arXiv:1412.2282v4 [stat.ME] 2 Nov 2015

Dirichlet Process Mixture Models for Nested Categorical Data

Jingchen Hu[1,*,??], Jerome P. Reiter[2,*,†,??,??] and Quanli Wang[2,†,??]

[1]*Department of Mathematics and Statistics, Vassar College, Box 27, Poughkeepsie, NY 12604 e-mail:* jihu@vassar.edu

[2]*Department of Statistical Science, Duke University, Durham, NC 27708-0251 e-mail:* jerry@stat.duke.edu

[3]*Department of Statistical Science, Duke University, Durham, NC 27708-0251 e-mail:* quanli@stat.duke.edu

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 with high analytic validity and low disclosure risks. Supplementary materials for this article are available online.

Keywords and phrases: 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 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 data for units in the same group as independent and identically distributed; 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 Dirichlet process mixture of products of multinomial distributions (NDPMPM), as we use a finite version of the nested Dirichlet process (Rodriguez et al., 2008) as the prior distribution for the class membership probabilities. 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. To the best of our knowledge, the models of Vermunt (2003, 2008) do not include group-level random variables—they allow group-level predictors of latent class assignments, but not group-level outcomes—nor do these models account for groups with physically impossible combinations of units.

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 de-

mographic data from the American Community Survey (ACS). In particular, we use posterior predictive distributions from the NDPMPM models to generate synthetic datasets (Rubin, 1993; Little, 1993), 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. 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. 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 $\phi_{gmc}^{(k)} = \Pr(X_{ijk} = c \mid (G_i, M_{ij}) = (g, m))$; here, $\phi_{gmc}^{(k)}$ 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 $\phi = \{\phi_{gmc}^{(k)} : c = 1, \ldots, d_k; k = 1, \ldots, p; g = 1, \ldots, F; m = 1, \ldots, S\}$.

We let both the q household-level variables and p individual-level variables follow independent, class-specific multinomial distributions. The model for the

data and corresponding latent classes in the NDPMPM is

$$
\begin{aligned}
X_{ik} \mid G_i, \lambda &\overset{ind}{\sim} \text{Multinomial}(\lambda_{G_i1}^{(k)}, \ldots, \lambda_{G_id_k}^{(k)}) \\
&\qquad \text{for all } i, k = p+1, \ldots, p+q \tag{1} \\
X_{ijk} \mid G_i, M_{ij}, n_i, \phi &\overset{ind}{\sim} \text{Multinomial}(\phi_{G_iM_{ij}1}^{(k)}, \ldots, \phi_{G_iM_{ij}d_k}^{(k)}) \\
&\qquad \text{for all } i, j, k = 1, \ldots, p \tag{2} \\
G_i \mid \pi &\sim \text{Multinomial}(\pi_1, \ldots, \pi_F) \ \text{ for all } i, \tag{3} \\
M_{ij} \mid G_i, n_i, \omega &\sim \text{Multinomial}(\omega_{G_i1}, \ldots, \omega_{G_iS}) \ \text{ for all } i, j, \tag{4}
\end{aligned}
$$

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 let one of the variables in X_{ik} correspond to the household size n_i, so that (1) — (4) can be interpreted as a generative model for households. In (4) the parameters do not depend on n_i to encourage borrowing strength across households of different sizes while simplifying computations.

As prior distributions on π and ϕ, we use the truncated stick breaking representation of the Dirichlet process (Sethuraman, 1994), following the approach described in Rodriguez et al. (2008). We have

$$
\pi_g = u_g \prod_{f<g}(1 - u_f) \ \text{ for } g = 1, \ldots, F \tag{5}
$$

$$
u_g \overset{iid}{\sim} \text{Beta}(1, \alpha) \ \text{ for } g = 1, \ldots, F-1, \quad u_F = 1 \tag{6}
$$

$$
\alpha \sim \text{Gamma}(a_\alpha, b_\alpha) \tag{7}
$$

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

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

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

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

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

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

We set $a_{k1} = \cdots = a_{kd_k} = 1$ for all k to correspond to uniform distributions. 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 Appendix A 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

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

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 $Pr(X_{ijk} = c, X_{ij'k} = c') \neq Pr(X_{ijk} = c) Pr(X_{ij'k} = c')$. We also note that, as argued in Rodriguez et al. (2008), the nested DP prior distribution *a priori* induces stronger associations among individuals in the same group than among individuals in different groups.

For values of F and S, we recommend that analysts start with modest values, say $F = 10$ and $S = 10$, to favor expedient computations. After convergence of the MCMC chain, analysts should 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, analysts should increase F. When this is not the case but the number of occupied individual-level classes hits S, analysts can 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, analysts can increase S, possibly in addition to F. We emphasize that these types of titrations are useful primarily to reduce computation time; analysts always can increase S and F simultaneously so that they exceed the number of occupied classes in initial runs.

It can be computationally convenient to set $\beta_g = \beta$ for all g in (10), as doing so reduces the number of parameters in the model. Allowing the 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 $\mathcal{C}$ represent all combinations of individuals and households, ignoring any impossible combinations. For any household with h individuals, let $\mathcal{S}_h \in \mathcal{C}$ be the set of combinations that should have zero probability, i.e., the set of impossible combinations for households of that size. Let $\mathcal{S} = \bigcup_{h \in \mathcal{H}} \mathcal{S}_h$, 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 $\mathcal{X}^* = (\mathbf{X}_1^*, \ldots, \mathbf{X}_n^*)$ be a sample comprising data from n households, the likelihood component of the truncated NDPMPM model can be written as $p(\mathcal{X}^*|\theta)$ proportional to

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

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 $\mathcal{X}^*$. From (14), we seek to compute the posterior distribution

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

Here, $\pi_{0h}(\theta) = Pr(\mathbf{X}_i \in \mathcal{S}_h|\theta)$, where again $\mathbf{X}_i$ is the random variable with unrestricted support.

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 (15). 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 $\mathcal{X} = (\mathcal{X}^*, \mathcal{X}^0)$ of unknown size. We assume $\mathcal{X}$ arises from a NDPMPM model with density function f 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 $\mathcal{X}$, we draw θ from the NDPMPM model as in Section 2, treating $\mathcal{X}$ as if it were collected data. Given a draw of θ, we draw $\mathcal{X}^0$ 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 $\mathcal{X}^0$ the generated households that fall in $\mathcal{S}$.

We now state a result, proved in Appendix B, that ensures draws of θ from this rejection scheme correspond to draws from (15). First, we note that each record in $\mathcal{X}$ is associated with a household-level and individual-level latent class assignment. Let $\mathbf{G}^* = (G_1^*, \ldots, G_n^*)$ and $\mathbf{M}^* = \{(M_{i1}^*, \ldots, M_{in_i}^*), i = 1, \ldots, n\}$ include all the latent class assignments corresponding to households and individuals in $\mathcal{X}^*$. Similarly, let $\mathbf{G}^0 = (G_1^0, \ldots, G_{n_0}^0)$ and $\mathbf{M}^0 = \{(M_{i1}^0, \ldots, M_{in_i}^0), i = 1, \ldots, n_0\}$ include all the latent class assignments corresponding to the n_0 cases in $\mathcal{X}^0$. Let n_{0h} be a random variable corresponding to the number of households of size h generated in $\mathcal{X}^0$. If we set $p(n_{0h} \mid n_{*h}) \propto 1/(n_{*h} + n_{0h})$ for all $h \in \mathcal{H}$, we can show that

$$p(\theta|\mathcal{X}^*) = \int f(\theta, \mathbf{G}^*, \mathbf{G}^0, \mathbf{M}^*, \mathbf{M}^0, \mathcal{X}^0, \{n_{0h} : h \in \mathcal{H}\} \mid \mathcal{X}^*) d\mathcal{X}^0 dG^* dG^0 dM^* dM^0 d\{n_{0h}\}.$$

(16)

Thus, we can obtain samples from the posterior distribution $p(\theta|\mathcal{X}^*)$ in the truncated NDPMPM model from the sampler for $f(\theta, \mathbf{G}^*, \mathbf{G}^0, \mathbf{M}^*, \mathbf{M}^0, \mathcal{X}^0, \{n_{0h}\} \mid \mathcal{X}^*)$ under the unrestricted NDPMPM model. The full conditionals for this Gibbs sampler are summarized in Appendix A.

4. Using the NDPMPM to Generate Synthetic Household Data

We now apply the NDPMPM to estimate joint distributions for subsets of household level and individual level variables in the American Community Survey (ACS). 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. To evaluate the performance of the models in these two scenarios, 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.

If we act like the observed ACS data are confidential and cannot be shared as is, the simulated datasets can be viewed as redacted public use files. 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 (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 Lon-

gitudinal 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. Indeed, in developing the truncated NDPMPM, one of our main motivations is to develop a method that the Census Bureau could use to generate a synthetic, public use file for the decennial census.

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 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 $Z^{(l)}$. 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.

In addition to assessing the quality of synthetic data generated from the NDPMPM, we also evaluate measures of disclosure risk. Specifically, we quantify the probabilities that intruders can learn the values of original data records given the released synthetic data. To save space, we present the disclosure analysis in the online supplement. To summarize very briefly, the analyses suggest that synthetic data generated from the NDPMPM have low disclosure risks.

4.1. Illustration without structural zeros

We use data comprising $n = 10000$ households and $N = 20504$ individuals randomly sampled from the 2012 ACS public use file (Ruggles et al., 2010). For illustration, we disregard the stratification by census tract and act as if the data are a simple random sample. We discuss approaches to incorporating the stratification in Section 5. We use the four household-level variables and ten individual-level variables summarized in Table 1. Household sizes range from one to nine, with $(n_{*1}, \ldots, n_{*9}) = (2528, 5421, 1375, 478, 123, 52, 16, 5, 2)$. 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 with a 95% central interval of $(25, 29)$. Within household-level classes, the posterior number of occupied individual-level classes ranges from 5 to 8. To monitor convergence of the MCMC sampler, we focus of π, α, and β, which are not subject to label switching.

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.

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, $(X_{ip+1}^{(l)}, \ldots, X_{ip+q}^{(l)})$, 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, $(X_{ij1}^{(l)}, \ldots, X_{ijp}^{(l)})$, 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.

To evaluate the quality of the NDPMPM model, we compare the relationships among the variables in the original and synthetic datasets. 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. Plots in Figure 1 display each $\bar{q}_5$ plotted against its corresponding empirical probability in the original data. 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.

We also examine several probabilities that depend on values for individuals in the same household. As evident in Table 2, for most quantities the syn-
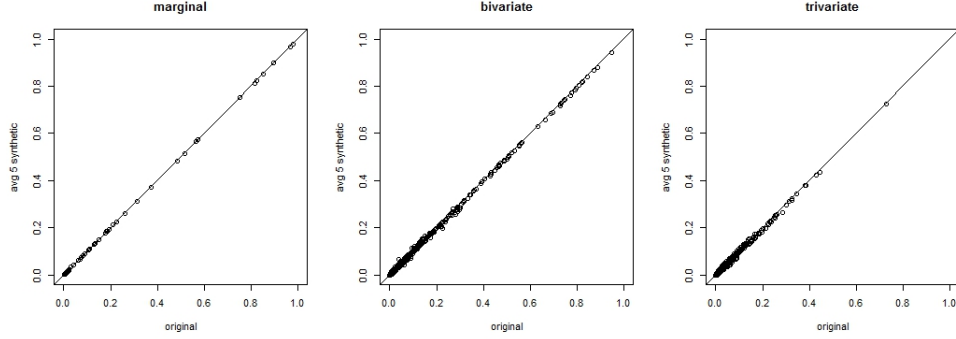
FIG 1. *Marginal, bivariate and trivariate probabilities computed in the original and synthetic datasets for the illustration without structural zeros.*

	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 full ACS public use sample of 308769 households.

thetic data point and interval estimates are similar to those based on the original data, 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 ACS public use file, 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.

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.

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. The DPMPM also struggles for other quantities involving racial compositions. Unlike the NDPMPM model, the non-nested 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

We now apply the truncated NDPMPM to data comprising $n = 10000$ households and $N = 26756$ individuals randomly sampled from the 2011 ACS public use file (Ruggles et al., 2010). We select variables to mimic those on the U. S. decennial census, including a variable indicating relationships among individuals within the same household. This 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 two household-level variables and five individual-level variables, summarized in Table 3. We ex-

clude 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 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. Household sizes are $(n_2, n_3, n_4) = (5370, 2504, 2126)$.

We run the MCMC sampler for the truncated NDPMPM model of Section 3 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 with a 95% central interval of (25, 31). 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 auto-correlations. 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.
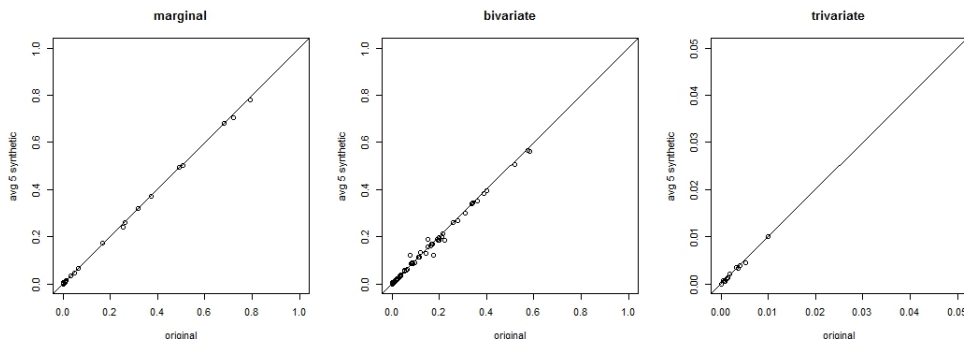

FIG 2. *Marginal, bivariate and trivariate distributions probabilities computed in the original and synthetic datasets in illustration with structural zeros. Plot restricted to cells with at least ten individuals.*

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. Plots in Figure 2 display each $\bar{q}_5$ plotted against its corresponding probability from the original data. The point estimates are quite similar, indicating that the NDPMPM captures relationships among the variables.

	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 ACS public use sample of 127685 households.

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 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, 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 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 nested 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 $\mathcal{X}^0$. 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 bottlenecks 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 upper bounds on n_{0h} (that are still much larger than n_{*h}). Although this results in an approximation to the Gibbs sampler, this still could yield reasonable inferences or synthetic datasets, particularly when many records in $\mathcal{X}^0$ 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 the probit specification of Albert and Chib (1993) or the rank likelihood of 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 (2014). Evaluating the properties of such models is a topic for future research.

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 (and 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. Kunihama et al. (2014) present a related and computationally convenient approach that uses only the survey weights for

sampled cases. We conjecture that a similar approach could be applied for nested categorical data.

Appendix A: Full conditional distributions for MCMC samplers

We present the full conditional distributions used in the Gibbs samplers for the versions of the NDPMPM with and without structural zeros. In both presentations, we assume common β for all household-level clusters.

NDPMPM without structural zeros

Step 1: Sample $G_i \in \{1, \ldots, F\}$ from a multinomial distribution with sample size one and probabilities

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

Step 2: Sample $M_{ij} \in \{1, \ldots, S\}$ given G_i from a multinomial distribution with sample size one and probabilities

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

Step 3: Set $u_F = 1$. Sample u_g from the Beta distribution for $g = 1, \ldots, F-1$, where

$$(u_g|-) \sim \text{Beta}(1 + \sum_{i=1}^{n} \mathbb{1}(G_i = g), \alpha + \sum_{f=g+1}^{F} \sum_{i=1}^{n} \mathbb{1}(G_i = f))$$

$$\pi_g = u_g \prod_{f<g} (1 - u_f).$$

Step 4: Set $v_{gM} = 1$. Sample v_{gm} from the Beta distribution for $m = 1, \ldots, S-1$, where

$$(v_{gm}|-) \sim \text{Beta}(1 + \sum_{i=1}^{n} \mathbb{1}(M_{ij} = m, G_i = g), \beta + \sum_{s=m+1}^{S} \sum_{i=1}^{n} \mathbb{1}(M_{ij} = s, G_i = g))$$

$$\omega_{gm} = v_{gm} \prod_{s<m} (1 - v_{gs}).$$

Step 5: Sample $\lambda_g^{(k)}$ from the Dirichlet distribution for $g = 1, \ldots, F$, and $k = p + 1, \ldots, q$, where

$$(\lambda_g^{(k)}|-) \sim \text{Dir}(a_{k1} + \sum_{i|G_i=g}^{n} \mathbb{1}(X_{ik} = 1), \ldots, a_{kd_k} + \sum_{i|G_i=g}^{n} \mathbb{1}(X_{ik} = d_k)).$$

Step 6: Sample $\phi_{gm}^{(k)}$ from the Dirichlet distribution for $g = 1, \ldots, F$, $m = 1, \ldots, S$ and $k = 1, \ldots, p$, where

$$(\phi_{gm}^{(k)}|-) \sim \text{Dir}(a_{k1} + \sum_{i,j|\substack{G_i=g, \\ M_{ij}=m}}^{n,n_i} \mathbb{1}(X_{ijk} = 1), \ldots, a_{kd_k} + \sum_{i,j|\substack{G_i=g, \\ M_{ij}=m}}^{n,n_i} \mathbb{1}(X_{ijk} = d_k)).$$

Step 7: Sample α from the Gamma distribution,

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

Step 8: Sample β from the Gamma distribution,

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

NDPMPM with structural zeros

Step 1: In each MCMC iteration, we have to sample $(n_{0h}, \{\mathcal{X}_i^0, G_i^0, M_{ij}^0 : i = 1, \ldots, n_0, j = 1, \ldots, h\})$ for each $h \in \mathcal{H}$. We do so by means of a rejection sampler. To begin, we initialize $\mathcal{X}^0 = \emptyset$ at each MCMC iteration. For each $h \in \mathcal{H}$, we repeat the following steps.

1a. Set $t_0 = 0$. Set $t_1 = 0$.
1b. Sample a value of G_i from a multinomial distribution with sample size one and $Pr(G_i = g|-) \propto Pr(X_{ik} = n_i \mid G_i = g)\pi_g$, where X_{ik} corresponds to the variable for household size.
1c. For $j = 1, \ldots, h$, sample a value of M_{ij} from a multinomial distribution with sample size one and $Pr(M_{ij} = m|-) = \omega_{G_{im}}$.
1d. Set $X_{ik} = h$. Sample remaining household level values and all individual level values using (1) and (2). Let $\mathcal{X}_i^0$ be the simulated value.
1e. If $\mathcal{X}_i^0 \in \mathcal{S}_h$, let $t_0 = t_0 + 1$ and $\mathcal{X}^0 = \mathcal{X}^0 \cup \mathcal{X}_i^0$. Otherwise set $t_1 = t_1 + 1$.
1f. If $t_1 < n_{*h}$, return to Step 1b. Otherwise set $n_{0h} = t_0$.

Step 2: For observations in $\mathcal{X}^*$, sample $G_i \in \{1, \ldots, F\}$ from a multinomial distribution with sample size one and

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

Step 3: For observations in $\mathcal{X}^*$, sample $M_{ij} \in \{1, \ldots, S\}$ given G_i from a multinomial distribution with sample size one and

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

Step 4: Set $u_F = 1$. Sample u_g from the Beta distribution for $g = 1, \ldots, F-1$, where

$$(u_g|-) \sim \text{Beta}(1 + \sum_{i=1}^{n} \mathbb{1}(G_i = g), \alpha + \sum_{f=g+1}^{F} \sum_{i=1}^{n+n_0} \mathbb{1}(G_i = f))$$

$$\pi_g = u_g \prod_{f<g}(1 - u_f).$$

Step 5: Set $v_{gM} = 1$. Sample v_{gm} from the Beta distribution for $m = 1, \ldots, S-1$, where

$$(v_{gm}|-) \sim \text{Beta}(1 + \sum_{i=1}^{n+n_0} \mathbb{1}(M_{ij} = m, G_i = g), \beta + \sum_{s=m+1}^{S} \sum_{i=1}^{n+n_0} \mathbb{1}(M_{ij} = s, G_i = g))$$

$$\omega_{gm} = v_{gm} \prod_{s<m}(1 - v_{gs}).$$

Step 6: Sample $\lambda_g^{(k)}$ from the Dirichlet distribution for $g = 1, \ldots, F$, and $k = p+1, \ldots, q$, where

$$(\lambda_g^{(k)}|-) \sim \text{Dir}(a_{k1} + \sum_{i|G_i=g}^{n+n_0} \mathbb{1}(X_{ik} = 1), \ldots, a_{kd_k} + \sum_{i|G_i=g}^{n+n_0} \mathbb{1}(X_{ik} = d_k)).$$

Step 7: Sample $\phi_{gm}^{(k)}$ from the Dirichlet distribution for $g = 1, \ldots, F$, $m = 1, \ldots, S$ and $k = 1, \ldots, p$, where

$$(\phi_{gm}^{(k)}|-) \sim \text{Dir}(a_{k1} + \sum_{\substack{i,j|_{M_{ij}=m}^{G_i=g,}}}^{n+n_0,n_i} \mathbb{1}(X_{ijk} = 1), \ldots, a_{kd_k} + \sum_{\substack{i,j|_{M_{ij}=m}^{G_i=g,}}}^{n+n_0,n_i} \mathbb{1}(X_{ijk} = d_k)).$$

Step 8: Sample α from the Gamma distribution,

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

Step 9: Sample β from the Gamma distribution,

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

Appendix B: Proof of result ([16]) in Section [3]

We now prove that one can obtain samples from $p(\theta|\mathcal{X}^*)$ in the truncated NDPMPM in ([15]) from a sampler for $f(\theta, \mathbf{G}^*, \mathbf{G}^0, \mathbf{M}^*, \mathbf{M}^0, \mathcal{X}^0, \{n_{0h} : h \in \mathcal{H}\} \mid \mathcal{X}^*)$ under an untruncated NDPMPM model. That is, we seek to show that

$$\int f(\theta, \mathbf{G}^*, \mathbf{G}^0, \mathbf{M}^*, \mathbf{M}^0, \mathcal{X}^0, \{n_{0h}\}|\mathcal{X}^*, \{n_{*h}\})d\mathcal{X}^0 d\mathbf{G}^* d\mathbf{G}^0 d\mathbf{M}^* d\mathbf{M}^0 d\{n_{0h}\}$$
$$= p(\theta \mid \mathcal{X}^*, \{n_{*h}\}). \tag{17}$$

Here, we use integration signs rather than summation signs to simplify notation. We write $\{n_{*h}\} = \{n_{*h} : h \in \mathcal{H}\}$ as part of the conditional information, even though it is already implicit in $\mathcal{X}^*$. This is to emphasize that the MCMC algorithm proceeds separately for each h, generating until reaching n_{*h} feasible households.

Let $\mathbf{G}_h = \{G_i : n_i = h\}$ be the household level latent class assignments of size h households and $\mathbf{M}_h = \{M_{ij} : n_i = h, j = 1, \ldots, n_i\}$ be the individual level latent class assignments associated with members of size h households. We split $\mathbf{G}_h$ into $\mathbf{G}_h^*$ and $\mathbf{G}_h^0$, representing the values for records in $\mathcal{X}^*$ and in $\mathcal{X}^0$ respectively. We similarly split $\mathbf{M}_h$ into $\mathbf{M}_h^*$ and $\mathbf{M}_h^0$. Let $\mathcal{X}_h^* = \{\mathbf{X}_i^* : n_i = h\}$, and let $\mathcal{X}_h^0 = \{\mathbf{X}_i^0 : n_i = h\}$. We emphasize that each $\mathbf{X}_i^*$ is observed, whereas each $\mathbf{X}_i^0$ is imputed in the MCMC sampler. Using this notation, we have

$$\int f(\theta, \mathbf{G}^*, \mathbf{G}^0, \mathbf{M}^*, \mathbf{M}^0, \mathcal{X}^0, \{n_{0h}\}|\mathcal{X}^*, \{n_{*h}\})d\mathcal{X}^0 d\mathbf{G}^* d\mathbf{G}^0 d\mathbf{M}^* d\mathbf{M}^0 d\{n_{0h}\}$$

$$\propto p(\theta) \prod_{h \in \mathcal{H}} \int f(\mathcal{X}_h^*, \mathbf{G}_h^*, \mathbf{M}_h^*, \mathcal{X}_h^0, \mathbf{G}_h^0, \mathbf{M}_h^0 \mid \theta, n_{0h})p(n_{0h} \mid n_{*h})$$
$$d\mathbf{G}_h^*, d\mathbf{M}_h^* d\mathcal{X}_h^0 d\mathbf{G}_h^0, d\mathbf{M}_h^0 dn_{0h}. \tag{18}$$

Extending the generative model in Section [2], we view each $\mathcal{X}_h^*$ as a truncated sample from the households in $\mathcal{X}$ of size h. This implies for any given value of

n_{0h} that

$$f(\mathcal{X}_h^*, \mathbf{G}_h^*, \mathbf{M}_h^*, \mathcal{X}_h^0, \mathbf{G}_h^0, \mathbf{M}_h^0 \mid \theta, n_{0h}) = \binom{n_{0h} + n_{*h}}{n_{0h}} \prod_{i:n_i=h} \mathbb{1}\{\mathbf{X}_i^* \notin \mathcal{S}_h\} f(\mathbf{X}_i^* | G_i^*, \mathbf{M}_i^*, \theta)$$

$$f(G_i^*, \mathbf{M}_i^0 | \theta) \prod_{i=1}^{n_{0h}} \mathbb{1}\{\mathbf{X}_i^0 \in \mathcal{S}_h\}$$

$$f(\mathbf{X}_i^0 \mid G_i^0, \mathbf{M}_i^0, \theta) f(G_i^0, \mathbf{M}_i^0 | \theta). \qquad (19)$$

Substituting (19) in (18) and expanding the integrals, we have

$$p(\theta) \prod_{h \in \mathcal{H}} \int f(\mathcal{X}_h^*, \mathbf{G}_h^*, \mathbf{M}_h^*, \mathcal{X}_h^0, \mathbf{G}_h^0, \mathbf{M}_h^0 \mid \theta, n_{0h}) p(n_{0h} \mid n_{*h}) d\mathbf{G}_h^*, d\mathbf{M}_h^* d\mathcal{X}_h^0 d\mathbf{G}_h^0, d\mathbf{M}_h^0 dn_{0h}$$

$$= p(\theta) \prod_{h \in \mathcal{H}} \prod_{i:n_i=h} \int \mathbb{1}\{\mathbf{X}_i^* \notin \mathcal{S}_h\} f(\mathbf{X}_i^* | G_i^*, \mathbf{M}_i^*, \theta) f(G_i^* | \theta) \prod_{j=1}^{h} f(M_{ij}^* | \theta) dG_i^* d\mathbf{M}_i^*$$

$$\prod_{h \in \mathcal{H}} \sum_{n_{0h}=0}^{\infty} p(n_{0h} | n_{*h}) \binom{n_{0h} + n_{*h}}{n_{0h}} \prod_{i=1}^{n_{0h}} \int \mathbb{1}\{\mathbf{X}_i^0 \in \mathcal{S}_h\} f(\mathbf{X}_i^0 \mid G_i^0, \mathbf{M}_i^0, \theta)$$

$$f(G_i^0, \mathbf{M}_i^0 | \theta) dG_i^0 d\mathbf{M}_i^0$$

$$\propto p(\theta) \prod_{h \in \mathcal{H}} \prod_{i:n_i=h} \int \mathbb{1}\{\mathbf{X}_i^* \notin \mathcal{S}_h\} f(\mathbf{X}_i^* | G_i^*, \mathbf{M}_i^*, \theta) f(G_i^* | \theta) \prod_{j=1}^{h} f(M_{ij}^* | \theta) d\mathbf{G}_h^* d\mathbf{M}_h^*$$

$$\prod_{h \in \mathcal{H}} \sum_{n_{0h}=0}^{\infty} \binom{n_{0h} + n_{*h} - 1}{n_{0h}} (\pi_{0h}(\theta))^{n_{0h}}$$

$$= p(\theta) \prod_{h \in \mathcal{H}} (\prod_{i:n_i=h} \int \mathbb{1}\{\mathbf{X}_i^* \notin \mathcal{S}_h\} f(\mathbf{X}_i^* | G_i^*, \mathbf{M}_i^*, \theta) f(G_i^* | \theta)$$

$$\prod_{j=1}^{h} f(M_{ij}^* | \theta) d\mathbf{G}_h^* d\mathbf{M}_h^* (1 - \pi_{0h}(\theta))^{-n_{*h}})$$

From (14) and (15), this expression is equivalent to $p(\theta) \prod_{i=1}^{n} p(\mathcal{X}_i^* | \theta)$, as desired.

Supplementary Materials

Supplementary Materials for "Dirichlet Process Mixture Models for Nested Categorical Data supplementary material" (DOI: ; .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.

Albert, J. H. and Chib, S. (1993). "Bayesian Analysis of Binary and Polychotomous Response Data." *Journal of the American Statistical Association*, 88: 669–679.

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." In *33rd International Colloquium on Automata, Languages, and Programming, part II*, 1–12. Berlin: Springer.

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

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

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

Hoff, P. D. (2009). *A First Course in Bayesian Statistical Methods*. Springer Dordrecht Heidelberg London New York.

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.

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

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

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.

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

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

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.

Manrique-Vallier, D. and Reiter, J. P. (2014). "Bayesian estimation of discrete multivariate latent structure models with strutural zeros." *Journal of Computational and Graphical Statistics*, 23: 1061 – 1079.

Murray, J. S. and Reiter, J. P. (2014). "Multiple Imputation of Missing Categorical and Continuous Values via Bayesian Mixture Models with Local Dependence." *arXiv:1410.0438*.

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

1471.

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

Rodriguez, A., Dunson, D. B., and Gelfand, A. E. (2008). "The Nested Dirichelt Process." *Journal of the American Statistical Association*, 103: 1131–1154.

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

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*.

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: 499–521.

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

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

Introduction

This supplement presents the results of the assessments of disclosure risks for the synthetic data illustrations in Section 4 of the main text. In the supplement, we describe the methodology for assessing risks in Section 2 and the computational methods we use in Section 3. We summarize the disclosure risk evaluations for the scenario without and with structural zeros in Section 4 and Section 5, respectively.

Disclosure risk measures

When synthesizing entire household compositions (but keeping household size distributions fixed), it is nonsensical for intruders to match the proposed synthetic datasets to external files, since there is no unique mapping of the rows (individuals) in the synthetic datasets $\mathbf{Z}$ to the rows in the original data $\mathbf{D}$, nor unique mapping of the households in $\mathbf{Z}$ to the households in $\mathbf{D}$ (except for household sizes with $n^h = 1$). We therefore consider questions of the form: can intruders accurately infer from Z that some individual or entire household with a particular set of data values is in the confidential data? When the combination of values is unique in the population (or possibly just the sample), this question essentially asks if intruders can determine whether or not a specific individual or household is in D (Hu et al., 2014).

To describe the disclosure risk evaluations, we follow the presentation of Hu et al. (2014). We consider two possible attacks on $\mathbf{Z}$, namely (i) the intruder seeks to learn whether or not someone with a particular combination of the p individual-level variables and the q household-level variables is in $\mathbf{D}$, and (ii) an intruder seeks to learn whether or not an entire household with a particular combination of household-level and individual-level characteristics is in $\mathbf{D}$. For the first scenario, we assume that the intruder knows the values in $\mathbf{D}$ for all individuals but the target individual, say individual ij. We use $\mathbf{D}_{-ij}$ to denote the data known to the intruder. For the second scenario, we assume that the intruder knows the values in D for all households but the target house, say household i. We use $\mathbf{D}_{-i}$ to denote the data known to the intruder. In many cases, assuming the intruder knows $\mathbf{D}_{-ij}$ or $\mathbf{D}_{-i}$ is conservative; for example, in random samples from large populations intruders are unlikely to know $N-1$ individuals or $n-1$ households selected in the sample. We adopt this strong assumption largely to facilitate computation. Risks deemed acceptable under this assumption should be acceptable for weaker intruder knowledge. We note that assuming the intruder knows all records but one is related to, but quite distinct from, the assumptions used in differential privacy (Dwork, 2006).

Let T_{ij} or T_i be the random variable corresponding to the intruder's guess about the true values of the target. Let t generically represent a possible guess at the target, where for simplicity of notation we use a common notation for individual and household targets. Let $\mathcal{I}$ represent any information known by the intruder about the process of generating Z, for example meta-data indicating the values of F, S and $(a_\alpha, b_\alpha, a_\beta, b_\beta)$ for the NDPMPM synthesizer.

For the first type of attack, we assume the intruder seeks the posterior probability,

$$
\rho_{ij}^t = p(T_{ij} = t \mid \mathbf{Z}, \mathbf{D}_{-ij}, \mathcal{I}) \;=\; \frac{p(\mathbf{Z} \mid T_{ij} = t, \mathbf{D}_{-ij}, \mathcal{I}) p(T_{ij} = t \mid \mathbf{D}_{-ij}, \mathcal{I})}{\sum_{t \in \mathcal{U}} p(\mathbf{Z} \mid T_{ij} = t, \mathbf{D}_{-ij}, \mathcal{I}) p(T_{ij} = t \mid \mathbf{D}_{-ij}, \mathcal{I})} \tag{20}
$$
$$
\propto \; p(\mathbf{Z} \mid T_{ij} = t, \mathbf{D}_{-ij}, \mathcal{I}) p(T_{ij} = t \mid \mathbf{D}_{-ij}, \mathcal{I}), \tag{21}
$$

where $\mathcal{U}$ represents the universe of all feasible values of t. Here, $p(\mathbf{Z} \mid T_{ij} = t, \mathbf{D}_{-ij}, \mathcal{I})$ is the likelihood of generating the particular set of synthetic data given that t is in the confidential data and whatever else is known by the intruder. The $p(T_{ij} = t \mid \mathbf{D}_{-ij}, \mathcal{I})$ can be considered the intruder's prior distribution on T_{ij} based on $(\mathbf{D}_{-ij}, \mathcal{I})$.

As described in Hu et al. (2014), intruders can use $p(T_{ij} = t \mid \mathbf{Z}, \mathbf{D}_{-ij}, \mathcal{I})$ to take guesses at the true value t_{ij}. For example, the intruder can find the t that offers the largest probability, and use that as a guess of t_{ij}. Similarly, agencies can use $p(T_{ij} = t \mid \mathbf{Z}, \mathbf{D}_{-ij}, \mathcal{I})$ in disclosure risk evaluations. For example, for each $t_{ij} \in \mathbf{D}$, they can rank each t by its associated value of $p(T_{ij} = t \mid \mathbf{Z}, \mathbf{D}_{-ij}, \mathcal{I})$, and evaluate the rank at the truth, $t = t_{ij}$. When the rank of t_{ij} is high (close to 1, which we define to be the rank associated with the highest probability), the agency may deem that record to be at risk under the strong intruder knowledge scenario. When the rank of t_{ij} is low (far from 1), the agency may deem the risks for that record to be acceptable.

When $\mathcal{U}$ is very large, computing the normalizing constant in (20) is impractical. To facilitate computation, we follow Hu et al. (2014) and consider as feasible candidates only those t that differ from t_{ij} in one variable, along with t_{ij} itself; we call this space $\mathcal{R}_{ij}$. Restricting to $\mathcal{R}_{ij}$ can be conceived as mimicking a knowledgeable intruder who searches in spaces near t_{ij}. As discussed by Hu et al. (2014), restricting support to $\mathcal{R}_{ij}$ results in a conservative ranking of the $t \in \mathcal{R}_{ij}$, in that ranks determined to be acceptably low when using $\mathcal{R}_{ij}$ also are acceptably low when using $\mathcal{U}$.

For T_i, we use a similar approach to risk assessment. We compute

$$
\rho_i^t = p(T_i = t \mid \mathbf{Z}, \mathbf{D}_{-i}, \mathcal{I}) \propto p(\mathbf{Z} \mid T_i = t, \mathbf{D}_{-i}, \mathcal{I}) p(T_i = t \mid \mathbf{D}_{-i}, \mathcal{I}). \tag{22}
$$

We consider only t that differ from t_i in either (i) one household-level variable for the entire household or (ii) one individual-level variable for one household member, along with t_i itself; we call this space $\mathcal{R}_i$.

Computational methods for risk assessment with the NDPMPM model

We describe the computational methods for computing (22) in detail. Methods for computing (21) are similar.

For any proposed t, let $\mathbf{D}_i^t = (T_i = t, \mathbf{D}_{-i})$ be the plausible confidential dataset when $T_i = t$. Because each $\mathbf{Z}^{(l)}$ is generated independently, we have

$$
P(\mathbf{Z} \mid \mathbf{D}_i^t, \mathcal{I}) = \prod_{l=1}^{L} P(\mathbf{Z}^{(l)} \mid \mathbf{D}_i^t, \mathcal{I}). \tag{23}
$$

Hence, we need to compute each $P(\mathbf{Z}^{(l)} \mid \mathbf{D}_i^t, \mathcal{I})$.

Let $\Theta = \{\pi, \omega, \lambda, \phi\}$ denote parameters from a NDPMPM model. We can write $P(\mathbf{Z}^{(l)} \mid \mathbf{D}_i^t, \mathcal{I})$ as

$$P(\mathbf{Z}^{(l)} \mid \mathbf{D}_i^t, \mathcal{I}) = \int p(\mathbf{Z}^{(l)} \mid \mathbf{D}_i^t, \mathcal{I}, \Theta) p(\Theta \mid \mathbf{D}_i^t, \mathcal{I}) d\Theta. \tag{24}$$

To compute (24), we could sample many values of Θ that could have generated $\mathbf{Z}^{(l)}$; that is, we could sample $\Theta^{(r)}$ for $r = 1, \ldots, R$. For each $\Theta^{(r)}$, we compute the probability of generating the released $\mathbf{Z}^{(l)}$. We then average these probabilities over the R draws of Θ.

Conceptually, to draw Θ replicates, we could re-estimate the NDPMPM model for each $\mathbf{D}_i^t$. This quickly becomes computationally prohibitive. Instead, we suggest using the sampled values of Θ from $p(\Theta \mid \mathbf{D})$ as proposals for an importance sampling algorithm. To set notation, suppose we seek to estimate the expectation of some function $g(\Theta)$, where Θ has density $f(\Theta)$. Further suppose that we have available a sample $(\Theta^{(1)}, \ldots, \Theta^{(R)})$ from a convenient distribution $f^*(\Theta)$ that slightly differs from $f(\Theta)$. We can estimate $E_f(g(\Theta))$ using

$$E_f(g(\Theta)) \approx \sum_{r=1}^R g(\Theta^{(r)}) \frac{f(\Theta^{(r)})/f^*(\Theta^{(r)})}{\sum_{r=1}^R f(\Theta^{(r)})/f^*(\Theta^{(r)})}. \tag{25}$$

Let $t_i^{*(l)}$ be the ith household's values of all variables, including household-level and individual-level variables, in synthetic dataset $\mathbf{Z}^{(l)}$, where $i = 1, \ldots, n$ and $l = 1, \ldots, L$. For each $\mathbf{Z}^{(l)}$ and any proposed t, we define the $g(\Theta)$ in (25) to equal $cP(\mathbf{Z}^{(l)} \mid \mathbf{D}_i^t, \mathcal{I})$. We approximate the expectation of each $g(\Theta)$ with respect to $f(\Theta) = f(\Theta \mid \mathbf{D}_i^t, \mathcal{I})$. In doing so, for any sampled $\Theta^{(r)}$ we use

$$g(\Theta^{(r)}) = P(\mathbf{Z}^{(l)} \mid \mathbf{D}_i^t, \mathcal{I}, \Theta^{(r)}) = \prod_{i=1}^n \left(\sum_{g=1}^F \pi_g^{(r)} \{ \prod_{k=p+1}^{p+q} \lambda_{g t_{ik}^{*(l)}}^{(k)(r)} (\prod_{j=1}^{n_i} \sum_{m=1}^S \omega_{gm}^{(r)} \prod_{k=1}^p \phi_{g m t_{ijk}^{*(l)}}^{(k)(r)}) \} \right). \tag{26}$$

We set $f^*(\Theta) = f(\Theta \mid \mathbf{D}, \mathcal{I})$, so that we can use R draws of Θ from its posterior distribution based on $\mathbf{D}$. Let these R draws be $(\Theta^{(1)}, \ldots, \Theta^{(R)})$. We note that one could use any $\mathbf{D}_i^t$ to obtain the R draws, so that intruders can use similar importance sampling computations. As evident in (1), (2), (3) and (4) in the main text,

the only differences in the kernels of $f(\Theta)$ and $f^*(\Theta)$ include (i) the components of the likelihood associated with record i and (ii) the normalizing constant for each density. Let $\mathbf{t} = \{(c_{p+1}, \ldots, c_{p+q}), (c_{j1}, \ldots, c_{jp}), j = 1, \ldots, n_i\}$, where each $c_k \in (1, \ldots, d_k)$, be a guess at T_i, for household-level and individual-level variables respectively. After computing the normalized ratio in (25) and canceling common terms from the numerator and denominator, we are left with $P(\mathbf{Z}^{(l)} \mid \mathbf{D}_i^t, \mathcal{I}) = \sum_{r=1}^R p_r q_r$ where

$$p_r = \prod_{i=1}^{n} \left(\sum_{g=1}^{F} \pi_g^{(r)} \{ \prod_{k=p+1}^{p+q} \lambda_{gt_{ik}^{*(l)}}^{(k)(r)} (\prod_{j=1}^{n_i} \sum_{m=1}^{S} \omega_{gm}^{(r)} \prod_{k=1}^{p} \phi_{gmt_{ijk}^{*(l)}}^{(k)(r)}) \} \right) \qquad (27)$$

$$q_r = \frac{\dfrac{\sum_{g=1}^{F} \pi_g^{(r)} \{ \prod_{k=p+1}^{p+q} \lambda_{gc_k}^{(k)(r)} (\prod_{j=1}^{n_i} \sum_{m=1}^{S} \omega_{gm}^{(r)} \prod_{k=1}^{p} \phi_{gme_{jk}}^{(k)(r)}) \}}{\sum_{g=1}^{F} \pi_g^{(r)} \{ \prod_{k=p+1}^{p+q} \lambda_{gt_{ik}}^{(k)(r)} (\prod_{j=1}^{n_i} \sum_{m=1}^{S} \omega_{gm}^{(h)} \prod_{k=1}^{p} \phi_{gmt_{ijk}}^{(k)(r)}) \}}}{\sum_{u=1}^{R} \left(\dfrac{\sum_{g=1}^{F} \pi_g^{(u)} \{ \prod_{k=p+1}^{p+q} \lambda_{gc_k}^{(k)(u)} (\prod_{j=1}^{n_i} \sum_{m=1}^{S} \omega_{gm}^{(u)} \prod_{k=1}^{p} \phi_{gme_{jk}}^{(k)(u)}) \}}{\sum_{g=1}^{F} \pi_g^{(u)} \{ \prod_{k=p+1}^{p+q} \lambda_{gt_{ik}}^{(k)(u)} (\prod_{j=1}^{n_i} \sum_{m=1}^{S} \omega_{gm}^{(u)} \prod_{k=1}^{p} \phi_{gmt_{ijk}}^{(k)(u)}) \}} \right)} \qquad (28)$$

We repeat this computation for each $\mathbf{Z}^{(l)}$, plugging the L results into (23).

Finally, to approximate ρ_i^t, we compute (23) for each $t \in \mathcal{R}_i$, multiplying each resulting value by its associated $P(T_i = t \mid \mathbf{D}_{-i}, \mathcal{I})$. In what follows, we presume an intruder with a uniform prior distribution over the support $t \in \mathcal{R}_i$. In this case, the prior probabilities cancel from the numerator and denominator of (20), so that risk evaluations are based only on the likelihood function for $\mathbf{Z}$. We discuss evaluation of other prior distributions in the illustrative application.

For risk assessment for T_{ij} in (21), we use a similar importance sampling approximation, resulting in

$$q_r = \frac{\dfrac{\sum_{g=1}^{F} \pi_g^{(r)} \{ \prod_{k=p+1}^{p+q} \lambda_{gc_k}^{(k)(r)} (\sum_{m=1}^{S} \omega_{gm}^{(r)} \prod_{k=1}^{p} \phi_{gme_{jk}}^{(k)(h)}) \}}{\sum_{g=1}^{F} \pi_g^{(r)} \{ \prod_{k=p+1}^{p+q} \lambda_{gt_{ik}}^{(k)(h)} (\sum_{m=1}^{S} \omega_{gm}^{(r)} \prod_{k=1}^{p} \phi_{gmt_{ijk}}^{(k)(h)}) \}}}{\sum_{u=1}^{R} \left(\dfrac{\sum_{g=1}^{F} \pi_g^{(u)} \{ \prod_{k=p+1}^{p+q} \lambda_{gc_k}^{(k)(u)} (\sum_{m=1}^{S} \omega_{gm}^{(u)} \prod_{k=1}^{p} \phi_{gme_{jk}}^{(k)(u)}) \}}{\sum_{g=1}^{F} \pi_g^{(u)} \{ \prod_{k=p+1}^{p+q} \lambda_{gt_{ik}}^{(k)(u)} (\sum_{m=1}^{S} \omega_{gm}^{(u)} \prod_{k=1}^{p} \phi_{gmt_{ijk}}^{(k)(u)}) \}} \right)}. \qquad (29)$$

Disclosure risk assessments for synthesis without structural zeros

To evaluate the disclosure risks for individuals, we drop each individual record in $\mathbf{D}$ one at a time. For each individual ij, we compute the resulting ρ_{ij}^t for all t in the reduced support $\mathcal{R}_{ij}$. Here, each $\mathcal{R}_{ij}$ is the union of the true t_{ij} plus the 39 other combinations of t obtained by changing one variable in t_{ij} to any possible outcome. For any two records ij and $i'j'$ such that $t_{ij} = t_{i'j'}$ in $\mathbf{D}$, $\rho_{ij}^t = \rho_{i'j'}^t$ for any possible t. Thus, we need only compute the set of ρ_{ij}^t for the 15280 combinations that appeared in $\mathbf{D}$. We use a uniform prior distribution over all $t \in \mathcal{R}_{ij}$, for each record ij.

Figure 3 displays the distribution of the rank of the true t_{ij} for each of the 15280 combinations. Here, a rank equal to 1 means the true t_{ij} has the highest probability of being the unknown T_{ij}, whereas a rank of 40 means the true t_{ij} has the lowest probability of being T_{ij}. As evident in the figures, even armed with $\mathbf{D}_{-ij}$ the intruder gives the top rank to the true t_{ij} for only 11 combinations. The intruder gives t_{ij} a ranking in the top three for only 194 combinations. We note that, even though 12964 combinations were unique in $\mathbf{D}$, the NDPMPM synthesizer involves enough smoothing that we do not recover the true t_{ij} in the overwhelming majority of cases.

Figure 4 displays a histogram of the corresponding probabilities associated with the true t_{ij} in each of the 15280 combinations. The largest probability is 0.2360. Only 1 probability exceeds 0.2, and 40 probabilities exceed 0.1. The majority of probabilities are in the 0.03 range. As we assumed a uniform prior distribution over the 40 possibilities in R_{ij}, the ratio of the posterior to prior probability is typically around one. Only a handful of combinations have ratios exceeding two. Thus, compared to random guesses over a close neighborhood of the true values, $\mathbf{Z}$ typically does not provide much additional information about t_{ij}. We also look at the disclosure risks for households. To do so, we drop each household record in $\mathbf{D}$ one at a time. For households of size 2, the reduced support $\mathcal{R}_i$ comprises the true t_i plus the 56 other combinations of t obtained by changing t_i in one variable. For the household-level variables, we change the entire variable for all members of the household. For the individual-level variable, we change one variable for each individual as before. We need only compute ρ_i^t for each of the 5375 combinations of households of size 2 that appear in $\mathbf{D}$. We use a uniform prior distribution over all $t \in \mathcal{R}_i$.
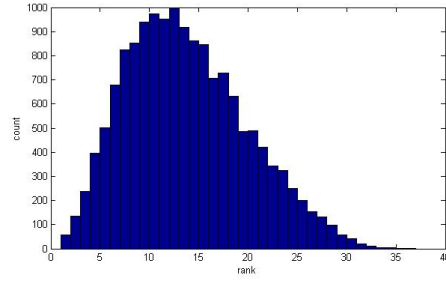
FIG 3. *Histogram of ranks of the probabilities associated with true t_{ij}. Data have no structural zeros.*
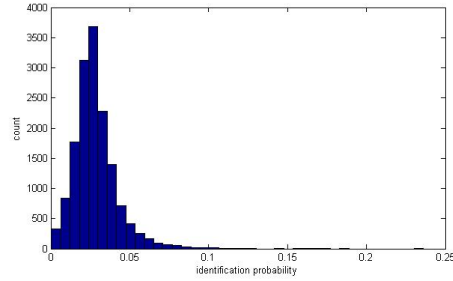
FIG 4. *Histogram of re-normalized probabilities associated with the true t_{ij}. Data have no structural zeros.*
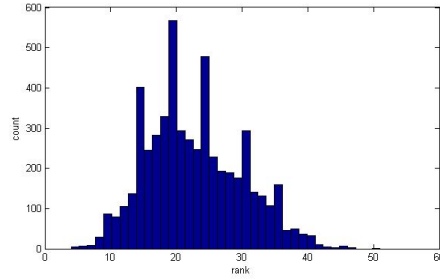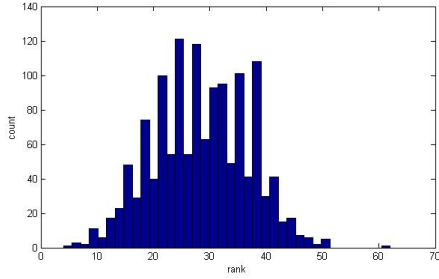

FIG 5. *Histogram of ranks of the probabilities associated with true t_i, for households of size 2. Data have no structural zeros.*
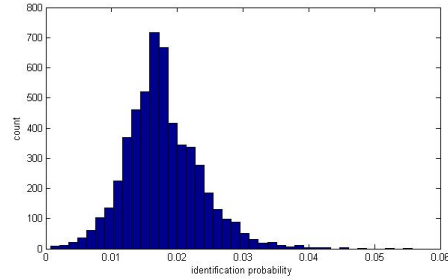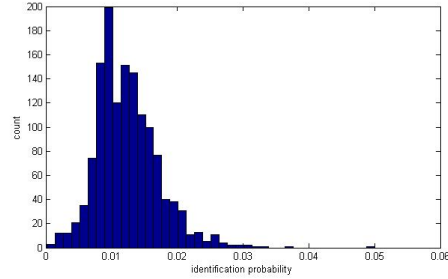
FIG 6. *Histogram of re-normalized probabilities associated with the true t_i, for households of size 2. Data have no structural zeros.*

Figure 5 displays the distribution of the rank of the true t_i for each of the 5375 combinations. Once again, even armed with $\mathbf{D}_{-i}$, the intruder never gives the top rank to the true t_i. the intruder gives the true t_i a ranking in the top three for only seven household combinations. We note that 5331 household combinations of size 2 were unique in $\mathbf{D}$.

Figure 6 displays a histogram of the corresponding probabilities associated with the true t_i in each of the 5375 combinations of households of size 2. The majority of probabilities are in the 0.02 range. As we assumed a uniform prior distribution over the 57 possibilities in the support, the ratio of the posterior to prior probability is typically around one. Thus, as with individuals, compared to random guesses over a close neighborhood of the true values, $\mathbf{Z}$ typically does not provide much additional information about t_i. The largest probability is 0.0557.

For households of size 3, the reduced support $\mathcal{R}_i$ comprises the true t_i plus 81 other combinations of t obtained by changing one variable at a time, as done for households of size 3. We need only compute ρ_i^t for each of the 1375 combinations that appear in $\mathbf{D}$. We use a uniform prior distribution over all $t \in \mathcal{R}_i$.

FIG 7. *Histogram of ranks of the probabilities associated with true t_i, for households of size 3. Data have no structural zeros.*

FIG 8. *Histogram of re-normalized probabilities associated with the true t_i, for households of size 3. Data have no structural zeros.*

Figure 7 displays the distribution of the rank of the true t_i for each of the 1375 combinations. Even armed with $\mathbf{D}_{-i}$, the intruder gives t_i a ranking in the top three for no combinations. We note that all these 1375 combinations were unique in $\mathbf{D}$, yet evidently the nested Dirichlet process synthesizer involves enough smoothing that we do not recover the true t_i in the overwhelming majority of cases.

Figure 8 displays a histogram of the corresponding probabilities associated with the true t_i in each of the 1375 combinations. The majority of probabilities are in the 0.010 range. As we assumed a uniform prior distribution over the 82 possibilities in the support, the ratio of the posterior to prior probability is typically one or less. Thus, compared to random guesses over a reasonably close neighborhood of the true values, $\mathbf{Z}$ typically does not provide much additional information about t_i. The largest probability is 0.0500.

For households of size 4, the reduced support $\mathcal{R}_i$ comprises the true t_i plus 106 other combinations of t obtained by changing one variable at a time, as with the other sizes. We do computations for each of the 478 combinations that appear in **D**. We use a uniform prior distribution over all $t \in \mathcal{R}_i$.
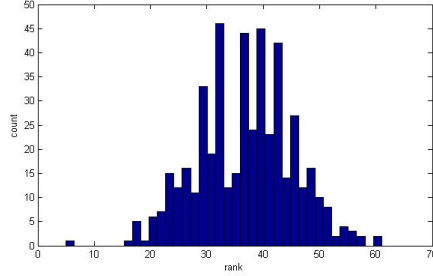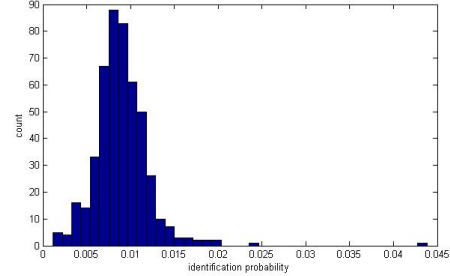

FIG 9. *Histogram of ranks of the probabilities associated with true t_i, for households of size 4. Data have no structural zeros.*

FIG 10. *Histogram of re-normalized probabilities associated with the true t_i, for households of size 4. Data have no structural zeros.*

Figure 9 displays the distribution of the rank of the true t_i for each of the 478 combinations. The intruder gives the true t_i a ranking in the top three for no combinations. All these 478 combinations were unique in **D**. Figure 10 displays a histogram of the corresponding probabilities associated with the true t_i in each of the 478 combinations. The majority of probabilities are in the 0.01 range. As we assumed a uniform prior distribution over the 107 possibilities in the support, the ratio of the posterior to prior probability is typically one or less. Once again, **Z** typically does not provide much additional information about t_i. The largest probability is 0.0438.

For households of size 5, the reduced support $\mathcal{R}_i$ comprises the true t_i plus 131 other combinations of t obtained by changing one variable at a time, as with the other sizes. We do computations for each of the 123 combinations that appear in **D**. We use a uniform prior distribution over all $t \in \mathcal{R}_i$.

Figure 11 displays the distribution of the rank of the true t_i for each of the 123 combinations. The intruder gives the true t_i a ranking in the top three for no combinations. All these 123 combinations were unique in **D**. Figure 12 displays a histogram of the corresponding probabilities associated with the true t_i in each of the 123 combinations. The majority of probabilities are in the 0.008 range. As we assumed a uniform prior distribution over the 132 possibilities in the support, the ratio of the posterior to prior probability is typically around one. Once again, **Z** typically does not provide much additional information about t_i. The largest probability is 0.0292.
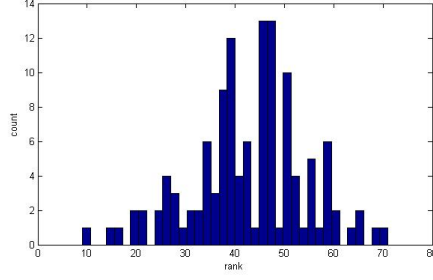
FIG 11. *Histogram of ranks of the probabilities associated with true t_i, for households of size 5. Data have no structural zeros.*
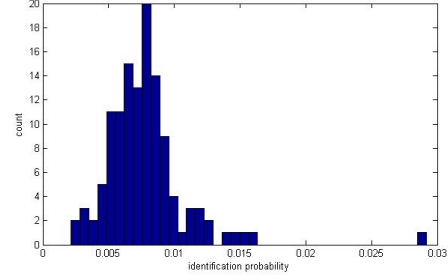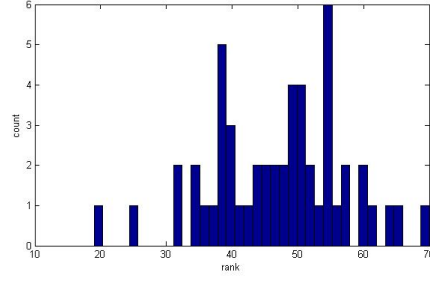
FIG 12. *Histogram of re-normalized probabilities associated with the true t_i, for households of size 5. Data have no structural zeros.*

FIG 13. *Histogram of ranks of the probabilities associated with true t_i, for households of size 6. Data have no structural zeros.*
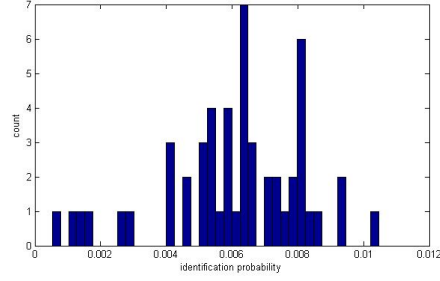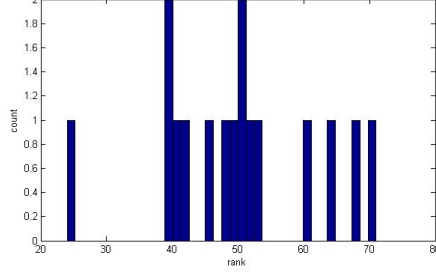
FIG 14. *Histogram of re-normalized probabilities associated with the true t_i, for households of size 6. Data have no structural zeros.*

For households of size 6, the reduced support $\mathcal{R}_i$ comprises the true t_i plus 156 other combinations of t obtained by changing one variable at a time, as with the other sizes. We do computations for each of the 52 combinations that appear in **D**. We use a uniform prior distribution over all $t \in \mathcal{R}_i$.

Figure 13 displays the distribution of the rank of the true t_i for each of the 52 combinations. The intruder gives the true t_i a ranking in the top three for no combinations. All these 52 combinations were unique in **D**. Figure 14 displays a histogram of the corresponding probabilities associated with the true t_i in each of the 52 combinations. The majority of probabilities are in the 0.007 range. As we assumed a uniform prior distribution over the 157 possibilities in the support, the ratio of the posterior to prior probability is typically around one. Once again, **Z** typically does not provide much additional information about t_i. The largest probability is 0.0105.

For households of size 7, the reduced support $\mathcal{R}_i$ comprises the true t_i plus 181 other combinations of t obtained by changing one variable at a time, as with the other sizes. We do computations for each of the 16 combinations that

appear in **D**. We use a uniform prior distribution over all $t \in \mathcal{R}_i$.

FIG 15. *Histogram of ranks of the probabilities associated with true t_i, for households of size 7. Data have no structural zeros.*
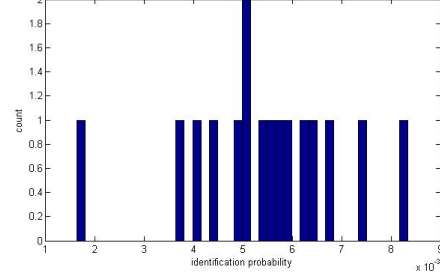
FIG 16. *Histogram of re-normalized probabilities associated with the true t_i, for households of size 7. Data have no structural zeros.*

Figure 15 displays the distribution of the rank of the true t_i for each of the 16 combinations. The intruder gives the true t_i a ranking in the top three for no combinations. All these 16 combinations were unique in **D**. Figure 16 displays a histogram of the corresponding probabilities associated with the true t_i in each of the 16 combinations. The majority of probabilities are in the 0.005 range. As we assumed a uniform prior distribution over the 182 possibilities in the support, the ratio of the posterior to prior probability is typically around one. Once again, **Z** typically does not provide much additional information about t_i. The largest probability is 0.0083.

For households of size 8, the reduced support $\mathcal{R}_i$ comprises the true t_i plus 206 other combinations of t obtained by changing one variable at a time, as with the other sizes. We do computations for each of the 5 combinations that appear in **D**. We use a uniform prior distribution over all $t \in \mathcal{R}_i$.

The ranks of the true t_i for each of the 5 combinations are $\{52, 39, 84, 57, 67\}$. The intruder gives the true t_i a ranking in the top three for no combinations. We note that all 5 household combinations of size 8 were unique in **D**. The corresponding probabilities associated with the true t_i in each of the 4 combinations are $\{0.0057, 0.0049, 0.0043, 0.0075, 0.0041\}$. As we assumed a uniform prior distribution over the 207 possibilities in the support, the ratio of the posterior to prior probability is typically around one. Once again, **Z** typically does not provide much additional information about t_i. The largest probability is 0.0075.

For households of size 9, the reduced support $\mathcal{R}_i$ comprises the true t_i plus 231 other combinations of t obtained by changing one variable at a time, as with the other sizes. We do computations for each of the 2 combinations that appear in **D**. We use a uniform prior distribution over all $t \in \mathcal{R}_i$.
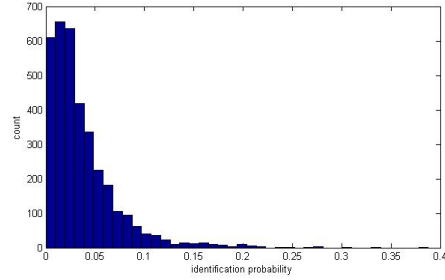
The ranks of the true t_i for each of the 2 combinations are $\{57, 66\}$. We note that both 2 household combinations of size 9 were unique in **D**. The corresponding probabilities associated with the true t_i in each of the 2 combinations are $\{0.0029, 0.0017\}$. As we assumed a uniform prior distribution over the 232

possibilities in the support, the ratio of the posterior to prior probability is less than one. Once again, **Z** typically does not provide much additional information about t_i.

Disclosure risk assessments for structural zeros example

We now turn to illustrating the assessment of disclosure risks for the synthesis with structural zeros, described in Section 4.2 of the main text. For individual disclosure risks, for each individual ij we compute the ρ_{ij}^t for all t in $\mathcal{R}_{ij}$ defined as the union of the true t_{ij} plus the 24 other combinations of t obtained by changing one variable at a time, keeping the relationship variable fixed as a computational convenience. We compute ρ_{ij}^t for each of the 2517 combinations that appear in **D**. We use a uniform prior distribution over all $t \in \mathcal{R}_{ij}$.
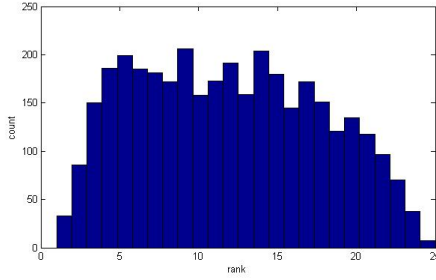
FIG 17. *Histogram of ranks of the probabilities associated with true t_{ij}. Data have structural zeros.*

FIG 18. *Histogram of re-normalized probabilities associated with the true t_{ij}, all individuals. Data have structural zeros.*

Figure 17 displays the distribution of the rank of the true t_{ij} for each of the 3517 combinations. Even armed with $\mathbf{D}_{-ij}$, the intruder gives the top rank to the true t_{ij} for only 33 combinations. The intruder gives the true t_{ij} a ranking in the top three for 269 combinations. We note that 1204 combinations were unique in **D**.

Figure 18 displays a histogram of the corresponding probabilities associated with the true t_i in each of the 3517 combinations. The majority of probabilities are in the 0.03 range. As we assumed a uniform prior distribution over the 25 possibilities in the support, the ratio of the posterior to prior probability is typically only slightly above one. Thus, compared to random guesses over a close neighborhood of the true values, **Z** typically does not provide much additional information about t_{ij}. The largest probability is 0.3878, and only 4 probabilities exceed 0.3, 27 probabilities exceed 0.2, and 183 probabilities exceed 0.1.

We also look at the disclosure risks for households. For households of size 2, the reduced support $\mathcal{R}_i$ consists of the true t_i plus 31 other combinations of t obtained by changing t_i in one variable. We need only do computations

for each of the 4070 combinations that appeared in **D**. We use a uniform prior distribution over all $t \in \mathcal{R}_i$.

Figure 19 displays the distribution of the rank of the true t_i for each of the 4070 combinations. Even armed with $\mathbf{D}_{-i}$, the intruder gives the top rank to the true t_i for no household combination, and gives t_i a ranking in the top three for only 18 combinations. We note that 3485 combinations were unique in **D**.
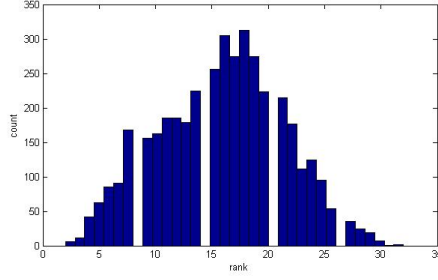

Fig 19. *Histogram of ranks of the probabilities associated with true t_i, for households of size 2. Data have structural zeros.*
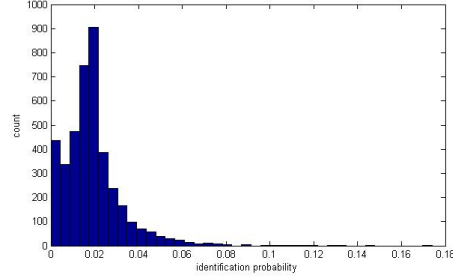
Fig 20. *Histogram of re-normalized probabilities associated with the true t_i, for households of size 2. Data have structural zeros.*

Figure 20 displays a histogram of the corresponding probabilities associated with the true t_i in each of the 4070 combinations. The majority of probabilities are in the 0.025 range. As we assumed a uniform prior distribution over the 32 possibilities in the support, the ratio of the posterior to prior probability is typically one or less. Thus, compared to random guesses over a reasonably close neighborhood of the true values, **Z** typically does not provide much additional information about t_i. The largest probability is 0.1740, and only 15 probabilities exceed 0.1.

For households of size 3, the reduced support $\mathcal{R}_i$ consists of the true t_i plus 46 other combinations of t. We need only do computations for each of the 2492 combinations that appeared in **D**. We use a uniform prior distribution over all $t \in \mathcal{R}_i$.

Figure 21 displays the distribution of the rank of the true t_i for each of the 2492 combinations. Even armed with $\mathbf{D}_{-i}$, the intruder gives the top rank to the true t_i for no combination and gives t_i a ranking in the top three for only 2 combinations. We note that 2480 combinations were unique in **D**.
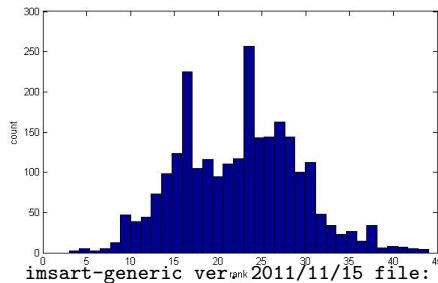
Fig 21. *Histogram of ranks of the probabilities associated with true t_i, for households of size 3. Data have structural zeros.*
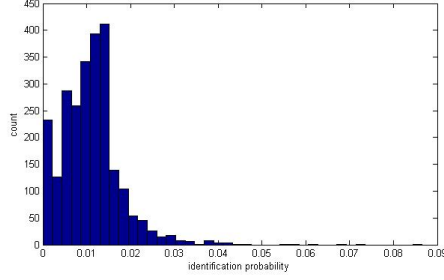
Fig 22. *Histogram of re-normalized probabilities associated with the true t_i, for households of size 3. Data have structural zeros.*
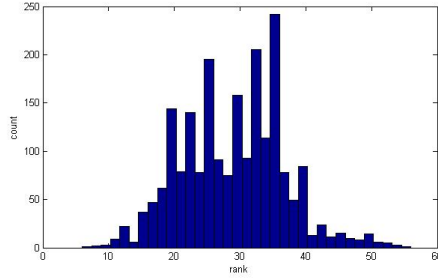

Fig 23. *Histogram of ranks of the probabilities associated with true t_i, for households of size 4. Data have structural zeros.*
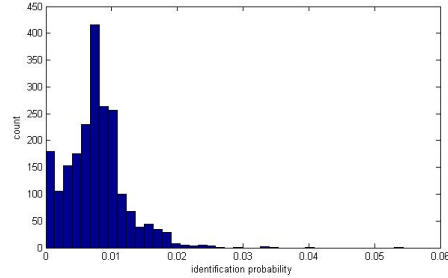
Fig 24. *Histogram of re-normalized probabilities associated with the true t_i, for households of size 4. Data have structural zeros.*

Figure 22 displays a histogram of the corresponding probabilities associated with the true t_i in each of the 2492 combinations. The majority of probabilities are in the 0.01 range. As we assumed a uniform prior distribution over the 47 possibilities in the support, the ratio of the posterior to prior probability is typically less than one. Thus, compared to random guesses over a reasonably close neighborhood of the true values, **Z** typically does not provide much additional information about t_i. The largest probability is 0.0866.

For households of size 4, the reduced support $\mathcal{R}_i$ consists of the true t_i plus 61 other combinations of t. We need only do computations for each of the 2124 combinations that appeared in **D**. We use a uniform prior distribution over all $t \in \mathcal{R}_i$.

Figure 23 displays the distribution of the rank of the true t_i for each of the 2124 combinations. Even armed with $\mathbf{D}_{-i}$, the intruder gives the top rank to the true t_i for no combination and gives t_i a ranking in the top three for no combinations. We note that 2122 combinations were unique in **D**.

Figure 24 displays a histogram of the corresponding probabilities associated

with the true t_i in each of the 2124 combinations. The majority of probabilities are in the 0.01 range. As we assumed a uniform prior distribution over the 62 possibilities in the support, the ratio of the posterior to prior probability is typically less than one. Thus, compared to random guesses over a reasonably close neighborhood of the true values, **Z** typically does not provide much additional information about t_i. The largest probability is 0.0544.