Bayesian Analysis (2018) **13**, Number 1, pp. 183–200

Dirichlet Process Mixture Models for Modeling and Generating Synthetic Versions of Nested Categorical Data

Jingchen Hu[*,§], Jerome P. Reiter[†,§,¶,‖], and Quanli Wang[‡,¶,‖]

Abstract. We present a Bayesian model for estimating the joint distribution of multivariate categorical data when units are nested within groups. Such data arise frequently in social science settings, for example, people living in households. The model assumes that (i) each group is a member of a group-level latent class, and (ii) each unit is a member of a unit-level latent class nested within its group-level latent class. This structure allows the model to capture dependence among units in the same group. It also facilitates simultaneous modeling of variables at both group and unit levels. We develop a version of the model that assigns zero probability to groups and units with physically impossible combinations of variables. We apply the model to estimate multivariate relationships in a subset of the American Community Survey. Using the estimated model, we generate synthetic household data that could be disseminated as redacted public use files. Supplementary materials (Hu et al., 2017) for this article are available online.

Keywords: confidentiality, disclosure, latent, multinomial, synthetic.

1 Introduction

In many settings, the data comprise units nested within groups (e.g., people within households), and include categorical variables measured at the unit level (e.g., individuals' demographic characteristics) and at the group level (e.g., whether the family owns or rents their home). A typical analysis goal is to estimate multivariate relationships among the categorical variables, accounting for the hierarchical structure in the data.

To estimate joint distributions with multivariate categorical data, many analysts rely on mixtures of products of multinomial distributions, also known as latent class models. These models assume that each unit is a member of an unobserved cluster, and that variables follow independent multinomial distributions within clusters. Latent class models can be estimated via maximum likelihood (Goodman, 1974) and Bayesian approaches (Ishwaran and James, 2001; Jain and Neal, 2007; Dunson and Xing, 2009). Of particular note, Dunson and Xing (2009) present a nonparametric Bayesian version

[*]Department of Mathematics and Statistics, Vassar College, Box 27, Poughkeepsie, NY 12604, jihu@vassar.edu
[†]Department of Statistical Science, Duke University, Durham, NC 27708-0251, jerry@stat.duke.edu
[‡]Department of Statistical Science, Duke University, Durham, NC 27708-0251, quanli@stat.duke.edu
[§]National Science Foundation CNS-10-12141.
[¶]National Science Foundation SES-11-31897.
[‖]Arthur P. Sloan Foundation G-2015-2-166003.

 https://doi.org/10.1214/16-BA1047

of the latent class model, using a Dirichlet process mixture (DPM) for the prior distribution. The DPM prior distribution is appealing, in that (i) it has full support on the space of joint distributions for unordered categorical variables, ensuring that the model does not restrict dependence structures *a priori*, and (ii) it fully incorporates uncertainty about the effective number of latent classes in posterior inferences.

For data nested within groups, however, standard latent class models may not offer accurate estimates of joint distributions. In particular, it may not be appropriate to treat the units in the same group as independent; for example, demographic variables like age, race, and sex of individuals in the same household are clearly dependent. Similarly, some combinations of units may be physically impossible to place in the same group, such as a daughter who is older than her biological father. Additionally, every unit in a group must have the same values of group-level variables, so that one cannot simply add multinomial kernels for the group-level variables.

In this article, we present a Bayesian mixture model for nested categorical data. The model assumes that (i) each group is a member of a group-level latent class, and (ii) each unit is a member of a unit-level latent class nested within its group-level latent class. This structure encourages the model to cluster groups into data-driven types, for example, households with children where everyone has the same race. This in turn allows for dependence among units in the same group. The nested structure also facilitates simultaneous modeling of variables at both group and unit levels. We refer to the model as the nested data Dirichlet process mixture of products of multinomial distributions (NDPMPM). We present two versions of the NDPMPM: one that gives support to all configurations of groups and units, and one that assigns zero probability to groups and units with physically impossible combinations of variables (also known as structural zeros in the categorical data analysis literature).

The NDPMPM is similar to the latent class models proposed by Vermunt (2003, 2008), who also uses two layers of latent classes to model nested categorical data. These models use a fixed number of classes as determined by a model selection criterion (e.g., AIC or BIC), whereas the NDPMPM allows uncertainty in the effective number of classes at each level. The NDPMPM also is similar to the latent class models in Bennink et al. (2016) for nested data, especially to what they call the "indirect model." The indirect model regresses a single group-level outcome on group-level and individual-level predictors, whereas the NDPMPM is used for estimation of the joint distribution of multiple group-level and individual-level variables. To the best of our knowledge, the models of Vermunt (2003, 2008) and Bennink et al. (2016) do not account for groups with physically impossible combinations of units.

One of our primary motivations in developing the NDPMPM is to develop a method for generating redacted public use files for household data, specifically for the variables on the United States decennial census. Public use files in which confidential data values are replaced with draws from predictive distributions are known in the disclosure limitation literature as synthetic datasets (Rubin, 1993; Little, 1993; Raghunathan et al., 2003; Reiter, 2005; Reiter and Raghunathan, 2007). Synthetic data techniques have been used to create several high-profile public use data products, including the Survey of Income and Program Participation (Abowd et al., 2006), the Longitudinal Business

Database (Kinney et al., 2011), the American Community Survey group quarters data (Hawala, 2008), and the OnTheMap application (Machanavajjhala et al., 2008). None of these products involve synthetic household data. In these products, the synthesis strategies are based on chains of generalized linear models for independent individuals, e.g., simulate variable x_1 from some parametric model $f(x_1)$, x_2 from some parametric model $f(x_2|x_1)$, etc. We are not aware of any synthesis models appropriate for nested categorical data like the decennial census variables.

As part of generating the synthetic data, we evaluate disclosure risks using the measures suggested in Hu et al. (2014). Specifically, we quantify the posterior probabilities that intruders can learn values from the confidential data given the released synthetic data, under assumptions about the intruders' knowledge and attack strategy. This is the only strategy we know of for evaluating statistical disclosure risks for nested categorical data. To save space, the methodology and results for the disclosure risk evaluations are presented in the supplementary material only. To summarize very briefly, the analyses suggest that synthetic data generated from the NDPMPM have low disclosure risks.

The remainder of this article is organized as follows. In Section 2, we present the NDPMPM model when all configurations of groups and units are feasible. In Section 3, we present a data augmentation strategy for estimating a version of the NDPMPM that puts zero probability on impossible combinations. In Section 4, we illustrate and evaluate the NDPMPM models using household demographic data from the American Community Survey (ACS). In particular, we use posterior predictive distributions from the NDPMPM models to generate synthetic datasets, and compare results of representative analyses done with the synthetic and original data. In Section 5, we conclude with discussion of implementation of the proposed models.

2 The NDPMPM Model

As a working example, we suppose the data include N individuals residing in only one of $n < N$ households, where n (but not N) is fixed by design. For $i = 1, \ldots, n$, let $n_i \geq 1$ equal the number of individuals in house i, so that $\sum_{i=1}^{n} n_i = N$. For $k = 1, \ldots, p$, let $X_{ijk} \in \{1, \ldots, d_k\}$ be the value of categorical variable k for person j in household i, where $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$. For $k = p+1, \ldots, p+q$, 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. We let one of the variables in X_{ik} correspond to the household size n_i; thus, N is a random variable. For now, we assume no impossible combinations of variables within individuals or households.

We assume that each household belongs to some group-level latent class, which we label with G_i, where $i = 1, \ldots, n$. Let $\pi_g = \Pr(G_i = g)$ for any class g; that is, π_g is the probability that household i belongs to class g for every household. For any $k \in \{p+1, \ldots, p+q\}$ and any value $c \in \{1, \ldots, d_k\}$, let $\lambda_{gc}^{(k)} = \Pr(X_{ik} = c \mid G_i = g)$ for any class g; here, $\lambda_{gc}^{(k)}$ is the same value for every household in class g. For computational expediency, we truncate the number of group-level latent classes at some sufficiently large value F. Let $\pi = \{\pi_1, \ldots, \pi_F\}$, and let $\lambda = \{\lambda_{gc}^{(k)} : c = 1, \ldots, d_k; k = p+1, \ldots, p+q; g = 1, \ldots, F\}$.

Within each household class, we assume that each individual member belongs to some individual-level latent class, which we label with M_{ij}, where $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$. Let $\omega_{gm} = \Pr(M_{ij} = m \mid G_i = g)$ for any class (g, m); that is, ω_{gm} is the conditional probability that individual j in household i belongs to individual-level class m nested within group-level class g, for every individual. For any $k \in \{1, \ldots, p\}$ and any value $c \in \{1, \ldots, d_k\}$, let $\phi^{(k)}_{gmc} = \Pr(X_{ijk} = c \mid (G_i, M_{ij}) = (g, m))$; here, $\phi^{(k)}_{gmc}$ is the same value for every individual in class (g, m). Again for computational expediency, we truncate the number of individual-level latent classes within each g at some sufficiently large number S that is common across all g. Thus, the truncation results in a total of $F \times S$ latent classes used in computation. Let $\omega = \{\omega_{gm} : g = 1, \ldots, F; m = 1, \ldots, S\}$, and let $\phi = \{\phi^{(k)}_{gmc} : c = 1, \ldots, d_k; k = 1, \ldots, p; g = 1, \ldots, F; m = 1, \ldots, S\}$.

We let both the q household-level variables and p individual-level variables follow independent, class-specific multinomial distributions. Thus, the model for the data and corresponding latent classes in the NDPMPM is

$$
\begin{aligned}
X_{ik} \mid G_i, \lambda &\sim \text{Multinomial}(\lambda^{(k)}_{G_i 1}, \ldots, \lambda^{(k)}_{G_i d_k}) \\
&\quad \text{for all } i,\ k = p+1, \ldots, p+q, && (1) \\
X_{ijk} \mid G_i, M_{ij}, n_i, \phi &\sim \text{Multinomial}(\phi^{(k)}_{G_i M_{ij} 1}, \ldots, \phi^{(k)}_{G_i M_{ij} d_k}) \\
&\quad \text{for all } i, j, k = 1, \ldots, p, && (2) \\
G_i \mid \pi &\sim \text{Multinomial}(\pi_1, \ldots, \pi_F) \ \text{ for all } i, && (3) \\
M_{ij} \mid G_i, n_i, \omega &\sim \text{Multinomial}(\omega_{G_i 1}, \ldots, \omega_{G_i S}) \ \text{ for all } i, j, && (4)
\end{aligned}
$$

where each multinomial distribution has sample size equal to one and number of levels implied by the dimension of the corresponding probability vector. We allow the multinomial probabilities for individual-level classes to differ by household-level class. One could impose additional structure on the probabilities, for example, force them to be equal across classes as suggested in Vermunt (2003, 2008); we do not pursue such generalizations here.

We condition on n_i in (2) and (4) so that the entire model can be interpreted as a generative model for households; that is, the size of the household could be sampled from (1), and once the size is known the characteristics of the household's individuals could be sampled from (2). The distributions in (2) and (4) do not depend on n_i other than to fix the number of people in the household; that is, within any G_i, the distributions of all parameters do not depend on n_i. This encourages borrowing strength across households of different sizes while simplifying computations.

As prior distributions on π and ω, we use the truncated stick breaking representation of the Dirichlet process (Sethuraman, 1994). We have

$$
\begin{aligned}
\pi_g &= u_g \prod_{f<g} (1 - u_f) \ \text{ for } g = 1, \ldots, F, && (5) \\
u_g &\sim \text{Beta}(1, \alpha) \ \text{ for } g = 1, \ldots, F-1, \quad u_F = 1, && (6) \\
\alpha &\sim \text{Gamma}(a_\alpha, b_\alpha), && (7)
\end{aligned}
$$

$$\omega_{gm} = v_{gm} \prod_{s<m} (1 - v_{gs}) \quad \text{for } m = 1, \ldots, S, \tag{8}$$

$$v_{gm} \sim \text{Beta}(1, \beta_g) \quad \text{for } m = 1, \ldots, S-1, \quad v_{gS} = 1, \tag{9}$$

$$\beta_g \sim \text{Gamma}(a_\beta, b_\beta). \tag{10}$$

The prior distribution in (5)–(10) is similar to the truncated version of the nested Dirichlet process prior distribution of Rodriguez et al. (2008) based on conditionally conjugate prior distributions (see Section 5.1 in their article). The prior distribution in (5)–(10) also shares characteristics with the enriched Dirichlet process prior distribution of Wade et al. (2011), in that (i) it gets around the limitations caused by using a single precision parameter α for the mixture probabilities, and (ii) it allows different mixture components for different variables.

As prior distributions on λ and ϕ, we use independent Dirichlet distributions,

$$\lambda_g^{(k)} = (\lambda_{g1}^{(k)}, \ldots, \lambda_{gd_k}^{(k)}) \sim \text{Dir}(a_{k1}, \ldots, a_{kd_k}), \tag{11}$$

$$\phi_{gm}^{(k)} = (\phi_{gm1}^{(k)}, \ldots, \phi_{gmd_k}^{(k)}) \sim \text{Dir}(a_{k1}, \ldots, a_{kd_k}). \tag{12}$$

One can use data-dependent prior distributions for setting each $(a_{k1}, \ldots, a_{k_{d_k}})$, for example, set it equal to the empirical marginal frequency. Alternatively, one can set $a_{k1} = \cdots = a_{kd_k} = 1$ for all k to correspond to uniform distributions. We examined both approaches and found no practical differences between them for our applications; see the supplementary material. In the applications, we present results based on the empirical marginal frequencies. Following Dunson and Xing (2009) and Si and Reiter (2013), we set $(a_\alpha = .25, b_\alpha = .25)$ and $(a_\beta = .25, b_\beta = .25)$, which represents a small prior sample size and hence vague specification for the Gamma distributions. We estimate the posterior distribution of all parameters using a blocked Gibbs sampler (Ishwaran and James, 2001; Si and Reiter, 2013); see the supplement for the relevant full conditionals.

Intuitively, the NDPMPM seeks to cluster households with similar compositions. Within the pool of individuals in any household-level class, the model seeks to cluster individuals with similar characteristics. Because individual-level latent class assignments are conditional on household-level latent class assignments, the model induces dependence among individuals in the same household (more accurately, among individuals in the same household-level cluster). To see this mathematically, consider the expression for the joint distribution for variable k for two individuals j and j' in the same household i. For any $(c, c') \in \{1, \ldots, d_k\}$, we have

$$Pr(X_{ijk} = c, X_{ij'k} = c') = \sum_{g=1}^{F} \left(\sum_{m=1}^{S} \phi_{gmc}^{(k)} \omega_{gm} \sum_{m=1}^{S} \phi_{gmc'}^{(k)} \omega_{gm} \right) \pi_g. \tag{13}$$

Since $Pr(X_{ijk} = c) = \sum_{g=1}^{F} \sum_{m=1}^{S} \phi_{gmc}^{(k)} \omega_{gm} \pi_g$ for any $c \in \{1, \ldots, d_k\}$, the $Pr(X_{ijk} = c, X_{ij'k} = c') \neq Pr(X_{ijk} = c) Pr(X_{ij'k} = c')$.

Ideally we fit enough latent classes to capture key features in the data while keeping computations as expedient as possible. As a strategy for doing so, we have found it

convenient to start an Markov Chain Monte Carlo (MCMC) chain with reasonably-sized values of F and S, say $F = S = 10$. After convergence of the MCMC chain, we check how many latent classes at the household-level and individual-level are occupied across the MCMC iterations. When the numbers of occupied household-level classes hits F, we increase F. When this is not the case but the number of occupied individual-level classes hits S, we try increasing F alone, as the increased number of household-level latent classes may sufficiently capture heterogeneity across households as to make S adequate. When increasing F does not help, for example there are too many different types of individuals, we increase S, possibly in addition to F. We emphasize that these types of titrations are useful primarily to reduce computation time; analysts always can set S and F both to be very large so that they are highly likely to exceed the number of occupied classes in initial runs.

It is computationally convenient to set $\beta_g = \beta$ for all g in (10), as doing so reduces the number of parameters in the model. Allowing β_g to be class-specific offers additional flexibility, as the prior distribution of the household-level class probabilities can vary by class. In our evaluations of the model on the ACS data, results were similar whether we used a common or distinct values of β_g.

3 Adapting the NDPMPM for Impossible Combinations

The models in Section 2 make no restrictions on the compositions of groups or individuals. In many contexts this is unrealistic. Using our working example, suppose that the data include a variable that characterizes relationships among individuals in the household, as the ACS does. Levels of this variable include household head, spouse of household head, parent of the household head, etc. By definition, each household must contain exactly one household head. Additionally, by definition (in the ACS), each household head must be at least 15 years old. Thus, we require a version of the NDPMPM that enforces zero probability for any household that has zero or multiple household heads, and any household headed by someone younger than 15 years.

We need to modify the likelihoods in (1) and (2) to enforce zero probability for impossible combinations. Equivalently, we need to truncate the support of the NDPMPM. To express this mathematically, let $\mathcal{C}_h$ represent all combinations of individuals and households of size h, including impossible combinations; that is, $\mathcal{C}_h$ is the Cartesian product $\Pi_{k=p+1}^{p+q}(1,\ldots,d_k)\left(\Pi_{j=1}^{h}\Pi_{k=1}^{p}(1,\ldots,d_k)\right)$. For any household with h individuals, let $\mathcal{S}_h \subset \mathcal{C}_h$ be the set of combinations that should have zero probability, i.e., $Pr(X_{ip+1},\ldots,X_{ip+q},X_{i11},\ldots,X_{ihp} \in \mathcal{S}_h) = 0$. Let $\mathcal{C} = \bigcup_{h\in\mathcal{H}}\mathcal{C}_h$ and $\mathcal{S} = \bigcup_{h\in\mathcal{H}}\mathcal{S}_h$, where $\mathcal{H}$ is the set of all household sizes in the observed data. We define a random variable for all the data for person j in household i as $\mathbf{X}_{ij}^* = (X_{ij1}^*,\ldots,X_{ijp}^*,X_{ip+1}^*,\ldots,X_{ip+q}^*)$, and a random variable for all data in household i as $\mathbf{X}_i^* = (\mathbf{X}_{i1}^*,\ldots,\mathbf{X}_{in_i}^*)$. Here, we write a superscript $*$ to indicate that the random variables have support only on $\mathcal{C}-\mathcal{S}$; in contrast, we use $\mathbf{X}_{ij}$ and $\mathbf{X}_i$ to indicate the corresponding random variables with unrestricted support on $\mathcal{C}$. Letting $\mathcal{X}^*$ be the sampled data from n households, i.e., a realization of $(\mathbf{X}_1^*,\ldots,\mathbf{X}_n^*)$, the likelihood component of the truncated NDPMPM model, $p(\mathcal{X}^*|\theta)$, can be written as proportional to

$$L(\mathcal{X}^* \mid \theta)$$
$$= \prod_{i=1}^{n} \sum_{h \in \mathcal{H}} \left(\mathbb{1}\{n_i = h\} \mathbb{1}\{\mathbf{X}_i^* \notin \mathcal{S}_h\} \sum_{g=1}^{F} \left(\prod_{k=p+1}^{p+q} \lambda_{gX_{ik}^*}^{(k)} \left(\prod_{j=1}^{h} \sum_{m=1}^{S} \prod_{k=1}^{p} \phi_{gmX_{ijk}^*}^{(k)} \omega_{gm} \right) \right) \pi_g \right), \tag{14}$$

where θ includes all parameters of the model described in Section 2. Here, $\mathbb{1}\{.\}$ equals one when the condition inside the $\{\}$ is true and equals zero otherwise.

For all $h \in \mathcal{H}$, let $n_{*h} = \sum_{i=1}^{n} \mathbb{1}\{n_i = h\}$ be the number of households of size h in $\mathcal{X}^*$. Let $\pi_{0h}(\theta) = Pr(\mathbf{X}_i \in \mathcal{S}_h | \theta)$, where $\mathbf{X}_i$ is the random variable with unrestricted support. The normalizing constant in the likelihood in (14) is $\prod_{h \in \mathcal{H}} (1 - \pi_{0h}(\theta))^{n_{*h}}$. Hence, we seek to compute the posterior distribution

$$p(\theta | \mathcal{X}^*, T(\mathcal{S})) \propto p(\mathcal{X}^* \mid \theta) p(\theta) = \frac{1}{\prod_{h \in \mathcal{H}} (1 - \pi_{0h}(\theta))^{n_{*h}}} L(\mathcal{X}^* \mid \theta) p(\theta). \tag{15}$$

The $T(\mathcal{S})$ emphasizes that the density is for the truncated NDPMPM, not the density from Section 2.

The Gibbs sampling strategy from Section 2 requires conditional independence across individuals and variables, and hence unfortunately is not appropriate as a means to estimate the posterior distribution. Instead, we follow the general approach of Manrique-Vallier and Reiter (2014). The basic idea is to treat the observed data $\mathcal{X}^*$, which we assume includes only feasible households and individuals (e.g., there are no reporting errors that create impossible combinations in the observed data), as a sample from an augmented dataset $\mathcal{X}$ of unknown size. We assume $\mathcal{X}$ arises from an NDPMPM model that does not restrict the characteristics of households or individuals; that is, all combinations of households and individuals are allowable in the augmented sample. With this conceptualization, we can construct a Gibbs sampler that appropriately assigns zero probability to combinations in $\mathcal{S}$ and results in draws of θ from (15). Given a draw of θ, we draw $\mathcal{X}$ using a negative binomial sampling scheme. For each stratum $h \in \mathcal{H}$ defined by unique household sizes in $\mathcal{X}^*$, we repeatedly simulate households with individuals from the untruncated NDPMPM model, stopping when the number of simulated feasible households matches n_{*h}. We make $\mathcal{X}$ comprise $\mathcal{X}^*$ and the generated households that fall in $\mathcal{S}$. Given a draw of $\mathcal{X}$, we draw θ from the NDPMPM model as in Section 2, treating $\mathcal{X}$ as if it were collected data. The full conditionals for this sampler, as well as a proof that it generates draws from (15), are provided in the supplement.

4 Using the NDPMPM to Generate Synthetic Household Data

We now illustrate the ability of the NDPMPM to estimate joint distributions for subsets of household level and individual level variables. Section 4.1 presents results for a scenario where the variables are free of structural zeros (i.e., $\mathcal{S} = \emptyset$), and Section 4.2 presents results for a scenario with impossible combinations.

We use subsets of variables selected from the public use files for the ACS. As brief background, the purpose of the ACS is to enable estimation of population demographics and housing characteristics for the entire United States. The questionnaire is sent to about 1 in 38 households. It includes questions about the individuals living in the household (e.g., their ages, races, incomes) and about the characteristics of the housing unit (e.g., number of bedrooms, presence of running water or not, presence of a telephone line or not). We use only data from non-vacant households.

In both simulation scenarios, we treat data from the public use files as populations, so as to have known population values, and take simple random samples from them on which we estimate the NDPMPM models. We use the estimated posterior predictive distributions to create simulated versions of the data, and compare analyses of the simulated data to the corresponding analyses based on the observed data and the constructed population values.

If we act like the samples from the constructed populations are confidential and cannot be shared as is, the simulated datasets can be viewed as redacted public use file, i.e., synthetic data. We generate L synthetic datasets, $\mathbf{Z} = (\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(L)})$, by sampling L datasets from the posterior predictive distribution of a NDPMPM model. We generate synthetic data so that the number of households of any size h in each $\mathbf{Z}^{(l)}$ exactly matches n_{*h}. This improves the quality of the synthetic data by ensuring that the total number of individuals and household size distributions match in $\mathbf{Z}$ and $\mathcal{X}^*$. As a result, $\mathbf{Z}$ comprises partially synthetic data (Little, 1993; Reiter, 2003), even though every released Z_{ijk} is a simulated value.

To make inferences with $\mathbf{Z}$ we use the approach in Reiter (2003). Suppose that we seek to estimate some scalar quantity Q. For $l = 1, \ldots, L$, let $q^{(l)}$ and $u^{(l)}$ be respectively the point estimate of Q and its associated variance estimate computed with $\mathbf{Z}^{(l)}$. Let $\bar{q}_L = \sum_l q^{(l)}/L$; $\bar{u}_L = \sum_l u^{(l)}/L$; $b_L = \sum_l (q^{(l)} - \bar{q}_L)^2/(L-1)$; and $T_L = \bar{u}_L + b_L/L$. We make inferences about Q using the $t-$distribution, $(\bar{q}_L - Q) \sim t_v(0, T_L)$, with $v = (L-1)(1 + L\bar{u}_L/b_L)^2$ degrees of freedom.

4.1 Illustration without Structural Zeros

For this scenario, we use data from the 2012 ACS public use file (Ruggles et al., 2010) to construct a population with 308769 households. From this we take a simple random sample of $n = 10000$ households. We use the four household-level variables and ten individual-level variables summarized in Table 1. We select these variables purposefully to avoid structural zeros. Household sizes range from one to nine, with $(n_{*1}, \ldots, n_{*9}) = (2528, 5421, 1375, 478, 123, 52, 16, 5, 2)$. This sample of n households includes $N = 20504$ individuals. We treat income and age as unordered categorical variables; we discuss adapting the model for ordered categorical variables in Section 5.

We run the MCMC sampler for the NDPMPM model of Section 2 for 10000 iterations, treating the first 5000 iterations as burn-in. We set $(F, S) = (30, 10)$ and use a common β. The posterior mean of the number of occupied household-level classes is 27 and ranges from 25 to 29. Within household-level classes, the posterior number of occupied individual-level classes ranges from 5 to 8. To monitor convergence of the

Description	Categories
Ownership of dwelling	1 = owned or being bought, 2 = rented
House acreage	1 = house on less than 10 acres,
	2 = house on 10 acres or more
Household income	1 = less than 25K, 2 = between 25K and 45K,
	3 = between 45K and 75K,
	4 = between 75K and 100K, 5 = more than 100K
Household size	1 = 1 person, 2 = 2 people, etc.
Age	1 = 18, 2 = 19, ..., 78 = 95
Gender	1 = male, 2 = female
Recoded general race code	1 = white alone, 2 = black alone,
	3 = American Indian/Alaska Native alone,
	4 = Asian or Pacific Islander alone,
	5 = other, 6 = two or more races
Speaks English	1 = does not speak English, 2 = speaks English
Hispanic origin	1 = not Hispanic, 2 = Hispanic
Health insurance coverage	1 = no, 2 = yes
Educational attainment	1 = less than high school diploma,
	2 = high school diploma/GED/alternative credential,
	3 = some college, 4 = bachelor's degree,
	5 = beyond bachelor's degree
Employment status	1 = employed, 2 = unemployed, 3 = not in labor force
Migration status, 1 year	1 = in the same house, 2 = moved within state,
	3 = moved between states, 4 = abroad one year ago
Marital status	1 = married spouse present,
	2 = married spouse absent, 3 = separated,
	4 = divorced, 5 = widowed,
	6 = never married/single

Table 1: Subset of variables in the empirical illustration without structural zeros. The first four variables are household-level variables, and the last ten variables are individual-level variables.

MCMC sampler, we focus of π, α, and β. As a check on the choice of (F, S), we also estimated the model with $(F, S) = (50, 50)$. We found similar results for both the number of occupied classes and the posterior predictive distributions; see the supplement for details.

We generate $\mathbf{Z}^{(l)}$ by sampling a draw of $(\mathbf{G}, \mathbf{M}, \lambda, \phi)$ from the posterior distribution. For each household $i = 1, \ldots, n$, we generate its synthetic household-level attributes, $(X^{(l)}_{ip+1}, \ldots, X^{(l)}_{ip+q})$, from (1) using G_i and the corresponding probabilities in λ. For each individual $j = 1, \ldots, n_i$ in each household, we generate the synthetic individual-level attributes, $(X^{(l)}_{ij1}, \ldots, X^{(l)}_{ijp})$, from (2) using M_{ij} and the corresponding probabilities in ϕ. We repeat this process $L = 5$ times, using approximately independent draws of parameters obtained from iterations that are far apart in the MCMC chain.

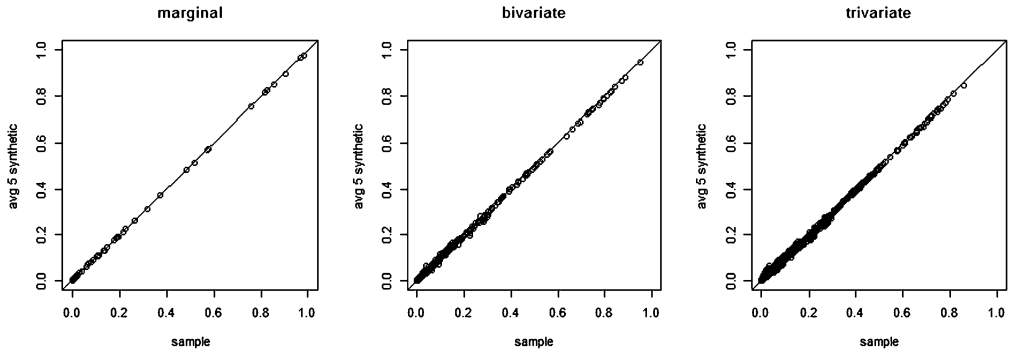

Figure 1: Marginal, bivariate and trivariate probabilities computed in the sample and synthetic datasets for the illustration without structural zeros. Restricted to categories with expected counts equal to at least 10. Point estimates from both sets of data are similar, suggesting that the NDPMPM fits the data well.

To evaluate the quality of the NDPMPM model, we compare the relationships among the variables in the original and synthetic datasets to each other, as is typical in synthetic data evaluations, as well as to the corresponding population values. We consider the marginal distributions of all variables, bivariate distributions of all possible pairs of variables, and trivariate distributions of all possible triplets of variables. We restrict the plot to categories where the expected count in samples of 10000 households is at least 10. Plots in Figure 1 display each $\bar{q}_5$ plotted against its corresponding empirical probability in the original data for all parameters. As evident in the figures, the synthetic point estimates are close to those from the original data, suggesting that the NDPMPM accurately estimates the relationships among the variables. Both sets of point estimates are close to the corresponding probabilities in the population, as we show in the supplement.

We also examine several probabilities that depend on values for individuals in the same household, that is, they are affected by within-household relationships. As evident in Table 2, and not surprisingly given the sample size, the point estimates from the original sampled data are close to the values in the constructed population. For most quantities the synthetic data point and interval estimates are similar to those based on the original sample, suggesting that the NDPMPM model has captured the complicated within household structure reasonably well. One exception is the percentage of households with everyone of the same race: the NDPMPM underestimates these percentages. Accuracy worsens as household size increases. This is partly explained by sample sizes, as $n_{*3} = 1375$ and $n_{*4} = 478$, compared to $n_{*2} = 5421$. We also ran a simulation with $n = 50000$ households comprising $N = 101888$ individuals sampled randomly from the same constructed population, in which $(n_{*1}, \ldots, n_{*10}) = (12804, 27309, 6515, 2414, 630, 229, 63, 26, 8, 2)$. For households with $n_i = 3$, the 95% intervals from the synthetic and original data are, respectively, (.870, .887) and (.901, .906); for households of size $n_i = 4$, the 95% intervals from the synthetic and original data are, respectively, (.826, .858) and (.889, .895). Results for the remaining probabilities in Table 2 are also improved.

	Q	Original	NDPMPM	DPMPM
All same race				
$n_i = 2$	.928	(.923, .933)	(.847, .868)	(.648, .676)
$n_i = 3$	.906	(.889, .901)	(.803, .845)	(.349, .407)
$n_i = 4$	.885	(.896, .908)	(.730, .817)	(.183, .277)
All white, rent	.123	(.115, .128)	(.110, .126)	(.052, .062)
All white w/ health insur.	.632	(.622, .641)	(.582, .603)	(.502, .523)
All married, working	.185	(.177, .192)	(.171, .188)	(.153, .168)
All w/ college degree	.091	(.086, .097)	(.071, .082)	(.067, .077)
All w/ health coverage	.807	(.800, .815)	(.764, .782)	(.760, .777)
All speak English	.974	(.969, .976)	(.959, .967)	(.963, .970)
Two workers in home	.291	(.282, .300)	(.289, .309)	(.287, .308)

Table 2: 95% confidence intervals in the original and synthetic data for selected probabilities that depend on within household relationships. Results for illustration without structural zeros. Intervals for probability that all family members are the same race are presented only for households of size two, three, and four because of inadequate sample sizes for $n_i > 4$. The quantity Q is the value in the constructed population of 308769 households.

As a comparison, we also generated synthetic datasets using a non-nested DPMPM model (Dunson and Xing, 2009) that ignores the household clustering. Not surprisingly, the DPMPM results in substantially less accuracy for many of the probabilities in Table 2. For example, for the percentage of households of size $n_i = 4$ in which all members have the same race, the DPMPM results in a 95% confidence interval of (.183, .277), which is quite unlike the (.896, .908) interval in the original data and far from the population value of .885. The DPMPM also struggles for other quantities involving racial compositions. Unlike the NDPMPM model, the DPMPM model treats each observation as independent, thereby ignoring the dependency among individuals in the same household. We note that we obtain similar results with nine other independent samples of 10000 households, indicating that the differences between the NDPMPM and DPMPM results in Table 2 are not reflective of chance error.

4.2 Illustration with Structural Zeros

For this scenario, we use data from the 2011 ACS public use file (Ruggles et al., 2010) to construct the population. We select variables to mimic those on the U. S. decennial census, per the motivation described in Section 1. These include a variable that explicitly indicates relationships among individuals within the same household. This variable creates numerous and complex patterns of impossible combinations. For example, each household can have only one head who must be at least 16 years old, and biological children/grandchildren must be younger than their parents/grandparents. We use the two household-level variables and five individual-level variables summarized in Table 3, which match those on the decennial census questionnaire. We exclude households with only one individual because these individuals by definition must be classified as household heads, so that we have no need to model the family relationship variable. To

Description	Categories
Ownership of dwelling	1 = owned or being bought (loan), 2 = rented
Household size	2 = 2 people, 3 = 3 people, 4 = 4 people
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/more major races
Hispanic origin (recoded)	1 = not Hispanic, 2 = Mexican, 3 = Puerto Rican, 4 = Cuban, 5 = other
Age (recoded)	1 = 0 (less then one year old), 2 = 1, . . . , 94 = 93
Relationship to the household head	1 = head/householder, 2 = spouse, 3 = child, 4 = child-in-law, 5 = parent, 6 = parent-in-law, 7 = sibling, 8 = sibling-in-law, 9 = grandchild, 10 = other relatives, 11 = partner, friend, visitor, 12 = other non-relatives

Table 3: Subset of variables used in the illustration with structural zeros. The first two variables are household-level variables, and the last five variables are individual-level variables.

generate synthetic data for households of size $n_i = 1$, one could use non-nested versions of latent class models (Dunson and Xing, 2009; Manrique-Vallier and Reiter, 2014). We also exclude households with $n_i > 4$ for presentational and computational convenience.

The constructed population comprises 127685 households, from which we take a simple random sample of $n = 10000$ households. Household sizes are $(n_2, n_3, n_4) = (5370, 2504, 2126)$. The 10000 households comprise $N = 26756$ individuals.

We fit the truncated NDPMPM model of Section 3, using all the variables in Table 3 as X_{ijk} or X_{ik} in the model. We run the MCMC sampler for 10000 iterations, treating the first 6000 iterations as burn-in. We set $(F, S) = (40, 15)$ and use a common β. The posterior mean of the number of household-level classes occupied by households in $\mathcal{X}^*$ is 28 and ranges from 23 to 36. Within household-level classes, the posterior number of individual-level classes occupied by individuals in $\mathcal{X}^*$ ranges from 5 to 10. To check for convergence of the MCMC chain, we look at trace plots of π, α, β, and n_0. The plots for (π, α, β) suggest good mixing; however, the plot for n_0 exhibits non-trivial autocorrelations. Values of n_0 are around 8.0×10^5 near the 6000th and 10000th iterations of the chain, with a minimum around 7.2×10^5 near the 6500th iteration and a maximum around 9.3×10^5 near the 9400th iteration. As a byproduct of the MCMC sampler, at each MCMC iteration we create n households that satisfy all constraints. We use these households to form each $\mathbf{Z}^{(l)}$, where $l = 1, \ldots, 5$, selecting from five randomly sampled, sufficiently separated iterations.

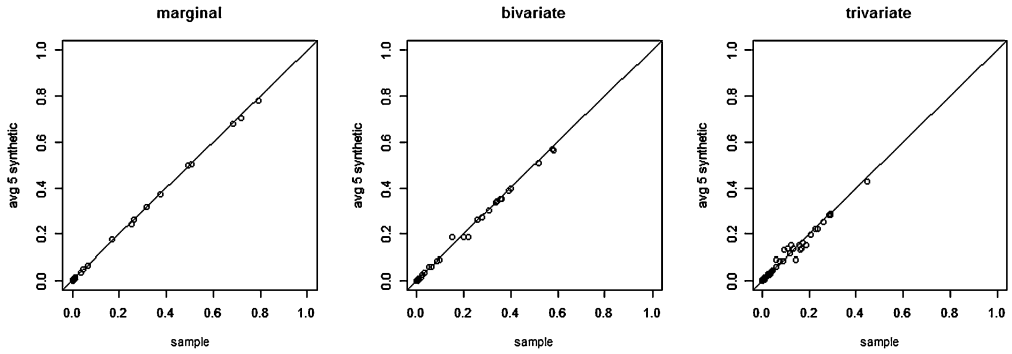

Figure 2: Marginal, bivariate and trivariate distributions probabilities computed in the sample and synthetic datasets in illustration with structural zeros. Restricted to categories with expected counts equal to at least 10. Point estimates from both sets of data are similar, suggesting that the truncated NDPMPM fits the data reasonably well.

As in Section 4.1, we evaluate the marginal distributions of all variables, bivariate distributions of all possible pairs of variables, and trivariate distributions of all possible triplets of variables, restricting to categories where the expected counts are at least 10. Plots in Figure 2 display each $\bar{q}_5$ plotted against its corresponding estimate from the original data, the latter of which are close to the population values (see the supplementary material). The point estimates are quite similar, indicating that the NDPMPM captures relationships among the variables.

Table 4 compares original and synthetic 95% confidence intervals for selected probabilities involving within-household relationships. We choose a wide range of household types involving multiple household level and individual level variables. We include quantities that depend explicitly on the "relationship to household head" variable, as these should be particularly informative about how well the truncated NDPMPM model estimates probabilities directly impacted by structural zeros. As evident in Table 4, estimates from the original sample data are generally close to the corresponding population values. Most intervals from the synthetic data are similar to those from the original data, indicating that the truncated NDPMPM model captures within-household dependence structures reasonably well. As in the simulation with no structural zeros, the truncated NDPMPM model has more difficulty capturing dependencies for the larger households, due to smaller sample sizes and more complicated within-household relationships.

For comparison, we also generate synthetic data using the NDPMPM model from Section 2, which does not account for the structural zeros. In the column labeled "NDPMPM untruncate", we use the NDPMPM model and completely ignore structural zeros, allowing the synthetic data to include households with impossible combinations. In the column labeled "NDPMPM rej samp", we ignore structural zeros when estimating model parameters but use rejection sampling at the data synthesis stage to ensure that no simulated households include physically impossible combinations. As seen in Table 4, the interval estimates from the truncated NDPMPM generally are more accurate than

	Q	Original	NDPMPM truncate	NDPMPM untruncate	NDPMPM rej samp
All same race					
$n_i = 2$	.906	(.900, .911)	(.858, .877)	(.824, .845)	(.811, .840)
$n_i = 3$	.869	(.871, .884)	(.776, .811)	(.701, .744)	(.682, .723)
$n_i = 4$	.866	(.863, .876)	(.756, .800)	(.622, .667)	(.614, .667)
Spouse present	.667	(.668, .686)	(.630, .658)	(.438, .459)	(.398, .422)
Spouse w/ white HH	.520	(.520, .540)	(.484, .510)	(.339, .359)	(.330, .356)
Spouse w/ black HH	.029	(.024, .031)	(.022, .029)	(.023, .030)	(.018, .025)
White cpl	.489	(.489, .509)	(.458, .483)	(.261, .279)	(.306, .333)
White cpl, own	.404	(.401, .421)	(.370, .392)	(.209, .228)	(.240, .266)
Same race cpl	.604	(.603, .622)	(.556, .582)	(.290, .309)	(.337, .361)
White-nonwhite cpl	.053	(.049, .057)	(.048, .058)	(.031, .039)	(.039, .048)
Nonwhite cpl, own	.085	(.079, .090)	(.068, .079)	(.025, .033)	(.024, .031)
Only mother	.143	(.128, .142)	(.103, .119)	(.113, .126)	(.201, .219)
Only one parent	.186	(.172, .187)	(.208, .228)	(.230, .247)	(.412, .435)
Children present	.481	(.473, .492)	(.471, .492)	(.472, .492)	(.566, .587)
Parents present	.033	(.029, .036)	(.038, .046)	(.035, .043)	(.011, .016)
Siblings present	.029	(.022, .028)	(.032, .041)	(.027, .034)	(.029, .039)
Grandchild present	.035	(.028, .035)	(.032, .041)	(.035, .043)	(.024, .031)
Three generations present	.043	(.036, .043)	(.042, .051)	(.051, .060)	(.028, .035)

Table 4: 95% confidence intervals in the original and synthetic data for selected probabilities that depend on within household relationships. Results for illustration with structural zeros. "NDPMPM truncate" uses the model from Section 3. "NDPMPM untruncate" uses the model from Section 2. "NDPMPM rej samp" uses the model from Section 2 but rejecting any proposed synthetic observation that fails to respect the structural zeros. "HH" means household head, and "cpl" means couple. The quantity Q is the value in the full constructed population of 127685 households.

those based on the other two approaches. When structural zeros most directly impact the probability, i.e., when the "relationship to household head" variable is involved, the performances of "NDPMPM untruncate" and "NDPMPM rej samp" are substantially degraded.

5 Discussion

The MCMC sampler for the NDPMPM in Section 2 is computationally expedient. However, the MCMC sampler for the truncated NDPMPM in Section 3 is computationally intensive. The primary bottlenecks in the computation arise from simulation of $\mathcal{X}$. When the probability mass in the region defined by $\mathcal{S}$ is large compared to the probability mass in the region defined by $\mathcal{C} - \mathcal{S}$, the MCMC can sample many households with impossible combinations before getting n feasible ones. Additionally, it can be time consuming to check whether or not a generated record satisfies all constraints in $\mathcal{S}$. These bottle-

necks can be especially troublesome when n_i is large for many households. To reduce running times, one can parallelize many steps in the sampler (which we did not do). As examples, the generation of augmented records and the checking of constraints can be spread over many processors. One also can reduce computation time by putting an upper bounds on the size of $\mathcal{X}$ (that is still much larger than n). Although this results in an approximation to the Gibbs sampler, this still could yield reasonable inferences or synthetic datasets, particularly when many records in $\mathcal{X}$ end up in clusters with few data points from $\mathcal{X}^*$.

Conceptually, the methodology can be readily extended to handle other types of variables. For example, one could replace the multinomial kernels with continuous kernels (e.g., Gaussian distributions) to handle numerical variables. For ordered categorical variables, one could use a probit specification Albert and Chib (1993) or the rank likelihood (Hoff, 2009, Ch. 12). For mixed data, one could use the Bayesian joint model for multivariate continuous and categorical variables developed in Murray and Reiter (forthcoming). Evaluating the properties of such models is a topic for future research.

We did not take advantage of prior information when estimating the models. Such information might be known, for example, from other data sources. Incorporating prior information in latent class models is tricky, because we need to do so in a way that does not distort conditional distributions. Schifeling and Reiter (2016) presented a simple approach to doing so for non-nested latent class models, in which the analyst appends to the original data partially complete, pseudo-observations with empirical frequencies that match the desired prior distribution. If one had prior information on household size jointly with some other variable, say individuals' races, one could follow the approach of Schifeling and Reiter (2016) and augment the collected data with partially complete households. When the prior information does not include household size, e.g., just a marginal distribution of race, it is not obvious how to incorporate the prior information in a principled way.

Like most joint models, the NDPMPM generally is not appropriate for estimating multivariate distributions with data from complex sampling designs. This is because the model reflects the distributions in the observed data, which might be collected by differentially sampling certain subpopulations. When design variables are categorical and are available for the entire population (not just the sample), analysts can use the NDPMPM as an engine for Bayesian finite population inference (Gelman et al., 2013, Ch. 8). In this case, the analyst includes the design variables in the NDPMPM, uses the implied, estimated conditional distribution to impute many copies of the non-sampled records' unknown survey values given the design variables, and computes quantities of interest on each completed population. These completed-population quantities summarize the posterior distribution. Absent this information, there is no consensus on the "best" way to incorporate survey weights in Bayesian joint mixture models. Kunihama et al. (2014) present a computationally convenient approach that uses only the survey weights for sampled cases. A similar approach could be applied for nested categorical data. Evaluating this approach, as well as other adaptations of ideas proposed in the literature, is a worthy topic for future research.

The truncated NDPMPM also assumes the observed data do not include errors that create theoretically impossible combinations of values. When such faulty values

are present, analysts should edit and impute corrected values, for example, using the Fellegi and Holt (1976) paradigm popular with statistical agencies. Alternatively, one could add a stochastic measurement error model to the truncated NDPMPM, as done by Kim et al. (2015) for continuous data and Manrique-Vallier and Reiter (forthcoming) for non-nested categorical data. While conceptually feasible, this is not a trivial extension. The NDPMPM is already computationally intensive; searching over the huge space of possible error localizations could increase the computational burden substantially. This suggests one would need alternatives to standard MCMC algorithms for model fitting.

Supplementary Material

Supplementary Materials for "Dirichlet Process Mixture Models for Modeling and Generating Synthetic Versions of Nested Categorical Data" (DOI: 10.1214/16-BA1047SUPP; .pdf).

References

Abowd, J., Stinson, M., and Benedetto, G. (2006). "Final Report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project." Technical report, U.S. Census Bureau Longitudinal Employer-Household Dynamics Program. Available at http://www.census.gov/sipp/synth_data.html. 184

Albert, J. H. and Chib, S. (1993). "Bayesian analysis of binary and polychotomous response data." *Journal of the American Statistical Association*, 88: 669–679. MR1224394. 197

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*, 10(2): 139–154. MR3505053. doi: https://doi.org/10.1007/s11634-016-0234-1. 184

Dunson, D. B. and Xing, C. (2009). "Nonparametric Bayes modeling of multivariate categorical data." *Journal of the American Statistical Association*, 104: 1042–1051. 183, 187, 193, 194

Fellegi, I. P. and Holt, D. (1976). "A systematic approach to automatic edit and imputation." *Journal of the American Statistical Association*, 71: 17–35. MR0371177. 198

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. London: Chapman & Hall. MR3235677. 197

Goodman, L. A. (1974). "Exploratory latent structure analysis using both identifiable and unidentifiable models." *Biometrika*, 61: 215–231. 183

Hawala, S. (2008). "Producing partially synthetic data to avoid disclosure." In *Proceedings of the Joint Statistical Meetings*. Alexandria, VA: American Statistical Association. 185

Hoff, P. D. (2009). *A First Course in Bayesian Statistical Methods*. New York: Springer. 197

Hu, J., Reiter, J. P., and Wang, Q. (2014). "Disclosure risk evaluation for fully synthetic categorical data." In Domingo-Ferrer, J. (ed.), *Privacy in Statistical Databases*, 185–199. Springer. 185

Hu, J., Reiter, J. P., and Wang, Q. (2017). "Supplementary Materials for "Dirichlet Process Mixture Models for Modeling and Generating Synthetic Versions of Nested Categorical Data"." *Bayesian Analysis*. doi: https://doi.org/10.1214/16-BA1047SUPP. 183

Ishwaran, H. and James, L. F. (2001). "Gibbs sampling methods for stick-breaking priors." *Journal of the American Statistical Association*, 161–173. MR1952729. doi: https://doi.org/10.1198/016214501750332758. 183, 187

Jain, S. and Neal, R. M. (2007). "Splitting and merging components of a nonconjugate Dirichlet process mixture model." *Bayesian Analysis*, 2: 445–472. 183

Kim, H. J., Cox, L. H., Karr, A. F., Reiter, J. P., and Wang, Q. (2015). "Simultaneous editing and imputation for continuous data." *Journal of the American Statistical Association*, 110: 987–999. MR3420678. doi: https://doi.org/10.1080/01621459.2015.1040881. 198

Kinney, S., Reiter, J. P., Reznek, A. P., Miranda, J., Jarmin, R. S., and Abowd, J. M. (2011). "Towards unrestricted public use business microdata: The synthetic Longitudinal Business Database." *International Statistical Review*, 79: 363–384. 185

Kunihama, T., Herring, A. H., Halpern, C. T., and Dunson, D. B. (2014). "Nonparametric Bayes modeling with sample survey weights." *arXiv:1409.5914*. 197

Little, R. J. A. (1993). "Statistical analysis of masked data." *Journal of Official Statistics*, 9: 407–426. 184, 190

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., and Vilhuber, L. (2008). "Privacy: Theory meets practice on the map." In *IEEE 24th International Conference on Data Engineering*, 277–286. 185

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. 189, 194

Manrique-Vallier, D. and Reiter, J. P. (forthcoming). "Bayesian simultaneous edit and imputation for multivariate categorical data." *Journal of the American Statistical Association*, to appear. doi: https://doi.org/10.1080/01621459.2016.1231612. 198

Murray, J. S. and Reiter, J. P. (forthcoming). "Multiple imputation of missing categorical and continuous values via Bayesian mixture models with local dependence." *Journal of the American Statistical Association*, to appear. doi: https://doi.org/10.1080/01621459.2016.1174132. 197

Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003). "Multiple imputation for statistical disclosure limitation." *Journal of Official Statistics*, 19: 1–16. 184

Reiter, J. and Raghunathan, T. E. (2007). "The multiple adaptations of multiple imputation." *Journal of the American Statistical Association*, 102: 1462–1471. MR2372542. doi: https://doi.org/10.1198/016214507000000932. 184

Reiter, J. P. (2003). "Inference for partially synthetic, public use microdata sets." *Survey Methodology*, 29: 181–189. 190

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. 184

Rodriguez, A., Dunson, D. B., and Gelfand, A. E. (2008). "The nested Dirichelt process." *Journal of the American Statistical Association*, 103: 1131–1154. MR2528831. doi: https://doi.org/10.1198/016214508000000553. 187

Rubin, D. B. (1993). "Discussion: Statistical disclosure limitation." *Journal of Official Statistics*, 9: 462–468. 184

Ruggles, S., Alexander, J. T., Genadek, K., Goeken, R., Schroeder, M. B., and Sobek, M. (2010). "Integrated Public Use Microdata Series: Version 5.0 [Machine-readable database]." *Minneapolis: University of Minnesota*. 190, 193

Schifeling, T. and Reiter, J. P. (2016). "Incorporating marginal prior information in latent class models." *Bayesian Analysis*, 2: 499–518. 197

Sethuraman, J. (1994). "A constructive definition of Dirichlet priors." *Statistica Sinica*, 4: 639–650. 186

Si, Y. and Reiter, J. P. (2013). "Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys." *Journal of Educational and Behavioral Statistics*, 38: 499–521. 187

Vermunt, J. K. (2003). "Multilevel latent class models." *Sociological Methodology*, 213–239. 184, 186

Vermunt, J. K. (2008). "Latent class and finite mixture models for multilevel data sets." *Statistical Methods in Medical Research*, 33–51. 184, 186

Wade, S., Mongelluzzo, S., and Petrone, S. (2011). "An enriched conjugate prior for Bayesian nonparametric inference." *Bayesian Analysis*, 6: 359–385. 187