Nonparametric Bayesian Multiple Imputation for Incomplete Categorical Variables in Large-Scale Assessment Surveys

Corresponding Author: Jerome P. Reiter

Mrs. Alexander Hehmeyer Associate Professor of Statistical Science

Department of Statistical Science, Duke University

Box 90251, Duke University, Durham, NC 27708

e-mail: jerry@stat.duke.edu

phone: 919-668-5227. fax: 919-684-8594.

Co-author: Yajuan Si

Biographical Note: Yajuan Si is a Postdoctoral Research Scholar in the Department of Statistics at Columbia University, New York, NY 10027; e-mail: sophie2012@gmail.com. Her research interests include methods for handling missing data and complex Bayesian modeling. Jerome P. Reiter is Mrs. Alexander Hehmeyer Associate Professor of Statistical Science, at Duke University, Box 90251, Duke University, Durham, NC 27708; email: jerry@stat.duke.edu. His research interests include methods for protecting confidentiality in public use data, methods for handling missing data, and Bayesian methods for complex surveys.

Abstract

In many surveys, the data comprise a large number of categorical variables that suffer from item nonresponse. Standard methods for multiple imputation, like log-linear models or sequential regression imputation, can fail to capture complex dependencies and can be difficult to implement effectively in high dimensions. We present a fully Bayesian, joint modeling approach to multiple imputation for categorical data based on Dirichlet process mixtures of multinomial distributions. The approach automatically models complex dependencies while being computationally expedient. The Dirichlet process prior distributions enable analysts to avoid fixing the number of mixture components at an arbitrary number. We illustrate repeated sampling properties of the approach using simulated data. We apply the methodology to impute missing background data in the 2007 Trends in International Mathematics and Science Study.

Key Words: Dirichlet Process, Latent Class, Missing, Mixture

1 Introduction

Large-scale surveys of educational progress, such as the National Assessment of Educational Progress, the Programme for International Student Assessment, and the Trends in International Mathematics and Science Study (TIMSS), typically collect many categorical variables on data subjects. These variables can include, for example, students' demographic information, interests, activities, and study habits, as well as information about teachers and schools. Such background variables are central for data analysis, modeling, and policy research (e.g., Thomas, 2002; Ballou *et al.*, 2004; von Davier and Sinharay, 2007, 2010). However, background variables often suffer from significant item nonresponse. For example, for the subset of the 2007 TIMSS data file that we analyze, only 4,385 out of 90,505 students have complete data on a set of 80 background variables.

As is well-known (Little and Rubin, 2002), using only the complete cases (all variables are

observed) or available cases (all variables for the particular analysis are observed) can cause problems for statistical inferences, even when variables are missing at random (Rubin, 1976). By tossing out cases with partially observed data, both approaches sacrifice information that could be used to increase precision. Further, using available cases complicates model comparisons, since different models could be estimated on different sets of cases; standard model comparison strategies do not account for such disparities. For educational data collected with complex survey designs, using available cases complicates survey-weighted inference, since the original weights are no longer meaningful for the available samples.

An alternative to complete/available cases is to fill in the missing items with multiple imputations (Rubin, 1987). The basic idea is to simulate values for the missing items by sampling repeatedly from predictive distributions. This creates $m > 1$ completed datasets that can be analyzed or, as relevant for many data producers, disseminated to the public. When the imputation models meet certain conditions (Rubin, 1987, Chapter 4), analysts of the m completed datasets can make valid inferences using complete-data statistical methods and software. Specifically, the analyst computes point and variance estimates of interest with each dataset and combines these estimates using simple formulas developed by Rubin (1987). These formulas serve to propagate the uncertainty introduced by missing data and imputation through the analyst's inferences. See Rubin (1996), Barnard and Meng (1999), and Reiter and Raghunathan (2007) for reviews of multiple imputation.

In this article, we present a fully Bayesian, joint modeling approach to multiple imputation for high-dimensional categorical data. The approach is motivated by missing values among background variables in TIMSS. In such high dimensions (80 categorical variables), typical multiple imputation methods for categorical data, like log-linear models and sequential regression strategies (Raghunathan *et al.*, 2001), can fail to capture complex dependencies and can be difficult to implement effectively. We model the implied contingency table of the background variables as a mixture of independent multinomial distributions, estimating

the mixture distributions nonparametrically with Dirichlet process prior distributions as in Dunson and Xing (2009). Our approach is related to those of Vermunt *et al.* (2008) and Gebregziabher and DeSantis (2010), who also use mixtures of multinomials for multiple imputation of categorical data. Both Vermunt *et al.* (2008) and Gebregziabher and DeSantis (2010) require an *ad hoc* selection of a fixed number of mixture components, whereas we avoid this difficulty via nonparametric modeling. Vermunt *et al.* (2008) use repeated maximum likelihood estimation on bootstrapped samples to approximate draws of imputation model parameters, and Gebregziabher and DeSantis (2010) use a routine for drawing parameters that appears to rely on maximum likelihood estimates (we were unable to determine their method exactly from their article). In contrast, we use a fully Bayesian approach that remains computationally efficient.

The reminder of this article is organized as follows. In Section 2, we review several approaches to multiple imputation for categorical data and describe their shortcomings in high dimensions. In Section 3, we present the nonparametric Bayesian multiple imputation approach, including an MCMC algorithm for computation. We also further contrast the fully Bayesian approach with the approach of Vermunt *et al.* (2008). In Section 4, we evaluate frequentist properties of the procedure with simulations. In Section 5, we apply the approach to create multiply-imputed background characteristics for students in the 2007 TIMSS data. We present results of imputation model diagnostics based on posterior predictive checks. In Section 6, we conclude with a brief discussion of future research directions.

2 Review of Existing Approaches

Vermunt *et al.* (2008) offer an excellent summary of the shortcomings of standard multiple imputation methods for categorical data in high dimensions. We augment their discussion here, focusing on imputation via log-linear models and sequential regression techniques as

these methods explicitly account for discrete data. Imputation methods that treat the categorical data as continuous, e.g., as multivariate normal, can work well for some problems but are known to fail in others, even in low dimensions (Graham and Schafer, 1999; Allison, 2000; Horton *et al.*, 2003; Ake, 2005; Bernaards *et al.*, 2007; Finch, 2010; Yucel *et al.*, 2011).

Log-linear models are a natural choice for imputation of categorical data (Schafer, 1997). However, log-linear models have known limitations in high dimensions (Erosheva *et al.*, 2002). Model selection becomes very challenging, as the number of possible models is enormous. With large dimensions it is impossible to enumerate all possible log-linear models, so that automated model selection procedures—which are complicated to implement with missing data—are necessary. With sparse tables, as is the case in practice with high dimensional categorical data, many cells of the observed contingency table randomly equal zero. Maximum likelihood estimates of the log-linear model coefficients corresponding to zero margins cannot be determined, so that one either has to assume that those cells have expected values equal to zero (Bishop *et al.*, 1975), which results in biased estimates of observed non-zero cell probabilities, or has to ensure that models do not include problematic cells, which artificially restricts the range of possible models. The latter can be problematic for missing data imputation, in that subsequent estimates of complex interactions could be attenuated due to insufficiently complex imputation models.

In sequential regression modeling (Raghunathan *et al.*, 2001), also called multiple imputation by chained equations (Van Buuren and Oudshoorn, 1999; Su *et al.*, 2010), the analyst constructs a series of univariate conditional models and imputes missing values sequentially with these models. These conditional models are in lieu of specifying a joint model for all variables. For categorical variables, the models typically are logistic or multinomial logistic regressions. Unfortunately, in large, sparse tables these conditional models suffer from similar model selection and estimation problems as log-linear models. Further, when the number of variables is large, the analyst needs to specify many conditional models, which if

done carefully is a time-consuming task. Hence, many users of chained equations use default settings that include main effects only in the conditional models. This failure to capture complex dependencies can lead to biased inferences (Vermunt *et al.*, 2008).

The model selection problem in chained equations can be obviated somewhat by using nonparametric methods as imputation engines. For example, Burgette and Reiter (2010) use classification and regression trees (CART) as the conditional models for imputation. They demonstrate improved performance over default, main-effects-only applications of multiple imputation by chained equations in simulation studies with complex dependencies. However, sequential CART imputation, and other fully conditional specifications like those in Raghunathan *et al.* (2001) and Van Buuren and Oudshoorn (1999), technically are not coherent models and thus are subject to odd behaviors. For example, the order in which variables are placed in the chain could impact the imputations (Baccini *et al.*, 2010; Li *et al.*, 2012). When automated routines are applied to selected conditional models, it is possible to have inconsistent sequences; for example, two variables are conditionally independent in one model but conditionally dependent in another, as could arise when some variable Y_a is included in the model (or tree) selected for some other variable Y_b, but Y_b is not included in the model (or tree) selected for Y_a. Inconsistent sequences generate conditional dependence relationships that depend on the order of imputation and when the chain is stopped.

3 Mixture Models for Multiple Imputation

When confronted with high dimensional categorical data with nontrivial item nonresponse, we desire a multiple imputation approach that (i) avoids the difficulties of model selection and estimation inherent in log-linear models, (ii) has theoretical grounding as a a coherent Bayesian joint model, and (iii) offers efficient computation. In seeking a Bayesian imputation model, we are following the advice of Rubin (1987), who argues that valid imputation

inferences require analysts to incorporate all sources of uncertainty, including parameter estimation. Bayesian models incorporate such uncertainty automatically.

We propose to use the Dirichlet process mixture of products of multinomial distributions model (DPMPM), which is a nonparametric Bayesian model for multivariate unordered categorical data. The DPMPM was proposed initially by Dunson and Xing (2009), who apply it to model the dependencies among nucleotides within DNA sequences involved in gene regulation. The DPMPM uses a prior distribution with full support on the space of distributions for multivariate unordered categorical variables (i.e., it includes any possible distribution), and it has been shown to be consistent for any complete table (Dunson and Xing, 2009). Further, it can be fit with a computationally efficient Gibbs sampler that scales readily and is not difficult to code.

To describe the DPMPM for multiple imputation, we first introduce a finite mixture model for multinomial data. This is essentially the latent class model used by Vermunt *et al.* (2008) and Gebregziabher and DeSantis (2010). It also is known as latent structure analysis (Lazarsfeld and Henry, 1968) and is a special case of the general diagnostic model (von Davier, 2008, 2010). The DPMPM can be viewed as a generalization of finite mixture models that allows the number of components (classes) to be infinite rather than a single best guess. We first introduce the finite mixture model and DPMPM without missing data, and then describe how to adapt each for missing data.

3.1 Finite Mixture of Products of Multinomials

Suppose that we have a complete dataset comprising n individuals and p categorical variables. Let X_{ij} be the value of variable j for individual (student) i, where $i = 1, \ldots, n$ and $j = 1, \ldots, p$. Let $X_i = (X_{i1}, \ldots, X_{ip})$. Without loss of generality, we assume that the possible values of X_{ij} are in $\{1, \ldots, d_j\}$, where $d_j \geq 2$ is the total number of categories for variable j. Let D be the contingency table formed from all levels of all p variables, so that D has

$d = d_1 \times d_2 \times \cdots \times d_p$ cells. We denote each cell in D as $(c_1, \ldots, c_p)$, where each $c_j \in \{1, \ldots, d_j\}$. For all cells in D, let $\theta_{c_1,\ldots,c_p} = \Pr(X_{i1} = c_1, \ldots, X_{ip} = c_p)$ be the probability that individual i is in cell $(c_1, \ldots, c_p)$. We require the $\sum_D \theta_{c_1,\ldots,c_p} = 1$. Let $\theta = \{\theta_{c_1,\ldots,c_p} : c_j \in (1, \ldots, d_j), j = 1 \ldots, p\}$ be the collection of all d cell probabilities.

In the finite mixture of multinomials model, we suppose that each individual i belongs to exactly one of $K < \infty$ latent classes. For $i = 1, \ldots, n$, let $z_i \in \{1, \ldots, K\}$ indicate the class of individual i, and let $\pi_h = \Pr(z_i = h)$. We assume that $\pi = (\pi_1, \ldots, \pi_K)$ is the same for all individuals. Within any class, we suppose that each of the p variables independently follows a class-specific multinomial distribution. This implies that individuals in the same latent class have the same cell probabilities. For any value x, let $\phi_{hjx} = \Pr(X_{ij} = x | z_i = h)$ be the probability of $X_{ij} = x$ given that individual i is in class h. Let $\phi = \{\phi_{hjx} : x = 1, \ldots, d_j, j = 1, \ldots, p, h = 1, \ldots, K\}$ be the collection of all ϕ_{hjx}. Mathematically, the finite mixture model can be expressed as

$$X_{ij} \mid z_i, \phi \overset{ind}{\sim} \text{Multinomial}(\phi_{z_i j 1}, \ldots, \phi_{z_i j d_j}) \text{ for all } i, j \tag{1}$$

$$z_i \mid \pi \sim \text{Multinomial}(\pi_1, \ldots, \pi_K) \text{ for all } i, \tag{2}$$

where each multinomial distribution has sample size equal to one and the number of levels is implied by the dimension of the corresponding probability vector. As a result, $p(X_i \mid z_i, \phi)$ is the product of p conditionally independent multinomial distributions. The model in (1) – (2) can be estimated via maximum likelihood estimation (e.g., Si *et al.*, 2010) or, with the addition of prior distributions on ϕ and π, a straightforward Gibbs sampler. An equivalent expression for the model after integrating out the class membership indicators is given by

$$X_i \mid \theta \sim \text{Multinomial}(\theta) \text{ for all } i \tag{3}$$

$$\theta_{c_1,\ldots,c_p} = \sum_{h=1}^{K} \pi_h \prod_{j=1}^{p} \phi_{hjc_j} \text{ for all } (c_1, \ldots, c_p) \in D, \tag{4}$$

where the multinomial distribution has sample size equal to one. This expression reveals the flexibility of using mixtures, since (4) theoretically allows θ to take any values.

For fixed K, Vermunt *et al.* (2008) turn (1) – (2) into a multiple imputation engine for missing X_{ij}. Let X_{obs} and X_{mis} represent, respectively, the observed and missing values in the $n \times p$ matrix of sampled data. Here, X_{obs} also includes any fully observed variables. Their procedure for imputing X_{mis} is as follows. First, they create m bootstrap replicates of n cases by sampling rows with replacement from the $n \times p$ matrix of sampled data. In each replicate, they compute the maximum likelihood estimates (or possibly a local mode) of (π, ϕ), so as to have m values $(\pi^{(l)}, \phi^{(l)})$, where $l = 1, \ldots, m$. These m draws incorporate uncertainty from parameter estimation in the imputation procedure. Second, using each $(\pi^{(l)}, \phi^{(l)})$, for each individual i with missing data they compute $\Pr(z_i^{(l)} = h \mid \pi^{(l)}, \phi^{(l)}, X_{obs})$ for all h. They randomly draw values of $z_i^{(l)}$ according to each individual's set of K probabilities. Third, using each drawn $z_i^{(l)}$, they draw each component of $X_{i,mis}$ independently from the relevant multinomial distributions in (1). The result is m completed datasets.

3.2 Infinite Mixture of Products of Multinomials

The mixture model in (1) – (2) presumes a fixed K, but in practice K is not known and must be specified by the analyst. When the analyst sets K to be too small, the mixture model could have insufficient flexibility to estimate complex dependencies. Using empirical examples, Vermunt *et al.* (2008) show that multiple imputation estimates can be sensitive to the choice of K, particularly for small values of K.

How does one select K for multiple imputation purposes so that it is not too small? Vermunt *et al.* (2008) suggest that analysts select K based on penalized likelihood statistics, such as BIC or AIC. Their examples indicate that these criteria can select quite different values of K. In their empirical analysis comprising 4292 individuals and 79 categorical variables, the BIC suggested $K = 8$ classes whereas the AIC had not yet settled on a

minimum at $K = 35$ classes. It should be noted, however, that the multiple imputation inferences for these data were similar when $K = 8$ or $K = 35$. Vermunt *et al.* (2008) do not offer theorems that such similarities are guaranteed to arise; indeed, this would be difficult to do mathematically even without missing data. To our knowledge, the performance of penalized likelihood model selection procedures in this context remains largely unstudied. It is not obvious that the AIC (or BIC) is guaranteed to pick sufficiently large K.

Even disregarding these issues, selection of a single K ignores the uncertainty about K. This results in underestimation of variance in parameter estimates, which goes against Rubin's (1987) recommendations for generating multiple imputations. In practical terms, underestimation of uncertainty could lead to unjustifiably precise multiple imputation inferences and reduced confidence interval coverage rates.

These limitations motivate the use of the DPMPM for multiple imputation. The DPMPM is essentially an infinite mixture of products of multinomial distributions. The prior distribution for the mixture probabilities, $\pi = (\pi_1, \ldots, \pi_\infty)$, is modeled using the stick-breaking representation of the Dirichlet process (Ferguson, 1973, 1974; Sethuraman, 1994; Ishwaran and James, 2001), shown in (7) – (8) below. We note that Dirichlet process mixture models (Antoniak, 1974) are used in many fields including, for example, econometrics (Chib and Hamilton, 2002; Hirano, 2002), social science (Kyung *et al.*, 2010) and finance (Rodríguez

and Dunson, 2011). In particular, for the DPMPM we have

$$
\begin{aligned}
X_{ij} \mid z_i, \phi &\overset{ind}{\sim} \text{Multinomial}(\phi_{z_i j 1}, \ldots, \phi_{z_i j d_j}) \text{ for all } i, j && (5)\\
z_i \mid \pi &\sim \text{Multinomial}(\pi_1, \ldots, \pi_\infty) \text{ for all } i && (6)\\
\pi_h &= V_h \prod_{g<h} (1 - V_g) \text{ for } h = 1, \ldots, \infty && (7)\\
V_h &\sim \text{Beta}(1, \alpha) && (8)\\
\alpha &\sim \text{Gamma}(a_\alpha, b_\alpha) && (9)\\
\phi_{hj} = (\phi_{hj1}, \ldots, \phi_{hjd_j}) &\sim \text{Dirichlet}(a_{j1}, \ldots, a_{jd_j}). && (10)
\end{aligned}
$$

The Gamma distribution is parametrized such that $E(\alpha \mid a_\alpha, b_\alpha) = a_\alpha / b_\alpha$. Here, (a_α, b_α) and each $(a_{j1}, \ldots, a_{jd_j})$ are analyst-supplied constants. We set each element of $(a_{j1}, \ldots, a_{jd_j})$ equal to one to correspond to uniform prior distributions. Following Dunson and Xing (2009), we set $(a_\alpha = .25, b_\alpha = .25)$. We recommend investigating the sensitivity of multiple imputation inferences to other choices of (a_α, b_α) that satisfy $a_\alpha + b_\alpha = .5$, which represents a small prior sample size, and hence vague specification for Gamma distributions. This ensures that the information from the data dominates the posterior distribution. In our applications, results were insensitive to different choices of (a_α, b_α). The specification of prior distributions in (7) – (9) encourages π_h to decrease stochastically with h. In fact, when α is very small, most of the probability in π is allocated to the first few components. The vague prior distribution for α in (9) also encourages the posterior distribution of π to be data-dominated (Escobar and West, 1998).

With these specifications, the DPMPM typically puts non-negligible posterior probability on only finite numbers of classes, even though it allows for an infinite number of them. Importantly, these finite numbers are not fixed *a priori* by the analyst, but determined by the data via the model. As a result, if one ignores classes with negligible mass, inferences

from a DPMPM effectively can be interpreted as averages over models with different finite values of K rather than conditional on a single finite K. In this way, the DPMPM takes uncertainty about K into account in inferences and multiple imputations.

3.3 Posterior Computation and Multiple Imputation

The joint posterior distribution of the parameters in (5) – (10) is not analytically tractable, but it can be approximated via MCMC. Here, we present an algorithm based on the blocked Gibbs sampler (Ishwaran and James, 2001), truncating the infinite stick-breaking probabilities at some large number H^*. Essentially, we approximate (6) and (7) using

$$z_i \mid \pi \sim \text{Multinomial}(\pi_1, \ldots, \pi_{H^*}) \text{ for all } i \tag{11}$$

$$\pi_h = V_h \prod_{g<h}(1 - V_g) \text{ for } h = 1, \ldots, H^*. \tag{12}$$

Here, one makes H^* as large as possible while still offering fast computation. Using an initial proposal for H^*, say $H^* = 20$, analysts can examine the posterior distributions of the sampled number of unique classes across MCMC iterates to diagnose if H^* is large enough. Significant posterior mass at a number of classes equal to H^* suggests that the truncation limit be increased. We note that one can use other MCMC algorithms to estimate the posterior distribution that avoid truncation, for example a slice sampler (Walker, 2007; Dunson and Xing, 2009) or an exact blocked sampler (Papaspiliopoulos, 2008).

We first present the full conditionals needed for the Gibbs sampler assuming no missing data. Equivalently, we condition on (X_{obs}, X_{mis}) in all steps. In what follows, we use a dash after the condition sign to represent all data and other parameters, using the most recently updated values in the cycles of the full conditionals in the Gibbs sampler.

S1. For $i = 1, \ldots, n$, sample $z_i \in \{1, \ldots, H^*\}$ from a multinomial distribution with sample

size one and probabilities

$$Pr(z_i = h|-) = \frac{\pi_h \prod_{j=1}^{p} \phi_{hjX_{ij}}}{\sum_{k=1}^{H^*} \pi_k \prod_{j=1}^{p} \phi_{kjX_{ij}}}. \tag{13}$$

S2. For $h = 1, \ldots, H^* - 1$, sample V_h from the Beta distribution,

$$(V_h|-) \sim \text{Beta}(1 + n_h, \alpha + \sum_{k=h+1}^{H^*} n_k), \tag{14}$$

where $n_h = \sum_{i=1}^{n} I(z_i = h)$ for all h. Here, $I(\cdot) = 1$ when the condition inside the parentheses is true and $I(\cdot) = 0$ otherwise. Set $V_{H^*} = 1$ per the truncation. From these H^* values, calculate each $\pi_h = V_h \prod_{g<h}(1 - V_g)$.

S3. For $h = 1, \ldots, H^*$ and $j = 1, \ldots, p$, sample a new value of $\phi_{hj} = (\phi_{hj1}, \ldots, \phi_{hjd_j})$ from the Dirichlet distribution

$$(\phi_{hj} \mid -) \sim \text{Dirichlet}\left(a_{j1} + \sum_{i:z_i=h} I(X_{ij} = 1), \ldots, a_{jd_j} + \sum_{i:z_i=h} I(X_{ij} = d_j)\right). \tag{15}$$

S4. Sample a new value of α from the Gamma distribution

$$(\alpha|-) \sim \text{Gamma}(a_\alpha + H^* - 1, b_\alpha - \log(\pi_{H^*})). \tag{16}$$

The Gibbs sampler proceeds by initializing the chain. We suggest initializing $\alpha = 1$, each V_h with an independent draw from Beta$(1, 1)$ which is identical to a uniform distribution on $(0, 1)$, and ϕ with the marginal frequency estimates from the observed data. Initializing $\alpha = 1$ is consistent with drawing V_h from a uniform distribution, which we select for convenience; typically one does not have intuition about values of V_h likely to have high density. Initializing ϕ with the marginal frequencies also is for convenience. We note, however, that

initializing the chain at values of ϕ that are unreasonable based on the data can result in slower convergence than starting them at plausible values (like marginal frequencies). After initialization, we proceed through each step, repeating many times until convergence.

The Gibbs sampler is easily modified for missing data. All we need is to sample from the full conditional distribution for each value in X_{mis} after step S4, and use the updated values in subsequent Gibbs cycles. Since we know each unit's latent class and all parameter values, the full conditional distribution of X_{mis} is given in (5). Thus, to account for missing data we simply add a fifth step to S1 – S4 as follows.

S5. For each X_{ij} value in X_{mis}, sample from

$$(X_{ij}|-) \overset{ind}{\sim} \text{Multinomial}(\phi_{z_i j 1}, \ldots, \phi_{z_i j d_j}). \tag{17}$$

To initialize X_{mis} for the MCMC sampler, we suggest sampling from the complete-data, empirical marginal distribution of each X_j. This ensures that the chain starts with values that are not unreasonable based on the data, which can speed convergence of the sampler.

To obtain m completed datasets for use in multiple imputation, analysts select m of the sampled X_{mis} after convergence of the Gibbs sampler. These datasets should be spaced sufficiently so as to be approximately independent (given X_{obs}). This involves thinning the MCMC samples so that the autocorrelations among parameters are close to zero.

Because this is an MCMC algorithm, it is essential to examine convergence diagnostics. Due to the complexity of the models, as well as the missing data, chains can get stuck in regions around local modes and therefore not fully explore the parameter space. We therefore recommend long runs with multiple chains. MCMC diagnostic checks can focus on the draws of θ computed via (4) rather than specific component parameters, since ultimately the implied θ determines the imputations. Further, specific component parameters are subject to label switching among the mixture components, which complicates interpretation of the

components and MCMC diagnostics; we note that label switching does not affect θ nor the multiple imputations. When the dimension of θ is large, as a quick-and-dirty diagnostic analysts can examine the marginal probabilities for all variables (or a random sample when p is very large) and several randomly selected cell probabilities.

4 Simulation Studies of Frequentist Performance

We now investigate the performance of the DPMPM multiple imputation method via simulation studies, focusing on repeated sampling properties. We consider two scenarios: a small number of variables ($p = 7$) generated via log-linear models, and a somewhat large number of variables ($p = 50$) generated via finite mixtures of multinomial distributions. In both we consider simulated bias of point estimates and coverage rates of 95% confidence intervals, all constructed using Rubin's (1987) methods of multiple imputation inference.

Before launching into the results, we briefly review multiple imputation inference (Rubin, 1987). For $l = 1, \ldots, m$, let $q^{(l)}$ and $u^{(l)}$ be respectively the estimate of some population quantity Q and the estimate of the variance of $q^{(l)}$ in completed dataset $(X_{obs}, X_{mis}^{(l)})$. Analysts use $\bar{q}_m = \sum_{l=1}^{m} q^{(l)}/m$ to estimate Q, and use $T_m = (1 + 1/m)b_m + \bar{u}_m$ to estimate $\text{var}(\bar{q}_m)$, where $b_m = \sum_{l=1}^{m}(q^{(l)} - \bar{q}_m)^2/(m-1)$ and $\bar{u}_m = \sum_{l=1}^{m} u^{(l)}/m$. For large samples, inferences for Q are obtained from the t-distribution, $(\bar{q}_m - Q) \sim t_{\nu_m}(0, T_m)$, where the degrees of freedom is $\nu_m = (m-1)\left[1 + \bar{u}_m / \left((1 + 1/m)b_m\right)\right]^2$. A better degrees of freedom for small samples is presented by Barnard and Rubin (1999). Tests of significance for multicomponent null hypotheses are derived by Li *et al.* (1991), Meng and Rubin (1992) and Reiter (2007).

4.1 Case 1: Small p and log-linear model data generation

With modest p, it is straightforward to write-down and simulate from distributions that encode complex dependence relationships. We generate data comprising $n = 5,000$ individuals

and $p = 7$ binary variables as follows. The first five variables are sampled independently from a multinomial distribution with probabilities governed by the log-linear model with

$$\begin{aligned} \log \Pr(X_1, X_2, X_3, X_4, X_5) &\propto \sum_{j=1}^{5} -2X_j + \sum_{j=1}^{4} \sum_{j'=j+1}^{5} X_j X_{j'} + X_1 X_2 X_3 \\ &- X_2 X_3 X_4 - 2X_3 X_4 X_5 + X_2 X_3 X_5 + X_1 X_4 X_5. \end{aligned} \quad (18)$$

We generate X_6 from Bernoulli distributions with probabilities governed by the logistic regression with

$$\begin{aligned} \text{logit } \Pr(X_6) &= -1 + X_1 + 2.2X_2 - 2.5X_3 + .9X_4 + 1.1X_5 - 2.8X_2X_3 + 2.3X_3X_4 \\ &- .5X_2X_4 - 2.4X_3X_5 + 1.55X_1X_4 - 2.1X_4X_5 + 1.2X_3X_4X_5. \end{aligned} \quad (19)$$

We generate X_7 from Bernoulli distributions with probabilities governed by the logistic regression with

$$\begin{aligned} \text{logit } \Pr(X_7) &= -.3 + 1.5X_1 - 2.15X_2 - 2.25X_3 + 1.6X_4 - .88X_5 + 1.11X_6 - .96X_2X_3 \\ &+ 2.3X_1X_3 - .5X_2X_6 - 2X_5X_6 + 1.21X_1X_5 - 2.7X_1X_2 + 1.5X_1X_2X_3. \end{aligned} \quad (20)$$

The inclusion of high order interactions in (18)–(20) represents the type of complex dependence structure that would be difficult to identify from the observed data, and hence difficult to capture with imputation approaches based on log-linear models or chained equations. The specific values of the coefficients are not particularly important, except for noting that the higher order interactions with non-zero coefficients generate complex dependencies.

We suppose that (X_1, X_2, X_7) have values missing at random. We let X_1 be missing with probabilities $(.1, .4, .4, .7)$, respectively, for each of the four combinations of (X_3, X_4). We let X_2 be missing with probabilities $(.7, .4, .4, .1)$, respectively, for each of the four combinations

of (X_5, X_6). We let X_7 be missing with probabilities $(.5, .2, .3, .7)$, respectively, for each of the four combinations of (X_5, X_6). About 99% of the units have at least one missing value.

After introducing missing data, we implement the DPMPM to create $m = 5$ multiply-imputed datasets. We set $H^* = 20$ and run the chains for 50,000 iterations, which via experimentation appears sufficient to ensure convergence and offer repeated simulation results in reasonable time. To provide a comparison of the DPMPM with an existing alternative, we also implement a default version of chained equations using the MICE software package in *R* (Van Buuren and Oudshoorn, 1999). We repeat the process of generating observed data, introducing missing values, and performing multiple imputations 500 times.

We evaluate the two approaches on regression coefficients for three models. The first model is the log-linear model in (18). The second and third models are the two logistic regressions in (19) and (20), excluding the three-way interactions. These are excluded so as to avoid problems caused by random zeros in the repeated simulations. Random zeros cause logistic regression coefficient estimates and standard errors to blow up, which in turn causes problems when fitting MICE and when estimating the logistic regressions from the completed datasets. We note that random zeros do not cause problems for the DPMPM imputation procedure, which is another advantage.

Figure 1 displays average point estimates and 95% confidence interval coverage rates across the 500 simulations. The average point estimates based on DPMPM are closer to the corresponding true values than those based on default MICE. Across all estimands and simulations, the average mean squared error of $\bar{q}_m$ equals .08 when using DPMPM, whereas it equals .13 (50% higher) when using default MICE. The simulated coverage rates of the 95% confidence intervals based on DPMPM generally are closer to 95% than those based on default MICE. Indeed, default MICE results in several rates below 20%. These belong to the three-way interaction terms from (18). The DPMPM, in contrast, has reasonable coverage rates for the three-way interactions. The simulated coverage rates below 80% for

the DPMPM belong to coefficients in the logistic regression for X_7. These rates (77.4%, 65.2%, 70.8%, 49.6%) for the most part are better than the corresponding ones from default MICE (46.8%, 30.4%, 96.4%, 16.6%). We note that the improved coverage rates for DPMPM do not result from unrealistic inflation of variances, as evidenced by the reasonable standard error bars for $\bar{q}_m$ in the left panel of Figure 1.

The DPMPM represents a substantial improvement over default MICE for these simulations. Of course, one can do better than main effects only conditional models when using MICE. Including interaction effects in MICE should improve coverage rates. However, we suspect that the complex dependencies in these data would be challenging to identify in practice when specifying the conditional regression models, and that many analysts would use the default application of MICE.

4.2 Case 2: Large p and mixture model data generation

In general, it is computationally cumbersome to simulate large contingency tables with complex dependence structure from log-linear models and logistic regressions. We therefore generate tables from a finite mixture model akin to the one in Section 3.1. We set $p = 50$ and allow the number of levels for each variable to be randomly chosen from 2 to 6. The final table has $d \approx 10^{30}$ cells. We sample $n = 1,000$ individuals; hence, the vast majority of the d cells in any sampled dataset are in fact empty. We use $K = 4$ classes such that $(\pi_1 = .3, \pi_2 = .2, \pi_3 = .4, \pi_4 = .1)$. Within any class h, we set ϕ to differ across (h, j, x) so as to induce complex dependence. Specifically, for each (h, j, x) we set

$$\phi_{hjx} = \max\left(\frac{h(d_j - 1)}{(h+1)d_j^2}, .05h\right), \quad \phi_{hjd_j} = 1 - \sum_{x=1}^{d_j - 1} \phi_{hjx}. \tag{21}$$

Although it is difficult to summarize succinctly the degree of dependence that results, we note that in one complete dataset randomly generated from this model only 103 of the

$\binom{50}{2} = 1,225$ bivariate χ^2 tests of independence had p-values exceeding .05.

In each dataset, we make each of the first 20 variables have 40% values missing completely at random. We implement multiple imputation using the DPMPM with $m = 5$, $H^* = 20$, and 100,000 MCMC iterations. We did not implement MICE, as it was computationally too expensive to run in repeated simulations with $p = 50$. We note that with the large numbers of random zeros that result from this simulation, we would be essentially forced to run MICE with main effects only to avoid (randomly) inestimable multinomial regression coefficients. We repeat the process of generating observed data, introducing missing values, and performing DPMPM multiple imputation 100 times. The smaller number of simulations than in Section 4.1 reflects the increased time for evaluation with $p = 50$.

For evaluation purposes, we examine twenty arbitrary conditional probabilities, including (i) four involving 2 variables both with missing data, (ii) four involving 2 variables with only one having missing data, (iii) four involving 3 variables all with missing data, (iv) four involving 3 variables with two having missing data, and (v) four involving 3 variables with one having missing data. Figure 2 displays average point estimates and 95% confidence interval coverage rates across the 100 simulations. The average point estimates based on DPMPM are close to the corresponding true values, and the simulated coverage rates are at least 93%, suggesting reasonable performance. We note that the conservative nature of the simulated coverage rates could be an artifact of the limited number of simulation runs.

5 Imputation of TIMSS Background Variables

The TIMSS is conducted by the International Association for the Evaluation of Educational Achievement. Data are collected on a four year cycle and made available for downloads via a dedicated TIMSS website (www.timssandpirls.bc.edu). The goal of TIMSS is to facilitate comparisons of student achievement in mathematics and science across countries. In

addition to domain-specific test questions, TIMSS data include background information on students including demographics, amount of educational resources at the home, time spent on homework, and attitudes towards mathematics and science. These background variables are all categorical.

We use data from the 2007 TIMSS that comprise 80 background variables on 90,505 students (88,129 in Grade 4 and 2,376 in Grade 5) from 22 countries. Among these 80 variables, most (68) have less than 10% missing values; six variables have between 10% and 30% missing values; only one variable has more than 75%. Missingness rates differ by country but not dramatically so. The TIMSS data file lists reasons for missingness, including omitted (student should have answered but did not), not administered (missing because of the rotation design or unintentional misprint), and not reached (incorrect responses). For purposes of multiple imputation, we do not distinguish response reasons and treat all item nonresponse as missing at random.

To create multiple imputations, we run DPMPMs separately in each country. Separate imputation avoids smoothing estimates towards common values, which seems prudent since TIMSS is intended for comparisons across countries. It is possible to extend the DPMPM to allow borrowing information across countries using the hierarchical Dirichlet process (Si, 2012, Chapter 2). When variables are not collected in a country, we remove them from the model for that country.

For the MCMC, we set $H^* = 20$. The posterior distribution of the number of classes among individuals (within any country) had nearly all of its mass below twenty, so that we do not expect truncation to impact the imputations materially. Imputations with $H^* = 50$ on a smaller set of countries resulted in similar performance. We also examined different vague prior specifications for a_α and b_α—including, for example, $(a_\alpha = 1, b_\alpha = .25)$, $(a_\alpha = 1, b_\alpha = 1)$, and $(a_\alpha = 1, b_\alpha = 2)$—and did not observe noticeable differences in the posterior distributions of θ. In each country, we ran the Gibbs sampler for at least 10,000

iterations. MCMC diagnostics of marginal and randomly selected joint components of θ suggest convergence. For the typical country, generating 10,000 iterations takes about 30 minutes on a standard desktop computer using a single CPU. Thus, by running the 22 countries in parallel, the entire TIMSS can be multiply-imputed within a few hours.

To assess the quality of the multiple imputations, we focus on one arbitrarily selected country in the data file. In this country, fifty-nine variables have 10% or less missing values, eleven variables have between 20% and 30% missing values, five variables have between 50% and 70% missing values, and one variable has more than 80% missing values. Only 32 out of 4,223 individuals have complete records. For purposes of evaluating the imputations, we increased the MCMC iterations to 500,000. This took roughly 10 hours to run.

Comparisons of the marginal distributions of the observed and imputed values (Gelman *et al.*, 2005) show similar distributions; these are not shown here to save space. While comforting, such diagnostics offer only partial insights into the quality of the imputations for multivariate relationships. We therefore consider posterior predictive checks that directly assess the ability of the imputation models to preserve associations, following the approach in He *et al.* (2010) and Burgette and Reiter (2010). The basic idea is to use the imputation model to generate not only X_{mis} but an entirely new full dataset, i.e., create a completed dataset $D^{(l)} = (X_{obs}, X_{mis}^{(l)})$ and a replicated dataset $R^{(l)}$ in which both X_{obs} and X_{mis} are simulated from the imputation model. After repeating the process of generating pairs $(D^{(l)}, R^{(l)})$ many times (we use $T = 500$), we compare each $R^{(l)}$ with its corresponding $D^{(l)}$ on statistics of interest. When the statistics are dissimilar, the diagnostic suggests that the imputation model does not generate replicated data that look like the completed data, so that it may not be generating plausible values for the missing data. When the statistics are not dissimilar, the diagnostic does not offer evidence of imputation model inadequacy (with respect to that statistic).

More formally, let S be the statistic of interest, such as a regression coefficient or joint

probability. Let $S_{D^{(l)}}$ and $S_{R^{(l)}}$ be the values of S computed with $D^{(l)}$ and $R^{(l)}$, respectively. For each S we compute the two-sided posterior predictive probability,

$$ppp = (2/T) * \min\left(\sum_{l=1}^{T} I(S_{D^{(l)}} - S_{R^{(l)}} > 0), \sum_{l=1}^{T} I(S_{R^{(l)}} - S_{D^{(l)}} > 0)\right). \quad (22)$$

We note that *ppp* is small when $S_{D^{(l)}}$ and $S_{R^{(l)}}$ consistently deviate from each other in one direction, which would indicate that the imputation model is systematically distorting the relationship captured by S. For S with small *ppp*, it is prudent to examine the distribution of $S_{R^{(l)}} - S_{D^{(l)}}$ to evaluate if the difference is practically important.

To obtain the pairs $(D^{(l)}, R^{(l)})$, we add a step to the MCMC that replaces all values of X_{mis} and X_{obs} using the parameter values at that iteration. This step is used only for computation of the *ppp*; the estimation of parameters continues to be based on X_{obs}. When autocorrelations among parameters are high, we recommend thinning the chain so that θ draws are approximately independent before creating the set of $R^{(l)}$. Further, we advise saving the T pairs of $(D^{(l)}, R^{(l)})$, so that they can be used repeatedly with different S.

We present posterior predictive checks for 36 coefficients in a multinomial logistic regression and 1,000 joint probabilities from a contingency table. The multinomial logistic regression predicts how much students agree that they like being in school (like a lot, like a little, dislike a little, and dislike a lot). It has 4.5% missing values. The predictor variables include how much students agree that they have tried their best (4 categories); how much students agree that teachers want students to do their best (4 categories); whether or not students have had something stolen from them at school (2 categories); whether students were hit or hurt by others at school (2 categories); whether students were made to do things by others at school (2 categories); whether or not students were made fun of or called names at school (2 categories); and, whether or not students were left out of activities at school (2 categories). Among all these predictors, the missing data rates range from 4.5% to 5.8%.

Investigations of the completed datasets indicate strong associations among the variables.

The contingency table includes whether or not students ever use a computer at home (2 categories); whether or not students ever use a computer at school (2 categories); whether or not students ever use a computer elsewhere (2 categories); how often students use a computer for mathematics homework (5 categories); how often students use a computer for science homework (5 categories); and, how often students spend time playing computer games (5 categories). Among all these variables, the missing data rates range from 10% to 63%.

We consider each coefficient and joint probability as separate S. The 1,036 values of *ppp* are displayed in Figure 3. For the contingency table, only 27 out of 1,000 values are below .05, suggesting that overall the DPMPM model generates replicated tables that look similar to the completed ones. For the multinomial regression, three out of 36 values of *ppp* are below .05 but above .01, and two are below .01. These suggest potential model mis-specification involving the associated variables. These five low values correspond to coefficients for the two four-category variables. These variables each have two levels with moderately low marginal probabilities (between 5.2% and 6.8%). Apparently, the DPMPM is somewhat inaccurate at replicating the associations with the outcome at these levels. We note, however, that the standard errors for these coefficients are large compared to the point estimates, so that the impact of modest imputation model mis-specification on multiple imputation inferences for these coefficients is likely to be swamped by sampling variability.

6 Concluding Remarks

The Dirichlet process mixture of products of multinomial distributions offers a fully Bayesian, joint model for multiple imputation of large-scale, incomplete categorical data. The approach is flexible enough to capture complex dependencies automatically and computationally ef-

ficient enough to be applied in large datasets. Although based on mixture models, the approach avoids *ad hoc* selection of a fixed number of classes, thereby reducing risks of using too few classes while fully estimating uncertainty in posterior distributions.

The methods presented here have utility for multiple audiences. Organizations that collect and disseminate large-scale categorical databases can use these imputation methods to create completed public-use files. This places the burden of handling missing data on the organization rather than the secondary analyst, who instead can concentrate on scientific modeling of the multiply-imputed data. These methods also can be applied by individual analysts to categorical data with missing values. Free software implementing the approach in MATLAB is available from the first author, and we expect an R package to be available on CRAN soon.

This approach can serve as the basis for additional methodological and applied developments. Many surveys include both categorical and continuous data. The nonparametric Bayesian approach could be extended to handle mixed data, for example by letting the continuous data be modeled as mixtures of independent normal distributions within latent classes. Many education surveys have data with hierarchical structures, such as students within teachers or schools within counties. The nonparametric Bayesian approach could be extended to include random effects that account for such hierarchical structure. Finally, many educational surveys employ multiple imputation to create "plausible values" of students' abilities (e.g., Mislevy *et al.*, 1992). To the best of our knowledge, current practice typically is to impute plausible values and missing background characteristics separately. Doing so simultaneously may offer improved accuracy when estimating associations between background variables and student proficiency.

References

Ake, C. F. (2005). Rounding after multiple imputation with non-binary categorical covariates (paper 112-30). In *Proceedings of the Thirtienth Annual SAS Users Group International Conference*, 1–11. Cary, NC: SAS Institute Inc.

Allison, P. D. (2000). Multiple imputation for missing data: A cautionary tale. *Sociological Methods and Research* **28**, 301–309.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to nonparametric problems. *The Annals of Statistics* **2**, 1152–1174.

Baccini, M., Cook, S., Frangakis, C. E., Li, F., Mealli, F., Rubin, D. B., and Zell, E. R. (2010). Multiple imputation in the anthrax vaccine research program. *CHANCE* **23**, 1, 16–23.

Ballou, D., Sanders, W., and Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics* **29**, 1, 37–65.

Barnard, J. and Meng, X. L. (1999). Applications of multiple imputation in medical studies: From AIDS to NHANES. *Statistical Methods in Medical Research* **8**, 17–36.

Barnard, J. and Rubin, D. B. (1999). Small-sample degrees of freedom with multiple-imputation. *Biometrika* **86**, 948–955.

Bernaards, C. A., Belin, T. R., and Schafer, J. L. (2007). Robustness of a multivariate normal approximation for imputation of binary incomplete data. *Statistics in Medicine* **26**, 1368–1382.

Bishop, Y., Feinberg, S., and Holland, P. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA.

Burgette, L. F. and Reiter, J. P. (2010). Multiple imputation via sequential regression trees. *American Journal of Epidemiology* **172**, 1070–1076.

Chib, S. and Hamilton, B. H. (2002). Semiparametric Bayes analysis of longitudinal data treatment models. *Journal of Econometrics* **110**, 67 – 89.

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

Erosheva, E. A., Fienberg, S. E., and Junker, B. W. (2002). Alternative statistical models and representations for large sparse multi-dimensional contingency tables. *Annales de la Faculté des Sciences de Toulouse* **11**, 4, 485–505.

Escobar, M. D. and West, M. (1998). Computing nonparametric hierarchical models. In D. D. Dey, P. Müller, and D. Sinha, eds., *Practical Nonparametric and Semiparametric Bayesian Statistics*, 1–16. Berlin: Springer-Verlag.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* **1**, 209–230.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics* **2**, 615–629.

Finch, W. H. (2010). Imputation methods for missing categorical questionnaire data: A comparison of approaches. *Journal of Data Science* **8**, 361–378.

Gebregziabher, M. and DeSantis, S. M. (2010). Latent class based multiple imputation approach for missing categorical data. *Journal of Statistical Planning and Inference* **140**, 3252–3262.

Gelman, A., Van Mechelen, I., Verbeke, G., Heitjan, D. F., and Meulders, M. (2005). Multiple imputation for model checking: Completed-data plots with missing and latent data. *Biometrics* **61**, 74–85.

Graham, J. W. and Schafer, J. L. (1999). On the performance of multiple imputation for multivariate data with small sample size. In R. H. Hoyle, ed., *Statistical Strategies for Small Sample Research*, 1–29. Thousand Oaks, CA: Sage.

He, Y., Zaslavsky, A. M., and Landrum, M. B. (2010). Multiple imputation in a large-scale complex survey: a guide. *Statistical Methods in Medical Research* **19**, 653–670.

Hirano, K. (2002). Semiparametric Bayesian inference in autoregressive panel data models. *Econometrica* **70**, 781–799.

Horton, N. J., Lipsitz, S. P., and Parzen, M. (2003). A potential for bias when rounding in multiple imputation. *The American Statistician* **57**, 229–232.

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

Kyung, M., Gill, J., and Casella, G. (2010). Estimation in Dirichlet random effects models. *Annals of Statistics* **38**, 979–1009.

Lazarsfeld, P. and Henry, N. (1968). *Latent Structure Analysis*. Boston: Houghton Mifflin Co.

Li, F., Yu, Y., and Rubin, D. B. (2012). Imputing missing data by fully conditional models: Some cautionary examples and guidelines. Tech. rep., Department of Statistical Science, Duke University.

Li, K. H., Raghunathan, T. E., and Rubin, D. B. (1991). Large sample significance levels from multiply imputed data using moment-based statistics and an f reference distribution. *Journal of the American Statistical Association* **86**, 1065–1073.

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

Meng, X. L. and Rubin, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika* **79**, 103–111.

Mislevy, R. J., Beaton, A. E., Kaplan, B., and Sheehan, K. M. (1992). Population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement* **29**, 133–162.

Papaspiliopoulos, O. (2008). A note on posterior sampling from Dirichlet mixture models. Tech. rep., Centre for Research in Statistical Methodology, University of Warwick.

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models technique for multiply imputing missing values using a series of regression models. *Survey Methodology* **27**, 85–96.

Reiter, J. P. (2007). Small-sample degrees of freedom for multi-component significance tests with multiple imputation for missing data. *Biometrika* **94**, 502–508.

Reiter, J. P. and Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association* **102**, 1462–1471.

Rodríguez, A. and Dunson, D. B. (2011). Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Analysis* **6**, 145–178.

Rubin, D. B. (1976). Inference and missing data (with discussion). *Biometrika* **63**, 581–592.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

Rubin, D. B. (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.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4**, 639–650.

Si, Y. (2012). *Nonparametric Bayesian Methods for Multiple Imputation of Large Scale Incomplete Categorical Data in Panel Studies*. Ph.D. thesis, Department of Statistical Science, Duke University.

Si, Y., von Davier, M., and Xu, X. (2010). Imputation for missing data on background variables in large-scale assessment surveys. Tech. rep., Educational Testing Service, Princeton, NJ.

Su, Y. S., Gelman, A., Hill, J., and Yajima, M. (2010). Multiple imputation with diagnostics (mi) in R: Opening windows into the black box. *Journal of Statistical Software* **45**, 2, 1–31.

Thomas, N. (2002). The role of secondary covariates when estimating latent trait population distributions. *Psychometrika* **67**, 33–48.

Van Buuren, S. and Oudshoorn, C. (1999). Flexible multivariate imputation by MICE. Tech. rep., Leiden: TNO Preventie en Gezondheid, TNO/VGZ/PG 99.054.

Vermunt, J. K., Ginkel, J. R. V., der Ark, L. A. V., and Sijtsma, K. (2008). Multiple imputation of incomplete categorical data using latent class analysis. *Sociological Methodology* **38**, 369–397.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology* **61**, 287–307.

von Davier, M. (2010). Hierarchical mixtures of diagnostic models. *Psychological Test and Assessment Modeling* **52**, 8–28.

von Davier, M. and Sinharay, S. (2007). An importance sampling em algorithm for latent regression models. *Journal of Educational and Behavioral Statistics* **32**, 3, 233–251.

von Davier, M. and Sinharay, S. (2010). Stochastic approximation methods for latent regression item response models. *Journal of Educational and Behavioral Statistics* **35**, 2, 174–193.

Walker, S. G. (2007). Sampling the Dirichlet mixture models with slices. *Computations in Statistics-Simulation and Computation* **36**, 45–54.

Yucel, R. M., He, Y., and Zaslavsky, A. M. (2011). Gaussian-based routines to impute categorical variables in health surveys. *Statistics in Medicine* **30**, 29, 3447–3460.

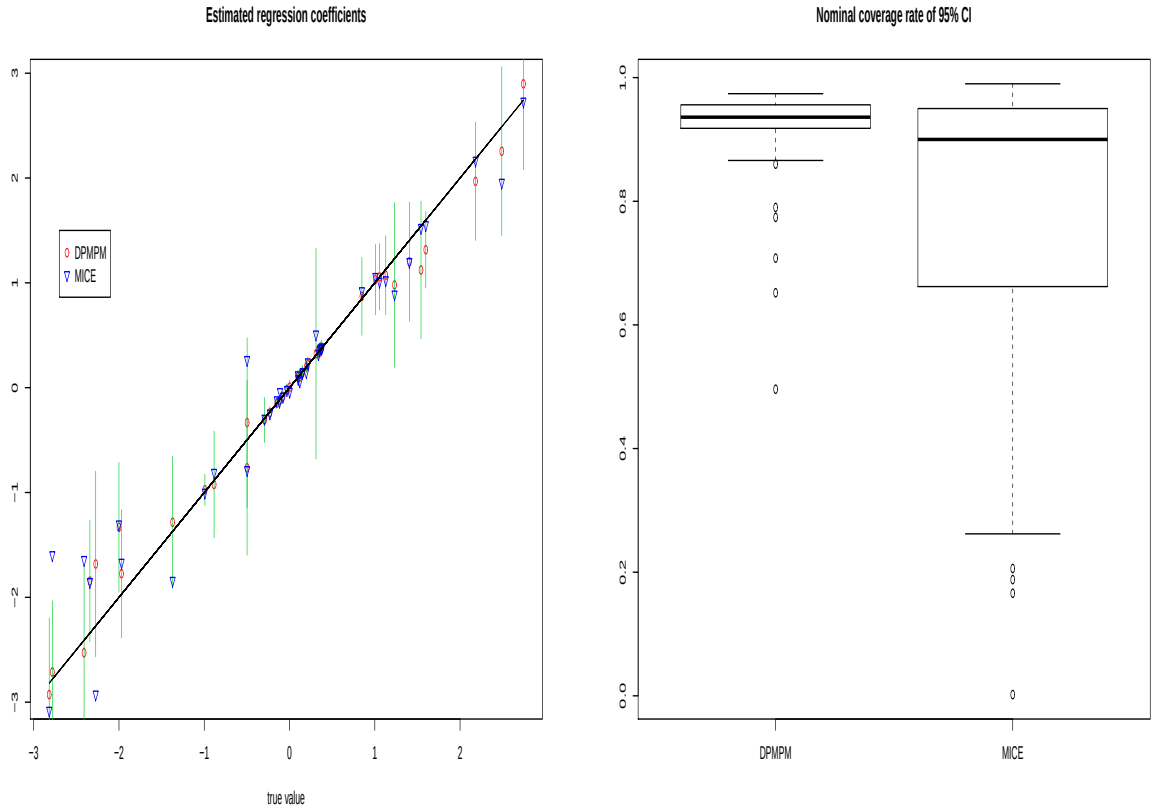

Figure 1: Small p simulation results. Simulated average point estimates and 95% confidence interval coverage rates for 45 regression coefficients. True values of coefficients (shown on solid line) obtained from a very large, complete dataset generated from the data models. Error bars for each $\bar{q}_m$ for DPMPM stretch $\pm 1.96\sqrt{\text{avg. } T_m}$, where "avg. T_m" is the average multiple imputation variance estimate across the 500 replications.

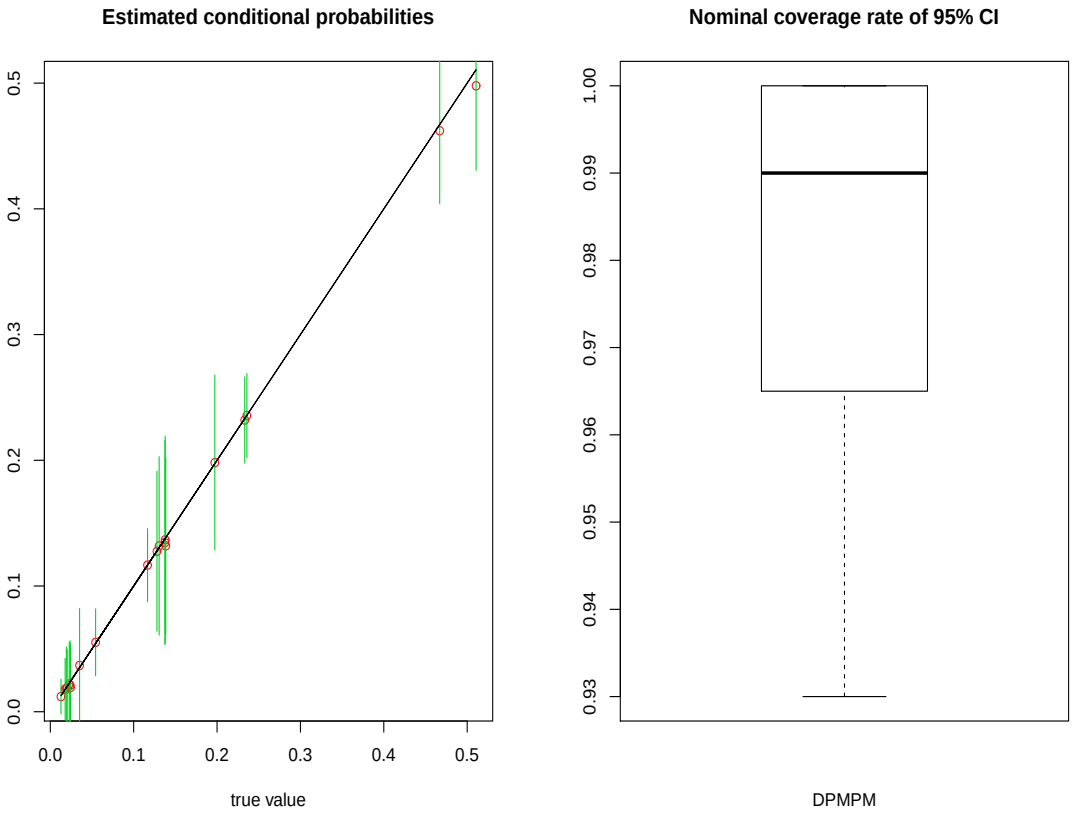

Figure 2: Large p simulation results. Simulated average point estimates and 95% confidence interval coverage rates for 20 conditional probabilities. True values of probabilities (shown on solid line) obtained by stacking the 100 complete datasets. Error bars for each $\bar{q}_m$ for DPMPM stretch $\pm 1.96\sqrt{\text{avg. } T_m}$, where "avg. T_m" is the average multiple imputation variance estimate across the 100 replications.

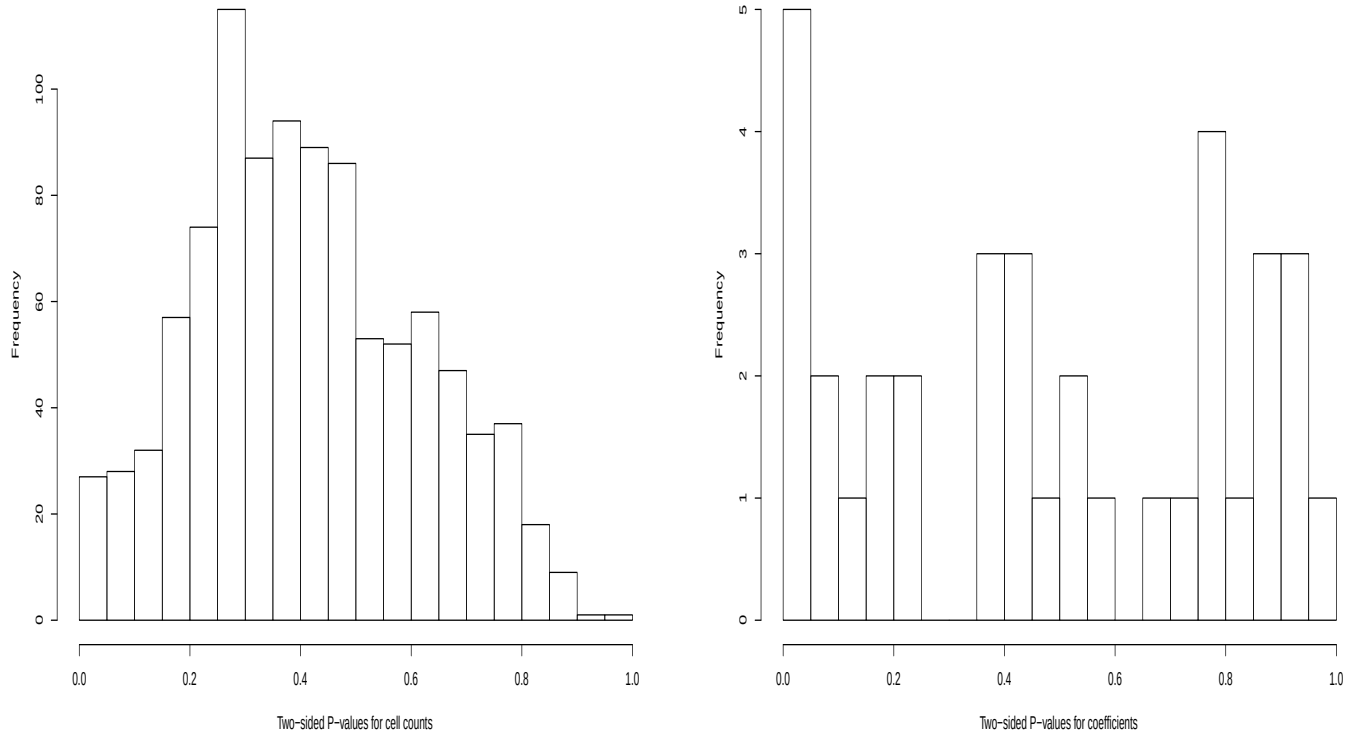

Figure 3: Frequency distributions of *ppp* for the TIMSS imputation. The left panel is for the 1000 cell probabilities, i.e., the completed-data counts over N, and the right panel is for the 36 multinomial regression coefficients.