Bayesian Models for Imputing Missing Data and Editing Erroneous Responses in Surveys

by

Olanrewaju Michael Akande

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Jerome P. Reiter, Supervisor

_____
Fan Li

_____
Alexander Volfovsky

_____
D. Sunshine Hillygus

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2019

Abstract

Bayesian Models for Imputing Missing Data and Editing Erroneous Responses in Surveys

by

Olanrewaju Michael Akande

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Jerome P. Reiter, Supervisor

_____
Fan Li

_____
Alexander Volfovsky

_____
D. Sunshine Hillygus

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University
2019

Abstract

This thesis develops Bayesian methods for handling unit nonresponse, item nonresponse, and erroneous responses in large scale surveys and censuses containing categorical data. I focus on applications to nested household data where individuals are nested within households and certain combinations of the variables are not allowed, such as the U.S. Decennial Census, as well as surveys subject to both unit and item nonresponse, such as the Current Population Survey.

The first contribution is a Bayesian model for imputing plausible values for item nonresponse in data nested within households, in the presence of impossible combinations. The imputation is done using a nested data Dirichlet process mixture of products of multinomial distributions model, truncated so that impossible household configurations have zero probability in the model. I show how to generate imputations from the Markov Chain Monte Carlo sampler, and describe strategies for improving the computational efficiency of the model estimation. I illustrate the performance of the approach with data that mimic the variables collected in the U.S. Decennial Census. The results indicate that my approach can generate high quality imputations in such nested data.

The second contribution extends the imputation engine in the first contribution to allow for the editing and imputation of household data containing faulty values. The approach relies on a Bayesian hierarchical model that uses the nested data Dirichlet process mixture of products of multinomial distributions as a model for the true

unobserved data, but also includes a model for the location of errors, and a reporting model for the observed responses in error. I illustrate the performance of the edit and imputation engine using data from the 2012 American Community Survey. I show that my approach can simultaneously estimate multivariate relationships in the data accurately, adjust for measurement errors, and respect impossible combinations in estimation and imputation.

The third contribution is a framework for using auxiliary information to specify nonignorable models that can handle both item and unit nonresponse simultaneously. My approach focuses on how to leverage auxiliary information from external data sources in nonresponse adjustments. This method is developed for specifying imputation models so that users can posit distinct specifications of missingness mechanisms for different blocks of variables, for example, a nonignorable model for variables with auxiliary marginal information and an ignorable model for the variables exclusive to the survey. I illustrate the framework using data on voter turnout in the Current Population Survey.

The final contribution extends the framework in the third contribution to complex surveys, specifically, handling nonresponse in complex surveys, such that we can still leverage auxiliary data while respecting the survey design through survey weights. Using several simulations, I illustrate the performance of my approach when the sample is generated primarily through stratified sampling.

Dedicated to my wonderful wife, Oluwatosin, my ever supportive family, and to God Almighty, by whose grace I was able to complete this work.

Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

Abbreviations

ACS	American Community Survey.
CPS	Current Population Survey.
MAAR	Missing Always at Random.
MAR	Missing at Random.
MACAR	Missing Always Completely At Random.
MCAR	Missing Completely At Random.
MCMC	Markov Chain Monte Carlo.
MI	Multiple Imputation.
MNAAR	Missing Not Always at Random.
MNAR	Missing Not at Random.
NDPMPM	Nested Data Dirichlet Process Mixture of Products of Multinomial Distributions.

Acknowledgements

I would like to thank everyone in the Department of Statistical Science for a wonderful experience. Special thanks to my advisor, Jerry Reiter, for being such an incredible mentor and for his guidance through graduate school. Thanks to Quanli Wang for his assistance throughout the research, and Gabriel Madson and D. Sunshine Hillygus for their inputs in our collaborative work. To my colleagues who helped in one way or the other, thank you so much. My sincere thanks to my parents and siblings for their support and encouragement. Special thanks to my wife, Oluwatosin, for always being there and supporting me; you are the best. Finally, I thank God for His grace and favor through which I was able to successfully complete this dissertation.

This research was supported by grants from the National Science Foundation (SES-1131897, SES-1733835) and the Alfred P. Sloan Foundation (G-2015-20166003).

1

Introduction

Large surveys and censuses often suffer from some form of nonresponse. This usually occurs in the form of unit nonresponse (for example, when a sampled individual does not participate in the survey), or item nonresponse (for example, when a responding sampled individual responds to some but not all of the survey questions). It is well known that statistical analyses based on only the observed data – that is, ignoring the missing data – can be problematic (Little and Rubin, 2002). Such analyses are often either inefficient, as they sacrifice information from individuals who partially responded, or biased, when there are systematic differences between the observed data and the missing data.

Handling nonresponse in multivariate categorical data nested within households, for example, individuals grouped within houses, can be particularly challenging. In such household data, imputations must preserve complex relationships within and across households, while also respecting certain structural zeros or edit constraints; for example, within any household, a child should not be older than his/her biological parent. In addition to missing values in household data, the observed data also can contain reported values that fail the edit constraints. Since recontacting survey

participants can be very expensive, data collecting agencies often supplement re-contact operations with variants of single error localization (Fellegi and Holt, 1976; Winkler, 1995; Winkler and Petkunas, 1997) and single imputation processes (de Waal and Coutinho, 2005; de Waal et al., 2011), which can severely underestimate uncertainty (Kim et al., 2015a; Manrique-Vallier and Reiter, 2018).

Another complication with missing data arises when unit and item nonresponse follow different missing data mechanisms, for example, an ignorable unit nonresponse mechanism but a nonignorable item nonresponse mechanism. In applied settings, the observed data alone does not usually contain enough information to fully distinguish between the two forms of nonresponse and identify the parameters necessary to impute missing values correctly. This is even more severe in complex surveys where imputations must respect the survey design. Here, it is possible to use information on marginal distributions from auxiliary sources to improve the imputations.

In this thesis, I present novel approaches to imputing missing and faulty data in the settings described above. In Chapter 2, I develop a Bayesian imputation engine for handling missing values in multivariate categorical data nested within households, in the presence of impossible combinations. The imputation engine relies on a nested data Dirichlet process mixture of products of multinomial distributions 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. I present a Gibbs sampler for estimating the model and generating imputations. I also describe strategies for improving the computational efficiency of the model estimation. I illustrate the performance of the approach with data that mimic the variables collected in the U.S. Decennial Census.

In Chapter 3, I extend the imputation engine in Chapter 2, to allow for editing and imputation of household data containing faulty values, based on a Bayesian

hierarchical model that includes (i) the nested data Dirichlet process mixture of products of multinomial distributions as the model for the true latent values of the data, (ii) a model for the location of errors, and (iii) a reporting model for the observed responses in error. This imputation engine propagates uncertainty due to unknown locations of errors and missing values, and generates plausible datasets that satisfy all edit constraints. I illustrate the approach using data from the 2012 American Community Survey.

In Chapter 4, I develop a framework for specifying imputation models for unit and item nonresponse so that (i) the models reflect realistic assumptions about the missing data mechanisms, (ii) the models take full advantage of auxiliary marginal information, and (iii) all parameters in the models can be uniquely estimated. I do so by focusing on how to leverage information from auxiliary data sources, such as administrative records and databases gathered by private-sector data aggregators, in nonresponse adjustments. The methodology is illustrated on an application examining voter turnout among subgroups of the population in the Current Population Survey (CPS), with data from government election statistics utilized as population-based auxiliary data. The information in the auxiliary margins is used to adjust the CPS data for nonresponse with a more reasonable set of assumption than previous analyses of voter turnout based on the CPS.

In Chapter 5, I present an approach for incorporating survey weights in the framework in Chapter 4. My approach ensures imputations for nonresponse respect the complex survey design when leveraging auxiliary marginal information. I use several simulation scenarios to illustrate the performance of my approach and show that it generates good design-based estimates of the joint relationship between the variables in a survey, under the settings of my simulations.

In the remainder of this chapter, I review some concepts that I use and refer to throughout this thesis.

1.1 Missing Data Mechanisms

Nonresponse or missing data can follow different mechanisms, all of which fall into one of two classes: (i) ignorable missingness mechanism (Rubin, 1976), where the missing data mechanism is unrelated to missing values, and (ii) non-ignorable missingness mechanism (Rubin, 1976), where the missing data mechanism could potentially be related to missing values. The exact missing data mechanisms are typically unknown, especially in large datasets. Thus, plausible assumptions about the mechanisms have to be made accordingly.

In this section, we introduce some notation to review the different missing data mechanisms. This review closely follows Mealli and Rubin (2015), who define and clarify these different missingness mechanisms.

Let $Y = (Y_1, \ldots, Y_n)$ represent notation for a single random variable in a collected dataset containing n units, where each Y_i is a scalar quantity, with $i = 1, \ldots, n$. Let $R_i = 1$ when Y_i is missing, and $R_i = 0$ otherwise. Let $R = (R_1, \ldots, R_n)$, that is, R is the vector of missing indicators for Y. Let $y = (y_1, \ldots, y_n)$ and $r = (r_1, \ldots, r_n)$ be realizations of Y and R, respectively. The missing data mechanism can be represented as $\Pr(R = r|Y = y, \phi)$, the conditional distribution of R given Y and a parameter ϕ.

The vector Y can be partitioned into the ordered subvectors $Y_{(1)} = (Y_i : r_i = 1)$ and $Y_{(0)} = (Y_i : r_i = 0)$. Similarly, y also can be partitioned into $y_{(1)} = (y_i : r_i = 1)$, the subvector of missing values, and $y_{(0)} = (y_i : r_i = 0)$, the subvector of observed values. Finally, let $\tilde{r}$ be a particular sample realization of R and $\tilde{y}_{(0)}$ be a particular sample realization of $\tilde{y}_{(0)}$. In this thesis, we sometimes use the superscript "obs" interchangeably with the subscript (0), and the superscript "miss" interchangeably with the subscript (1).

In this section, we present definitions for the most common missingness mecha-

nisms. Other missing data mechanisms, such as the sequentially additive nonignorable models of Sadinle and Reiter (2019) and the itemwise conditionally independent nonresponse models of Sadinle and Reiter (2017), also exists within the missing data literature. We do not cover the definitions for those mechanisms here, and instead refer readers to Sadinle and Reiter (2017), Linero and Daniels (2018) and Sadinle and Reiter (2019).

1.1.1 Missing completely at random and missing always completely at random

Data are missing completely at random (MCAR) when

$$\Pr(R = \tilde{r}|Y = y, \phi) = \Pr(R = \tilde{r}|\phi) \quad \forall\ y \text{ and } \phi.$$

Data are missing always completely at random (MACAR) when

$$\Pr(R = r|Y = y, \phi) = \Pr(R = r|\phi) \quad \forall\ r, y \text{ and } \phi.$$

MCAR and MACAR (both ignorable missingness mechanisms) imply that the reason for missingness in a variable is unrelated to any realization of that variable (or to the realizations of other measured variables). The missing data mechanism is fully characterized by the missingness parameter ϕ.

1.1.2 Missing at random and missing always at random

Analogously, data are missing at random (MAR) when

$$\Pr(R = \tilde{r}|Y_{(0)} = \tilde{y}_{(0)}, Y_{(1)} = y_{(1)}, \phi) = \Pr(R = \tilde{r}|Y_{(0)} = \tilde{y}_{(0)}, Y_{(1)} = y'_{(1)}, \phi),$$

for all $y_{(1)}$, $y'_{(1)}$ and ϕ. Data are missing always at random (MAAR) when

$$\Pr(R = r|Y_{(0)} = y_{(0)}, Y_{(1)} = y_{(1)}, \phi) = \Pr(R = r|Y_{(0)} = y_{(0)}, Y_{(1)} = y'_{(1)}, \phi),$$

for all $y_{(0)}$, $y_{(1)}$, $y'_{(1)}$, r and ϕ. MAR and MAAR (both ignorable missingness mechanisms) imply that the reason for missingness is unrelated to the missing values of

that variable but could depend on the observed values (and possibly on realizations of other measured variables). MAR (or MAAR) is a stronger assumption than MCAR (or MACAR).

1.1.3 Missing not at random and missing not always at random

Finally, data are missing not at random (MNAR) when

$$\Pr(R = \tilde{r}|Y_{(0)} = \tilde{y}_{(0)}, Y_{(1)} = y_{(1)}, \phi) \neq \Pr(R = \tilde{r}|Y_{(0)} = \tilde{y}_{(0)}, Y_{(1)} = y'_{(1)}, \phi),$$

for some ϕ, and some $y_{(1)} \neq y'_{(1)}$. Data are missing not always at random (MNAAR) when

$$\Pr(R = r|Y_{(0)} = y_{(0)}, Y_{(1)} = y_{(1)}, \phi) \neq \Pr(R = r|Y_{(0)} = y_{(0)}, Y_{(1)} = y'_{(1)}, \phi),$$

for some $y_{(0)}$, r, ϕ, and some $y_{(1)} \neq y'_{(1)}$. MNAR and MNAAR (both nonignorable missingness mechanisms) imply that the reason for missingness could be systematically related to the actual missing values in addition to the observed values (and realizations of other measured variables). MNAR (or MNAAR) is an stronger assumption than MAR (or MAAR).

MACAR, MAAR and MNAAR all define the missing data mechanisms in terms of all possible values of R and $Y_{(0)}$, whereas MCAR, MAR and MNAR characterize the missing data mechanisms in terms of only the observed values $\tilde{r}$ and $\tilde{y}_{(0)}$, that is, only on the observed data/sample. Intuitively, the difference between these two groups is whether or not analysts prefer to assume a missingness mechanism for all possible realizations of the missing data pattern (which includes the observed realization), versus assuming a weaker version of the mechanisms only for the specific observed realization.

1.2 Multiple Imputation

This review closely follows Akande et al. (2017), who compare different multiple imputation methods for categorical data.

Multiple imputation (MI) is a common approach for handling missing data. In MI, analysts create $L > 1$ completed datasets, $\mathbf{Z} = (\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(L)})$, by replacing missing values in the data with draws from the (posterior) predictive distributions of models estimated based on the observed data. Analysts then can compute sample estimates for estimands of interest in each completed dataset, and combine them using MI inferences developed by Rubin (1987).

When implementing MI, most analysts do so either through joint modeling (JM) or fully conditional specification (FCS) (van Buuren, 2007). When implementing JM, analysts specify a joint distribution for all variables in the data. Imputations are then sampled from the implied predictive distributions of the variables with missing data, given all other variables. The JM strategy is appealing because it aligns with the theory in Rubin (1987). On the other hand, when implementing FCS, analysts directly specify and sample from univariate distributions for each variable conditional on all other variables, without first forming a proper joint distribution. FCS is appealing because of its simplicity and flexibility — one can tailor the predictive models for individual variables. However, the specified univariate conditional distributions, in some implementations of the FCS strategy, may be potentially incompatible (Arnold and Press, 1989; Gelman and Speed, 1993); that is, the set of the conditional distributions may not correspond to any joint distribution. The methodology developed in this thesis uses the JM strategy.

To review the combining rules for MI, let q be the completed-data point estimator of some estimand Q, and let u be the estimator of the 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 $\mathbf{Z}^{(l)}$.

The MI point estimate of Q is

$$\bar{q}_L = \sum_{l=1}^{L} \frac{q^{(l)}}{L},$$

and the corresponding MI estimate of the variance of $\bar{q}_L$ is given by

$$T_L = \left(1 + \frac{1}{L}\right) b_L + \bar{u}_L,$$

where

$$b_L = \sum_{l=1}^{L} \frac{(q^{(l)} - \bar{q}_L)^2}{L - 1},$$

and

$$\bar{u}_L = \sum_{l=1}^{L} \frac{u^{(l)}}{L}.$$

Inferences about Q can be made using

$$(\bar{q}_L - Q) \sim t_v(0, T_L),$$

where t_v is a t-distribution with

$$v = (L - 1) \left[\frac{1 + \bar{u}_L}{\left(1 + \frac{1}{L}\right) b_L} \right]^2$$

degrees of freedom.

For in-depth reviews of MI as well as comparisons of MI methods, see Rubin (1996), Schafer (1997), Barnard and Meng (1999), Harel and Zhou (2007), Reiter and Raghunathan (2007) and Akande et al. (2017).

1.3 Partially Synthetic Data

In this thesis, we also use models to generate partially synthetic data and use combining rules from Reiter (2003). The combining rules for synthetic data are similar to the MI combining rules. Let q be the point estimator of some estimand Q, and let u be the estimator of the 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 $\mathbf{Z}^{(l)}$. The quantities needed for inferences are

$$\bar{q}_L = \sum_{l=1}^{L} \frac{q_l}{L}$$

$$b_L = \sum_{l=1}^{L} \frac{(q_l - \bar{q}_L)^2}{L-1}$$

$$\bar{u}_L = \sum_{l=1}^{L} \frac{u_l}{L}.$$

The analyst uses $\bar{q}_L$ to estimate Q and

$$T_L = \bar{u}_L + \frac{b_L}{L}$$

to estimate its variance. Inferences about Q are based on the t-distribution

$$(\bar{q}_L - Q) \sim t_v(0, T_L)$$

with

$$v = (L-1)\left[1 + \frac{L\bar{u}_L}{b_L}\right]^2$$

degrees of freedom.

For reviews of inference on partially synthetic data, see Reiter (2003, 2005).

2

Multiple Imputation of Missing Values in Household Data with Structural Zeros

This presentation in this chapter closely follows Akande et al. (2019), where the research first appeared.

2.1 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.

First, there is no consensus on how to apply hot deck or how to select the best metrics for matching donors to recipients in household data. Even if the analyst can come up with a reasonable metric for matching donors, applying hot decks in household data can become complicated when structural zeros exist since hot decks do not necessarily preserve them for the imputations. When it is important to preserve these structural zeros, the hot deck imputations need to be adjusted. Second, hot deck is often used as a single imputation method, so that the completed datasets are analyzed as if they had no missing values (Marker et al., 2002). Single imputation methods underestimate uncertainty since the extra variability due to nonresponse is often ignored (Rubin, 1987). Third, when the sample size is small, rare covariates' values for recipients can lead to sparseness of donors for them and good matches for such recipients are either hard to find or become overused; finding good matches is more likely in large samples. Although hot deck methods have been shown to be consistent when data are missing completely at random, useful conditions for consistency when data are missing at random seems lacking. See Andridge and Little (2010) for more discussions on hot deck methods and their drawbacks.

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 et al. (2018), which assumes that (i) each household is a member of a household-level latent class, and (ii) each individual is a 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 et al. (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 chapter. 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, (ii) using an approximation to the likelihood function, and (iii) using a hybrid method that randomly samples between

new imputations and a pre-generated set of possible completions. 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 chapter is organized as follows. In Section 2.2, we review the NDPMPM model in the presence of structural zeros and the MCMC sampler for fitting the model without missing data. In Section 2.3, we extend the MCMC sampler for the NDPMPM model to allow for missing data. In Section 2.4, we present the two strategies for speeding up the MCMC sampler. In Section 2.5, 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 2.6, we discuss findings, caveats and future work.

2.2 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).

2.2.1 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 $\mathbf{X}_i = (X_{i(p+1)}, \ldots, X_{i(p+q)}, X_{i11}, \ldots, X_{in_ip})$ 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 $\phi^{(k)}_{gmc} = \Pr(X_{ijk} = c|(G_i, M_{ij}) = (g, m))$ for the class pair (g, m), where $\phi^{(k)}_{gmc}$ 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 $\phi = \{\phi^{(k)}_{gmc} : c = 1, \ldots, d_k; k = 1, \ldots, p; m = 1, \ldots, S; g = 1, \ldots, F\}$.

For purposes of the Gibbs sampler in Section 2.2.2, 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, $\mathbf{X}^1_i$ 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 $(\mathbf{X}^1_1, \ldots, \mathbf{X}^1_n)$. The kernel of the NDPMPM, $\Pr(\mathcal{X}^1|\theta)$, is

$$\mathrm{L}(\mathcal{X}^1|\theta) = \prod_{i=1}^{n} \sum_{h \in \mathcal{H}} \mathbb{1}\{n_i = h\} \mathbb{1}\{\mathbf{X}^1_i \notin \mathcal{S}_h\} \left[\sum_{g=1}^{F} \pi_g \prod_{k=p+1}^{p+q} \lambda^{(k)}_{gX^1_{ik}} \prod_{j=1}^{h} \sum_{m=1}^{S} \omega_{gm} \prod_{k=1}^{p} \phi^{(k)}_{gmX^1_{ijk}} \right], \tag{2.1}$$

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

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

where $T(\mathcal{S})$ 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

$$\begin{aligned} X_{ik}|G_i, \lambda &\sim \text{Discrete}(\lambda^{(k)}_{G_i 1}, \ldots, \lambda^{(k)}_{G_i d_k}) \\ &\forall i = 1, \ldots, n \text{ and } k = p+1, \ldots, p+q \end{aligned} \tag{2.3}$$

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

$$\begin{aligned} G_i|\pi &\sim \text{Discrete}(\pi_1, \ldots, \pi_F) \\ &\forall i = 1, \ldots, n \end{aligned} \tag{2.5}$$

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

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),

$$\lambda^{(k)}_g = (\lambda^{(k)}_{g1}, \ldots, \lambda^{(k)}_{gd_k}) \sim \text{Dirichlet}(1, \ldots, 1) \tag{2.7}$$

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

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

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

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

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

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

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

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).

2.2.2 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 $G_i^0 \in \{1, \ldots, F\} \sim \text{Discrete}(\pi_1^{\star\star}, \ldots, \pi_F^{\star\star})$ where $\pi_g^{\star\star} \propto \lambda_{gh}^{(k)} \pi_g$ 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 $G_i \in \{1, \ldots, F\} \sim \text{Discrete}(\pi_1^\star, \ldots, \pi_F^\star)$ for $i = 1, \ldots, n$, where

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

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

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

$$\omega_{G_i^1 m}^\star = \Pr(M_{ij} = m|-) = \frac{\omega_{G_i^1 m} \prod_{k=1}^{p} \phi_{G_i^1 m X_{ijk}^1}^{(k)}}{\sum_{s=1}^{S} \omega_{G_i^1 s} \prod_{k=1}^{p} \phi_{G_i^1 s X_{ijk}^1}^{(k)}}$$

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

S3. Set $u_F = 1$. Sample

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

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

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

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

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

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

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

S5. Sample

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

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

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

S6. Sample

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

$$\text{where } \nu_{gmc}^{(k)} = \sum_{\substack{i,j| \\ M_{ij}^1=m}}^{\substack{n \\ G_i^1=g,}} \mathbb{1}(X_{ijk}^1 = c) + \sum_{\substack{i,j| \\ M_{ij}^0=m}}^{\substack{n_0 \\ G_i^0=g,}} \mathbb{1}(X_{ijk}^0 = c)$$

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

S7. Sample

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

S8. Sample

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

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

2.3 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 $\mathbf{a}_i = (a_{i(p+1)}, \ldots, a_{i(p+q)})$ 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 $\mathbf{b}_{ij} = (b_{ij1}, \ldots, b_{ijp})$ 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 $\mathbf{X}_i^1 = (\mathbf{X}_i^{\text{obs}}, \mathbf{X}_i^{\text{mis}})$, 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

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

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 $P^{\star}_{\mathbf{X}_i^{\text{mis}}} = \pi_{G_i^1} \prod_{k|a_{ik}=1}^{p+q} \lambda^{(k)}_{G_i^1 X_{ik}^1} (\prod_{j=1}^{n_i} \omega_{G_i^1 M_{ij}^1} \prod_{k|b_{ijk}=1}^{p} \phi^{(k)}_{G_i^1 M_{ij}^1 X_{ijk}^1})$, until we obtain a valid sample that satisfies $\mathbf{X}_i^1 \notin \mathcal{S}_h$. Notice that since $P^{\star}_{\mathbf{X}_i^{\text{mis}}}$ 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 $\mathbf{X}_i^{\text{mis}}$ 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_{ik}^1 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_{ijk}^1 using (2.4).

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

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

To initialize each $\mathbf{X}_i^{\text{mis}}$, 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 $\mathbf{X}_i^1 \notin \mathcal{S}_h$.

2.3.1 Proof that step S9′ 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. We prove that the rejection sampling scheme does result 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

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

where

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

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

1. $w(\mathbf{X}_i^{\text{mis}}) = p(\mathbf{X}_i^{\text{mis}})/g(\mathbf{X}_i^{\text{mis}}) = \mathbb{1}\{\mathbf{X}_i^1 \notin \mathcal{S}_h\}/\Pr(\mathbf{X}_i \notin \mathcal{S}_h|\theta) \leqslant 1/\Pr(\mathbf{X}_i \notin \mathcal{S}_h|\theta)$, and $0 < \Pr(\mathbf{X}_i \notin \mathcal{S}_h|\theta) < 1 \ \Rightarrow \ 1 < 1/\Pr(\mathbf{X}_i \notin \mathcal{S}_h|\theta) < \infty$ 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 $w(\mathbf{X}_i^{\text{mis}})/M = \mathbb{1}\{\mathbf{X}_i^1 \notin \mathcal{S}_h\}$.

2.4 Strategies for Speeding Up the MCMC Sampler

The rejection sampling step in the Gibbs sampler in Section 2.2.2 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.

2.4.1 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 $n = 1000$ households of size two, resulting in a total of $N = 2000$ 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 $(.86/.14) * 1000 \approx 6,143$ 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 2.2.2.

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 2.2.2; rather, it is a data reformatting technique that can be combined with the sampler.

2.4.2 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 2.2.2 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 2.4.1. 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 $3,072 = \lceil 6143 \times 0.5 \rceil$ impossible households in every iteration of the MCMC sampler. The probability of generating one impossible household is $3072/(1000 + 3072) = 0.75$, 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 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

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

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

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

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

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

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

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

S5*. Sample

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

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

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

S6*. Sample

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

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

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

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.

2.5 Empirical Study

To evaluate the performance of the NDPMPM as an imputation method, as well as the performance of the speed up strategies, we use data from the public use

Table 2.1: 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

microdata files from the 2011 and 2012 ACS, available for download from the United States Census Bureau (`http://www2.census.gov/acs2011_1yr/pums/` and `http://www2.census.gov/acs2012_1yr/pums/`).

2.5.1 Empirical study of the speedup approaches

To evaluate the performance of the two speedup approaches, we construct a population of $857,018$ households of sizes $\mathcal{H} = \{2, 3, 4, 5, 6\}$ from the 2011 ACS, and we sample $n = 10,000$ households comprising $N = 29,117$ individuals from the constructed population. We work with the variables described in Table 2.1. The structural zeros involve ages and relationships of individuals in the same house; see Appendix A for a full list of rules that we used. We evaluate the approaches using probabilities that depend on within household relationships and the household head.

We consider the NDPMPM using two approaches, both moving the values of the

household head to the household level as in Section 2.4.1 and also using the cap-and-weight approach in Section 2.4.2. 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, 1993b; Reiter, 2003), even though every released Z_{ijk} is a simulated value. We combine the estimates using the combining rules for synthetic data (Reiter, 2003) in Section 1.3.

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 2.2 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.

Table 2.2: 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 2.2.2, "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				
$n_i = 2$	(.939, .951)	(.918, .932)	(.912, .928)	(.910, .925)
$n_i = 3$	(.896, .920)	(.859, .888)	(.845, .875)	(.844, .874)
$n_i = 4$	(.885, .912)	(.826, .860)	(.813, .848)	(.817, .852)
$n_i = 5$	(.879, .922)	(.786, .841)	(.786, .841)	(.777, .834)
$n_i = 6$	(.831, .910)	(.701, .803)	(.718, .819)	(.660, .768)
SP present	(.693, .711)	(.678, .697)	(.676, .695)	(.677, .695)
SP with white HH	(.589, .608)	(.577, .597)	(.576, .595)	(.575, .595)
SP with black HH	(.036, .043)	(.035, .043)	(.034, .042)	(.034, .042)
White couple	(.570, .589)	(.560, .580)	(.553, .573)	(.552, .572)
White couple, own	(.495, .514)	(.468, .488)	(.461, .481)	(.463, .483)
Same race couple	(.655, .673)	(.636, .655)	(.626, .645)	(.625, .644)
White-nonwhite couple	(.028, .035)	(.028, .035)	(.034, .041)	(.036, .044)
Nonwhite couple, own	(.057, .067)	(.047, .056)	(.045, .053)	(.045, .054)
Only mother present	(.017, .022)	(.014, .019)	(.014, .019)	(.013, .018)
Only one parent present	(.021, .026)	(.026, .032)	(.026, .033)	(.027, .033)
Children present	(.507, .527)	(.493, .512)	(.517, .537)	(.511, .531)
Siblings present	(.022, .028)	(.027, .034)	(.027, .033)	(.027, .033)
Grandchild present	(.041, .049)	(.051, .060)	(.049, .058)	(.050, .059)
Three generations present	(.036, .044)	(.037, .045)	(.042, .050)	(.040, .048)
White HH, older than SP	(.309, .327)	(.283, .301)	(.294, .313)	(.302, .321)
Nonhisp HH	(.882, .894)	(.875, .888)	(.879, .891)	(.876, .889)
White, Hisp HH	(.071, .082)	(.074, .085)	(.072, .082)	(.073, .084)
Same age couple	(.087, .098)	(.027, .034)	(.023, .029)	(.024, .031)

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 2.3: Description of variables used in the missing data illustration. "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

2.5.2 Empirical study of missing data imputation under nonignorable missingness

To evaluate the performance of the NDPMPM as an imputation method, we construct a population of 764,580 households of sizes $\mathcal{H} = \{2, 3, 4\}$ again from the 2012 ACS, and we sample $n = 5,000$ households comprising $N = 13,181$ individuals from the constructed population. We work with the variables described in Table 2.3, which mimic those in the U. S. decennial census. The structural zeros remain the same as those in Appendix A. We move the household head to the household level as in Section 2.4.1 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

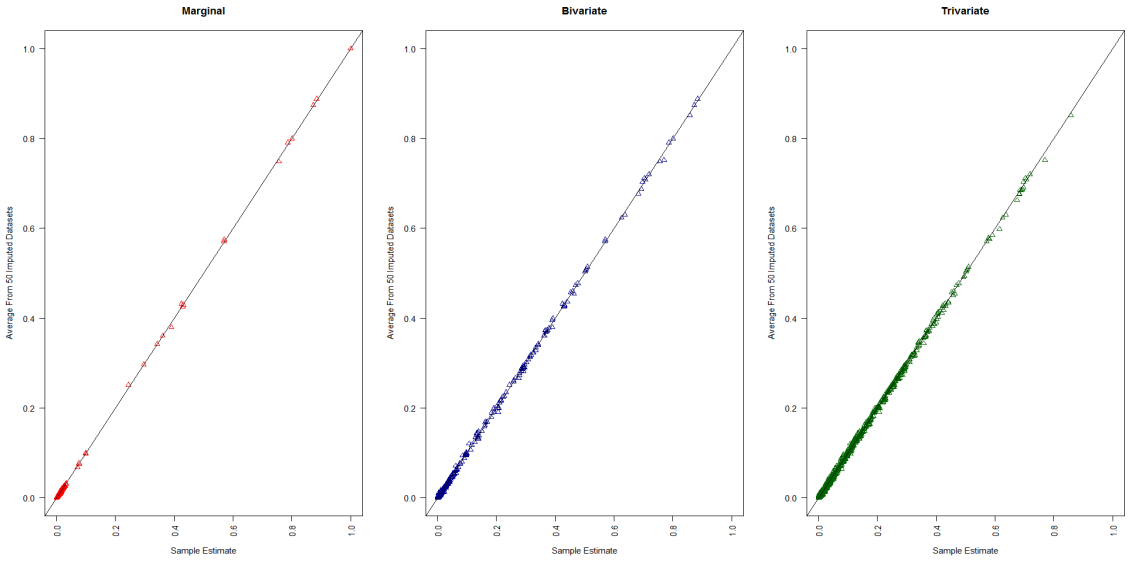


FIGURE 2.1: Marginal, bivariate and trivariate probabilities computed in the sample and imputed datasets from the truncated NDPMPM with the rejection sampler. Household heads' data values moved to the household level.

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 $\{x : x \leqslant 20\}$, $\{x : 20 < x \leqslant 50\}$, $\{x : 50 < x \leqslant 70\}$, 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 nonignorable, 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.

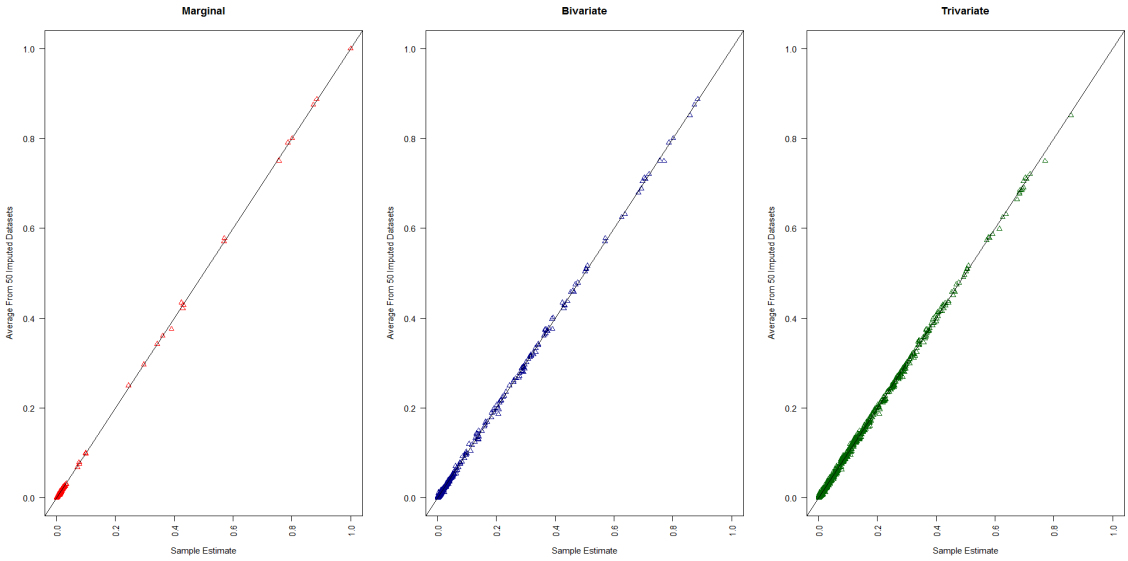


FIGURE 2.2: Marginal, bivariate and trivariate probabilities computed in the sample and imputed datasets from the truncated NDPMPM using the cap-and-weight approach. Household heads' data values to the household level.

We estimate the NDPMPM using two approaches, both using the rejection step S9′ in Section 2.3. 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. As we did in Section 2.5.1, 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 the multiple imputation combining rules of Rubin (1987) in Section 1.2.

Figures 2.1 and 2.2 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 2.1 shows the results for the NDPMPM with the rejection sampler, and Figure 2.2 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 2.2 in particular also shows that the cap-and-weight approach did not degrade the estimates.

Table 2.4 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

Table 2.4: 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:				
$n_i = 2$	.942	(.932, .949)	(.891, .917)	(.884, .911)
$n_i = 3$	.908	(.907, .937)	(.843, .890)	(.821, .870)
$n_i = 4$	.901	(.879, .917)	(.793, .851)	(.766, .828)
SP present	.696	(.682, .707)	(.695, .722)	(.695, .722)
Same race CP	.656	(.641, .668)	(.640, .669)	(.634, .664)
SP present, HH is White	.600	(.589, .616)	(.603, .632)	(.604, .634)
White CP	.580	(.569, .596)	(.577, .606)	(.574, .604)
CP with age difference less than five	.488	(.465, .492)	(.341, .371)	(.324, .355)
Male HH, home owner	.476	(.456, .484)	(.450, .479)	(.451, .480)
HH over 35, no CH present	.462	(.441, .468)	(.442, .470)	(.443, .471)
At least one biological CH present	.437	(.431, .458)	(.430, .459)	(.428, .456)
HH older than SP, White HH	.322	(.309, .335)	(.307, .339)	(.311, .343)
Adult female w/ at least one CH under 5	.078	(.070, .085)	(.062, .078)	(.061, .077)
White HH with Hisp origin	.066	(.064, .078)	(.062, .079)	(.062, .078)
Non-White CP, home owner	.058	(.050, .063)	(.038, .052)	(.037, .051)
Two generations present, Black HH	.057	(.053, .066)	(.052, .066)	(.052, .067)
Black HH, home owner	.052	(.046, .058)	(.044, .058)	(.044, .059)
SP present, HH is Black	.039	(.032, .042)	(.032, .044)	(.031, .043)
White-nonwhite CP	.034	(.029, .039)	(.038, .053)	(.043, .059)
Hisp HH over 50, home owner	.029	(.025, .034)	(.023, .034)	(.024, .034)
One grandchild present	.028	(.023, .033)	(.024, .035)	(.023, .035)
Adult Black female w/ at least one CH under 18	.027	(.028, .038)	(.025, .036)	(.025, .036)
At least two generations present, Hisp CP	.027	(.022, .031)	(.022, .032)	(.023, .033)
Hisp CP with at least one biological CH	.025	(.020, .028)	(.019, .029)	(.020, .030)
At least three generations present	.023	(.020, .028)	(.017, .026)	(.017, .026)
Only one parent	.020	(.016, .024)	(.013, .021)	(.013, .021)
At least one stepchild	.019	(.018, .026)	(.019, .030)	(.019, .030)
Adult Hisp male w/ at least one CH under 10	.018	(.017, .025)	(.014, .022)	(.014, .022)
At least one adopted CH, White CP	.008	(.005, .010)	(.004, .010)	(.004, .011)
Black CP with at least two biological children	.006	(.003, .007)	(.003, .007)	(.003, .007)
Black HH under 40, home owner	.005	(.005, .009)	(.006, .013)	(.007, .013)
Three generations present, White CP	.005	(.004, .008)	(.004, .010)	(.004, .009)
White HH under 25, home owner	.003	(.002, .005)	(.003, .007)	(.003, .007)

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

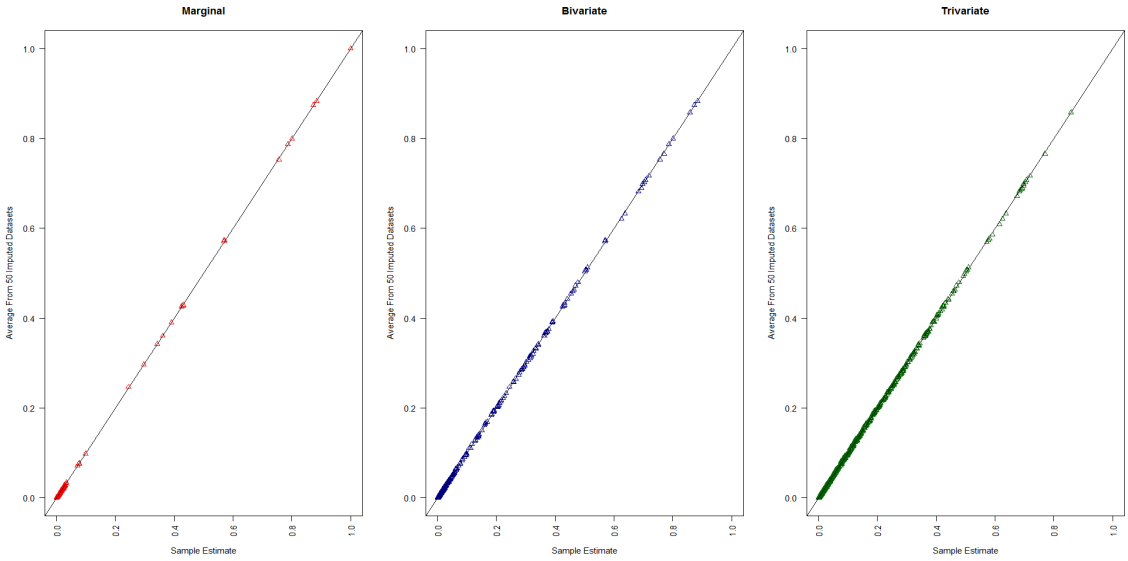


FIGURE 2.3: Marginal, bivariate and trivariate probabilities computed in the sample and imputed datasets under MCAR from the truncated NDPMPM with the rejection sampler. Household heads' data values moved to the household level.

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.

2.5.3 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 and

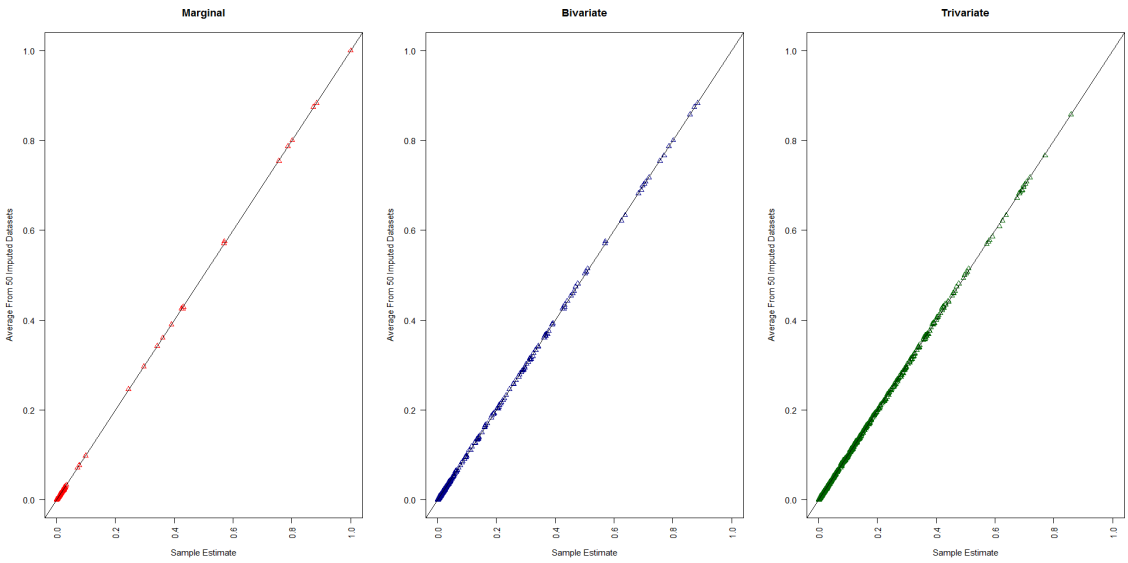


FIGURE 2.4: Marginal, bivariate and trivariate probabilities computed in the sample and imputed datasets under MCAR from the truncated NDPMPM using the cap-and-weight approach. Household heads' data values to the household level.

setup as in Section 2.5.2. We also compare the same methods as in Section 2.5.2. 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. This results in missingness rates in the 10% range across all the variables. We use these low rates to mimic the actual rates of item nonresponse in census data.

Figures 2.3 and 2.4 display each estimated marginal, bivariate and trivariate probability $\bar{q}_{50}$ plotted against its corresponding estimate from the original data, without missing values. Figure 2.3 shows the results for the NDPMPM with the rejection sampler, and Figure 2.4 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

Table 2.5: 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:				
$n_i = 2$	.942	(.932, .949)	(.924, .944)	(.925, .946)
$n_i = 3$	.908	(.907, .937)	(.887, .924)	(.890, .925)
$n_i = 4$	.901	(.879, .917)	(.854, .900)	(.855, .900)
SP present	.696	(.682, .707)	(.683, .709)	(.683, .709)
Same race CP	.656	(.641, .668)	(.637, .664)	(.638, .665)
SP present, HH is White	.600	(.589, .616)	(.590, .618)	(.590, .618)
White CP	.580	(.569, .596)	(.568, .596)	(.568, .597)
CP with age difference less than five	.488	(.465, .492)	(.422, .451)	(.422, .450)
Male HH, home owner	.476	(.456, .484)	(.455, .483)	(.456, .485)
HH over 35, no CH present	.462	(.441, .468)	(.438, .466)	(.438, .466)
At least one biological CH present	.437	(.431, .458)	(.432, .460)	(.432, .460)
HH older than SP, White HH	.322	(.309, .335)	(.308, .335)	(.306, .333)
Adult female w/ at least one CH under 5	.078	(.070, .085)	(.068, .084)	(.067, .083)
White HH with Hisp origin	.066	(.064, .078)	(.064, .079)	(.064, .079)
Non-White CP, home owner	.058	(.050, .063)	(.048, .061)	(.048, .061)
Two generations present, Black HH	.057	(.053, .066)	(.053, .066)	(.053, .067)
Black HH, home owner	.052	(.046, .058)	(.046, .059)	(.046, .059)
SP present, HH is Black	.039	(.032, .042)	(.032, .043)	(.032, .042)
White-nonwhite CP	.034	(.029, .039)	(.032, .044)	(.032, .044)
Hisp HH over 50, home owner	.029	(.025, .034)	(.025, .035)	(.025, .035)
One grandchild present	.028	(.023, .033)	(.024, .034)	(.024, .034)
Adult Black female w/ at least one CH under 18	.027	(.028, .038)	(.027, .037)	(.027, .037)
At least two generations present, Hisp CP	.027	(.022, .031)	(.022, .031)	(.022, .031)
Hisp CP with at least one biological CH	.025	(.020, .028)	(.019, .028)	(.019, .028)
At least three generations present	.023	(.020, .028)	(.019, .028)	(.019, .028)
Only one parent	.020	(.016, .024)	(.016, .024)	(.016, .024)
At least one stepchild	.019	(.018, .026)	(.018, .027)	(.018, .027)
Adult Hisp male w/ at least one CH under 10	.018	(.017, .025)	(.016, .025)	(.016, .025)
At least one adopted CH, White CP	.008	(.005, .010)	(.005, .010)	(.005, .010)
Black CP with at least two biological children	.006	(.003, .007)	(.003, .007)	(.003, .007)
Black HH under 40, home owner	.005	(.005, .009)	(.005, .010)	(.005, .011)
Three generations present, White CP	.005	(.004, .008)	(.004, .010)	(.004, .009)
White HH under 25, home owner	.003	(.002, .005)	(.004, .009)	(.004, .009)

short, the results are very similar to those in Section 2.5.2, though more accurate.

Table 2.5 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 Section 2.5.2. 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.

2.6 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 pre-

sented in this chapter. 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 2.2.2. 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.

3

Simultaneous Edit and Imputation For Household Data with Structural Zeros

This presentation in this chapter closely follows Akande et al. (2018), where the research first appeared.

3.1 Introduction

In addition to the item nonresponse often present in household data as discussed in Chapter 2, the reported data also often includes erroneous values that fail structural zeros or edit rules. Such erroneous values can arise due to data processing errors, e.g., when the age of an individual is erroneously recorded by the data collecting agency as 5 instead of 50, or respondent errors, e.g., when a household head responding to a survey accidentally selects the "relationship to household head" status of his/her biological parent as a child.

Agencies generally prefer not to analyze or disseminate data with overt errors. The errors can affect inferences, potentially resulting in misleading conclusions. When included in data releases, errors can undermine data users' confidence in the quality of the data. On the other hand, inferences based only on the subset of data

without overt errors can be inefficient or even biased, depending on the reasons why values are in error (Rubin, 1976; Little and Rubin, 2002). It is therefore prudent for agencies to edit faulty data in hopes of improving quality before analysis or dissemination.

When confronted with errors, ideally the agency can re-contact respondents to ascertain their true responses. However, this can be expensive and impractical to do for all respondents, especially in the context of censuses or large surveys. Many agencies therefore supplement re-contact operations with a process known as automatic edit-imputation. In the edit step, agencies specify an error localization process to determine a set of values that are in error for each record. This is usually done using a variant of the error localization suggested by Fellegi and Holt (1976), where the values are selected by minimizing the number of fields necessary to turn an erroneous record into a theoretically valid one (Winkler, 1995; Winkler and Petkunas, 1997). In the imputation step, the values selected in the error localization are replaced with plausibly valid entries (de Waal and Coutinho, 2005; de Waal et al., 2011). This is usually done by some form of hot deck imputation (Kalton and Kasprzyk, 1986; Andridge and Little, 2010).

Kim et al. (2015a) and Manrique-Vallier and Reiter (2018) describe some shortcomings of edit-imputation approaches based on the Fellegi-Holt paradigm for non-nested data. In particular, they highlight two problems, namely that (i) the error localization process typically does not fully take advantage of multivariate relationships in the data, and (ii) the selection of a single error localization coupled with a single imputation underestimates uncertainty. These shortcomings can be relevant in household data in complicated ways. To illustrate, suppose the reported data include a household with a head, a spouse, and three biological children, two of whom are reported as age 6 and 8 and a third reported as age 30. The reported age 30 exceeds the age of the household head. Most likely, the reported 30 year old has at least one

field in error; the person's age or relationship to household head is likely erroneous. Agencies following the Fellegi-Holt paradigm would change one of these two fields based on some heuristic, e.g., change the variable that is more likely to have errors according to experience in similar datasets. Now suppose the data also inform us that everyone except the reported 30 year child has the same race of Asian, and the 30 year old reports a race of black. The data may well indicate that biologically-related families with two Asian parents, two Asian children, and one black child are highly unlikely. Thus, it may be more plausible to leave age alone and change the relationship value to "unrelated" rather than leave relationship alone and change age. Of course, it may well be that multiple fields are in error, including race. In any case, we would like to incorporate the uncertainty in the error localization by averaging over plausible localizations and corrected values.

In this chapter, we present an edit-imputation approach intended to address these shortcomings. Specifically, we follow the approach of Kim et al. (2015a) and Manrique-Vallier and Reiter (2018) and handle the edit and imputation processes simultaneously with a Bayesian hierarchical model. The hierarchical model includes (i) a multivariate model for the true latent values of the data which has support only on theoretically possible households, (ii) a model for locations of errors given the latent true values, and (iii) a model for the reported values for fields in error. For the multivariate model of the true values, we again use the truncated NDPMPM in Chapter 2. For the error location and reporting models, we adapt the measurement error model used for non-nested categorical data in Manrique-Vallier and Reiter (2018). We discuss alternative measurement error models in Section 3.2.2 and Section 3.5. We use a Markov chain Monte Carlo sampler to fit the hierarchical model, which generates plausible datasets without errors as byproducts. These can be analyzed or disseminated as multiple imputations (Rubin, 1987; Ghosh-Dastidar and Schafer, 2003).

The remainder of this chapter is organized as follows. In Section 3.2, we present the Bayesian edit-imputation model for household data, which we refer to as the EIHD model. In Section 3.3, we describe the MCMC algorithm for fitting the EIHD. In Section 3.4, we report the results of simulation studies used to illustrate the performance of the EIHD. The simulations are based on a subset of data from the 2012 American Community Survey (ACS) public use files. In Section 3.5, we discuss findings, caveats and future work. The EIHD is implemented in the R package, "NestedCategBayesImpute," available on CRAN. Source code is at `https://github.com/akandelanre/Edit-Imputation-for-Nested-Data/tree/master`.

3.2 The EIHD Model

In describing the EIHD, we use the notation from Chapter 2 and Manrique-Vallier and Reiter (2018). Let n, n_i and N be defined as in Chapter 2. For $k = p+1, \ldots, p+q$, let $X^1_{ik} \in \{1, \ldots, d_k\}$ be the true unobserved value of household level variable k for household i, which is assumed to be identical for all n_i individuals in household i. Similarly, for $k = 1, \ldots, p$ and $j = 1, \ldots, n_i$, for each household let $X^1_{ijk} \in \{1, \ldots, d_k\}$ be the true unobserved value of individual level variable k for person j in household i. We associate each household i with the true unobserved $\mathbf{X}^1_i = (X^1_{ip+1}, \ldots, X^1_{ip+q}, X^1_{i11}, \ldots, X^1_{in_ip})$, which includes all household level and individual level variables for the n_i individuals in the household.

Let $\mathcal{X}^1 = (\mathbf{X}^1_1, \ldots, \mathbf{X}^1_n)$. We do not observe $\mathcal{X}^1$; instead, we observe the reported data $\mathcal{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$. Each $\mathbf{Y}_i$ is a potentially contaminated version of its corresponding $\mathbf{X}^1_i$. We assume that each $\mathbf{Y}_i$ is generated conditional on $\mathbf{X}^1_i$ through a measurement error model, with density $f_Y(\mathbf{Y}|\mathbf{X}^1, \theta_\mathbf{Y})$. Whenever it is the case that $\mathbf{Y}_i \neq \mathbf{X}^1_i$, we say that the observed data for the i-th household contains errors.

Let $\mathcal{H}$, $\mathcal{C}$, $\mathcal{C}_h$, $\mathcal{S}$ and $\mathcal{S}_h$ all be defined as in Chapter 2. Although true responses are such that $\mathbf{X}^1_i \notin \mathcal{S}_{n_i}$, reported responses potentially can violate the edit rules, i.e.,

$\Pr(\mathbf{Y}_i \in \mathcal{S}_{n_i}) > 0$. Whenever $\mathbf{Y}_i \in \mathcal{S}_{n_i}$, we know for sure that $\mathbf{Y}_i$ contains errors. We refer to such errors as detectable. Though the errors may be detectable, the exact location of the errors is usually unknown. For example, suppose a household contains a male household head who is 35 years old and his biological child who is 60 years old. Certainly, this household contains errors but we cannot say for sure whether the error is in the ages of at least one of the individuals, the relationship between them (since the 60 year old could in fact be a parent instead of a biological child), or both.

It may be possible that $\mathbf{Y}_i \notin \mathcal{S}_{n_i}$ while $\mathbf{Y}_i \neq \mathbf{X}_i^1$. We refer to such errors as undetectable. In this chapter, we make the simplifying assumption that the only errors in the data are detectable ones. While this assumption can be viewed as somewhat unrealistic, it is consistent with the practice of most statistical agencies that use automatic edit-imputation algorithms, including Fellegi-Holt systems (de Waal et al., 2011; Kim et al., 2015a). It stems from a philosophy that agencies should change as few respondents' reported values as possible. We describe how to relax this assumption at the end of Section 3.2.2.

Finally, we assume that each unobserved $\mathbf{X}_i^1$ is stochastically generated from a common data generating process with density $f_X(\mathbf{X}^1|\theta_{\mathbf{X}})$ and support restricted to $\mathbf{X}^1 \in \mathcal{C} - \mathcal{S}$, so that the realized values for $\mathcal{X}^1$ must satisfy the structural zero rules. Under this setup, the objective is to estimate the joint distribution of the underlying true data $\mathcal{X}^1$ and the erroneous data $\mathcal{Y}$, and obtain posterior predictive samples of $\mathcal{X}^1$ from it.

3.2.1 True response model

In theory, $f_X(\mathbf{X}^1|\theta_{\mathbf{X}})$ can be any multivariate categorical data model that adequately describes the joint distribution of all the variables, has support restricted to $\mathcal{C}-\mathcal{S}$, and captures the relevant structure in $\mathcal{X}^1$. For household data, the truncated NDPMPM

model of Hu et al. (2018) presented in Chapter 2.2 has those properties. That is the model we select as $f_X(\mathbf{X}^1|\theta_{\mathbf{X}})$.

3.2.2 Measurement error model

For the measurement error model, we introduce a series of binary indicator variables to help with model specification. Let $Z_i = 1$ if household i has an error and $Z_i = 0$ otherwise. Let $E_{ik} = 1$ if household level variable k is in error for household i, and $E_{ik} = 0$ otherwise. Let $E_{ijk} = 1$ if individual level variable k is in error for person j in household i, and $E_{ijk} = 0$ otherwise. By design, $Z_i = 0$ implies that $E_{ik} = E_{ijk} = 0$ for all j and k corresponding to household i, whereas $Z_i = 1$ implies that at least one of the E_{ik} and E_{ijk} corresponding to household i equal one. Since we assume no undetectable errors, each Z_i is observed rather than latent. This saves computational time, since whenever $\mathbf{Y}_i \notin \mathcal{S}_{n_i}$, we set $\mathbf{X}_i^1 = \mathbf{Y}_i$ and do not need to sample a new plausible value for $\mathbf{X}_i^1$. Finally, let $\mathbf{E} = (\mathbf{E}_1, \ldots, \mathbf{E}_n)$, where each $\mathbf{E}_i = (E_{ip+1}, \ldots, E_{ip+q}, E_{i11}, \ldots, E_{in_ip})$.

We formulate the measurement error model as two sub-models, namely (i) a reporting model for $\mathbf{Y}_i$ conditional on $\mathbf{X}_i^1$ and $\mathbf{E}_i$, and (ii) an error location model for $\mathbf{E}_i$ conditional on the Z_i. For the reporting model for any Y_{ijk} or Y_{ik}, we have

$$Y_{ijk}|X_{ijk}^1 = c, E_{ijk} = e \sim \begin{cases} \delta_{X_{ijk}^1} & \text{if } e = 0 \\ \text{Discrete}(\{1, \ldots, d_k\} \backslash \{c\}; \{q_k, \ldots, q_k\}) & \text{if } e = 1 \end{cases} \tag{3.1}$$

$$\forall\ i, j, k = 1, \ldots, p$$

$$Y_{ik}|X_{ik}^1 = c, E_{ik} = e \sim \begin{cases} \delta_{X_{ik}^1} & \text{if } e = 0 \\ \text{Discrete}(\{1, \ldots, d_k\} \backslash \{c\}; \{q_k, \ldots, q_k\}) & \text{if } e = 1 \end{cases} \tag{3.2}$$

$$\forall\ i, k = p+1, \ldots, p+q$$

where $q_k = 1/(d_k - 1)$. This model assumes that each Y_{ijk} (and Y_{ik}) in error is generated uniformly from the set of all possible values for variable k, excluding the

true value X^1_{ijk} (and X^1_{ik}). It implies that when people make reporting errors, they do so completely randomly and independently across variables. One could replace these assumptions with informative models, should such information be available from other data sources or previous experience. For example, one could specify reporting models that put higher probability on categories adjacent to the true X^1_{ijk}, reflecting response errors where people mistakenly select nearby categories on a survey form.

For the error location model for any E_{ijk} or E_{ik}, we have

$$E_{ijk}|Z_i = z, \epsilon_k \sim \begin{cases} \delta_0 & \text{if } z = 0 \\ \text{Bernoulli}(\epsilon_k) & \text{if } z = 1 \end{cases} \quad \forall\ i, j, k = 1, \ldots, p \tag{3.3}$$

$$E_{ik}|Z_i = z, \epsilon_k \sim \begin{cases} \delta_0 & \text{if } z = 0 \\ \text{Bernoulli}(\epsilon_k) & \text{if } z = 1 \end{cases} \quad \forall\ i, k = p+1, \ldots, p+q \tag{3.4}$$

This model assumes that the error indicators are independent across variables, which again accords with people making errors at random although possibly with different rates for different variables. One could replace these assumptions with models conditional on the true values, e.g., people who are older are more likely to make errors on certain variables. When conditioning on true values that are latent, this creates a nonignorable faulty data mechanism.

We complete the specification with prior distributions for $\boldsymbol{\epsilon} = \{\epsilon_k : k = 1, \ldots, p+q\}$. In the empirical study with the ACS data, we use conjugate beta priors for each ϵ_k, primarily for computational convenience. We have

$$\epsilon_k \sim \text{Beta}(a_{\epsilon_k}, b_{\epsilon_k}) \ \ \forall\ k = 1, \ldots, p, p+1, \ldots, p+q. \tag{3.5}$$

We set $a_{\epsilon_k} = b_{\epsilon_k} = 1$ for each $k \in \{1, \ldots, p, p+1, \ldots, p+q\}$ to represent complete ignorance about the true error rates. In applied contexts, we suggest setting each $(a_{\epsilon_k}, b_{\epsilon_k})$ to reflect prior beliefs on the error rates whenever reasonable prior information is available.

The measurement error model can be extended to allow for undetectable errors by letting Z_i be latent for $\mathbf{Y}_i \notin \mathcal{S}$. We continue to let $Z_i = 1$ for cases where $\mathbf{Y}_i \in \mathcal{S}$. For example, we can let $Z_i|\rho \sim \text{Bernoulli}(\rho)$, with $\rho \sim \text{Beta}(a_\rho, b_\rho)$ reflecting prior beliefs about the fraction of cases with errors. We leave investigation of this model for future research.

The model also can handle missing values simultaneously with faulty data. One sets $Z_i = 1$ for households that contain at least one missing entry and $E_{ik} = E_{ijk} = 1$ for all variables that have missing values, forcing $\mathbf{X}_i^1$ to be imputed for those households. This presumes that (i) the values are missing at random, and (ii) the same measurement error and true response models apply for the households with error and the households with missing data.

3.3 MCMC Estimation

We use a Gibbs sampler to estimate the posterior distribution of the parameters in the EIHD model. Given the data $\mathcal{Y}$ and a draw of $(\mathbf{G}^1, \mathbf{G}^0, \mathbf{M}^1, \mathbf{M}^0, \mathcal{X}^0, \theta_\mathbf{X}, n_0)$, we update $(\mathcal{X}^1, \mathbf{E}, \boldsymbol{\epsilon})$. We outline these updates in Section 3.3.1. We then update $(\mathbf{G}^1, \mathbf{G}^0, \mathbf{M}^1, \mathbf{M}^0, \mathcal{X}^0, \theta_\mathbf{X}, n_0)$ given a draw of $\mathcal{X}$ using the modified cap-and-weight Gibbs sampler in Chapter 2.4.2 – we only make a small change to the notation by letting n_h^1 be the number of households of size h in $\mathcal{X}^1$, so that $n = \sum_h n_h^1$; we do so to make the notation similar to n_h^0, the number of households of size h in $\mathcal{X}^0$.

Upon convergence of the Gibbs sampler, analysts can obtain posterior inferences for parameters of interest or treat the posterior samples of $\mathcal{X}^1$ as multiply imputed datasets (Rubin, 1987). For the latter, analysts can select a modest number L of datasets (usually, $L \geqslant 5$), reasonably spaced so that they are approximately independent.

3.3.1 *Sampling* $(\mathcal{X}^1, \mathbf{E}, \boldsymbol{\epsilon})$

Sampling ϵ_k for each variable k is straightforward since error rates are independent across variables. We use the following step in the sampler.

S1. For $k = 1, \ldots, p, p+1, \ldots, p+q$, sample $\epsilon_k | \ldots \sim \text{Beta}(a_{\epsilon_k} + a^\star_{\epsilon_k}, b_{\epsilon_k} + b^\star_{\epsilon_k})$, where

$$a^\star_{\epsilon_k} = \sum_{i|Z_i=1} \mathbb{1}(E_{ik} = 1), \quad b^\star_{\epsilon_k} = \sum_{i|Z_i=1} \mathbb{1}(E_{ik} = 0) \quad \text{if } k \in \{p+1, \ldots, p+q\}; \quad \text{and}$$

$$a^\star_{\epsilon_k} = \sum_{i|Z_i=1} \sum_{j=1}^{n_i} \mathbb{1}(E_{ijk} = 1), \quad b^\star_{\epsilon_k} = \sum_{i|Z_i=1} \sum_{j=1}^{n_i} \mathbb{1}(E_{ijk} = 0) \quad \text{if } k \in \{1, \ldots, p\}.$$

Sampling $(\mathcal{X}^1, \mathbf{E})$ is more involved. Since each $\mathbf{E}_i$ is completely determined by $\mathbf{X}^1_i$ and $\mathbf{Y}_i$, we cannot form independent Gibbs steps for each using the full conditionals $\Pr(\mathbf{X}^1_i | \ldots)$ and $\Pr(\mathbf{E}_i | \ldots)$. Instead, we sample directly from $\Pr(\mathbf{X}^1_i, \mathbf{E}_i | \ldots)$ using the factorization

$$\Pr(\mathbf{X}^1_i, \mathbf{E}_i | \ldots) = \Pr(\mathbf{X}^1_i | \ldots, -\{\mathbf{E}_i\}) \times \Pr(\mathbf{E}_i | \ldots)$$

where $\Pr(\mathbf{X}^1_i | \ldots, -\{\mathbf{E}_i\})$ is the conditional pmf of $\mathbf{X}^1_i$ given all other random variables in the model except $\mathbf{E}_i$. We therefore sample $(\mathcal{X}^1, \mathbf{E})$ using the following steps.

S2. For $i = 1, \ldots, n$, set $\mathbf{X}^1_i = \mathbf{Y}_i$ if $Z_i = 0$. If $Z_i = 1$, sample $\mathbf{X}^1_i$ from

$$\Pr(\mathbf{X}^1_i | \ldots, -\{\mathbf{E}_i\}) \propto \mathbb{1}\{\mathbf{X}^1_i \notin \mathcal{S}_h\} \pi_{G^1_i} \prod_{k=p+1}^{p+q} \lambda^{(k)(\star)}_{G^1_i X^1_{ik}} \left(\prod_{j=1}^{n_i} \omega_{G^1_i M^1_{ij}} \prod_{k=1}^{p} \phi^{(k)(\star)}_{G^1_i M^1_{ij} X^1_{ijk}} \right),$$

where

$$\lambda^{(k)(\star)}_{G^1_i X^1_{ik}} = \lambda^{(k)}_{G^1_i X^1_{ik}} (1 - \epsilon_k)^{\mathbb{1}\{Y_{ik} = X^1_{ik}\}} (q_k \epsilon_k)^{\mathbb{1}\{Y_{ik} \neq X^1_{ik}\}}, \quad \text{and}$$

$$\phi^{(k)(\star)}_{G^1_i M^1_{ij} X^1_{ijk}} = \phi^{(k)}_{G^1_i M^1_{ij} X^1_{ijk}} (1 - \epsilon_k)^{\mathbb{1}\{Y_{ijk} = X^1_{ijk}\}} (q_k \epsilon_k)^{\mathbb{1}\{Y_{ijk} \neq X^1_{ijk}\}}.$$

Sampling from this conditional distribution is nontrivial because of the dependence among variables induced by the structural zero rules in each $\mathcal{S}_h$. However, we can generate the samples through the following rejection sampling scheme.

(a) Set $X^1_{ik} = n_i$ for the k corresponding to the household size. For the remaining household level variables $k \in \{p+1, \ldots, p+q\}$, sample $X^1_{ik} \sim \text{Discrete}(\lambda^{(k)(\star)}_{G^1_i 1}, \ldots, \lambda^{(k)(\star)}_{G^1_i d_k})$.

(b) Sample $X^1_{ijk} \sim \text{Discrete}(\phi^{(k)(\star)}_{G^1_i M^1_{ij} 1}, \ldots, \phi^{(k)(\star)}_{G^1_i M^1_{ij} d_k})$ for each $k \in \{1, \ldots, p\}$ and $j = 1, \ldots, n_i$.

(c) Set the sampled household level and individual level values to $\mathbf{X}^{1\star}_i$.

(d) If $\mathbf{X}^{1\star}_i \notin \mathcal{S}_h$, set $\mathbf{X}^1_i = \mathbf{X}^{1\star}_i$, otherwise, return to step (a).

We then set $\mathcal{X}^1 = (\mathbf{X}^1_1, \ldots, \mathbf{X}^1_n)$. We then set E_i deterministically as follows.

S3. For $i = 1, \ldots, n$ and $k \in \{p+1, \ldots, p+q\}$, set $E_{ik} = 1$ if $X^1_{ik} \neq Y_{ik}$ and $E_{ik} = 0$ otherwise. Similarly, for $i = 1, \ldots, n$, $j = 1, \ldots, n_i$ and $k \in \{1, \ldots, p\}$, set $E_{ijk} = 1$ if $X^1_{ijk} \neq Y_{ijk}$ and $E_{ijk} = 0$ otherwise.

3.4 Empirical Study

To illustrate the performance of the EIHD, we again use data from the 2012 ACS. We construct a population of 842746 households from which we sample $n = 3000$ households comprising $N = 8686$ individuals. The sample includes households with two to six people, and $\{n^1_2, \ldots, n^1_6\} = \{1541, 630, 525, 210, 94\}$. We again work with the variables described in Table 2.3, although the variable "Household size" now include two extra categories for households of sizes five and six. The structural zeros remain the same as those in Appendix A. The Census Bureau purges the 2012 ACS public use microdata file of detectable errors. Therefore, for each household i, we

treat its values on the public use file as error-free $\mathbf{X}_i^1$, and we purposefully introduce errors and missing values to the complete data file, under four different simulation scenarios.

3.4.1 Empirical study 1: Uniform substitution model with $\rho = 0.2$ and 20% missing data.

In the first simulation scenario, we randomly generate each Z_i, where $i = 1, \ldots, 3000$, from a Bernoulli(ρ) distribution, where $\rho = 0.2$. For each household with $Z_i = 0$, we let $\mathbf{X}_i^1 = \mathbf{Y}_i$. For each household with $Z_i = 1$, we sample error locations using (3.3) and (3.4), and sample reported values from (3.1) and (3.2). As we allow only detectable errors, we create errors only in variables used in the exemplary definitions of the structural zeros. These include the gender and age of the household head, and gender, age and relationship to household head for the remaining household members. We set $\epsilon = (0.65, 0.80, 0.70, 0.85, 0.90)$ for these five variables. For each household with $Z_i = 1$, we repeatedly sample values until the household fails the structural zero rules. This results in approximately 17% overall error rate for each variable across the 3000 sampled households. Although editing rates can be smaller in some contexts (Jackson, 2010), we view the 17% error rate as a challenging but reasonable stress test for the EIHD. Finally, we introduce missing values for all variables, except household size, not subject to errors. To do so, we randomly and independently blank 20% of the values for each variable.

The method of generating the reported values implies that the true substitution probabilities q_k in (3.1) and (3.2) do not equal $1/(d_k - 1)$ for all levels of variable k. For example, in the contaminated data that we generated, given that $E_{ijk} = 1$ for the relationship to household head variable, the probability of a spouse being wrongly reported as a parent is 0.177 whereas the probability of a spouse being wrongly reported as a sibling is 0.112. However, we still set $q_k = 1/(d_k - 1)$ when

fitting our model to the data, mirroring a scenario where an agency uses a default application of EIHD and unknown true substitution rates.

We put the data for the household head as household level variables as suggested in Chapter 2.4.1. This offers computational gains relative to modeling the household head variables at the individual level as discussed in Chapter 2. For the cap-and-weight approach, we consider two choices for the weights, namely $(\psi_2, \psi_3, \psi_4, \psi_5, \psi_6) = (1, 1, 1, 1, 1)$, which corresponds to the original sampler, and $(\psi_2, \psi_3, \psi_4, \psi_5, \psi_6) = (1/2, 1/2, 1/3, 1/3, 1/3)$, which is computationally efficient according to preliminary runs of the Gibbs sampler.

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 $F = 20$ and $S = 15$ based on initial tuning runs. For convergence, we examined trace plots of α, β, and a random sample of the marginal probabilities of the variables in (2.3) and (2.4). The posterior number of occupied household-level clusters usually ranges from 10 to 14 for $(\psi_2, \psi_3, \psi_4, \psi_5, \psi_6) = (1, 1, 1, 1, 1)$ and from 13 to 19 for $(\psi_2, \psi_3, \psi_4, \psi_5, \psi_6) = (1/2, 1/2, 1/3, 1/3, 1/3)$, while the posterior number of occupied individual-level clusters across all household-level clusters ranges from 3 to 8 with $(\psi_2, \psi_3, \psi_4, \psi_5, \psi_6) = (1, 1, 1, 1, 1)$ and from 3 to 7 with $(\psi_2, \psi_3, \psi_4, \psi_5, \psi_6) = (1/2, 1/2, 1/3, 1/3, 1/3)$. Thus, it appears that $F = 20$ and $S = 15$ are adequate (Hu et al., 2018). For each of the two choices of ψ, we create $L = 50$ multiply imputed datasets, $\mathbf{Z} = (\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(50)})$, from the posterior samples of $\mathcal{X}^1$. From the imputed datasets, we estimate the marginal probabilities of all the variables, bivariate probabilities of all possible pairs of variables, and trivariate probabilities of all possible triplets of variables. There are 229 marginal probabilities, 18135 bivariate probabilities and 623173 trivariate probabilities. As was the case in Chapter 2.5, we also estimate selected probabilities that depend on within-household relationships and characteristics of the household head. We combine the estimates

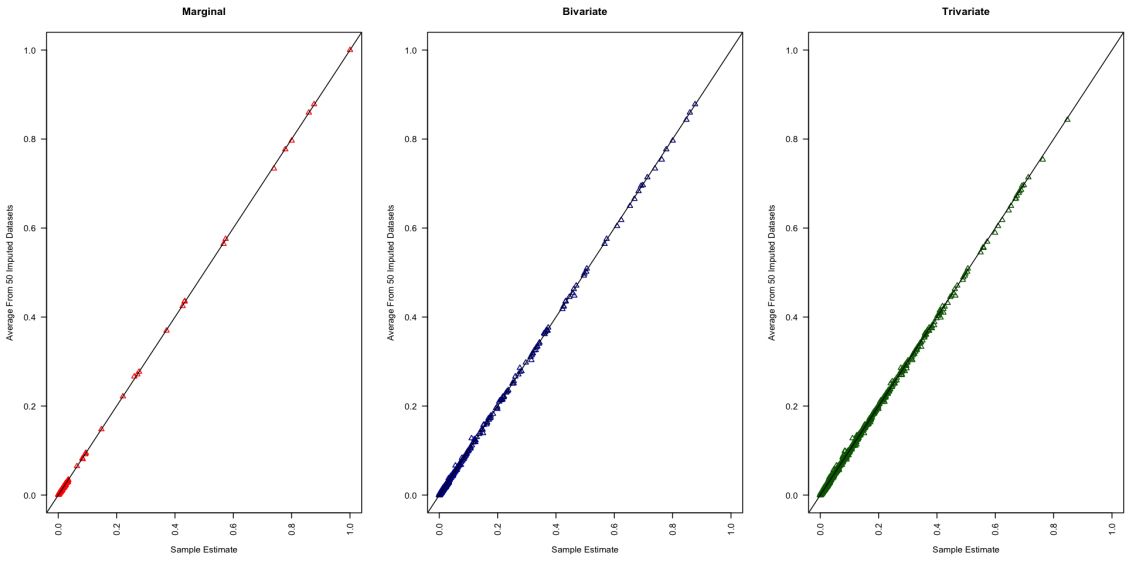


FIGURE 3.1: **Empirical study 1:** Marginal, bivariate and trivariate probabilities computed in the original and imputed datasets from the EIHD with $(\psi_2, \psi_3, \psi_4, \psi_5, \psi_6) = (1, 1, 1, 1, 1)$.

using the multiple imputation combining rules.

We only examine the performance of the multiple imputation inferences for the EIHD model. We are not aware of any publicly available Fellegi-Holt editing systems for household level data. For example, the "editrules" package (de Jonge and van der Loo, 2015) in *R* applies Fellegi-Holt editing only for independent individuals. We also cannot easily compare to complete case analysis due to the rate of missingness. About 80% of the households have at least one missing entry, with the proportion rising to about 85% when one adds the households with faulty data.

Figures 3.1 and 3.2 display the estimated probabilities obtained from the multiple imputation combining rules for the EIHD with $(\psi_2, \psi_3, \psi_4, \psi_5, \psi_6) = (1, 1, 1, 1, 1)$ and $(\psi_2, \psi_3, \psi_4, \psi_5, \psi_6) = (1/2, 1/2, 1/3, 1/3, 1/3)$, respectively. In both versions of EIHD, the point estimates are very close to those from the original data (i.e., before the introduction of missing and erroneous values), suggesting that the EIHD edit-imputations captured these features of the joint distribution of the variables. Table

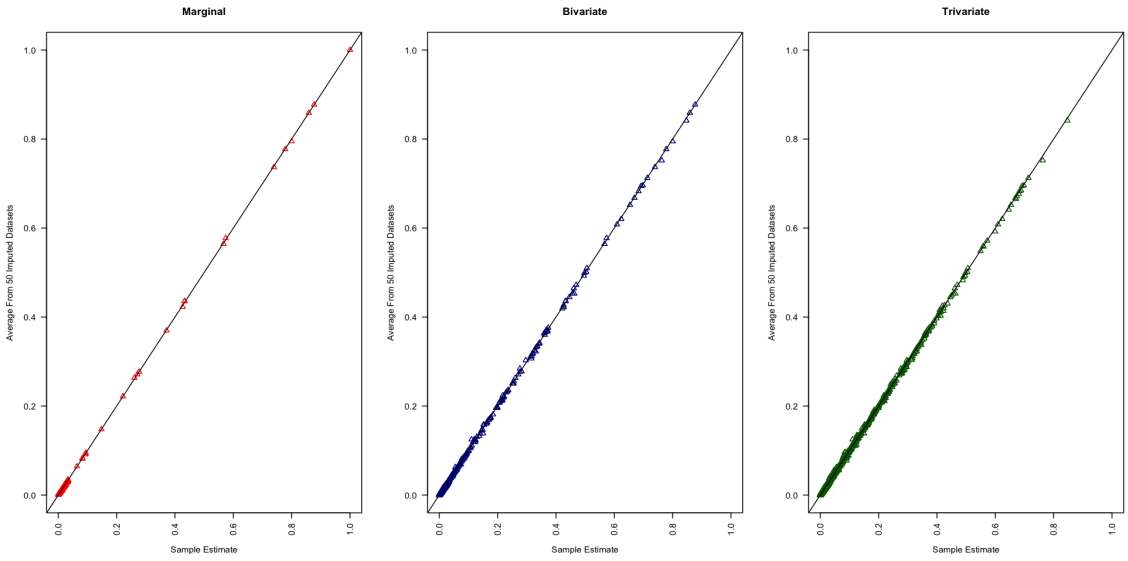


FIGURE 3.2: **Empirical study 1:** Marginal, bivariate and trivariate probabilities computed in the original and imputed datasets from the EIHD with $(\psi_2, \psi_3, \psi_4, \psi_5, \psi_6) = (1/2, 1/2, 1/3, 1/3, 1/3)$.

3.1 displays multiple imputation 95% confidence intervals for the selected probabilities involving within-household relationships, as well as the values in the full population of 842746 households. For the most part, the intervals from both versions of EIHD are quite close to those based on the original data. One exception is when estimating the proportion of households where everyone is the same race, especially for larger households. Our illustrations in Chapter 2 also identified biases for the same estimands when using the NDPMPM to generate fully synthetic data or impute missing data. With so many levels, it is difficult to estimate within household relationships involving age with high accuracy. We revisit this issue in Section 3.5.

3.4.2 Empirical study 2: Uniform substitution model with $\rho = 0.4$ and 30% missing data.

In this simulation scenario, we simulate a higher rate of households that violate the constraints, setting $\rho = 0.4$ and contaminate the households as in Section 3.4.1. For gender and age of household head, and for gender, age and relationship to house-

Table 3.1: **Empirical study 1:** Confidence intervals for selected probabilities that depend on within-household relationships in the original and imputed datasets. "Sample" is based on the sampled data before introducing errors and missing values, "EIHD" is the EIHD with $(\psi_2, \psi_3, \psi_4, \psi_5, \psi_6) = (1, 1, 1, 1, 1)$, and "EIHD Capped" is the EIHD with $(\psi_2, \psi_3, \psi_4, \psi_5, \psi_6) = (1/2, 1/2, 1/3, 1/3, 1/3)$. "HH " means household head, "SP" means spouse, "CH" means child, and "CP" means couple. "Q" is the value in the full population of 842, 746 households.

	Q	Sample	EIHD	EIHD w/caps
All same race household:				
$n_i = 2$	.942	(.936, .958)	(.909, .938)	(.901, .934)
$n_i = 3$	.908	(.863, .912)	(.830, .890)	(.820, .884)
$n_i = 4$	.901	(.890, .938)	(.839, .906)	(.827, .894)
$n_i = 5$	.887	(.848, .933)	(.772, .892)	(.770, .888)
$n_i = 6$	.871	(.766, .914)	(.693, .875)	(.647, .846)
SP present	.704	(.687, .720)	(.682, .718)	(.682, .719)
Same race CP	.663	(.645, .679)	(.628, .667)	(.624, .662)
SP present, HH is White	.599	(.590, .625)	(.585, .624)	(.585, .623)
White CP	.579	(.572, .607)	(.563, .602)	(.561, .600)
CP with age difference less than five	.494	(.472, .508)	(.405, .442)	(.403, .441)
At least one biological CH present	.490	(.473, .509)	(.481, .519)	(.479, .517)
Male HH, home owner	.472	(.455, .491)	(.442, .482)	(.445, .485)
HH over 35, no CH present	.416	(.397, .432)	(.390, .428)	(.390, .428)
HH older than SP, White HH	.320	(.318, .352)	(.308, .345)	(.309, .347)
Adult female w/ at least one CH under 5	.093	(.078, .098)	(.077, .099)	(.077, .098)
White HH with Hisp origin	.076	(.059, .077)	(.055, .074)	(.055, .074)
Non-White CP, home owner	.063	(.048, .064)	(.039, .056)	(.039, .057)
Two generations present, Black HH	.059	(.043, .059)	(.045, .063)	(.045, .063)
Black HH, home owner	.052	(.039, .053)	(.038, .054)	(.038, .054)
At least three generations present	.041	(.032, .046)	(.032, .047)	(.031, .047)
SP present, HH is Black	.041	(.028, .042)	(.028, .043)	(.028, .043)
At least two generations present, Hisp CP	.040	(.028, .040)	(.027, .041)	(.027, .041)
Hisp CP with at least one biological CH	.038	(.027, .040)	(.026, .040)	(.026, .049)
White-nonwhite CP	.035	(.027, .039)	(.032, .050)	(.034, .052)
One grandchild present	.032	(.021, .032)	(.023, .038)	(.024, .038)
Hisp HH over 50, home owner	.030	(.023, .035)	(.023, .036)	(.023, .038)
Adult Black female w/ at least one CH under 18	.030	(.022, .034)	(.020, .033)	(.020, .033)
Adult Hisp male w/ at least one CH under 10	.027	(.015, .025)	(.014, .024)	(.014, .025)
At least one stepchild	.026	(.021, .032)	(.020, .033)	(.020, .033)
Only one parent	.023	(.015, .025)	(.018, .030)	(.017, .030)
Three generations present, White CP	.013	(.008, .016)	(.007, .016)	(.007, .016)
At least one adopted CH, White CP	.010	(.008, .015)	(.007, .015)	(.007, .016)
Black CP with at least two biological children	.009	(.004, .010)	(.005, .012)	(.005, .012)
Black HH under 40, home owner	.006	(.003, .008)	(.005, .014)	(.005, .013)
White HH under 25, home owner	.003	(.000, .002)	(.002, .009)	(.001, .007)

hold head for remaining household members, we set $\epsilon = (0.25, 0.60, 0.20, 0.85, 0.50)$, respectively. This results in overall error rates in the five variables of approximately

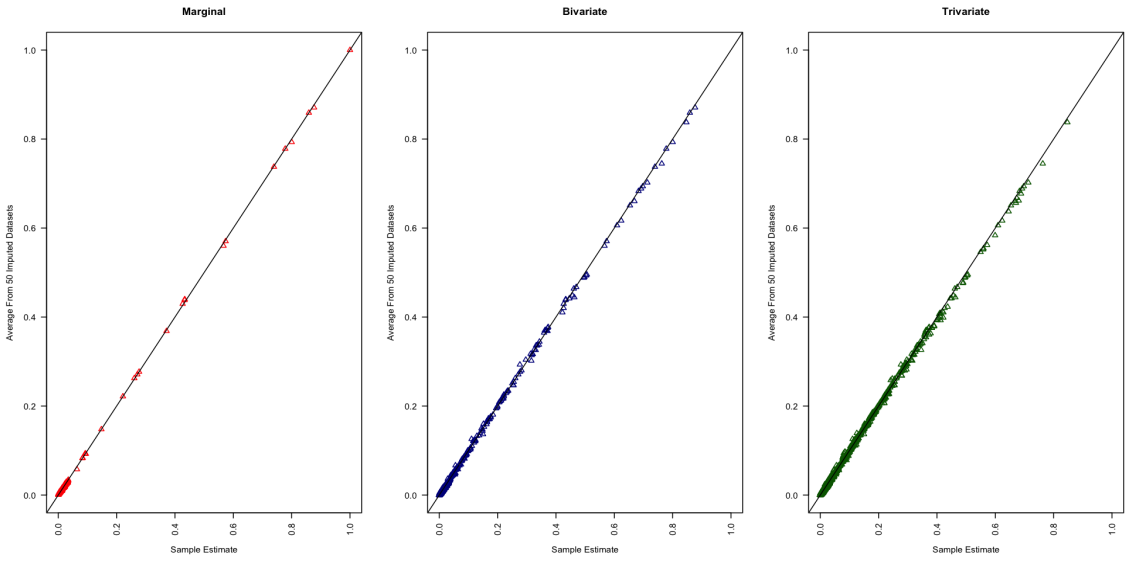


FIGURE 3.3: **Empirical study 2:** Marginal, bivariate and trivariate probabilities computed in the original and imputed datasets from the EIHD with $(\psi_2, \psi_3, \psi_4, \psi_5, \psi_6) = (1, 1, 1, 1, 1)$.

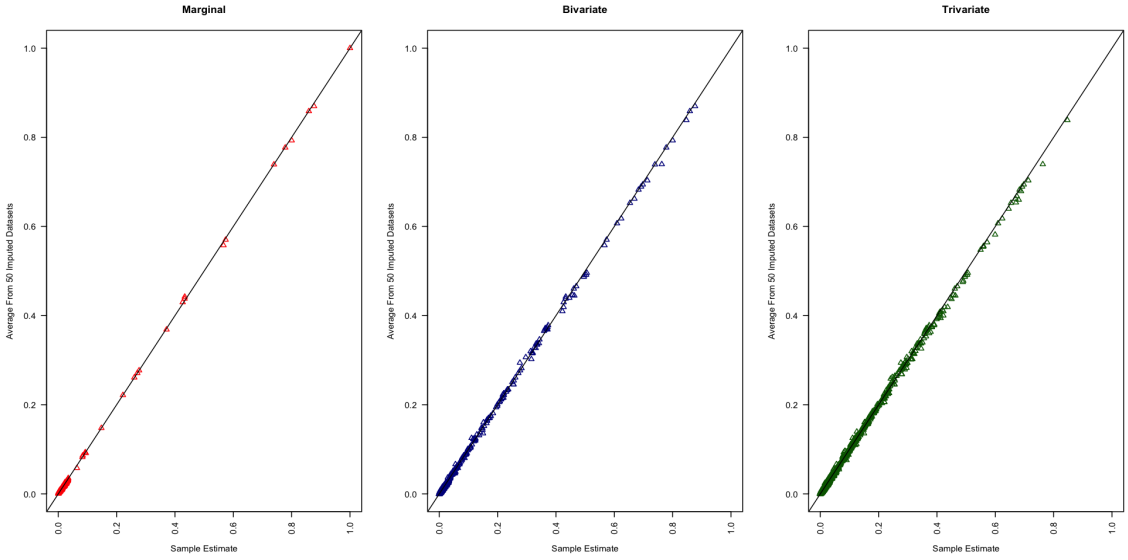


FIGURE 3.4: **Empirical study 2:** Marginal, bivariate and trivariate probabilities computed in the original and imputed datasets from the EIHD with $(\psi_2, \psi_3, \psi_4, \psi_5, \psi_6) = (1/2, 1/2, 1/3, 1/3, 1/3)$.

$(17\%, 29\%, 12\%, 36\%, 24\%)$ across the 3,000 sampled households. We also increase the proportion of missing data to 30% for the remaining variables, except household

Table 3.2: **Empirical study 2:** Confidence intervals for selected probabilities that depend on within-household relationships in the original and imputed datasets. "Sample" is based on the sampled data before introducing errors and missing values, "EIHD" is the EIHD with $(\psi_2, \psi_3, \psi_4, \psi_5, \psi_6) = (1, 1, 1, 1, 1)$, and "EIHD w/caps" is the EIHD with $(\psi_2, \psi_3, \psi_4, \psi_5, \psi_6) = (1/2, 1/2, 1/3, 1/3, 1/3)$. "HH " means household head, "SP" means spouse, "CH" means child, and "CP" means couple. "Q" is the value in the full population of 842,746 households.

	Q	Sample	EIHD	EIHD w/caps
All same race household:				
$n_i = 2$	.942	(.936, .958)	(.880, .919)	(.875, .915)
$n_i = 3$	.908	(.863, .912)	(.798, .872)	(.784, .857)
$n_i = 4$	.901	(.890, .938)	(.797, .879)	(.771, .860)
$n_i = 5$	.887	(.848, .933)	(.730, .861)	(.694, .837)
$n_i = 6$	.871	(.766, .914)	(.622, .851)	(.580, .805)
SP present	.704	(.687, .720)	(.679, .718)	(.678, .720)
Same race CP	.663	(.645, .679)	(.618, .659)	(.611, .655)
SP present, HH is White	.599	(.590, .625)	(.580, .620)	(.579, .621)
White CP	.579	(.572, .607)	(.555, .597)	(.551, .592)
CP with age difference less than five	.494	(.472, .508)	(.348, .388)	(.347, .386)
At least one biological CH present	.490	(.473, .509)	(.484, .524)	(.484, .523)
Male HH, home owner	.472	(.455, .491)	(.443, .487)	(.443, .489)
HH over 35, no CH present	.416	(.397, .432)	(.387, .427)	(.389, .427)
HH older than SP, White HH	.320	(.318, .352)	(.302, .344)	(.301, .343)
Adult female w/ at least one CH under 5	.093	(.078, .098)	(.072, .095)	(.071, .094)
White HH with Hisp origin	.076	(.059, .077)	(.063, .083)	(.060, .083)
Non-White CP, home owner	.063	(.048, .064)	(.035, .053)	(.034, .053)
Two generations present, Black HH	.059	(.043, .059)	(.044, .062)	(.044, .063)
Black HH, home owner	.052	(.039, .053)	(.036, .053)	(.036, .053)
At least three generations present	.041	(.032, .046)	(.034, .052)	(.033, .051)
SP present, HH is Black	.041	(.028, .042)	(.030, .047)	(.031, .048)
At least two generations present, Hisp CP	.040	(.028, .040)	(.027, .041)	(.028, .042)
Hisp CP with at least one biological CH	.038	(.027, .040)	(.025, .039)	(.026, .040)
White-nonwhite CP	.035	(.027, .039)	(.037, .060)	(.043, .066)
One grandchild present	.032	(.021, .032)	(.032, .050)	(.029, .049)
Hisp HH over 50, home owner	.030	(.023, .035)	(.023, .037)	(.023, .038)
Adult Black female w/ at least one CH under 18	.030	(.022, .034)	(.018, .032)	(.018, .032)
Adult Hisp male w/ at least one CH under 10	.027	(.015, .025)	(.014, .026)	(.014, .026)
At least one stepchild	.026	(.021, .032)	(.020, .036)	(.021, .038)
Only one parent	.023	(.015, .025)	(.019, .035)	(.018, .034)
Three generations present, White CP	.013	(.008, .016)	(.008, .019)	(.009, .019)
At least one adopted CH, White CP	.010	(.008, .015)	(.006, .015)	(.006, .016)
Black CP with at least two biological children	.009	(.004, .010)	(.004, .010)	(.004, .010)
Black HH under 40, home owner	.006	(.003, .008)	(.005, .013)	(.005, .014)
White HH under 25, home owner	.003	(.000, .002)	(.002, .010)	(.002, .011)

size. Figures 3.3 and 3.4 display the estimated probabilities obtained from the multiple imputation combining rules for the EIHD with $(\psi_2, \psi_3, \psi_4, \psi_5, \psi_6) = (1, 1, 1, 1, 1)$

and $(\psi_2, \psi_3, \psi_4, \psi_5, \psi_6) = (1/2, 1/2, 1/3, 1/3, 1/3)$, respectively. Table 3.2 displays multiple imputation 95% confidence intervals for the selected probabilities involving within-household relationships, as well as the value in the full population of 842,746 households. The figures and table show results that are slightly degraded for some of the estimands in comparison with those in Section 3.4.1. However, this can be attributed to the high proportion of erroneous households as well as the high proportion of missing data. Here there simply is not enough "uncontaminated data" to properly estimate all estimands.

3.4.3 Empirical study 3: Non-uniform substitution model with fixed error rates, $\rho = 0.2$ and 20% missing data.

Next, we repeat the simulation with $\rho = 0.2$ but deviate from the assumption of uniform substitution rates when generating the data. We make the substitution probabilities q_k in 3.1 and 3.2 non-uniform by setting them proportional to samples from a Gamma$(40, 1)$ distribution when generating the data. However, we still fit 3.1 and 3.2 to the data as before. For gender and age of household head, and for gender, age and relationship to household head for remaining household members, we set the error rates to $\epsilon = (0.65, 0.80, 0.70, 0.85, 0.90)$ as in Section 3.4.1. This results in overall error rates in the five variables of approximately $(16\%, 18\%, 15\%, 18\%, 18\%)$ across the 3,000 sampled households. In the resulting contaminated dataset, the probability of a spouse being wrongly reported as a parent becomes approximately 0.130, whereas the probability of a spouse being wrongly reported as a sibling becomes approximately 0.095, indicating that the substitution probabilities are indeed different from the ones in Section 3.4.1. We also set the proportion of missing data to 20% for the remaining variables, except household size.

Figures 3.5 and 3.6 display the estimated probabilities obtained from the multiple imputation combining rules for the EIHD with $(\psi_2, \psi_3, \psi_4, \psi_5, \psi_6) = (1, 1, 1, 1, 1)$

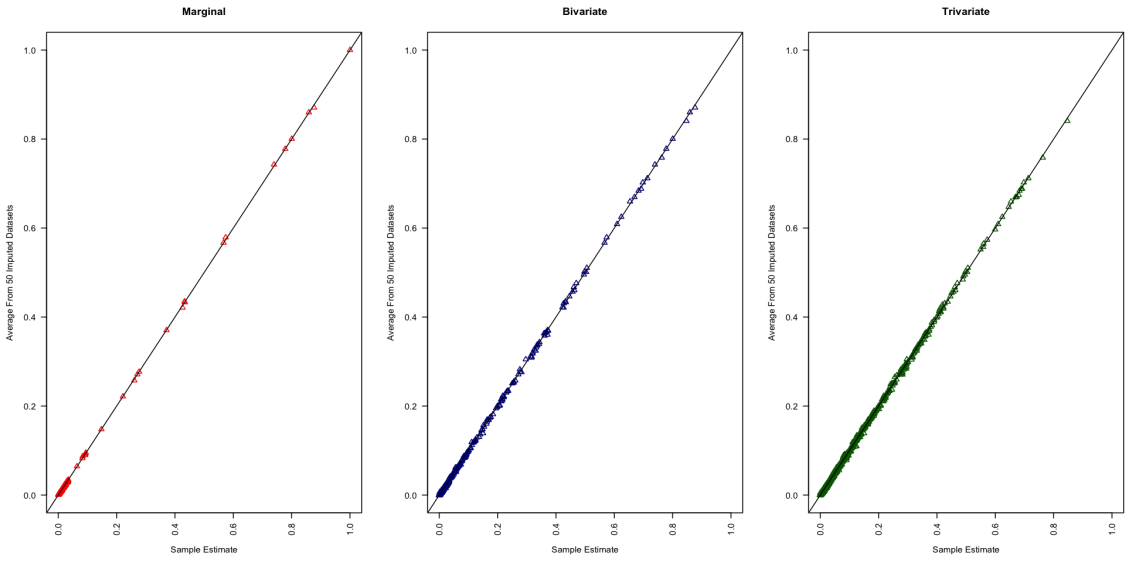


FIGURE 3.5: **Empirical study 3:** Marginal, bivariate and trivariate probabilities computed in the original and imputed datasets from the EIHD with $(\psi_2, \psi_3, \psi_4, \psi_5, \psi_6) = (1, 1, 1, 1, 1)$.

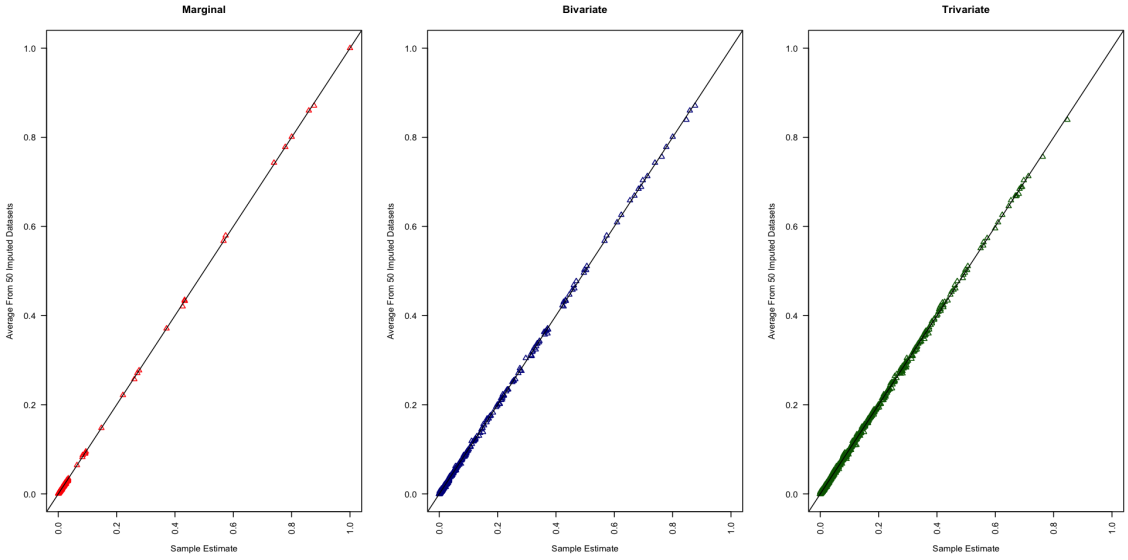


FIGURE 3.6: **Empirical study 3:** Marginal, bivariate and trivariate probabilities computed in the original and imputed datasets from the EIHD with $(\psi_2, \psi_3, \psi_4, \psi_5, \psi_6) = (1/2, 1/2, 1/3, 1/3, 1/3)$.

and $(\psi_2, \psi_3, \psi_4, \psi_5, \psi_6) = (1/2, 1/2, 1/3, 1/3, 1/3)$, respectively. Table 3.3 displays multiple imputation 95% confidence intervals for the selected probabilities involving

Table 3.3: **Empirical study 3:** Confidence intervals for selected probabilities that depend on within-household relationships in the original and imputed datasets. "Sample" is based on the sampled data before introducing errors and missing values, "EIHD" is the EIHD with $(\psi_2, \psi_3, \psi_4, \psi_5, \psi_6) = (1, 1, 1, 1, 1)$, and "EIHD w/caps" is the EIHD with $(\psi_2, \psi_3, \psi_4, \psi_5, \psi_6) = (1/2, 1/2, 1/3, 1/3, 1/3)$. "HH " means household head, "SP" means spouse, "CH" means child, and "CP" means couple. "Q" is the value in the full population of 842, 746 households.

	Q	Sample	EIHD	EIHD w/caps
All same race household:				
$n_i = 2$	.942	(.936, .958)	(.906, .935)	(.903, .935)
$n_i = 3$	.908	(.863, .912)	(.827, .888)	(.821, .884)
$n_i = 4$	.901	(.890, .938)	(.847, .910)	(.840, .908)
$n_i = 5$	.887	(.848, .933)	(.778, .897)	(.770, .885)
$n_i = 6$	.871	(.766, .914)	(.705, .888)	(.702, .887)
SP present	.704	(.687, .720)	(.684, .720)	(.684, .733)
Same race CP	.663	(.645, .679)	(.633, .671)	(.632, .672)
SP present, HH is White	.599	(.590, .625)	(.588, .626)	(.590, .628)
White CP	.579	(.572, .607)	(.567, .606)	(.567, .606)
CP with age difference less than five	.494	(.472, .508)	(.406, .443)	(.407, .444)
At least one biological CH present	.490	(.473, .509)	(.479, .516)	(.477, .516)
Male HH, home owner	.472	(.455, .491)	(.454, .492)	(.455, .492)
HH over 35, no CH present	.416	(.397, .432)	(.391, .428)	(.390, .429)
HH older than SP, White HH	.320	(.318, .352)	(.311, .349)	(.313, .350)
Adult female w/ at least one CH under 5	.093	(.078, .098)	(.077, .099)	(.077, .099)
White HH with Hisp origin	.076	(.059, .077)	(.059, .079)	(.059, .079)
Non-White CP, home owner	.063	(.048, .064)	(.044, .061)	(.042, .059)
Two generations present, Black HH	.059	(.043, .059)	(.045, .062)	(.045, .063)
Black HH, home owner	.052	(.039, .053)	(.038, .054)	(.038, .054)
At least three generations present	.041	(.032, .046)	(.031, .047)	(.031, .047)
SP present, HH is Black	.041	(.028, .042)	(.028, .043)	(.028, .044)
At least two generations present, Hisp CP	.040	(.028, .040)	(.027, .041)	(.027, .041)
Hisp CP with at least one biological CH	.038	(.027, .040)	(.026, .040)	(.026, .040)
White-nonwhite CP	.035	(.027, .039)	(.028, .044)	(.030, .048)
One grandchild present	.032	(.021, .032)	(.023, .038)	(.022, .037)
Hisp HH over 50, home owner	.030	(.023, .035)	(.024, .038)	(.024, .037)
Adult Black female w/ at least one CH under 18	.030	(.022, .034)	(.020, .032)	(.020, .033)
Adult Hisp male w/ at least one CH under 10	.027	(.015, .025)	(.015, .026)	(.015, .026)
At least one stepchild	.026	(.021, .032)	(.020, .033)	(.020, .033)
Only one parent	.023	(.015, .025)	(.017, .030)	(.017, .030)
Three generations present, White CP	.013	(.008, .016)	(.007, .016)	(.007, .016)
At least one adopted CH, White CP	.010	(.008, .015)	(.007, .016)	(.007, .016)
Black CP with at least two biological children	.009	(.004, .010)	(.005, .012)	(.005, .012)
Black HH under 40, home owner	.006	(.003, .008)	(.004, .011)	(.004, .011)
White HH under 25, home owner	.003	(.000, .002)	(.001, .007)	(.001, .007)

within-household relationships, as well as the value in the full population of 842,746 households. The results are also consistent with the results in Sections 3.4.1 and

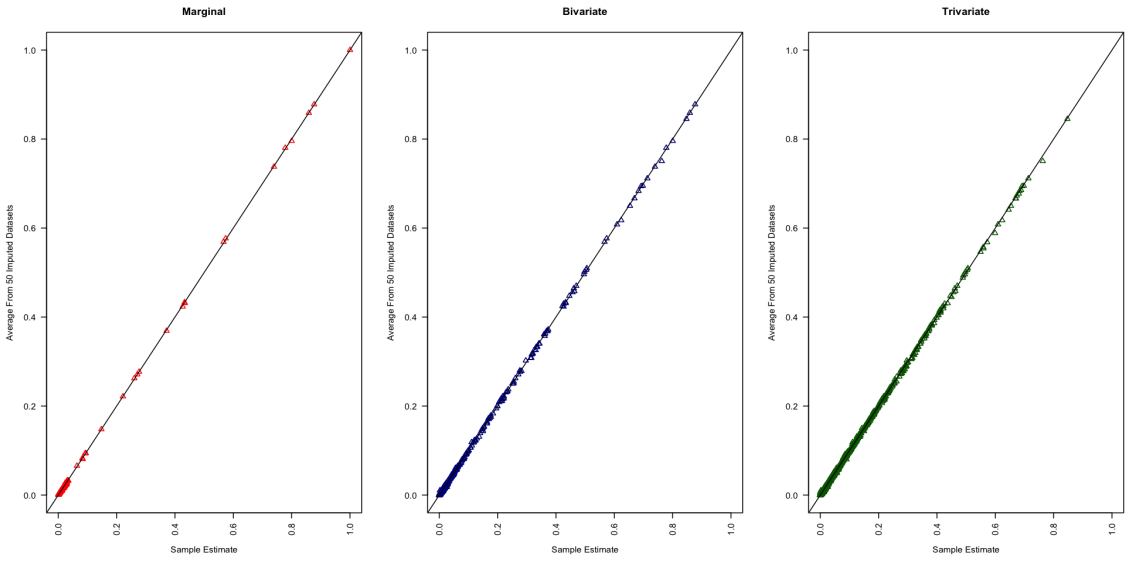


FIGURE 3.7: **Empirical study 4:** Marginal, bivariate and trivariate probabilities computed in the original and imputed datasets from the EIHD with $(\psi_2, \psi_3, \psi_4, \psi_5, \psi_6) = (1, 1, 1, 1, 1)$.

3.4.2; the EIHD again performs well.

3.4.4 Empirical study 4: Non-uniform substitution model with Beta distributed error rates, $\rho = 0.2$ and 20% missing data.

Finally, we repeat the simulation in Section 3.4.3 but we sample error rates per individual for gender and age of household head, and for gender, age and relationship to household head for remaining household members from the Beta$(30, 5)$ distribution. This results in approximately 18% overall error rate for each variable across the 3,000 sampled households. We set the proportion of missing data to 20% for the remaining variables, except household size, as in Section 3.4.3.

Figures 3.7 and 3.8 display the estimated probabilities obtained from the multiple imputation combining rules for the EIHD with $(\psi_2, \psi_3, \psi_4, \psi_5, \psi_6) = (1, 1, 1, 1, 1)$ and $(\psi_2, \psi_3, \psi_4, \psi_5, \psi_6) = (1/2, 1/2, 1/3, 1/3, 1/3)$, respectively. Table 3.4 displays multiple imputation 95% confidence intervals for the selected probabilities involving within-household relationships, as well as the value in the full population of 842,746

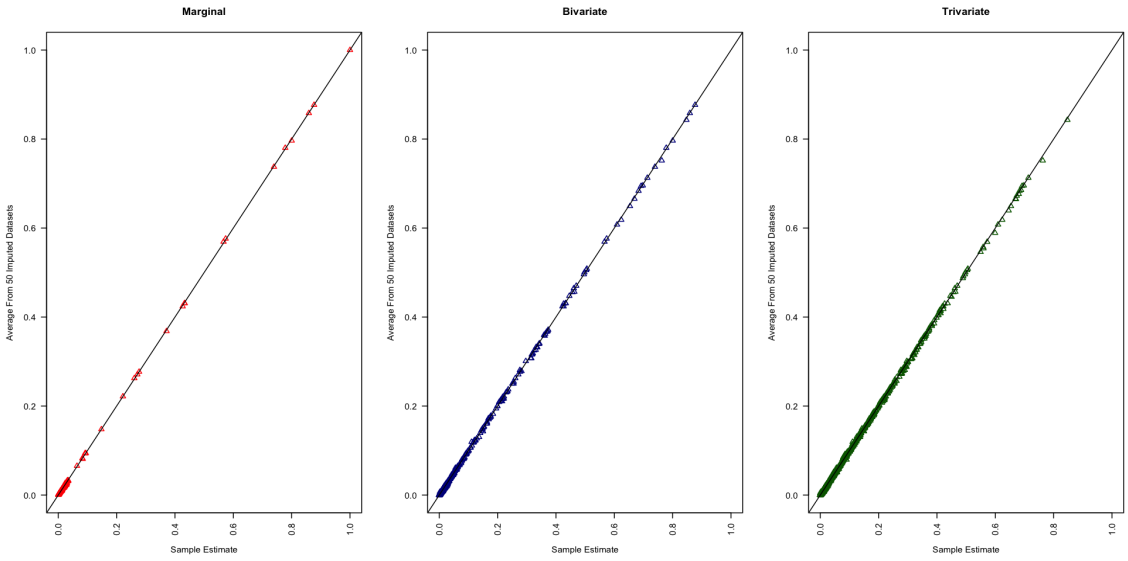


FIGURE 3.8: **Empirical study 4:** Marginal, bivariate and trivariate probabilities computed in the original and imputed datasets from the EIHD with $(\psi_2, \psi_3, \psi_4, \psi_5, \psi_6) = (1/2, 1/2, 1/3, 1/3, 1/3)$.

households. The figures and table show results that are again consistent with the previous results.

Overall, the performance of the EIHD is similar qualitatively across all four simulation studies.

3.5 Discussion

Simultaneously estimating multivariate relationships accurately, capturing within-household relationships, adjusting for measurement errors, and respecting structural zeros in estimation and imputation is a challenging task. The simulation results here suggest that the EIHD does a good job at that task, at least when the measurement error modeling assumptions are approximately true. As with any imputation strategy applied on genuine data, it does not capture all associations perfectly. In particular, quantities that involve many individuals within the same household, such as combinations of races, or that are based on many categories, such as multivariate

Table 3.4: **Empirical study 4:** Confidence intervals for selected probabilities that depend on within-household relationships in the original and imputed datasets. "Sample" is based on the sampled data before introducing errors and missing values, "EIHD" is the EIHD with $(\psi_2, \psi_3, \psi_4, \psi_5, \psi_6) = (1, 1, 1, 1, 1)$, and "EIHD w/caps" is the EIHD with $(\psi_2, \psi_3, \psi_4, \psi_5, \psi_6) = (1/2, 1/2, 1/3, 1/3, 1/3)$. "HH " means household head, "SP" means spouse, "CH" means child, and "CP" means couple. "Q" is the value in the full population of 842,746 households.

	Q	Sample	EIHD	EIHD w/caps
All same race household:				
$n_i = 2$	.942	(.936, .958)	(.901, .934)	(.897, .931)
$n_i = 3$	.908	(.863, .912)	(.824, .888)	(.826, .889)
$n_i = 4$	.901	(.890, .938)	(.809, .883)	(.825, .896)
$n_i = 5$	.887	(.848, .933)	(.726, .854)	(.742, .872)
$n_i = 6$	.871	(.766, .914)	(.601, .812)	(.622, .844)
SP present	.704	(.687, .720)	(.681, .718)	(.680, .716)
Same race CP	.663	(.645, .679)	(.622, .660)	(.620, .660)
SP present, HH is White	.599	(.590, .625)	(.584, .623)	(.585, .623)
White CP	.579	(.572, .607)	(.560, .600)	(.559, .598)
CP with age difference less than five	.494	(.472, .508)	(.397, .433)	(.403, .440)
At least one biological CH present	.490	(.473, .509)	(.474, .512)	(.474, .511)
Male HH, home owner	.472	(.455, .491)	(.450, .487)	(.449, .486)
HH over 35, no CH present	.416	(.397, .432)	(.394, .431)	(.393, .430)
HH older than SP, White HH	.320	(.318, .352)	(.313, .351)	(.310, .349)
Adult female w/ at least one CH under 5	.093	(.078, .098)	(.074, .096)	(.074, .096)
White HH with Hisp origin	.076	(.059, .077)	(.056, .076)	(.055, .075)
Non-White CP, home owner	.063	(.048, .064)	(.040, .057)	(.039, .056)
Two generations present, Black HH	.059	(.043, .059)	(.045, .062)	(.045, .062)
Black HH, home owner	.052	(.039, .053)	(.036, .052)	(.036, .052)
At least three generations present	.041	(.032, .046)	(.032, .047)	(.032, .048)
SP present, HH is Black	.041	(.028, .042)	(.029, .044)	(.029, .044)
At least two generations present, Hisp CP	.040	(.028, .040)	(.028, .042)	(.028, .042)
Hisp CP with at least one biological CH	.038	(.027, .040)	(.026, .040)	(.028, .042)
White-nonwhite CP	.035	(.027, .039)	(.036, .054)	(.037, .056)
One grandchild present	.032	(.021, .032)	(.026, .040)	(.026, .040)
Hisp HH over 50, home owner	.030	(.023, .035)	(.021, .035)	(.022, .036)
Adult Black female w/ at least one CH under 18	.030	(.022, .034)	(.020, .033)	(.020, .032)
Adult Hisp male w/ at least one CH under 10	.027	(.015, .025)	(.016, .027)	(.016, .028)
At least one stepchild	.026	(.021, .032)	(.021, .035)	(.022, .035)
Only one parent	.023	(.015, .025)	(.013, .024)	(.013, .025)
Three generations present, White CP	.013	(.008, .016)	(.008, .017)	(.009, .019)
At least one adopted CH, White CP	.010	(.008, .015)	(.007, .015)	(.007, .016)
Black CP with at least two biological children	.009	(.004, .010)	(.004, .011)	(.005, .011)
Black HH under 40, home owner	.006	(.003, .008)	(.003, .010)	(.003, .010)
White HH under 25, home owner	.003	(.000, .002)	(.001, .007)	(.001, .007)

probabilities involving ages, can be difficult to estimate. In large part this is a result of inadequate sample size for the model to capture such quantities in the larger

households. For example, our sample data contains 1541 households of size two but only 94 households of size six. One possible solution is to redefine variables with many categories to reduce the number of parameters. For example, if acceptable one can use age intervals rather than discrete values of age, or possibly replace age with a variable capturing the difference from the household head's age.

The EIHD can be computationally expensive due to the rejection sampling steps. It can take time to generate feasible imputations for the faulty households, especially when the true error rates are high or when the proportion of households with detectable errors is high. Fortunately, the rejection sampling step can be easily parallelized. This should speed up the sampler by a factor of roughly the number of processors available. The rejection steps are very expensive for large households sizes (e.g., 10 or more people) because the size of $\mathcal{S}_h$ grows exponentially in h. For such households, which usually are not present in large numbers, we suggest exploring ad-hoc versions of the EIHD; for example, imputations for each large household with faulty values can be generated from a fixed set of proposals generated using a variation of hot-deck or cold-deck imputation.

The EIHD model can be used to provide disclosure limitation in public release files. In particular, it can be an engine for creating synthetic data (Raghunathan et al., 2003; Reiter, 2005), enabling agencies to handle the disclosure protection and edit-imputation in one integrated approach (Kim et al., 2015b, 2018). Simply, once the agency has draws of the EIHD model parameters estimated with the faulty data, the agency uses the rejection sampler to generate the synthetic households following the steps in Hu et al. (2018).

The EIHD model also can be used in conjunction with the disclosure control technique PRAM (Gouweleeuw et al., 1998). In PRAM, the agency purposefully introduces measurement errors to categorical values using what is essentially the measurement error model from Section 3.2.2 with fixed ϵ_k. To illustrate an applica-

tion of PRAM, for each individual (ij), we keep each Y_{ijk} at its collected value with probability 0.6 and reset Y_{ijk} to a random draw from the other $d_k - 1$ values in variable k with probability 0.4. Agencies first could use PRAM to perturb confidential values, then use the EIHD synthetic data engine with ϵ_k fixed at the PRAM probabilities (0.6 in this example) and estimated with the perturbed data. The resulting synthetic datasets would satisfy all edit constraints as well as propagate uncertainty due to the perturbations and synthesis. This integration of PRAM and EIHD also generates synthetic household data that satisfy differential privacy (Dwork, 2006), since PRAM applied to all the variables satisfies differential privacy—albeit with a privacy budget that scales with the dimension of the table—and applying the EIHD synthesis engine to the perturbed data is a post-processing step that does not negatively affect the privacy guarantees of PRAM. We leave investigation of the use of EIHD for disclosure limitation as a topic for future research.

Finally, as with all empirical evaluations of new methods, the results presented here are based on limited simulations. In particular, as seen in related work in Kim et al. (2015a), we expect nonignorable missingness or error mechanisms to degrade the performance of the EIHD compared to the presentation here. More informative measurement error models are necessary for any method, including EIHD and Fellegi-Holt approaches, to be effective for such mechanisms. An important future research topic is to incorporate nonignorable measurement errors in the EIHD approach, as well as to assess the sensitivity of inferences from the completed datasets to different specifications of the measurement error models.

4

Leveraging Auxiliary Information on Marginal Distributions in Nonignorable Models for Item and Unit Nonresponse in Surveys

4.1 Introduction

Many surveys, including even the highest quality government surveys, have seen a steep decline in unit and item response rates (Brick and Williams, 2013; Curtin et al., 2005). Recognizing this, government agencies and survey organizations—henceforth all called agencies—often take steps to account for missing values in the data products that they disseminate. Typically, agencies consider item nonresponse and unit nonresponse separately and with distinct statistical approaches. Most commonly, this involves adjusting survey weights to handle unit nonresponse (Brick and Kalton, 1996) and variants of imputation (Andridge and Little, 2010; Kim, 2011; Rubin, 1987) to handle item nonresponse.

Recently, there has been a push among agencies to leverage information in auxiliary data, such as administrative records, to improve the quality of missing data adjustments. For example, in two recent reports, the National Academy of Sciences

recommended that the Census Bureau make greater use of administrative records when accounting for missing values (among other uses) in the Survey of Income and Program Participation (National Research Council, 2009) and the American Community Survey (National Research Council, 2015). It is easy to imagine the potential benefits of leveraging auxiliary information, particularly when using imputation approaches to handle nonresponse. To illustrate, suppose a simple random sample has no unit nonresponse but has item nonresponse on the survey question asking the sex of the respondent. If 70% of participants report male, and we know from auxiliary data that the target population includes 50% men and 50% women, we likely should impute more women than men for the missing values in the survey question asking the sex of the respondent. Of course, we do not want to use solely the population margin; we also should take advantage of observed information in other variables, so as to preserve multivariate relationships as best as possible.

Broadly, auxiliary data for nonresponse adjustments can be classified into two types: sample-based information, which is available for individual sample members, and population-based information, which is available for the target population without being directly linked to the individual sample members. Much of the recent research on the use of auxiliary information for nonresponse adjustment has focused on sample-based auxiliary data. The typical setting has the agency attaching additional variables to individual responses (e.g., Krueger and West, 2014; Sakshaug and Kreuter, 2012; West and Little, 2013). For example, agencies might link survey records to administrative data, either by matching on unique IDs or using probabilistic record linkage techniques (Fellegi and Sunter, 1969; Herzog et al., 2007), to get auxiliary variables for use in imputation modeling. Although there is a growing body of work examining the potential of sample-based auxiliary data, there are only limited opportunities to use such data for effective nonresponse adjustment. Much of the available sample-based auxiliary data are not ideal for such approaches.

Effective survey nonresponse adjustment variables should be highly correlated with both the propensity to respond to a survey and the survey variables of interest, and demographic variables are often insufficient on those grounds (Groves, 2006; Little and Vartivarian, 2005; Peytcheva and Groves, 2009). There are cases in which data rich sample frames are available—for example, some countries have population registries that can be used as a sample frame—but such sample frames are not typically available for U. S. surveys (and even when available, access is often restricted). The accuracy and completeness of sample-based auxiliary data can limit their effectiveness, even more so because this can vary across respondents and nonrespondents (Sinibaldi et al., 2014). The linking of individual survey responses to auxiliary data also can raise questions about the quality of the match, as well as concerns about privacy and informed consent (Sakshaug and Kreuter, 2012). Paradata, which are data about the interview process, offer another potential sample-based auxiliary data source, but extant work finds mixed results about the usefulness of most available paradata (e.g., Biemer et al., 2013; Kreuter et al., 2010; West and Kreuter, 2013).

Instead, we propose to make use of population-based auxiliary data, in particular information on the marginal distributions of subsets of variables in the survey. Population-based auxiliary data can be more readily available than sample-based auxiliary data; for example, state and local government agencies, and even private-sector data aggregators, may be reluctant to share individual-level administrative data with other agencies, but they may be willing to share summary statistics on variables common to their database and the survey. Although population-based auxiliary data are routinely used in post-stratification adjustments for unit nonresponse, survey researchers have done far less work on how to leverage population-based auxiliary data in imputation approaches for handling unit and item nonresponse.

We develop methodology within the framework of multiple imputation (Reiter and Raghunathan, 2007; Rubin, 1987; Schafer, 1997), which is a convenient and flex-

ible approach for creating completed-data analysis files that can be shared with the public. The multiple imputation framework has appealing features for nonresponse adjustment (Alanya et al., 2015; Little and Vartivarian, 2005; Peytchev, 2012). It can handle unit nonresponse and item nonresponse simultaneously, without the typical inflation in variances and often arbitrary decisions about weight trimming that arise with weighting adjustments. One can tailor the imputation models and missingness mechanisms to each variable, whereas weighting adjustments are based on a single set of assumptions about nonresponse that apply for all variables. Finally, the framework facilitates incorporating all sources of uncertainty in inferences.

Our framework for specifying imputation models will leverage auxiliary marginal information. The framework allows distinct specifications of missingness mechanisms for different blocks of variables, e.g., a nonignorable model for variables with auxiliary marginal information and an ignorable model for the variables exclusive to the survey. We fuse features of Bayesian modeling and multiple imputation to propagate uncertainty from not only the missing data, but also from using auxiliary marginal information with non-trivial variance.

We apply the methodology by leveraging auxiliary government election statistics to estimate voter turnout across elections and among population subgroups in the Current Population Survey (CPS). The CPS is the premier data source for voter turnout research, but the survey suffers from considerable item and unit nonresponse that biases turnout estimates. We use the auxiliary data to create multiply-imputed versions of the CPS, which we use for substantive empirical analyses examining the dynamics of voter turnout among various population subgroups (age, sex, and state).

The remainder of this chapter is organized as follows. In section 4.2 we review the additive nonignorable (AN) model of Hirano et al. (1998, 2001) in the context of general missing data problems. We base our methodology on extensions of the AN model. In section 4.3, we present the framework for specifying nonignorable

models for both unit and item nonresponse using information obtained from auxiliary sources. In section 4.4, we present an application of the methods using the 2012 CPS data. In section 4.5, we conclude and discuss some possible extensions of the framework.

4.2 The AN Model

Hirano et al. (2001) developed the AN model in the context of refreshment samples when the primary variable of interest has missing values. Other research, for example Nevo (2003), Bhattacharya (2008), Das et al. (2013), Deng et al. (2013), and Schifeling et al. (2015) have since applied the AN model in different refreshment samples settings. Schifeling et al. (2015) in particular applied the AN model to unit nonresponse but still within the context of refreshment samples. We review the model in the context of item nonresponse in a two-variable missing data example by viewing the information from a refreshment sample as auxiliary data, when such auxiliary information is available.

4.2.1 Notation

Let $\mathcal{D}$ comprise data from the survey of $i = 1, \ldots, n$ individuals, and $\mathcal{A}$ comprise data from the auxiliary database. Let $X = (X_1, \ldots, X_p)$ represent the p variables in both $\mathcal{A}$ and $\mathcal{D}$, where each $X_k = (X_{1k}, \ldots, X_{nk})^T$ for $k = 1, \ldots, p$. Let $Y = (Y_1, \ldots, Y_q)$ represent the q variables in $\mathcal{D}$ but not in $\mathcal{A}$, where each $Y_k = (Y_{1k}, \ldots, Y_{nk})^T$ for $k = 1, \ldots, q$. We disregard variables in $\mathcal{A}$ but not $\mathcal{D}$ as they are not of primary interest. The information in $\mathcal{A}$ can be general; for example, $\mathcal{A}$ might be an individual-level dataset of n_A individuals that provide the joint distribution of X. Throughout this chapter however, we assume that $\mathcal{A}$ only contains sets of marginal distributions/probabilities for variables in X, summarized from some external database.

We introduce indicators to account for nonresponse. For each $k = 1, \ldots, p$,

let $R^x_{ik} = 1$ if individual i would not respond to the question on X_k in survey $\mathcal{D}$ and $R^x_{ik} = 0$ otherwise. Similarly, for each $k = 1, \ldots, q$, let $R^y_{ik} = 1$ if individual i would not respond to the question on Y_k in survey $\mathcal{D}$ and $R^y_{ik} = 0$ otherwise. Let $R^x = (R^x_1, \ldots, R^x_p)$ and $R^y = (R^y_1, \ldots, R^y_q)$ be the vectors of item nonresponse indicators for variables in X and Y respectively, where each $R^x_k = (R^x_{1k}, \ldots, R^x_{nk})^T$, and $R^y_k = (R^y_{1k}, \ldots, R^y_{nk})^T$. Finally, for simplicity, we use generic notations such as f and η for technically different functions and parameters respectively, although their actual meanings should be clear within each context. For example, f, η_0, and η_1 need not be the same in the conditional probability mass functions $\Pr(X_1 = 1|Y_1) = f(\eta_0 + \eta_1 Y_1)$ and $\Pr(Y_1 = 1|X_1) = f(\eta_0 + \eta_1 X_1)$.

4.2.2 Model specification

To make the development easy to follow, we work with an example containing two binary variables, where there are no unit nonrespondents in the data. Suppose we have two binary variables X_1 and Y_1. Following our notation, X_1 and Y_1 are contained in $\mathcal{D}$, but we only have the auxiliary marginal distribution for X_1 from $\mathcal{A}$, and no auxiliary information for Y_1. Also, suppose X_1 suffers from item nonresponse but Y_1 is fully observed, so that R^x_1 is the fully observed vector of item nonresponse indicators for X_1. There is no need to include a model for R^y_1 since there is no nonresponse in Y_1. We also note that there is no need to include auxiliary margins for fully observed variables in our framework. Auxiliary information containing only the marginal distribution of a fully observed variable will not necessarily increase the number of estimable parameters.

The observed data for our two-variable example, along with the auxiliary data, takes the form shown in Table 4.1a, where "✓" represents the observed components and "**?**" represents the missing components. The incomplete contingency table representing the joint distribution of (X_1, Y_1, R^x_1), with observed marginal probabilities

excluded, is shown in Table 4.1b. Tables 4.1a and 4.1b both visualize the extent of the sparsity in the full joint distribution. We first show how identifiability restrictions influences model selection using similar arguments to those of Hirano et al. (1998, 2001), Deng et al. (2013) and Schifeling et al. (2015). We then proceed to show how to leverage the marginal information about X_1 to specify a more flexible nonignorable model (the AN model) than would have been otherwise possible.

The joint distribution of (X_1, Y_1, R_1^x) can be fully described by the eight cell probabilities $\Pr(X_1 = x, Y_1 = y, R_1^x = r)$ for $x, y, r \in \{0, 1\}^3$. We can factorize the joint distribution as

$$\begin{aligned}\Pr(X_1 = x, Y_1 = y, R_1^x = r) = \Pr(X_1 = x | Y_1 = y, R_1^x = r) \\ \times \Pr(Y_1 = y | R_1^x = r) \Pr(R_1^x = r).\end{aligned} \tag{4.1}$$

This factorization, known as a pattern mixture model factorization (Glynn et al., 1986; Little, 1993a), can be fully parameterized using seven parameters: $\theta_{yr} = \Pr(X_1 = 1 | Y_1 = y, R_1^x = r)$, $\pi_r = \Pr(Y_1 = 1 | R_1^x = r)$ and $q = \Pr(R_1^x = 1)$. Here, q, π_r for $r \in \{0, 1\}$, and θ_{y0} for $y \in \{0, 1\}$, can be directly estimated from the observed data alone, as long as the data is a large representative sample of the underlying population. Thus, we can uniquely estimate five of the seven parameters. Unfortunately, we cannot learn the values of θ_{01} and θ_{11} since the observed data alone contain no information about them.

Although θ_{01} and θ_{11} are not identifiable from the observed data, the auxiliary information about the marginal distribution of X_1 does provide a constraint about both parameters to help identify either one of them or some combination of both. This auxiliary margin restricts the possible values of the unobserved joint cell probabilities in the contingency table and consequently, adds the following linear constraint on both θ_{01} and θ_{11}.

$$\Pr(X_1 = 1) - \Pr(X_1 = 1, Y_1 = y, R_1^x = 0) = q\left[\theta_{01}(1 - \pi_1) + \theta_{11}\pi_1\right], \tag{4.2}$$

Table 4.1: Two binary variables Y_1 and X_1: Y_1 is fully observed, X_1 suffers from item nonresponse and the data contains no unit nonrespondents.

(a) Data

	X_1	Y_1	R_1^x
Original data	✓	✓	0
	?		1
Auxiliary margin →	✓	?	?

(b) Contigency table

	$R_1^x = 0$		$R_1^x = 1$	
	$X_1 = 0$	$X_1 = 1$	$X_1 = 0$	$X_1 = 1$
$Y_1 = 0$	✓	✓	?	?
$Y_1 = 1$	✓	✓	?	?

where the probabilities on the left side of (4.2) can be estimated from the observed data. One linear constraint isn't enough to learn about both θ_{01} and θ_{11}. However, the linear constraint does increase the number of estimable parameters from five to six. Therefore, we need to either add an additional constraint to achieve full parametric specification or specify a joint model with at most six parameters.

There are a few potential models/mechanisms within the missing data literature for specifying the additional constraint. One option is MAR which sets $\theta_{01} = \theta_{00}$ and $\theta_{11} = \theta_{10}$. This allows for the estimation of all seven parameters. Only five of those are unique, and MAR does not maximize all the information available; it only uses the information available from the observed data alone and does not take advantage of the auxiliary margin. A flexible approach should allow for the estimation of six unique parameters to maximize all available information.

A second option is to set either θ_{01} or θ_{11} equal to zero, but both constraints imply very strong assumptions. Consider an example where X_1 represents sex (where 0 = male and 1 = female) and Y_1 represents age (where 0 = less than 18 and 1 = 18+). Then setting $\theta_{01} = 0$ and $\theta_{11} \neq 0$ implies that all nonrespondents cannot have an X_1 value equal to one whenever $Y_1 = 0$. In the context of our age and sex example, $\theta_{01} = 0$ and $\theta_{11} \neq 0$ implies that all nonrespondents younger than 18 years must be

male. Similarly, setting $\theta_{11} = 0$ but $\theta_{01} \neq 0$ implies that all nonrespondents cannot have an X_1 value equal to one whenever $Y_1 = 1$. That is, all nonrespondents older than 18 years must be male. Both constraints seem rather restrictive and we do not recommend adopting them unless the specific application at hand demands and justifies such strong assumptions.

A third option is to set $\theta_{01} = \theta_{11} + b$ for some constant b or quite simply, to set $\theta^\star = \theta_{01} = \theta_{11}$ with $b = 0$, which then simplifies (4.2) to

$$\theta^\star = \frac{\Pr(X_1 = 1) - (1 - q)\left[\theta_{00}(1 - \pi_0) + \theta_{10}\pi_0\right]}{q}. \tag{4.3}$$

To further understand this constraint, suppose for example that we use the following logistic regression model to characterize $\Pr(X_1 = 1|Y_1 = y, R_1^x = r)$.

$$X_{i1}|Y_{i1}, R_{i1}^x \sim \ Bern(\pi_{X_{i1}}); \ \ logit(\pi_{X_{i1}}) = \alpha_0 + \alpha_1 Y_{i1} + \alpha_2 R_{i1}^x + \alpha_3 Y_{i1} R_{i1}^x. \tag{4.4}$$

It is easy to show that the constraint $\theta^\star = \theta_{01} = \theta_{11}$ implies that $\alpha_3 = -\alpha_1$ in (4.4), which reduces (4.4) to

$$X_{i1}|Y_{i1}, R_{i1}^x \sim \ Bern(\pi_{X_{i1}}); \ \ logit(\pi_{X_{i1}}) = \alpha_0 + \alpha_1 Y_{i1}(1 - R_{i1}^x) + \alpha_2 R_{i1}^x. \tag{4.5}$$

This constraint implies conditional independence between Y_1 and X_1 for nonrespondents (when $R_1^x = 1$) which is still a strong assumption, and adopting it would require practical justifications.

All three options, as well as other options which we do not mention here, are either restrictive or do not maximize all available information. The AN model on the other hand provides a more flexible option that maximizes the available information without forcing analysts to make as many untestable assumptions as we have to make under most of the other mechanisms/model specifications. Writing the AN model in the form suggested by Hirano et al. (2001) requires a selection model factorization

(Little, 1995) of the joint distribution of (Y_1, X_1, R_1^x), instead of the pattern mixture factorization in (4.1). We write this as

$$(X_1, Y_1) \sim f(X_1, Y_1|\Theta) \tag{4.6}$$

$$\Pr(R_1^x = 1|X_1, Y_1) = h_1(\eta_0 + \eta_1 X_1 + \eta_2 Y_1), \tag{4.7}$$

where Θ and (η_0, η_1, η_2) represents the set of parameters in f and h_1 respectively, and $h_1(a)$ is a strictly increasing function satisfying $\lim_{a\to-\infty} h_1(a) = 0$ and $\lim_{a\to\infty} h_1(a) = 1$. Hirano et al. (2001) prove that the AN model is likelihood-identified for general distributions. Common examples of functions which satisfy the conditions on $h_1(a)$ include the logit and probit link functions.

The AN model assumes that the reason for item nonresponse in X_1 depends on X_1 and Y_1 through a function that is additive in X_1 and Y_1; that is, an interaction term between X_1 and Y_1 is not allowed, as additivity is in fact necessary to enable identification of the model parameters. Additivity of the model in X_1 and Y_1 may be reasonable in practical contexts. Also, special cases of the model are informative. For example, $(\eta_1 = 0, \eta_2 = 0)$ results in an MCAR mechanism, $(\eta_1 \neq 0, \eta_2 = 0)$ results in an MAR mechanism, whereas $\eta_2 \neq 0$ results in an MNAR mechanism. In fact, $(\eta_1 = 0, \eta_2 \neq 0)$ in particular results in the nonignorable model of Hausman and Wise (1979); henceforth, we refer to this specification as the HW model. For more details regarding the AN model, see Hirano et al. (1998, 2001). For comparisons between the AN, MAR and HW models as well as simulations exploring the bias due to an interaction term, when the true nonresponse mechanism includes the interaction, see Deng et al. (2013).

The AN model provides a flexible framework that allows the data determine the appropriate mechanism in similar scenarios, so that analysts can avoid making an untestable choice between MCAR, MAR, and MNAR. While the AN model can be very useful when dealing with one nonresponse mechanism, it is not directly obvious

how the assumptions of the AN model need to change when item nonresponse occurs for more than one variable, when unit and item nonresponse occur simultaneously, or when more marginal information exists. Our framework builds on and extends the AN model to address these shortcomings.

4.3 The SCINN Framework

Our approach builds on and combines nonignorable nonresponse mechanisms — such as the AN model, the itemwise conditionally independent nonresponse model of Sadinle and Reiter (2017), which we henceforth refer to as ICIN, and MNAR — with ignorable nonresponse mechanisms — such as MAR and MCAR. Specifically, we propose characterizing the joint distribution of all variables in $\mathcal{D}$ as well as the nonresponse indicators using a sequential factorization of conditional distributions, and specifying an identifiable conditional model for each conditional distribution, where the models for the unit and item nonresponse indicators are some combination of nonignorable and ignorable models. Sadinle and Reiter (2019) present various identification results for such sequential factorizations and we leverage those results here.

Our framework is based on a step-by-step construction for specifying identifiable models that maximize all available information, including both the observed data and auxiliary margins. First, we begin by specifying an identifiable model for the observed data alone. Since nonignorable nonresponse models are often unidentifiable without external information, this first step usually results in ignorable models, such as MCAR and MAR, for all the nonresponse indicators. In this chapter, we specifically build the observed data using a sequence of identifiable conditional models. This step also often results in default choices for handling nonresponse in the missing data literature. Next, we add more estimable parameters to the observed data model using the available auxiliary margins. Each auxiliary margin allows for the

estimation of a parameter related directly to the corresponding variable. Therefore, this step involves identifying the set of parameters that can be estimable from each margin, and making a decision about which ones to include. We present the ideas within our framework by first rethinking the construction the AN model. We then proceed to show how to implement our framework using three simple scenarios.

As we show later in this section, the number of parameters estimable in the models depends on the observed information and the amount of information in $\mathcal{A}$, just as is the case with the AN model. We develop and demonstrate our framework using two binary variables, as we did in Section 4.2.2, as a running example across different scenarios; we discuss extensions to more variables at the end of the section. In each scenario, we first show how identifiability restrictions influence model selection using similar arguments to those in Section 4.2.2.

4.3.1 Rethinking the construction of the AN model

We first begin by rethinking and re-presenting the AN model within our framework. Our framework, which we refer to as SCINN (sequential conditional ignorable and nonignorable nonresponse models), can be used to build flexible identifiable models using ideas from the AN model. Broadly, SCINN is based on the following set of modeling steps/choices.

Step 1: Observed data model. Write down a model identifiable from the observed data alone. Preferably, the model should allow for the maximum number of parameters identifiable from the observed data alone, without any auxiliary information. This can be done using either a selection model or a pattern mixture specification, and analysts can decide based on preference.

For the scenario in Section 4.2.2 for example, the most obvious choice for a selection model is the model that excludes parameters capturing the relationship between X_1 and R_1^x. Since X_1 and R_1^x are never fully observed jointly, the relationship be-

tween them cannot be estimated from the observed data alone without any additional information or constraints. The most obvious choice is thus

$$(X_1, Y_1) \sim f(X_1, Y_1|\Theta) \tag{4.8}$$

$$\Pr(R_1^x = 1|X_1, Y_1) = h_1(\eta_0 + \eta_1 Y_1), \tag{4.9}$$

which results in an MAR model with up to five unique parameters, the maximum possible based on the observed data alone. η_0 is identifiable since the marginal probability of R_1^x is known from the observed data alone, and η_1 is identifiable since the joint probabilities between R_1^x and Y_1 are also known from the observed data alone. Given the conditional independence assumption between R_1^x and X_1 in (4.9), Θ would also be identifiable for common choices of f, such as the multinomial distribution or a sequence of logistic models. Note that the HW model is not an option here because the relationship between R_1^x and X_1 is not estimable from the observed data alone. An MCAR model could be an option, but we would not recommend it since it only allows for up to four unique parameters.

Similarly, should analysts prefer a pattern mixture factorization, the most obvious specification is

$$(Y_1, R_1^x) \sim f(Y_1, R_1^x|\Phi) \tag{4.10}$$

$$\Pr(X_1 = 1|Y_1, R_1^x) = h_1(\eta_0 + \eta_1 Y_1), \tag{4.11}$$

which also has up to five unique parameters. Again, we must exclude parameters capturing the relationship between X_1 and R_1^x. As discussed in Section 4.2.2, η_0, η_1, and Φ would clearly be identifiable from the observed data alone, especially for common choices of f.

Step 2: Incorporate auxiliary margins. Use each auxiliary margin to estimate one main effects parameter relating directly to the variable with the margin (each margin by itself does not contain information about other variables and there-

fore must be used for the corresponding variable), in addition to the model already chosen in Step 1.

For the selection model factorization in (4.8) and (4.9) for example, that implies using the auxiliary margin $\Pr(X_1 = 1)$ of X_1 to add $\eta_2 X_1$ to (4.9), resulting in the AN model in (4.6) to (4.7) with up to six identifiable parameters. As discussed in Section 4.2.2, we can only estimate as many extra parameters, in addition to Step 1, as the auxiliary information available. For example, auxiliary information about $\Pr(X_1 = 1)$ alone in this scenario allows for the estimation of $\eta_2 X_1$, whereas auxiliary joint distribution $\Pr(X_1, Y_1)$ of (X_1, Y_1) would allow for the estimation of both $\eta_1 X_1$ and $\eta_3 X_1 Y_1$. If an analyst instead prefers the pattern mixture factorization in (4.10) and (4.11), then the corresponding action is to add $\eta_2 R_1^x$ to (4.11), resulting in

$$\Pr(X_1 = 1|Y_1, R_1^x) = h_1(\eta_0 + \eta_1 Y_1 + \eta_2 R_1^x). \tag{4.12}$$

Step 1 alone often results in models that would be the default choices in the missing data literature, without using auxiliary data to identify more parameters. The presence of auxiliary information in the form of marginal distributions alone improves upon those default choices. Unless the auxiliary information available encodes joint distributions instead of marginal distributions, we cannot estimate interactions involving nonresponse indicators and the variables of interest. Therefore, we make the same assumptions about $h_1(a)$ as the AN model. That is, (i) $h_1(a)$ is a strictly increasing function satisfying $\lim_{a\to-\infty} h_1(a) = 0$, $\lim_{a\to\infty} h_1(a) = 1$, and (ii) interaction terms are not allowed in h_1.

The steps above are quite intuitive and provide direction for extending the AN model. In particular, when the data includes unit nonrespondents or item nonresponse in more than one variable, Step 2 provides flexibility in how to use the auxiliary information. We now show to use this framework for more general cases than the AN model, to allow for flexibility in modeling nonresponse in scenarios

with unit nonresponse or item nonresponse in more than one variable. We show how these steps in the framework can provide more flexibility than current methods when modeling different combinations of item and unit nonresponse. To make the framework easy to follow, we focus on three scenarios which are extensions of the two-variable scenario in Section 4.2.2.

4.3.2 Two variables suffering from item nonresponse but no unit nonrespondents

First we extend the scenario in Section 4.2.2 so that both variables suffer from item nonresponse. We assume the data does not contain any unit nonrespondents; we include unit nonrespondents in Sections 4.3.3 and 4.3.4. Suppose X_1 and Y_1 are defined in the same way as in Section 4.2.2, where X_1 again suffers from item nonresponse and R_1^x is defined as before, but Y_1 now suffers from item nonresponse, so that R_1^y is the fully observed vector of item nonresponse indicators for Y_1. Again, we have the auxiliary margin for X_1 but no auxiliary information for Y_1. The observed data, along with the auxiliary data, takes the form shown in Table 4.2a and the incomplete contingency table representing only the joint distribution of (X_1, Y_1, R_1^x, R_1^y) is shown in Table 4.2b. Again we have a sparse contingency table as we only observe four out of 16 possible cell probabilities.

As before, we first write the joint distribution of (X_1, Y_1, R_1^x, R_1^y), fully described by the 16 cell probabilities $\Pr(X_1 = x, Y_1 = y, R_1^x = r^x, R_1^y = r^y)$ for $x, y, r^x, r^y \in \{0, 1\}^4$, as the following pattern mixture factorization

$$\begin{aligned}\Pr(X_1 = x, Y_1 = y, R_1^x = r^x, R_1^y = r^y) &= \Pr(X_1 = x | Y_1 = y, R_1^x = r^x, R_1^y = r^y) \\ &\quad \times \Pr(Y_1 = y | R_1^x = r^x, R_1^y = r^y) \\ &\quad \times \Pr(R_1^x = r^x | R_1^y = r^y) \Pr(R_1^y = r^y).\end{aligned} \tag{4.13}$$

We can fully parameterize this factorization using $\theta_{yr^xr^y} = \Pr(X_1 = 1 | Y_1 = y, R_1^x = r^x, R_1^y = r^y)$, $\pi_{r^xr^y} = \Pr(Y_1 = 1 | R_1^x = r^x, R_1^y = r^y)$, $q_{r^y} = \Pr(R_1^x = 1 | R_1^y =$

Table 4.2: Two binary variables X_1 and Y_1: both suffer from item nonresponse and the data contains no unit nonrespondents.

(a) Data

	X_1	Y_1	R_1^x	R_1^y
Original data	✓	✓	0	0
	?	✓	1	0
	✓	?	0	1
	?	?	1	1
Auxiliary margin →	✓	?	?	?

(b) Contigency table

		$R_1^y = 0$		$R_1^y = 1$	
		$Y_1 = 0$	$Y_1 = 1$	$Y_1 = 0$	$Y_1 = 1$
R_1^x	$X_1 = 0$	✓	✓	?	?
	$X_1 = 1$	✓	✓	?	?
R_1^x	$X_1 = 0$	?	?	?	?
	$X_1 = 1$	?	?	?	?

r^y) and $p = \Pr(R_1^y = 1)$ resulting in a total of 15 parameters. Here, p, q_{r^y} for $r^y \in \{0, 1\}$, $\pi_{r^x 0}$ for $r^x \in \{0, 1\}$, and θ_{y00} for $y \in \{0, 1\}$, can be estimated from the observed data alone, so that we can estimate seven of the 15 parameters directly from the data. With the auxiliary margin $\Pr(X_1 = 1)$ of X_1, we have the following linear constraint on the eight unobserved parameters

$$\begin{aligned} &\Pr(X_1 = 1) - \Pr(X_1 = 1, Y_1 = y, R_1^x = 0, R_1^y = 0) \\ &\qquad = pq_1 \left[\theta_{011}(1 - \pi_{11}) + \theta_{111}\pi_{11}\right] \\ &\qquad + p(1 - q_1) \left[\theta_{001}(1 - \pi_{01}) + \theta_{101}\pi_{01}\right] \\ &\qquad + (1 - p)q_0 \left[\theta_{010}(1 - \pi_{10}) + \theta_{110}\pi_{10}\right], \end{aligned} \tag{4.14}$$

allowing us to estimate one more parameter, resulting in a total of eight estimable parameters. In other to specify a model with eight identifiable parameters, we use our framework to build a sequence of identifiable models.

We use our framework to specify a selection model. First we follow step Step 1 in Section 4.3 and specify a model with the maximum number of parameters identifiable from the observed data alone. Since the maximum number of identifiable parameters

without auxiliary data is seven, the most obvious specification is

$$(X_1, Y_1) \sim f(X_1, Y_1|\Theta) \tag{4.15}$$

$$\Pr(R_1^x = 1|X_1, Y_1) = h_1(\eta_0 + \eta_1 Y_1) \tag{4.16}$$

$$\Pr(R_1^y = 1|X_1, Y_1, R_1^x) = k_1(\zeta_0 + \zeta_1 X_1) \tag{4.17}$$

where $f(X_1, Y_1|\Theta)$ has three parameters since that is the number of parameters needed to fully estimate the joint distribution of X_1 and Y_1. As another option, it is also possible to replace (4.17) with

$$\Pr(R_1^y = 1|X_1, Y_1, R_1^x) = k_1(\zeta_0 + \zeta_2 R_1^x) \tag{4.18}$$

resulting in a slightly different variation of the model within the same framework. In practice however, it is often more interesting to investigate and usually more plausible to assume that item nonresponse for a variable depends on either the variable itself or other variables in the data, rather than other nonresponse indicators. Therefore, due to the bound on the number of parameters estimable, we will for the most part assume that nonresponse indicators are conditionally independent given all other variables. Although we adopt this assumption for most of this chapter, clearly analysts can choose still (4.18) instead of (4.17) should they prefer; that is part of the flexibility in options our framework offers.

Next, we follow Step 2, which implies that we can only add a term involving X_1 since this is the variable we have auxiliary information on. Since (4.17) already has an X_1 term, we add $\eta_2 X_1$ to (4.16), so that (4.16) becomes

$$\Pr(R_1^x = 1|X_1, Y_1) = h_1(\eta_0 + \eta_1 Y_1 + \eta_2 X_1) \tag{4.19}$$

and our full model includes (4.15), (4.17) and (4.19).

Alternatively, should an analyst prefer (4.18) to (4.17) as mentioned before, then Step 2 gives us two options: (i) add $\zeta_2 X_1$ to (4.18), so that our nonresponse models becomes a combination of (4.16) and

$$\Pr(R_1^y = 1|X_1, Y_1, R_1^x) = k_1(\zeta_0 + \zeta_1 X_1 + \zeta_2 R_1^x); \tag{4.20}$$

or (ii) add $\eta_2 X_1$ to (4.16) as before, resulting in (4.18) and (4.19) as the nonresponse models. In all cases, we are able to specify models with eight identifiable parameters as desired.

Note that (4.19) is the AN model discussed in Section 4.2.2, with the caveat that Y_1 now suffers from item nonresponse but the parameters are still identifiable using the same argument. Here, (4.17), (4.18) and (4.20) are all special cases of the ICIN mechanism of Sadinle and Reiter (2017) who show that the parameters are identifiable under the scenario we have here. In particular (4.18) can be viewed as MAR, since R_1^x is fully observed, whereas (4.17) and (4.20) are not MAR because they depend on X_1 which suffers from item nonresponse. MAR models should depend only on fully observed variables and not values of other variables that are themselves missing. However, MAR can be viewed as a special case of the ICIN mechanism in this particular setup (Sadinle and Reiter, 2017). To the best of our knowledge, AN and ICIN mechanisms have not been combined when specifying models for handling missing data.

The same procedure followed here can used to specify a pattern mixture model here instead. Specifically, Step 1 gives us

$$(R_1^y, R_1^x) \sim f(R_1^y, R_1^x|\Phi) \tag{4.21}$$

$$\Pr(Y_1 = 1|R_1^y, R_1^x) = h_1(\eta_0 + \eta_2 R_1^x) \tag{4.22}$$

and either

$$\Pr(X_1 = 1|R_1^y, R_1^x, Y_1) = k_1(\zeta_0 + \zeta_1 R_1^y) \tag{4.23}$$

or

$$\Pr(X_1 = 1|R_1^y, R_1^x, Y_1) = k_1(\zeta_0 + \zeta_2 Y_1). \tag{4.24}$$

We note that ζ_2 in (4.24) is identifiable because of the conditional independence between X_1 and R_1^x implied in (4.24). As before, Step 2 then gives us multiple

options on the extra parameter to estimate. We suggest using the auxiliary margin to estimate $\zeta_2 R_1^x$ in (4.24) to allow for a nonignorable mechanism. We do not fully explore the different modeling options under this pattern mixture factorization as we use a selection model to continue to emphasize the connection to the AN model.

We now use simulation studies to illustrate how each option under the selection model factorization affects the estimation of the joint relationship between X_1 and Y_1. We also use the logistic regression as a default choice for the functions f, h_1 and k_1. The conclusions presented here extend to other choices such as the probit function. We simulate data containing $n = 5000$ individuals from the following sequence of logistic models.

$$X_{i1} \sim Bern(\pi_{X_{i1}}); \quad logit(\pi_{X_{i1}}) = \alpha_0 \tag{4.25}$$

$$Y_{i1}|X_{i1} \sim Bern(\pi_{Y_{i1}}); \quad logit(\pi_{Y_{i1}}) = \beta_0 + \beta_1 X_{i1} \tag{4.26}$$

where $\alpha_0 = 0.65$ and $(\beta_0, \beta_1) = (0.5, -1)$.

We then generate missing data using the following five nonresponse mechanisms previously discussed.

1. The ICIN models in (4.16) and (4.17) which we refer to as ICIN and specify using the following logistic regressions

$$R_{i1}^x|X_{i1}, Y_{i1} \sim Bern(\pi_{R_{i1}^x}); \; logit(\pi_{R_{i1}^x}) = \eta_0 + \eta_1 Y_{i1} \tag{4.27}$$

$$R_{i1}^y|X_{i1}, Y_{i1}, R_{i1}^x \sim Bern(\pi_{R_{i1}^y}); \; logit(\pi_{R_{i1}^y}) = \zeta_0 + \zeta_1 X_{i1}, \tag{4.28}$$

   where $(\eta_0, \eta_1) = (-0.4, -1.1)$ and $(\zeta_0, \zeta_1) = (-0.6, -0.4)$.

2. The ICIN plus MAR models in (4.16) and (4.18) which we refer to as ICIN+MAR and specify using (4.27) and

$$R_{i1}^y|X_{i1}, Y_{i1}, R_{i1}^x \sim Bern(\pi_{R_{i1}^y}); \; logit(\pi_{R_{i1}^y}) = \zeta_0 + \zeta_2 R_{i1}^x, \tag{4.29}$$

   where $(\zeta_0, \zeta_2) = (-0.9, 0.3)$.

3. The AN plus ICIN models in (4.17) and (4.19) which we refer to as SCINN1 and specify using

$$R^x_{i1}|X_{i1}, Y_{i1} \sim Bern(\pi_{R^x_{i1}});\ \ logit(\pi_{R^x_{i1}}) = \eta_0 + \eta_1 Y_{i1} + \eta_2 X_{i1} \tag{4.30}$$

and (4.28), where $(\eta_0, \eta_1, \eta_2) = (-0.4, 0.25, -0.95)$.

4. The ICIN models in (4.16) and (4.20) which we refer to as SCINN2 and specify using (4.27) and

$$R^y_{i1}|X_{i1}, Y_{i1}, R^x_{i1} \sim Bern(\pi_{R^y_{i1}});\ \ logit(\pi_{R^y_{i1}}) = \zeta_0 + \zeta_1 X_{i1} + \zeta_2 R^x_{i1}, \tag{4.31}$$

where $(\zeta_0, \zeta_1, \zeta_2) = (-0.2, -1.40, -0.65)$.

5. The AN plus MAR models in (4.18) and (4.19) which we refer to as SCINN3 and specify using (4.29) and (4.30).

In each simulation, we fit the true generating model for X_1 and Y_1 (that is, (4.25) and (4.26)) plus all five nonresponse models back to the generated data. In each case, we re-estimate parameters to examine how accurately each model estimates the joint distribution of X_1 and Y_1. We fit all models using Bayesian MCMC with non-informative priors for all parameters, to appropriately capture uncertainty; in our CPS application in Section 4.4, we combine Bayesian inference with multiple imputation. We run the MCMC sampler for 10,000 iterations, discarding the first 5,000 as burn-in, resulting in 5,000 posterior samples. We set the values for the parameters in all five models so that we have approximately 30% missing data in each of X_1 and Y_1. We select 30% as a default and moderate level of missing data.

To incorporate the marginal information contained in $\mathcal{A}$ in our simulations and illustrations, we follow the approach of Schifeling and Reiter (2016). We augment the observed data with synthetic observations so that the empirical distribution of each variable in X matches the margins in A, with the remaining variables left completely missing for the synthetic observations. We also treat the auxiliary margins available

Table 4.3: Posterior means and standard errors for simulation study when all five nonignorable models are fitted to data generated under ICIN and ICIN+MAR. "Par" is parameter.

				Fitted Model									
				ICIN		ICIN+MAR		SCINN1		SCINN2		SCINN3	
		Par.	Truth	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
Generative Model	ICIN	α_0	.65	.66	.03	.66	.04	.62	.01	.63	.01	.62	.02
		β_0	.50	.64	.08	.64	.07	.61	.08	.65	.07	.61	.08
		β_1	-1.00	-1.13	.10	-1.12	.09	-1.09	.10	-1.13	.09	-1.10	.09
		η_0	-.40	-.39	.04	-.39	.04	-.28	.11	-.40	.04	-.26	.12
		η_1	-1.10	-1.04	.08	-1.07	.08	-1.08	.09	-1.03	.08	-1.11	.09
		η_2	—	—	—	—	—	-.16	.14	—	—	-.17	.14
		ζ_0	-.60	-.63	.06	-.90	.04	-.64	.06	-.66	.06	-.90	.04
		ζ_1	-.40	-.38	.08	—	—	.38	.08	-.39	.08	—	—
		ζ_2	—	—	—	.06	.07	—	—	.08	.07	.06	.07
	ICIN+ MAR	α_0	.65	.66	.04	.66	.04	.62	.02	.62	.01	.62	.02
		β_0	.50	.57	.07	.58	.07	.55	.07	.57	.07	.55	.07
		β_1	-1.00	-1.05	.09	-1.06	.09	-1.03	.09	-1.04	.08	-1.03	.09
		η_0	-.40	-.39	.04	-.39	.04	-.31	.11	-.39	.04	-.31	.11
		η_1	-1.10	-1.06	.08	-1.06	.08	-1.09	.09	-1.05	.08	-1.08	.09
		η_2	—	—	—	—	—	-.11	.14	—	—	-.11	.14
		ζ_0	-.90	-.87	.06	-.97	.04	-.86	.06	-.99	.07	-.97	.04
		ζ_1	—	-.07	.08	—	—	.05	.08	.03	.09	—	—
		ζ_2	.30	—	—	.44	.07	—	—	.44	.07	.44	.07

about X as having negligible standard error. When using synthetic observations to incorporate the auxiliary information, it is straightforward to generate synthetic data that accounts for the standard errors. See Schifeling and Reiter (2016) for more details on this approach. As Schouten et al. (2018) showed, analysts can also incorporate auxiliary data directly in prior specification when using Bayesian methods to handle nonresponse in surveys. R_1^x and R_1^y are observed for the n individuals in $\mathcal{D}$ but not for the individuals in the augmented synthetic observations.

Here, we augment with $n^\star = 20000$ synthetic observations so that the empirical distribution X_1 matches the true $\Pr(X_1 = 1)$ with negligible standard error. The remaining variables are left completely missing for the synthetic observations. ICIN and ICIN+MAR represent default choices when auxiliary information is absent. That

Table 4.4: Posterior means and standard errors for simulation study when all five non-ignorable models are fitted to data generated under SCINN1, SCINN2 and SCINN3. "Par" is parameter.

				Fitted Model									
				ICIN		ICIN+MAR		SCINN1		SCINN2		SCINN3	
		Par.	Truth	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
Generative Model	SCINN1	α_0	.65	.96	.04	.97	.04	.62	.01	.67	.01	.62	.01
		β_0	.50	.65	.08	.64	.08	.60	.07	.67	.08	.59	.07
		β_1	-1.00	-1.03	.09	-1.03	.10	-1.07	.09	-1.09	.10	-1.07	.10
		η_0	-.40	-1.07	.05	-1.06	.05	-.29	.10	-1.09	.05	-.28	.09
		η_1	.25	.48	.07	.46	.07	.23	.08	.51	.07	.21	.08
		η_2	-.95	—	—	—	—	-1.08	.13	—	—	-1.08	.12
		ζ_0	-.60	-.61	.07	-.89	.04	-.64	.06	-.61	.07	-.89	.04
		ζ_1	-.40	-.38	.08	—	—	-.37	.08	-.40	.08	—	—
		ζ_2	—	—	—	.04	.07	—	—	.01	.07	.05	.07
	SCINN2	α_0	.65	.70	.04	.64	.04	.65	.01	.62	.01	.62	.01
		β_0	.50	.53	.08	.41	.09	.52	.09	.42	.08	.40	.08
		β_1	-1.00	-1.01	.10	-.89	.10	-1.00	.10	-.89	.10	-.88	.10
		η_0	-.40	-.40	.04	-.40	.04	-.83	.12	-.42	.04	.35	.11
		η_1	-1.10	-1.04	.08	-1.09	.08	-.93	.08	-1.02	.08	-1.10	.08
		η_2	—	—	—	—	—	.53	.14	—	—	-.07	.13
		ζ_0	-.20	.09	.06	-.63	.04	.53	.14	.24	.06	-.63	.04
		ζ_1	-1.40	-1.50	.08	—	—	-1.55	.07	-1.46	.08	—	—
		ζ_2	-.65	—	—	-.72	.07	—	—	.72	.08	-.72	.07
	SCINN3	α_0	.65	.97	.04	.97	.04	.62	.01	.67	.01	.62	.02
		β_0	.50	.57	.08	.58	.07	.55	.07	.60	.08	.55	.07
		β_1	-1.00	-.93	.09	-.94	.09	-1.00	.09	-.99	.10	-1.00	.09
		η_0	-.40	-1.07	.05	-1.07	.05	-.29	.09	-1.09	.05	-.31	.10
		η_1	.25	.48	.08	.48	.08	.26	.08	.52	.08	.26	.09
		η_2	-.95	—	—	—	—	-1.11	.13	—	—	-1.08	.13
		ζ_0	-.90	-.85	.07	-.92	.04	-.78	.06	-.95	.07	-.92	.04
		ζ_1	—	.04	.09	—	—	-.08	.08	.04	.08	—	—
		ζ_2	.30	—	—	.30	.07	—	—	.31	.07	.30	.06

is, based on current missing data literature, these are the two identifiable models analysts are most likely to fit for the data in this scenario. Therefore, we do not augment the data with synthetic observations when fitting them. On the other hand, SCINN1, SCINN2 and SCINN3 all represent variations of our framework on how best to use the auxiliary margin. We augment the data with synthetic observations when fitting these three models.

The estimated posterior means and standard errors of the parameters are shown in Tables 4.3 and 4.4. From Tables 4.3 and 4.4, our framework is able to recover estimates as well as ICIN and ICIN+MAR when the data is generated according to both of them. However, when the true nonresponse model is either one of SCINN1, SCINN2 and SCINN3, fitting ICIN and ICIN+MAR sometimes results in less accurate estimates than all three within our framework. In fact, SCINN1 appears to be the most consistent across all simulation scenarios for estimating the joint distribution between X_1 and Y_1. This confirms our previous assertion that assuming nonresponse indicators are conditionally independent given all other variables can be a reasonable choice. Although we only present results in one simulation run here, we rerun the simulation exercise four more times to verify our results; qualitatively, the conclusions are the same. We also take the same approach of running each simulation five times, in Sections 4.3.4 and 4.3.4.

Our framework provides more flexibility than current methods that do not leverage auxiliary data in that it encompasses those methods as special cases. It is clear that parameter estimates for the methods based on our framework will be biased if nonzero interactions exists in the true model. However, analysts can view those interactions as sensitivity parameters to test the additivity assumptions. Sensitivity analysis can be done using ideas akin to those used in the simulations studies by Deng et al. (2013).

In applying this method, analysts will need to make several implementation decisions. We have the following recommendations based on multiple simulation exercises (including those presented here). First, we suggest using the AN model for the variable that has auxiliary information and an ICIN model for the other variable, as is the case with SCINN1. This seems to be the most robust to model misspecification in the nonresponse models based on all our simulation scenarios. Second, when specifying the sequence of models, one also needs to decide the ordering of the variables.

In our experience and based on simulations, the ordering of the nonresponse variables does not seem to affect the conclusions. However, for the models for the variables of interest themselves, we suggest starting with the variable that is fully observed since the marginal distribution would be identifiable from the observed data alone. If none of the variables are fully observed, we suggest starting with the variable that has auxiliary information available, for the same reason.

Extending this scenario to incorporate auxiliary information about Y_1 is straightforward. We can relabel Y_1 as X_2 to indicate that it is now in X. We can also relabel the item nonresponse indicator for X_2 as R_{i2}^X accordingly. Based on our recommendation of the SCINN1 model, the auxiliary margin of Y_1 would allow us to estimate one more parameter in (4.28), so that we can have (4.30) and either

$$R_{i2}^X | X_{i1}, X_{i2}, R_{i1}^x \sim Bern(\pi_{R_{i2}^X}); \quad logit(\pi_{R_{i2}^X}) = \zeta_0 + \zeta_1 X_{i1} + \zeta_2 X_{i2} \tag{4.32}$$

or

$$R_{i2}^X | X_{i1}, X_{i2}, R_{i1}^x \sim Bern(\pi_{R_{i2}^X}); \quad logit(\pi_{R_{i2}^X}) = \zeta_0 + \zeta_1 X_{i1} + \zeta_2 R_{i1}^x \tag{4.33}$$

for the nonresponse mechanism for X_2. A nonzero value for ζ_2 in (4.32) results in an MNAR mechanism (AN model) for R_{i2}^X, whereas a nonzero value for ζ_2 in (4.33) results in an ICIN mechanism for R_{i2}^X. We again suggest using the margin to estimate an AN model for the nonresponse mechanism for the variable with the margin. Therefore, we recommend (4.32) instead of (4.33).

4.3.3 Two variables, with one fully observed and unit nonresponse included

We now show how modeling item nonresponse for one variable and unit nonresponse works within our framework by directly extending the scenario in Section 4.2.2 to include unit nonresponse. Suppose X contains two binary variables X_1 and X_2 but let Y be empty. Suppose X_2 suffers from item nonresponse but X_1 is fully observed, so that R_2^x is the fully observed vector of item nonresponse indicators for X_2 but

there is no need to account for R_1^x since it is a vector of zeros. Also, suppose the data contains unit nonrespondents. Let $U = (U_1, \ldots, U_n)$, where each U_i is the unit nonresponse indicator for each individual i. That is, $U_i = 1$ if individual i would not respond to the survey $\mathcal{D}$ at all and $U_i = 0$ otherwise. Here, U is fully observed. Following our notation, we have auxiliary margins for both X_1 and X_2. The data, along with the auxiliary margins, can be represented by Table 4.5a. The incomplete contingency table representing the joint distribution of (X_1, X_2, R_2^x, U) is shown in Table 4.5b. Clearly, we again only observe four out of 16 possible cell probabilities.

As we did before, we first characterize the joint distribution of (X_1, X_2, R_2^x, U) fully described by the 16 cell probabilities $\Pr(X_1 = x_1, X_2 = x_2, R_2^x = r, U = u)$, for $x_1, x_2, r, w \in \{0, 1\}^4$, using the following pattern mixture factorization

$$\begin{aligned}\Pr(X_1 = x_1, X_2 = x_2, R_2^x = r, U = u) = \Pr(X_2 = x_2 | X_1 = x_1, R_2^x = r, U = u) \\ \times \Pr(X_1 = x_1 | R_2^x = r, U = u) \\ \times \Pr(R_2^x = r | U = u) \Pr(U = u).\end{aligned} \tag{4.34}$$

We can fully parameterize this factorization using $\theta_{xru} = \Pr(X_2 = 1 | X_1 = x, R_2^x = r, U = u)$, $\pi_{ru} = \Pr(X_1 = 1 | R_2^x = r, U = u)$, $q_u = \Pr(R_2^x = 1 | U = u)$ and $p = \Pr(U = 1)$ resulting in a total of 15 parameters. Here, p, q_0, π_{r0} for $r \in \{0, 1\}$, and θ_{x00} for $x \in \{0, 1\}$, can be estimated from the data alone, so that we can estimate six of the 15 parameters directly from the data. With the auxiliary margins $\Pr(X_1 = 1)$ of X_1 and $\Pr(X_2 = 1)$ of X_2, we have the following linear constraints on the nine unobserved parameters

$$\Pr(X_1 = 1) - \Pr(X_1 = 1, R_2^x = r, U = 0) = p\left[\pi_{01}(1 - q_1) + \pi_{11} q_1\right] \tag{4.35}$$

Table 4.5: Two binary variables X_1 and X_2: X_1 is fully observed, X_2 suffers from item nonresponse and the data contains unit nonrespondents.

(a) Data

	X_1	X_2	R_2^x	U
Original data	✓	✓	0	0
		?	1	
	?	?	?	1
Auxiliary margin	✓	?	?	?
	?	✓	?	?

(b) Contigency table

		$R_2^x = 0$		$R_2^x = 1$	
		$X_2 = 0$	$X_2 = 1$	$X_2 = 0$	$X_2 = 1$
$U = 0$	$X_1 = 0$	✓	✓	?	?
	$X_1 = 1$	✓	✓	?	?
$U = 1$	$X_1 = 0$	?	?	?	?
	$X_1 = 1$	?	?	?	?

$$\begin{aligned}
&\Pr(X_2 = 1) - \Pr(X_2 = 1, X_1 = x_1, R_2^x = 0, U = 0) \\
&\quad = pq_1 \left[\theta_{011}(1 - \pi_{11}) + \theta_{111}\pi_{11}\right] \\
&\quad + p(1 - q_1) \left[\theta_{001}(1 - \pi_{01}) + \theta_{101}\pi_{01}\right] \\
&\quad + (1 - p)q_0 \left[\theta_{010}(1 - \pi_{10}) + \theta_{110}\pi_{10}\right],
\end{aligned} \tag{4.36}$$

allowing us to estimate two more parameters, to make a total of eight parameters. In other to specify a model with only eight identifiable parameters, we once again use our framework to specify a selection model. As mentioned before, the same steps can be followed to specify a pattern mixture model instead but we continue to focus on a selection model factorization.

According to Step 1 in Section 4.3, we need to specify a model with six parameters, the maximum number of parameters identifiable from the observed data alone. The obvious choice is

$$(X_1, X_2) \sim f(X_1, X_2|\Theta) \tag{4.37}$$

$$\Pr(U = 1|X_1, X_2) = g(\eta_0) \tag{4.38}$$

$$\Pr(R_2^x = 1|X_1, X_2, U) = h_2(\zeta_0 + \zeta_1 X_1). \tag{4.39}$$

This specification results in an MAR model on the nonresponse for X_2 and an MCAR model on the unit nonresponse. The models in (4.37) to (4.39) could be the default choice without using auxiliary data. η_0 is identifiable since the marginal probability of U is known from the observed data alone, while both ζ_0 and ζ_2 are identifiable since the marginal distribution of R_2^x as well as the joint distribution between R_2^x and X_1 can be estimated from the observed data alone. Given the conditional independence assumption between X_2 and R_2^x in (4.39), Θ would also be identifiable.

According to Step 2, we need to use the auxiliary margins $\Pr(X_1 = 1)$ and $\Pr(X_2 = 1)$ to estimate two more parameters in (4.38) or (4.39) relating directly to X_1 and X_2. Since (4.39) already contains X_1, we can only use $\Pr(X_1 = 1)$ to add X_1 to (4.38), so that (4.38) becomes

$$\Pr(U = 1|X_1, X_2) = g(\eta_0 + \eta_1 X_1). \tag{4.40}$$

We do however have some flexibility in how to use $\Pr(X_2 = 1)$. We can add X_2 to (4.39) so that the nonresponse models become (4.40) and

$$\Pr(R_2^x = 1|X_1, X_2, U) = h_2(\zeta_0 + \zeta_1 X_1 + \zeta_2 X_2), \tag{4.41}$$

or we can add X_2 to (4.40) so that the nonresponse models become (4.39) and

$$\Pr(U = 1|X_1, X_2) = g(\eta_0 + \eta_1 X_1 + \eta_2 X_2). \tag{4.42}$$

Note that we cannot uniquely estimate the relationship between U and R_1^x, since R_1^x is never observed whenever $U = 1$, and there is no auxiliary information about R_1^x to help provide a constraint on the relationship between them. As a result, we have to assume conditional independence between U and R_1^x. This is coherent with our assumption of conditional independence between all nonresponse indicators. Due to this independence assumption, changing the order between U and R_1^x in the sequence of models does not affect inference.

An important takeaway of our approach here is that due to the identifiability restrictions, we cannot specify nonignorable (specifically, MNAR) models for both

unit and item nonresponse, given the amount of information available in the data. However, we can still specify a nonignorable model for one of the nonresponse mechanisms under our framework. We have flexibility to either specify an ignorable (MAR) model for item nonresponse and a nonignorable (MNAR) model for unit nonresponse, or a nonignorable model for item nonresponse and an ignorable model for unit nonresponse.

As in Section 4.3.2, we use simulation studies to show how each option affects the estimation of the joint relationship between X_1 and Y_1. Again, we use logistic regressions default choices for the functions f, g, and h_2. We simulate data containing $n = 5000$ individuals from the following sequence of logistic models.

$$X_{i1} \sim Bern(\pi_{X_{i1}}); \quad logit(\pi_{X_{i1}}) = \alpha_0 \tag{4.43}$$

$$X_{i2}|X_{i1} \sim Bern(\pi_{X_{i2}}); \quad logit(\pi_{X_{i2}}) = \beta_0 + \beta_1 X_{i1} \tag{4.44}$$

where $\alpha_0 = 0.65$ and $(\beta_0, \beta_1) = (0.5, -1)$. We then generate missing data using the following three nonresponse mechanisms discussed.

1. The MCAR plus MAR models in (4.38) and (4.39) which we refer to as MCAR+MAR and specify using the following logistic regressions

$$U_i|X_{i1}, X_{i2} \sim Bern(\pi_{U_i}); \quad logit(\pi_{U_i}) = \eta_0 \tag{4.45}$$

$$R^x_{i2}|X_{i1}, X_{i2}, U_i \sim Bern(\pi_{R^x_{i2}}); \quad logit(\pi_{R^x_{i2}}) = \zeta_0 + \zeta_1 X_{i1} \tag{4.46}$$

   where $\eta_0 = -0.9$ and $(\zeta_0, \zeta_1) = (-0.25, -0.95)$, which would be a default approach without using auxiliary information.

2. The MAR plus MNAR models in (4.40) and (4.41) which we refer to as SCINN1 and specify using

$$U_i|X_{i1}, X_{i2} \sim Bern(\pi_{U_i}); \quad logit(\pi_{U_i}) = \eta_0 + \eta_1 X_{i1} \tag{4.47}$$

$$R^x_{i2}|X_{i1}, X_{i2}, U_i \sim Bern(\pi_{R^x_{i2}}); \quad logit(\pi_{R^x_{i2}}) = \zeta_0 + \zeta_1 X_{i1} + \zeta_2 X_{i2} \tag{4.48}$$

   where $(\eta_0, \eta_1) = (-0.6, -0.45)$ and $(\zeta_0, \zeta_1, \zeta_2) = (0.2, -0.8, -1.2)$.

Table 4.6: Posterior means and standard errors for simulation study when all three nonignorable models are fitted to data generated under each of them.

				Fitted Model					
				MCAR+MAR		SCINN1		SCINN2	
		Parameter	Truth	Mean	SE	Mean	SE	Mean	SE
Generative Model	MAR+ MCAR	α_0	.65	.65	.04	.62	.01	.62	.02
		β_0	.50	.63	.08	.45	.07	.51	.06
		β_1	-1.00	-1.11	.09	-1.01	.10	-1.10	.10
		η_0	-.90	-.86	.03	-.80	.09	-.57	.14
		η_1	—	—	—	-.10	.13	-.20	.15
		η_2	—	—	—	—	—	-.39	.17
		ζ_0	-.25	-.24	.06	-.01	.10	-.25	.06
		ζ_1	-.95	-.98	.08	-1.09	.09	-.98	.07
		ζ_2	—	—	—	-.39	.15	—	—
	SCINN1	α_0	.65	.80	.04	.62	.01	.62	.01
		β_0	.50	.96	.08	.39	.06	.33	.06
		β_1	-1.00	-1.13	.10	-.91	.08	-.83	.09
		η_0	-.60	-.86	.03	-.50	.07	.37	.11
		η_1	-.45	—	—	-.58	.12	-1.00	.14
		η_2	—	—	—	—	—	-1.61	.19
		ζ_0	.20	-.60	.07	.23	.10	-.60	.06
		ζ_1	-.80	-.36	.08	-.75	.09	-.36	.08
		ζ_2	-1.20	—	—	-1.53	.19	—	—
	SCINN2	α_0	.65	.49	.03	.62	.01	.62	.01
		β_0	.50	.80	.08	.34	.06	.51	.06
		β_1	-1.00	-1.09	.09	-.84	.09	-1.11	.09
		η_0	-.90	-.85	.03	-1.15	.10	-.56	.13
		η_1	.45	—	—	.44	.14	.18	.15
		η_2	-.60	—	—	—	—	-1.06	.18
		ζ_0	-.25	-.26	.06	.31	.10	-.26	.05
		ζ_1	-.95	-.95	.08	-1.20	.09	-.95	.07
		ζ_2	—	—	—	-1.00	.17	—	—

3. The MNAR plus MAR models in (4.39) and (4.42) which we refer to as SCINN2 and specify using

$$U_i|X_{i1}, X_{i2} \sim Bern(\pi_{U_i});\ \ logit(\pi_{U_i}) = \eta_0 + \eta_1 X_{i1} + \eta_2 X_{i2} \tag{4.49}$$

and (4.46), where $(\eta_0, \eta_1, \eta_2) = (-0.9, 0.45, -0.6)$.

In each simulation, we again fit the true generating model for X_1 and Y_1 (that is, (4.43) and (4.44)) plus all three nonresponse models back to the generated data.

In each case, we re-estimate parameters to examine how accurately each model estimates the joint distribution of X_1 and Y_1 as we did before. We fit all models using Bayesian MCMC, with uninformative priors for all parameters. We again run the MCMC sampler for 10,000 iterations, discarding the first 5,000 as burn-in, resulting in 5,000 posterior samples. The estimated posterior means and standard errors of the parameters are shown in Table 4.6. We set the values for the parameters in all three models so that we again have approximately 30% missing data for X_1 and Y_1 for the same reasons as before. We also incorporate the auxiliary margins by augmenting the data with synthetic observations as previously discussed. U is observed for the n individuals in $\mathcal{D}$ but not for the individuals in the augmented synthetic observations.

Similar to our results in Section 4.3.2, our framework is able to recover estimates as well as MCAR+MAR when the data is generated according to it. However, when the true nonresponse model is either one of SCINN1 or SCINN2, MCAR+MAR results in less accurate estimates than SCINN1 and SCINN2. There can be some weak identifiability issues with SCINN2, that is, when the unit nonresponse model depends on variables that also suffer from item nonresponse. This phenomenon is unsurprising because the observed data itself contains no information whatsoever about the unit nonresponse mechanism; all that information must come from the constraints implied from the auxiliary data. This is an important caveat to keep in mind. It suggests one should recommend SCINN1 in scenarios similar to this. However, SCINN2 still appears to be consistent across all simulation scenarios for estimating the joint distribution between X_1 and Y_1.

4.3.4 Two variables, with both suffering from item nonresponse and unit nonresponse included

Finally, we show how to model, within our framework, data where both variables have item nonresponse, and where the data also includes unit nonrespondents. We

define X_1 and X_2 as in Section 4.3.3 but with X_1 now allowed to suffer from item nonresponse, so that R_1^x is the item nonresponse indicator for X_1 and R_2^x is the item nonresponse indicator for X_2 as before. Also, let U be the unit nonresponse indicator as before. The data, along with the auxiliary margins, can be represented by Table 4.7a. The incomplete contingency table representing the joint distribution of $(X_1, X_2, R_1^x, R_2^x, U)$ is shown in Table 4.7b. Clearly, we have a severely sparse table as we only observe four out of all 32 possible cell probabilities.

Following our previous approaches, we first describe the joint distribution of $(X_1, X_2, R_1^x, R_2^x, W)$ fully described by the 32 cell probabilities $\Pr(X_1 = x_1, X_2 = x_2, R_1^x = r_1, R_2^x = r_2, U = u)$ for $x_1, x_2, r_1, r_2, w \in \{0,1\}^5$, using the following pattern mixture factorization. We have

$$\begin{aligned}
&\Pr(X_1 = x_1, X_2 = x_2, R_1^x = r_1, R_2^x = r_2, U = u) \\
&\qquad = \Pr(X_2 = x_2 | X_1 = x_1, R_2^x = r_2, R_1^x = r_1, U = u) \\
&\qquad \times \Pr(X_1 = x_1 | R_2^x = r_2, R_1^x = r_1, U = u) \\
&\qquad \times \Pr(R_2^x = r_2 | R_1^x = r_1, U = u) \\
&\qquad \times \Pr(R_1^x = r_1 | U = u) \Pr(U = u).
\end{aligned} \tag{4.50}$$

We can fully parameterize this factorization using $\theta_{xr_2r_1w} = \Pr(X_2 = 1 | X_1 = x, R_2^x = r_2, R_1^x = r_1, U = u)$, $\pi_{r_2r_1w} = \Pr(X_1 = 1 | R_2^x = r_2, R_1^x = r_1, U = u)$, $q_{r_1w} = \Pr(R_2^x = 1 | R_1^x = r_1, U = u)$, $s_u = \Pr(R_1^x = 1 | U = u)$ and $p = \Pr(U = 1)$ resulting in a total of 31 parameters. Here, p, s_0, q_{r_10} for $r_1 \in \{0,1\}$, π_{r_200} for $r_2 \in \{0,1\}$, and θ_{x000} for $x \in \{0,1\}$, can be estimated from the data alone. Thus, we can estimate eight of the 31 parameters directly from the data. With auxiliary information about the marginal distributions of X_1 and X_2, we would be able to estimate two more parameters, resulting in a total of ten parameters. We do not write out the constraints explicitly but they follow directly from the constraints in 4.14, 4.35 and 4.36.

To use our framework, we need to re-characterize the joint distribution of (X_1, X_2, R_2^x, U)

Table 4.7: Two binary variables X_1 and X_2: both suffer from item nonresponse and the data contains unit nonrespondents.

(a) Data

	X_1	X_2	R_1^x	R_2^x	U
Original data	✓	✓	0	0	0
	?	✓	1	0	
	✓	?	0	1	
	?	?	1	1	
	?	?	?	?	1
Auxiliary margin	✓	?	?	?	?
	?	✓	?	?	?

(b) Contigency table

$U = 0$		$R_2^x = 0$		$R_2^x = 1$	
		$X_2 = 0$	$X_2 = 1$	$X_2 = 0$	$X_2 = 1$
$R_1^x = 0$	$X_1 = 0$	✓	✓	?	?
	$X_1 = 1$	✓	✓	?	?
$R_1^x = 1$	$X_1 = 0$	?	?	?	?
	$X_1 = 1$	?	?	?	?

$U = 1$		$R_2^x = 0$		$R_2^x = 1$	
		$X_2 = 0$	$X_2 = 1$	$X_2 = 0$	$X_2 = 1$
$R_1^x = 0$	$X_1 = 0$	?	?	?	?
	$X_1 = 1$	?	?	?	?
$R_1^x = 1$	$X_1 = 0$	?	?	?	?
	$X_1 = 1$	?	?	?	?

via a selection model specification. There are a few options here but based on our discussions in Sections 4.3.2 and 4.3.3, we should specify some variant of the combination of SCINN1 from Section 4.3.2 and SCINN1/SCINN2 from Section 4.3.3. First, suppose the joint distribution of (X_1, X_2) is modeled as

$$(X_1, X_2) \sim f(X_1, X_2|\Theta). \tag{4.51}$$

Then, we can consider three choices for the nonresponse indicators using Steps 1 and 2 within our framework.

1. Use an MNAR model for the unit nonresponse indicator and ICIN models for the item nonresponse indicators. That is,

$$\Pr(U = 1|X_1, X_2) = g(\eta_0 + \eta_1 X_1 + \eta_2 X_2) \tag{4.52}$$

$$\Pr(R_1^x = 1|X_1, X_2, U) = h_1(\zeta_0 + \zeta_1 X_2) \tag{4.53}$$

$$\Pr(R_2^x = 1|X_1, X_2, U, R_1^x) = h_2(\gamma_0 + \gamma_1 X_1). \tag{4.54}$$

2. Use an ICIN model for the unit nonresponse indicator and one of the item nonresponse indicators, plus an MNAR model for the other item nonresponse indicator. That is either

$$\Pr(U = 1|X_1, X_2) = g(\eta_0 + \eta_1 X_1) \tag{4.55}$$

$$\Pr(R_1^x = 1|X_1, X_2, U) = h_1(\zeta_0 + \zeta_1 X_2) \tag{4.56}$$

$$\Pr(R_2^x = 1|X_1, X_2, U, R_1^x) = h_2(\gamma_0 + \gamma_1 X_1 + \gamma_2 X_2) \tag{4.57}$$

or

$$\Pr(U = 1|X_1, X_2) = g(\eta_0 + \eta_2 X_2) \tag{4.58}$$

$$\Pr(R_1^x = 1|X_1, X_2, U) = h_1(\zeta_0 + \zeta_1 X_2 + \zeta_2 X_1) \tag{4.59}$$

$$\Pr(R_2^x = 1|X_1, X_2, U, R_1^x) = h_2(\gamma_0 + \gamma_1 X_1) \tag{4.60}$$

3. Use an MCAR model for the unit nonresponse indicator and MNAR models for the item nonresponse indicators. That is,

$$\Pr(U = 1|X_1, X_2) = g(\eta_0) \tag{4.61}$$

$$\Pr(R_1^x = 1|X_1, X_2, U) = h_1(\zeta_0 + \zeta_1 X_2 + \zeta_2 X_1) \tag{4.62}$$

$$\Pr(R_2^x = 1|X_1, X_2, U, R_1^x) = h_2(\gamma_0 + \gamma_1 X_1 + \gamma_2 X_2) \tag{4.63}$$

Any one of the choices would be an improvement on current methods. Essentially, analysts have two choices for each variable that has an auxiliary margin available: (i) use the margin to estimate the relationship between the variable and its nonresponse indicator, or (ii) use the margin to estimate the relationship between the variable and the unit nonresponse indicator. This is a direct extension of our findings in Sections 4.3.2 and 4.3.3. We do not include simulation results comparing these options as the conclusions remain the same as in Sections 4.3.2 and 4.3.3.

It is quite possible for a subset of variables in a survey to not contain enough information to distinguish between item nonresponse on all of the questions being analyzed (that is, $R_1^x = R_2^x = 1$) and unit nonrespondents (that is, $U = 1$). In all our example scenarios, this is not a problem by design since we have all the information necessary to separate all item nonresponse and unit nonresponse. However, when there is no information to distinguish between them, it is necessary to make extra assumptions and incorporate them into the model as constraints. It could be plausible to assume that individuals who do not respond to any of the questions being analyzed and unit nonrespondents actually behave the same way. That is, we can think of people who do not respond to any of the questions being analyzed as unit nonrespondents – we do not have information to separate the two groups anyway. Therefore, when writing down models for the nonresponse indicators, we must constrain the probability of each indicator being equal one. For example, we could set $\Pr(R_2^x = 1)$ to zero whenever $R_1^x = 1$ and vice-versa, so that U completely captures all unit nonrespondents plus item nonrespondents who do not respond to any questions.

4.3.5 Extension to other scenarios

We have so far only focused on two binary variables, but extending our approach to more variables as well as categorical variables with more than two levels is conceptually straightforward. The data can still be represented using contingency tables as we have done in all three scenarios, so that it is easy to visualize the observed data and write down models for the observed data alone. Although writing down the constraints implied by the auxiliary information can be a daunting task with a large number of variables, we have showed using our simple examples that the core steps needed to specify identifiable nonignorable models do not require explicitly writing down the constraints themselves. Following our proposed steps and recommenda-

tions will still work in those settings. We have yet to explore the applications of our framework to continuous data, but this remains an important topic of future research.

We now proceed to apply our framework to the 2012 Current Population Survey (CPS), to show how to use our proposed steps to specify nonignorable identifiable models in application settings.

4.4 Application to CPS Data

Scholars and policymakers often lament over low voter turnout in the United States. Turnout rates typically fall just below 60% in presidential elections, 40% in midterm elections, and are lower still in off-year local elections – among the worst participation rates in advanced democracies (Powell, 1986). Turnout is also unequal among demographic subgroups. Turnout is especially low among young Americans, who consistently vote at rates 20-30 percentage points lower than older citizens (Holbein and Hillygus, 2016). Although there is widespread recognition of low and unequal electoral participation, it turns out that estimating turnout rates—especially for subgroups, can be difficult.

Given that we have official government counts of ballots cast in an election, it might seem puzzling that calculating turnout rates is at all complicated. There are two major reasons for this complication. First, estimates of ballots cast are often not available for population subgroups of interest. The ability to calculate votes by demographic subgroups depends on the information available in voter registration records, which varies considerably, especially by state. Second, to calculate turnout rate, we need an estimate of the denominator – the voting eligible population. Although there are official estimates of the voting age population, not all adults living in a locale are in fact eligible to vote. The discrepancy between voting age population and vote eligible population is mostly due to disenfranchised felons and non-citizens. This

discrepancy has grown over time, and varies considerably across states and across subgroups of the population (McDonald and Popkin, 2001). Given these issues with official election records, many researchers rely instead on survey-based estimates of turnout.

Among surveys, the CPS is considered the gold standard for estimating voter turnout. Every year the CPS November Supplement asks sampled respondents a variety of questions about voter registration and turnout. Response rates far exceed those of other surveys, approaching 90 percent. Additionally, the CPS is one of the few surveys with sufficient sample size to be able to make turnout estimates by state since the sample size exceeds 75,000 voting-age citizens, stratified by state.

Nonetheless, the CPS voter turnout measure is plagued by high levels of item nonresponse. In its official reports, the CPS treats "Don't Know", "Refused", and "No Response" as indicating that the respondent did not vote: "Nonrespondents and people who reported that they did not know if they voted were included in the 'did not vote' class because of the general overreporting by other respondents in the sample". This means the official Census Bureau estimates treat item and unit nonresponders in the November supplement as nonvoters. Hur and Achen (2013) highlight an example of how this creates biased estimates. In the 2008 U.S. Presidential Election between Barack Obama and John McCain, despite a historic number of ballots cast, the official CPS estimate reported a turnout rate that was slightly lower than Hur and Achen (2013) 2004 estimates. Although, Hur and Achen (2013) claim that imputing all missing responses as nonvoters helps reduce bias in the overall turnout estimates, it is unclear if it might increase bias of particular subgroups. For example, young people are both more likely to be survey nonrespondents and to be non-voters, raising the possibility that CPS estimates of youth turnout could be biased without corrections for unit and item non-response.

Our application estimates turnout among different subgroups of the population

in four U.S. states using data from the 2012 CPS. We compare subgroup turnout estimates using our framework to estimates found using the CPS approach as well as a second approach by Hur and Achen (2013) that drops all instances of non-response and re-weights voters and non-voters in a given state based on auxiliary data. For the purposes of our application, we assume that the data does not suffer from reporting errors; we discuss extensions for incorporating reporting errors in Section 4.5.

4.4.1 Data

To illustrate our proposed framework, we estimate voter turnout for demographic subgroups from 2012 CPS. We restrict our analysis to four states: Florida (FL), Georgia (GA), North Carolina (NC) and South Carolina (SC). We use four variables described in Table 4.8. Including two categorical variables (state and age) in our application shows that our framework can also be applied to categorical variables with more than two levels. The resulting dataset contains $n = 10,800$ individuals with data missing according to the nonresponse rates in Table 4.9. Although item nonresponse in sex is trivial and item nonresponse in age is fairly low, the severe nonresponse rates in both our variables of primary interest, vote, and unit nonresponse, means that 30% of the observations in this data suffer from some form of missing data.

We use the voter-eligible population (VEP) for highest office as marginal information for voter turnout by state: FL = 62.8%, GA = 59.0%, NC = 64.8 % and SC = 56.3%. This aggregate-level information was obtained from The United States Elections Project (USEP) (McDonald, 2008), which compiles government data to create election year estimates of the voting eligible population from the American Community Survey and Department of Justice felon estimates. This auxillary information is the same information used by Hur and Achen (2013) for reweighting purposes. For comparison, the unweighted estimates of voter turnout from the data

Table 4.8: Description of variables used in CPS illustration.

Variable	Categories
State	1 = Florida, 2 = Georgia, 3 = North Carolina, 4 = South Carolina
Sex	0 = Male, 1 = Female
Age	1 = 18 - 29, 2 = 30 - 49, 3 = 50 - 69, 4 = 70+
Vote	0 = Did not vote; 1 = Voted

Table 4.9: Unit and item nonresponse rates by state.

	Unit	Item		
		Vote	Sex	Age
FL	.18	.18	.00	.07
GA	.11	.16	.00	.05
NC	.14	.11	.00	.03
SC	.16	.10	.00	.03

7 total cases of missing sex

for complete cases by state are: FL = 75.2%, GA = 73.4%, NC = 77.4% and SC = 72.6%; the large differences between estimates from the complete cases and the VEP show the severe upward bias when using self-reported turnout data alone (DeBell et al., 2018). We also use marginal information for the age groups by state from the 2010 census – the proportions are given in Table 4.10.

4.4.2 Model

Let S_i, G_i, A_i and V_i represent the state, sex, age and vote of the $i = 1, \ldots, n$ individuals in the data. We use this notation for our variables instead of the notation for the variables in section 4.2.1, that is X_{ik} and Y_{ik}, to make the variables easier to identify. Following the notation in section 4.2.1 however, we note that all four variables are contained in $\mathcal{D}$ but G_i is not contained in $\mathcal{A}$, since we have no auxiliary information about sex. As before, let U_i represent the unit nonresponse indicator for individual i, where $U_i = 1$ if the individual did not respond to the survey and $U_i = 0$ otherwise. Also, let R_i^G, R_i^A and R_i^V be item nonresponse indicators for individual i, for sex, age and vote respectively, where each equals one if the corresponding variable

Table 4.10: Distribution of age by state from the 2010 census.

	Age			
	18-29	30-49	50-69	70+
FL	.20	.34	.31	.16
GA	.23	.39	.29	.09
NC	.22	.37	.30	.11
SC	.22	.34	.32	.12

Table 4.11: Monotone nonresponse in the three variables (sex, age and vote) suffering from item nonresponse across all states.

G_i	A_i	V_i
✓	✓	✓
	?	?
?		

is missing and equals zero otherwise.

Following our exploratory analysis, the data has monotone nonresponse in the three variables suffering from item nonresponse across all states: sex is always missing when age and vote are missing, and age is always missing when vote is missing. The monotone item nonresponse pattern is shown in Table 4.11. We recommend that analysts must first do exploratory analysis to find missing data patterns that should be included in the specified models, before proceeding to specify the nonresponse models. We therefore include constraints in our model to address the monotone nonresponse pattern.

Given the identifiability constraints discussed in Sections 4.2 and 4.3, it is not possible to specify a fully saturated model for all the variables and nonresponse indicators. We follow our framework in Section 4.3 and specify a selection model. Unit nonresponse is defined by state in this data, so that the relationships between S_i and the other variables as well as the nonresponse indicators are always observed, including the joint relationship between S_i and U_i in particular. That also implies that we do not have to include a model for S_i; we can simply view it as a covariate

in all the models.

First we specify the following sequence of models for the joint distribution of G_i, A_i and V_i.

$$\begin{aligned} G_i|S_i &\sim Bern(\pi_i^G); \\ & \quad logit(\pi_i^G) = \beta_1 + \beta_{2j}\mathbb{1}[S_i = j] \end{aligned} \tag{4.64}$$

$$\begin{aligned} A_i|G_i, S_i &\sim Cat(Pr[A_i \leqslant k] - Pr[A_i \leqslant k-1]); \\ & \quad logit(Pr[A_i \leqslant k]) = \phi_{1,k} + \phi_{2j}\mathbb{1}[S_i = j] + \phi_3 G_i \end{aligned} \tag{4.65}$$

$$\begin{aligned} V_i|A_i, G_i, S_i &\sim Bern(\pi_i^V); \\ & \quad logit(\pi_i^V) = \ \nu_1 + \nu_{2j}\mathbb{1}[S_i = j] + \nu_{3j}\mathbb{1}[A_i = j] + \nu_4 G_i \\ & \quad + \nu_{5jk}\mathbb{1}[S_i = j, A_i = k]. \end{aligned} \tag{4.66}$$

We exclude all interaction terms except $\mathbb{1}[S_i = j, A_i = k]$ to keep the model parsimonious – all interaction terms excluded were not significant based on step-wise model selection. In scenarios where the interaction terms are actually significant or where they are of inferential interest, they should be included since they can be estimated based on our discussions in Section 4.3.

Next we specify models for the nonresponse indicators. Following our recommendations from all scenarios in Section 4.3, we have several options within our framework. Since we are most interested in estimating turnout, we prefer to use the margin available for vote to estimate an AN model for R_i^V. Also, since we have such a high unit nonresponse rate, we prefer to use the margin for age in the model for U_i, so that we can learn as much as possible about the unit nonresponse. That implies we cannot have an AN model for R_i^A anymore. The most we can do, while still desiring a nonignorable model for R_i^A, is an ICIN model. Finally, since we do not have a margin for sex, we cannot include G_i in the models for R_i^G or U_i. Since the nonresponse rate for sex is so low, we adopt an MAR model for R_i^G to keep that model parsimonious as well – even though we can still include A_i and V_i in the model

for R_i^G. Given all these choices, we therefore specify nonignorable models for unit nonresponse as well as for item nonresponse on both age and vote, but an ignorable model for item nonresponse on sex. We write our specification formally as

$$\begin{aligned} U_i|\ldots &\sim Bern(\pi_i^U); \\ & \quad logit(\pi_i^U) = \gamma_1 + \gamma_{2j}\mathbb{1}[S_i = j] + \gamma_{3j}\mathbb{1}[A_i = j] \end{aligned} \tag{4.67}$$

$$\begin{aligned} R_i^G|U_i,\ldots &\sim Bern(\pi_i^{R^G}); \\ & \quad logit(\pi_i^{R^G}) = \eta_1 + \eta_{2j}\mathbb{1}[S_i = j] \end{aligned} \tag{4.68}$$

$$\begin{aligned} R_i^A|R_i^G,U_i,\ldots &\sim Bern([\pi_i^{R^A}]^{(1-R_i^G)}); \\ & \quad logit(\pi_i^{R^A}) = \alpha_1 + \alpha_{2j}\mathbb{1}[S_i = j] + \alpha_3 G_i + \alpha_4 V_i \end{aligned} \tag{4.69}$$

$$\begin{aligned} R_i^V|R_i^A,R_i^G,U_i,\ldots &\sim Bern([\pi_i^{R^V}]^{(1-R_i^A)}); \\ & \quad logit(\pi_i^{R^V}) = \psi_1 + \psi_{2j}\mathbb{1}[S_i = j] + \psi_{3j}\mathbb{1}[A_i = j] \\ & \quad + \psi_4 G_i + \psi_5 V_i. \end{aligned} \tag{4.70}$$

In each model, "..." represents conditioning on V_i, A_i, G_i, and S_i. We recommend that analysts take the same approach when applying our framework. That is, they should also compare the different choices and specify the models best suited to their statistical goals within our framework – our approach provides the flexibility to do so.

We fit all models using Bayesian MCMC, with non-informative priors for all parameters. We run the MCMC sampler for 10,000 iterations, discarding the first 5,000 as burn-in, resulting in 5,000 posterior samples. We incorporate marginal information by augmenting the observed data with $n^\star = 3n$ synthetic observations for each of the two variables with available auxiliary margins, resulting in a total of $64,800$ synthetic observations added to the observed data. For each margin, we augment with three times the size of the observed data so that the empirical margins match the auxiliary information with negligible standard error. We create $L = 50$

multiply imputed datasets, $\mathbf{Z} = (\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(50)})$ from the posterior samples. We estimate our estimands of interest from each imputed dataset and combine them using the multiple imputation rules (Rubin, 1987).

Our final joint model, that is (4.64) to (4.70), actually has less parameters than the maximum allowable. Although we can still estimate more parameters if desired, certain parameters remain inestimable. For example, we still cannot add $\gamma_4 V_i$ to (4.67) because this information can only come from the margin for vote (V_i and U_i are never observed together) and we have chosen to instead use the margin to estimate $\psi_5 V_i$ in (4.70) as V_i and R_i^V are also never observed together. Analysts should therefore keep this restriction on how the auxiliary margins can be used in mind. The restriction on which parameters are estimable comes not only from the number of parameters estimable (from combining the observed data with the auxiliary information) but also from the specific variable that carries the auxiliary information. Should analysts prefer to still include V_i in the unit nonresponse model, we can modify (4.67) and (4.70) as follows

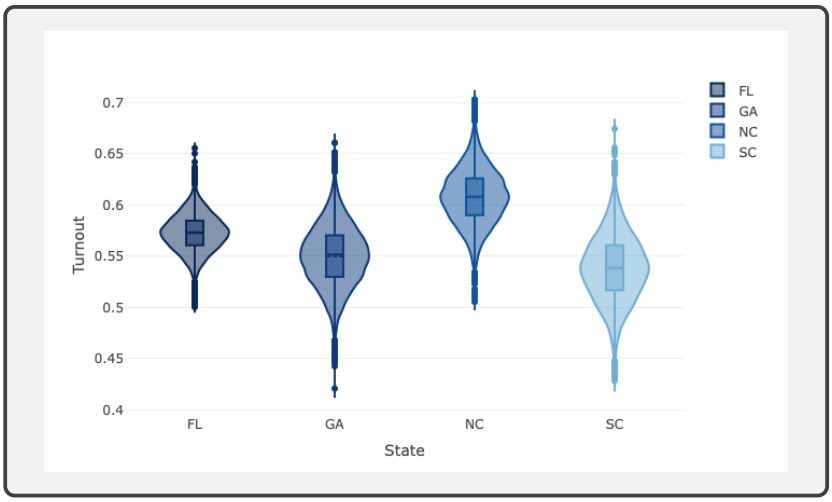
$$\begin{aligned} U_i|\ldots &\sim Bern(\pi_i^U); \\ & \quad logit(\pi_i^U) = \gamma_1 + \gamma_{2j}\mathbb{1}[S_i = j] + \gamma_{3j}\mathbb{1}[A_i = j] + \gamma_4 V_i \end{aligned} \tag{4.71}$$

$$\begin{aligned} R_i^V|\ldots &\sim Bern([\pi_i^{R^V}]^{(1-R_i^A)}); \\ & \quad logit(\pi_i^{R^V}) = \psi_1 + \psi_{2j}\mathbb{1}[S_i = j] + \psi_{3j}\mathbb{1}[A_i = j] + \psi_4 G_i. \end{aligned} \tag{4.72}$$

4.4.3 Results

The results presented in this section are based on the models in (4.64) to (4.70). The quantities we choose to investigate are based on collaborative work with political scientists. Figure 4.1 shows that our approach predicts more than 40% of unit non-respondents in the data to be voters. In comparison, the CPS method treats all unit nonrespondents as non-voters, and Hur and Achen (2013) drop all nonrespondents

FIGURE 4.1: Posterior predicted turnout among unit nonrespondents.



from the calculation of their estimates.

The point estimates for vote turnout, and the corresponding standard errors are shown in Table 4.12. Table 4.13 allows us to make a variety of comparisons between our method and the other methods' estimates of turnout. At the state-level, SCINN is comparable to the CPS estimates across all four states, and is comparable to the Hur and Achen (2013) method in Florida and Georgia. SCINN subgroup turnout estimates differ quite a bit from estimates using the Hur and Achen (2013) method and, to a lesser extent, the CPS method. Estimates of turnout for those under the age of 30 are around 5 percentage points higher for SCINN compared to Hur and Achen (2013) across the four states, whereas SCINN estimates are around 2 to 3 percentage points lower than CPS estimates. These trends continue when we further breakdown voters under 30 by sex. As expected, these subgroup estimates are the most dramatically different between the methods, likely because youth voters are known to experience both low levels of turnout and high levels of survey non-response compared to other age groups. In total, we still find that using more reasonable assumptions about missingness creates different turnout estimates from the Hur and Achen (2013) for subgroups where we would expect both low turnout

Table 4.12: Turnout estimates of subpopulations by state for the SCINN framework. M is male and F is female. Standard errors are in parenthesis.

	FL	**GA**	**NC**	**SC**
Full	.61 (.01)	.61 (.01)	.68 (.01)	.64 (.01)
M	.59 (.01)	.59 (.02)	.66 (.02)	.59 (.02)
F	.63 (.01)	.63 (.01)	.69 (.01)	.68 (.02)
<30	.45 (.02)	.46 (.02)	.53 (.03)	.53 (.03)
30-49	.59 (.01)	.62 (.02)	.68 (.02)	.61 (.02)
50-69	.68 (.01)	.71 (.02)	.74 (.02)	.72 (.02)
70+	.71 (.02)	.64 (.04)	.78 (.02)	.68 (.04)
<30(M)	.43 (.03)	.40 (.04)	.49 (.04)	.47 (.05)
30-49(M)	.56 (.02)	.59 (.03)	.64 (.03)	.57 (.03)
50-69(M)	.66 (.02)	.71 (.03)	.74 (.03)	.65 (.03)
70+(M)	.73 (.03)	.70 (.06)	.81 (.04)	.69 (.06)
<30(F)	.47 (.03)	.51 (.03)	.57 (.04)	.60 (.05)
30-49(F)	.62 (.02)	.65 (.02)	.71 (.03)	.64 (.03)
50-69(F)	.70 (.02)	.71 (.03)	.74 (.02)	.77 (.03)
70+(F)	.70 (.03)	.60 (.05)	.76 (.03)	.68 (.05)

and survey non-response to be common (young voters).

While benchmark comparisons do not exist for turnout among subgroups, we are able to provide a validation for our turnout estimates by comparing the demographic composition of voters in our sample to voter files in those states. We rely on a 1% sample of the Catalist database of 280 million individuals to attempt to qualify our estimates. Catalist compiles all state and county election lists, standardizes those lists, and checks the accuracy of the information against other sources, such as National Change of Address registry and the Postal Service list of valid addresses. Catalist keeps the records of those who have been purged (i.e., removed from official registration records due to nonvoting, registration elsewhere, or death), accounts for duplicate listings, and seeks to identify movers via comparison to information from the U. S. Post Office. The Catalist data offers fully observed estimates (with negligible standard error) of the joint probabilities of turnout rates for each subgroup variable we included in our analysis. However, reliable auxiliary marginal information in the Catalist data is available only for those registered to vote, and not for the entire voting eligible population.Figure 4.2 compares the composition of voters in the

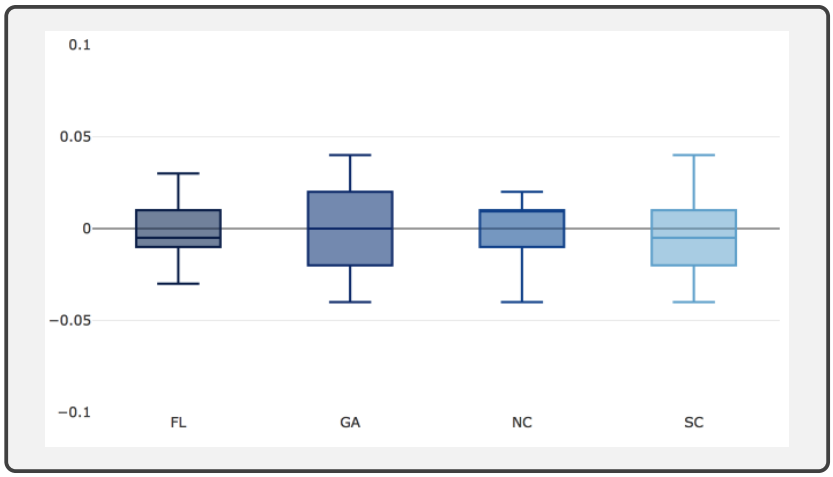
Table 4.13: Comparing turnout estimates of subpopulations by state for different methods. M is male and F is female. SCINN is our proposed method, CPS is the current weighted CPS estimates using the U.S. Census Bureau's method, and H&A uses the Hur and Achen (2013) method.

	FL			**GA**			**NC**			**SC**		
	SCINN	CPS	H&A	SCINN	CPS	H&A	SCINN	CPS	H&A	SCINN	CPS	H&A
Full	.61	.61	.62	.61	.62	.59	.68	.69	.64	.64	.65	.56
M	.59	.59	.58	.59	.59	.55	.66	.67	.62	.59	.60	.50
F	.63	.63	.65	.63	.64	.62	.69	.70	.66	.68	.69	.61
<30	.45	.46	.40	.46	.47	.40	.53	.56	.49	.53	.56	.47
30-49	.59	.58	.60	.62	.61	.60	.68	.69	.64	.61	.62	.53
50-69	.68	.67	.71	.71	.72	.72	.74	.73	.71	.72	.71	.63
70+	.71	.70	.74	.64	.66	.61	.78	.77	.74	.68	.70	.59
<30(M)	.43	.45	.38	.40	.39	.33	.49	.53	.45	.47	.49	.40
30-49(M)	.56	.56	.55	.59	.57	.55	.64	.65	.58	.57	.58	.50
50-69(M)	.66	.64	.68	.71	.72	.72	.74	.73	.72	.65	.64	.54
70+(M)	.73	.72	.78	.70	.73	.68	.81	.80	.79	.69	.70	.61
<30(F)	.47	.47	.42	.51	.54	.47	.57	.59	.52	.60	.63	.55
30-49(F)	.62	.61	.65	.65	.65	.65	.71	.72	.70	.64	.65	.56
50-69(F)	.70	.70	.74	.71	.71	.73	.74	.73	.70	.77	.76	.71
70+(F)	.70	.68	.71	.60	.61	.57	.76	.75	.71	.68	.69	.57

Catalist data and SCINN's estimates in each of the four states. SCINN produces a composition of voters nearly identical to estimates produced using Catalist's verified voters in these states.

When we replace (4.67) and (4.70) with (4.71) and (4.72), the overall conclusions remain the same, with the difference being that almost all unit nonrespondents are now predicted to be non-voters, but more than half of item nonrespondents for vote are instead predicted to be voters. This phenomenon is due to the large differences between the turnout rates from the complete cases and those from the VEP. Essentially, for the model to make up for the large differences, a lot of nonrespondents must be predicted as nonvoters, and whether those would be unit or item nonrespondents depends on whether the margin for vote is used in the unit or item nonresponse models. In fact, our results in Table 4.13 are very similar to the CPS estimates even though we make different assumptions. There is so much of an upward bias in vote turnout that one way or another, a large number of missing values in the vote

FIGURE 4.2: Distributions of deviations in SCINN estimates for voter composition from Catalist estimates.



variable must be predicted as nonvoters anyway. We posit that this would not be the case if the rates in the complete cases were much closer to the true margins and our model would yield slightly different results to those based on the CPS method. In both cases however, our approach still predicts that a good proportion of nonrespondents are in fact voters, which does not align with the assumptions of either the CPS method of Hur and Achen (2013).

4.5 Discussion

Fitting identifiable nonignorable models that simultaneously handle unit and unit nonresponse can be challenging. Many parameters in nonignorable models for item nonresponse alone can often be unidentifiable, with this problem becoming more complicated when including unit nonresponse. As our results show, the SCINN framework provides a flexible framework for tackling these two forms of nonresponse simultaneously when auxiliary data is available. Specifically, we have showed how our framework uses the marginal information from auxiliary data sources to identify extra parameters that would have been otherwise unidentifiable, at least under the

simulations and application settings explored in this article. With nonresponse rates on the rise, and budgets for nonresponse follow-up on the decline, agencies need ways to take advantage of information in auxiliary data sources. The SCINN framework provides implementable options for doing so.

Opportunities for further improvement of our framework exist as future research topics. First, our approach of specifying a sequence of parametric regressions within the SCINN framework is challenging with large numbers of variables. With a large number of variables, there are an enormous number of possible model specifications, especially interaction terms. Model selection is further complicated by the need to worry about identification constraints. As most surveys often contain a large number of variables, an important future research topic is to explore how to use flexible conditional nonparametric models, for example, classification and regression tress, and Bayesian additive regression trees.

Second, the SCINN framework can be extended to handle reporting error as well. In large surveys, reporting errors can result in significantly biased inference. In the CPS, for example, respondents are often inclined to say they voted even though they did not, because it is socially desirable to vote. One recent validation study of the American National Election Study found that 5.5 percent of the sample said they voted when they did not in fact do so (Jackman and Spahn, 2014). Future work would extend our model-based framework to handle reporting error simultaneously with missing data imputation through a hierarchical specification. This can be done by adding a reporting model to explain how reported values are generated from the true unobserved values.

Finally, the framework presented here does not incorporate survey design variables or survey weights in complex surveys. When such design variables exist, they often contain important information that should be incorporated in imputation models. When survey weights are available, they can be incorporated for example, by

imposing the requirement that the completed datasets results in design-unbiased estimates of the margins of X that are plausible realizations given the sampling design and the auxiliary margin. I explore this extension in Chapter 5.

5

Incorporating Survey Weights when Leveraging Auxiliary Information in Multiple Imputation for Complex Surveys

5.1 Introduction

In survey applications, the observed sample is rarely a simple random sample from the true population. In fact, more complex sampling schemes are typically used, resulting in unequal inclusion probabilities and survey weights. When such complex sampling methods are used, information on the survey design is often encoded in the survey design variables or survey weights. Analysts should account for this information, for example, by using design-based estimation (Lohr, 2010).

In Chapter 4, we developed the framework for leveraging auxiliary marginal information under the assumption of simple random sampling. When dealing with nonresponse in complex surveys, the imputation process may be complicated by the survey design. We should incorporate the survey design into modeling. A key challenge, then, is how to ensure imputations for nonresponse respect the complex survey design when leveraging auxiliary marginal information.

To illustrate, suppose the design-unbiased estimate of the percentage of men based on the complete cases in a sample is 70%. However, suppose we know from auxiliary data that the target population includes 50% men and 50% women. Then, not only should we likely impute more women than men when imputing the question asking the sex of the respondent, we also need to ensure that the weights play a role in the imputation. Essentially, we need to ensure that the imputations result in a plausible design-unbiased estimate of the proportion of men in the completed dataset. Specifically, the design-unbiased estimate should be close to 50%, given the particular survey design and nonresponse mechanism.

When using multiple imputation for handling nonresponse in complex surveys, a common recommendation is to incorporate survey design variables in the imputation models (Reiter et al., 2006). Other approaches are to include survey weights in the imputation models, to use design-based methods to estimate the imputation models, and undoing the effects of the design to make simple random samples (Zhou et al., 2016). None of these methods take advantage of auxiliary margins.

We propose to impose the requirement that the completed datasets result in design-unbiased estimates of the margins of the variables of interest that are plausible realizations, given the sampling design and the auxiliary margins. We do so by fusing the ideas of the SCINN framework in Chapter 4 with large sample results under frequentist (survey-weighted) paradigms. Our approach ensures that imputations are influenced by relationships in the data and the auxiliary information, while being faithful to the survey design through survey weights.

In this chapter, we work with base weights, which usually only contain information on the survey design, instead of more complex "adjusted" weights, which are often inflated to adjust for nonresponse or post-stratification. Since we take a model-based approach to handling survey nonresponse, there is no obvious justification for using adjusted weights that already account for the nonresponse. In fact,

using such adjusted weights assumes that the weights are fixed, which is not often true as pointed out by Fienberg (2010). We also do not focus on the construction of survey weights or their limitations (Si et al., 2015). Our primary focus is on imputation given base survey weights; specifically, using the base weights in addition to the auxiliary information to ensure imputations respect the survey design.

The remainder of this chapter is organized as follows. In section 5.2, we present our proposed method. In section 5.3, we use simulations to illustrate the performance of the proposed approach. In section 5.4, we conclude and discuss possible extensions.

5.2 Methods

5.2.1 Notation

We use some of the notation from Chapter 4. As a brief review, $\mathcal{D}$ comprises data from the survey of $i = 1, \ldots, n$ individuals, and $\mathcal{A}$ comprises data from the auxiliary database. Let $X = (X_1, \ldots, X_p)$ represent the p variables in both $\mathcal{A}$ and $\mathcal{D}$, where each $X_k = (X_{1k}, \ldots, X_{nk})^T$ for $k = 1, \ldots, p$. Let $Y = (Y_1, \ldots, Y_q)$ represent the q variables in $\mathcal{D}$ but not in $\mathcal{A}$, where each $Y_k = (Y_{1k}, \ldots, Y_{nk})^T$ for $k = 1, \ldots, q$. Let $X_i = (X_{i1}, \ldots, X_{ip})$ and $Y_i = (Y_{i1}, \ldots, Y_{iq})$. As we did in Chapter 4, we continue to use generic notations such as f and η for technically different functions and parameters.

Let N represent the number of units in the population from which the n survey units in $\mathcal{D}$ are sampled. Let $W = (w_1, \ldots, w_n)$, where each w_i is the base weight for the ith unit in the sample survey $\mathcal{D}$. Here, w_i is the inverse probability of selection π_i of the ith unit, without calibration or nonresponse adjustments. Let the superscript "*pop*" represent the population counterparts of the survey variables already defined. For example, X^{pop} and Y^{pop} represent the population based counterparts of X and Y respectively, where each $X_i \in X^{pop}$ and $Y_i \in Y^{pop}$. We do not observe values of X^{pop} or Y^{pop} for all non-sampled units in the population. In this chapter, we assume that

Table 5.1: Two binary variables Y_1 and X_1 with Y_1 fully observed and X_1 containing item nonresponse.

(a) Observed data

W	X_1	Y_1	R_1^x
✓	✓	✓	0
	?		1

(b) Contigency table

	$R_1^x = 0$		$R_1^x = 1$	
	$X_1 = 0$	$X_1 = 1$	$X_1 = 0$	$X_1 = 1$
$Y_1 = 1$	✓	✓	?	?
$Y_1 = 2$	✓	✓	?	?
$Y_1 = 3$	✓	✓	?	?

W is the only available variable containing information related to the survey design. The w_i's are known for all individuals in the survey $\mathcal{D}$ but not for the non-sampled units in the population.

5.2.2 Our proposed approach

To make the development easy to follow, we work with a two-variable example similar to the one in Section 4.2.2. Let X_1 be a binary variable and Y_1 be a categorical variable with three levels, that is, $Y_1 \in \{1, 2, 3\}$. For now, we assume that the data does not contain any unit nonrespondents; we discuss extensions to scenarios including unit nonrespondents in Section 5.4. X_1 and Y_1 (in X and Y respectively) are contained in $\mathcal{D}$. We only have auxiliary information for X_1 from $\mathcal{A}$, and no auxiliary information for Y_1. Suppose X_1 suffers from item nonresponse, so that we need to specify a model for R_1^x, the fully observed vector of item nonresponse indicators for X_1. Suppose Y_1 is fully observed and we do not need to include a model for R_1^y.

The observed data for this two-variable example takes the form shown in Table 5.1a. The incomplete contingency table representing the joint distribution of (X_1, Y_1, R_1^x), with observed marginals and weights excluded, is shown in Table 5.1b. Clearly, from both tables, we cannot fit a fully saturated model to the data. We therefore follow our SCINN framework in Chapter 4 and build a sequence of identifiable models to the data. First, we apply Step 1 of the SCINN framework from

Section 4.3 and specify the following model without any auxiliary information. The model is

$$Y_1 \sim f(\theta) \tag{5.1}$$

$$\Pr(X_1 = 1|Y_1) = g(\alpha_0 + \alpha_{1j}\mathbb{1}[Y_1 = j]) \tag{5.2}$$

$$\Pr(R_1^x = 1|X_1, Y_1) = h(\gamma_0 + \gamma_{1j}\mathbb{1}[Y_1 = j]), \tag{5.3}$$

resulting in an ignorable (MAR) mechanism, where $j = 1, 2, 3$. We set $\alpha_{1j} = 0$ and $\gamma_{1j} = 0$ to ensure the model is identifiable. We seek to fit a nonignorable model that includes $\gamma_2 X_1$ in (5.3), so that (5.3) becomes the AN model

$$\Pr(R_1^x = 1|X_1, Y_1) = h(\gamma_0 + \gamma_{1j}\mathbb{1}[Y_1 = j] + \gamma_2 X_1). \tag{5.4}$$

To do so, we need the auxiliary margin for X_1. This is necessary to force an extra constraint on the parameters, as we did in Chapter 4. However, when the survey design is complex, it may not be sufficient to use the auxiliary margin as we did in Chapter 4 as that approach does not incorporate the survey weights directly.

In practice, the most common marginal information is the population total (or mean) of some of the variables. For example, for totals, we know that

$$T_X = \sum_{i=1}^{N} X_{i1}^{pop} = N \times \Pr(X_1^{pop} = 1), \tag{5.5}$$

where $\Pr(X_1^{pop} = 1)$ is the true auxiliary marginal probability. A classical survey-unbiased estimator of T_X in this case is the Horvitz-Thompson estimator (Horvitz and Thompson, 1952), henceforth referred to as HT estimator, which is

$$\hat{T}_X = \sum_{i=1}^{n} \frac{X_{i1}}{\pi_i} = \sum_{i=1}^{n} w_i X_{i1}. \tag{5.6}$$

When the data contain nonresponse, it could then be desirable to require that imputed values result in values of $\hat{T}_X$ that are plausible under the sampling design and

the true T_X. Therefore, we need to impose a constraint that ties (5.6) in the imputed datasets directly to T_X. It is also desirable to make the constraint probabilistic, so as to incorporate uncertainty about the fact that various combinations of imputed values could be plausible.

For all $i \in \mathcal{D}$, let $X_{i1}^{\star} = X_{i1}$ when $R_{i1}^{x} = 0$, and let $X_{i1}^{\star}$ be the unknown true value when $R_{i1}^{x} = 1$. We propose the following probabilistic constraint

$$\sum_{i \in \mathcal{D}} w_i X_{i1}^{\star} \sim N(T_X, V_X). \tag{5.7}$$

We seek imputations consistent with (5.7) when generating imputed values for X under the posterior predictive distribution implied by (5.1), (5.2) and (5.4). Here, V_X is the estimated design-based variance associated with $\sum_{i \in \mathcal{D}} w_i X_{i1}^{\star}$, for example, based on previous knowledge or an average of estimates from preliminary sets of completed data. Throughout this chapter, we assume that V_X is pre-specified and treated as known. The constraint in (5.7) is based on frequentist (survey-weighted) paradigms; specifically, it takes the form of the central limit theorem result for the HT estimator in this case (Fuller, 2009).

We incorporate (5.7) into our MCMC sampler through a Metropolis algorithm. At each MCMC iteration t, let the current draw of each X_{i1} be $X_{i1}^{(t)}$. We use the following sampler.

S1. For all $i = 1, \ldots, n$, set $X_{i1}^{\star} = X_{i1}$ if $R_{i1}^{x} = 0$. If $R_{i1}^{x} = 1$, generate a candidate $X_{i1}^{\star}$ for the missing X_{i1} from the implied posterior predictive distribution implied by (5.1), (5.2) and (5.4).

S2. Let $\hat{T}_X^{\star} = \sum_{i \in \mathcal{D}} w_i X_{i1}^{\star}$ and $\hat{T}_X^{(t)} = \sum_{i \in \mathcal{D}} w_i X_{i1}^{(t)}$. Calculate the acceptance ratio

$$p = \frac{N(\hat{T}_X^{\star}; T_X, V_X)}{N(\hat{T}_X^{(t)}; T_X, V_X)}. \tag{5.8}$$

S3. Let $u \sim Unif(0,1)$. If $u \leqslant p$, accept all the proposed candidates $X_{i1}^{\star}, \ldots, X_{in}^{\star}$ and set $X_{i1}^{(t+1)} = X_{i1}^{\star}$ for $i = 1, \ldots, n$. Otherwise, reject all the proposed candidates and set $X_{i1}^{(t+1)} = X_{i1}^{(t)}$ for $i = 1, \ldots, n$.

Intuitively, these steps reject completed datasets that yield highly improbable design-based estimates, while simultaneously allowing us to estimate $\gamma_2 X_1$ in (5.4), since the accepted values are constrained by the auxiliary total. Although (5.7) provides a constraint on a distribution, whereas using the auxiliary margins as in Chapter 4 forms linear constraints, $\gamma_2 X_1$ is still estimable when using (5.7), as we illustrate later.

We recommend that analysts monitor the acceptance ratio of the missing data sampler in Steps S1 to S3, just as is the case with standard Metropolis samplers. In cases where the acceptance ratio is considerably low, analysts can inflate or tune V_X or even consider other methods of generating more "plausible" imputations from the implied posterior predictive distribution. In our simulation scenarios, however, there is no need to do so as our samplers mix properly. We do not worry about cases where the acceptance ratio is high because we view (5.7) as a constraint rather than a "target distribution", as is the case with standard Metropolis samplers. Therefore, we interpret a high acceptance ratio as meaning that the sampler is doing a good job of generating imputations that respect the survey design, as desired.

5.3 Simulations

We illustrate our approach using several simulation scenarios. In all the scenarios, we use probit regression as the default choice for the functions g and h, instead of the probit regressions of Chapter 4. We do so for computational reasons; the MCMC sampler mixes much better when including the constraint in (5.7) in a sequence of probit regressions than in a sequence of probit regressions. We focus on stratified

sampling, using eight simulation scenarios. We discuss extensions to scenarios with unequal weights and other sampling schemes in Section 5.4.

In each scenario, we first create a population of size $N = 50000$ and split them into two strata: 70% of units ($N_1 = 35000$) in stratum 1 and 30% of units ($N_2 = 15000$) in stratum 2. We generate data for Y_1 and X_1 using the following generative model.

$$Y_{i1} \sim \text{Discrete}(\theta_1, \theta_2, \theta_3) \tag{5.9}$$

$$X_{i1}|Y_{i1} \sim \text{Bernoulli}(\pi_{X_{i1}}); \quad \Phi^{-1}(\pi_{X_{i1}}) = \alpha_0 + \alpha_{1j}\mathbb{1}[Y_{i1} = j], \tag{5.10}$$

for $j \in \{2, 3\}$, where $\pi_{X_{i1}} = \Pr[X_{i1} = 1|Y_{i1}]$. Here, the Discrete distribution refers to the multinomial distribution with sample size equal to one, and Φ^{-1} is the inverse cumulative distribution function of the standard normal distribution. In all scenarios, we set $\theta = (\theta_1, \theta_2, \theta_3) = (0.5, 0.15, 0.35)$ in stratum 1, and $\theta = (0.1, 0.45, 0.45)$ in stratum 2. This ensures that the joint distribution of Y_1 and X_1 differ across strata. We set different values for α_0 and $\alpha_1 = \{\alpha_{12}, \alpha_{13}\}$ under each simulation scenario to explore how the strength of the relationship between X_1 and Y_1 affects our results.

We sample $n = 5000$ observations from the population using stratified simple random sampling. We sample $n_1 = 1500$ units from stratum 1 and $n_2 = 3500$ units from stratum 2. We set n_1 and n_2 to ensure that the base weights matter in the estimation. Based on this setup, the probability of selection $\pi_i = n_1/N_1$ for each unit in stratum 1, and $\pi_i = n_2/N_2$ for units in stratum 2. Thus, $w_i = N_1/n_1 = 35000/1500 = 23.33$ for units in stratum 1, but $w_i = N_2/n_2 = 15000/3500 = 4.29$ for units in stratum 2. Finally, we add item nonresponse by specifying an AN model for X_1. We have

$$R^x_{i1}|Y_{i1}, X_{i1} \sim \text{Bernoulli}(\pi_{R^x_{i1}}); \quad \Phi^{-1}(\pi_{R^x_{i1}}) = \gamma_0 + \gamma_{1j}\mathbb{1}[Y_{i1} = j] + \gamma_2 X_{i1}. \tag{5.11}$$

We set different values for γ_0, $\gamma_1 = \{\gamma_{12}, \gamma_{13}\}$ and γ_2 under the simulation scenarios to explore how the strength of nonignorability in the item nonresponse for X_1 affects

our conclusions. In each case, we set the values for γ_0, γ_1 and γ_2 so that we have approximately 30% missing data in X_1.

We examine the following approaches for imputing the item nonresponse in X_1. In all approaches, we fit (5.9) to (5.10) to the simulated data, but estimate the item nonresponse mechanism and incorporate both the weights and auxiliary information differently.

1. SCINN: We follow our SCINN framework in Chapter 4 by fitting the true nonignorable nonresponse model in (5.11). We incorporate the auxiliary margin by augmenting the sample with $n^\star = 2n = 10000$ synthetic observations so that the empirical distribution X_1 matches the population $\Pr(X_1^{pop} = 1)$ with negligible standard error.

2. AN+Constraint: We use our proposed method in Section 5.2 by again fitting the true nonignorable nonresponse model in (5.11), but we incorporate the auxiliary total and weights through the constraint in (5.7). We do not augment with synthetic observations.

3. AN+Weight: Here we incorporate the survey weights by fitting (5.11) and adding w_i as a factor variable to (5.10). Since there is a one-to-mapping between weights and strata, we incorporate w_i by adding an indicator variable S_i for strata, so that we have

$$\begin{aligned} X_{i1}|Y_{i1} &\sim \text{Bernoulli}(\pi_{X_{i1}}); \\ \Phi^{-1}(\pi_{X_{i1}}) &= \alpha_0 + \alpha_{1j}\mathbb{1}[Y_{i1} = j] + \alpha_2\mathbb{1}[S_i = 2] \end{aligned} \tag{5.12}$$

   as the model for X_1 instead of (5.10). This approach represents one of the default approaches analysts could fit for the data in this scenario, without using either our framework in Chapter 4 or our approach in this chapter. We do not augment the data with synthetic observations when using this method.

4. AN+Constraint+Weight: We also explore a combination of the AN+Constraint and AN+Weight approaches. Essentially, we follow the AN+Constraint method but then use (5.12) instead of (5.10) to further control for the weights. Again we do not augment the data with synthetic observations here.

5. MAR+Weight: Without auxiliary data in the AN+Weight approach, $\gamma_2 X_{i1}$ in (5.11) actually cannot be identified using the observed data alone. Therefore, we also consider a variation of the AN+Weight approach that uses (5.12) but excludes $\gamma_2 X_{i1}$ in (5.11), so that we have

$$R^x_{i1}|Y_{i1}, X_{i1} \sim \text{Bernoulli}(\pi_{R^x_{i1}}); \quad \Phi^{-1}(\pi_{R^x_{i1}}) = \gamma_0 + \gamma_{1j}\mathbb{1}[Y_{i1} = j], \tag{5.13}$$

which is an MAR model for the nonresponse.

We also explore results for a combination of the SCINN and AN+Constraint approaches. Specifically, we augment the data with synthetic data generated according to the auxiliary population $\Pr(X_1^{pop} = 1)$, but we also include the constraint in (5.7). However, we exclude those results because qualitatively, the results are worse off than the AN+Constraint approach by itself.

We fit all models using Bayesian MCMC with non-informative priors for all parameters. We use a data augmentation scheme for the probit regressions. We run the MCMC sampler for 10,000 iterations, discarding the first 5,000 as burn-in, resulting in 5,000 posterior samples. We create $L = 50$ multiply imputed datasets, $\mathbf{Z} = (\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(50)})$, every 100^{th} of the posterior samples. From each completed dataset $\mathbf{Z}^{(l)}$, we re-estimate the population total T_X and compute the design-based estimates of α_0, α_{12}, α_{13}, γ_0, γ_1 and γ_2, along with the corresponding standard errors, using the survey-weighted generalized linear models option in the R package, "survey". Although there could be many challenges associated with using survey weights in regression models (Pfeffermann, 1993; Gelman, 2007), as we do with the "survey" package, we proceed to do so only for illustrative purposes. We combine

all the estimates across all multiply-imputed datasets using the multiple imputation rules (Rubin, 1987). We repeat all simulations $M = 10$ times to further account for Monte Carlo errors and average all MI estimates across the 10 runs. That is, we use $\sum_{m=1}^{M} \bar{q}_L/M$ as the point estimate of each estimand Q, and $\sqrt{\sum_{m=1}^{M} T_L/M}$ as the standard error associated with it, where $\bar{q}_L$ and T_L are the MI quantities needed for inference as presented in Chapter 1.2.

5.3.1 Overall margin for X_1, strong relationship between Y_1 and X_1 and strong nonignorable nonresponse

In this scenario, we set $\alpha_0 = 0.5$, $(\alpha_{12}, \alpha_{13}) = (-0.5, -1)$, $\gamma_0 = -0.25$, $(\gamma_{12}, \gamma_{13}) = (0.1, 0.3)$, and $\gamma_2 = -1.1$ to reflect a strong relationship between Y_1 and X_1 and a strong nonignorable nonresponse mechanism. Table 5.2a shows the HT estimates for T_X and the standard error under each method, both averaged across all 10 separate MCMC runs. AN+Constraint and AN+Constraint+Weight give the most accurate estimates, with SCINN not too far behind. It appears that controlling for the weight in the model for X_1 as in the AN+Constraint+Weight method also decreases the standard error, and increases the range of the acceptance ratio in comparison to AN+Constraint. The AN+Weight and MAR+Weight methods give the least accurate estimates of all five methods. In fact, the standard error associated with AN+Weight is much higher than all other methods. This is in part due to the identifiability issues associated with using the AN model without any auxiliary information, resulting in greater uncertainty from the nonresponse mechanism.

Table 5.2b shows survey-weighted estimates of α_0, α_{12}, α_{13}, γ_0, γ_{12}, γ_{13} and γ_2, along with the corresponding standard errors, also averaged across the 10 runs. Here, both AN+Constraint and AN+Constraint+Weight give nearly identical results and closely estimate the true parameter estimates, whereas SCINN offers slightly less accurate estimates. AN+Weight and MAR+Weight again give the least accurate

Table 5.2: **Scenario one**: overall margin for X_1, strong relationship between Y_1 and X_1 and strong nonignorable nonresponse.

(a) HT estimates for T_X under each method, the corresponding standard errors, and acceptance ratios. "Population" is the value of T_X in the population of $N = 50000$ individuals. "No Missing Data" is the weighted estimate based on the sampled data before introducing item nonresponse.

	T_X		Acceptance Ratio	
Method	Mean	SE	Mean	Range
Population	25026	—	—	—
Mo Missing Data	25275	582	—	—
MAR+Weight	30579	670	—	—
AN+Weight	28222	2789	—	—
AN+Constraint	24993	741	.82	[.79, .84]
AN+Constraint+Weight	25019	718	.83	[.80, .86]
SCINN	26905	719	—	—

(b) Survey-weighted estimates of α_0, α_{12}, α_{13}, γ_0, γ_{12}, γ_{13} and γ_2, along with the corresponding standard errors. "MAR+W" is MAR+Weight, "AN+W" is AN+Weights, "AN+C" is AN+Constraint, and "AN+C+W" is "AN+Constraint+Weight".

		MAR+W		AN+W		AN+C		AN+C+W		SCINN	
Par.	Truth	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
α_0	.50	.74	.05	.63	.13	.49	.05	.49	.05	.59	.05
α_{12}	-.50	-.45	.07	-.47	.07	-.49	.07	-.49	.06	-.49	.07
α_{13}	-1.00	-.88	.07	-.92	.10	-.98	.06	-.98	.06	-.98	.07
γ_0	-.25	-.88	.04	-.63	.35	-.22	.07	-.23	.07	-.42	.06
γ_{12}	.10	.29	.05	.21	.11	.10	.06	.10	.06	.17	.06
γ_{13}	.30	.63	.05	.48	.17	.27	.07	.27	.07	.40	.06
γ_2	-1.10	—	—	-.48	.57	-1.15	.14	-1.15	.13	-.69	.08

results. The AN+Constraint and AN+Constraint+Weight approaches outperforms the other choices under this scenario.

5.3.2 Overall margin for X_1, strong relationship between Y_1 and X_1 and weak nonignorable nonresponse

In this scenario, we again set $\alpha_0 = 0.5$ and $(\alpha_{12}, \alpha_{13}) = (-0.5, -1)$ to continue to reflect a strong relationship between Y_1 and X_1. However, we set $\gamma_0 = -1$, $(\gamma_{12}, \gamma_{13}) = (-0.6, 1.4)$, and $\gamma_2 = -0.2$ to reflect a weak nonignorable nonresponse mechanism. Table 5.3a shows the HT estimates for T_X and the standard error under each method. Table 5.3b shows survey-weighted estimates of α_0, α_{12}, α_{13}, γ_0, γ_{12},

γ_{13} and γ_2, along with the corresponding standard errors.

AN+Constraint and AN+Constraint+Weight once again outperform the other methods, although AN+Constraint now has a slightly smaller standard error for T_X than AN+Constraint+Weight. Interestingly, MAR+W actually outperforms SCINN somewhat here. In the presence of a weak nonignorable nonresponse mechanism, there appears to be little degradation when using an MAR model instead of the true AN model. In addition, whatever degradation or bias that should have been attributed to the survey design appears to be taken care of by including the strata indicator in the model for X_1. AN+Weight continues to perform worse than all other four methods. Unlike before, AN+Weight actually underestimates rather than overestimate T_X in this scenario. Overall, the range of acceptance ratios has decreased slightly from the previous scenario.

5.3.3 Overall margin for X_1, weak relationship between Y_1 and X_1 and strong nonignorable nonresponse

In these simulations, we weaken the relationship between Y_1 and X_1 but return to the case of a strong nonignorable nonresponse mechanism. We set $\alpha_0 = 0.15$ and $(\alpha_{12}, \alpha_{13}) = (-0.45, -0.15)$, but set $\gamma_0 = -0.25$, $(\gamma_{12}, \gamma_{13}) = (0.1, 0.3)$ and $\gamma_2 = -1.1$ as in Section 5.3.1. The results are presented in Tables 5.4a and 5.4b. The results and conclusions are quite similar to those in Section 5.3.1. AN+Constraint and AN+Constraint+Weight continue to remain the most accurate methods, whereas AN+Weight continues to perform worse than other methods. In fact, the standard errors are slightly higher for AN+Weight due to the weaker relationship between X_1 and Y_1.

Table 5.3: **Scenario two**: overall margin for X_1, strong relationship between Y_1 and X_1 and weak nonignorable nonresponse.

(a) HT estimates for T_X under each method, the corresponding standard errors, and acceptance ratios. "Population" is the value of T_X in the population of $N = 50000$ individuals. "No Missing Data" is the weighted estimate based on the sampled data before introducing item nonresponse.

	T_X		Acceptance Ratio	
Method	Mean	SE	Mean	Range
Population	25012	—	—	—
Mo Missing Data	25002	578	—	—
MAR+Weight	26027	673	—	—
AN+Weight	23380	2160	—	—
AN+Constraint	24979	706	.80	[.77, .81]
AN+Constraint+Weight	24985	712	.79	[.76, .81]
SCINN	26694	723	—	—

(b) Survey-weighted estimates of α_0, α_{12}, α_{13}, γ_0, γ_{12}, γ_{13} and γ_2, along with the corresponding standard errors. "MAR+W" is MAR+Weight, "AN+W" is AN+Weights, "AN+C" is AN+Constraint, and "AN+C+W" is "AN+Constraint+Weight".

		MAR+W		AN+W		AN+C		AN+C+W		SCINN	
Par.	Truth	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
α_0	.50	.51	.05	.41	.09	.48	.05	.48	.05	.53	.05
α_{12}	-.50	-.49	.06	-.42	.08	-.47	.06	-.47	.06	-.50	.06
α_{13}	-1.00	-.88	.07	-1.08	.18	-.97	.08	-.97	.08	-.83	.07
γ_0	-1.00	-1.11	.05	-.82	.24	-.99	.08	-.99	.08	-1.19	.07
γ_{12}	-.60	-.59	.07	-.73	.12	-.64	.07	-.63	.07	-.57	.07
γ_{13}	1.40	1.45	.06	1.27	.14	1.38	.06	1.39	.06	1.49	.06
γ_2	-.20	—	—	-.72	.59	-.19	.09	-.19	.09	-.11	.08

5.3.4 Overall margin for X_1, weak relationship between Y_1 and X_1 and weak nonignorable nonresponse

We again set $\alpha_0 = 0.15$ and $(\alpha_{12}, \alpha_{13}) = (-0.45, -0.15)$ as in Section 5.3.3, to continue to reflect a weak relationship between Y_1 and X_1. However, we now set $\gamma_0 = -1$, $(\gamma_{12}, \gamma_{13}) = (-0.6, 1.4)$ and $\gamma_2 = -0.2$ as in Section 5.3.2 to reflect a weak nonignorable nonresponse mechanism. The results are presented in Tables 5.5a and 5.5b. The overall conclusions are qualitatively similar to those in Section 5.3.2.

For the remaining simulation scenarios in Sections 5.3.5 to 5.3.8, we explore the performance of the same four approaches as before, but we do so in the context of

Table 5.4: **Scenario three**: overall margin for X_1, weak relationship between Y_1 and X_1 and strong nonignorable nonresponse.

(a) HT estimates for T_X under each method, the corresponding standard errors, and acceptance ratios. "Population" is the value of T_X in the population of $N = 50000$ individuals. "No Missing Data" is the weighted estimate based on the sampled data before introducing item nonresponse.

	T_X		Acceptance Ratio	
Method	Mean	SE	Mean	Range
Population	24695	—	—	—
Mo Missing Data	24816	571	—	—
MAR+Weight	30913	670	—	—
AN+Weight	27995	3170	—	—
AN+Constraint	24699	730	.82	[.80, .85]
AN+Constraint+Weight	24660	722	.82	[.79, .84]
SCINN	25860	714	—	—

(b) Survey-weighted estimates of α_0, α_{12}, α_{13}, γ_0, γ_{12}, γ_{13} and γ_2, along with the corresponding standard errors. "MAR+W" is MAR+Weight, "AN+W" is AN+Weights, "AN+C" is AN+Constraint, and "AN+C+W" is "AN+Constraint+Weight".

		MAR+W		AN+W		AN+C		AN+C+W		SCINN	
Par.	Truth	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
α_0	.15	.42	.05	.29	.15	.15	.05	.14	.05	.20	.05
α_{12}	-.45	-.41	.07	-.42	.07	-.44	.06	-.44	.06	-.44	.07
α_{13}	-.15	-.05	.07	-.09	.08	-.15	.06	-.15	.06	-.13	.06
γ_0	-.25	-.73	.04	-.52	.33	-.21	.06	-.21	.06	-.29	.05
γ_{12}	.10	.26	.05	.18	.10	.09	.06	.09	.06	.12	.06
γ_{13}	.30	.30	.05	.29	.06	.28	.06	.28	.06	.28	.05
γ_2	-1.10	—	—	-.60	.68	-1.15	.12	-1.16	.12	-.89	.08

auxiliary information by strata. In stratified sampling settings, just like we have here, we may know the auxiliary margins and totals by strata. When such information is available, it can help increase the precision of estimates. In this case, it is possible to implement the constraint in (5.7) for each stratum. Under our stratification simulations, the weight assigned to each unit is $w_i = N_s/n_s$, where N_s and n_s represent the number of observations assigned to stratum s in the population and sample respectively, with $s = 1, 2$. Then, for each stratum s, we require that

$$\sum_{\substack{S_i=s;\\ i\in\mathcal{D}}} w_i X^{\star}_{i1} = \frac{N_s}{n_s} \sum_{\substack{S_i=s;\\ i\in\mathcal{D}}} X^{\star}_{i1} \sim N(T_X^{(s)}, V_X^{(s)}), \tag{5.14}$$

Table 5.5: **Scenario four**: overall margin for X_1, weak relationship between Y_1 and X_1 and weak nonignorable nonresponse.

(a) HT estimates for T_X under each method, the corresponding standard errors, and acceptance ratios. "Population" is the value of T_X in the population of $N = 50000$ individuals. "No Missing Data" is the weighted estimate based on the sampled data before introducing item nonresponse.

	T_X		Acceptance Ratio	
Method	Mean	SE	Mean	Range
Population	24677	—	—	—
Mo Missing Data	24742	570	—	—
MAR+Weight	26098	662	—	—
AN+Weight	23705	2519	—	—
AN+Constraint	24666	698	.79	[.77, .81]
AN+Constraint+Weight	24653	705	.79	[.77, .81]
SCINN	25974	728	—	—

(b) Survey-weighted estimates of α_0, α_{12}, α_{13}, γ_0, γ_{12}, γ_{13} and γ_2, along with the corresponding standard errors. "MAR+W" is MAR+Weight, "AN+W" is AN+Weights, "AN+C" is AN+Constraint, and "AN+C+W" is "AN+Constraint+Weight".

		MAR+W		AN+W		AN+C		AN+C+W		SCINN	
Par.	Truth	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
α_0	.15	.19	.05	.12	.08	.15	.05	.15	.05	.18	.05
α_{12}	-.45	-.48	.06	-.43	.08	-.45	.06	-.45	.06	-.47	.06
α_{13}	-.15	-.04	.07	-.23	.21	-.15	.07	-.16	.07	-.05	.07
γ_0	-1.00	-1.12	.05	-.97	.21	-1.00	.06	-1.00	.06	-1.11	.06
γ_{12}	-.60	-.57	.07	-.64	.10	-.61	.07	-.61	.07	-.57	.07
γ_{13}	1.40	1.42	.06	1.41	.06	1.42	.06	1.42	.06	1.42	.06
γ_2	-.20	—	—	-.44	.47	-.23	.08	-.23	.08	-.02	.07

where $T_X^{(s)}$ is the auxiliary total of X_1^{pop} for strata s, and $V_X^{(s)}$ is the corresponding variance associated with it. For the AN+Constraint and AN+Constraint+Weight methods, we implement this constraint by applying the Metropolis steps S1 to S3 in Section 5.2 within each stratum.

For the SCINN method, we augment with the auxiliary information within strata by generating the synthetic observations by stratum. That is, we ensure that the empirical conditional distribution X_1 in each stratum s matches the conditional probability $\Pr(X_1^{pop} = 1|S_i = s)$ in the population.

Table 5.6: **Scenario five**: margin for X_1 within each stratum, strong relationship between Y_1 and X_1 and strong nonignorable nonresponse.

(a) HT estimates for T_X under each method, the corresponding standard errors, and acceptance ratios by strata. "Population" is the value of T_X in the population of $N = 50000$ individuals. "No Missing Data" is the weighted estimate based on the sampled data before introducing item nonresponse.

	T_X		Acceptance Ratio			
			Stratum 1		Stratum 2	
Method	Mean	SE	Mean	Range	Mean	Range
Population	24994	—	—	—	—	—
Mo Missing Data	25043	580	—	—	—	—
MAR+Weight	30447	668	—	—	—	—
AN+Weight	28488	3034	—	—	—	—
AN+Constraint	25062	665	.81	[.66, .91]	.80	[.74, .83]
AN+Constraint+Weight	25070	667	.80	[.61, .90]	.79	[.74, .84]
SCINN	25105	702	—	—	—	—

(b) Survey-weighted estimates of α_0, α_{12}, α_{13}, γ_0, γ_{12}, γ_{13} and γ_2, along with the corresponding standard errors. "MAR+W" is MAR+Weight, "AN+W" is AN+Weights, "AN+C" is AN+Constraint, and "AN+C+W" is "AN+Constraint+Weight".

		MAR+W		AN+W		AN+C		AN+C+W		SCINN	
Par.	Truth	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
α_0	.50	.74	.05	.64	.13	.50	.05	.50	.05	.50	.05
α_{12}	-.50	-.45	.07	-.46	.08	-.50	.07	-.50	.07	-.50	.07
α_{13}	-1.00	-.89	.07	-.90	.12	-1.00	.06	-1.00	.07	-1.00	.06
γ_0	-.25	-.89	.04	-.73	.44	-.27	.06	-.27	.06	-.27	.06
γ_{12}	.10	.30	.05	.22	.11	.12	.06	.12	.06	.12	.06
γ_{13}	.30	.65	.05	.52	.19	.31	.06	.31	.06	.31	.06
γ_2	-1.10	—	—	-.41	.69	-1.08	.09	-1.08	.09	-1.08	.10

5.3.5 Margin for X_1 within each stratum, strong relationship between Y_1 and X_1 and strong nonignorable nonresponse

First, we set the parameters exactly as in Section 5.3.1 to reflect a strong relationship between Y_1 and X_1, and strong nonignorable nonresponse mechanism. Table 5.6a shows the HT estimates for T_X, the standard error under each method and the acceptance ratios by strata. Table 5.6b again shows survey-weighted estimates of α_0, α_{12}, α_{13}, γ_0, γ_{12}, γ_{13} and γ_2, and the corresponding standard errors. The overall conclusions are qualitatively similar to those in Section 5.3.1. In particular, incorporating

the auxiliary margin by strata in AN+Constraint and AN+Constraint+Weight reduces the standard errors. Augmenting the data by strata in the SCINN method also greatly improves the accuracy of its estimates; SCINN estimates are much closer to those from AN+Constraint and AN+Constraint+Weight here, although with slightly larger standard errors. The range of acceptance ratios are much wider suggesting that there is a smaller set of combinations of imputed values that fulfill the constraints within each stratum, than with the combined constraint.

5.3.6 Margin for X_1 within each stratum, strong relationship between Y_1 and X_1 and weak nonignorable nonresponse

We set the parameters exactly as in Section 5.3.2 to reflect a strong relationship between Y_1 and X_1, but a weak nonignorable nonresponse mechanism. The results are presented in Tables 5.7a and 5.7b, and the conclusions are qualitatively similar to those in Sections 5.3.2 and 5.3.5. As before, implementing the constraint by strata reduces the standard errors for AN+Constraint and AN+Constraint+Weight.

5.3.7 Margin for X_1 within each stratum, weak relationship between Y_1 and X_1 and strong nonignorable nonresponse

We set the parameters as in Section 5.3.3 to reflect a weak relationship between Y_1 and X_1, but a strong nonignorable nonresponse mechanism. The results are presented in Tables 5.8a and 5.8b, and they are consistent with the results from all the previous simulations.

5.3.8 Margin for X_1 within each stratum, weak relationship between Y_1 and X_1 and weak nonignorable nonresponse

Finally, we set the parameters as in Section 5.3.4 to reflect a weak relationship between Y_1 and X_1, and a weak nonignorable nonresponse mechanism as well. We present the results in Tables 5.9a and 5.9b, and they are also consistent with the results from all the other simulations.

Table 5.7: **Scenario six**: margin for X_1 within each stratum, strong relationship between Y_1 and X_1 and weak nonignorable nonresponse.

(a) HT estimates for T_X under each method, the corresponding standard errors, and acceptance ratios by strata. "Population" is the value of T_X in the population of $N = 50000$ individuals. "No Missing Data" is the weighted estimate based on the sampled data before introducing item nonresponse.

	T_X		Acceptance Ratio			
			Stratum 1		Stratum 2	
Method	Mean	SE	Mean	Range	Mean	Range
Population	24976	—	—	—	—	—
Mo Missing Data	24970	578	—	—	—	—
MAR+Weight	26305	673	—	—	—	—
AN+Weight	24056.25	2523.97	—	—	—	—
AN+Constraint	24907	671	.84	[.74, .89]	.78	[.73, .81]
AN+Constraint+Weight	24894	662	.81	[.67, .88]	.76	[.69, .80]
SCINN	24940	729	—	—	—	—

(b) Survey-weighted estimates of α_0, α_{12}, α_{13}, γ_0, γ_{12}, γ_{13} and γ_2, along with the corresponding standard errors. "MAR+W" is MAR+Weight, "AN+W" is AN+Weights, "AN+C" is AN+Constraint, and "AN+C+W" is "AN+Constraint+Weight".

		MAR+W		AN+W		AN+C		AN+C+W		SCINN	
Par.	Truth	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
α_0	.50	.56	.05	.47	.10	.52	.05	.52	.05	.52	.05
α_{12}	-.50	-.54	.06	-.48	.09	-.52	.06	-.52	.06	-.52	.06
α_{13}	-1.00	-.93	.07	-1.10	.21	-1.05	.07	-1.05	.07	-1.04	.08
γ_0	-1.00	-1.12	.05	-.87	.30	-.95	.07	-.95	.07	-.95	.07
γ_{12}	-.60	-.56	.07	-.68	.14	-.62	.07	-.62	.07	-.61	.07
γ_{13}	1.40	1.48	.06	1.31	.17	1.39	.06	1.39	.06	1.39	.06
γ_2	-.20	—	—	-.61	.66	-.26	.07	-.26	.07	-.26	.09

5.4 Discussion

The results presented across all simulation scenarios are based on limited scenarios, but they suggest that our approach provides a promising approach for incorporating survey weights and auxiliary information when imputing nonresponse in complex surveys. In particular, AN+Constraint and AN+Constraint+Weight appear to outperform both SCINN and the default option of controlling for the weights in the joint model for the variables in $\mathcal{D}$. The different simulation settings also show when the other methods might be reasonable. For example, the MAR+Weight approach

Table 5.8: **Scenario seven**: margin for X_1 within each stratum, weak relationship between Y_1 and X_1 and strong nonignorable nonresponse.

(a) HT estimates for T_X under each method, the corresponding standard errors, and acceptance ratios by strata. "Population" is the value of T_X in the population of $N = 50000$ individuals. "No Missing Data" is the weighted estimate based on the sampled data before introducing item nonresponse.

	T_X		Acceptance Ratio			
			Stratum 1		Stratum 2	
Method	Mean	SE	Mean	Range	Mean	Range
Population	24683	—	—	—	—	—
Mo Missing Data	24609	568	—	—	—	—
MAR+Weight	30660	665	—	—	—	—
AN+Weight	29494	3025	—	—	—	—
AN+Constraint	24643	659	.78	[.67, .89]	.79	[.74, .84]
AN+Constraint+Weight	24650	663	.74	[.61, .88]	.78	[.72, .84]
SCINN	24876	708	—	—	—	—

(b) Survey-weighted estimates of α_0, α_{12}, α_{13}, γ_0, γ_{12}, γ_{13} and γ_2, along with the corresponding standard errors. "MAR+W" is MAR+Weight, "AN+W" is AN+Weights, "AN+C" is AN+Constraint, and "AN+C+W" is "AN+Constraint+Weight".

		MAR+W		AN+W		AN+C		AN+C+W		SCINN	
Par.	Truth	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
α_0	.15	.42	.05	.37	.14	.16	.05	.16	.05	.17	.05
α_{12}	-.45	-.43	.07	-.42	.08	-.45	.07	-.45	.06	-.46	.06
α_{13}	-.15	-.08	.07	-.09	.08	-.18	.06	-.18	.06	-.17	.06
γ_0	-.25	-.74	.04	-.72	.35	-.23	.05	-.23	.05	-.25	.05
γ_{12}	.10	.25	.05	.22	.10	.08	.06	.08	.06	.09	.06
γ_{13}	.30	.31	.05	.30	.05	.28	.06	.28	.06	.29	.06
γ_2	-1.10	—	—	-.18	.61	-1.10	.08	-1.10	.09	-1.05	.09

yields decent results when the nonresponse mechanism is only weakly nonignorable; this method should perform even better for fully ignorable nonresponse mechanisms. However, the results based on AN+Constraint and AN+Constraint+Weight are the most consistent across the different scenarios.

Opportunities for extensions of this approach exist as future research topics. First, the constraint we propose can be incorporated into scenarios with unit nonresponse. This can be done by following our SCINN framework in Chapter 4, but replacing the synthetic data augmentation step used to incorporate the auxiliary

Table 5.9: **Scenario eight**: margin for X_1 within each stratum, weak relationship between Y_1 and X_1 and weak nonignorable nonresponse.

(a) HT estimates for T_X under each method, the corresponding standard errors, and acceptance ratios by strata. "Population" is the value of T_X in the population of $N = 50000$ individuals. "No Missing Data" is the weighted estimate based on the sampled data before introducing item nonresponse.

	T_X		Acceptance Ratio			
			Stratum 1		Stratum 2	
Method	Mean	SE	Mean	Range	Mean	Range
Population	24724	—	—	—	—	—
Mo Missing Data	24613	569	—	—	—	—
MAR+Weight	25969	669	—	—	—	—
AN+Weight	23551	3038	—	—	—	—
AN+Constraint	24710	651	.79	[.60, .87]	.76	[.68, .79]
AN+Constraint+Weight	24689	672	.77	[.59, .86]	.75	[.66, .78]
SCINN	24811	710	—	—	—	—

(b) Survey-weighted estimates of α_0, α_{12}, α_{13}, γ_0, γ_{12}, γ_{13} and γ_2, along with the corresponding standard errors. "MAR+W" is MAR+Weight, "AN+W" is AN+Weights, "AN+C" is AN+Constraint, and "AN+C+W" is "AN+Constraint+Weight".

		MAR+W		AN+W		AN+C		AN+C+W		SCINN	
Par.	Truth	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
α_0	.15	.18	.05	.12	.09	.15	.05	.15	.05	.15	.05
α_{12}	-.45	-.48	.06	-.44	.08	-.46	.06	-.46	.06	-.47	.06
α_{13}	-.15	-.05	.07	-.24	.25	-.14	.07	-.14	.07	-.13	.07
γ_0	-1.00	-1.09	.05	-.97	.26	-.98	.06	-.98	.06	-.99	.06
γ_{12}	-.60	-.60	.07	-.68	.12	-.64	.07	-.64	.07	-.63	.07
γ_{13}	1.40	1.38	.05	1.37	.06	1.38	.06	1.38	.06	1.38	.05
γ_2	-.20	—	—	.50	.62	-.21	.07	-.21	.07	-.19	.08

margins, with the constraint proposed in this chapter. The steps in the SCINN framework still should be followed to ensure that the specified models are identifiable.

Second, our simulations only focus on stratified data. Future work would explore extensions to other sampling techniques, as well as weights with many unique values. Preliminary simulations, not shown here, show that our approach could work for many valued, unequal weights. Also, in the case of many valued, unequal weights in particular, generating plausible imputations that satisfy the constraint can be chal-

lenging whenever the set of combinations of imputed values that pass the constraint is small, compared to the set of all possible combinations. When this is the case, it is necessary to develop a more efficient sampler for generating proposals for the imputations. One possible solution could be to generate proposals for missing entries within weight bins. Essentially, units might likely behave more like other units with similar weights, than every other unit, so that it could be plausible to leverage this relationship when proposing imputations. Future work would explore this idea.

Finally, when applying our approach to data suffering from unit nonresponse, survey weights must be available for unit nonrespondents. Data collection agencies often only release them for respondents. When this is the case, one possibility for generating weights for the nonrespondents could be to include a model for the weights to allow us impute them for unit nonrespondents. This can be done using methods akin to those proposed by Si et al. (2015). In practice, agencies also often release adjusted weights (adjusted for unit nonresponse or poststratification) instead of the base weights. Future work would explore the extension of our approach to adjusted weights as well.

6

Conclusions

In this thesis, I present Bayesian approaches for handling missing and erroneous data in surveys. Chapter 2 presented an imputation engine for dealing with missing values in household data containing structural zero. Chapter 3 developed an edit and imputation model for dealing with erroneous responses in household data when structural zero are again present. Chapter 4 presented a framework for leveraging auxiliary data in specifying conditional nonignorable models for unit and item nonresponse. Chapter 5 presented an approach for incorporating survey design variables, specifically survey weights, in complex surveys, within the framework of Chapter 4.

Looking to the future, the methodology presented in Chapters 2 to 5 can be eventually combined into a coherent imputation engine that handles missing data and reporting error, leverages auxiliary information on marginal distributions, and incorporates survey weights when specifying nonignorable models for unit and item nonresponse. This imputation engine can be implemented within the multiple imputation framework to allow data collection agencies release multiply-imputed datasets to the public. To the best of my knowledge, there is currently no imputation methodology that can handle all these challenges simultaneously.

The work presented in this thesis also suggests directions for future work, in addition to those already mentioned in each chapter. First, the NDPMPM model used for imputing missing data in Chapter 2, and by extension the EIHD model in Chapter 3, have limitations that affect their scalability. I use multiple rejection sampling steps in the Gibbs samplers in both chapters, and these steps become expensive as the household sizes increases. Future work could investigate solutions to this problem, in addition to the parallelization already discussed in the chapters. One potential solution is to implement a hybrid method, for example, generating imputations for the large households with missing or faulty data from a fixed set of proposals generated using an ad-hoc imputation method such as a hot deck. Additionally, variables that have a large number of categories, such as age, also slow down the sampler. One could develop other alternatives to treating such variables as categorical. With age for example, an alternative is to treat it as a continuous variable. Treating age as a continuous variable requires replacing the NDPMPM model with a model suitable for mixed data types.

Second, so far, I have only used the NDPMPM and EIHD models to handle ignorable missing data and measurement error mechanisms. In Chapter 3, for example, I use independent uniform distributions for the reporting model and independent Bernoulli distributions for the error locations to represent a default model for lack of knowledge about the nature of the measurement error. However, datasets may contain nonignorable mechanisms. For example, in clinical data, patients may drop out of a drug trial because the drug itself causes harm to their health. In such cases, the missing values clearly depend on the value of the responses themselves. Future work would develop nonignorable extensions to the NDPMPM and EIHD.

Third, the constraint proposed in Chapter 5 is only one way to incorporate the survey design. Future work would investigate including the weights directly in the likelihood function. For example, one possibility could be by raising the contribution

from each unit in the sample to the probability of including that unit in the survey, that is the inverse of its corresponding base weight, similar to ideas used in weighted likelihood methods (Newton and Raftery, 1994).

Finally, the focus of the work presented in this thesis is on categorical data. Most of these ideas should extend to continuous and mixed data.

Appendix A

List of structural zeros

I fit the models in Chapters 2 and 3 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 ACS by identifying combinations involving the relationship variable that do not appear in the constructed populations. This list should not be interpreted as a "true" list of impossible combinations in census data. They only reflect what was present in the data.

Table A.1: List of structural zeros.

	Description
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.
8.	The household head must be older than the oldest biological child by at least 7.
9.	The household head must be older than the oldest adopted child by at least 11.
10.	The household head must be older than the oldest stepchild by at least 9.

Bibliography

Akande, O., Li, F., and Reiter, J. (2017), "An Empirical Comparison of Multiple Imputation Methods for Categorical Data," *The American Statistician*, 71, 162–170.

Akande, O., Barrientos, A., and Reiter, J. P. (2018), "Simultaneous Edit and Imputation For Household Data with Structural Zeros," *Journal of Survey Statistics and Methodology*.

Akande, O., Reiter, J., and Barrientos, A. F. (2019), "Multiple Imputation of Missing Values in Household Data with Structural Zeros," *Survey Methodology*.

Alanya, A., Wolf, C., and Sotto, C. (2015), "Comparing multiple imputation and propensity score weighting in unit nonresponse adjustments," *Public Opinion Quarterly*, 79, 635–661.

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

Arnold, B. C. and Press, S. J. (1989), "Compatible Conditional Distributions," *Journal of the American Statistical Association*, 84, 152–156.

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

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*.

Bhattacharya, D. (2008), "Inference in panel data models under attrition caused by unobservables," *Journal of Econometrics*, 144, 430–446.

Biemer, P. P., Chen, P., and Wang, K. (2013), "Using level-of-effort paradata in non-response adjustments with application to field surveys," *Journal of the Royal Statistical Society: Series A*, 176, 147–168.

Brick, J. M. and Kalton, G. (1996), "Handling missing data in survey research," *Statistical Methods in Medical Research*, 5, 215–238.

Brick, J. M. and Williams, D. (2013), "Explaining rising nonresponse rates in cross-sectional surveys," *The ANNALS of the American Academy of Political and Social Science*, 645, 36–59.

van Buuren, S. (2007), "Multiple imputation of discrete and continuous data by fully conditional specification," *Statistical Methods in Medical Research*, 16, 219–242.

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

Curtin, R., Presser, S., and Singer, E. (2005), "Changes in telephone survey nonresponse over the past quarter century," *Public Opinion Quarterly*, 69, 87–98.

Das, M., Toepel, V., and van Soest, A. (2013), "Nonparametric Tests of Panel Conditioning and Attrition Bias in Panel Surveys," *Sociological Methods and Research*, 40, 32–56.

DeBell, M., Krosnick, J. A., Gera, K., Yeager, D. S., and McDonald, M. P. (2018), "The Turnout Gap in Surveys: Explanations and Solutions," *Sociological Methods & Research*, 0.

Deng, Y., Hillygus, D. S., Reiter, J. P., Si, Y., and Zheng, S. (2013), "Handling Attrition in Longitudinal Studies: The Case for Refreshment Samples," *Statist. Sci.*, 28, 238–256.

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

Dwork, C. (2006), *Differential privacy*, Automata, languages and programming, Springer.

Fellegi, I. P. and Holt, D. (1976), "A Systematic Approach to Automatic Edit and Imputation," *Journal of the American Statistical Association*, 71, 17–35.

Fellegi, I. P. and Sunter, A. B. (1969), "A theory for record linkage," *Journal of the American Statistical Association*, 64, 1183–1210.

Fienberg, S. E. (2010), "The Relevance or Irrelevance of Weights for Confidentiality and Statistical Analyses," *Journal of Privacy and Confidentiality*, 1, 183–195.

Fuller, W. A. (2009), *Probability Sampling from a Finite Universe*, chap. 1, pp. 1–93, John Wiley & Sons, Ltd.

Gelman, A. (2007), "Struggles with Survey Weighting and Regression Modeling," *Statist. Sci.*, 22, 153–164.

Gelman, A. and Speed, T. P. (1993), "Characterizing a joint probability distribution by conditionals," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 55, 185–188.

Ghosh-Dastidar, B. and Schafer, J. L. (2003), "Multiple Edit/Multiple Imputation for Multivariate Continuous Data," *Journal of the American Statistical Association*, 98, 807–817.

Glynn, R. J., Laird, N. M., and Rubin, D. B. (1986), *Selection Modeling Versus Mixture Modeling with Nonignorable Nonresponse*, pp. 115–142, Springer New York, New York, NY.

Gouweleeuw, J. M., Kooiman, P., Willenborg, L. C. R. J., and de Wolf, P.-P. (1998), "Post Randomisation for Statistical Disclosure Control: Theory and Implementation," *Journal of Official Statistics*, 14, 463–478.

Groves, R. M. (2006), "Nonresponse rates and nonresponse bias in household surveys," *Public Opinion Quarterly*, 70, 646–675.

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

Hausman, J. and Wise, D. (1979), "Attrition bias in experimental and panel data: The Gary income maintenance experiment," *Econometrica*, 47, 455–473.

Herzog, T. N., Scheuren, F. J., and Winkler, W. E. (2007), *Data Quality and Record Linkage*, Springer.

Hirano, K., Imbens, G., Ridder, G., and Rubin, D. (1998), "Combining panel data sets with attrition and refreshment samples," *Technical Working Paper 230*.

Hirano, K., Imbens, G., Ridder, G., and Rubin, D. (2001), "Combining panel data sets with attrition and refreshment samples," *Econometrica*, 69, 1645–1659.

Holbein, J. B. and Hillygus, D. S. (2016), "Making young voters: the impact of preregistration on youth turnout," *American Journal of Political Science*, 60, 364–382.

Horvitz, D. G. and Thompson, D. J. (1952), "A Generalization of Sampling Without Replacement from a Finite Universe," *Journal of the American Statistical Association*, 47, 663–685.

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.

Hur, A. and Achen, C. H. (2013), "Coding voter turnout responses in the Current Population Survey," *Public Opinion Quarterly*, 77, 985–993.

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

Jackman, S. and Spahn, B. (2014), "Why Does the American National Election Study Overestimate Voter Turnout?" *Political Analysis*, pp. 1–15.

Jackson, A. (2010), "2010 Census Item Nonresponse and Imputation Assessment Report," *U.S. Census Bureau*.

de Jonge, E. and van der Loo, M. (2015), "editrules: Parsing, Applying, and Manipulating Data Cleaning Rules," *The Comprehensive R Archive Network*.

Kalton, G. and Kasprzyk, D. (1986), "The treatment of missing survey data," *Survey Methodology*, 12, 1–16.

Kim, H. J., Cox, L. H., Karr, A. F., Reiter, J. P., and Wang, Q. (2015a), "Simultaneous Edit-Imputation for Continuous Microdata," *Journal of the American Statistical Association*, 110, 987–999.

Kim, H. J., Karr, A. F., and Reiter, J. P. (2015b), "Statistical disclosure limitation in the presence of edit rules," *Journal of Official Statistics*, 31, 121–138.

Kim, H. J., Reiter, J. P., and Karr, A. F. (2018), "Simultaneous edit-imputation and disclosure limitation for business establishment data," *Journal of Applied Statistics*, 45, 63–82.

Kim, J. K. (2011), "Parametric fractional imputation for missing data analysis," *Biometrika*, 98, 119–132.

Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T. M., Casas-Cordero, C., Lemay, M., Peytchev, A., Groves, R. M., and Raghunathan, T. E. (2010), "Using proxy measures and other correlates of survey outcomes to adjust for non-response: examples from multiple surveys," *Journal of the Royal Statistical Society: Series A*, 173, 389–407.

Krueger, B. S. and West, B. T. (2014), "Assessing the potential of paradata and other auxiliary information for nonresponse adjustments," *Public Opinion Quarterly*, 78, 795–831.

Linero, A. R. and Daniels, M. J. (2018), "Bayesian Approaches for Missing Not at Random Outcome Data: The Role of Identifying Restrictions," *Statist. Sci.*, 33, 198–213.

Little, R. J. A. (1993a), "Pattern-mixture models for multivariate incomplete data," *Journal of the American Statistical Association*, 88, 125–134.

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

Little, R. J. A. (1995), "Modeling the Drop-Out Mechanism in Repeated-Measures Studies," *Journal of the American Statistical Association*, 90, 1112–1121.

Little, R. J. A. and Rubin, D. B. (2002), *Statistical analysis with missing data*, New York: John Wiley & Sons, New Jersey.

Little, R. J. A. and Vartivarian, S. (2005), "Does weighting for nonresponse increase the variance of survey means?" *Survey Methodology*, 31, 161–168.

Lohr, S. L. (2010), *Sampling: Design and Analysis*, Brooks/Cole, Cengage Learning, Boston, MA, USA, 2 edn.

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.

Manrique-Vallier, D. and Reiter, J. P. (2018), "Bayesian simultaneous edit and imputation for multivariate categorical data," *Journal of the American Statistical Association*, 112, 1708–1719.

Marker, D. A., Judkins, D. R., and Wingless, M. (2002), "Large-scale imputation for complex surveys," *Survey Nonresponse*, pp. 329–341.

McDonald, M. (2008), "United States election project," *United States Elections Project*.

McDonald, M. P. and Popkin, S. L. (2001), "The myth of the vanishing voter," *American Political Science Review*, 95, 963–974.

Mealli, F. and Rubin, D. B. (2015), "Clarifying missing at random and related definitions, and implications when coupled with exchangeability," *Biometrika*, 102, 995–1000.

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*.

National Research Council (2009), *Reengineering the Survey of Income and Program Participation*, Panel on the Census Bureau Reengineered Survey of Income and Program Participation, Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington DC: The National Academies Press.

National Research Council (2015), *Realizing the Potential of the American Community Survey*, Panel on Addressing Priority Technical Issues for the Next Decade of the American Community Survey, Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington DC: The National Academies Press.

Nevo, A. (2003), "Using weights to adjust for sample selection when auxiliary information is available," *Journal of Business and Economic Statistics*, 21, 43–52.

Newton, M. A. and Raftery, A. E. (1994), "Approximate Bayesian Inference with the Weighted Likelihood Bootstrap," *Journal of the Royal Statistical Society. Series B (Methodological)*, 56, 3–48.

Peytchev, A. (2012), "Multiple imputation for unit nonresponse and measurement error," *Public Opinion Quarterly*, 76, 214–237.

Peytcheva, E. and Groves, R. M. (2009), "Using variation in response rates of demographic subgroups as evidence of nonresponse bias in survey estimates," *Journal of Official Statistics*, 25, 193–201.

Pfeffermann, D. (1993), "The Role of Sampling Weights When Modeling Survey Data," *International Statistical Review*, 61, 317–337.

Powell, G. B. (1986), "American voter turnout in comparative perspective," *American Political Science Review*, 80, 17–43.

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

Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003), "Multiple imputation for statistical disclosure limitation," *Journal of Official Statistics*, 19, 1–16.

Reiter, J. P. (2003), "Inference for partially synthetic, public use microdata sets," *Survey Methodology*, 29, 181–189.

Reiter, J. P. (2005), "Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study," *Journal of the Royal Statistical Society, Series A*, 168, 185–205.

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

Reiter, J. P., Raghunathan, T. E., and Kinney, S. (2006), "The importance of modeling the sampling design in multiple imputation for missing data," *Survey Methodology*, 32, 143–150.

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

Rubin, D. B. (1987), *Multiple imputation for nonresponse in surveys*, John Wiley & Sons, New York.

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

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

Sadinle, M. and Reiter, J. P. (2017), "Itemwise conditionally independent nonresponse modelling for incomplete multivariate data," *Biometrika*, 104, 207–220.

Sadinle, M. and Reiter, J. P. (2019), "Sequentially additive nonignorable missing data modeling using auxiliary marginal information," *Biometrika*.

Sakshaug, J. W. and Kreuter, F. (2012), "Assessing the magnitude of non-consent bias in linked survey and administrative data," *Survey Research Methods*, 6, 113–122.

Savitsky, T. D. and Toth, D. (2016), "Bayesian estimation under informative sampling," *Electronic Journal of Statistics*, 10, 1677–1708.

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

Schifeling, T. S. and Reiter, J. P. (2016), "Incorporating marginal prior information into latent class models," *Bayesian Analysis*, 11, 499–518.

Schifeling, T. S., Cheng, C., Reiter, J. P., and Hillygus, D. S. (2015), "Accounting for nonignorable unit nonresponse and attrition in panel studies with refreshment samples," *Journal of Survey Statistics and Methodology*, 3, 265–295.

Schouten, B., Mushkudiani, N., Shlomo, N., Durrant, G., Lundquist, P., and Wagner, J. (2018), "A Bayesian Analysis of Design Parameters in Survey Data Collection," *Journal of Survey Statistics and Methodology*, 6, 431–464.

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, 199–521.

Si, Y., Pillai, N. S., and Gelman, A. (2015), "Bayesian Nonparametric Weighted Sampling Inference," *Bayesian Anal.*, 10, 605–625.

Sinibaldi, J., Trappmann, M., and Kreuter, F. (2014), "Which is the better investment for nonresponse adjustment: Purchasing commercial auxiliary data or collecting interviewer observations?" *Public Opinion Quarterly*, 78, 440–473.

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

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

de Waal, T. and Coutinho, W. (2005), "Automatic Editing for Business Surveys: An Assessment of Selected Algorithms," *International Statistical Review*, 73, 73–102.

de Waal, T., Pannekoek, J., and Scholtus, S. (2011), *Handbook of Statistical Data Editing and Imputation*, John Wiley & Sons, Inc.

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*.

West, B. T. and Kreuter, F. (2013), "Factors affecting the accuracy of interviewer observations evidence from the National Survey of Family Growth," *Public Opinion Quarterly*, 77, 522–548.

West, B. T. and Little, R. J. A. (2013), "Non-response adjustment of survey estimates based on auxiliary variables subject to error," *Journal of the Royal Statistical Society, Series C*, 62, 213–231.

Winkler, W. (1995), "Editing Discrete Data," in *Proceedings of the Section on Survey Research Methods*, pp. 108–113, American Statistical Association.

Winkler, W. and Petkunas, T. F. (1997), "The DISCRETE edit system," *Statistical Data Editing*, 2, 52–62.

Zhou, H., R., E. M., and Raghunathan, T. E. (2016), "Synthetic multiple imputation procedure for multistage complex samples," *Journal of Official Statistics*, 32, 231–236.

Biography

Olanrewaju Michael Akande received his BSc in Mathematics and Statistics from the University of Lagos, Nigeria in 2010. Prior to attending Duke University, He worked as an analyst at KPMG Professional Services, Nigeria between 2011 and 2012. He also obtained an MSc in Statistical and Economic modeling from Duke University in 2015. He plans to graduate with his PhD in Statistical Science in May 2019, under the supervision of Professor Jerome P. Reiter. He also plans to graduate with a Certificate in College Teaching program in May 2019. After graduation, he will join the faculty of the Masters in Interdisciplinary Data Science program, within the Social Science Research Institute, at Duke University.