Bayesian Estimation of Discrete Multivariate Latent Structure Models with Structural Zeros

Daniel Manrique-Vallier and Jerome P. Reiter*

Abstract

In multivariate categorical data, models based on conditional independence assumptions, such as latent class models, offer efficient estimation of complex dependencies. However, Bayesian versions of latent structure models for categorical data typically do not appropriately handle impossible combinations of variables, also known as structural zeros. Allowing non-zero probability for impossible combinations results in inaccurate estimates of joint and conditional probabilities, even for feasible combinations. We present an approach for estimating posterior distributions in Bayesian latent structure models with potentially many structural zeros. The basic idea is to treat the observed data as a truncated sample from an augmented dataset, thereby allowing us to exploit the conditional independence assumptions for computational expediency. As part of the approach, we develop an algorithm for collapsing a large set of structural zero combinations into a much smaller set of disjoint marginal conditions, which greatly speeds computation. We apply the approach to sample from a semi-parametric version of the latent class model with structural zeros in the context of a key issue faced by national statistical agencies seeking to disseminate confidential data to the public: estimating the number of records in a sample that are unique in the population on a

*Daniel Manrique-Vallier is Assistant Professor at the Department of Statistics, Indiana University, (e-mail: dmanriqu@indiana.edu); and Jerome P. Reiter is Mrs. Alexander Hehmeyer Professor of Statistical Science, Duke University, Durham, NC 27708-0251, (e-mail: jerry@stat.duke.edu). This research was supported by a grant from the National Science Foundation (SES 1131897).

set of publicly available categorical variables. The latent class model offers remarkably accurate estimates of population uniqueness, even in the presence of a large number of structural zeros.

KEY WORDS: Contingency table; Confidentiality; Disclosure Risk; Latent class; Dirichlet Process; Multinomial.

1 Introduction

For multivariate categorical data $\mathbf{x}$, models based on latent conditional independence assumptions enable analysts to estimate joint and conditional probabilities with computationally efficient algorithms. Such models include, for example, latent class models (Goodman, 1974; Lazarsfeld and Henry, 1968), Rasch models (Rasch, 1980), and Grade of Membership models (Woodbury et al., 1978). These can be viewed as mixture models that use an auxiliary (latent) variable $\mathbf{z}$ to decouple the dependence structure among components of $\mathbf{x}$; that is, given $\mathbf{z}$ the components of $\mathbf{x}$ are assumed independent (Holland and Rosenbaum, 1986). By averaging component-specific probabilities over many components, latent structure models can encode arbitrarily complex dependence structure in $\mathbf{x}$ (Suppes and Zanotti, 1981; Sijtsma and Junker, 2006).

Many categorical datasets include structural zeros (Goodman, 1968; Bishop et al., 1975). These arise when certain combinations of responses are impossible, for example, pregnant men or married children. They also arise when particular combinations of categories are unobservable by design, for example, non-recorded individuals in multiple recapture-experiments (Fienberg, 1972). Fitting latent structure models (LSMs) without explicitly accounting for structural zeros can result in inaccurate estimates of joint and conditional probabilities, even for feasible combinations. This is because allowing non-zero probabilities for impossible combinations can pull probability mass away from feasible combinations. Models and estimation routines that account for structural zeros have been developed for non-Bayesian versions of LSMs (e.g., Vermunt, 1997);

however, to our knowledge, these have not been developed for Bayesian versions of LSMs. We note that structural zeros can be incorporated in other approaches for estimating contingency table probabilities, including methods based on loglinear models (Bishop et al., 1975), Markov bases (e.g., Dobra, 2012), and importance sampling (e.g., Dinwoodie and Chen, 2011).

In this article, we present a general approach and MCMC samplers for Bayesian estimation of latent structure models when the data include structural zeros. The basic idea is to view the data for the feasible combinations as a truncated version of an augmented dataset of unknown size comprising all combinations of $\mathbf{x}$. The MCMC sampler proceeds by iteratively (i) imputing counts to create a completed, augmented dataset and (ii) drawing parameters from the model based on the completed data. Key to our approach is a computational algorithm that collapses many structural zero combinations into much smaller sets of marginal conditions. This algorithm allows us to exploit the conditional independence structure to handle even large numbers of structural zeros in a computationally efficient manner. We apply the truncated latent structure models to estimate the number of records in a sample that are unique in the parent population, which is an important quantity when estimating the risks of re-identification in public use datasets. The truncated latent class models offer estimates that are very close to true values, whereas ignoring the structural zeros results in large bias.

The remainder of the article is organized as follows. In Section 2, we formally define Bayesian truncated latent structure models (TLSMs) for handling structural zeros. In Section 3, we derive an MCMC sampler for estimation of Bayesian TLSMs for modest-sized sets of structural zeros. In Section 4, we outline the computational algorithm for reducing structural zero combinations into sets of marginal conditions, which allows us to extend to large numbers of structural zeros. In Section 5, we present the MCMC sampler for latent class models. In Section 6, we apply a truncated latent class model to estimate population uniqueness in data derived from the year 2000 New York Public

Use Microdata Sample. In Section 7, we conclude with a brief discussion of settings other than disclosure risk estimation in which TLSMs can be applied.

2 Truncated Latent Structure Models

Let $\mathbf{x} = (x_1, x_2, \ldots, x_J)$ be a discrete multivariate response variable, where each component, x_j, can take values from a finite set of L_j levels. We label these levels with consecutive numbers from 1 to L_j. Thus, $\mathbf{x} \in \mathcal{C} = \prod_{j=1}^{J} \{1, \ldots, L_j\}$. Following standard contingency tables terminology, we call each of the values of $\mathbf{x}$ a cell.

Latent structure models rely on the conditional independence assumption,

$$f^{LSM}(\mathbf{x}|\theta) = p(\mathbf{x} \mid \theta) = \int h(\mathbf{z} \mid \theta) \prod_{j=1}^{J} g(x_j \mid \mathbf{z}, \theta) d\mathbf{z}, \tag{1}$$

where $\mathbf{z}$ is a finite-dimensional latent variable, and θ is the vector of parameters specific to the LSM. Here, $g(x_j \mid \mathbf{z}, \theta)$ denotes the probability mass function (pmf) of x_j conditional on θ and $\mathbf{z}$, and $h(\mathbf{z} \mid \theta)$ denotes the density (or pmf) of $\mathbf{z}$ given θ. For example, Rasch models for binary data specify $\mathbf{z} = \alpha \in \mathbb{R}$, $\theta = (\beta_1, \ldots, \beta_J) \in \mathbb{R}^J$, and $g(x_j \mid \mathbf{z}, \theta) = \text{Bernoulli}(x_j \mid 1/(1 + \exp(\alpha + \beta_j)))$. Latent class models set $\mathbf{z} = k \in \{1, 2, \ldots, K\}$, with $g(x_j \mid \mathbf{z}, \theta) = \text{Discrete}_{1:L_j}(\lambda_{jk}[1], \ldots, \lambda_{jk}[L_j])$ and $\sum_{l=1}^{L_j} \lambda_{jk}[l] = 1$. In what follows, we use the notation $p(\cdot)$ to indicate the density or pmf of the argument determined from context, except in cases of potential ambiguity.

When the data include structural zeros, we should restrict the support of $\mathbf{x}$ to an appropriate subset of $\mathcal{C}$. Let $S \subsetneq \mathcal{C}$ be the set of cells to be excluded. Let $\mathbf{x}^*$ be restricted to the set $\mathcal{C} \setminus S$. We define the *truncated latent structure model* (TLSM) by assuming

$$f^{TLSM}(\mathbf{x}^*|\theta) = p(\mathbf{x}^* \mid \theta, \mathrm{T}(S)) \propto 1\{\mathbf{x}^* \notin S\} \int g(\mathbf{x}^* \mid \mathbf{z}, \theta) h(\mathbf{z} \mid \theta) d\mathbf{z}, \tag{2}$$

where $1\{\cdot\}$ equals one when the condition inside the $\{\}$ is true and equals zero otherwise. Here, we use $\mathrm{T}(S)$ to indicate that $\mathbf{x}^*$ is distributed according to the TLSM, not the LSM. In writing (2), we purposefully use the same θ as in (1) to facilitate a computational strategy for incorporating structural zeros; we explain this further in Section 3. Finally, we let $g(\mathbf{x}^* \mid \mathbf{z}, \theta) = \prod_{j=1}^{J} g(x_j^* \mid \mathbf{z}, \theta)$.

Let $\mathcal{X}^* = (\mathbf{x}_1^*, \mathbf{x}_2^*, \ldots, \mathbf{x}_n^*)$ be a sample of n variates generated from the TLSM. We seek to compute the posterior distribution $p(\theta \mid \mathcal{X}^*, \mathrm{T}(S))$, where $\mathcal{X}^* = (\mathbf{x}_1^*, \mathbf{x}_2^*, \ldots, \mathbf{x}_n^*)$. For many mixture models—including but not limited to LSMs—there exist efficient algorithms based on data augmentation (Tanner, 1996): given a sample $\mathcal{X}$, draw samples from the joint posterior $p(\theta, \mathcal{Z} \mid \mathcal{X})$, where $\mathcal{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$, and obtain the desired posterior through marginalization. Examples of these algorithms include Patz and Junker (1999), Ishwaran and James (2001), and Erosheva et al. (2007), among others. Unfortunately, these algorithms do not directly transfer to the truncated case in (2). Instead of $\prod_{i=1}^{n} f^{LSM}(\mathbf{x}^* | \theta)$, the joint distribution of the iid sample $\mathcal{X}^*$ is

$$p(\mathcal{X}^* \mid \theta, \mathrm{T}(S)) = \prod_{i=1}^{n} f^{TLSM}(\mathbf{x}_i^* | \theta) = \prod_{i=1}^{n} \frac{\int g(\mathbf{x}_i^* \mid \mathbf{z}, \theta) h(\mathbf{z} \mid \theta) d\mathbf{z}}{\sum_{\mathbf{x} \notin S} \int g(\mathbf{x} \mid \mathbf{z}, \theta) h(\mathbf{z} \mid \theta) d\mathbf{z}} 1\{\mathbf{x}_i^* \notin S\}. \quad (3)$$

Therefore, with a prior distribution $p(\theta)$, we have

$$p(\theta \mid \mathcal{X}^*, \mathrm{T}(S)) \propto \frac{p(\theta)}{(1 - \pi_0(\theta))^n} \prod_{i=1}^{n} 1\{\mathbf{x}_i^* \notin S\} \int g(\mathbf{x}_i^* \mid \mathbf{z}, \theta) h(\mathbf{z} \mid \theta) d\mathbf{z} \quad (4)$$

where $\pi_0(\theta) = \Pr(\mathbf{x} \in S \mid \theta) = \sum_{\mathbf{x} \in S} f^{LSM}(\mathbf{x} | \theta)$. Because of the presence of $(1 - \pi_0(\theta))^n$ in the denominator of (4), existing methods for estimating the posterior distribution of θ based on (1) are no longer applicable.

3 A Sample Augmentation Strategy for TLSMs

To simulate the posterior distribution in (4) we use a sample augmentation strategy. Heuristically, the idea is to "complete" the sample with a number of units, n_0, in a way that allows us to apply methods that ignore the truncation—specifically through exploitation of the conditional independence structure—while yielding the same posterior distribution of parameters that we would have obtained under the truncated model. To do this we assume that $\mathcal{X}^*$ was realized through a two-step process: generate an iid sample of N realizations from the model in (1), and truncate the sample by throwing away all the data points that fall in S. In doing so, we treat N as an unknown parameter and n, the number of retained samples, as observed data. This strategy is related to other approaches for handling truncated data; see, for example, Gelman et al. (2004, p. 235), Meng and Zaslavsky (2002), and O'Malley and Zaslavsky (2008).

The sample augmentation strategy actually involves two distinct models: a computationally convenient LSM and the target TLSM. Because of this, it is not obvious in general that the sample augmentation strategy offers appropriate posterior distributions for the TLSM parameters. However, as we prove in Theorem 1 below, if we set $p(N) \propto 1/N$ and use the sample augmentation algorithm, the posterior distribution of parameters conditional on the realizations that we did not "throw away" matches the posterior distribution of the TLSM parameters conditional on the observed data.

Formally, let $\mathcal{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$ be a hypothetical complete data sample generated from the LSM process:

$$\mathbf{x}_i \overset{indep}{\sim} g(\mathbf{x}_i \mid \mathbf{z}_i, \theta, N), \quad \text{where } i = 1, \ldots, N \tag{5}$$

$$\mathbf{z}_i \overset{iid}{\sim} h(\mathbf{z}_i \mid \theta, N), \text{where} \quad i = 1, \ldots, N \tag{6}$$

$$\theta \sim p(\theta). \tag{7}$$

We view $\mathcal{X}$ as composed of two pieces: an "observable" part, $\mathcal{X}^1 = (\mathbf{x}_1^1, \ldots, \mathbf{x}_n^1)$,

comprising the n samples that did not fall into S and following some arbitrary sequence (e.g., their order of appearance in $\mathcal{X}$); and an "unobservable" part, $\mathcal{X}^0 = (\mathbf{x}_1^0, \ldots, \mathbf{x}_{n_0}^0)$, comprising the remaining $n_0 = N - n$ observations and also following an arbitrary sequence. Every $\mathbf{x}_i$ in $\mathcal{X}$ has an associated $\mathbf{z}_i$. We collect all $\mathbf{z}_i$ corresponding to cases in $\mathcal{X}^1$ in the vector $\mathcal{Z}^1 = (\mathbf{z}_1^1, \mathbf{z}_2^1, \ldots, \mathbf{z}_n^1)$, with labels matching those in $\mathcal{X}^1$. We define similarly $\mathcal{Z}^0 = (\mathbf{z}_1^0, ..., \mathbf{z}_{n_0}^0)$ as all $\mathbf{z}_i$s corresponding to $\mathcal{X}^0$. Note that according to this definition the length of $\mathcal{X}^1$ is a random quantity.

As the prior distribution for the unknown N, we set $p(N) \propto 1/N$. Although this $p(N)$ is improper, using it allows us to state Theorem 1, which is proved in the Appendix.

Theorem 1. *Let $\mathcal{X}^*$ comprise n iid samples from the TLSM in (2). Let $\mathcal{X}^1$ be generated from the LSM in (5) – (7) so that no element of $\mathcal{X}^1 \in S$. Assume that $\mathcal{X}^* = \mathcal{X}^1$. Let $p(N) \propto 1/N$ be the prior distribution of N, independent of the prior for θ, and $n_0 = N - n$. Then,*

$$p(\theta \mid \mathcal{X}^*, T(S)) = \int p(\theta, \mathcal{Z}^1, \mathcal{Z}^0, \mathcal{X}^0, n_0 \mid \mathcal{X}^1) d\mathcal{X}^0 d\mathcal{Z}^1 d\mathcal{Z}^0 dn_0. \tag{8}$$

Setting $p(N) \propto 1/N$ is necessary for the theorem to hold, as is implicit in the proof of the theorem.

In practical terms, Theorem 1 shows that given a dataset $\mathcal{X}^*$ containing n iid samples from the TLSM in (2), we can obtain samples from $p(\theta \mid \mathcal{X}^*, n, \mathrm{T}(S))$ from a sampler for $p(\theta, \mathcal{Z}^1, \mathcal{Z}^0, \mathcal{X}^0, n_0 \mid \mathcal{X}^*)$. Since $p(\theta, \mathcal{Z}^1, \mathcal{Z}^0, \mathcal{X}^0, n_0 \mid \mathcal{X}^1, n)$ involves the LSM, this theorem allows us to exploit the conditional independence structure of the LSM to derive an efficient Gibbs sampler.

Specifically, a Gibbs sampler for sampling $p(\theta, \mathcal{Z}^1, \mathcal{Z}^0, \mathcal{X}^0, n_0 \mid \mathcal{X}^1, n)$ for an arbitrary exclusion set $S = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_C\} \subsetneq \mathcal{C}$ can be constructed as follows.

1. Sample $\mathbf{z}_i^1$ for $i = 1, \ldots, n$ from its full conditional distribution,

$$p(\mathbf{z}_i^1 \mid \ldots) \propto h(\mathbf{z}_i^1 \mid \theta) g(\mathbf{x}_i^1 \mid \mathbf{z}_i^1, \theta). \tag{9}$$

2. Sample θ from its full conditional distribution,

$$p(\theta \mid \ldots) \propto p(\theta) \prod_{i=1 \ldots n_0} g(\mathbf{x}_i^0 \mid \mathbf{z}_i^0, \theta) h(\mathbf{z}_i^0 \mid \theta) \prod_{i=1 \ldots n} g(\mathbf{x}_i^1 \mid \mathbf{z}_i^1, \theta) h(\mathbf{z}_i^1 \mid \theta). \tag{10}$$

3. Sample $(\mathcal{X}^0, \mathcal{Z}^0, n_0)$ from their full joint conditional distribution based on the factorization

$$p(\mathcal{X}^0, \mathcal{Z}^0, n_0 \mid \ldots) = p(n_0 \mid \mathcal{X}^1, \theta) p(\mathcal{X}^0 \mid n_0, \mathcal{X}^1, \theta) p(\mathcal{Z}^0 \mid n_0, \mathcal{X}^0, \mathcal{X}^1, \theta). \tag{11}$$

This factorization avoids the full conditional of n_0, which includes $\mathcal{X}^0$ and $\mathcal{Z}^0$ in the conditioning. Since the length of $\mathcal{Z}^0$ and $\mathcal{X}^0$ is exactly n_0, we cannot sample n_0 conditional on $(\mathcal{X}^0, \mathcal{Z}^0)$ without making the chain reducible. This issue was identified by Basu and Ebrahimi (2001), who used a similar factorization in the context of multiple-recapture estimation. Additionally, all the information needed from $\mathcal{X}^1$ in (11) is its length, n. Thus we can replace all references to $\mathcal{X}^1$ in (11) by n.

To facilitate sampling from (11), we define $(\mathbf{s}_1, ..., \mathbf{s}_C)$ to be an enumeration of the elements of S and $\mathbf{n} = (n_1, n_2, \ldots, n_C)$ their respective counts, i.e., $n_c = \#\{\mathbf{x} \in \mathcal{X}^0 : \mathbf{x} = \mathbf{s}_c\}$ for $c = 1, 2, \ldots, C$. Note that $\sum_{c=1}^{C} n_c = n_0$. Factoring $p(\mathcal{Z}^0, \mathbf{n} \mid \ldots) \propto p(\mathbf{n} \mid \theta, n) p(\mathcal{Z}^0 \mid \mathbf{n}, \theta, n)$, the partial conditional distribution of $\mathbf{n}$ (integrating out $\mathcal{Z}^0$) is

$$p(\mathbf{n} \mid \theta, n) \propto p(N) \frac{(n + n_0)!}{n! n_1! \ldots n_C!} p_d^n \prod_{c=1}^{C} p_c^{n_c}, \tag{12}$$

where $p_d = \Pr(\mathbf{x} \notin S \mid \theta)$ and $p_c = \Pr(\mathbf{x} = \mathbf{s}_c \mid \theta)$, for $c = 1, \ldots, C$. Replacing

$p(N) \propto 1/N = 1/(n + n_0)$, we have

$$p(\mathbf{n} \mid \theta, n) = \frac{\Gamma(n + n_0)}{\Gamma(n) \prod_{c=1}^{C} n_c!} p_d^n \prod_{c=1}^{C} p_c^{n_c} = \mathrm{NM}(\mathbf{n} \mid n, p_1, \ldots, p_C). \tag{13}$$

This resulting distribution is negative multinomial (NM), a multivariate generalization of the negative binomial distribution. NM-distributed variates can be obtained from compounding a multinomial distribution by a negative binomial or from the composition of independent Poisson variates by a Gamma distribution (Sibuya et al., 1964). Thus, we can sample from the full conditional distribution of $(n_0, \mathcal{Z}^0, \mathcal{X}^0)$ in three steps.

(a) Sample $\mathbf{n} \sim NM(n, p_1, \ldots, p_C)$. Make $n_0 = \sum_{c=1}^{C} n_c$.

(b) Sample the n_0 elements of $\mathcal{Z}^0$ in C steps. On each step sample a total of n_c iid variates $\mathbf{z}_{ci}$ from $p(\mathbf{z}_{ci} \mid \theta, n_c) \propto h(\mathbf{z}_{ci} \mid \theta) g(\mathbf{s}_c \mid \mathbf{z}_{ci}, \theta)$. Form the vector $\mathcal{Z}^0$ concatenating all n_0 of them.

(c) Similarly, form $\mathcal{X}^0$ by concatenating the n_0 elements

$$\mathcal{X}^0 = (\underbrace{\mathbf{s}_1, \ldots, \mathbf{s}_1}_{n_1 \text{times}}, \ldots, \underbrace{\mathbf{s}_C, \ldots, \mathbf{s}_C}_{n_C \text{times}}). \tag{14}$$

Of course, for any particular TLSM this algorithm requires being able to implement samplers for the full conditionals of $\mathbf{z}, \theta$, and $\mathbf{x}$. Some of these steps may not be possible for particular TLSMs. Fortunately, they are feasible for several common latent structure models, including IRT models (Patz and Junker, 1999), grade of membership models (Erosheva et al., 2007), truncated Dirichlet mixtures of multinomials (Ishwaran and James, 2001), and latent class models, the last of which we present in Section 5.

As with any truncated model, the draws of θ can be somewhat complicated to interpret. They are at the same time mixture model parameters for a hypothetical, augmented sample and the parameters of the TLSM (3). However, we can use the

draws of θ to simulate readily interpretable quantities. For example, to obtain posterior predictive distributions of cell probabilities in the contingency table, one can sample from f^{TLSM} according to (2). In Section 6, we use posterior predictive simulation in assessments of disclosure risks.

Since LSMs and TLSMs are different models, the interpretation of θ in the context of a particular LSM does not necessarily carry to its corresponding TLSM. This means that $p(\theta)$ for TLSMs should be specified with the truncated model in mind rather than the augmented LSM. Specifying informative prior distributions for truncated mixture models can be a challenging task due to the complexity of the models. In our experience, using diffuse prior distributions often applied in LSMs by default allows the likelihood to dominate the prior distribution in TLSMs.

4 Truncation Specified by Marginal Conditions

A potential limitation of the algorithm in Section 3 is the need to compute $\Pr(\mathbf{x} = \mathbf{s}_c \mid \theta)$ for all $\mathbf{s}_c \in S$ under the untruncated model. Even if efficient and fast algorithms for computing these individual probabilities were available, when the size of the set S is large—e.g., more than a few tens of thousands—this computational requirement can render this approach prohibitively expensive.

Situations in which the set S is extremely large are frequent. For instance, in a demographic dataset one might want to exclude all the cells that include any combination of $Sex = Male$ and $Pregnant = Yes$. If the dataset comprises several other variables, this condition alone could specify an extremely large number of cells. We call conditions derived from fixing the levels of a subset of the response components *marginal conditions*.

A convenient feature of discrete multivariate models based on conditional independence (given latent variables) is that computing marginal probabilities is simple. Let $\boldsymbol{\mu} = \{\mathbf{x} : x_{j_1} = a_{j_1}, \ldots, x_{j_M} = a_{j_M}\}$ where $j_1, j_2, \ldots, j_M$ is a subsequence of $1, 2, \ldots, J$.

This is a set defined by a marginal condition. Then, we have

$$\Pr(\mathbf{x} \in \boldsymbol{\mu} \mid \theta) = \sum_{\mathbf{x}' : \mathbf{x}' \in \boldsymbol{\mu}} \int h(\mathbf{z} \mid \theta) \prod_{j=1}^{J} Pr(x_j = x'_j \mid \mathbf{z}, \theta) d\mathbf{z} \tag{15}$$

$$= \int h(\mathbf{z} \mid \theta) \prod_{m=1}^{M} \Pr(x_{j_m} = a_{j_m} \mid \mathbf{z}, \theta) d\mathbf{z}. \tag{16}$$

Note that while the outermost sum in (15) involves $\#\boldsymbol{\mu}$ terms, the expression in (16) involves only one term.

We can adapt the data augmentation algorithm in Steps 1 – 3 to take advantage of this property in situations where we have the exclusion set S defined as the union of a collection of sets specified by marginal conditions. To facilitate this development, we introduce a special notation for marginal conditions. We denote marginal conditions through J-dimensional vectors $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_J)$ with components taking values on the enlarged set, $\mu_j \in \{1, 2, \ldots, L_j\} \cup \{*\}$. We interpret the vector $\boldsymbol{\mu}$ as the marginal conditions that define the set of cells $\{\mathbf{x} \in \mathcal{C} : x_j = \mu_j, \text{ if } \mu_j \neq *\}$. In a slight abuse of notation, we use the symbol $\boldsymbol{\mu}$ to represent both the marginal conditions and the set of cells represented by those conditions. For instance, $\boldsymbol{\mu} = (*, *, 1, 2, *, *, *, *)$ represents the $L_1 \times L_2 \times L_5 \times L_6 \times L_7 \times L_8$ cells such that $x_3 = 1$ and $x_4 = 2$. Thus, $p(\boldsymbol{\mu} \mid \theta) = \Pr(\mathbf{x} \in \boldsymbol{\mu} \mid \theta) = \Pr(x_3 = 1, x_4 = 2 \mid \theta)$. We call the entries with the symbol * *placeholder components* and the others *fixed components*. In the example, $\boldsymbol{\mu}$ has six placeholder components (1, 2, 5, 6, 7, and 8) and two fixed components (3 and 4). This notation enables easy evaluation of the intersection of two equal-dimensional marginal conditions, $\boldsymbol{\mu}^A = (\mu_j^A)$ and $\boldsymbol{\mu}^B = (\mu_j^B)$, as either the empty set if $\mu_j^A, \mu_j^B \neq *$ and $\mu_j^A \neq \mu_j^B$ for some $j = 1, \ldots, J$, or as the marginal condition $\boldsymbol{\mu}^{A \cap B} = (\mu_j^{A \cap B})$ where

$$\mu_j^{A \cap B} = \begin{cases} \mu_j^B & \text{if } \mu_j^A = * \text{ and } \mu_j^B \neq * \\ \mu_j^A & \text{if } \mu_j^A \neq * \text{ and } \mu_j^B = * \\ \mu_j^A & \text{if } \mu_j^A = \mu_j^B. \end{cases} \tag{17}$$

4.1 Case 1: Non-Overlapping Marginal Conditions

The simplest situation of exclusion sets defined by marginal conditions involves collections of non-overlapping conditions. In practice, such situations involve structural zeros specified by the union of distinct marginal conditions involving the same variables at different levels. For example, $\boldsymbol{\mu}_1 = (*, 2, 1, *, 1)$ and $\boldsymbol{\mu}_2 = (*, 3, *, *, 1)$ specify disjoint sets because they differ on the second coordinate. However, as we will see Section 4.2, the real importance of this case lies in the fact that general patterns of structural zeros defined by marginal conditions always can be reduced to non-overlapping marginal conditions.

Assume that S is defined to be the union of C' mutually exclusive sets defined by marginal conditions $S = \bigcup_{c'=1}^{C'} \boldsymbol{\mu}_{c'}$, with $\boldsymbol{\mu}_{c'_1} \cap \boldsymbol{\mu}_{c'_2} = \emptyset$ for $c'_1 \neq c'_2$. Here, $c' = 1, \ldots, C'$ indexes marginal conditions, not individual cells as before. We adapt the basic truncation-augmentation algorithm of Section 3 by redefining the vector $\mathbf{n} = (n_{c'})$ where $n_{c'} = \#\{\mathbf{x} \in \mathcal{X} : \mathbf{x} \in \boldsymbol{\mu}_{c'}\}$, and replace steps 3.a, 3.b, and 3.c as follows.

3.a' Sample $\mathbf{n} \sim NM(n; p_1, \ldots, p_{C'})$, where $p_{c'} = \Pr(\mathbf{x} \in \boldsymbol{\mu}_{c'} \mid \theta)$ for $c' = 1, \ldots, C'$.

3.b' For each $c' = 1, \ldots, C'$, sample $n_{c'}$ independent variates

$$p(\mathbf{z}_{c'i} \mid \theta, \boldsymbol{\mu}_{c'}) \propto h(\mathbf{z}_{c'i} \mid \theta) \sum_{\mathbf{x} \in \boldsymbol{\mu}_{c'}} g(\mathbf{x} \mid \mathbf{z}_{c'i}, \theta) = h(\mathbf{z}_{c'i} \mid \theta) \prod_{\{j:\mu_j \neq *\}} \Pr(x_j = \mu_j \mid \mathbf{z}_{c'i}, \theta), \tag{18}$$

and use each of them to sample $p(\mathbf{x}_{c'i} \mid \mathbf{z}_{c'i}, \boldsymbol{\mu}_{c'}, \theta) \propto g(\mathbf{x}_{c'i} \mid \mathbf{z}_{c'i}, \theta) 1\{\mathbf{x}_{c'i} \in \boldsymbol{\mu}_{c'}\}$. Sampling from arbitrarily truncated discrete distributions when the size of the truncated region is large or has high probability is often a major computational challenge (e.g., Dobra, 2012). However, the special structure of both $f^{LSM}(\mathbf{x} \mid \theta)$ and the regions defined by $\boldsymbol{\mu}'_c$ offer an efficient component-wise sampling strategy. For $j = 1 \ldots J$, sample $x_j \sim p(x_j \mid \mathbf{z}, \theta)$ if $\mu_j = *$; otherwise, make $x_j = \mu_j$.

3.c' Construct $\mathcal{Z}^0$ and $\mathcal{X}^0$ by concatenating all $\mathbf{z}_{c'i}$ and $\mathbf{x}_{c'i}$ in the same order.

A specific instantiation of these three steps is steps 4 - 6 of the MCMC sampler for

latent class models presented in Section 5.

4.2 Case 2: General Marginal Conditions

The situation is more complicated when we specify S through combinations of different variables, so that the marginal conditions define overlapping regions of $\mathcal{C}$. For example, the two margin conditions $(1, 2, *, *, *, *)$ and $(1, *, 3, *, *, *)$ define overlapping sets, as any cell with $(x_1 = 1, x_2 = 2, x_3 = 3)$ satisfies both conditions. Our approach is to find a representation of S in terms of non-overlapping marginal conditions so that we can apply the techniques from Section 4.1.

In principle, it is trivial to find such an equivalent representation: expand every marginal condition into the cells it represents, and eliminate the duplicates. Unfortunately, this solution can be extremely inefficient when S is large. A better solution can be devised by noting that any single marginal condition $\boldsymbol{\mu}$ that represents more than one cell, i.e., has one or more placeholder components, can be expanded into sets of several smaller non-overlapping marginal conditions that represent the same set of cells but with more fixed components. To do this, we select a set of placeholder components from the original $\boldsymbol{\mu}$ and expand all combination of their levels. For each of the expanded combinations, we create a new marginal condition that has the original fixed levels of $\boldsymbol{\mu}$ plus the expanded combination. To illustrate, let $\boldsymbol{\mu} = (*, *, 1, 2, *, *, *)$ with $L_6 = 2$ and $L_7 = 2$. The *expansion* of $\boldsymbol{\mu}$ with respect to components 6 and 7 is

$$\begin{aligned}\text{Expan}(\boldsymbol{\mu}, \{6, 7\}) = \{&(*, *, 1, 2, *, 1, 1), (*, *, 1, 2, *, 1, 2),\\ &(*, *, 1, 2, *, 2, 1), (*, *, 1, 2, *, 2, 2)\}.\end{aligned}$$

It is easy to see that both $\boldsymbol{\mu}$ and the union of the marginal conditions in $\text{Expan}(\boldsymbol{\mu}, \{6, 7\})$ represent the same set of cells.

Using this expansion operation, we develop an algorithm to transform a set of marginal conditions into a set of disjoint marginal conditions that represent the same

set of cells. For any pair of marginal conditions, we identify and remove the region in which the two conditions overlap by expanding the placeholder components from the second that are fixed components in the first. Consider, for example, the two overlapping conditions $\boldsymbol{\mu}_1 = (1, 1, *, *)$ and $\boldsymbol{\mu}_2 = (1, *, 2, *)$ with $L_2 = 2$. Expanding $\boldsymbol{\mu}_2$ with respect to its second component (which is a placeholder in $\boldsymbol{\mu}_2$ but fixed in $\boldsymbol{\mu}_1$), we see that $\boldsymbol{\mu}_2$ is equivalent to the union of $(1, 1, 2, *)$ and $(1, 2, 2, *)$. Since $(1, 2, 2, *) \in \boldsymbol{\mu}_1$, we remove it from consideration, yielding the equivalent representation of $\boldsymbol{\mu}_1 \cup \boldsymbol{\mu}_2$ as the disjoint union $(1, 1, *, *) \cup (1, 1, 2, *)$.

Generalizing to an arbitrary list of margin conditions, S_u, the following algorithm transforms S_u into a collection of disjoint marginal conditions, S_d, that represent the same collection of cells.

1. LET *Pending* be a list containing all the marginal conditions in S_u sorted in decreasing order according to the number of cells they represent. $Pending[1]$ is the marginal condition that represents the largest number of cells.
2. INITIALIZE $S_d \leftarrow \{Pending[1]\}$. REMOVE $Pending[1]$ from *Pending*.
3. WHILE *Pending* is not empty DO:
   - LET $\boldsymbol{\mu} \leftarrow Pending[1]$. Remove $\boldsymbol{\mu}$ from *Pending*.
   - LET $ComparList \leftarrow$ {all elements of *Pending* that are not disjoint with $\boldsymbol{\mu}$}.
   - LET $Cols \leftarrow$ {Indexes of all fixed components of elements of ComparList that are placeholder components on $\boldsymbol{\mu}$}.
   - IF (*ComparList* or *Cols* are empty) THEN
     - LET $S_d \leftarrow S_d \cup \{\boldsymbol{\mu}\}$
   - ELSE
     - LET $S_d \leftarrow S_d \cup$ {all elements of $Expan(\boldsymbol{\mu}, Cols)$ that are disjoint with every element of *ComparList*}.

We then can insert the resulting set of disjoint marginal conditions into the sampler

in Section 4.1 to obtain samples from the posterior distribution of θ. The algorithm always reduces the number of conditions compared to the implied total number of structural-zero cells, thus always improving computational running times.

This algorithm is a pre-processing step that takes place only once, before running the MCMC sampler. Thus, in the context of sampling from the posterior distribution of θ, its execution generally does not significantly impact computing costs. We note, however, that execution of the MCMC sampler can be computationally expensive when the resulting number of marginal conditions is large (e.g., more than 100,000 with typical current computing power).

5 Bayesian Latent Class Model with Structural Zeros

To illustrate these methods, in this section we develop an MCMC algorithm for obtaining samples from the posterior distribution of parameters from a specific TLSM, the finite mixture of product-multinomial distributions. This is expressed as

$$p(\mathbf{x} \mid \boldsymbol{\lambda}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} \lambda_{jk}[x_j], \tag{19}$$

where $\boldsymbol{\lambda} = (\lambda_{jk}[l])$, with all $\lambda_{jk}[l] > 0$ and $\sum_{l=1}^{L_j} \lambda_{jk}[l] = 1$. Here, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ with $\sum_{k=1}^{K} \pi_k = 1$. This is known as a latent class model (Goodman, 1974; Lazarsfeld and Henry, 1968). These models are often used to discover and characterize latent subpopulations based on observable discrete multivariate characteristics $\mathbf{x}$ (Clogg, 1995). They also can be used as a general-purpose contingency table smoothing tool, as well as engines for imputation of missing data (Vermunt et al., 2008; Gebregziabher and DeSantis, 2010; Si and Reiter, 2013). We only consider models for fixed K, as letting $K \to \infty$ as in Dunson and Xing (2009) results in break-downs of the algorithm in

Section 3. In particular, the infinite-dimensional structure of the mixture prevents us from directly sampling $\mathcal{Z}$ in steps 1 and 3.b of the algorithm.

We use the following prior distributions.

$$\lambda_{jk}[\cdot] \overset{indep}{\sim} \text{Dirichlet}(\mathbf{1}_{L_j}) \tag{20}$$

$$\boldsymbol{\pi} \sim SB_K(\alpha) \tag{21}$$

$$\alpha \sim \text{Gamma}(a, b). \tag{22}$$

Here, $SB_K(\alpha)$ is a K-dimensional finite stick breaking process (Sethuraman, 1994; Ishwaran and James, 2001). Let $V_K = 1$ and $V_k \sim \text{Beta}(1, \alpha)$ for $k = 1, ..., K-1$. We say that $\boldsymbol{\pi} \sim SB_K(\alpha)$ if $\pi_k = V_k \prod_{h<k}(1 - V_k)$. This prior distribution on $\boldsymbol{\pi}$ induces sparsity, thereby reducing computation and avoiding over-fitting for large K. As α decreases, the probability that all z_is take values from a proper subset of $\{1, \ldots, K\}$ increases. Thus, effectively the prior distribution for α selects reasonable values of K, while simultaneously accounting for the uncertainty associated with it. This Bayesian version of the LCM, without structural zeros, is essentially a truncated version of the infinite mixture of product-multinomial distributions developed by Dunson and Xing (2009).

Let $\mathcal{X}^1 = (\mathbf{x}_1^1, \ldots, \mathbf{x}_n^1)$ be a sample from the truncated latent class model,

$$p(\mathbf{x}_i^1 \mid \boldsymbol{\lambda}, \boldsymbol{\pi}, \mathrm{T}(S)) \propto 1\{\mathbf{x}_i^1 \notin S\} \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} \lambda_{jk}[x_j], \tag{23}$$

where $S \subsetneq \mathcal{C}$ is the set of cells to exclude from the support of the distribution. Let $S_d = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{C'}\}$ be the collection of disjoint marginal conditions such that $S = \cup_{c'=1}^{C'} \boldsymbol{\mu}_{c'}$, obtained by applying the algorithm in Section 4.2. Applying Section 4.1, we obtain samples from the posterior distribution of $(\boldsymbol{\lambda}, \boldsymbol{\pi}, \alpha)$ under model (23) as follows.

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

$\xi_{jkl} = 1 + \sum_{i=1}^{n} 1\{x_{ij}^1 = l, z_i^1 = k\} + \sum_{i=1}^{n_0} 1\{x_{ij}^0 = l, z_i^0 = k\}$.

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

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

5. Sample $(n_1, \ldots, n_{C'}) \sim NM(n, \omega_1, \ldots, \omega_{C'})$, and let $n_0 = \sum_{c'=1}^{C} n_{c'}$.

6. Let $\kappa \leftarrow 1$. Repeat the following for each $c' = 1, \ldots, C'$.

   (a) Compute the normalized vector $(p_1, \ldots, p_K)$, where $p_k \propto \pi_k \prod_{j:\mu_{c'j} \neq *} \lambda_{jk}[\mu_{c'j}]$.

   (b) Repeat the following three steps $n_{c'}$ times:

      i. Sample $z_\kappa^0 \sim \text{Discrete}(p_1, \ldots, p_k)$,

      ii. For $j = 1, \ldots, J$ sample

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

      where $\delta_{\mu_{c'j}}$ is a point mass distribution at $\mu_{c'j}$,

      iii. Let $\kappa \leftarrow \kappa + 1$.

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

As discussed in Section 3, this MCMC algorithm produces joint samples from the posterior distribution $p(\boldsymbol{\lambda}, \boldsymbol{\pi}, \alpha, \mathcal{Z}^1, \mathcal{Z}^0, \mathcal{X}^0, n_0 \mid \mathcal{X}^1)$ under the complete data model, from which we can obtain samples from $p(\boldsymbol{\lambda}, \boldsymbol{\pi}, \alpha \mid \mathcal{X}^1, S)$ after marginalizing $(\mathcal{Z}^1, \mathcal{Z}^0, \mathcal{X}^0, n_0)$.

6 Application: Estimating Disclosure Risks

When sharing data with the public, statistical agencies are ethically and often legally obligated to protect the confidentiality of data subjects' identities. Removing direct

identifiers like names and addresses may not suffice to protect confidentiality. For example, ill-intentioned users may be able to link records in the released data to other databases (that include direct identifiers) by matching on variables common to the two databases. Agencies are particularly concerned about individuals with combinations of publicly available variables, called *keys*, that appear only once in the sample. When such combinations also are unique in the population, attackers who find matches for these keys are guaranteed to make re-identifications (assuming no errors in the matching process or data sources).

Therefore, as part of their disclosure risk assessments, most statistical agencies seek to measure the number of records with keys that are unique in the sample that are also unique on those keys in the population; see, for example, Skinner and Shlomo (2008), Manrique-Vallier and Reiter (2012), and the references therein. Typically, the keys include multiple discrete variables such as race, sex, marital status, age (integer or interval-reported), tenure (own home or not), and number of people in household. In demographic data, keys typically contain many structural zeros. These often result from impossible pairwise combinations; for example, in the U.S., datasets should not contain individuals under age 15 years who are married.

In practical contexts, the number of population uniques in a sample is not known by the agency, since it does not observe the entire population. Hence, the agency must estimate this number from the sample. Typical estimation approaches are based on cell probabilities in the table of keys, estimated with log-linear models (Skinner and Shlomo, 2008). However, log-linear models can yield biased estimates of cell probabilities for sparse contingency tables with many (random) zero counts. This bias can result in unreliable estimates of probabilities of uniqueness and, hence, misrepresentations of disclosure risks. These shortcomings were illustrated by Manrique-Vallier and Reiter (2012), who proposed instead to estimate the cell probabilities using Bayesian grade of membership models. These models offer the potential for improved accuracy over state-of-the-art log-linear models; however, Manrique-Vallier and Reiter (2012) do not

handle structural zeros, and their estimation routines are computationally expensive.

We propose to use the truncated latent class model from Section 5 to estimate the number of population uniques in a sample on a set of discrete keys with structural zeros. In addition to handling the structural zeros, estimation with this TLSM is orders of magnitude faster than estimation with the Bayesian grade of membership model. To our knowledge, this represents the first time a Bayesian truncated latent structure model has been applied for this disclosure risk estimation context.

6.1 General Framework

We use a framework similar to that of Manrique-Vallier and Reiter (2012), who use methods from Bayesian finite population inference to estimate the number of population uniques. Suppose that an agency observes a simple random sample, $\mathcal{X}$, of n records collected from a finite population of size H. Let $\mathbf{F_x}$ be the number of times that pattern $\mathbf{x}$ appears in the population, and let $\mathbf{f_x}$ be the corresponding quantity for the sample. Hence, $\sum_{\mathbf{x}\in\mathcal{C}} \mathbf{f_x} = n$ and $\sum_{\mathbf{x}\in\mathcal{C}} \mathbf{F_x} = H$. The agency seeks to estimate the number of elements that are unique in both the sample and the population,

$$\tau = \tau(\mathbf{F}, \mathbf{f}) = \sum_{\mathbf{x}\in\mathcal{C}} 1\{\mathbf{f_x} = 1, \mathbf{F_x} = 1\}. \tag{24}$$

We assume that the (unobserved) population is an iid sample of H elements generated from a super-population model defined by the truncated latent class model in (23). Thus, given observed data $\mathcal{X}^*$ and the set of structural zeros S, we seek to estimate the posterior distribution of τ,

$$p(\tau \mid \mathcal{X}^*, \mathrm{T}(S)) = \int p(\tau(\mathbf{F}, \mathbf{f}) \mid \mathcal{X}^*, \theta, \mathrm{T}(S)) p(\theta \mid \mathcal{X}^*, \mathrm{T}(S)) d(\mathbf{F}, \theta). \tag{25}$$

This expression involves integration over the parameter space and over all possible populations of size H, which is analytically intractable. However, given a sample

from the posterior distribution of parameters, $\theta^{(m)}$, we can obtain a sample from the posterior distribution of τ, $\tau^{(m)}$, as follows.

1. Let U be the number of sample uniques, i.e., cells for which $\mathbf{f_x} = 1$, and let $(\mathbf{x}_1, \ldots, \mathbf{x}_U)$ be the vector containing the cells corresponding to the sample uniques, labelled according to some arbitrary order. Compute $(p_1, \ldots, p_U)$, where $p_i \propto \Pr(\mathbf{x} = \mathbf{x}_i \mid \theta^{(m)}, \mathrm{T}(S))$, for $i = 1, \ldots, U$.
2. Sample $(H_1, \ldots, H_U) \sim \text{Multinomial}(H - n, p_1, \ldots, p_U)$
3. Let $\tau^{(m)} = \sum_{i=1}^{U} 1\{H_i = 0\}$

This algorithm generates synthetic population counts for all the cells corresponding to sample uniques and as a byproduct offers the number of population uniques. In step 3, we count the number of empty cells instead of cells with only one element because we generate the variate $\tau^{(m)}$ conditional on the sample, which already includes one element in each of those cells. As noted by an associate editor, the algorithm also offers an estimate of $p(\tau = 0 \mid \mathcal{X}^*, \mathrm{T}(S))$, i.e., no sample uniques are population uniques. This could be computed as the average of $p(H_1 > 0, \ldots, H_U > 0 \mid \theta^{(m)}, \mathrm{T}(S))$ over the simulated values of $\theta^{(m)}$, where each probability can be determined directly from the multinomial distribution.

6.2 Description of Data

We use data from the 5% public use microdata sample (PUMS) of the 2000 U. S. census for the state of New York (Ruggles et al., 2010). We treat all $H = 953,076$ individuals from the sample as a population from which we take a random sample of $n = 5,000$ individuals. We select ten categorical variables as keys: ownership of dwelling (OWNERSHP: 3 levels), mortgage status (MORTGAGE: 4 levels), age (AGE: 9 levels), sex (SEX: 2 levels), marital status (MARST: 6 levels), single race identification (RACESING: 5 levels), educational attainment (EDUC: 11 levels), employment status (EMPSTAT: 4 levels), work disability status (DISABWRK: 3 levels),

and veteran status (VETSTAT: 3 levels). For all variables except AGE, we use the original coding in the PUMS. These collection of variables define a contingency table with 2,566,080 cells.

We categorize AGE into 9 groups: 0–14, 15, 16, 17, 18–24, 25–35, 36–50, 51–70, and 71+. We included distinct categories for ages 15, 16, and 17 to describe combinations of levels that are excluded by design and have to be treated as structural zeros. For example, EMPSATAT is defined only for people 16 or older, whereas MARST is defined only for people 15 or older. Parsing all the impossible combinations of levels, driven by the pairwise combinations OWNERSHP-MORTGAGE and AGE-(MARST, EMPSTAT, EDUC, DISABWRK and VETSTAT), we end up with 60 overlapping marginal conditions that represent 2,317,030 cells—essentially the entire table. After applying the algorithm in Section 4.2 to find an equivalent representation of this set, we dramatically reduce to 567 disjoint marginal conditions. We note that the algorithm ran in a fraction of second using a standard desktop computer.

6.3 Results

We take 100 independent samples with $n = 5000$, each time estimating the posterior distribution of τ based on the method in Section 6.1. To illustrate the effect of ignoring the structural zeros on the estimates, we also fit the regularized latent class model from (19) without the correction for truncation to the same set of 100 samples. For all cases, we use uniform prior distributions for all $\lambda_{jk}[\cdot]$. Following the advice from Dunson and Xing (2009) we use distribution $\alpha \sim \text{Gamma}(0.25, 0.25)$ in shape/inverse scale parametrization as a default non-informative prior.

In all cases, we set the maximum number of mixture components to $K = 50$. The sparsity-inducing prior specification of $\boldsymbol{\pi}$ results in an effective number of components, i.e., those comprising at least one individual, typically around 22. This is evident in Figure 1, which displays the posterior distribution of the effective number of com-

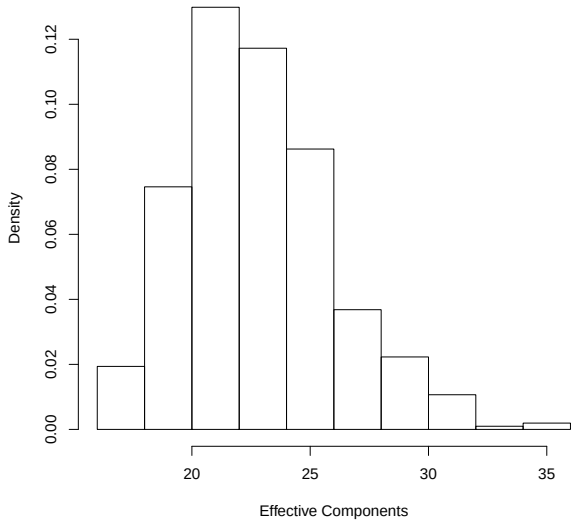

Figure 1: Posterior distribution of the effective number of components for a truncated regularized latent class model with $K = 50$, fitted to a sample of $n = 5,000$ from the NY data.

ponents for one of the samples. Further experimentation shows that the posterior distribution of the effective number of components is relatively insensitive to other vague prior specification of α and larger values of K. See Si and Reiter (2013) for further discussion of choosing K.

Figure 2 shows posterior 95% equal tail credible intervals and posterior medians of τ given each sample, re-centered at the true values of τ as computed from the population and each sample. We note that there is no single value of τ, since it is a function of both the population and each sample. When using the truncation, most of the posterior distributions of τ are appropriately centered near zero, so that each τ is closely estimated across the 100 replications. The average of τ over the 100 trials is 56.8 (sd = 7.43), and the average of the 100 posterior estimates (the posterior medians) of the corresponding τs is 57.1 (sd = 4.89). In contrast, when we ignore the effect of the truncation, the resulting estimates of the τs over-estimate the disclosure

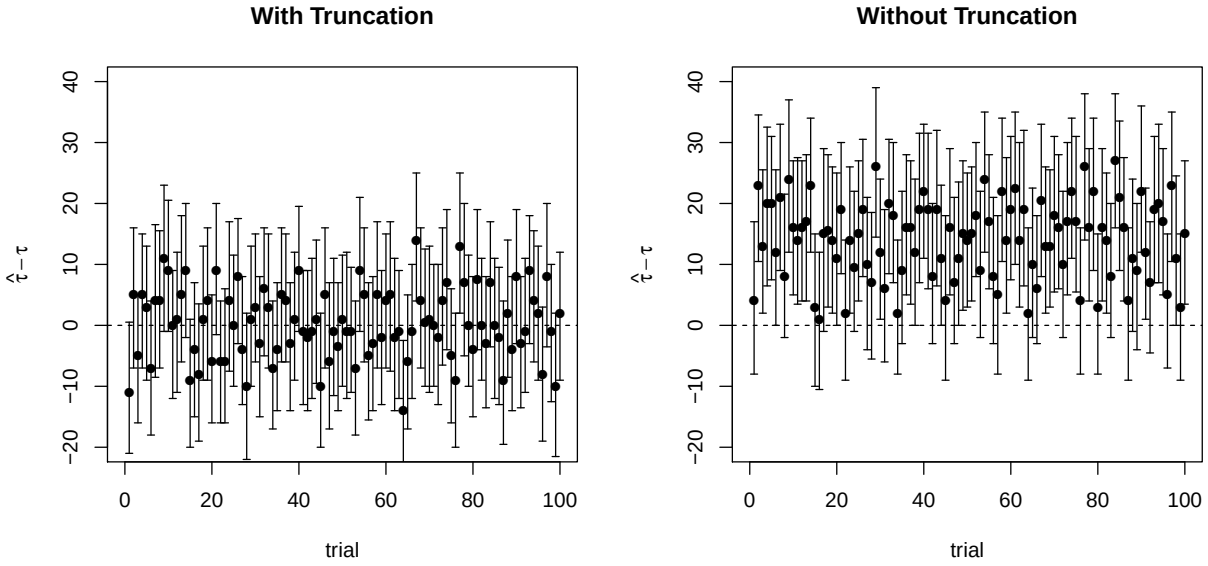

Figure 2: Posterior estimates of τ, re-centered at the true value of τ, for 100 random samples with $n = 5,000$ from the NY census data using regularized Bayesian latent class models with (left) and without (right) correction for truncation. Each bar marks the 0.025, 0.5 and 0.975 quantiles.

risk measure substantially, with a mean over the 100 trials of 70.9 (sd = 6.28). As a result, agencies basing risk calculations on the untruncated model could apply more disclosure treatment to the data, e.g., aggregation or data swapping (Reiter, 2012), than needed, thereby unnecessarily reducing the quality of the release data.

Although the results obviously are specific to the New York PUMS data, the truncated latent class model offers remarkably accurate estimates of τ. Given its flexibility, scalability, and computationally efficiency—for each sample the MCMC takes only a few minutes of running time on a standard desktop computer—these results suggest that statistical agencies could benefit from using the truncated latent class model in disclosure risk assessments.

7 Concluding Remarks

We conclude with a brief discussion of applications of TLSMs beyond estimating probabilities of uniqueness for disclosure risks. Basically, these models could apply to any setting involving truncated contingency tables. For example, TLSMs could be highly useful as engines for multiple imputation of missing values (Rubin, 1987) in surveys comprising many categorical items with extensive skip patterns, which represent structural zeros. Typical approaches to multiple imputation of categorical data based on log-linear models or chained equations (Raghunathan et al., 2001) can be impractical and unreliable, as it can be difficult to identify and accurately estimate important high-order interactions with these models. In contrast, a TLSM could capture complicated structure automatically while respecting structural zeros. Related, statistical agencies could use the posterior predictive distributions resulting from TLSMs to simulate realistic replicates of sampled or population-level categorical data, and release those replicates as public use files. This idea has been used for numerical data under the name synthetic data (Reiter and Raghunathan, 2007). TLSMs are directly applicable when certain combinations have been wholly removed from the sample, for example

by statistical agencies seeking to reduce disclosure risks or by data analysts seeking to exclude certain subgroups from contingency table analyses. Equivalently, TLSMs are useful when certain combinations have been effectively eliminated from the sample by design. A special case of this setup are Bayesian formulations of multiple-recapture population size estimation using latent variables (see e.g. Fienberg et al., 1999; Basu and Ebrahimi, 2001; Manrique-Vallier and Fienberg, 2008), in which the objective is to estimate the size of the unobserved portion of a population given joint observation patterns in multiple partial lists. Here, the TLSM offers estimates of the size of the population $(n + n_0)$.

References

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

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

Clogg, C. (1995), "Latent Class Models," in *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, eds. Arminger, G., Clogg, C., and M.E., S., New York: Plenum Press, pp. 311–360.

Dinwoodie, I. H. and Chen, Y. (2011), "Sampling large tables with constraints," *Statistica Sinica*, 21, 1591–1609.

Dobra, A. (2012), "Dynamic Markov bases," *Journal of Computational and Graphical Statistics*, 21, 496–517.

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

Erosheva, E., Fienberg, S., and Joutard, C. (2007), "Describing disability through individual-level mixture models for multivariate binary data," *Annals of Applied Statistics*, 1, 502–537.

Fienberg, S. (1972), "The Multiple recapture census for closed populations and incomplete 2^k contingency tables," *Biometrika*, 59, 591–603.

Fienberg, S., Johnson, M., and Junker, B. (1999), "Classical multilevel and Bayesian approaches to population size estimation using multiple lists," *Journal of the Royal Statistical Society. Series A*, 162, 383–406.

Gebregziabher, M. and DeSantis, S. M. (2010), "Latent class based multiple imputation approach for missing categorical data," *Journal of Statistical Planning and Inference*, 140, 3252–3262.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004), *Bayesian Data Analysis*, London: Chapman & Hall.

Goodman, L. (1968), "The analysis of cross-classified data: Independence, quasi-independence, and interactions in contingency tables with or without missing entries," *Journal of the American Statistical Association*, 63, 1091–1131.

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

Holland, P. W. and Rosenbaum, P. R. (1986), "Conditional association and unidimensionality in monotone latent variable models," *Annals of Statistics*, 14, 1523–1543.

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

Lazarsfeld, P. and Henry, N. (1968), *Latent structure analysis*, Houghton Mifflin Co.

Manrique-Vallier, D. and Fienberg, S. (2008), "Population size estimation using individual level mixture models," *Biometrical Journal*, 50, 1051–1063.

Manrique-Vallier, D. and Reiter, J. P. (2012), "Estimating Identification Disclosure Risk Using Mixed Membership Models," *Journal of the American Statistical As-*

sociation, (forthcoming).

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

O'Malley, A. J. and Zaslavsky, A. M. (2008), "Domain-Level Covariance Analysis for Multilevel Survey Data With Structured Nonresponse," *Journal of the American Statistical Association*, 103, 1405–1418.

Patz, R. J. and Junker, B. W. (1999), "A Straightforward Approach to Markov Chain Monte Carlo Methods for Item Response Models," *Journal of Educational and Behavioral Statistics*, 24, 146–178.

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., and Solenberger, P. (2001), "A multivariate technique for multiply imputing missing values using a series of regression models," *Survey Methodology*, 27, 85–96.

Rasch, G. (1980), *Probabilistic Models for Some Intelligence and Attainment Tests*, Chicago: University of Chicago Press, expanded edition of the 1960 work, with forward and afterward by B.D. Wright.

Reiter, J. P. (2012), "Statistical approaches to protecting confidentiality for microdata and their effects on the quality of statistical inferences," *Public Opinion Quarterly*, 76, 163–181.

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

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

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

Sethuraman, J. (1994), "A constructive definition of Dirichlet measures," *Statistica Sinica*.

Si, Y. and Reiter, J. P. (2013), "Nonparametric Bayesian Multiple Imputation for Incomplete Categorical Variables in Large-Scale Assessment Surveys," *Journal of Educational and Behavioral Statistics*, forthcoming.

Sibuya, M., Yoshimura, I., and Shimizu, R. (1964), "Negative multinomial distribution," *Annals of the Institute of Statistical Mathematics*, 16, 409–426.

Sijtsma, K. and Junker, B. (2006), "Item response theory: Past performance, present developments, and future expectations," *Behaviormetrika*, 33, 75–102.

Skinner, C. and Shlomo, N. (2008), "Assessing Identification Risk in Survey Microdata Using Log-Linear Models," *Journal of the American Statistical Association*, 103, 989–1001.

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

Tanner, M. (1996), *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, New York: Springer Verlag, 3rd ed.

Vermunt, J. K. (1997), *LEM: A General Program for the Analysis of Categorical Data*, Department of Methodology and Statistics, Tilburg University.

Vermunt, J. K., Ginkel, J. R. V., der Ark, L. A. V., and Sijtsma, K. (2008), "Multiple imputation of incomplete categorical data using latent class analysis," *Sociological Methodology*, 38, 369–397.

Woodbury, M., Clive, J., and Garson Jr, A. (1978), "Mathematical typology: A grade of membership technique for obtaining disease definition," *Computers in Biomedical Research*, 11, 277–98.

A Proof of Theorem 1

Following the generative model from Section 3, the joint distribution of $(\mathcal{X}^0, \mathcal{X}^1, \mathcal{Z}^0, \mathcal{Z}^1)$ conditional on N and θ is

$$p(\mathcal{X}^0,\mathcal{X}^1, \mathcal{Z}^0, \mathcal{Z}^1|\theta, N) = \binom{N}{n} 1\{N \geq n\} \prod_{i=1}^{n} g(\mathbf{x}_i^1 \mid \mathbf{z}_i^1) h(\mathbf{z}_i^1|\theta) 1\{\mathbf{x}_i^1 \notin S\} \times \prod_{i=1}^{N-n} g(\mathbf{x}_i^0 \mid \mathbf{z}_i^0) h(\mathbf{z}_i^0|\theta) 1\{\mathbf{x}_i^0 \in S\}$$

where n is the length of $\mathcal{X}^1$. Assuming $p(N) \propto 1/N$ independent of $p(\theta)$ and replacing $n_0 = N - n \geq 0$, we have

$$\begin{aligned}
&p(\theta,\mathcal{Z}^0, \mathcal{Z}^1, \mathcal{X}^0, N \mid \mathcal{X}^1) \\
&\quad \propto p(\mathcal{X}^0, \mathcal{X}^1, \mathcal{Z}^0, \mathcal{Z}^1|\theta, N) p(\theta) p(N) \\
&\quad \propto p(\theta) \binom{N-1}{n-1} 1\{N \geq n\} \prod_{i=1}^{n} g(\mathbf{x}_i^1 \mid \mathbf{z}_i^1) h(\mathbf{z}_i^1|\theta) 1\{\mathbf{x}_i^1 \notin S\} \times \prod_{i=1}^{n_0} g(\mathbf{x}_i^0 \mid \mathbf{z}_i^0) h(\mathbf{z}_i^0|\theta) 1\{\mathbf{x}_i^0 \in S\}.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&\int p(\theta,\mathcal{Z}^1, \mathcal{Z}^0, \mathcal{X}^0, n_0 \mid \mathcal{X}^1) d\mathcal{X}^0 d\mathcal{Z}^1 d\mathcal{Z}^0 dn_0 \\
&\quad \propto p(\theta) \prod_{i=1}^{n} \int g(\mathbf{x}_i^1 \mid \mathbf{z}, \theta) 1\{\mathbf{x}_i^1 \notin S\} h(\mathbf{z} \mid \theta) d\mathbf{z} \\
&\qquad \times \sum_{n_0=0}^{\infty} \binom{n_0+n-1}{n-1} \prod_{i=1}^{n_0} \sum_{\mathbf{x} \in \mathcal{C}} \int g(\mathbf{x} \mid \mathbf{z}, \theta) h(\mathbf{z} \mid \theta) 1\{\mathbf{x} \in S\} d\mathbf{z} \\
&\quad = p(\theta) \prod_{i=1}^{n} 1\{\mathbf{x}_i^1 \notin S\} \int g(\mathbf{x}_i^1 \mid \mathbf{z}\theta) h(\mathbf{z} \mid \theta) d\mathbf{z} \times \sum_{n_0=0}^{\infty} \binom{n_0+n-1}{n-1} \left[\sum_{\mathbf{x} \in S} f^{TSM}(\mathbf{x}|\theta) \right]^{n_0} \\
&\quad = p(\theta) \prod_{i=1}^{n} 1\{\mathbf{x}_i^1 \notin S\} \int g(\mathbf{x}_i^1 \mid \mathbf{z}, \theta) h(\mathbf{z} \mid \theta) d\mathbf{z} \times (1 - \pi_0(\theta))^{-n} \\
&\quad = \frac{p(\theta)}{(1 - \pi_0(\theta))^n} \prod_{i=1}^{n} 1\{\mathbf{x}_i^1 \notin S\} \int g(\mathbf{x}_i^1 \mid \mathbf{z}, \theta) h(\mathbf{z} \mid \theta) d\mathbf{z} \\
&\quad = p(\theta \mid \mathcal{X}^1, \mathrm{T}(S)) = p(\theta \mid \mathcal{X}^*, \mathrm{T}(S)),
\end{aligned}$$

completing the proof.