Published in final edited form as:
J Am Stat Assoc. 2012 December 1; 107(500): 1385–1394. doi:10.1080/01621459.2012.710508.

Estimating Identification Disclosure Risk Using Mixed Membership Models

Daniel Manrique-Vallier and
Postdoctoral Associate at the Social Science Research Institute and the Department of Statistical Science, Duke University, Durham, NC 27708-0251

Jerome P. Reiter
Mrs. Alexander Hehmeyer Associate Professor of Statistical Science, Duke University, Durham, NC 27708-0251

Daniel Manrique-Vallier: dman200@stat.duke.edu; Jerome P. Reiter: jerry@stat.duke.edu

Abstract

Statistical agencies and other organizations that disseminate data are obligated to protect data subjects' confidentiality. For example, ill-intentioned individuals might link data subjects to records in other databases by matching on common characteristics (keys). Successful links are particularly problematic for data subjects with combinations of keys that are unique in the population. Hence, as part of their assessments of disclosure risks, many data stewards estimate the probabilities that sample uniques on sets of discrete keys are also population uniques on those keys. This is typically done using log-linear modeling on the keys. However, log-linear models can yield biased estimates of cell probabilities for sparse contingency tables with many zero counts, which often occurs in databases with many keys. This bias can result in unreliable estimates of probabilities of uniqueness and, hence, misrepresentations of disclosure risks. We propose an alternative to log-linear models for datasets with sparse keys based on a Bayesian version of grade of membership (GoM) models. We present a Bayesian GoM model for multinomial variables and offer an MCMC algorithm for fitting the model. We evaluate the approach by treating data from a recent US Census Bureau public use microdata sample as a population, taking simple random samples from that population, and benchmarking estimated probabilities of uniqueness against population values. Compared to log-linear models, GoM models provide more accurate estimates of the total number of uniques in the samples. Additionally, they offer record-level predictions of uniqueness that dominate those based on log-linear models.

Keywords

Contingency table; Confidentiality; Disclosure; Grade of membership; Latent class

Correspondence to: Jerome P. Reiter, jerry@stat.duke.edu.

1 Introduction

Many organizations view sharing record-level data with others as an integral part of their mission. For example, federal statistical agencies disseminate public use files on individual persons, businesses, schools, farms, etc. Principal investigators running large data collection efforts share their data with researchers who are not on the original investigative team. Funding agencies like the National Institutes of Health and National Science Foundation mandate that their grantees make data available to others to promote new discoveries and reproducible research.

When sharing data, organizations are ethically and often legally obligated to protect the confidentiality of data subjects' identities and sensitive attributes. Thus, organizations must assess the risks that ill-intentioned individuals, henceforth called intruders, can learn confidential information from any proposed data release. These risks arise when intruders can link records in the released data to other databases (that include direct identifiers) by matching on variables common to the two databases. For example, Sweeney (2001) famously showed that 97% of the records in publicly available voter registration lists in Cambridge, MA, could be uniquely identified using birth date and nine digit zip code. By matching on the information in these lists, she was able to identify Governor William Weld in an anonymized medical database.

In this article, we consider a key aspect of disclosure risk assessment: estimating whether or not a record has a set of discrete characteristics, say **Y**, that are unique in the population. Population uniques are at comparatively high risk of identification disclosure; for example, when an intruder matches a released record to a record in an external database based on **Y**, the intruder's match is guaranteed to be correct (assuming no errors in the matching process or data sources) when the record is a population unique on those characteristics. The number of population uniques in the sample is used as a file-level measure of the disclosure risk associated with releasing the sampled data (Bethlehem et al., 1990; Greenberg and Zayatz, 1992; Skinner, 1992; Skinner et al., 1994; Chen and Keller-McNulty, 1998; Fienberg and Makov, 1998; Samuels, 1998; Pannekoek, 1999; Dale and Elliot, 2001; Elamir and Skinner, 2006; Forster and Webb, 2007; Skinner and Shlomo, 2008). Population uniqueness also factors into record-level disclosure risk measures for data that have been altered to protect confidentiality (Reiter, 2005; Drechsler and Reiter, 2008; Shlomo and Skinner, 2010).

Typically in sampled data, the organization does not know which records are population uniques on **Y**; rather, it only knows that records are unique, or not, in the sample. A variety of approaches have been developed to estimate probabilities that records are population uniques given that they are sample uniques. Most are based on cell probabilities in the table of **Y** that are estimated with log-linear models; see Skinner and Shlomo (2008) and the references therein.

Log-linear models have potential shortcomings for estimating population uniqueness. When tables are large and sparse, as is often the case with high-dimensional **Y**, estimates of cell probabilities from log-linear models can be distorted by the many random zero counts in the table. Typically, this results in overestimation of the population counts and, thus,

underestimation of the true risks of identification disclosures (Skinner and Shlomo, 2008). This sensitivity to sparsity is an intrinsic limitation of maximum likelihood estimation for log-linear models (Bishop et al., 1975; Erosheva et al., 2002); we discuss it further in Section 6. Additionally, it is not obvious how to select the terms, e.g., the order of interactions, to include in the models. However, estimates of the number of population uniques can change dramatically for different specifications (Skinner and Shlomo, 2008). Finally, when one considers the possible inclusion of high-order interactions for high-dimensional $\mathbf{Y}$, the number of potential log-linear models can be so large that evaluating all of them becomes infeasible.

To facilitate the process of log-linear model choice, Skinner and Shlomo (2008) developed principled criteria specific to the estimation of population uniqueness. They employ these criteria in stepwise model searches to identify log-linear models for disclosure risk estimation. Another perspective on model selection was offered by Forster and Webb (2007), who restrict the log-linear models to the sub-class of decomposable graphical models and use Bayesian model averaging to account for model uncertainty.

We propose to use a Bayesian version of the grade of membership (GoM) models (Woodbury et al., 1978; Erosheva et al., 2007) to estimate probabilities of population uniqueness in large, sparse contingency tables. GoM models have been shown to be particularly effective for estimation of cell probabilities in such tables; for example, they have been applied on 2^{16} sparse contingency tables in studies of disability among elders (Erosheva, 2002; Erosheva et al., 2007). Using an empirical study based on public use microdata from the state of California, we find that the GoM models deliver superior estimates of the number of population uniques compared to estimates based on log-linear models with the criteria of Skinner and Shlomo (2008). The GoM models also result in more accurate record-level predictions of uniqueness than the log-linear models, resulting in fewer false negatives (record estimated not to be population unique when it is) with similar rates of false positives (record estimated to be population unique when it is not).

The remainder of the article is organized as follows. In Section 2, we describe measures of identification disclosure risk based on population uniques that we use throughout. In Section 3, we present a GoM model for multinomial variables and an MCMC sampler for its Bayesian estimation. In Section 4, we show how to compute the risk measures using the GoM models. In Section 5, we empirically compare the performance of the GoM model to several approaches based on log-linear models using the California data. In Section 6, we conclude with a discussion of implementing the GoM model for disclosure risk assessment in practice.

2 Identification Disclosure Risk Measures

To provide context for the measures of identification disclosure risk, we begin by framing the setting of interest. An organization has collected a sample of n records from a finite population of size N. For each record $i = 1, \ldots, N$, let $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iJ})$ be the individual $J \times 1$ vector of key variables. The key variables are those from the sample deemed by the organization to be available to intruders in other data sources. Specifying the key variables is

a challenging task; we refer readers to the literature on identification disclosure risk assessment for advice on key specification (e.g., Duncan et al., 2011, Chapter 2). We suppose that, for $j = 1, \ldots, J$, the components Y_{ij} are discrete random variables with $n_j \geq 2$ levels. For convenience, we label the levels of each Y_{ij} using consecutive integers from one to n_j. Thus, $\boldsymbol{Y}_i \in \mathscr{C} = \prod_{j=1}^{J} \{1, 2, \ldots, n_j\}$ for all i. Following the literature on identification disclosure risk, we do not consider continuous key variables as they essentially imply that every record is a population unique. We ignore all variables in the sample that are not keys, since they are not available to intruders for matching.

We consider each $\mathbf{Y}_i$ in the population to be a realization from a common parametric super-population distribution indexed by θ, $\boldsymbol{Y}_i \overset{iid}{\sim} P_\theta$. The population can be summarized by a contingency table, $\mathbf{F}$, with counts in each cell $\mathbf{x} \in \mathcal{C}$ of the table defined as $F_{\mathbf{x}} = \sum_{i=1}^{N} I(\boldsymbol{Y}_i = \mathbf{x})$, where $I(\cdot) = 1$ if the argument is true and $I(\cdot) = 0$ otherwise. For every $\mathbf{x} \in \mathcal{C}$, let $P_\theta^{\mathbf{x}} = \Pr(\boldsymbol{Y} = \mathbf{x})$ be the probability of being in cell $\mathbf{x}$. The population cell count distribution is then multinomial with sample size N and probabilities $\{P_\theta^{\mathbf{x}}\}$. For large N and small cell probabilities, which is generally the case in practice, a reasonable approximation is $F_{\mathbf{x}} \sim \text{Poisson}(NP_\theta^{\mathbf{x}})$.

Moving now to the sample, we suppose that it is collected from the population following a Bernoulli sampling design with cell-homogeneous selection probabilities, $\pi_{\mathbf{x}}$. The sample also can be summarized as a contingency table, $\mathbf{f}$, with cell counts $f_{\mathbf{x}}$. Using the Poisson approximation for the population cell count distribution, we have $f_{\mathbf{x}} \sim \text{Poisson}(\pi_{\mathbf{x}} N P_\theta^{\mathbf{x}})$. As J and the number of levels n_j in the keys increase, the size of the number of cells in $\mathbf{f}$ increases exponentially. This has the effect of producing tables where the number of cells greatly exceeds the sample size n and, consequently, many cells have $f_{\mathbf{x}} = 0$, i.e., the sample table is sparse.

The organization has only $\mathbf{f}$ and does not know $\mathbf{F}$; hence, it does not know which records are unique in the population on $\mathbf{Y}$. Therefore, following common practice in disclosure risk assessment (e.g., Skinner and Holmes, 1998; Skinner and Shlomo, 2008), we seek to estimate the probability of being unique in the population conditional on being unique in the sample,

$$\mu_{\mathbf{x}} = \Pr(F_{\mathbf{x}} = 1 | f_{\mathbf{x}} = 1), \quad (1)$$

for all $\mathbf{x}$ such that $f_{\mathbf{x}} = 1$. Under the Poisson distributions for $\mathbf{F}$ and $\mathbf{f}$, we have

$$\mu_{\mathbf{x}} = \exp\left(-NP_\theta^{\mathbf{x}}(1 - \pi_{\mathbf{x}})\right). \quad (2)$$

Thus, an estimate of $P_\theta^{\mathbf{x}}$, e.g., from a log-linear model on $\mathbf{f}$, results in an estimate of $\mu_{\mathbf{x}}$.

For a measure of file-level risk, we consider the number of sample uniques that are also population uniques,

$$\tau = \sum_{\mathbf{x} \in C} I(F_{\mathbf{x}} = 1, f_{\mathbf{x}} = 1). \quad (3)$$

The quantities τ and $\mu_{\mathbf{x}}$ are related to each other through the identity

$$E[\tau | \mathbf{f}] = \sum_{\{\mathbf{x}: f_{\mathbf{x}} = 1\}} \mu_{\mathbf{x}}, \quad (4)$$

so that the estimates of μ_x can be used to estimate τ.

When using model-based frameworks such as the one here, population quantities like $F_{\mathbf{x}}$ and τ are random variables; that is, they are generated from a super-population process. Thus, when estimating the variability of $F_{\mathbf{x}}$ and τ, we should account for two sources of uncertainty: the estimation of the super-population model from **F** and the estimation of **F** from **f**. We refer readers to Rinott and Shlomo (2007) for discussion of this issue for (non-Bayesian) log-linear modeling contexts.

3 Grade of Membership Models

In GoM models, we assume the existence of a small number of latent classes, which can be interpreted as typical cases or *extreme profiles* of individuals in the population. However, rather than force each individual to be fully characterized by only one extreme profile, GoM models assume that individuals lie somewhere in between extremes. They allow individuals to share characteristics from each of the extreme profiles simultaneously with varying degrees determined by their position with respect to the extreme profiles. GoM models belong to the general class of mixed membership models (Erosheva, 2002; Erosheva et al., 2004), in which individuals are allowed membership in more than one class simultaneously. Models from this family have been proposed for a wide range of applications, including the study of disability among elders (Manton et al., 1994; Erosheva et al., 2007), network analysis (Airoldi et al., 2008), electoral preferences analysis (Gormley, 2006; Gormley and Murphy, 2008), estimation of judgment accuracy (Cooil and Varki, 2003), estimation of population sizes (Manrique-Vallier and Fienberg, 2008), genetic composition analysis (Pritchard et al., 2000) and text classification (Erosheva et al., 2004; Blei et al., 2003; Blei and Lafferty, 2007). Like other latent structure models, mixed membership models offer an approach to the analysis of large, sparse contingency tables with complex interactions. In contrast, log-linear models sometimes can be inadequate for analysis of such tables, because sparsity often prevents the estimation of structurally required high order terms.

To construct a GoM model, we first set the number of extreme profiles, K. We characterize these profiles as follows. For any individual that is a full member of the kth extreme profile, i.e., does not belong to any profile but k, we require that $\Pr(Y_{ij} = l | i\text{th individual in } k\text{th class}) = \lambda_{jk[l]}$, where $l \in \{1, 2\ldots, n_j\}$ and $\sum_{l=1}^{n_j} \lambda_{jk[l]} = 1$. Hence, all individuals of an extreme profile have the same probability distributions for **Y**. It is important to emphasize that the extreme profiles are latent and idealized; in reality there might not be any individuals who are full members of extreme profiles.

We next characterize the sampled individuals by associating each of them with its own K-dimensional membership vector, $\mathbf{g}_i = (g_{i1}, g_{i2}, \ldots, g_{iK})$, representing how much of a member of each class this particular individual is. Membership scores are restricted so that all $g_{ik} > 0$ and $\sum_{k=1}^{K} g_{ik} = 1$. We call the geometry implied by the possible values of $\mathbf{g}_i$ and extreme profiles the $K - 1$ dimensional unit simplex, and denote it by Δ_{K-1} or simply Δ when no ambiguity arises. We introduce the idea of partial membership by setting the distribution of each Y_{ij} given $\mathbf{g}_i$ as the convex combination,

$$p(y_{ij}|\mathbf{g}_i) = \Pr(Y_{ij} = y_{ij}|\mathbf{g}_i) = \sum_{k=1}^{K} g_{ik}\lambda_{jk[y_{ij}]}. \quad (5)$$

This characterization based on convexity is the defining characteristic of the GoM model. Geometrically, it implies that all individuals lie in the convex hull defined by the extreme profiles, with relative positions expressed by their membership vectors in barycentric coordinates with respect to the vertices. For additional study of the geometry of the GoM model, see Erosheva (2005).

We further assume that the J variables in $\mathbf{Y}_i$ are conditionally independent given $\mathbf{g}_i$. This condition, sometimes referred as latent conditional independence or local independence (Holland and Rosenbaum, 1986; Sijtsma and Junker, 2006), expresses the idea that the membership vector completely explains the dependence structure among the J manifest variables. With this assumption, we can construct the conditional joint distribution as

$$p(\mathbf{y}_i|\boldsymbol{\lambda}, \mathbf{g}_i) = \prod_{j=1}^{J} \sum_{k=1}^{K} g_{ik}\lambda_{jk[y_{ij}]}. \quad (6)$$

Assuming that the membership vectors all come from a common population-level distribution G with support in Δ ($g_i \overset{iid}{\sim} G$), we have the GoM model likelihood

$$p(\mathbf{y}_i|\boldsymbol{\lambda}, G) = \int_{\Delta} \prod_{j=1}^{J} \sum_{k=1}^{K} \gamma_k \lambda_{jk[y_{ij}]} G(d\gamma), \quad (7)$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_k)$.

The net effect is to represent the potentially huge contingency table (with $n_1 \times n_2 \times \ldots \times n_J$ cells) using only K extreme profiles, individually mixed according to a common distribution G. Thus, the GoM model offers significant reductions of dimensionality.

3.1 The GoM model as a latent class model

If we restrict the support of G from Δ to just its vertices (so that if $\boldsymbol{\gamma} \sim G$, then $\gamma_k = 1$ for some k and $\gamma_{k'} = 0$ for any $k' \neq k$), the expression in (7) simplifies to

$$p(\mathbf{y}_i|\boldsymbol{\lambda}, G)=\sum_{k=1}^{K}\pi_k\prod_{j=1}^{J}\lambda_{jz_j[y_{ij}]} \quad (8)$$

where $\pi_k = \Pr(\gamma_k = 1)$. This expression corresponds to a finite mixture of multinomial distributions, i.e., a latent class model (Goodman, 1974). Thus, we can view GoM models as a generalization of latent class models, where we allow membership vectors to take values from any position in the unit simplex instead of just its vertices.

Somewhat paradoxically, GoM models also can be characterized as a subclass of certain restricted latent class models (Haberman, 1995; Erosheva et al., 2007). To see this, consider again the expression in (7). Under the assumption $\mathbf{g}_i \overset{iid}{\sim} G$, we have

$$p(\mathbf{y}_i|\boldsymbol{\lambda}, G)=\int_{\Delta}\prod_{j=1}^{J}\sum_{k=1}^{K}\gamma_k\lambda_{jk[y_{ij}]}G(d\boldsymbol{\gamma})=\sum_{\mathbf{z}\in\mathcal{Z}}\pi_{\mathbf{z}}\prod_{j=1}^{J}\lambda_{jk[y_{ij}]}, \quad (9)$$

with $\mathbf{z} = (z_1, \ldots, z_J) \in \mathcal{Z} = \{1, 2, \ldots, K\}^J$ and $\pi_{\mathbf{z}}=E_G\left[\prod_{j=1}^{J}\gamma_{z_j}\right]$. This expression, typical of a discrete mixture model, suggests the augmented data representation of the GoM model,

$$p(\mathbf{y}_i, \mathbf{z}_i|\boldsymbol{\lambda}, \mathbf{g}_i)=\prod_{j=1}^{J}\prod_{k=1}^{K}\left(g_{ik}\lambda_{jk[y_{ij}]}\right)^{I(z_{ij}=k)}, \quad (10)$$

where now we are considering individual $z_i = (z_{i1}\ldots, z_{iJ}) \in \mathcal{Z}$. For details, see Erosheva et al. (2007).

3.2 MCMC sampler for the multinomial GoM model

Erosheva (2002) and Erosheva et al. (2007) use the latent class representation in (10) to develop an MCMC algorithm for sampling from the posterior distribution of a full Bayesian GoM model for binary Y_{ij}. We now extend it to datasets with some $n_j > 2$, as is typical for key variables in practice.

To begin, we specify the distribution G for membership vectors as

$$p(\mathbf{g}_i|\boldsymbol{\alpha})=\text{Dirichlet}(\mathbf{g}_i|\boldsymbol{\alpha}), \quad (11)$$

with parameter $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$. Based on the likelihood in (10), the joint posterior distribution of the GoM parameters is

$$p(\mathbf{a}, \boldsymbol{\lambda}, \mathbf{g}|\mathbf{y}, \mathbf{z}) \propto p(\boldsymbol{\lambda})p(\boldsymbol{\alpha})\left(\prod_{i=1}^{n}p(\mathbf{g}_i|\boldsymbol{\alpha})\right)\times\prod_{i=1}^{n}\prod_{j=1}^{J}\prod_{k=1}^{K}\left(g_{ik}\lambda_{jk[y_{ij}]}\right)^{I(z_{ij}=k)}. \quad (12)$$

As prior distributions for λs, we use $p(\lambda_{jk[\cdot]}) = \text{Dirichlet}(\mathbf{1}_{n_j})$, i.e., a uniform on the simplex Δ_{n_j-1}. Following Erosheva et al. (2007), we adopt the well known Dirich-let parametrization $a_0 = \Sigma_k\, a_k$ and $\boldsymbol{\xi} = (a_1/a_0, \ldots, a_K/a_0)$, and specify independent prior distributions for a_0 and $\boldsymbol{\xi}$. Parameter $\boldsymbol{\xi}$ is the expected value of the distribution. Parameter a_0 governs the concentration of the distribution. Small values of a_0 concentrate the distribution towards the vertices of Δ in proportions given by $\boldsymbol{\xi}$; large values concentrate the probability towards the expected value $\boldsymbol{\xi}$.

We use the hyperprior distributions $a_0 \sim \text{Gamma}(2, 1)$ and $\boldsymbol{\xi} \sim \text{Dirichlet}(\mathbf{1}_K)$, where the gamma distribution is in shape/inverse-scale parametrization. These distributions express *a priori* ignorance about $\boldsymbol{\xi}$ and a slight preference for small values of a_0. This preference for small values is based more on structural modeling decisions than expressions of prior knowledge: we regard individuals as belonging principally to a single extreme profile, but with influence from the others. In our empirical work, we have found that posterior distributions of a_0 tend to be strongly data-dominated (see the online appendix for evidence), but prior specifications with increased mass towards higher values of a_0, e.g., Gamma[2, 0.1], cause the Markov chains to exhibit relatively poor convergence properties.

To sample from the posterior distribution of the parameters, we use Gibbs sampling with Metropolis-Hastings steps when needed. The MCMC sampler can be implemented with successive applications of the following four steps.

1. For every $i \in \{1 \ldots n\}$ and $j \in \{1 \ldots J\}$, sample

$$(z_{ij}|\ldots) \sim \text{Discrete}(p_1, p_2, \ldots, p_K) \quad (13)$$

   with $p_k \propto g_{ik} \cdot \lambda_{jk[y_{ij}]}$ for all $k \in \{1, \ldots, K\}$.

2. Sample each $\lambda_{jk[\cdot]}$ from its full conditional distribution,

$$p(\lambda_{jk[\cdot]}|\ldots) \propto p(\lambda_{jk[\cdot]}) \times \prod_{i=1}^{n} \lambda_{jk[y_{jk}]}^{I(z_{ij}=k)} = \text{Dirichlet}(\lambda_{jk}|\mathbf{1}_{n_j}) \times \prod_{i=1}^{n}\prod_{l=1}^{n_j} \lambda_{jk[l]}^{I(z_{ij}=k, y_{jk}=l)} \quad (14)$$

$$\propto \text{Dirichlet}\left(1+\sum_{i=1}^{n} I(y_{ij}=1, z_{ij}=k), \ldots, 1+\sum_{i=1}^{n} I(y_{ij}=n_j, z_{ij}=k)\right) \quad (15)$$

3. Sample each $\mathbf{g}_i$ independently from its full conditional distribution,

$$(\mathbf{g}_i|\ldots) \sim \text{Dirichlet}\left(\alpha_1+\sum_{j=1}^{J} I(z_{ij}=1), \alpha_2+\sum_{j=1}^{J} I(z_{ij}=2), \ldots, \alpha_K+\sum_{j=1}^{J} I(z_{ij}=K)\right). \quad (16)$$

4. To sample $\boldsymbol{a}$, we first note that its full conditional distribution,

$$p(\boldsymbol{\alpha}|\ldots) \propto \text{Gamma}(\alpha_0|\tau,\eta) \times \text{Dirichlet}(\boldsymbol{\xi}|\mathbf{1}_K) \times \prod_{i=1}^{n} \text{Dirichlet}(\mathbf{g}_i|\boldsymbol{\alpha}) \quad (17)$$

$$\propto \alpha_0^{\tau-1}\exp[-\alpha_0\eta] \times \left[\frac{\Gamma(\alpha_0)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\right]^n \prod_{k=1}^{K}\left[\prod_{i=1}^{n} g_{ik}\right]^{\alpha_k}, \quad (18)$$

does not have a recognizable form. Thus, we use a Metropolis-Hastings within Gibbs step. We treat the vector a as a block and use the logarithmic scale, Gaussian random walk Metropolis-Hastings step proposed in Manrique-Vallier and Fienberg (2008):

a. Sample each component of $\boldsymbol{\alpha}^*=(\alpha_1^*, \alpha_2^*, \ldots, \alpha_K^*)$, as independent lognormal variates from

$$\alpha_k^* \overset{indep.}{\sim} \text{lognormal}(\log\alpha_k, \sigma^2). \quad (19)$$

b. Let $\alpha_0^*=\sum_{k=1}^{K}\alpha_k^*$, and compute

$$r=\min\left\{1, \exp\left[-\eta(\alpha_0^*-\alpha_0)\right]\left(\prod_{k=1}^{K}\frac{\alpha_k^*}{\alpha_k}\right)\left(\frac{\alpha_0^*}{\alpha_0}\right)^{\tau-1} \times \left[\frac{\Gamma(\alpha_0^*)}{\Gamma(\alpha_0)}\prod_{k=1}^{K}\frac{\Gamma(\alpha_k)}{\Gamma(\alpha_k^*)}\right]^n \prod_{k=1}^{K}\left(\prod_{i=1}^{n} g_{ik}\right)^{\alpha_k^*-\alpha_k}\right\}, \quad (20)$$

and update the chain, from step *m* to step *m* + 1, according to the rule

$$\boldsymbol{\alpha}^{(m+1)}=\begin{cases} \boldsymbol{\alpha}^* & \text{with probability } r \\ \boldsymbol{\alpha}^{(m)} & \text{with probability } 1-r. \end{cases} \quad (21)$$

This estimation algorithm does not take the data collection design into consideration; thus, it is appropriate only for ignorable sampling designs such as simple random sampling. We comment further on fitting GoM models with complex sampling designs in Section 6.

4 GoM Models for Disclosure Risk Estimation

In this section, we describe how to apply the GoM model to estimate the disclosure risk quantities $\mu_{\mathbf{x}}$ and τ from the sampled data. In what follows, we assume that **F** has been generated from a GoM model with implied probabilities $\{P_\theta^{\mathbf{x}}\}$, where $\theta=(\boldsymbol{\alpha}, \boldsymbol{\lambda})$.

From (2), the posterior distribution of each $\mu_{\mathbf{x}}$ is a function of its corresponding cell probability, $P^{\mathbf{x}}_{(\boldsymbol{\alpha},\boldsymbol{\lambda})}$, so that

$$p(\mu_{\mathbf{x}}|\mathbf{f})=p\left(\exp[-NP^{\mathbf{x}}_{(\boldsymbol{\alpha},\boldsymbol{\lambda})}(1-\pi_{\mathbf{x}})]|\mathbf{f}\right). \quad (22)$$

The cell probabilities are themselves functions of a and λ, whereby

$$P^{\mathbf{x}}_{(\boldsymbol{\alpha},\boldsymbol{\lambda})} = \int_{\Delta} \prod_{j=1}^{J} \sum_{k=1}^{K} \gamma_k \lambda_{jk[x_j]} G_{\boldsymbol{\alpha}}(d\boldsymbol{\gamma}). \quad (23)$$

Generalizing a result from Manrique-Vallier and Fienberg (2008), given α and λ, these probabilities can be computed with the closed-form formula,

$$P^{\mathbf{x}}_{(\boldsymbol{\alpha},\boldsymbol{\lambda})} = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + J)} \sum_{z \in \mathscr{Z}} \left(\prod_{k=1}^{K} \frac{\Gamma\left(\alpha_k + \sum_j I(z_j = k)\right)}{\Gamma(\alpha_k)} \times \prod_{j=1}^{J} \lambda_{j z_j [x_j]} \right). \quad (24)$$

This suggests that we sample θ from its posterior distribution and compute (24) to get a draw of $\mu_{\mathbf{x}}$ based on (22). However, (24) contains a sum over K^J terms which, while finite, quickly becomes unmanageable for models with a large number of extreme profiles or datasets with a large number of variables. In practice, therefore, we approximate (24) for given θ using Monte Carlo integration. In the empirical examples in Section 5, we find that Monte Carlo sample sizes as small as 5, 000 result in reasonably accurate estimates of even very small cell probabilities.

As evident in (3), the file-level risk measure, τ, is a function of both the observed sample, **f**, and the unobserved population, **F**, so that its posterior distribution requires integration over the posterior distribution of populations of size N. This is analytically intractable. However, as with $\mu_{\mathbf{x}}$, we can use a Monte Carlo approach to sample τ given draws of θ. Let $(\boldsymbol{\alpha}^{(m)}, \boldsymbol{\lambda}^{(m)})$ be the mth sample from the posterior distribution of θ. Then, to obtain the mth draw of τ, we apply the following algorithm.

1. Let $\mathbf{F}^{(m)} = \mathbf{f}$, i.e., initialize the population table as the sample table
2. For each remaining individual in the population $i \in \{n + 1, n + 2, \ldots, N\}$,
   a. Sample membership vector $\mathbf{g}_i^{(m)} \sim \text{Dirichlet}(\alpha^{(m)})$
   b. For each variable $j \in \{1, 2, \ldots, J\}$
      i. sample $z_{ij}^{(m)} \sim \text{Discrete}(\mathbf{g}_i^{(m)})$
      ii. sample $y_{ij}^{(m)} \sim \text{Discrete}(\lambda^{(m)}_{j z_{ij}^{(m)}})$
   c. Let $F_x^{(m)} = F_x^{(m)} + 1$, where x corresponds to the simulated cell value of $\mathbf{Y}_i$
3. Let $\tau^{(m)} = \sum_{\mathbf{x} \in \mathscr{C}} I(F_{\mathbf{x}}^{(m)} = 1, f_{\mathbf{x}} = 1)$

This procedure generates a contingency table, $\mathbf{F}^{(m)}$, from a simulated population comprising the n individuals from the sample and $N - n$ responses sampled from the GoM model conditional on $(\boldsymbol{\alpha}^{(m)}, \boldsymbol{\lambda}^{(m)})$. This process is akin to methods for Bayesian 3-nite population inference (Gelman et al., 2004, Chapter 7). Although we use these simulated populations only to compute τ, in principle they can be used to generate posterior predictive samples from any finite-population quantity of interest, including μ_x and population counts for **x** such

that $f_{\mathbf{x}} > 1$. They might also be used to create synthetic public use files (Rubin, 1993; Reiter and Raghunathan, 2007).

Another estimate of τ can be obtained by replacing $\mu_{\mathbf{x}}$ in (4) with some point estimate $\hat{\tau_{\mathbf{x}}}$—we use posterior medians in the empirical application—resulting in

$$\hat{\tau} = \sum_{\{\mathbf{x}: f_{\mathbf{x}}=1\}} \hat{\mu}_{\mathbf{x}}. \quad (25)$$

This is the basic form of the estimator employed by Skinner and Shlomo (2008), who use log-linear models to compute $\hat{\mu_x}$. The estimator in (25) has the computational advantage of reusing estimates of $\mu_{\mathbf{x}}$ that may have been produced otherwise. However, because (25) is actually an estimator of $E[\tau|\mathbf{f}]$ rather than of τ, posterior intervals or other estimates of uncertainty associated with (25) do not incorporate the uncertainty from sampling $\mathbf{F}$ from a super-population. In our empirical evaluations, the estimator in (25) and the Monte Carlo simulation procedure result in nearly identical point estimates of τ. We take this as further evidence of the quality of the Monte Carlo procedure used to approximate (24). However, since the population simulation procedure produces full posterior distributions and not just point estimates, we prefer it to (25) for estimation purposes.

5 Application to Census Data

To evaluate the performance of GoM models, we use data from the 5% public use microdata sample of the U.S. 2000 census for the state of California. The data are available via the Integrated Public Microdata Services (IPUMS, Ruggles et al., 2010). We treat all $N = 1, 150, 934$ individuals older than 21 years as the population, and specify ten key variables for $\mathbf{Y}$ (variable label and n_j in parentheses): number of children (A, 10 levels), Age (B, 10 levels), sex (C, 2 levels), marital status (D, 6 levels), race (E, 5 levels), education (F, 5 levels), employment status (G, 3 levels), income (H, 10 levels), disability (I, 2 levels) and veteran status (J, 2 levels). For all but age and income, the levels are defined in the IPUMS file. We categorize age and income into deciles. The resulting contingency tables for both population and samples have a total of 3,600,000 cells. Because we include only individuals older than 21, there are no impossible combinations of responses, such as married 3 year olds.

We select three independent, simple random samples of sizes $n \in \{1000, 5000, 10000\}$, and in each sample estimate the number of sample uniques that are population uniques. Table 1 displays relevant summaries of the three samples. The vast majority of the cells in each $\mathbf{f}$ are empty; for instance, even with the largest sample, the resulting contingency table has only 5, 478 of the 3, 600, 000 cells with non-zero counts. We note that simple random sampling differs slightly from Bernoulli sampling described in Section 2, but for small sampling fractions like those here the Bernoulli sampling with $\pi_{\mathbf{x}} = n/N$ is a reasonable approximation to simple random sampling.

In each sample, we compare the GoM model estimates with estimates obtained from the application of three log-linear models on sampled Y: the independence model, the no second order interaction (NSI) model, and the log-linear model chosen via the application of the

method proposed by Skinner and Shlomo (2008) (henceforth the SS approach). Independence models are rarely rich enough for disclosure risk estimation purposes and we include them here primarily for reference purposes. NSI models have been found to produce reasonable results in many cases (Fienberg and Makov, 1998; Elamir and Skinner, 2006; Skinner and Shlomo, 2008) and so represent a default modeling position. The SS approach is more complicated to employ than the Independence or NSI models, but the empirical evaluations of Skinner and Shlomo (2008) suggest that it outperforms both the Independence model and the NSI model; thus, it is the primary benchmark against which we will judge GoM model estimates.

Skinner and Shlomo (2008) present two criteria for evaluating log-linear models that are tuned to estimate disclosure risk quantities. The first—recommended by Skinner and Shlomo (2008) and used here—is to select the log-linear model that minimizes a standardized estimate of the bias of an estimator of $\tau_2 = \Sigma_{\mathbf{x}} E[1/F_{\mathbf{x}} | f_{\mathbf{x}} = 1]$ ($B_2/\sqrt{\nu}$, in the notation of Skinner and Shlomo, 2008). To find this model, we follow the SS approach by running a forward stepwise search procedure based on the minimization of $B_2/\sqrt{\nu}$, stopping when there is no evidence of underfitting. Further details including the sequence of fitted models, test statistics, and estimates of τ in the stepwise searches are provided in the online appendix. The second criterion uses a test statistic associated with τ ($B_1/\sqrt{\nu}$ in the notation of Skinner and Shlomo, 2008). We found similar results with either criterion and so report only results with $B_2/\sqrt{\nu}$.

For the GoM approach, analysts need to specify the number of extreme profiles *K*. In our experience, and as evident in Figure 1, as *K* increases the posterior medians of the disclosure risk quantities tend to stabilize on what can be considered the best GoM estimates. This behavior suggests an empirical method for choosing *K*. Starting from a simple model with a small number of extreme profiles, progressively fit more complex models and compute the posterior predictive estimate of τ for each of them. Once the estimates have stabilized around a particular value, choose the model with the smallest *K* that leads to that estimate. For computational expediency we recommend incrementing *K* by multiples of three to five in initial explorations, with greater multiples for larger *n*.

Following these guidelines, we selected models with $K = 6$, $K = 20$ and $K = 30$ for the samples with $n = 1000$, $n = 5000$, and $n = 10000$, respectively. Table 2 displays the estimates of τ for these models together with the results from the three log-linear models. The GoM model estimates are a clear improvement over those obtained through log-linear modeling. We note that past the point of stabilization, the exact number of extreme profiles does not make any practical difference. For instance, the estimates for the sample with $n = 10,000$ obtained from the selected model with $K = 30$ are essentially the same as those obtained with $K = 35$ and $K = 45$ and very similar to those obtained with $K = 40$. This is true for the individual measures of risk $\mu_{\mathbf{x}}$ as well.

To evaluate the performance of the GoM approach at the individual-record level, we consider the disclosure risk prediction as a classification problem: given a sample unique ($f_{\mathbf{x}} = 1$, observable) we wish to classify it as a population unique ($F_{\mathbf{x}} = 1$, not observable) or a

population multiple ($F_{\mathbf{x}} > 1$, not observable). A natural way of achieving this is to consider decision rules akin to those in record linkage problems (Fellegi and Sunter, 1969), e.g., a sample unique is classified as a population unique (a positive) when $\hat{\mu_{\mathbf{x}}} > \kappa_1$ and is classified as not a population unique (a negative) when $\hat{\mu_{\mathbf{x}}} < \kappa_2$, where $\kappa_2 < \kappa_1$. The cutoffs κ_1 and κ_2 are set by the agency to calibrate the misclassification rate to tolerable levels. A procedure for finding optimal values of (κ_1, κ_2) is an open problem that goes beyond the scope of the present work.

Figure 2 displays the fractions of false positives, (# of false positives)/(Sample uniques − τ), and false negatives, (# of false negatives)/τ, as a function of (κ_1, κ_2) for the SS log-linear and GoM models. For $n = 1,000$, both methods perform similarly, which is expected given the similarity of the estimates for τ in Table 2. For $n = 5,000$ and $n = 10,000$, the GoM models clearly outperform the SS log-linear models in terms of false negative rates no matter which κ_1 we choose, while the false positive rates are very similar. Arguably, false negatives are more problematic than false positives for organizations releasing data. Hence, for these samples, we conclude that the individual uniqueness assessment using the GoM approach offers improvements (in terms of accuracy of disclosure risk estimates) over the same task using log-linear models.

To illustrate the advantage of the GoM approach from another perspective, Figure 3 displays false negative rates implied by given false positive rates when setting $\kappa_1 = \kappa_2$, which forces every record to be classified. When $n = 1000$ both techniques again perform similarly. When $n = 5000$ or $n = 10000$, the implied false negative rate is always smaller for GoM models for any possible false positive level.

6 Discussion

The empirical evaluations show that using GoM models to estimate disclosure risk quantities can improve on previously suggested methods based on log-linear models, including what we consider to be the state of the art model selection technology of Skinner and Shlomo (2008). This improvement can be observed in both aggregated as well as individual measures of disclosure risk. Of course, any empirical evaluation is necessarily limited in breadth, and there may well be other settings where the GoM model does not perform as well as the best log-linear models.

As noted by Skinner and Shlomo (2008) and verified in our empirical work, different log-linear models can yield very different estimates of the disclosure risks. This applies for the class of GoM models as well: different numbers of extreme profiles yield different estimates of the measures of interest, typically monotonically decreasing as K increases. However, this decreasing trajectory appears to stabilize after reaching a specific threshold in the number of extreme profiles. Past that threshold, all GoM models with a larger number of extreme profiles yield essentially the same estimates, which are either comparable or a clear improvement over those obtained from log-linear smoothing. In the empirical evaluations, we found similar patterns of stabilization with K for posterior estimates of other quantities useful for disclosure risk estimation; see the online supplement for details. We note that

Manrique-Vallier and Fienberg (2008) observed related stabilizing behavior in the context of finite population size estimation with GoM models.

An important advantage of Bayesian implementations of GoM models over the use MLE-based log-linear estimates—which we believe helps explain the superior accuracy of GoM estimates in this context—is their tolerance to sparsity. In log-linear modeling, it is well known that certain patterns of zeros in contingency tables lead to the nonexistence of maximum likelihood estimates (see Rinaldo, 2005; Eriksson et al., 2006; Dobra et al., 2008); for example, patterns causing margins corresponding to the highest order terms in hierarchical log-linear models to have zeros. This is a common situation in datasets with a large number of variables with several levels, whereby the number of cells greatly exceeds that of the individual records. For instance, in the datasets used in Section 5, nearly all of the 3.6 million cells are empty; in fact, even in the most populated table we have 99.85% empty cells. This in turn causes many low order marginal tables—even some two-way tables—to have random zeros. Parameters for cells with those zero margins are inestimable, so that those cells have to be assumed to have expected values equal to zero (Bishop et al., 1975). However, this basically treats the responses corresponding to those cells as logical impossibilities, assigning them a probability of zero. In turn, this artificially increases the estimated probability of all other cells, thus inducing an under-estimation of the true risks (Skinner and Shlomo, 2008). Alternatively, one could consider only log-linear models that avoid the problematic cells—that is, ensure that no margin required for the model fitting has zeros—but this limits the range of available models and thus can engender bias due to over-smoothing. The Bayesian GoM model, in contrast, assumes that all cells have a positive probability regardless of the model complexity, so that it is more tolerant to sparsity.

A different situation arises when some cells in the contingency table correspond to impossible combinations of variables, so that their expected counts must be exactly zero. These structural zeros are easy to account for during maximum likelihood estimation of hierarchical log-linear models using algorithms such as Iterative Proportional Fitting (see Bishop et al., 1975). No such direct approaches exist for models based on conditional exchangeability like GoM models. Our current implementation of the Bayesian GoM model does not treat structural zeros differently from random zeros. This can result in biased estimates in the presence of many structural zeros. Hence, for the moment, we recommend that agencies use the GoM approach only for datasets, or subsets of datasets, for which structural zeros are not present. For example, in the California PUMS data we estimated disclosure risks only for adults, so that the (massive) numbers of zeros in the sample tables are all random and not structural. We currently are working on extensions to the GoM MCMC sampler algorithms to account for structural zeros, generalizing ideas from Manrique-Vallier and Fienberg (2008) that we believe can be directly applied to the problem of estimating population uniques.

Fitting the GoM models via MCMC can be computationally intensive, especially with large samples and large numbers of extreme profiles. In the empirical study, we typically needed around 400, 000 MCMC iterations before reaching stationary distributions, after which long runs—typically around 300,000 iterations—were needed to obtain adequate posterior samples due to high autocorrelations in the chains. This contrasts with the application of log-

linear models, for which efficient MLE algorithms like Iterative Proportional Fitting exist. With the most demanding dataset (n = 10, 000), our personal computers required around 21 hours to fit a single GoM model. The entire stepwise model selection with log-linear models took around 9 hours. We sped up the GoM model selection (i.e., fit GoM for multiple values of K) by running the required chains in parallel, using idle capacity on networked computers. For practical applications with similarly large datasets, we recommend running the MCMC chains in parallel, taking advantage of multicore processors and computer clusters.

Given the disparity in model complexity and running times, a referee of this article questioned whether or not the gains in accuracy from the GoM approach are worth the effort. While this judgment obviously sits with the agency estimating disclosure risks, in many cases we believe that the answer should be affirmative. Accurate disclosure risk estimation is important: significant underestimation of risk could lead to releases with unnecessarily high risks of disclosure, and significant overestimation could lead to unnecessary disclosure treatment. Additionally, in many studies, it can take many months and even years to complete the cycle from data collection to data dissemination, including time for quality checking, preparing data documentation, etc. With computing times on the order of just several hours, we suspect that many agencies would be more than willing to let a computer run longer to realize the accuracy gains in risk estimation apparent from the GoM approach.

We investigated a method for approximate inference with the GoM approach that is less computationally intensive based on mean field variational techniques. Similar methods have been successfully applied in other large scale, mixed membership settings (e.g., Blei et al., 2003; Blei and Lafferty, 2007; Airoldi et al., 2007). The estimates of disclosure risk quantities obtained from the variational approximations were unreliable, often being very far from the true values of τ. We believe that this is due to the fact that these variational techniques rely on a global approximation to the likelihood using a surrogate distribution—which minimizes the KL-divergence with the target distribution within a tractable family—that does not necessarily represent well all regions of the target likelihood, particularly with modest sample sizes. This is especially problematic for disclosure risk estimation, since the relevant measures depend on accurately estimating cells with very small probability. Hence, we do not generally recommend variational approximations for this setting.

Our application of the GoM model to the estimation of disclosure risk quantities assumes simple random sampling. It must be modified for samples obtained through complex designs. For stratified simple random sample designs with just a few strata, one way of accounting for the design is to introduce the strata labels as covariate information through the population level common distribution of membership scores, similar to the method for incorporating cohort information in Manrique-Vallier (2010). Alternatively, when sample sizes are large enough within strata, GoM models can be estimated separately in each stratum. For more general complex designs, Bertolet (2008) has proposed a set of extensions to the basic GoM model with binary responses that we believe can be extended for multinomial responses.

As a final remark, we note that improved methodology for disclosure risk estimation in simple random samples has benefits for broader data collection contexts. For example, determining population uniqueness is a major concern of government agencies considering the release of data collected without random sampling, such as administrative records and disease registries (henceforth all called registry data). These agencies seek to prevent intruders from identifying individuals in the registry by matching on known keys. However, since registry data typically are not representative of the whole population, agencies cannot simply fit log-linear or GoM models to the registry data and expect to get reasonable estimates of disclosure risk quantities. Instead, as suggested by Yu et al. (2011), agencies can identify large surveys with keys that are also in the registry data, like the Census Public Use Microdata Samples (PUMS) and American Community Survey (ACS), and estimate population counts for the table of common keys. The agency then can determine how many and which records in the registry match on the keys corresponding to the estimated population uniques. GoM models can be readily applied in this setting, since the PUMS and ACS are essentially stratified simple random samples. Since the tables of survey data for the specific registry sites will be large and sparse, we anticipate that GoM models would offer more accurate estimates of population uniqueness than log-linear models, and hence lead to better decisions about what registry data can be released to the public.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by grants from the National Institutes of Health (R21 AG032458-02) and National Science Foundation (SES-11-31897).

References

Airoldi EM, Blei DM, Fienberg SE, Xing EP. Mixed membership stochastic blockmodels. The Journal of Machine Learning Research. 2008; 9:1981–2014.

Airoldi EM, Fienberg SE, Joutard C, Love TM. Discovering Latent Patterns with Hierarchical Bayesian Mixed-Membership Models. Data Mining Patterns: New Methods and Applications. 2007:240–275.

Bertolet, M. PhD thesis. Carnegie Mellon University; 2008. To Weight Or Not To Weight? Incorporating Sampling Designs Into Model-Based Analyses.

Bethlehem JG, Keller WJ, Pannekoek J. Disclosure control of microdata. Journal of the American Statistical Association. 1990; 85:38–45.

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

Blei DM, Lafferty JD. A correlated topic model of *Science*. Annals of Applied Statistics. 2007; 1:17–35.

Blei DM, Ng A, Jordan MI. Latent Dirichlet allocation. Journal of Machine Learning Research. 2003; 3:993–1022.

Chen G, Keller-McNulty S. Estimation of Identification Disclosure Risk in Microdata. Journal of Official Statistics. 1998; 14:79–95.

Cooil B, Varki S. Using the conditional Grade-of-Membership model to assess judgment accuracy. Psychometrika. 2003; 68:453–471.

Dale A, Elliot M. Proposals for 2001 samples of anonymized records: an assessment of disclosure risk. Journal of the Royal Statistical Society, Series A. 2001; 164:427–447.

Dobra, A.; Fienberg, SE.; Rinaldo, A.; Slavkovic, AB.; Zhou, Y. Emerging Applications of Algebraic Geometry. New York: Springer; 2008. p. 63-88.chap. Algebraic statistics and contingency table problems: Log-linear models, likelihood estimation, and disclosure limitation., IMA Series in Applied Mathematics

Drechsler, J.; Reiter, JP. Accounting for intruder uncertainty due to sampling when estimating identification disclosure risks in partially synthetic data. In: Domingo-Ferrer, J.; Saygin, Y., editors. Privacy in Statistical Databases (LNCS 5262). New York: Springer-Verlag; 2008. p. 227-238.

Duncan, GT.; Elliott, M.; Salazar-Gonzalez, JJ. Statistical Confidentiality: Principles and Practice. Berlin: Springer; 2011.

Elamir E, Skinner CJ. Record level measures of disclosure risk for survey microdata. Journal of Official Statistics. 2006; 22:525–539.

Eriksson N, Fienberg SE, Rinaldo A, Sullivant S. Polyhedral conditions for the nonexistence of the MLE for hierarchical log-linear models. Journal of Symbolic Computation. 2006; 41:222–233.

Erosheva, E. PhD thesis. Department of Statistics. Carnegie Mellon University; 2002. Grade of membership and latent structures with application to disability survey data.

Erosheva E. Comparing Latent Structures of the Grade of Membership, Rasch, and Latent Class Models. Psychometrika. 2005; 70:619–628.

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

Erosheva E, Fienberg S, Junker B. Alternative statistical models and representations for large sparse multi-dimensional contingency tables. Annales de la faculté des sciences de Toulouse Sér 6. 2002; 11:485–505.

Erosheva E, Fienberg SE, Lafferty JD. Mixed-membership models of scientific publications. Proceedings of the National Academy of Sciences. 2004; 101:5220–5227.

Fellegi IP, Sunter AB. A theory for record linkage. Journal of the American Statistical Association. 1969; 64:1183–1210.

Fienberg SE, Makov UE. Confidentiality, uniqueness, and disclosure limitation for categorical data. Journal of Official Statistics. 1998; 14:361–372.

Forster J, Webb E. Bayesian disclosure risk assessment: predicting small frequencies in contingency tables. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2007; 56:551–570.

Gelman, A.; Carlin, JB.; Stern, HS.; Rubin, DB. Bayesian Data Analysis. London: Chapman & Hall; 2004.

Goodman LA. Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika. 1974; 61:215–231.

Gormley, C. PhD thesis. Department of Statistics, University of Dublin, Trinity College; 2006. Statistical Models For Rank Data.

Gormley I, Murphy T. A mixture of experts model for rank data with applications in election studies. The Annals of Applied Statistics. 2008; 2:1452–1477.

Greenberg BV, Zayatz LV. Strategies for measuring risk in public use microdata files. Statistica Neerlandica. 1992; 46:33–48.

Haberman SJ. Review: Statistical applications using fuzzy sets, by K. Man-ton, M. Woodbury and H. Tolley. Journal of the American Statistical Association. 1995; 90:1131–1133.

Holland PW, Rosenbaum PR. Conditional association and unidimensionality in monotone latent variable models. Annals of Statistics. 1986; 14:1523–1543.

Manrique-Vallier, D. PhD thesis. Department of Statistics, Carnegie Mellon University; 2010. Logitudinal Mixed Membership Models with Applications to Survey Disability Data.

Manrique-Vallier D, Fienberg S. Population size estimation using individual level mixture models. Biometrical Journal. 2008; 50:1051–1063. [PubMed: 19035548]

Manton, KG.; Woodbury, MA.; Tolley, HD. Statistical Applications Using Fuzzy Sets. New York: Wiley; 1994.

Pannekoek J. Statistical methods for some simple disclosure limitation rules. Statistica Neerlandica. 1999; 53:55–67.

Pritchard JK, Stephens M, Donnelly P. Inference of Population Structure Using Multilocus Genotype Data. Genetics. 2000; 155:945–959. [PubMed: 10835412]

Reiter JP. Estimating identification risks in microdata. Journal of the American Statistical Association. 2005; 100:1103–1113.

Reiter JP, Raghunathan TE. The multiple adaptations of multiple imputation. Journal of the American Statistical Association. 2007; 102:1462–1471.

Rinaldo, A. PhD thesis. Department of Statistics, Carnegie Mellon University; 2005. Maximum likelihood estimates in large sparse contingency tables.

Rinott, Y.; Shlomo, N. Variances and confidence intervals for sample disclosure risk measures. Proceedings of the 56th Session of the ISI; ISI, Lisbon. 2007. p. 22-29.

Rubin DB. Discussion: Statistical disclosure limitation. Journal of Official Statistics. 1993; 9:462–468.

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

Samuels SM. A Bayesian species-sampling-inspired approach to the uniques problem in microdata. Journal of Official Statistics. 1998; 14:373–384.

Shlomo N, Skinner CJ. Assessing the protection provided by misclassification-based disclosure limitation methods for survey microdata. Annals of Applied Statistics. 2010; 4:1291–1310.

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

Skinner C, Holmes D. Estimating the re-identification risk per record in microdata. Journal of Official Statistics. 1998; 14:361–372.

Skinner C, Marsh C, Openshaw S, Wymer C. Disclosure control for census microdata. Journal of Official Statistics. 1994; 10:31–51.

Skinner CJ. On identification disclosure and prediction disclosure for microdata. Statistica Neerlandica. 1992; 46:21–32.

Skinner CJ, Shlomo N. Assessing identification risk in survey microdata using log-linear models. Journal of the American Statistical Association. 2008; 103:989–1001.

Sweeney, L. PhD thesis. Massachusetts Institute of Technology; 2001. Computational Disclosure Control: Theory and Practice.

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

Yu, M.; Stinchcomb, D.; Cronin, K. Disclosure risk assessment for population-based cancer microdata. JSM Proceedings, Survey Research Methods Section; Miami Beach, FL: American Statistical Association; 2011. p. 2609-2622.

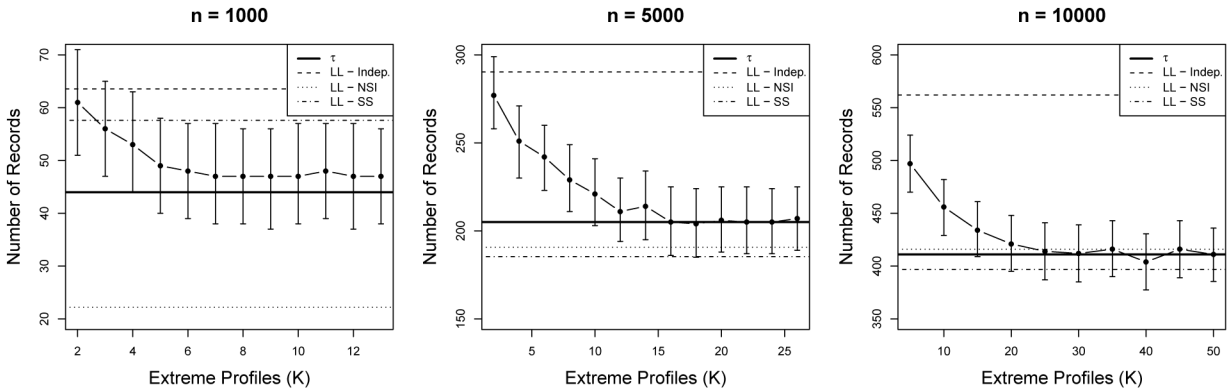

Figure 1.

Posterior medians of τ obtained from GoM models for different K for the three samples, along with error bars indicating 95% equal-tail posterior predictive credible intervals. Also included are the actual value of τ and the estimates obtained from the application of log-linear models using independence (Indep), no second order interactions (NSI), and the Skinner and Shlomo (2008) model selection criterion (SS).

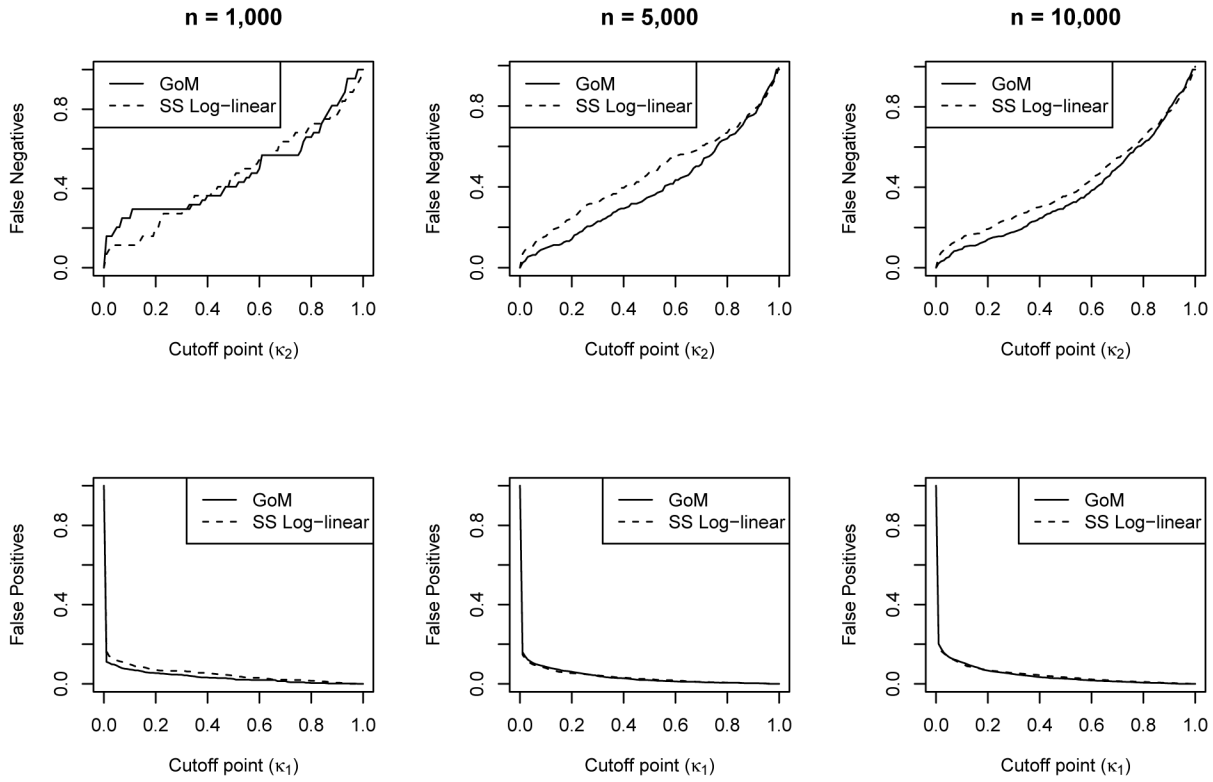

Figure 2.

Fraction of false positives and false negatives in the three samples for classifications obtained with SS log-linear and GoM models as a function of the classification cutoff points.

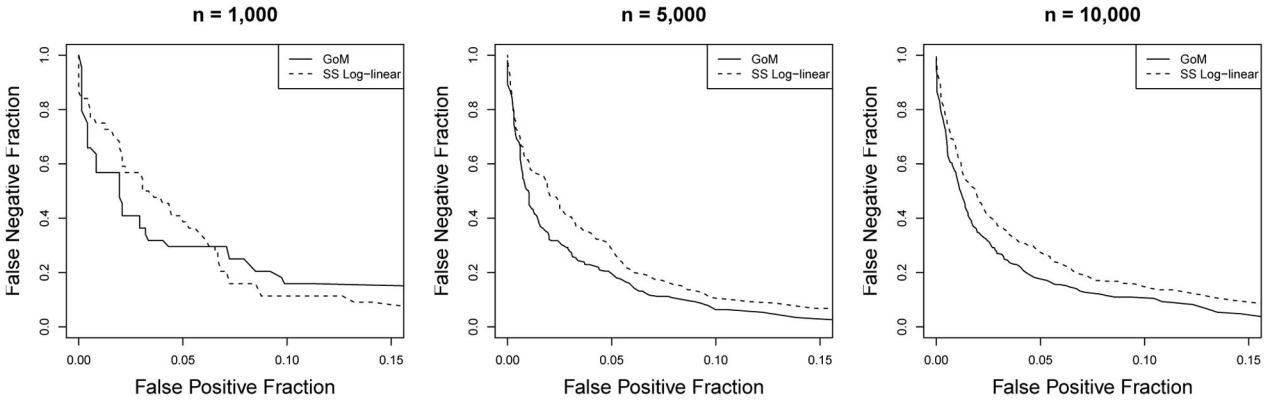

Figure 3.

Fraction of false positives vs. fraction of false negatives in the three samples for full classifications ($\kappa_1 = \kappa_2$) obtained with SS log-linear and GoM models.

Table 1

Three samples from the CA file. The number of individuals in the population is $N = 1, 150, 934$, and the total number of cells in the contingency tables is 3, 600, 000.

n	Populated Cells	Sample Uniques	τ
1, 000	863 (.02%)	763	44
5, 000	3,230 (.09%)	2,518	205
10, 000	5,478 (.15%)	4,015	411

Table 2

Estimates of the number of sample uniques that are also population uniques using log-linear models and GoM models for the three samples. Point estimates for GoM models are posterior medians; 95% intervals are equal tail posterior credible intervals.

n	Model	τ	$\hat{\tau}$	95%-Interval
1,000	GoM (K = 6)	44	48	[39, 57]
	Log-linear - Indep.		63.6	
	Log-linear - NSI		22.1	
	Log-linear - SS[1]		57.6	
5,000	GoM (K = 20)	205	206	[188, 225]
	Log-linear - Indep.		289.5	
	Log-linear - NSI		190.2	
	Log-linear - SS[2]		185.4	
10,000	GoM (K = 30)	411	412	[385, 439]
	Log-linear - Indep.		559.2	
	Log-linear - NSI		414.6	
	Log-linear - SS[3]		396.9	

[1] Indep. + [BD]

[2] NSI + [DGH]

[3] NSI + [BEH][DGH]