Survey Methodology

Multiple imputation of missing values in household data with structural zeros

by Olanrewaju Akande, Jerome Reiter and Andrés F. Barrientos

SURVEY
METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2005

VOLUME 31

NUMBER 1
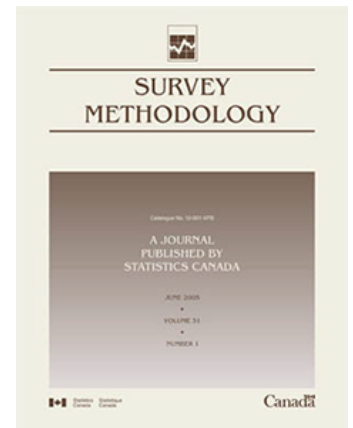
Statistics   Statistique
Canada   Canada

Canadä

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at STATCAN.infostats-infostats.STATCAN@canada.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service   1-800-263-1136
- National telecommunications device for the hearing impaired   1-800-363-7629
- Fax line   1-514-283-9350

Depository Services Program

- Inquiries line   1-800-635-7943
- Fax line   1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Multiple imputation of missing values in household data with structural zeros

Olanrewaju Akande, Jerome Reiter and Andrés F. Barrientos[1]

Abstract

We present an approach for imputation of missing items in multivariate categorical data nested within households. The approach relies on a latent class model that (i) allows for household-level and individual-level variables, (ii) ensures that impossible household configurations have zero probability in the model, and (iii) can preserve multivariate distributions both within households and across households. We present a Gibbs sampler for estimating the model and generating imputations. We also describe strategies for improving the computational efficiency of the model estimation. We illustrate the performance of the approach with data that mimic the variables collected in typical population censuses.

Key Words: Categorical; Census; Edit; Latent; Mixture; Nonresponse.

1 Introduction

In many population censuses and demographic surveys, statistical agencies collect data on individuals grouped within houses. In the U.S. decennial census, for example, the Census Bureau collects the age, race, sex, and relationship to the household head for every individual in the household, as well as whether or not the residents own the house. After collection, agencies share these datasets for secondary analysis, either as tabular summaries, public use microdata samples, or restricted access files.

When creating these data products, agencies typically have to deal with item nonresponse both for individual-level variables and household-level variables. They typically do so using some type of imputation procedure. Ideally, these procedures satisfy three desiderata. First, the imputations preserve the joint distribution of the variables as best as possible. As part of this, the procedure should preserve relationships within households. For example, the missing race of a spouse likely, but certainly not definitely, matches the race of the household head; the imputation procedure should reflect that. Second, the imputations respect structural zeros. For example, a daughter's age cannot exceed her biological mother's age. The imputations should not create impossible combinations of individuals in the same household. Third, the imputation procedure allows for appropriate uncertainty to be propagated in subsequent analyses of the data.

Typical approaches to imputation of missing household items use some variant of hot deck imputation (Kalton and Kasprzyk, 1986; Andridge and Little, 2010). However, depending on how the hot deck is implemented, it may not satisfy one or more of the desiderata. Indeed, we are not aware of any hot deck imputation procedure for household data that satisfies all three explicitly. An alternative is to estimate a model that describes the joint distribution of all the variables, and impute missing values from the implied predictive distributions in the model. For household data, one such model is the nested data Dirichlet process mixture of products of multinomial distributions (NDPMPM) model of Hu, Reiter and Wang (2018), which assumes that (i) each household is a member of a household-level latent class, and (ii) each individual is a

---

1. Olanrewaju M. Akande is Ph.D. Candidate, Department of Statistical Science, Duke University, Durham, NC 27708. E-mail: olanrewaju.akande@duke.edu; Jerome P. Reiter is Professor of Statistical Science, Duke University, Durham, NC 27708. E-mail: jerry@stat.duke.edu; Andrés F. Barrientos is Postdoctoral Associate, Department of Statistical Science, Duke University, Durham, NC 27708. E-mail: anfebar@stat.duke.edu.

member of an individual-level latent class nested within its household-level latent class. The model assigns zero probability to combinations corresponding to structural zeros, and also handles both household-level and individual-level variables simultaneously. The NDPMPM is appealing as an imputation engine, as it can preserve multivariate associations while avoiding imputations that result in impossible households. The NDPMPM is related to models proposed by Vermunt (2003, 2008) and Bennink, Croon, Kroon and Vermunt (2016), although these are used for regression rather than multivariate imputation and do not deal with structural zeros.

Hu et al. (2018) use the NDPMPM to generate synthetic datasets (Rubin, 1993; Raghunathan and Rubin, 2001; Reiter and Raghunathan, 2007) for statistical disclosure limitation, but they do not describe how to use it for imputation of missing data. We do so in this article. With structural zeros in the NDPMPM, the conditional distributions of the missing values given the observed values are not available in closed form. We therefore add a rejection sampling step to the Gibbs sampler used by Hu et al. (2018), which generates completed datasets as byproducts of the Markov chain Monte Carlo (MCMC) algorithms used to estimate the model. These completed datasets can be analyzed using multiple imputation inferences (Rubin, 1987). We also present two new strategies for speeding up the computations with NDPMPMs, namely (i) turning data for the household head into household-level variables rather than individual-level variables, and (ii) using an approximation to the likelihood function. These scalable innovations are necessary, as the NDPMPM is computationally quite intensive even without missing data. The speed-up strategies also can be employed when using the NDPMPM to generate synthetic data.

The remainder of this article is organized as follows. In Section 2, we review the NDPMPM model in the presence of structural zeros and the MCMC sampler for fitting the model without missing data. In Section 3, we extend the MCMC sampler for the NDPMPM model to allow for missing data. In Section 4, we present the two strategies for speeding up the MCMC sampler. In Section 5, we present results of simulation studies used to examine the performance of the NDPMPM as a multiple imputation engine, using the two strategies for speeding up the run time. In Section 6, we discuss findings, caveats and future work.

2 Review of the NDPMPM model

Hu et al. (2018) present the NDPMPM model including motivation for how it can preserve associations across variables and account for structural zeros. Here, we summarize the model without detailed motivations, referring the reader to Hu et al. (2018) for more information. We begin with notation needed to understand the model and the Gibbs sampler, assuming complete data. The presentation closely follows that in Hu et al. (2018).

2.1 Notation and model specification

Suppose the data contain n households. Each household $i = 1, \ldots, n$ contains n_i individuals, so that there are $\sum_{i=1}^{n} n_i = N$ individuals in the data. Let $X_{ik} \in \{1, \ldots, d_k\}$ be the value of categorical variable k for household i, which is assumed to be identical for all n_i individuals in household i, where $k = p + 1, \ldots, p + q$. Let $X_{ijk} \in \{1, \ldots, d_k\}$ be the value of categorical variable k for person j in

household i, where $j = 1, \ldots, n_i$ and $k = 1, \ldots, p$. Let $\mathbf{X}_i = \left(X_{i(p+1)}, \ldots, X_{i(p+q)}, X_{i11}, \ldots, X_{in_i p} \right)$ include all household-level and individual-level variables for the n_i individuals in household i.

Let $\mathcal{H}$ be the set of all household sizes that are possible in the population. For all $h \in \mathcal{H}$, let $\mathcal{C}_h$ represent the set of all combinations of individual-level and household-level variables for households of size h, including impossible combinations; that is, $\mathcal{C}_h = \prod_{k=p+1}^{p+q} \{1, \ldots, d_k\} \prod_{j=1}^{h} \prod_{k=1}^{p} \{1, \ldots, d_k\}$. Let $\mathcal{S}_h \subset \mathcal{C}_h$ represent the set of impossible combinations, i.e., those that are structural zeros, for households of size h. These include combinations of variables within any individual, e.g., a three year old person cannot be a spouse, or across individuals in the same household, e.g., a person cannot be older than his biological parents. Let $\mathcal{C} = \bigcup_{h \in \mathcal{H}} \mathcal{C}_h$ and $\mathcal{S} = \bigcup_{h \in \mathcal{H}} \mathcal{S}_h$.

Although the NDPMPM model we use restricts the support of $\mathbf{X}_i$ to $\mathcal{C} - \mathcal{S}$, it is helpful for understanding the model to begin with no restrictions on the support of $\mathbf{X}_i$. Each household i belongs to one of F classes representing latent household types. For $i = 1, \ldots, n$, let $G_i \in \{1, \ldots, F\}$ indicate the household class for household i. Let $\pi_g = \Pr(G_i = g)$ be the probability that household i belongs to class g. Within any class, all household-level variables follow independent, multinomial distributions. For any $k \in \{p+1, \ldots, p+q\}$ and any $c \in \{1, \ldots, d_k\}$, let $\lambda_{gc}^{(k)} = \Pr(X_{ik} = c \,|\, G_i = g)$ for any class g, where $\lambda_{gc}^{(k)}$ is the same value for every household in class g. Let $\pi = \{\pi_1, \ldots, \pi_F\}$, and $\lambda = \{\lambda_{gc}^{(k)}: c = 1, \ldots, d_k; \ k = p+1, \ldots, p+q; \ g = 1, \ldots, F\}$.

Within each household class, each individual belongs to one of S individual-level latent classes. For $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$, let M_{ij} represent the individual-level latent class of individual j in household i. Let $\omega_{gm} = \Pr(M_{ij} = m \,|\, G_i = g)$ be the probability that individual j in household i belongs to individual-level class m nested within household-level class g. Within any individual-level class, all individual-level variables follow independent, multinomial distributions. For any $k \in \{1, \ldots, p\}$ and any $c \in \{1, \ldots, d_k\}$, let $\phi_{gmc}^{(k)} = \Pr(X_{ijk} = c \,|\, (G_i, M_{ij}) = (g, m))$ for the class pair (g, m), where $\phi_{gmc}^{(k)}$ is the same value for every individual in the class pair (g, m). Let $\omega = \{\omega_{gm}: g = 1, \ldots, F; \ m = 1, \ldots, S\}$, and $\phi = \{\phi_{gmc}^{(k)}: c = 1, \ldots, d_k; \ k = 1, \ldots, p; \ m = 1, \ldots, S; \ g = 1, \ldots, F\}$.

For purposes of the Gibbs sampler in Section 2.2, it is useful to distinguish values of $\mathbf{X}_i$ that satisfy all the structural zero constraints from those that do not. Let the superscript "1" indicate that a random variable has support only on $\mathcal{C} - \mathcal{S}$. For example, $\mathbf{X}_i^1$ represents data for a household with values restricted only on $\mathcal{C} - \mathcal{S}$, i.e., not an impossible household, whereas $\mathbf{X}_i$ represents data for a household with any values in $\mathcal{C}$. Let $\mathcal{X}^1$ be the observed data comprising n households, that is, a realization of $(\mathbf{X}_1^1, \ldots, \mathbf{X}_n^1)$. The kernel of the NDPMPM, $\Pr(\mathcal{X}^1 \,|\, \theta)$, is

$$L(\mathcal{X}^1 \,|\, \theta) = \prod_{i=1}^{n} \sum_{h \in \mathcal{H}} \mathbb{1}\{n_i = h\} \mathbb{1}\{\mathbf{X}_i^1 \notin \mathcal{S}_h\} \left[\sum_{g=1}^{F} \pi_g \prod_{k=p+1}^{p+q} \lambda_{gX_{ik}^1}^{(k)} \prod_{j=1}^{h} \sum_{m=1}^{S} \omega_{gm} \prod_{k=1}^{p} \phi_{gmX_{ijk}^1}^{(k)} \right], \qquad (2.1)$$

where θ includes all the parameters, and $\mathbb{1}\{.\}$ equals one when the condition inside the $\{\}$ is true and equals zero otherwise.

For all $h \in \mathcal{H}$, let $n_{1h} = \sum_{i=1}^{n} \mathbb{1}\{n_i = h\}$ be the number of households of size h in $\mathcal{X}^1$ and $\pi_{0h}(\theta) = \Pr(\mathbf{X}_i \in \mathcal{S}_h \,|\, \theta)$. As stated in Hu et al. (2018), the normalizing constant in the likelihood in (2.1) is $\prod_{h \in \mathcal{H}} (1 - \pi_{0h}(\theta))^{n_{1h}}$. Therefore, the posterior distribution is

$$\Pr\left(\theta \,|\, \mathcal{X}^1, T\left(\mathcal{S}\right)\right) \propto \Pr\left(\mathcal{X}^1 \,|\, \theta\right) \Pr\left(\theta\right) = \frac{1}{\prod_{h \in \mathcal{H}} \left(1 - \pi_{0h}\left(\theta\right)\right)^{n_{1h}}} L\left(\mathcal{X}^1 \,|\, \theta\right) \Pr\left(\theta\right) \tag{2.2}$$

where $T\left(\mathcal{S}\right)$ emphasizes that the density is for the NDPMPM with support restricted to $\mathcal{C} - \mathcal{S}$.

The likelihood in (2.1) can be written as a generative model of the form

$$X_{ik} \,|\, G_i, \lambda \sim \text{Discrete}\left(\lambda_{G_i 1}^{(k)}, \ldots, \lambda_{G_i d_k}^{(k)}\right)$$
$$\forall i = 1, \ldots, n \ \text{ and } \ k = p+1, \ldots, p+q \tag{2.3}$$

$$X_{ijk} \,|\, G_i, M_{ij}, \phi, n_i \sim \text{Discrete}\left(\phi_{G_i M_{ij} 1}^{(k)}, \ldots, \phi_{G_i M_{ij} d_k}^{(k)}\right)$$
$$\forall i = 1, \ldots, n \,, \ j = 1, \ldots, n_i \ \text{ and } \ k = 1, \ldots, p \tag{2.4}$$

$$G_i \,|\, \pi \sim \text{Discrete}\left(\pi_1, \ldots, \pi_F\right)$$
$$\forall i = 1, \ldots, n \tag{2.5}$$

$$M_{ij} \,\big|\, G_i, \omega, n_i \sim \text{Discrete}\left(\omega_{G_i 1}, \ldots, \omega_{G_i S}\right)$$
$$\forall i = 1, \ldots, n \ \text{ and } \ j = 1, \ldots, n_i \tag{2.6}$$

where the Discrete distribution refers to the multinomial distribution with sample size equal to one. We restrict the support of each $\mathbf{X}_i$ to ensure the model assigns zero probability to all combinations in $\mathcal{S}$ as desired. The model in (2.3) to (2.6) can be used without restricting the support to $\mathcal{C} - \mathcal{S}$. This ignores all structural zeros. While not appropriate for the joint distribution of household data, this model turns out to useful for the Gibbs sampler. We refer to the generative model in (2.3) to (2.6) with support on all of $\mathcal{C}$ as the untruncated NDPMPM. For contrast, we call the model in (2.1) the truncated NDPMPM.

For prior distributions, we follow the recommendations of Hu et al. (2018). We use independent uniform Dirichlet distributions as priors for λ and ϕ, and the truncated stick-breaking representation of the Dirichlet process as priors for π and ω (Sethuraman, 1994; Dunson and Xing, 2009; Si and Reiter, 2013; Manrique-Vallier and Reiter, 2014),

$$\lambda_g^{(k)} = \left(\lambda_{g1}^{(k)}, \ldots, \lambda_{gd_k}^{(k)}\right) \sim \text{Dirichlet}\left(1, \ldots, 1\right) \tag{2.7}$$

$$\phi_{gm}^{(k)} = \left(\phi_{gm1}^{(k)}, \ldots, \phi_{gmd_k}^{(k)}\right) \sim \text{Dirichlet}\left(1, \ldots, 1\right) \tag{2.8}$$

$$\pi_g = u_g \prod_{f < g} \left(1 - u_f\right) \ \text{ for } \ g = 1, \ldots, F \tag{2.9}$$

$$u_g \sim \text{Beta}\left(1, \alpha\right) \ \text{ for } \ g = 1, \ldots, F - 1, u_F = 1 \tag{2.10}$$

$$\alpha \sim \text{Gamma}\left(0.25, 0.25\right) \tag{2.11}$$

$$\omega_{gm} = v_{gm} \prod_{s < m} \left(1 - v_{gs}\right) \ \text{ for } \ m = 1, \ldots S \tag{2.12}$$

$$v_{gm} \sim \text{Beta}\left(1, \beta_g\right) \ \text{ for } \ m = 1, \ldots, S - 1, v_{gS} = 1 \tag{2.13}$$

$$\beta_g \sim \text{Gamma}\,(0.25,\ 0.25).$$ (2.14)

We set the parameters for the Dirichlet distributions in (2.7) and (2.8) to $\mathbf{1}_{d_k}$ (a d_k – dimensional vector of ones) and the parameters for the Gamma distributions in (2.11) and (2.14) to 0.25 to represent vague prior specifications. We also set $\beta_g = \beta$ for computational expedience. For further discussion on prior specifications, see Hu et al. (2018).

Conceptually, the latent household-level classes can be interpreted as clusters of households with similar compositions, e.g., households with children or households in which no one is related. Similarly, the latent individual-level classes can be interpreted as clusters of individuals with similar characteristics, e.g., older male spouses or young female children. However, for purposes of imputation, we do not care much about interpreting the classes, as they serve mainly to induce dependence across variables and individuals in the joint distribution.

It is important to select F and S to be large enough to ensure accurate estimation of the joint distribution. However, we also do not want to make F and S so large as to produce many empty classes in the model estimation. Allowing many empty classes increases computational running time without any corresponding increase in estimation accuracy. This can be especially problematic in the Gibbs sampler for the truncated NDPMPM, as these empty classes can introduce mass in regions of the space where impossible combinations are likely to be generated. This slows down the convergence of the Gibbs sampler.

We therefore recommend following the strategy in Hu et al. (2018) when setting (F, S). Analysts can start with moderate values for both, say between 10 and 15, in initial tuning runs. After convergence, analysts examine posterior samples of the latent classes to check how many individual-level and household-level latent classes are occupied. Such posterior predictive checks can provide evidence for the case that larger values for F and S are needed. If the numbers of occupied household-level classes hits F, we suggest increasing F. If the number of occupied individual-level classes hits S, we suggest increasing F first but then increasing S, possibly in addition to F, if increasing F alone does not suffice. When posterior predictive checks do not provide evidence that larger values of F and S are needed, analysts need not increase the number of classes, as doing so is not expected to improve the accuracy of the estimation. We note that similar logic is used in other mixture model contexts (Walker, 2007; Si and Reiter, 2013; Manrique-Vallier and Reiter, 2014; Murray and Reiter, 2016).

2.2 MCMC sampler for the NDPMPM

Hu et al. (2018) use a data augmentation strategy (Manrique-Vallier and Reiter, 2014) to estimate the posterior distribution in (2.2). They assume that the observed data $\mathcal{X}^1$, which includes only feasible households, is a subset from a hypothetical sample $\mathcal{X}$ of $(n + n_0)$ households directly generated from the untruncated NDPMPM. That is, $\mathcal{X}$ is generated on the support $\mathcal{C}$ where all combinations are possible and structural zeros rules are not enforced, but we only observe the sample of n households $\mathcal{X}^1$ that satisfy the structural zero rules and do not observe the sample of n_0 households $\mathcal{X}^0 = \mathcal{X} - \mathcal{X}^1$ that fail the rules.

We use the strategy of Hu et al. (2018) and augment the data as follows. For each $h \in \mathcal{H}$, we simulate $\mathcal{X}$ from the untruncated NDPMPM, stopping when the number of simulated feasible households in $\mathcal{X}$

directly matches n_{1h} for all $h \in \mathcal{H}$. We replace the simulated feasible households in $\mathcal{X}$ with $\mathcal{X}^1$, thus, assuming that $\mathcal{X}$ already contains $\mathcal{X}^1$ and we only need to generate the part $\mathcal{X}^0$ that fall in $\mathcal{S}$. Given a draw of $\mathcal{X}$, we draw θ from posterior distribution defined by the untruncated NDPMPM, treating $\mathcal{X}$ as the observed data. This posterior distribution can be estimated using a blocked Gibbs sampler (Ishwaran and James, 2001; Si and Reiter, 2013).

　　We now present the full MCMC sampler for fitting the truncated NDPMPM. Let $\mathbf{G}^0$ and $\mathbf{M}^0$ be vectors of the latent class membership indicators for the households in $\mathcal{X}^0$ and n_{0h} be the number of households of size h in $\mathcal{X}^0$, with $n_0 = \sum_h n_{0h}$. In each full conditional, let "–" represent conditioning on all other variables and parameters in the model. At each MCMC iteration, we do the following steps.

S1. Set $\mathcal{X}^0 = \mathbf{G}^0 = \mathbf{M}^0 = \varnothing$. For each $h \in \mathcal{H}$, repeat the following:

　　(a) Set $t_0 = 0$ and $t_1 = 0$.

　　(b) Sample $G_i^0 \in \{1, \ldots, F\} \sim \text{Discrete}(\pi_1^{**}, \ldots, \pi_F^{**})$ where $\pi_g^{**} \propto \lambda_{gh}^{(k)}\pi_g$ and k is the index for the household-level variable "household size".

　　(c) For $j = 1, \ldots, h$, sample $M_{ij}^0 \in \{1, \ldots, S\} \sim \text{Discrete}(\omega_{G_i^0 1}, \ldots, \omega_{G_i^0 S})$.

　　(d) Set $X_{ik}^0 = h$, where X_{ik}^0 corresponds to the variable for household size. Sample the remaining household-level and individual-level values using the likelihoods in (2.3) and (2.4). Set the household's simulated value to $\mathbf{X}_i^0$.

　　(e) If $\mathbf{X}_i^0 \in \mathcal{S}_h$, let $t_0 = t_0 + 1$, $\mathcal{X}^0 = \mathcal{X}^0 \cup \mathbf{X}_i^0$, $\mathbf{G}^0 = \mathbf{G}^0 \cup G_i^0$ and $\mathbf{M}^0 = \mathbf{M}^0 \cup \{M_{i1}^0, \ldots, M_{ih}^0\}$. Otherwise set $t_1 = t_1 + 1$.

　　(f) If $t_1 < n_{1h}$, return to step (b). Otherwise, set $n_{0h} = t_0$.

S2. For observations in $\mathcal{X}^1$,

　　(a) Sample $G_i \in \{1, \ldots, F\} \sim \text{Discrete}(\pi_1^*, \ldots, \pi_F^*)$ for $i = 1, \ldots, n$, where

$$\pi_g^* = \text{Pr}(G_i = g \,|\, -) = \frac{\pi_g \left[\prod_{k=p+1}^q \lambda_{gX_{ik}^1}^{(k)} \left(\prod_{j=1}^{n_i} \sum_{m=1}^S \omega_{gm} \prod_{k=1}^p \phi_{gmX_{ijk}^1}^{(k)} \right) \right]}{\sum_{f=1}^F \pi_f \left[\prod_{k=p+1}^q \lambda_{fX_{ik}^1}^{(k)} \left(\prod_{j=1}^{n_i} \sum_{m=1}^S \omega_{gm} \prod_{k=1}^p \phi_{fmX_{ijk}^1}^{(k)} \right) \right]}$$

　　　for $g = 1, \ldots, F$. Set $G_i^1 = G_i$.

　　(b) Sample $M_{ij} \in \{1, \ldots, S\} \sim \text{Discrete}(\omega_{G_i^1 1}^*, \ldots, \omega_{G_i^1 S}^*)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$, where

$$\omega_{G_i^1 m}^* = \text{Pr}(M_{ij} = m \,|\, -) = \frac{\omega_{G_i^1 m} \prod_{k=1}^p \phi_{G_i^1 mX_{ijk}^1}^{(k)}}{\sum_{s=1}^S \omega_{G_i^1 s} \prod_{k=1}^p \phi_{G_i^1 sX_{ijk}^1}^{(k)}}$$

　　　for $m = 1, \ldots, S$. Set $M_{ij}^1 = M_{ij}$.

S3. Set $u_F = 1$. Sample

$$u_g \mid - \sim \text{Beta}\left(1 + U_g, \alpha + \sum_{f=g+1}^{F} U_f\right), \quad \pi_g = u_g \prod_{f<g}\left(1 - u_f\right)$$

where

$$U_g = \sum_{i=1}^{n} \mathbb{1}\left(G_i^1 = g\right) + \sum_{i=1}^{n_0} \mathbb{1}\left(G_i^0 = g\right)$$

for $g = 1, \ldots, F - 1$.

S4. Set $v_{gM} = 1$ for $g = 1, \ldots, F$. Sample

$$v_{gm} \mid - \sim \text{Beta}\left(1 + V_{gm}, \beta + \sum_{s=m+1}^{S} V_{gs}\right), \quad \omega_{gm} = v_{gm} \prod_{s<m}\left(1 - v_{gs}\right)$$

where

$$V_{gm} = \sum_{i=1}^{n} \mathbb{1}\left(M_{ij}^1 = m, G_i^1 = g\right) + \sum_{i=1}^{n_0} \mathbb{1}\left(M_{ij}^0 = m, G_i^0 = g\right)$$

for $m = 1, \ldots, S - 1$ and $g = 1, \ldots, F$.

S5. Sample

$$\lambda_g^{(k)} \mid - \sim \text{Dirichlet}\left(1 + \eta_{g1}^{(k)}, \ldots, 1 + \eta_{gd_k}^{(k)}\right)$$

where

$$\eta_{gc}^{(k)} = \sum_{i \mid G_i^1 = g}^{n} \mathbb{1}\left(X_{ik}^1 = c\right) + \sum_{i \mid G_i^0 = g}^{n_0} \mathbb{1}\left(X_{ik}^0 = c\right)$$

for $g = 1, \ldots, F$ and $k = p + 1, \ldots, q$.

S6. Sample

$$\phi_{gm}^{(k)} \mid - \sim \text{Dirichlet}\left(1 + v_{gm1}^{(k)}, \ldots, 1 + v_{gmd_k}^{(k)}\right)$$

where

$$v_{gmc}^{(k)} = \sum_{i,j \mid G_i^1 = g, M_{ij}^1 = m}^{n} \mathbb{1}\left(X_{ijk}^1 = c\right) + \sum_{i,j \mid G_i^0 = g, M_{ij}^0 = m}^{n_0} \mathbb{1}\left(X_{ijk}^0 = c\right)$$

for $g = 1, \ldots, F$, $m = 1, \ldots, S$ and $k = 1, \ldots, p$.

S7. Sample

$$\alpha \mid - \sim \text{Gamma}\left(a_\alpha + F - 1, b_\alpha - \sum_{g=1}^{F-1} \log\left(1 - u_g\right)\right).$$

S8. Sample

$$\beta \mid - \sim \text{Gamma}\left(a_\beta + F \times (S-1), \, b_\beta - \sum_{m=1}^{S-1}\sum_{g=1}^{F} \log\left(1 - v_{gm}\right)\right).$$

This Gibbs sampler is implemented in the R software package "NestedCategBayesImpute" (Wang, Akande, Hu, Reiter and Barrientos, 2016). The software can be used to generate synthetic versions of the original data, but it requires all data to be complete.

3 Handling missing data using the NDPMPM

We modify the Gibbs sampler for the truncated NDPMPM to incorporate missing data. For $i = 1, \ldots, n$, let $\mathbf{a}_i = \left(a_{i(p+1)}, \ldots, a_{i(p+q)}\right)$ be a vector with $a_{ik} = 1$ when household-level variable $k \in \{p + 1, \ldots, p + q\}$ in $\mathbf{X}_i^1$ is missing, and $a_{ik} = 0$ otherwise. For $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$, let $\mathbf{b}_{ij} = \left(b_{ij1}, \ldots, b_{ijp}\right)$ be a vector with $b_{ijk} = 1$ when individual-level variable $k \in \{1, \ldots, p\}$ for individual $j \in \{1, \ldots, n_i\}$ in $\mathbf{X}_i^1$ is missing, and $b_{ijk} = 0$ otherwise. For each household i, let $\mathbf{X}_i^1 = \left(\mathbf{X}_i^{\text{obs}}, \mathbf{X}_i^{\text{mis}}\right)$, where $\mathbf{X}_i^{\text{obs}}$ comprise all data values corresponding to $a_{ik} = 0$ and $b_{ijk} = 0$, and $\mathbf{X}_i^{\text{mis}}$ comprises all data values corresponding to $a_{ik} = 1$ and $b_{ijk} = 1$. We assume that the data are missing at random (Rubin, 1976).

To incorporate missing values in the Gibbs sampler, we need to sample from the full conditional of each variable in $\mathbf{X}_i^{\text{mis}}$, conditioned on the variables for which $a_{ik} = 0$ and $b_{ijk} = 0$, at every iteration. Thus, we add the ninth step,

S9. For $i = 1, \ldots, n$, sample $\mathbf{X}_i^{\text{mis}}$ from its full conditional distribution

$$\Pr\left(\mathbf{X}_i^{\text{mis}} \mid -\right) \propto \mathbb{1}\left\{\mathbf{X}_i^1 \notin \mathcal{S}_h\right\}\left(\pi_{G_i^1} \prod_{k \mid a_{ik}=1}^{p+q} \lambda_{G_i^1 X_{ik}^1}^{(k)} \prod_{j=1}^{n_i} \omega_{G_i^1 M_{ij}^1} \prod_{k \mid b_{ijk}=1}^{p} \phi_{G_i^1 M_{ij}^1 X_{ijk}^1}^{(k)} \right).$$

Sampling from this conditional distribution is nontrivial because of the dependence among variables induced by the structural zero rules in each $\mathcal{S}_h$. Because of the dependence, we cannot simply sample each variable independently using the likelihoods in (2.3) and (2.4). If we could generate the set of all possible completions for all households with missing entries, conditional on the observed values, then calculating the probability of each one and sampling from the set would be straightforward. Unfortunately, this approach is not practical when the size of each $\mathcal{S}_h$ is large. Even when the size of each $\mathcal{S}_h$ is modest, each household could have different sets of completions, necessitating significant computing, storage, and memory requirements.

However, the full conditional in S9 takes a similar form as the kernel of the truncated NDPMPM in (2.1), so that we can generate the desired samples through a second rejection sampling scheme. Essentially, we sample from an untruncated version of the full conditional $P_{\mathbf{X}_i^{\text{mis}}}^* = \pi_{G_i^1} \prod_{k \mid a_{ik}=1}^{p+q} \lambda_{G_i^1 X_{ik}^1}^{(k)} \left(\prod_{j=1}^{n_i} \omega_{G_i^1 M_{ij}^1} \prod_{k \mid b_{ijk}=1}^{p} \phi_{G_i^1 M_{ij}^1 X_{ijk}^1}^{(k)} \right)$, until we obtain a valid sample that satisfies $\mathbf{X}_i^1 \notin \mathcal{S}_h$;

see the Appendix for a proof that this rejection sampling scheme results in a valid Gibbs sampler. Notice that since $P^*_{\mathbf{X}^{\text{mis}}_i}$ itself is untruncated, we can generate samples from it by sampling each variable independently using (2.3) and (2.4). We therefore replace step S9 with S9'.

S9'. For $i = 1, \ldots, n$, sample $\mathbf{X}^{\text{mis}}_i$ as follows.

(a) For each missing household-level variable, that is, each variable where $k \in \{p + 1, \ldots, p + q\}$ with $a_{ik} = 1$, sample X^1_{ik} using (2.3).

(b) For each missing individual-level variable, that is, each variable where $j = 1, \ldots, n_i$ and $k \in \{1, \ldots, p\}$ with $b_{ijk} = 1$, sample X^1_{ijk} using (2.4).

(c) Set the sampled household-level and individual-level values to $\mathbf{X}^{\text{mis}*}_i$.

(d) Combine $\mathbf{X}^{\text{mis}*}_i$ with the observed $\mathbf{X}^{\text{obs}}_i$, that is, set $\mathbf{X}^{1*}_i = (\mathbf{X}^{\text{obs}}_i, \mathbf{X}^{\text{mis}*}_i)$. If $\mathbf{X}^{1*}_i \notin \mathcal{S}_h$, set $\mathbf{X}^{\text{mis}}_i = \mathbf{X}^{\text{mis}*}_i$, otherwise, return to step (9'a).

To initialize each $\mathbf{X}^{\text{mis}}_i$, we suggest sampling from the empirical marginal distribution of each variable k using the available cases for each variable, and requiring that the household satisfies $\mathbf{X}^1_i \notin \mathcal{S}_h$.

4 Strategies for speeding up the MCMC sampler

The rejection sampling step in the Gibbs sampler in Section 2.2 can be inefficient when $\mathcal{S}$ is large (Manrique-Vallier and Reiter, 2014; Hu et al., 2018), as the sampler tends to generate many impossible households before getting enough feasible ones. In addition, it takes computing time to check whether or not each sampled household satisfies all the structural zero rules. These computational costs are compounded when the sampler also incorporates missing values. In this section, we present two strategies that can reduce the number of impossible households that the algorithm generates, thereby speeding up the sampler. The Appendix includes simulation studies showing that both strategies can speed up the MCMC significantly.

4.1 Moving the household head to the household level

Many datasets include a variable recording the relationship of each individual to the household head. There can be only one household head in any household. This restriction can account for a large proportion of the combinations in $\mathcal{S}$. As a simple working example, consider a dataset that contains $n = 1,000$ households of size two, resulting in a total of $N = 2,000$ individuals. Suppose the data contain no household-level variables and two individual-level variables, age and relationship to household head. Also, suppose age has 100 levels while relationship to household head has 13 levels, which include household head, spouse of the household head, etc. Then, $\mathcal{C}$ contains $13^2 \times 100^2 = 1.69 \times 10^6$ combinations. Suppose the rule, "each household must contain exactly one head", is the only structural zero rule defined on the dataset. Then, $\mathcal{S}$ contains 1.45×10^6 impossible combinations, approximately 86% the size of $\mathcal{C}$. If, for example, the model assigns uniform probability to all combinations in $\mathcal{C}$, we would expect to sample about

$(0.86/0.14) * 1{,}000 \approx 6{,}143$ impossible households at every iteration to augment the n feasible households.

Instead, we treat the variables for the household head as a household-level characteristic. This eliminates structural zero rules defined on the household head alone. Using the working example, moving the household head to the household level results in one new household-level variable, age of household head, which has 100 levels. The relationship to household head variable can be ignored for household heads. For others in the household, the relationship to household head variable now has 12 levels, with the level corresponding to "household head" removed. Thus, $\mathcal{C}$ contains $12 \times 100^2 = 1.20 \times 10^5$ combinations, and $\mathcal{S}$ contains zero impossible combinations. We wouldn't even need to sample impossible households in the Gibbs sampler in Section 2.2.

In general, this strategy can reduce the size of $\mathcal{S}$ significantly, albeit usually not to zero as in the simple example here since $\mathcal{S}$ usually contains combinations resulting from other types of structural zero rules. This strategy is not a replacement for the rejection sampler in Section 2.2; rather, it is a data reformatting technique that can be combined with the sampler.

4.2 Setting an upper bound on the number of impossible households to sample

To reduce computation time, we can put an upper bound on the number of sampled cases in $\mathcal{X}^0$. One way to achieve this is to replace n_{1h} in step S1(f) of Section 2.2 with $\lceil n_{1h} \times \psi_h \rceil$, for some ψ_h such that $1/\psi_h$ is a positive integer, so that we sample only approximately $\lceil n_{0h} \times \psi_h \rceil$ impossible households for each $h \in \mathcal{H}$. However, doing so underestimates the actual probability mass assigned to $\mathcal{S}$ by the model. We can illustrate this using the simple example of Section 4.1. Suppose the model assigns uniform probability to all combinations in $\mathcal{C}$ as before. We set $\psi_2 = 0.5$, so that we sample approximately $3{,}072 = \lceil 6{,}143 \times 0.5 \rceil$ impossible households in every iteration of the MCMC sampler. The probability of generating one impossible household is $3{,}072/(1{,}000 + 3{,}072) = 0.75$, a decrease from the actual value of 0.86. Therefore, we would underestimate the true contribution of $\{\mathcal{X}^0, \mathbf{G}^0, \mathbf{M}^0\}$ to the likelihood.

To use the cap-and-weight approach, we need to apply a correction that re-weights the contribution of $\{\mathcal{X}^0, \mathbf{G}^0, \mathbf{M}^0\}$ to the full joint likelihood. We do so using ideas akin to those used by Chambers and Skinner (2003) and Savitsky and Toth (2016), approximating the likelihood of the full unobserved data with a "pseudo" likelihood using weights (the $1/\psi_h$'s). The impossible households only contribute to the full joint likelihood through the discrete distributions in (2.3) to (2.6). The sufficient statistics for estimating the parameters of the discrete distributions in (2.3) to (2.6) are the observed counts for the corresponding variables in the set $\{\mathcal{X}^1, \mathbf{G}^1, \mathbf{M}^1, \mathcal{X}^0, \mathbf{G}^0, \mathbf{M}^0\}$, within each latent class for the household-level variables and within each latent class pair for the individual-level variables. Thus, for each $h \in \mathcal{H}$, we can re-weight the contribution of impossible households by multiplying the observed counts for households of size h in $\{\mathcal{X}^0, \mathbf{G}^0, \mathbf{M}^0\}$ by $1/\psi_h$ for the corresponding variable and latent classes. This raises the likelihood contribution of impossible households of size h to the power of $1/\psi_h$. Clearly, $1/\psi_h$ need not be a positive integer. We require that only to make its multiplication with the observed counts free of decimals. We

modify the Gibbs sampler to incorporate the cap-and-weight approach by replacing steps S1, S3, S4, S5 and S6; see the Appendix for the modified steps.

Setting each $\psi_h = 1$ corresponds to the original rejection sampler, so that the two approaches should provide very similar results when ψ_h near 1. Based on our experience, results of the cap-and-weight approach become significantly less accurate than the regular rejection sampler when $\psi_h < 1/4$. The time gained using this speedup approach in comparison to the regular sampler depends on the features of the data and the specified values for the weights $\{\psi_h: h \in \mathcal{H}\}$. To select the ψ_h's, we suggest trying out different values – starting with values close to one – in initial runs of the MCMC sampler on a small random sample of the data. Analysts should examine the convergence and mixing behavior of the chains in comparison to the chain with all the ψ_h's set to one, and select values that offer reasonable speedup while preserving convergence and mixing. This can be done quickly by comparing trace plots of a random set of parameters from the model that are not subject to label switching, such as α and β, or by examining marginal, bivariate and trivariate probabilities estimated from synthetic data generated from the MCMC.

5 Empirical study

To evaluate the performance of the NDPMPM as an imputation method, as well as the speed up strategies, we use data from the public use microdata files from the 2012 ACS, available for download from the United States Census Bureau (http://www2.census.gov/acs2012_1yr/pums/). We construct a population of 764,580 households of sizes $\mathcal{H} = \{2, 3, 4\}$, from which we sample $n = 5,000$ households comprising $N = 13,181$ individuals. We work with the variables described in Table 5.1, which mimic those in the U.S. decennial census. The structural zeros involve ages and relationships of individuals in the same house; see the Appendix for a full list of rules that we used. We move the household head to the household level as in Section 4.1 to take advantage of the computational gains.

We introduce missing values using the following scenario. We let household size and age of household heads be fully observed. We randomly and independently blank 30% of each variable for the remaining household-level variables. For individuals other than the household head, we randomly and independently blank 30% of the values for gender, race and Hispanic origin. We make age missing with rates 50%, 20%, 40% and 30% for values of the relationship variable in the sets $\{2\}$, $\{3, 4, 5, 10\}$, $\{7, 9\}$ and $\{6, 8, 11, 12, 13\}$, respectively. We make the relationship variable missing with rates 40%, 25%, 10%, and 55% for values of age in the sets $\{x: x \leq 20\}$, $\{x: 20 < x \leq 50\}$, $\{x: 50 < x \leq 70\}$, and $\{x: x > 70\}$, respectively. This results in approximately 30% missing values for both variables. About 8% of the individuals in the sample are missing both the age and relationship variable, and 2% are missing gender, age, and relationship jointly. This mechanism results in data that technically are not missing at random, but we use the NDPMPM approach regardless to examine its potential in a complicated missingness mechanism. Actual rates of item nonresponse in census data tend to be smaller than what we use here, but we use high rates to put the NDPMPM through a challenging stress test. We also introduce missing values

using a missing completely at random scenario with rates in the 10% range across all the variables. In short, the results are similar to those here, though more accurate due to the lower rates of missingness. See the Appendix for the results.

Table 5.1
Description of variables used in the study. "HH" means household head

Description of variable		Categories
Household-level variables	Ownership of dwelling	1 = owned or being bought, 2 = rented
	Household size	2 = 2 people, 3 = 3 people, 4 = 4 people
	Gender of HH	1 = male, 2 = female
	Race of HH	1 = white, 2 = black, 3 = American Indian or Alaska native, 4 = Chinese, 5 = Japanese, 6 = other Asian/Pacific islander, 7 = other race, 8 = two major races, 9 = three or more major races
	Hispanic origin of HH	1 = not Hispanic, 2 = Mexican, 3 = Puerto Rican, 4 = Cuban, 5 = other
	Age of HH	1 = less than one year old, 2 = 1 year old, 3 = 2 years old, ..., 96 = 95 years old
Individual-level variables	Gender	same as "Gender of HH"
	Race	same as "Race of HH"
	Hispanic origin	same as "Hispanic origin of HH"
	Age	same as "Age of HH"
	Relationship to head of household	1 = spouse, 2 = biological child, 3 = adopted child, 4 = stepchild, 5 = sibling, 6 = parent, 7 = grandchild, 8 = parent-in-law, 9 = child-in-law, 10 = other relative, 11 = boarder, roommate or partner, 12 = other non-relative or foster child

We estimate the NDPMPM using two approaches, both using the rejection step S9' in Section 3. The first approach considers $\psi_2 = \psi_3 = \psi_4 = 1$, i.e., without using the cap-and-weight approach, while the second approach considers $\psi_2 = \psi_3 = 1/2$ and $\psi_4 = 1/3$. For each approach, we run the MCMC sampler for 10,000 iterations, discarding the first 5,000 as burn-in and thinning the remaining samples every five iterations, resulting in 1,000 MCMC post burn-in iterates. We set $F = 30$ and $S = 15$ for each approach based on initial tuning runs. Across the approaches, the effective number of occupied household-level clusters usually ranges from 13 to 16 with a maximum of 25, while the effective number of occupied individual-level clusters across all household-level clusters ranges from 3 to 5 with a maximum of 10. For convergence, we examined trace plots of α, β, and weighted averages of a random sample of the multinomial probabilities in (2.3) and (2.4) (since the multinomial probabilities themselves are prone to label switching).

For both methods, we generate $L = 50$ completed datasets, $\mathbf{Z} = (\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(50)})$, using the posterior predictive distribution of the NDPMPM, from which we estimate all marginal distributions, bivariate distributions of all possible pairs of variables, and trivariate distributions of all possible triplets of variables.

We also estimate several probabilities that depend on within household relationships and the household head to investigate the performance of the NDPMPM in estimating complex relationships. We obtain confidence intervals using multiple imputation inferences (Rubin, 1987). As a brief review, let q be the completed-data point estimator of some estimand Q, and let u be the estimator of variance associated with q. For $l = 1, \ldots, L$, let $q^{(l)}$ and $u^{(l)}$ be the values of q and u in completed dataset $\mathbf{Z}^{(l)}$. We use $\bar{q}_L = \sum_{l=1}^{L} q^{(l)} / L$ as the point estimate of Q. We use $T_L = (1 + 1/L) b_L + \bar{u}_L$ as the estimated variance of $\bar{q}$, where $b_L = \sum_{l=1}^{L} (q^{(l)} - \bar{q}_L)^2 / (L - 1)$ and $\bar{u}_L = \sum_{l=1}^{L} u^{(l)} / L$. We make inference about Q using $(\bar{q}_L - Q) \sim t_v (0, T_L)$, where t_v is a $t-$distribution with $v = (L - 1)(1 + \bar{u}_L / [(1 + 1/L) b_L])^2$ degrees of freedom.

Figures 5.1 and 5.2 display the value of $\bar{q}_{50}$ for each estimated marginal, bivariate and trivariate probability plotted against its corresponding estimate from the original data, without missing values. Figure 5.1 shows the results for the NDPMPM with the rejection sampler, and Figure 5.2 shows the results for the NDPMPM using the cap-and-weight approach. For both approaches, the point estimates are close to those from the data before introducing missing values, suggesting that the NDPMPM does a good job of capturing important features of the joint distribution of the variables. Figure 5.2 in particular also shows that the cap-and-weight approach did not degrade the estimates.

Table 5.2 displays 95% confidence intervals for several probabilities involving within-household relationships, as well as the value in the full population of 764,580 households. The intervals include the two based on the NDPMPM imputation engines and the interval from the data before introducing missingness. For the latter, we use the usual Wald interval, $\hat{p} \pm 1.96 \sqrt{\hat{p}(1 - \hat{p})/n}$, where $\hat{p}$ is the corresponding sample percentage. For the most part, the intervals from the NDPMPM with the full rejection sampling are close to those based on the data without any missingness. They tend to include the true population quantity. The NDPMPM imputation engine results in noticeable downward bias for the percentages of households where everyone is the same race, with bias increasing as the household size gets bigger. This is a challenging estimand to estimate accurately via imputation, particularly for larger households. Hu et al. (2018) identified biases in the same direction when using the NDPMPM (with household head data treated as individual-level variables) to generate fully synthetic data, noting that the bias gets smaller as the sample size increases. The NDPMPM fits the joint distribution of the data better and better as the sample size grows. Hence, we expect the NDPMPM imputation engine to be more accurate with larger sample sizes, as well as with smaller fractions of missing values.

The interval estimates from the cap-and-weight method are generally similar to those for the full rejection sampler, with some degradation particularly for the percentages of same race households by household size. This degradation comes with a benefit, however. Based on MCMC runs on a standard laptop, the NDPMPM using the cap-and-weight approach and moving household heads' data values to the household level is about 42% faster than the NDPMPM with household heads' data values moved to the household level.
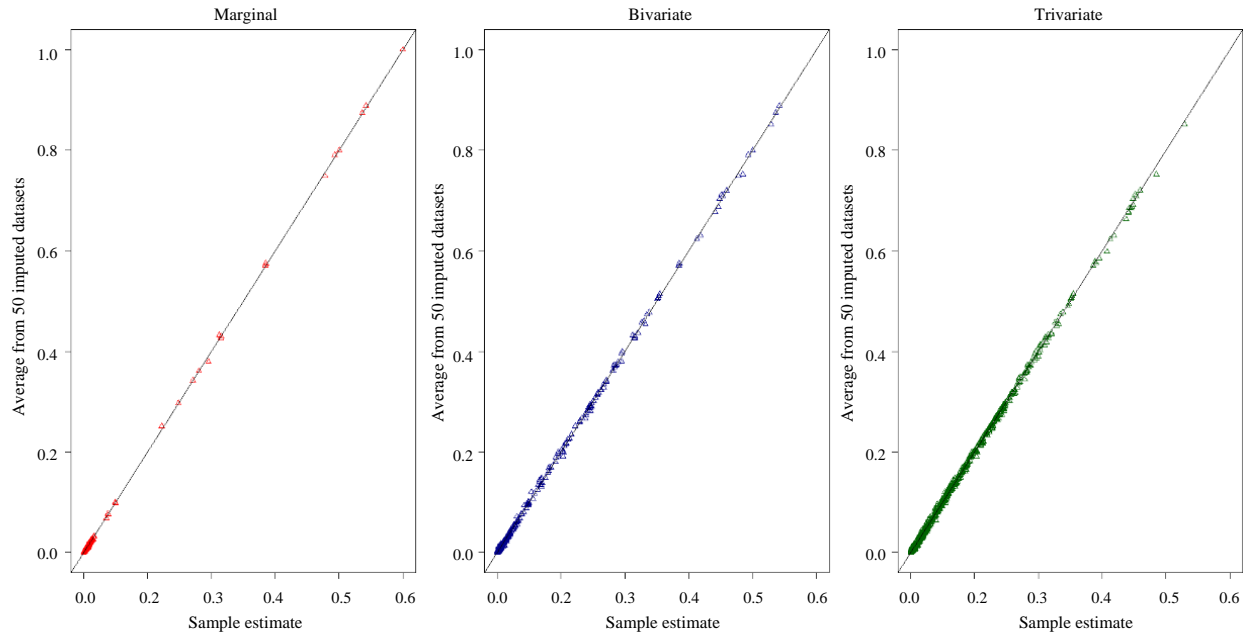
Figure 5.1 **Marginal, bivariate and trivariate probabilities computed in the sample and imputed datasets from the truncated NDPMPM with the rejection sampler. Household heads' data values moved to the household level.**
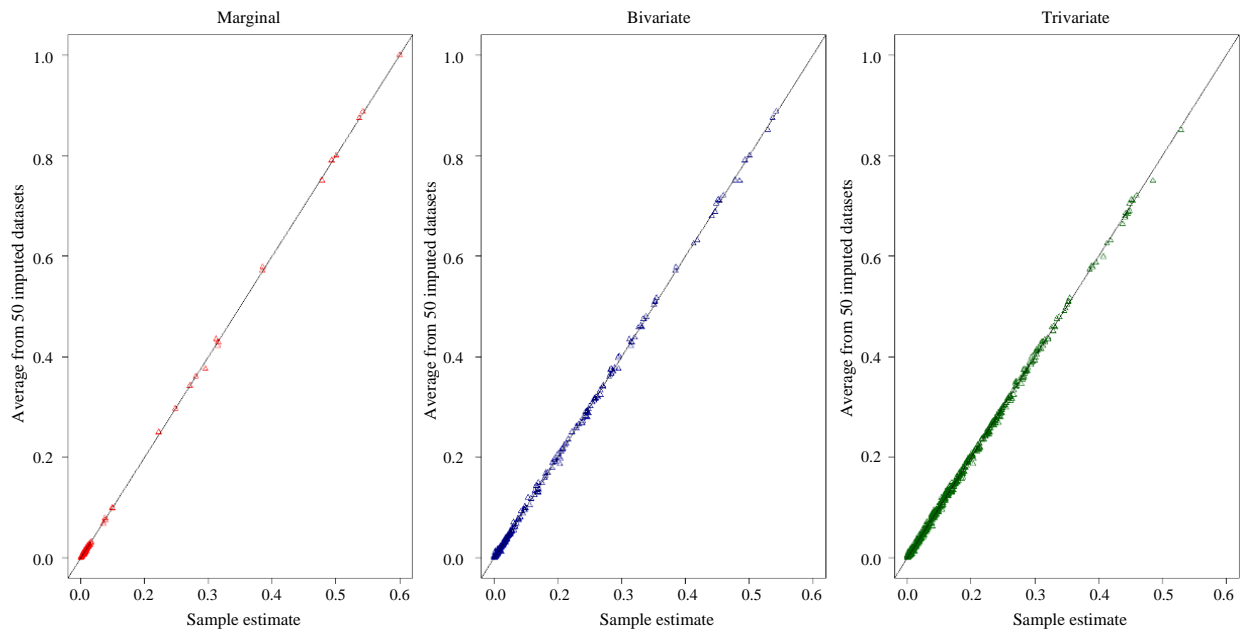

Figure 5.2 **Marginal, bivariate and trivariate probabilities computed in the sample and imputed datasets from the truncated NDPMPM using the cap-and-weight approach. Household heads' data values to the household level.**

Table 5.2
Confidence intervals for selected probabilities that depend on within-household relationships in the original and imputed datasets. "No missing" is based on the sampled data before introducing missing values, "NDPMPM" uses the truncated NDPMPM, moving household heads' data values to the household level, and "NDPMPM Capped" uses the truncated NDPMPM with the cap-and-weight approach and moving household heads' data values to the household level. "HH" means household head, "SP" means spouse, "CH" means child, and "CP" means couple. Q is the value in the full population of 764,580 households

		Q	No Missing	NDPMPM	NDPMPM Capped
All same race household:	$n_i = 2$	0.942	(0.932, 0.949)	(0.891, 0.917)	(0.884, 0.911)
	$n_i = 3$	0.908	(0.907, 0.937)	(0.843, 0.890)	(0.821, 0.870)
	$n_i = 4$	0.901	(0.879, 0.917)	(0.793, 0.851)	(0.766, 0.828)
SP present		0.696	(0.682, 0.707)	(0.695, 0.722)	(0.695, 0.722)
Same race CP		0.656	(0.641, 0.668)	(0.640, 0.669)	(0.634, 0.664)
SP present, HH is White		0.600	(0.589, 0.616)	(0.603, 0.632)	(0.604, 0.634)
White CP		0.580	(0.569, 0.596)	(0.577, 0.606)	(0.574, 0.604)
CP with age difference less than five		0.488	(0.465, 0.492)	(0.341, 0.371)	(0.324, 0.355)
Male HH, home owner		0.476	(0.456, 0.484)	(0.450, 0.479)	(0.451, 0.480)
HH over 35, no CH present		0.462	(0.441, 0.468)	(0.442, 0.470)	(0.443, 0.471)
At least one biological CH present		0.437	(0.431, 0.458)	(0.430, 0.459)	(0.428, 0.456)
HH older than SP, White HH		0.322	(0.309, 0.335)	(0.307, 0.339)	(0.311, 0.343)
Adult female w/ at least one CH under 5		0.078	(0.070, 0.085)	(0.062, 0.078)	(0.061, 0.077)
White HH with Hisp origin		0.066	(0.064, 0.078)	(0.062, 0.079)	(0.062, 0.078)
Non-White CP, home owner		0.058	(0.050, 0.063)	(0.038, 0.052)	(0.037, 0.051)
Two generations present, Black HH		0.057	(0.053, 0.066)	(0.052, 0.066)	(0.052, 0.067)
Black HH, home owner		0.052	(0.046, 0.058)	(0.044, 0.058)	(0.044, 0.059)
SP present, HH is Black		0.039	(0.032, 0.042)	(0.032, 0.044)	(0.031, 0.043)
White-nonwhite CP		0.034	(0.029, 0.039)	(0.038, 0.053)	(0.043, 0.059)
Hisp HH over 50, home owner		0.029	(0.025, 0.034)	(0.023, 0.034)	(0.024, 0.034)
One grandchild present		0.028	(0.023, 0.033)	(0.024, 0.035)	(0.023, 0.035)
Adult Black female w/ at least one CH under 18		0.027	(0.028, 0.038)	(0.025, 0.036)	(0.025, 0.036)
At least two generations present, Hisp CP		0.027	(0.022, 0.031)	(0.022, 0.032)	(0.023, 0.033)
Hisp CP with at least one biological CH		0.025	(0.020, 0.028)	(0.019, 0.029)	(0.020, 0.030)
At least three generations present		0.023	(0.020, 0.028)	(0.017, 0.026)	(0.017, 0.026)
Only one parent		0.020	(0.016, 0.024)	(0.013, 0.021)	(0.013, 0.021)
At least one stepchild		0.019	(0.018, 0.026)	(0.019, 0.030)	(0.019, 0.030)
Adult Hisp male w/ at least one CH under 10		0.018	(0.017, 0.025)	(0.014, 0.022)	(0.014, 0.022)
At least one adopted CH, White CP		0.008	(0.005, 0.010)	(0.004, 0.010)	(0.004, 0.011)
Black CP with at least two biological children		0.006	(0.003, 0.007)	(0.003, 0.007)	(0.003, 0.007)
Black HH under 40, home owner		0.005	(0.005, 0.009)	(0.006, 0.013)	(0.007, 0.013)
Three generations present, White CP		0.005	(0.004, 0.008)	(0.004, 0.010)	(0.004, 0.009)
White HH under 25, home owner		0.003	(0.002, 0.005)	(0.003, 0.007)	(0.003, 0.007)

6 Discussion

The empirical study suggests that the NDPMPM can provide high quality imputations for categorical data nested within households. To our knowledge, this is the first parametric imputation engine for nested multivariate categorical data. The study also illustrates that, with modest sample sizes, agencies should not expect the NDPMPM to preserve all features of the joint distribution. Of course, this is the case with any imputation engine. For the NDPMPM, agencies may be able to improve accuracy for targeted quantities by recoding the data used to fit the model. For example, one can create a new household-level variable that equals one when everyone has the same race and equals zero otherwise, and replace the individual race variable with a new variable that has levels "1 = race is the same as race of household head", "2 = race is white and differs from race of household head", "3 = race is black and differs from race of household head", and so on. The NDPMPM would be estimated with the household-level same race variable and the new individual-level race variable. This would encourage the NDPMPM to estimate the percentages with the

same race very accurately, as it would be just another household-level variable like home ownership. It also would add structural zeros involving race to the computation. Evaluating the trade offs in accuracy and computational costs of such recodings is a topic for future research.

The NDPMPM can be computationally expensive, even with the speed-ups presented in this article. The expensive parts of the algorithm are the rejection sampling steps. Fortunately, these can be done easily by parallel processing. For example, we can require each processor to generate a fraction of the impossible cases in Section 2.2. We also can spread the rejection steps for the imputations over many processors. These steps should cut run time by a factor roughly equal to the number of processors available.

The empirical study used households up to size four. We have run the model on data with households up to size seven in reasonable time (a few hours on a standard laptop). Accuracy results are similar qualitatively. As the household sizes get large, the model can generate hundreds or even thousands times as many impossible households as there are feasible ones, slowing the algorithm. In such cases, the cap-and-weight approach is essential for practical applications.

Acknowledgements

Appendix

This is an Appendix to the paper. It contains proof that the rejection sampling step S9' in Section 3 generates samples from the correct posterior distribution. It also contains the modified Gibbs sampler for the cap-and-weight approach and a list of the structural zero rules used in fitting the NDPMPM model. Finally, we include empirical results for the speedup approaches mentioned in the paper, using synthetic data, and additional results for handling missing data using the NDPMPM under a missing completely at random scenario.

A.1 Proof that the rejection sampling step S9' in Section 3 generates samples from the correct posterior distribution

The X_{ik}^1 and X_{ijk}^1 values generated using the rejection sampler in Step S9' are generated from the full conditionals, resulting in a valid Gibbs sampler. The proof follows from the properties of rejection sampling (or simple accept reject). The target distribution is the full conditional for $\mathbf{X}_i^{\text{mis}}$. It can be re-expressed as

$$p\left(\mathbf{X}_i^{\text{mis}}\right) = \frac{\mathbb{1}\left\{\mathbf{X}_i^1 \notin \mathcal{S}_h\right\}}{\Pr\left(\mathbf{X}_i \notin \mathcal{S}_h \mid \theta\right)} g\left(\mathbf{X}_i^{\text{mis}}\right)$$

where

$$g\left(\mathbf{X}_i^{\mathrm{mis}}\right) = \pi_{G_i^1} \prod_{k\,|\,a_{ik}=1}^{p+q} \lambda_{G_i^1 X_{ik}^1}^{(k)} \left(\prod_{j=1}^{n_i} \omega_{G_i^1 M_{ij}^1} \prod_{k\,|\,b_{ijk}=1}^{p} \phi_{G_i^1 M_{ij}^1 X_{ijk}^1}^{(k)}\right).$$

Our rejection scheme uses $g\left(\mathbf{X}_i^{\mathrm{mis}}\right)$ as a proposal for $p\left(\mathbf{X}_i^{\mathrm{mis}}\right)$. To show that the draws are indeed from $p\left(\mathbf{X}_i^{\mathrm{mis}}\right)$, we need to verify that $w\left(\mathbf{X}_i^{\mathrm{mis}}\right) = p\left(\mathbf{X}_i^{\mathrm{mis}}\right)/g\left(\mathbf{X}_i^{\mathrm{mis}}\right) < M$, where $1 < M < \infty$, and that we are accepting each sample with probability $w\left(\mathbf{X}_i^{\mathrm{mis}}\right)/M$. In our case,

1. $w\left(\mathbf{X}_i^{\mathrm{mis}}\right) = p\left(\mathbf{X}_i^{\mathrm{mis}}\right)/g\left(\mathbf{X}_i^{\mathrm{mis}}\right) = \mathbb{1}\{\mathbf{X}_i^1 \notin \mathcal{S}_h\}/\Pr\left(\mathbf{X}_i \notin \mathcal{S}_h \mid \theta\right) \leq 1/\Pr\left(\mathbf{X}_i \notin \mathcal{S}_h \mid \theta\right)$, and $0 < \Pr\left(\mathbf{X}_i \notin \mathcal{S}_h \mid \theta\right) < 1 \Rightarrow 1 < 1/\Pr\left(\mathbf{X}_i \notin \mathcal{S}_h \mid \theta\right) < \infty$ necessarily.

2. By sampling until we obtain a valid sample that satisfies $\mathbf{X}_i^1 \notin \mathcal{S}_h$, we are indeed sampling with probability $w\left(\mathbf{X}_i^{\mathrm{mis}}\right)/M = \mathbb{1}\{\mathbf{X}_i^1 \notin \mathcal{S}_h\}$.

A.2 Modified Gibbs sampler for the cap-and-weight approach

The modified Gibbs sampler for the cap-and-weight approach replaces steps S1, S3, S4, S5 and S6 of the Gibbs sampler in the main text as follows.

S1*. For each $h \in \mathcal{H}$, repeat steps S1(a) to S1(e) as before but modify step S1(f) to: if $t_1 < \lceil n_{1h} \times \psi_h \rceil$, return to step (b). Otherwise, set $n_{0h} = t_0$.

S3*. Set $u_F = 1$. Sample

$$u_g \mid - \sim \mathrm{Beta}\left(1 + U_g, \alpha + \sum_{f=g+1}^{F} U_f\right), \quad \pi_g = u_g \prod_{f<g}\left(1 - u_f\right)$$

where

$$U_g = \sum_{i=1}^{n} \mathbb{1}\left(G_i^1 = g\right) + \sum_{h\in\mathcal{H}} \frac{1}{\psi_h} \sum_{i\,|\,n_i^0=h} \mathbb{1}\left(G_i^0 = g\right)$$

for $g = 1, \ldots, F - 1$.

S4*. Set $v_{gM} = 1$ for $g = 1, \ldots, F$. Sample

$$v_{gm} \mid - \sim \mathrm{Beta}\left(1 + V_{gm}, \beta + \sum_{s=m+1}^{S} V_{gs}\right), \quad \omega_{gm} = v_{gm} \prod_{s<m}\left(1 - v_{gs}\right)$$

where

$$V_{gm} = \sum_{i=1}^{n} \mathbb{1}\left(M_{ij}^1 = m, G_i^1 = g\right) + \sum_{h\in\mathcal{H}} \frac{1}{\psi_h} \sum_{i\,|\,n_i^0=h} \mathbb{1}\left(M_{ij}^0 = m, G_i^0 = g\right)$$

for $m = 1, \ldots, S - 1$ and $g = 1, \ldots, F$.

S5*. Sample

$$\lambda_g^{(k)} \big| - \sim \text{Dirichlet}\left(1 + \eta_{g1}^{(k)}, \dots, 1 + \eta_{gd_k}^{(k)}\right)$$

where

$$\eta_{gc}^{(k)} = \sum_{i \,|\, G_i^1 = g}^{n} \mathbb{1}\left(X_{ik}^1 = c\right) + \sum_{h \in \mathcal{H}} \frac{1}{\psi_h} \sum_{i \,|\, n_i^0 = h, \, G_i^0 = g} \mathbb{1}\left(X_{ik}^0 = c\right)$$

for $g = 1, \dots, F$ and $k = p + 1, \dots, q$.

S6*. Sample

$$\phi_{gm}^{(k)} \big| - \sim \text{Dirichlet}\left(1 + \nu_{gm1}^{(k)}, \dots, 1 + \nu_{gmd_k}^{(k)}\right)$$

where

$$\nu_{gmc}^{(k)} = \sum_{i \,|\, G_i^1 = g, \, M_{ij}^1 = m}^{n} \mathbb{1}\left(X_{ijk}^1 = c\right) + \sum_{h \in \mathcal{H}} \frac{1}{\psi_h} \sum_{i \,|\, n_i^0 = h, \, G_i^0 = g, \, M_{ij}^0 = m} \mathbb{1}\left(X_{ijk}^0 = c\right)$$

for $g = 1, \dots, F$, $m = 1, \dots, S$ and $k = 1, \dots, p$.

A.3 List of structural zeros

We fit the NDPMPM model using structural zeros which involve ages and relationships of individuals in the same house. The full list of the rules used is presented in Table A.1. These rules were derived from the 2012 ACS by identifying combinations involving the relationship variable that do not appear in the constructed population. This list should not be interpreted as a "true" list of impossible combinations in census data.

Table A.1
List of structural zeros

Description
Rules common to generating both the synthetic and imputed datasets
1. Each household must contain exactly one head and he/she must be at least 16 years old.
2. Each household cannot contain more than one spouse and he/she must be at least 16 years old.
3. Married couples are of opposite sex, and age difference between individuals in the couples cannot exceed 49.
4. The youngest parent must be older than the household head by at least 4.
5. The youngest parent-in-law must be older than the household head by at least 4.
6. The age difference between the household head and siblings cannot exceed 37.
7. The household head must be at least 31 years old to be a grandparent and his/her spouse must be at least 17. Also, He/she must be older than the oldest grandchild by at least 26.
Rules specific to generating the synthetic datasets
8. The household head must be older than the oldest child by at least 7.
Rules specific to generating the imputed datasets
9. The household head must be older than the oldest biological child by at least 7.
10. The household head must be older than the oldest adopted child by at least 11.
11. The household head must be older than the oldest stepchild by at least 9.

A.4 Empirical study of the speedup approaches

We evaluate the performance of the two speedup approaches mentioned in the main text using synthetic data. We use data from the public use microdata files from the 2012 ACS, available for download from the United States Census Bureau (http://www2.census.gov/acs2012_1yr/pums/) to construct a population of 857,018 households of sizes $\mathcal{H} = \{2, 3, 4, 5, 6\}$, from which we sample $n = 10,000$ households comprising $N = 29,117$ individuals. We work with the variables described in Table A.2. We evaluate the approaches using probabilities that depend on within household relationships and the household head.

Table A.2
Description of variables used in the synthetic data illustration

Description of variable		Categories
Household-level variables	Ownership of dwelling	1 = owned or being bought, 2 = rented
	Household size	2 = 2 people, 3 = 3 people, 4 = 4 people, 5 = 5 people, 6 = 6 people
Individual-level variables	Gender	1 = male, 2 = female
	Race	1 = white, 2 = black, 3 = American Indian or Alaska native, 4 = Chinese, 5 = Japanese, 6 = other Asian/Pacific islander, 7 = other race, 8 = two major races, 9 = three or more major races
	Hispanic origin	1 = not Hispanic, 2 = Mexican, 3 = Puerto Rican, 4 = Cuban, 5 = other
	Age	1 = less than one year old, 2 = 1 year old, 3 = 2 years old, ..., 96 = 95 years old
	Relationship to head of household	1 = household head, 2 = spouse, 3 = child, 4 = child-in-law, 5 = parent, 6 = parent-in-law, 7 = sibling, 8 = sibling-in-law, 9 = grandchild, 10 = other relative, 11 = partner/friend/visitor, 12 = other non-relative

We consider the NDPMPM using two approaches, both moving the values of the household head to the household level as in Section 4.1 of the main text and also using the cap-and-weight approach in Section 4.2 of the main text. The first approach considers $\psi_2 = \psi_3 = \psi_4 = \psi_5 = \psi_6 = 1$ while the second approach considers $\psi_2 = \psi_3 = 1/2$ and $\psi_4 = \psi_5 = \psi_6 = 1/3$. We compare these approaches to the NDPMPM as presented in Hu et al., 2018. For each approach, we create $L = 50$ synthetic datasets, $\mathbf{Z} = (\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(50)})$. We generate the synthetic datasets so that the number of households of size $h \in \mathcal{H}$ in each $\mathbf{Z}^{(l)}$ exactly matches n_h from the observed data. Thus, $\mathbf{Z}$ comprises partially synthetic data (Little, 1993; Reiter, 2003), even though every released Z_{ijk} is a simulated value. We combine the estimates using using the approach in Reiter (2003). As a brief review, let q be the point estimator of some estimand Q, and let u be the estimator of variance associated with q. For $l = 1, \ldots, L$, let q_l and u_l be the values of q and u in synthetic dataset $\mathbf{Z}^{(l)}$. We use $\bar{q} = \sum_{l=1}^{L} q_l \big/ L$ as the point estimate of Q and $T = \bar{u} + b/L$ as the estimated variance of $\bar{q}$, where $b = \sum_{l=1}^{L} (q_l - \bar{q})^2 \big/ (L-1)$ and $\bar{u} = \sum_{l=1}^{L} u_l \big/ L$. We make inference about Q using $(\bar{q} - Q) \sim t_v(0, T)$, where t_v is a $t-$distribution with $v = (L-1)(1 + L\bar{u}/b)^2$ degrees of freedom.

For each approach, we run the MCMC sampler for 20,000 iterations, discarding the first 10,000 as burn-in and thinning the remaining samples every five iterations, resulting in 2,000 MCMC post burn-in iterates. We create the $L = 50$ synthetic datasets by randomly sampling from the 2,000 iterates. We set $F = 40$ and $S = 15$ for each approach based on initial tuning runs. For convergence, we examined trace plots of α, β and weighted averages of a random sample of the multinomial probabilities in the NDPMPM likelihood. Across the approaches, the effective number of occupied household-level clusters usually ranges from 20 to 33 with a maximum of 38, while the effective number of occupied individual-level clusters across all household-level clusters ranges from 5 to 9 with a maximum of 12.

Based on MCMC runs on a standard laptop, moving household heads' data values to the household level alone results in a speedup of about 63% on the default rejection sampler while the cap-and-weight approach alone results in a speedup of about 40%.

Table A.3 shows the 95% confidence intervals for each approach. Essentially, all three approaches result in similar confidence intervals, suggesting not much loss in accuracy from the speedups. Most intervals also are reasonably similar to confidence intervals based on the original data, except for the percentage of same age couples. The last row is a rigorous test of how well each method can estimate a probability that can be fairly difficult to estimate accurately. In this case, the probability that a household head and spouse are the same age can be difficult to estimate since each individual's age can take 96 different values. All three approaches are thus off from the estimate from the original data in this case. These results suggest that we can significantly speedup the sampler with minimal loss in accuracy of estimates and confidence intervals of population estimands.

Table A.3
Confidence intervals for selected probabilities that depend on within-household relationships in the original and synthetic datasets. "Original" is based on the sampled data, "NDPMPM" is the default MCMC sampler described in Section 2.2 of the main text, "NDPMPM w/ HH moved" is the default sampler, moving household heads' data values to the household level, "NDPMPM capped w/ HH moved" uses the cap-and-weight approach and moving household heads' data values to the household level. "HH" means household head and "SP" means spouse

		Original	NDPMPM	NDPMPM w/ HH moved	NDPMPM capped w/ HH moved
All same race	$n_i = 2$	(0.939, 0.951)	(0.918, 0.932)	(0.912, 0.928)	(0.910, 0.925)
	$n_i = 3$	(0.896, 0.920)	(0.859, 0.888)	(0.845, 0.875)	(0.844, 0.874)
	$n_i = 4$	(0.885, 0.912)	(0.826, 0.860)	(0.813, 0.848)	(0.817, 0.852)
	$n_i = 5$	(0.879, 0.922)	(0.786, 0.841)	(0.786, 0.841)	(0.777, 0.834)
	$n_i = 6$	(0.831, 0.910)	(0.701, 0.803)	(0.718, 0.819)	(0.660, 0.768)
SP present		(0.693, 0.711)	(0.678, 0.697)	(0.676, 0.695)	(0.677, 0.695)
SP with white HH		(0.589, 0.608)	(0.577, 0.597)	(0.576, 0.595)	(0.575, 0.595)
SP with black HH		(0.036, 0.043)	(0.035, 0.043)	(0.034, 0.042)	(0.034, 0.042)
White couple		(0.570, 0.589)	(0.560, 0.580)	(0.553, 0.573)	(0.552, 0.572)
White couple, own		(0.495, 0.514)	(0.468, 0.488)	(0.461, 0.481)	(0.463, 0.483)
Same race couple		(0.655, 0.673)	(0.636, 0.655)	(0.626, 0.645)	(0.625, 0.644)
White-nonwhite couple		(0.028, 0.035)	(0.028, 0.035)	(0.034, 0.041)	(0.036, 0.044)
Nonwhite couple, own		(0.057, 0.067)	(0.047, 0.056)	(0.045, 0.053)	(0.045, 0.054)
Only mother present		(0.017, 0.022)	(0.014, 0.019)	(0.014, 0.019)	(0.013, 0.018)
Only one parent present		(0.021, 0.026)	(0.026, 0.032)	(0.026, 0.033)	(0.027, 0.033)
Children present		(0.507, 0.527)	(0.493, 0.512)	(0.517, 0.537)	(0.511, 0.531)
Siblings present		(0.022, 0.028)	(0.027, 0.034)	(0.027, 0.033)	(0.027, 0.033)
Grandchild present		(0.041, 0.049)	(0.051, 0.060)	(0.049, 0.058)	(0.050, 0.059)
Three generations present		(0.036, 0.044)	(0.037, 0.045)	(0.042, 0.050)	(0.040, 0.048)
White HH, older than SP		(0.309, 0.327)	(0.283, 0.301)	(0.294, 0.313)	(0.302, 0.321)
Nonhisp HH		(0.882, 0.894)	(0.875, 0.888)	(0.879, 0.891)	(0.876, 0.889)
White, Hisp HH		(0.071, 0.082)	(0.074, 0.085)	(0.072, 0.082)	(0.073, 0.084)
Same age couple		(0.087, 0.098)	(0.027, 0.034)	(0.023, 0.029)	(0.024, 0.031)

A.5 Empirical study of missing data imputation under MCAR

We also evaluate the performance of the NDPMPM as an imputation method under a missing completely at random (MCAR) scenario. We use the same data as in Section 5 of the main text. As a reminder, the data contains $n = 5,000$ households of sizes $\mathcal{H} = \{2, 3, 4\}$, comprising $N = 13,181$ individuals. We introduce missing values using a MCAR scenario. We randomly select 80% households to be complete cases for all variables. For the remaining 20%, we let the variable "household size" be fully observed and randomly – and independently – blank 50% of each variable for the remaining household-level and individual-level variables. We use these low rates to mimic the actual rates of item nonresponse in census data.

Similar to the main text, we estimate the NDPMPM using two approaches, both combining the rejection step in Section 4.1 of the main text with the cap-and-weight approach in Section 4.2 of the main text. The first approach considers $\psi_2 = \psi_3 = \psi_4 = 1$ while the second approach considers $\psi_2 = \psi_3 = 1/2$ and $\psi_4 = 1/3$. For each approach, we run the MCMC sampler for 10,000 iterations, discarding the first 5,000 as burn-in and thinning the remaining samples every five iterations, resulting in 1,000 MCMC post burn-in iterates. We set $F = 30$ and $S = 15$ for each approach based on initial tuning runs. We monitor convergence as in the main text. For both methods, we generate $L = 50$ completed datasets, $\mathbf{Z} = (\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(50)})$, using the posterior predictive distribution of the NDPMPM, from which we estimate the same probabilities as in the main text.

Figures A.1 and A.2 display each estimated marginal, bivariate and trivariate probability $\overline{q}_{50}$ plotted against its corresponding estimate from the original data, without missing values. Figure A.1 shows the results for the NDPMPM with the rejection sampler, and Figure A.2 shows the results for the NDPMPM using the cap-and-weight approach. For both approaches, the NDPMPM does a good job of capturing important features of the joint distribution of the variables as the point estimates are very close to those from the data before introducing missing values. In short, the results are very similar to those in the main text, though more accurate.

Table A.4 displays 95% confidence intervals for selected probabilities involving within-household relationships, as well as the value in the full population of 764,580 households. The intervals include the two based on the NDPMPM imputation engines and the interval from the data before introducing missingness. The intervals are generally more accurate than those presented in the main text. This is expected since we use lower rates of missingness in the MCAR scenario. For the most part, the intervals from the NDPMPM with the two approaches tend to include the true population quantity. Again, the NDPMPM imputation engine results in downward bias for the percentages of households where everyone is the same race. As mentioned in the main text, this is a challenging estimand to estimate accurately via imputation, particularly for larger households.
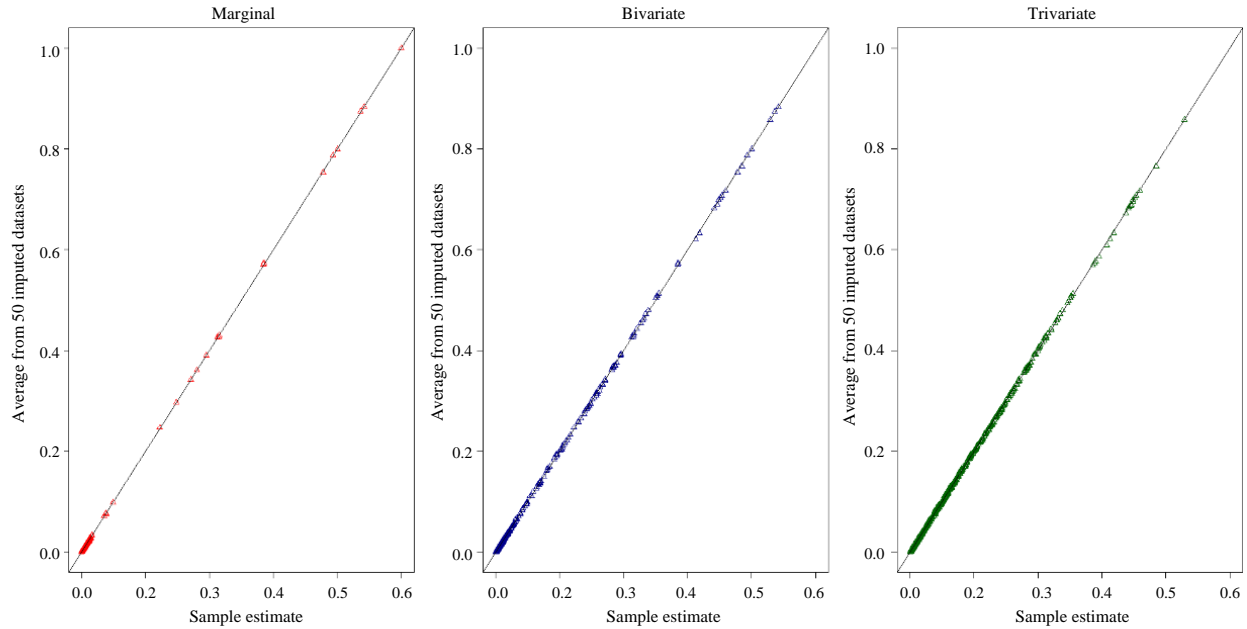
Figure A.1 Marginal, bivariate and trivariate probabilities computed in the sample and imputed datasets under MCAR from the truncated NDPMPM with the rejection sampler. Household heads' data values moved to the household level.
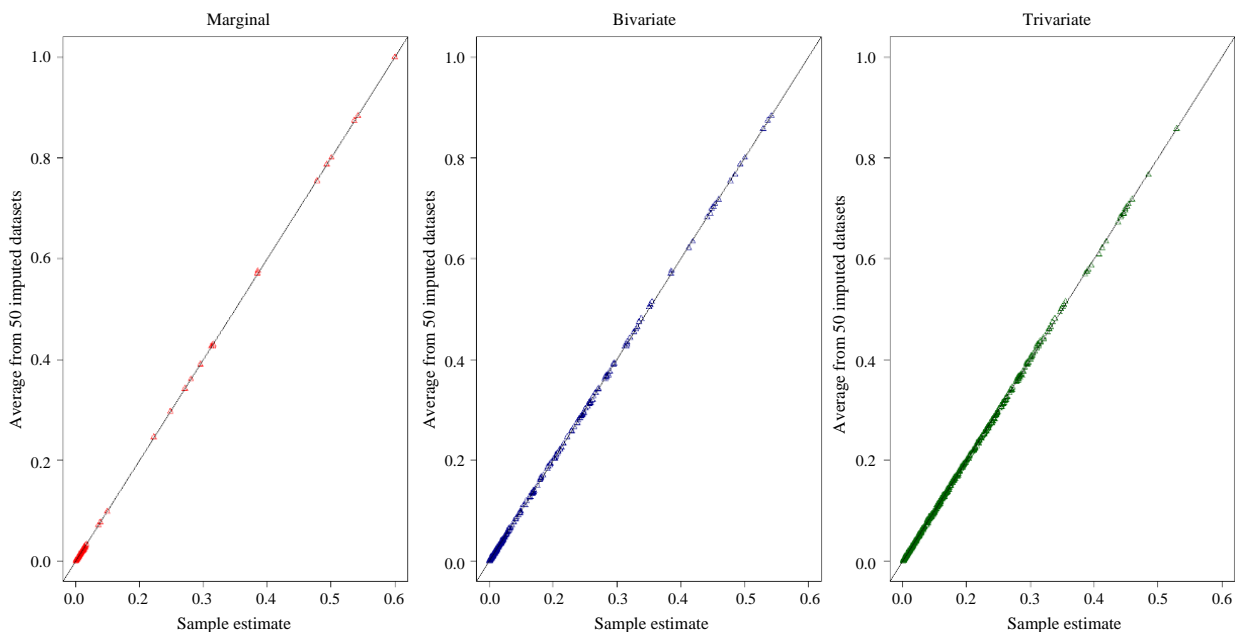

Figure A.2 Marginal, bivariate and trivariate probabilities computed in the sample and imputed datasets under MCAR from the truncated NDPMPM using the cap-and-weight approach. Household heads' data values to the household level.

Table A.4

Confidence intervals for selected probabilities that depend on within-household relationships in the original and imputed datasets under MCAR. "No missing" is based on the sampled data before introducing missing values, "NDPMPM" uses the truncated NDPMPM, moving household heads' data values to the household level, and "NDPMPM Capped" uses the truncated NDPMPM with the cap-and-weight approach and moving household heads' data values to the household level. "HH " means household head, "SP" means spouse, "CH" means child, and "CP" means couple. Q is the value in the full population of 764,580 households

		Q	No Missing	NDPMPM	NDPMPM Capped
All same race household:	$n_i = 2$	0.942	(0.932, 0.949)	(0.924, 0.944)	(0.925, 0.946)
	$n_i = 3$	0.908	(0.907, 0.937)	(0.887, 0.924)	(0.890, 0.925)
	$n_i = 4$	0.901	(0.879, 0.917)	(0.854, 0.900)	(0.855, 0.900)
SP present		0.696	(0.682, 0.707)	(0.683, 0.709)	(0.683, 0.709)
Same race CP		0.656	(0.641, 0.668)	(0.637, 0.664)	(0.638, 0.665)
SP present, HH is White		0.600	(0.589, 0.616)	(0.590, 0.618)	(0.590, 0.618)
White CP		0.580	(0.569, 0.596)	(0.568, 0.596)	(0.568, 0.597)
CP with age difference less than five		0.488	(0.465, 0.492)	(0.422, 0.451)	(0.422, 0.450)
Male HH, home owner		0.476	(0.456, 0.484)	(0.455, 0.483)	(0.456, 0.485)
HH over 35, no CH present		0.462	(0.441, 0.468)	(0.438, 0.466)	(0.438, 0.466)
At least one biological CH present		0.437	(0.431, 0.458)	(0.432, 0.460)	(0.432, 0.460)
HH older than SP, White HH		0.322	(0.309, 0.335)	(0.308, 0.335)	(0.306, 0.333)
Adult female w/ at least one CH under 5		0.078	(0.070, 0.085)	(0.068, 0.084)	(0.067, 0.083)
White HH with Hisp origin		0.066	(0.064, 0.078)	(0.064, 0.079)	(0.064, 0.079)
Non-White CP, home owner		0.058	(0.050, 0.063)	(0.048, 0.061)	(0.048, 0.061)
Two generations present, Black HH		0.057	(0.053, 0.066)	(0.053, 0.066)	(0.053, 0.067)
Black HH, home owner		0.052	(0.046, 0.058)	(0.046, 0.059)	(0.046, 0.059)
SP present, HH is Black		0.039	(0.032, 0.042)	(0.032, 0.043)	(0.032, 0.042)
White-nonwhite CP		0.034	(0.029, 0.039)	(0.032, 0.044)	(0.032, 0.044)
Hisp HH over 50, home owner		0.029	(0.025, 0.034)	(0.025, 0.035)	(0.025, 0.035)
One grandchild present		0.028	(0.023, 0.033)	(0.024, 0.034)	(0.024, 0.034)
Adult Black female w/ at least one CH under 18		0.027	(0.028, 0.038)	(0.027, 0.037)	(0.027, 0.037)
At least two generations present, Hisp CP		0.027	(0.022, 0.031)	(0.022, 0.031)	(0.022, 0.031)
Hisp CP with at least one biological CH		0.025	(0.020, 0.028)	(0.019, 0.028)	(0.019, 0.028)
At least three generations present		0.023	(0.020, 0.028)	(0.019, 0.028)	(0.019, 0.028)
Only one parent		0.020	(0.016, 0.024)	(0.016, 0.024)	(0.016, 0.024)
At least one stepchild		0.019	(0.018, 0.026)	(0.018, 0.027)	(0.018, 0.027)
Adult Hisp male w/ at least one CH under 10		0.018	(0.017, 0.025)	(0.016, 0.025)	(0.016, 0.025)
At least one adopted CH, White CP		0.008	(0.005, 0.010)	(0.005, 0.010)	(0.005, 0.010)
Black CP with at least two biological children		0.006	(0.003, 0.007)	(0.003, 0.007)	(0.003, 0.007)
Black HH under 40, home owner		0.005	(0.005, 0.009)	(0.005, 0.010)	(0.005, 0.011)
Three generations present, White CP		0.005	(0.004, 0.008)	(0.004, 0.010)	(0.004, 0.009)
White HH under 25, home owner		0.003	(0.002, 0.005)	(0.004, 0.009)	(0.004, 0.009)

References

Andridge, R.R., and Little, R.J.A. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, 78(1), 40-64.

Bennink, M., Croon, M.A., Kroon, B. and Vermunt, J.K. (2016). Micro-macro multilevel latent class models with multiple discrete individual-level variables. *Advances in Data Analysis and Classification*.

Chambers, R., and Skinner, C. (2003). Analysis of Survey Data, Wiley Series in Survey Methodology, Wiley.

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

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.

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

Kalton, G., and Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1, 1-16. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1986001/article/14404-eng.pdf.

Little, R.J.A. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, 9, 407-426.

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.

Murray, J.S., and Reiter, J.P. (2016). Multiple imputation of missing categorical and continuous values via Bayesian mixture models with local dependence (forthcoming). *Journal of the American Statistical Association*.

Raghunathan, T.E., and Rubin, D.B. (2001). Multiple imputation for statistical disclosure limitation. Technical Report.

Reiter, J.P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29, 2, 181-188. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2003002/article/6785-eng.pdf.

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

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, Inc.

Rubin, D.B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, 9, 462-468.

Savitsky, T.D., and Toth, D. (2016). Bayesian estimation under informative sampling. *Electronic Journal of Statistics*, 10.1, 1677-1708.

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

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*, 38.5, 199-521.

Vermunt, J.K. (2003). Multilevel latent class models. *Sociological Methodology*, 213-239.

Vermunt, J.K. (2008). Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research*, 33-51.

Walker, S.G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics - Simulation and Computation*, 1, 45-54.

Wang, Q., Akande, O., Hu, J., Reiter, J. and Barrientos, A. (2016). NestedCategBayesImpute: Modeling and Generating Synthetic Versions of Nested Categorical Data in the Presence of Impossible Combinations. *The Comprehensive R Archive Network*.