical study.

Before presenting results, we note that we repeated both studies on a second random sample of 1,026 qualifying household heads. The patterns are very similar to the ones presented here; see Chapter 4 of [20] for details.

Table 1: Variables in ACS empirical study. First four variables are for households; the remainder are for the head of the household. *Cts* is short for continuous, and *Cat* is short for categorical. # Levels is the number of levels of the categorical variable. PROPTX99 is categorical with a large number of levels, and is modeled as such. It is treated as continuous when we report results.

Name	Label	Cts./Cat.[#Levels]
PROPTX99	Annual property taxes	Cat[67]
COSTELEC	Annual electricity cost	Cts
COSTGAS	Annual gas cost	Cts
COSTWATR	Annual water cost	Cts
AGE	Age	Cts
SEX	Sex	Cat[2]
MARST	Marital status	Cat[6]
RACE	Race	Cat[7]
HCOVANY	Any health insurance coverage	Cat[2]
EDUC	Educational attainment	Cat[9]
SCHLTYPE	Public or private school	Cat[3]
INCTOT	Total personal income	Cts
OCCSCORE	Occupational income score	Cts
PWTYPE	Place of work: metropolitan status	Cat[5]
MIGRATE1	Migration status, 1 year	Cat[4]
DIFFSENS	Vision or hearing difficulty	Cat[2]

4.2. Studies

As the measure to determine which variables to move into $\mathcal{F}$, we use the relative mutual information. For any two continuous variables A and B, the mutual information is

$$I(A, B) = \int_B \int_A p(a, b)\ log\left(\frac{p(a, b)}{p(a)p(b)}\right) da\, db. \tag{4.1}$$

The relative mutual information with respect to a variable A is a ratio of $I(A, B)$ over $I(A, A)$. For categorical variables, we replace the integrals with summations.

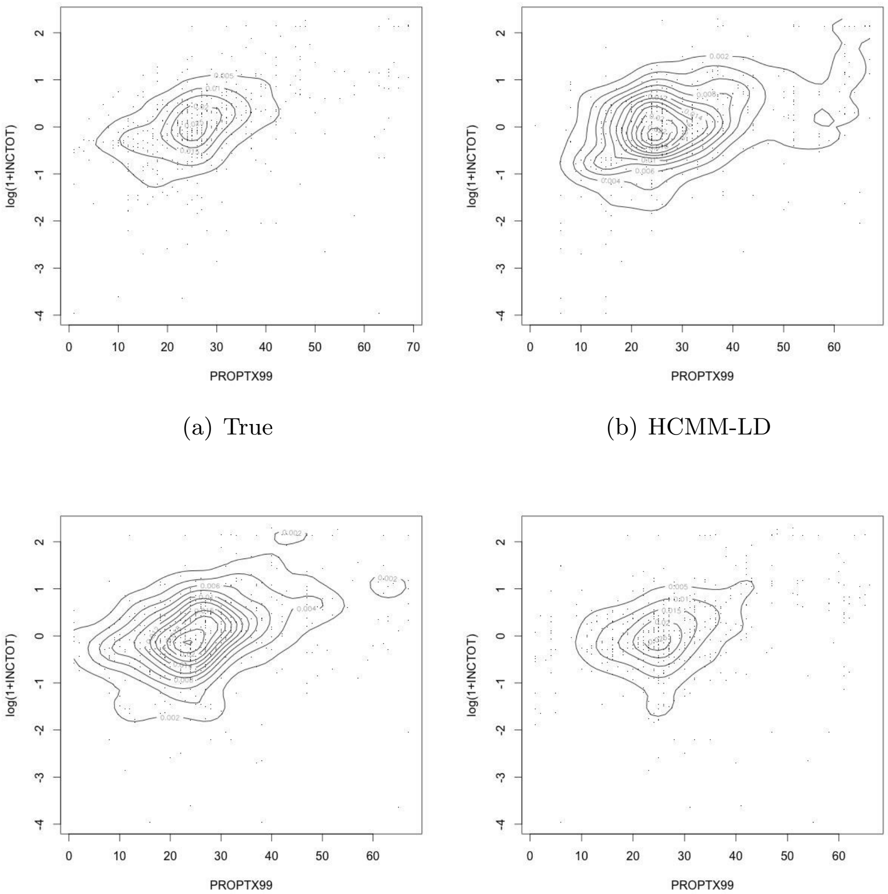
We run two studies, which we call the high and low mutual information studies. In each study, we impute the missingness in the working dataset using three models: HCMM-FNF with *Stay*, HCMM-FNF with *Move*, and the mixture model of [15], which we label HCMM-LD. HCMM-LD does not use any focused clustering, essentially putting all variables in $\mathcal{F}$. We use the performance of HCMM-LD as a benchmark for *Stay* and *Move*.

High Mutual Information (HMI) Study

We begin with a study in which variables in $\mathcal{NF}_0$ are predictive of variables in $X^{(F_0)}$, i.e., they share high amounts of mutual information. From the categorical variables in Table 1, we assign EDUC and PROPTX99 to have 50% values MCAR and thus to be in $\mathcal{F}_I$ automatically. We assign INCTOT, OCCSCORE, AGE, COSTELEC, COSTGAS, and COSTWATR as $Y^{(NF_0)}$, and the remaining variables as $X^{(NF_0)}$. Variables in $\mathcal{NF}_0$ have 1% values MCAR.

INCTOT and OCCSCORE have relatively high mutual information with EDUC and PROPTX99 with values at 0.26 and 0.22, respectively. All other values are 0.11 or lower, with all but two being below 0.05. Thus, we add INCTOT and OCCSCORE to the focus variables under *Move*. We analyze the marginal probabilities of PROPTX99 and EDUC, and pay special attention to associations between the variables in $\mathcal{F}$ after *Move*.

Figure 6 displays contour plots from the kernel density estimates of the standardized values of $log(1 + INCTOT)$ and PROPTX99 for the missing observations. The true density is unimodal, concentrated in the area with PROPTX99 from $(5, 45)$ and $log(1 + INCTOT)$ from $(-1.5, 1.2)$. By comparison, the completed data density estimates under HCMM-LD and *Stay* have a large spread and distorted contours. The density estimate under *Move* looks most similar to the truth.

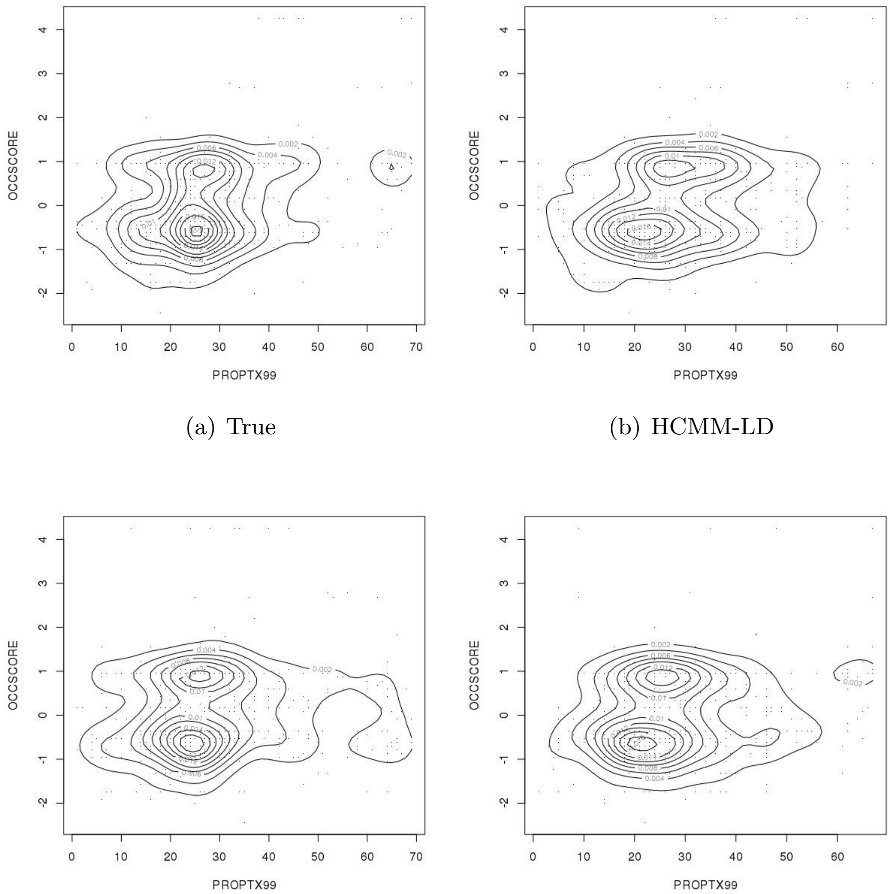


(a) True (b) HCMM-LD

(c) *Stay* (d) *Move*

Figure 6: Contour plots from the kernel density estimates of $log(1 + INCTOT)$ (standardized) and PROPTX99 for the missing observations in the *HMI* study. Each completed-data plot is from one randomly selected dataset.

Figure 7 displays the kernel density estimate of the standardized OCCSCORE and PROPTX99 for the missing observations. The true density has two high density, connected modes and one low density, isolated mode. The small mode reflects household heads whose occupational score is around 1 (41 on the original scale) and pay a high amount for their property taxes. Both HCMM-LD and *Stay* have trouble capturing this isolated mode; *Move* captures it more effectively than the other models. There are no significant differences among the three models for other quantities, including the marginal cell counts of EDUC and the bivariate associations involving EDUC. Details can be found in Chapter 4 of [20].



(a) True (b) HCMM-LD

(c) *Stay* (d) *Move*

Figure 7: Contour plots from the kernel density of OCCSCORE (standardized) and PROPTX99 for the missing observations in the *HMI* study. Each completed-data plot is from one randomly selected dataset.

Low Mutual Information (LMI) Study

We next consider a study where we treat EDUC and DIFFSENS as $X^{(F_0)}$, INCTOT and OCCSCORE as $Y^{(F_0)}$, PROPTX99, SEX, RACE, MARST, MIGRATE1, HCOVANY, and PWTYPE as $X^{(NF_0)}$, and the remaining variables as $Y^{(NF_0)}$. We again make 50% of values MCAR for variables in $\mathcal{F}_0$ and 1% of values MCAR for variables in $\mathcal{NF}_0$. The four variables in $\mathcal{F}_0$ frequently are used to assess socioeconomic status, which motivates why we create a simulation where they are the variables with high rates of missing data.

PROPTX99 has high relative mutual information with INCTOT and OCCSCORE as described previously. It also has relative mutual information values of 0.16 for EDUC and DIFFSENS, the two categorical focus variables. Other relationships are comparatively weak, with only one value exceeding 0.10 (AGE and DIFFSENS at 0.13). Thus, we add only PROPTX99 to the focus variables under *Move*.

Based on results in Section 3, we do not expect moving PROPTX99 to $\mathcal{F}$ to improve the quality of imputations substantially. In the simulations of Scenario 3 where we moved categorical variables highly associated with continuous $Y^{(F_0)}$, which most closely matches the characteristics of the *LMI* setting, *Move* and *Stay* had similar performances. The results from *LMI* bear this out. We compare the marginal probability densities of INCTOT and OCCSCORE, the marginal cell counts of EDUC and DIFFSENS, the joint distributions of (INCTOT, OCCSCORE), (INCTOT, PROPTX99), and (OCCSCORE, PROPTX99), and the associations of (INCTOT, EDUC), (OCCSCORE, EDUC), (PROPTX99, EDUC), (INCTOT, DIFFSENS), (OCCSCORE, DIFFSENS), and (PROPTX99, DIFFSENS). We find that *Stay* and *Move* perform very similarly. They also are not very different from HCMM-LD. To save space, we do not present these results here; details are in Chapter 4 of [20].

5. CONCLUSION

In general, the results of the artificial data simulations and the empirical study tell a consistent story. Compared to *Stay*, *Move* can improve estimation of the distribution of focus categorical variables, particularly for their associations with the variables moved to $\mathcal{F}$. *Move* improved the estimate of the association between INCTOT and PROPTX99, as well as OCCSCORE and PROPTX99, in *HMI*. The degree of improvement depends on the strength of the association between $X^{(F_0)}$ and $\mathcal{NF}_0$. This is evident in the result that *Move* did not substantially improve the accuracy of estimates involving EDUC in both *HMI* and *LMI*, as well as those involving DIFFSENS in *LMI*. For continuous variables in $\mathcal{F}_0$, *Stay* and *Move* performed similarly, suggesting that *Move* does not help much in terms of accuracy when the initial focus variables are continuous.

As a final comment, we note that *Move* and *Stay* can offer computational advantages over HCMM-LD. With HCMM-LD, one models all continuous variables with a multivariate normal distribution, which can result in a large number of covariance parameters when there are many continuous variables. In contrast, both *Stay* and *Move* assume that $Y^{(NF)}$ are locally independent, thereby removing them from the multivariate normal distributions.

ACKNOWLEDGMENTS

This work has been supported by the grant NSF SES 1131897.

REFERENCES

[1] Banerjee, A.; Murray, J. and Dunson, D.B. (2013). Bayesian learning of joint distributions of objects, *Journal of Machine Learning Research Workshop and Conference Proceedings*, **31**, 1–9.

[2] DeYoreo, M. and Kottas, A. (2015). A fully nonparametric modeling approach to binary regression, *Bayesian Analysis*, **10**, 821–847.

[3] DeYoreo, M. and Kottas, A. (2017). Bayesian nonparametric modeling for multivariate ordinal regression, *Journal of Computational and Graphical Statistics*, Forthcoming.

[4] DeYoreo, M.; Reiter, J.P. and Hillygus, D.S. (2017). Nonparametric Bayesian models with focused clustering for mixed ordinal and nominal data, *Bayesian Analysis*, **12**, 679–703.

[5] Dunson, D.B. and Xing, C. (2009). Nonparametric Bayes modeling of multivariate categorical data, *Journal of the American Statistical Association*, **104**, 1042–1051.

[6] Hannah, L.A.; Blei, D.M. and Powell, W.B. (2011). Dirichlet process mixtures of generalized linear models, *Journal of Machine Learning Research*, **12**, 1923–1953.

[7] Hu, J.; Reiter, J.P. and Wang, Q. (2018). Dirichlet process mixture models for modeling and generating synthetic versions of nested categorical data, *Bayesian Analysis*, **13**, 183–200.

[8] Ishwaran, H. and James, L.F. (2001). Gibbs sampling methods for stick-breaking priors, *Journal of the American Statistical Association*, **96**, 161–173.

[9] Kim, H.J.; Reiter, J.P.; Wang, Q.; Cox, L.H. and Karr, A.F. (2014). Multiple imputation of missing or faulty values under linear constraints, *Journal of Business & Economic Statistics*, **32**, 375–386.

[10] Kim, H.J.; Cox, L.H.; Karr, A.F.; Reiter, J.P. and Wang, Q. (2015). Simultaneous edit-imputation for continuous microdata, *Journal of the American Statistical Association*, **110**, 987–999.

[11] Manrique-Vallier, D. and Reiter, J.P. (2013). Bayesian multiple imputation for large-scale categorical data with structural zeros, *Survey Methodology*, **40**, 125–134.

[12] Manrique-Vallier, D. and Reiter, J.P. (2014). Bayesian estimation of discrete multivariate latent structure models with structural zeros, *Journal of Computational and Graphical Statistics*, **23**, 1061–1079.

[13] Müller, P. and Quintana, F.A. (2004). Nonparametric Bayesian data analysis, *Statistical Science*, **19**, 95–110.

[14] Müller, P. and Mitra, R. (2013). Bayesian nonparametric inference–why and how, *Bayesian Analysis*, **8**, 269–302.

[15] Murray, J.S. and Reiter, J.P. (2016). Multiple imputation of missing categorical and continuous values via Bayesian mixture models with local dependence, *Journal of the American Statistical Association*, **111**, 1466–1479.

[16] Rubin, D.B. (1987). Multiple imputation for nonresponse in surveys, *Wiley Series in Probability and Mathematical Statistics: Applied probability and statistics*, Wiley & Sons, New York.

[17] Sethuraman, Jayaram (1994). A constructive definition of Dirichlet priors, *Statistica Sinica*, **4**, 639–650.

[18] Si, Y. and Reiter, J.P. (2013). Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys, *Journal of Educational and Behavioral Statistics*, **38**, 499–521.

[19] Wade, S.; Dunson, D.B.; Petrone, S. and Trippa, L. (2014). Improving prediction from Dirichlet process mixtures via enrichment, *Journal of Machine Learning Research*, **15**, 1041–1071.

[20] Wei, L. (2016). Methods for Imputing Missing Values and Synthesizing Confidential Values for Continuous and Magnitude Data, *Duke University*.