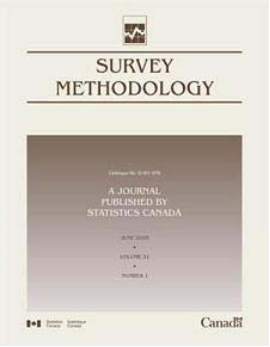
Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Bayesian multiple imputation for large-scale categorical data with structural zeros



by Daniel Manrique-Vallier and Jerome P. Reiter

June 2014

Statistics Canada Statistique Canada

Canada

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 infostats@statcan.gc.ca,

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-877-287-4369

Depository Services Program

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

To access this product

This product, Catalogue no. 12-001-X, is available free in electronic format. To obtain a single issue, visit our website, www.statcan.gc.ca, and browse by “Key resource” > “Publications.”

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 “About us” > “The agency” > “Providing services to Canadians.”

Published by authority of the Minister responsible for Statistics Canada

Cette publication est aussi disponible en français.

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.

Standard symbols

The following symbols are used in Statistics Canada publications:

.	not available for any reference period
..	not available for a specific reference period
...	not applicable
0	true zero or a value rounded to zero
0^s	value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p}	preliminary
^{r}	revised
x	suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E}	use with caution
F	too unreliable to be published
*	significantly different from reference category ($p < 0.05$)

Survey Methodology, June 2014
Vol. 40, No. 1, pp. 125-134
Statistics Canada, Catalogue No. 12-001-X

Bayesian multiple imputation for large-scale categorical data with structural zeros

Daniel Manrique-Vallier and Jerome P. Reiter[1]

Abstract

We propose an approach for multiple imputation of items missing at random in large-scale surveys with exclusively categorical variables that have structural zeros. Our approach is to use mixtures of multinomial distributions as imputation engines, accounting for structural zeros by conceiving of the observed data as a truncated sample from a hypothetical population without structural zeros. This approach has several appealing features: imputations are generated from coherent, Bayesian joint models that automatically capture complex dependencies and readily scale to large numbers of variables. We outline a Gibbs sampling algorithm for implementing the approach, and we illustrate its potential with a repeated sampling study using public use census microdata from the state of New York, U.S.A.

Key Words: Latent class; Log-linear; Missing; Mixture; Multinomial; Nonresponse.

1 Introduction

Many agencies collect surveys comprising large numbers of exclusively categorical variables. Inevitably, these surveys suffer from item nonresponse that, when left unattended, can reduce precision or increase bias (Little and Rubin 2002). To handle item nonresponse, one approach is multiple imputation (Rubin 1987), in which the agency fills in the missing items by sampling repeatedly from predictive distributions. This creates $M > 1$ completed datasets that can be analyzed or 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. For reviews of multiple imputation, see Rubin (1996), Barnard and Meng (1999), Reiter and Raghunathan (2007), and Harel and Zhou (2007).

Multiple imputation typically is implemented via one of two strategies. The first is to posit a joint model for all variables and estimate the model using Bayesian techniques, usually involving data augmentation and Markov chain Monte Carlo (MCMC) sampling. Common joint models include the multivariate normal for continuous data and log-linear models for categorical data (Schafer 1997). The second strategy is to use approaches based on chained equations (Van Buuren and Oudshoorn 1999; Raghunathan, Lepkowski, van Hoewyk and Solenberger 2001; White, Royston and Wood 2011). The analyst estimates a series of univariate conditional models and imputes missing values sequentially with these models. Typical conditional models include normal regressions for continuous dependent variables and logistic or multinomial logistic regressions for categorical dependent variables.

As noted by Vermunt, Ginkel, der Ark and Sijtsma (2008) and Si and Reiter (2013), chained equation strategies are not well-suited for large categorical datasets with complex dependencies. For any conditional (multinomial) logistic regression, the number of possible models is enormous once one considers potential interaction effects. Carefully specifying each conditional model is a very

1. Daniel Manrique-Vallier is Assistant Professor at the Department of Statistics, Indiana University, Bloomington, IN 47408. E-mail: dmanriqu@indiana.edu; Jerome P. Reiter is Mrs. Alexander Hehmeyer Professor of Statistical Science, Duke University, Durham, NC 27708-0251. E-mail: jerry@stat.duke.edu.

time-consuming task with no guarantee of a theoretically coherent set of models; indeed, for this reason many practitioners of chained equations use default settings that include main effects only in the conditional models. By excluding interactions, analysts risk generating completed datasets that yield biased estimates. We note that similar model selection difficulties plague approaches based on log-linear models.

To avoid these issues, Si and Reiter (2013) propose a fully Bayesian, joint modeling approach to multiple imputation for high-dimensional categorical data based on latent class models. The idea is to model the implied contingency table of the categorical variables as a mixture of independent multinomial distributions, estimating the mixture distributions nonparametrically with Dirichlet process prior distributions. Mixtures of multinomials can describe arbitrarily complex dependencies and are computationally expedient, so that they are effective general purpose multiple imputation engines. For example, Si and Reiter (2013) applied their models to impute missing values in 80 categorical variables in the Trends in International Mathematics and Science Study.

The approach of Si and Reiter (2013) does not deal with an important and prevalent complication in survey data: certain combinations of variables may not be possible *a priori*. These are called structural zeros (Bishop, Fienberg and Holland 1975). For example, in the United States it is impossible for children under age 15 to be married. Structural zeros also can arise from skip patterns in surveys. The imputation algorithms of Si and Reiter (2013), if applied directly, allow non-zero probability for structural zeros, which in turn biases estimates of probabilities for feasible combinations.

In this article, we present a fully Bayesian, joint modeling approach to multiple imputation of large categorical datasets with structural zeros. Our approach blends the latent class imputation model of Si and Reiter (2013) with the approach to handling structural zeros developed by Manrique-Vallier and Reiter (forthcoming 2014). Using simulations, we show that the approach generates multiply-imputed datasets that do not violate structural zero conditions and can have well-calibrated repeated sampling properties.

2 Bayesian latent class imputation model with structural zeros

Suppose that we have a sample of n individuals measured on J categorical variables. Each individual has an associated response vector $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$, whose components take values from a set of L_j levels. For convenience, we label these levels using consecutive numbers, $x_{ij} \in \{1, \ldots, L_j\}$, so that $\mathbf{x}_i \in \mathcal{C} = \prod_{j=1}^{J} \{1, \ldots, L_j\}$. Note that $\mathcal{C}$ includes all combinations of the J variables, including structural zeros, and that each combination $\mathbf{x}$ can be viewed as a cell in the contingency table formed by $\mathcal{C}$. Let $\mathbf{x}_i = (\mathbf{x}_i^{\text{obs}}, \mathbf{x}_i^{\text{mis}})$, where $\mathbf{x}_i^{\text{obs}}$ includes the variables with observed values and $\mathbf{x}_i^{\text{mis}}$ includes the variables with missing values. Finally, let $S = \{s_1, \ldots, s_C\}$, where $s_c \in \mathcal{C}$ and $c = 1, \ldots, C < |S|$, be the set of structural zero cells, *i.e.*, $\Pr(\mathbf{x}_i \in S) = 0$.

2.1 Latent class models

As an initial step, we describe the Bayesian latent class model without any concerns for structural zeros and without any missing data, *i.e.*, $\mathbf{x}_i = \mathbf{x}_i^{\text{obs}}$. This model is a finite mixture of product-multinomial distributions,

$$p(\mathbf{x}\,|\,\boldsymbol{\lambda},\boldsymbol{\pi}) = f^{\,\mathrm{LCM}}(\mathbf{x}\,|\,\boldsymbol{\lambda},\boldsymbol{\pi}) = \sum_{k=1}^{K}\pi_k\prod_{j=1}^{J}\lambda_{jk}\left[x_j\right], \tag{2.1}$$

where $\boldsymbol{\lambda} = \left(\lambda_{jk}[l]\right)$, with all $\lambda_{jk}[l] > 0$ and $\sum_{l=1}^{L_j}\lambda_{jk}[l] = 1$. Here, $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_K)$ with $\sum_{k=1}^{K}\pi_k = 1$. This model corresponds to the generative process,

$$x_{ij}\,|\,z_i \overset{\text{indep}}{\sim} \text{Discrete}_{1:L_j}\left(\lambda_{jz_i}[1],\ldots,\lambda_{jz_i}\left[L_j\right]\right) \text{ for all } i \text{ and } j \tag{2.2}$$

$$z_i\,|\,\boldsymbol{\pi} \overset{\text{iid}}{\sim} \text{Discrete}_{1:K}(\pi_1,\ldots,\pi_K) \text{ for all } i. \tag{2.3}$$

As notation, let $(\mathcal{X},\mathcal{Z})$ be a sample of n variates obtained from this process, with $\mathcal{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_n)$ and $\mathcal{Z} = (z_1,\ldots,z_n)$. For K large enough, (2.1) can represent arbitrary joint distributions for $\mathbf{x}$ (Suppes and Zanotti 1981; Dunson and Xing 2009). And, using the conditional independence representation in (2.2) and (2.3), the model can be estimated and simulated from efficiently even for large J.

For prior distributions on $\boldsymbol{\pi}$, we follow Si and Reiter (2013) and Manrique-Vallier and Reiter (forthcoming 2014). We have

$$\lambda_{jk}[\cdot] \overset{\text{indep}}{\sim} \text{Dirichlet}\left(\mathbf{1}_{L_j}\right) \tag{2.4}$$

$$\pi_k = V_k\prod_{h<k}(1-V_h) \tag{2.5}$$

$$V_k \overset{\text{iid}}{\sim} \text{Beta}(1,\alpha) \text{ for } k = 1,\ldots,K-1; V_K = 1 \tag{2.6}$$

$$\alpha \sim \text{Gamma}(0.25, 0.25) \tag{2.7}$$

The prior distributions in (2.4) are equivalent to uniform distributions over the support of the $J \times K$ multinomial conditional probabilities and hence represent vague prior knowledge. The prior distribution for $\boldsymbol{\pi}$ in (2.5)-(2.7) is an example of a finite-dimensional stick-breaking prior distribution (Sethuraman 1994; Ishwaran and James 2001). As discussed in Dunson and Xing (2009) and Si and Reiter (2013), it typically allocates $\mathcal{Z}$ to fewer than K classes, thereby reducing computation and avoiding over-fitting. For further discussion and justification of this model as an imputation engine, see Si and Reiter (2013).

2.2 Truncated latent class models

The latent class model in (2.1) does not naturally specify cells with structural zeros *a priori*, because it assumes a positive probability for each cell. Thus, to represent tables with structural zeros, we need to truncate the model so that

$$f^{\,\mathrm{TLCM}}(\mathbf{x}\,|\,\boldsymbol{\lambda},\boldsymbol{\pi},S) \propto 1\{\mathbf{x} \notin S\}\sum_{k=1}^{K}\pi_k\prod_{j=1}^{J}\lambda_{jk}\left[x_j\right]. \tag{2.8}$$

As Manrique-Vallier and Reiter (forthcoming 2014) show, obtaining samples from the posterior distribution of parameters $(\boldsymbol{\lambda}, \boldsymbol{\pi})$, conditional on a sample $\mathcal{X}^1 = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, can be greatly facilitated by adopting a sample augmentation strategy akin to those in Basu and Ebrahimi (2001) and O'Malley and Zaslavsky (2008). We consider $\mathcal{X}^1$ to be the portion of variates that did not fall into the set S from a larger sample, $\mathcal{X}$, generated directly from (2.1). Let n_0, $\mathcal{X}^0$, and $\mathcal{Z}^0$ be the the (unknown) sample size, response vectors, and latent class labels for the portion of $\mathcal{X}$ that did fall into S. Using a prior distribution from Meng and Zaslavsky (2002), Manrique-Vallier and Reiter (forthcoming 2014) show that if $p(N) \propto 1/N$, where $N = n_0 + n$, the posterior distribution of $(\boldsymbol{\lambda}, \boldsymbol{\pi})$ under the truncated model (2.8) can be obtained by integrating the posterior distribution under the augmented sample model over $\left(n_0, \mathcal{X}^0, \mathcal{Z}^0, \mathcal{Z}^1\right)$.

In doing so, Manrique-Vallier and Reiter (forthcoming 2014) develop a computationally efficient algorithm for dealing with large sets of structural zeros when they can be expressed as the union of sets defined by *margin conditions*. These are sets defined by fixing some levels of a subset of the categorical variables, for example, the set of all cells such that $\{\mathbf{x} \in \mathcal{C} : x_3 = 1, x_6 = 3\}$. Manrique-Vallier and Reiter (forthcoming 2014) introduce a vector notation to denote margin conditions, which we use here as well. Let $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_J)$ where, for $j = 1, \ldots, J$, we let $\mu_j = x_j$ whenever x_j is fixed at some level and $\mu_j = *$ otherwise, where $*$ is special notation for a placeholder. Using this notation and assuming $J = 8$, the conditions that define the example set above ($x_3 = 1$ and $x_6 = 3$) correspond to the vector $(*, *, 1, *, *, 3, *, *)$. To avoid cluttering the notation, we use the vectors $\boldsymbol{\mu}$ to represent both the margin conditions and the cells defined by those margin conditions, determined from context.

2.3 Estimation and multiple imputation

We now discuss how the model in Section 2.2 can be estimated, and subsequently converted into a multiple imputation engine, when some items are missing at random. The basic strategy is to use a Gibbs sampler. Given a completed dataset $\left(\mathbf{x}^{\text{obs}}, \mathbf{x}^{\text{mis}}\right)$, we take a draw of the parameters using the algorithm from Manrique-Vallier and Reiter (forthcoming 2014). Given a draw of the parameters, we take a draw of $\mathbf{x}^{\text{mis}}$ as described below.

Formally, the algorithm proceeds as follows. Suppose that the set of structural zeros can be defined as the union of C disjoint margin conditions, $S = \cup_{c=1}^{C} \boldsymbol{\mu}_c$, and that we use the priors for α, $\boldsymbol{\lambda}$ and $\boldsymbol{\pi}$ defined in Section 2.1. Given $\mathbf{x}_i = \left(\mathbf{x}_i^{\text{obs}}, \mathbf{x}_i^{\text{mis}}\right)$ for $i = 1, \ldots, n$, the algorithm of Manrique-Vallier and Reiter (forthcoming 2014) samples parameters as follows.

1. For $i = 1, \ldots, n$, sample $z_i^1 \sim \text{Discrete}_{1:K}(p_1, \ldots, p_k)$, with $p_k \propto \pi_k \prod_{j=1}^{J} \lambda_{jk\left[x_{ij}^1\right]}$.
2. For $j = 1, \ldots, J$ and $k = 1, \ldots, K$, sample $\boldsymbol{\lambda}_{jk[\cdot]} \sim \text{Dirichlet}\left(\xi_{jk1}, \ldots, \xi_{jkL_j}\right)$, with $\xi_{jkl} = 1 + \sum_{i=1}^{n} 1\{x_{ij}^1 = l, z_i^1 = k\} + \sum_{i=1}^{n_0} 1\{x_{ij}^0 = l, z_i^0 = k\}$.

3. For $k = 1,\ldots,K-1$ sample $V_k \sim \text{Beta}\left(1+\nu_k, a+\sum_{h=k+1}^{K}\nu_k\right)$ where $\nu_k = \sum_{i=1}^{n} 1\{z_i^1 = k\} + \sum_{i=1}^{n_0} 1\{z_i^0 = k\}$. Let $V_K = 1$ and make $\pi_k = V_k \prod_{h<k}(1-V_h)$ for all $k = 1,\ldots,K$.

4. For $c = 1,\ldots,C$, compute $\omega_c = \Pr(\mathbf{x} \in \boldsymbol{\mu}_c \mid \boldsymbol{\lambda}, \boldsymbol{\pi}) = \sum_{k=1}^{K}\pi_k \prod_{\mu_{cj}\neq *}\lambda_{jk}[\mu_{cj}]$.

5. Sample $(n_1,\ldots,n_C) \sim NM(n,\omega_1,\ldots,\omega_C)$, where NM is the negative multinomial distribution, and let $n_0 = \sum_{c=1}^{C} n_c$.

6. Let $\kappa \leftarrow 1$. Repeat the following for each $c = 1,\ldots,C$.
   (a) Compute the normalized vector $(p_1,\ldots,p_K)$, where $p_k \propto \pi_k \prod_{j:\mu_{cj}\neq *}\lambda_{jk}[\mu_{cj}]$.
   (b) Repeat the following three steps n_c times:
      i. Sample $z_\kappa^0 \sim \text{Discrete}(p_1,\ldots,p_k)$,
      ii. For $j = 1,\ldots,J$ sample

$$x_{\kappa j}^0 \sim \begin{cases} \text{Discrete}_{1:L_j}\left(\lambda_{jz_\kappa^0}[1],\ldots,\lambda_{jz_\kappa^0}[L_j]\right) & \text{if } \mu_{cj} = * \\ \delta_{\mu_{jc}} & \text{if } \mu_{cj} \neq * \end{cases}$$

      where $\delta_{\mu_{cj}}$ is a point mass distribution at μ_{cj},
      iii. Let $\kappa \leftarrow \kappa + 1$.

7. Sample $\alpha \sim \text{Gamma}(a-1+K, b-\log\pi_K)$.

Having sampled parameters, we now need to take a draw of $\mathbf{x}^{\text{mis}}$. For $i = 1,\ldots,n$, let $\mathbf{m}_i = (m_{i1},\ldots,m_{iJ})$ be a vector such that $m_{ij} = 1$ if component j in $\mathbf{x}_i$ is missing and $m_{ij} = 0$ otherwise. Assuming that data are missing at random, we need to sample only the components of each $\mathbf{x}_i$ for which $m_{ij} = 1$, conditional on the components for which $m_{ij} = 0$. Thus, we add an eighth step to the algorithm.

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

$$p\left(\mathbf{x}_i^{\text{mis}} \mid \ldots\right) \propto 1\{\mathbf{x}_i \notin S\} \prod_{j:m_{ij}=1}\lambda_{jz_i}[x_{ij}]. \tag{2.9}$$

In the absence of structural zeros, the x_{ij} to be imputed are conditionally independent given z_i, making the imputation task a routine multinomial sampling exercise (Si and Reiter 2013). However, the structural zeros in S induce dependency between the components. Thus, we cannot simply sample the components independently of one another. A naive approach is to use an acceptance-rejection scheme, sampling repeatedly from the proposal distribution $p\left(\mathbf{x}^{\text{mis}*}\right) = \prod_{j:m_{ij}=1}\lambda_{jz_i}[x_{ij}]$ until obtaining a variate such that $\mathbf{x}^{\text{mis}*} \notin S$. However, when the rejection region is large or has a high probability, this approach can be very inefficient.

Instead we suggest forming additional Gibbs sampling steps, computing the conditional distributions of all missing components so that they can be sampled individually. Let $\text{Rep}(\mathbf{x}_i, j, l)$ be the vector that results from replacing component j in $\mathbf{x}_i$ by an arbitrary value $l \in \{1, 2, \ldots, L_j\}$. The full conditional distribution of missing component j of $\mathbf{x}_i$ (when $m_{ij} = 1$) is $p(x_{ij} | \ldots) \propto 1\{\text{Rep}(\mathbf{x}_i, j, x_{ij}) \notin S\} \lambda_{jz_i}[x_{ij}]$. Thus, we replace step 8 in the algorithm with

8'. For each $(i, j) \in \{(i, j) : m_{ij} = 1\}$, sample $x_{ij} \sim \text{Discrete}_{1:L_j}(p_1, \ldots, p_{L_j})$, where $p_l \propto \lambda_{jz_i}[l] 1\{\text{Rep}(\mathbf{x}_i, j, l) \notin S\}$.

The definition of p_l implies trimming the support of the full conditional distribution of x_{ij} from $\{1, \ldots, L_j\}$ to only values that avoid $\mathbf{x}_i \in S$, given current values of $\{x_{ij'} : \text{all } j' \neq j\}$.

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

3 Simulation study

To illustrate empirically the performance of this imputation engine, we conducted a repeated sampling experiment using an extract of the 5% public use microdata sample from the 2000 U.S. census data for the state of New York (Ruggles, Alexander, Genadek, Goeken, Schroeder and Sobek 2010). The data include $H = 953{,}076$ individuals and ten categorical variables: ownership of dwelling (3 levels), mortgage status (4 levels), age (9 levels), sex (2 levels), marital status (6 levels), single race identification (5 levels), educational attainment (11 levels), employment status (4 levels), work disability status (3 levels), and veteran status (3 levels). These variables define a contingency table with 2,566,080 cells, of which 2,317,030 correspond to structural zeros.

We treat the H records as a population from which we take 500 independent samples of size $n =$ 1,000. For each sample, we impose missing data by randomly blanking 30% of the recorded item-level values of each variable. We then estimate the truncated latent class model of Section 2.3, using 10,000 MCMC iterates and discarding the first 5,000 as burn-in. From each remaining chain we create $M = 50$ completed datasets via a systematic sample of every 100 iterations. In all 500 simulation runs we use a maximum number of latent classes $K = 50$. The effective number of components, *i.e.*, those comprising at least one individual, are typically between 10 and 15 (depending on the particular sub-sample) and not larger than 26.

As estimands, we use all three-way probabilities with values exceeding 0.1 in the population (the $H =$ 953,076 individuals). This equates to 279 estimands. In each sample, we estimate 95% confidence intervals for each of the 279 probabilities using the multiple imputation combining rules of Rubin (1987). We also compute the corresponding intervals with the data before introducing missing values, which we call the complete data.

Figure 3.1 shows the percentages of the five hundred 95% confidence intervals that cover their population values. For the most part, the simulated coverage rates for multiple imputation are within Monte Carlo error of the nominal level. A few intervals based on multiple imputation have low coverage rates; in particular, three are below 85% while their counterparts with complete data are closer to the nominal level. However, as evident in Figure 3.2, the absolute magnitudes of the biases in the point estimates of these quantities tend to be modest. These encouraging results are in accord with the results in Si and Reiter (2013), whose simulations included up to 50 variables (without any structural zeros).

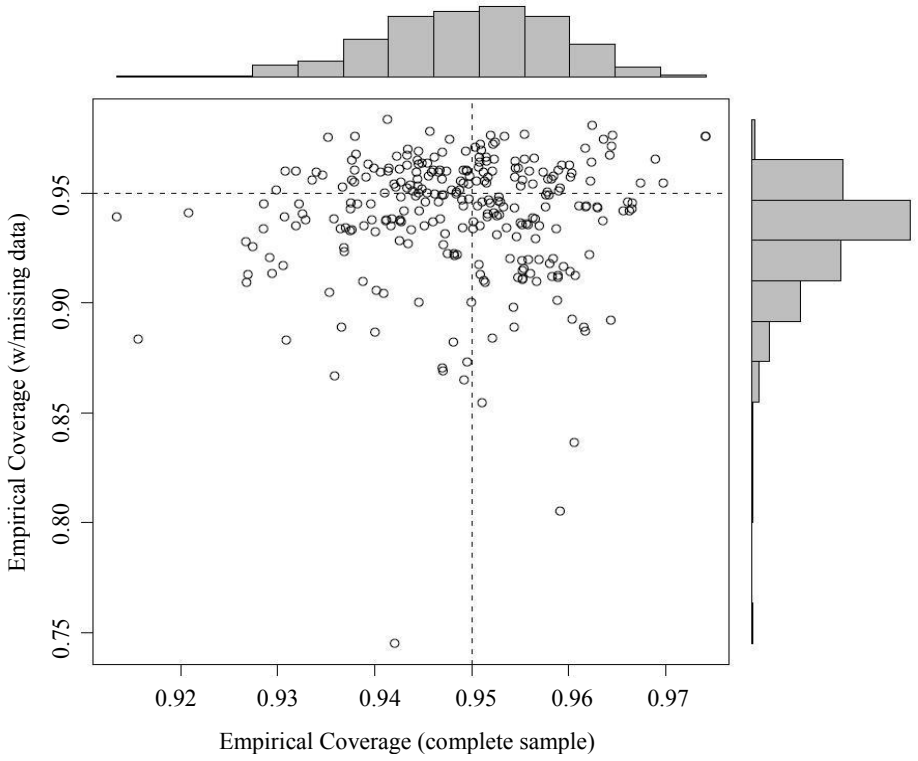


Figure 3.1 Comparison of empirical coverage rates (over 500 trials) of confidence intervals for three-way marginal probability estimates computed from the complete samples *vs.* multiply imputed datasets. Discontinuous lines indicate nominal coverage level. Random Unif(-0.004, 0.004) noise added for clarity.

For each estimand, we also compute the mean estimated fraction of missing information (FMI Rubin 1987, page 77) over the 500 trials. These are displayed in Figure 3.3. Most mean FMIs are close to the missing item rate of 30% that we imposed on every variable in the simulation design. However, many of the mean FMIs are significantly smaller than 30%, including four exactly equal to zero. The estimands with mean FMIs significantly below 0.30 correspond to entries of 3-way marginal probability tables where structural zeros severely restrict the possible imputations. In effect, the structural zeros reduce the information loss due to missingness. For example, the four estimands with mean FMI = 0 correspond to combinations of variables where restrictions leave only one possible imputation pattern to choose from; thus, no information is lost even though data values are actually missing. By incorporating the structural

zeros, we automatically impute such cases appropriately and can take advantage of the information supplied by the structural zero restrictions.

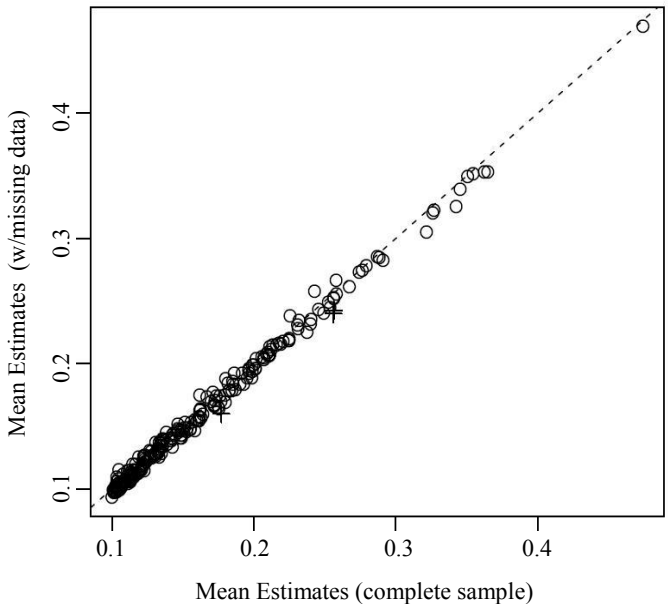


Figure 3.2 Mean (over 500 trials) three-way marginal probability estimates computed from the multiple imputed datasets *vs.* computed from the complete samples. Points marked with crosses are estimates for which the empirical coverage of the multiple-imputation based 95% confidence intervals fell below 85%.

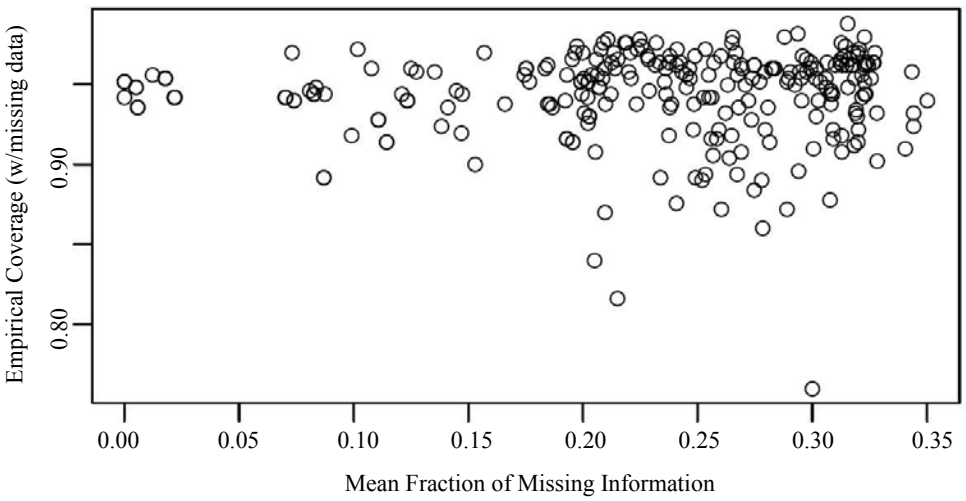


Figure 3.3 Empirical coverage rates (over 500 trials) of confidence intervals for 279 three-way marginal probability estimates computed from the multiply imputed datasets *vs.* their corresponding mean (over the 500 trials) estimated fraction of missing information.

4 Concluding remarks

Structural zero restrictions are an important feature of many surveys, *e.g.*, impossible combinations and skip patterns. They also play a key role in imputation. Ignoring structural zeros when estimating models can result in severe biases when estimating quantities that depend on joint or conditional probabilities. This translates to generating imputed values that do not accurately reflect the dependency structure in the data, and subsequently can lead to biased multiple imputation inferences. Additionally, structural zeros often function as consistency rules. Not enforcing them in imputation could result in completed datasets with inconsistent responses—like widowed toddlers or non-homeowners paying property taxes—that many agencies would be reluctant to release and many public users would find difficult to analyze. The approach suggested here based on Bayesian truncated latent class models offers survey researchers a way to avoid such problems, leading to multiple imputations from theoretically coherent and computationally expedient models that can capture complex dependencies, and simultaneously reducing the labor and guesswork in model specification that often accompanies traditional approaches to multiple imputation for categorical data. Computer code in C++ and R implementing the algorithms in this article can be obtained directly from the authors.

Acknowledgements

This research was supported by a grant from the National Science Foundation (SES 11-31897).

References

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

Basu, S., and Ebrahimi, N. (2001). Bayesian capture-recapture methods for error detection and estimation of population size: Heterogeneity and dependence. *Biometrika*, 88, 269-279.

Bishop, Y., Fienberg, S. and Holland, P. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press, reprinted in 2007, New York: Springer-Verlag.

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

Harel, O., and Zhou, X.H. (2007). Multiple imputation: review of theory, implementation and software. *Statistics in Medicine*, 26, 3057-3077.

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

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

Manrique-Vallier, D., and Reiter, J.P. (forthcoming 2014). Bayesian estimation of discrete multivariate truncated latent structure models. *Journal of Computational and Graphical Statistics*.

Meng, X.L., and Zaslavsky, A.M. (2002). Single observation unbiased priors. *The Annals of Statistics*, 30, 1345-1375.

O'Malley, A.J., and Zaslavsky, A.M. (2008). Domain-level covariance analysis for multilevel survey data with structured nonresponse. *Journal of the American Statistical Association*, 103, 1405-1418.

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. *Survey Methodology*, 27, 1, 85-95.

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

Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473-489.

Ruggles, S., Alexander, T., Genadek, K., Goeken, R., Schroeder, M.B. and Sobek, M. (2010). Integrated Public Use Microdata Series: Version 5.0 [Machine-readable database]. University of Minnesota, Minneapolis. http://usa.ipums.org.

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., and Reiter, J.P. (2013). Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *Journal of Educational and Behavioral Statistics*, forthcoming.

Suppes, P., and Zanotti, M. (1981). When are probabilistic explanations possible? *Synthese*, 48, 191-199.

Van Buuren, S., and Oudshoorn, C. (1999). Flexible multivariate imputation by MICE. Technical report, 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.

White, I.R., Royston, P. and Wood, A.M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30, 377-399.