Statistics and Computing (2025) 35:158
https://doi.org/10.1007/s11222-025-10693-9

ORIGINAL PAPER

Disclosure risk assessment with Bayesian non-parametric hierarchical modelling

Marco Battiston[1] · **Lorenzo Rimella**[2]

Received: 3 February 2025 / Accepted: 22 July 2025

Abstract
Micro and survey datasets often contain private information about individuals, like their health status, income, or political preferences. Previous studies have shown that, even after data anonymization, a malicious intruder could still be able to identify individuals in the dataset by matching their variables to external information. Disclosure risk measures are statistical measures meant to quantify how big such a risk is for a specific dataset. One of the most common measures is the number of sample unique values that are also population unique. Mixed membership models can provide very accurate estimates of this measure. A limitation of this approach is that the number of extreme profiles has to be chosen by the modeller. In this article, we propose a non-parametric version of the model, based on the Hierarchical Dirichlet Process (HDP). The proposed approach does not require any tuning parameter or model selection step and provides accurate estimates of the disclosure risk measure, even with samples as small as 1% of the population size. Moreover, a data augmentation scheme to address the presence of structural zeros is presented. The proposed methodology is tested on a real dataset from the New York microdata.

Keywords Disclosure risk measures · Contingency tables · Privacy · Latent class models · Bayesian non-parametrics · Hierarchical Dirichlet process

1 Introduction

Statistical agencies routinely collect and disseminate to the public record-level and microdata on individual persons and businesses. This data may contain private information about individuals, like their income, political or sexual preferences, or health conditions. This creates a serious concern for privacy breaches and the need to protect individuals' anonymity. Previous studies have shown that, even after removing names from the data, a malicious intruder could still be able to identify individuals by matching some of their variables in the dataset to external data. Indeed, in a famous example, Sweeney (2001) was able to identify 97% of the records in a voter registration list by using just their birth date and zip code.

Disclosure risk assessment refers to a broad range of statistical techniques that can be used to assess whether record-level or file-level data has to be considered at risk of disclosing private information. A popular disclosure risk measure, proposed by Skinner et al. (1994), is the number of sample unique values that are also population unique, denoted by τ_1. Individuals having a rare or unique combination of values in some variables in the dataset are those most at risk of identification because if their variables are matched using another dataset, it results in a perfect match. Therefore, if the estimated measures of disclosure risk are high, additional privacy-preserving techniques, like, for example, variable anonymization, data swapping, and addition of noise or cell suppression, should be applied to the dataset before its release to the public.

Among the most popular models to estimate τ_1 or similar disclosure risk measures are log-linear, Skinner and Shlomo (2008), and mixed-membership (also referred to as grade of membership) models, Manrique-Vallier and Reiter (2012). Log-linear models are computationally very efficient, but their estimates may deteriorate with the presence of many

Marco Battiston and Lorenzo Rimella contributed equally to this work.

✉ Marco Battiston
m.battiston@lancaster.ac.uk

Lorenzo Rimella
lorenzo.rimella@unibg.it

[1] School of Mathematical Sciences, Lancaster University, Lancaster LA1 4YF, UK

[2] Department of Economics, Università degli Studi di Bergamo, Bergamo 24127, Italy

Published online: 29 July 2025

structural zeros. The mixed membership models, as proposed in Manrique-Vallier and Reiter (2012), seem to provide very accurate estimates of τ_1, even with samples as small as 1% or less of the entire population. However, a limitation of this model is the practitioner needs to select a number of extreme profiles K to use for a specific dataset. This has two drawbacks. Firstly, it affects the running time of the methodology, since the model needs to be fitted for different values of K to evaluate differences in the estimates of τ_1. Secondly, the choice of a suitable K depends on the value of τ_1, which in real data scenarios is not available to the practitioner.

In this article, we propose a non-parametric version of the mixed-membership model of Manrique-Vallier and Reiter (2012) to perform disclosure risk assessment. The proposed model is formulated as a Hierarchical Dirichlet Process, Teh et al. (2006), and allows a potentially unbounded number of extreme profiles. This number is then estimated directly from the data, hence resulting in a tuning-free modelling approach. We describe how to estimate τ_1 within the MCMC using both a population sampling approach, as in Manrique-Vallier and Reiter (2012), and a much faster Monte Carlo approximation, which can speed up the computational cost of the algorithm substantially.

A common problem with modelling contingency tables is the presence of many structural zeros. These are combinations of categorical variables that lead to impossible values, i.e. values that are known to be zero in the population, for example, a pregnant male. In real-data applications, structural zeros can account for a very large proportion of possible combinations, and, if their presence is not properly accounted for in the statistical analysis, the performance of the model can deteriorate dramatically. In this article, we also describe how to extend the proposed non-parametric model to deal with the presence of many structural zeros, following the data augmentation idea presented in Manrique-Vallier and Reiter (2014). This latter paper presents an approach to handle structural zeros in tabular data, with an application in disclosure risk using a mixture model, in which the dimension K is fixed, see also (Manrique-Vallier and Hu 2018).

To sum up, the paper is organized as follows. Section 2 reviews the disclosure risk problem and introduces the disclosure risk measure τ_1. Section 3 presents the non-parametric generalization of the mixed membership model of Manrique-Vallier and Reiter (2012), using the Hierarchical Dirichlet Process, and describes the Markov Chain Monte Carlo (MCMC) algorithm to make inference on the model parameters and to estimate τ_1. The extension of the model to include structural zeroes is discussed in Section 4. In Section 5, some empirical illustrations are presented to show the performance of the proposed methodology on synthetic and a real-data example with New York microdata from the American Community Survey. Finally, some background material, derivations, additional information on the experiments, and more experiments are included in the Online Supplementary Material.

2 Disclosure Risk Problem

Disclosure risk problems for record-level data usually involve two distinct classes of variables: 1) one set of variables, usually called *sensitive variables*, that contain private information, e.g. health status or salary; 2) another class of identifying categorical variables, usually called *key variables*, e.g. gender, age, job, and more general demographic information. Disclosure risk arises because a malicious intruder could potentially identify individuals in the dataset by cross-classifying their key variables and matching them to some external source of information, like publicly available census data. If these matches are correct, the intruder will be able to identify individuals' identities and disclose information contained in their sensitive variables. Disclosure risk measures are statistical measures that try to quantify how easy it is to identify individuals based on the values of their key variables.

In order to formalize the problem, let us assume that J categorical key variables in the dataset have been observed for a sample of n individuals, sampled from a population of size N. The j-th key variable has n_j possible categories, labeled, without loss of generality, from 1 up to n_j. Focusing only on key variables, observation for individual i, denoted $X_i = (X_{i1}, \ldots, X_{iJ})$, therefore takes values in the state space $\mathcal{C} := \times_{j=1}^{J}\{1 \ldots, n_j\}$. This set has $|\mathcal{C}| = \prod_{j=1}^{J} n_j$ values, corresponding to all possible cross-classification of the J key variables. Information about the sample is usually given through the sample frequency vector $(f_1, \ldots, f_{|\mathcal{C}|})$, where f_c counts how many individuals out of the n in the sample have a particular combination of cross-classified key variables, corresponding to cell $c \in \mathcal{C}$. $(F_1, \ldots, F_{|\mathcal{C}|})$ denotes the corresponding vector of frequencies in the whole population of N individuals, i.e. F_c is the number of individuals in the population belonging to cell c.

The earliest papers to consider disclosure risk problems include (Bethlehem et al. 1990; Duncan and Lambert 1986, 1989; Lambert 1993). These works propose different measures of disclosure risk and ways to estimate them, under specific model assumptions for $(f_1, \ldots, f_{|\mathcal{C}|})$ and $(F_1, \ldots, F_{|\mathcal{C}|})$. Skinner and Elliot (2002), Skinner et al. (1994) provide reviews of the most popular measures of disclosure risk. Disclosure risk measures depend on the sample frequencies $(f_1, \ldots, f_{|\mathcal{C}|})$ and often focus on small frequencies, especially on cells having frequency 1, called *sample uniques*. Individuals belonging to these cells are those at the highest risk of having their sensitive information disclosed. This is because if any of these sample unique values are also unique values in the population, called *population*

uniques, any match of their key variables with information from another dataset will produce a perfect match, i.e. perfect certainty about the identity of that specific record, and their sensitive information will be therefore disclosed. For a review of disclosure risk problems, the reader is referred to Matthews and Harel (2011).

We usually distinguish between two groups of *measures of disclosure risk*:

1. **Record-Level** (or per-record) measures: they assign a measure of risk to each data point or specific cell values. Among the most popular ones, there are

$$r_{1c} = \mathbb{P}(F_c = 1 | f_c = 1), \quad r_{2c} = \mathbb{E}(1/F_c | f_c = 1). \quad (1)$$

   $c \in \{1, \ldots, |\mathcal{C}|\}$. The first measure provides the probability that a sample unique is also population unique. The second one gives the probability that, given a sample unique c, we guess her identity correctly, by choosing one of the F_c values in the population uniformly at random. In general, the first measure is less conservative and is always smaller than the second.
2. **File-level** measures: they provide an overall measure of risk for an entire sample or dataset. File-level measures are usually defined by aggregating the corresponding record-level ones. Popular examples are

$$\tau_1 = \sum_{c \in \mathcal{C}: f_c = 1} r_{1c}, \quad \tau_2 = \sum_{c \in \mathcal{C}: f_c = 1} r_{2c}. \quad (2)$$

In the disclosure risk literature, τ_1 is a popular measure of disclosure risk, Bethlehem et al. (1990), Skinner et al. (1994), Skinner and Shlomo (2008), Skinner and Shlomo (2008), Carota et al. (2015), Carota et al. (2022), Manrique-Vallier and Reiter (2012), Manrique-Vallier and Reiter (2014), Reiter (2005), Rinott and Shlomo (2006) and, in the rest of the paper, we will focus on its estimation using the data $(f_1, \ldots, f_{|\mathcal{C}|})$. In the literature, the most popular modelling choices for this task are *log-linear* and *mixed membership* models. Regarding the former ones, the main references are (Skinner and Shlomo 2008), Shlomo and Skinner (2010), in which indexes (1) and (2) are derived in closed form and estimated using plug-in MLE estimators. Regarding the latter class of models, Manrique-Vallier and Reiter (2012) proposed the use of mixed membership models, which resulted in very accurate estimates for (2), even for sample sizes n much smaller than the population size N.

If the estimated values of (1) and (2) are too high, then the data curator should apply a disclosure limitation technique to the dataset before releasing it to the public. Some possibilities are, for example, rounding, data swapping, cell suppression of extreme values or entire variables, subsampling, or perturbation techniques. See (Willenborg and De Waal 2012) for a review of different disclosure limitation techniques.

3 Mixed Membership models

In this section, we extend the mixed membership model of Manrique-Vallier and Reiter (2012), reviewed in the Online Supplementary Material, to its non-parametric version. Then, we summarize both the MCMC sampler to perform posterior inference and describe how to estimate (2) within the sampler using either population sampling, as in Manrique-Vallier and Reiter (2012), or a faster Monte Carlo approximation.

In terms of background about the Hierarchical Dirichlet Process and its properties (Stick Breaking construction, Chinese Restaurant Franchise, Posterior representation), the reader is referred to Teh and Jordan (2010)

3.1 Non-parametric Mixed Membership Model

Mixed Membership models are generalizations of mixture models to model multiple groups of observations. In their parametric version, they assume K extreme profiles (alias mixture components), having weights in the population regulated by a K-dimensional probability vector $\boldsymbol{g}_0$. Within each group, some heterogeneity from the common proportions $\boldsymbol{g}_0$ is allowed by introducing a group-specific partial affiliation vector $\boldsymbol{g}_i$. In the model used in Manrique-Vallier and Reiter (2012), the i-th group of observations corresponds to the J observations of key variables of the i-th individual.

In order to allow an unbounded number of extreme profiles, we select $G_0 \sim \text{DP}(\alpha_0, H)$, where DP is a Dirichlet Process, Ferguson (1973), with concentration parameter α_0 and base measure H. The base measure H is a probability measure on the space of all arrays with J rows, having a n_j-dimensional probability vector in the j-th row. From the stick-breaking representation of the Dirichlet Process, G_0 can be represented as

$$G_0 = \sum_{k=1}^{\infty} g_{0,k} \delta_{\theta^{(k)}}$$

$\left(\theta^{(k)}\right)_{k=1}^{\infty}$ are independent and identically distributed arrays, sampled from the base measure H, representing the likelihood of the possibly unbounded extreme profiles, while the sequence $\left(g_{0,k}\right)_{k=1}^{\infty}$ is such that all entries $0 \leqslant g_{0,k} \leqslant 1$ and $\sum_k g_{0,k} = 1$, and is sampled following a stick breaking distribution of parameter α_0, Sethuraman (1994). As in the parametric case, $g_{0,k}$ can be thought of as the popularity of extreme profile $\theta^{(k)}$ in the population.

Given G_0, each individual i selects her own affiliation distribution G_i, representing her partial affiliation to each

possible extreme profile, according to $G_i|G_0 \sim \text{DP}(\alpha_i, G_0)$. Given the almost sure discreteness of G_0, each G_i is supported on the same atoms of G_0 and can be represented as

$$G_i = \sum_{k=1}^{\infty} g_{i,k}\delta_{\theta^{(k)}}$$

for a sequence of probability weights $\left(g_{i,k}\right)_{k=1}^{\infty}$, see pages 161-162 of Teh and Jordan (2010). The parameter α_i regulates the variability of the weights $\left(g_{i,k}\right)_{k=1}^{\infty}$ around their mean value $\left(g_{0,k}\right)_{k=1}^{\infty}$. The higher α_i, the more heterogeneous individual i is from the rest of the population.

Given the individual specific affiliation vector G_i, individual i will select her j-th key variable from the infinite mixture model

$$X_{i,j}|G_i \sim \sum_{k=1}^{\infty} g_{i,k}\theta_{j,\cdot}^{(k)}$$

where $\theta_{j,\cdot}^{(k)}$ denotes the j-th row of $\theta^{(k)}$.

As in the finite-dimensional case, it is computationally convenient to introduce the mixture classification variables $Z_{i,j}$, taking integer values, and summarize the model as follows,

$$\begin{aligned}
X_{i,j}|\left(Z_{i,j}=k\right), \left(\theta_k\right)_{k=1}^{\infty} &\sim \theta_{j,\cdot}^{(k)} \quad i=1,\ldots,n, \quad j=1,\ldots,J \\
\mathbb{P}\left(Z_{i,j}=k|G_i\right) &= g_{i,k} \quad k\in\mathbb{N}, \quad i=1,...,n, \quad j=1,...,J \\
G_i|\alpha, G_0 &\sim \text{DP}(\alpha_i, G_0) \quad i=1,\ldots,n \\
G_0 &\sim \text{DP}(\alpha_0, H) \\
\alpha_i &\sim \text{Ga}(a,b) \quad i=1,\ldots,n \\
\alpha_0 &\sim \text{Ga}(a_0,b_0).
\end{aligned}$$

where the base measure H is chosen to assign $\text{Dir}\left(\mathbb{I}_{n_j}\right)$ prior to the j-th row, for each $j \in \{1,\ldots,J\}$, where $\mathbb{I}_{n_j}$ is a vector of dimension n_j with all entries equal to 1. Finally, Ga denotes a Gamma distribution, and a, b, a_0, b_0 are positive hyperparameters. In all the experiments of Section 5, the hyperparameters are set $a = a_0 = 2$ and $b = b_0 = 1$.

3.2 Posterior Inference

3.2.1 MCMC sampler

Posterior inference of the model parameters can be performed using the Direct Assignment algorithm for the Hierarchical Dirichlet Process, pages 196-199 of Teh and Jordan (2010). In the sampler, m_{ik} denotes the number of tables in individual i assigned to mixture component k. At any stage of the algorithm, we denote by K_n the number of active mixture components, i.e. components $\theta^{(k)}$ with at least one $Z_{i,j}$ assigned to them. At step 3, the sampler resamples $\left(g_{0,k}\right)_{k=1}^{\infty}$, by drawing a probability vector $\left(g_{0,k}\right)_{k=0}^{K_n}$ (using the posterior representation of G_0, formula 5.9 in Teh and Jordan (2010)), where $g_{0,0}$ represent the probability of a new mixture, i.e. $g_{0,0} = 1 - \sum_{k=1}^{K_n} g_{0,k}$. Similarly for $\left(g_{i,k}\right)_{k=1}^{\infty}$ at step 4. For ease of notation, we will simply write $\left(g_{0,k}\right)$ and $\left(g_{i,k}\right)$, where the index is over k and ranges from 0 to K_n. The sampler iterates over the following steps.

1. Sample $Z_{i,j}$: for $i \in \{1,\ldots,n\}$ and $j \in \{1,\ldots,J\}$, sample $Z_{i,j}$ from

$$Z_{i,j} = \begin{cases} k & \text{with prob} \quad \propto \quad g_{i,k}\theta_{j,x_{ij}}^{(k)} \\ k^{\text{new}} & \text{with prob} \quad \propto \quad g_{i,0}\frac{1}{n_j} \end{cases} \tag{3}$$

for $k \in \{1,\ldots,K_n\}$, where the factor $1/n_j$ is the marginal probability of $Z_{i,j}$ being sampled from a new mixture $\theta^{(K_n+1)}$, when $\theta^{(K_n+1)}$ is distributed according to H.
If $Z_{i,j} = k^{\text{new}}$, draw $\theta^{(k^{\text{new}})}$ from (6), and update $\left(g_{0,k}\right)$ and $\left(g_{i,k}\right)$ as follows

$$\begin{aligned}
\nu_0|\alpha_0 &\sim \text{Beta}(\alpha_0, 1) \\
\left(g_{0,0}^{\text{new}}, g_{0,K_n+1}^{\text{new}}\right) &= \left(g_{0,0}\nu_0, g_{0,0}(1-\nu_0)\right) \\
\nu_i|g_{0,0}, \alpha, \nu_0 &\sim \text{Beta}\left(\alpha g_{0,0}\nu_0, \alpha g_{0,0}(1-\nu_0)\right) \\
\left(g_{i,0}^{\text{new}}, g_{i,K_n+1}^{\text{new}}\right) &= \left(g_{i,0}\nu_i, g_{i,0}(1-\nu_i)\right)
\end{aligned}$$

for every $i = 1,\ldots,n$. Finally, set $Z_{i,j} = K_n + 1$ and increment K_n by 1.
2. Sample m_{ik}: for $i \in \{1,\ldots,n\}$ and $k \in \{1,\ldots,K_n\}$, compute $n_{i\cdot k} = \sum_{j=1}^{J} \mathbb{I}\left(Z_{i,j}=k\right)$, and sample m_{ik} from

$$\mathbb{P}\left(m_{ik}=m|n_{i\cdot k}, g_{0,k}, \alpha_0\right) = \frac{\Gamma\left(\alpha_0 g_{0,k}\right)}{\Gamma\left(\alpha_0 g_{0,k}+n_{i\cdot k}\right)} s\left(n_{i\cdot k}, m\right)\left(\alpha_0 g_{0,k}\right)^m$$

for $m \in \{1,\ldots,n_{i\cdot k}\}$, and where $s(n,m)$ are the unsigned Stirling numbers of the first kind, which can be pre-computed outside the sampler from the recursion, $s(0,0) = s(1,1) = 1$, $s(n,0) = 0$ for $n > 0$ and $s(n,m) = 0$ for $m > n$ and $s(n+1,m) = s(n,m-1) + n s(n,m)$.
As an alternative, m_{ik} can also be computed by drawing a Chinese Restaurant Process with $n_{i\cdot k}$ customers and concentration parameter $\alpha_0 g_{0,k}$, and setting m_{ik} equal to the number of resulting tables. This approach is incredibly fast when the number J of categorical variables is small, and it is the approach we considered in the experiments.
3. Sample $\left(g_{0,k}\right)$: compute $m_{\cdot k} = \sum_{i=1}^{n} m_{ik}$ for $k = 1,\ldots,K_n$, and resample $\left(g_{0,k}\right)$ from

$$\text{Dir}\left(\alpha_0, m_{\cdot 1}, \ldots, m_{\cdot K_n}\right) \tag{4}$$

4. Sample $(g_{i,k})$: for $i \in \{1, \ldots, n\}$, resample $(g_{i,k})$ from

$$\text{Dir}\left(\alpha_i g_{0,0}, \alpha_i g_{0,1} + n_{i\cdot 1}, \ldots, \alpha_i g_{0,K_n} + n_{i\cdot K}\right) \quad (5)$$

5. Sample $\theta^{(k)}$: for $k \in \{1, \ldots, K_n\}$ and $j \in \{1, \ldots, J\}$ sample $\theta^{(k)}_{j,\cdot}$ according to,

$$\text{Dir}\left(1 + \sum_{i=1}^{n} \mathbb{I}\left(Z_{ij} = k, X_{ij} = 1\right), ..., 1 + \sum_{i=1}^{n} \mathbb{I}\left(Z_{ij} = k, X_{ij} = n_j\right)\right) \quad (6)$$

6. Sample α_0, α_i: Using the augmentation from the Appendix of Teh et al. (2006), let $m_{\cdot\cdot} = \sum_{k=1}^{K_n} m_{\cdot k}$, then sample α_0 according to

$$\begin{aligned}
\eta_0|\alpha_0, m_{\cdot\cdot} &\sim \text{Beta}\,(\alpha_0 + 1, m_{\cdot\cdot}) \\
s_0|m_{\cdot\cdot}, \eta_0, K_n &\sim \text{Bern}\left(\frac{m_{\cdot\cdot}\,(b_0 - \log \eta_0)}{K_n + a_0 - 1 + m_{\cdot\cdot}\,(b_0 - \log \eta_0)}\right) \\
\alpha_0|\eta_0, s_0, K_n &\sim \text{Gamma}\,(a_0 + K_n - s_0, b_0 - \log \eta_0)
\end{aligned}$$

and α_i, for $i \in \{1, \ldots, n\}$ according to

$$\begin{aligned}
\eta_i|\alpha_i, J &\sim \text{Beta}(\alpha_i + 1, J) \quad i = 1, ..., n \\
s_i|m_{i\cdot}, \eta_i &\sim \text{Bern}\left(\frac{J(b - \log \eta_i)}{m_{i\cdot} + a - 1 + J(b - \log \eta_i)}\right) \quad i = 1, ..., n \\
\alpha_i|\eta_i, m_{i\cdot} &\sim \text{Gamma}(a + m_{i\cdot} - s_i, b - \log \eta_i) \quad i = 1, ..., n.
\end{aligned}$$

3.2.2 Estimation of τ_1

In this section, we describe two approaches to estimate the disclosure risk measure τ_1, formula (2), within the sampler described in 3.2.

The first approach follows (Manrique-Vallier and Reiter 2012) and relies on the simulation of the unobserved individuals in the population. Specifically, remember that $f = (f_1, \ldots, f_{|\mathcal{C}|})$ denotes the vector of frequencies of each cell $c \in \mathcal{C}$ in the sample of size n, and F the corresponding vector in the population of size N. Then, at iteration m of the MCMC sampler, the m-th draw of $\tau_1^{(m)}$ can be obtained by applying the following algorithm.

1. Let $F^{(m)} = f$, i.e. initialize the population vector using the sample vector.
2. For $i = n + 1, \ldots, N$:
   (a) draw $(g_{i,k})$ from (5);
   (b) for $j = 1, \ldots, J$:
      (i) sample $Z_{i,j}|\,(g_{i,k})$ from (3);
      (ii) sample $X_{i,j} \sim \theta^{(Z_{i,j})}_{j,\cdot}$;
   (c) set $F_c^{(m)} = F_c^{(m)} + 1$, where c is the cell corresponding to the sampled X_i.
3. Set $\tau_1^{(m)} = \sum_{c \in \mathcal{C}} \mathbb{I}\left(F_c^m = 1, f_c = 1\right)$, where $\mathbb{I}$ denotes the indicator function.

Point estimates and credible intervals of τ_1 can then be obtained from the empirical quantities. This approach is computationally intensive when the population N is large.

An alternative approach, computationally much faster, relies on a Monte Carlo approximation. Specifically, let us recall that $\mathcal{C} := \times_{j=1}^{J}\{1 \ldots, n_j\}$ denotes the state space of the observations. Given a sample $X_{1:n} := (X_1, \ldots, X_n)$, and denoting by $\tilde{\mathcal{C}}_{X_{1:n}}$ the set $\tilde{\mathcal{C}}_{X_{1:n}} := \{c \in \mathcal{C} : \sum_{c \in \mathcal{C}} \mathbb{I}\left(\sum_{i=1}^{n} \mathbb{I}\,(X_i = c) = 1\right)\}$ the set of combinations appearing with frequency 1 in the sample. Then, τ_1 can be estimated within the MCMC using the following algorithm.

1. For $t = 1, \ldots T$: draw $(g_{t,k})$ from (5).
2. For $c \in \tilde{\mathcal{C}}_{X_{1:n}}$: Compute the Monte Carlo approximation,

$$\mathbb{P}\left(\{X_{n+1} = c\}|G_0, \alpha_0\right) \approx \frac{1}{T}\sum_{t=1}^{T}\prod_{j=1}^{J}\left(\sum_{k=1}^{K_n} g_{t,k}\theta^{(k)}_{j,c_j} + g_{t,0}\frac{1}{n_j}\right).$$

3. Set $\tau_1^{(m)} = \sum_{c \in \tilde{\mathcal{C}}_{X_{1:n}}} \left(1 - \mathbb{P}\left(\{X_{n+1} = c\}|G_0, \alpha_0\right)\right)^{N-n}$.

In the algorithm, T is the number of Monte Carlo draws to approximate $\mathbb{P}\left(\{X_{n+1} = c\}|G_0, \alpha_0\right)$. The algorithm is easily parallelizable both in t and c. The derivations of this approximation and the corresponding formula for the algorithm with structural zeros of Section 4, can be found in the Online Supplementary Material. This Monte Carlo approximation of τ_1 is also a novel contribution in itself and significantly improves the running time of the algorithm.

A final remark is that record-level measures of disclosure risk, such as r_{1c} in formula (1), can be easily estimated with the same approach. Indeed, both algorithms presented in this subsection, at bullet points 2. of each iteration, estimate the record-level quantities r_{1c} for each sample unique cell c. Then, at bullet point 3., they sum over all $c \in \tilde{\mathcal{C}}_{X_{1:n}}$ to obtain the corresponding file-level estimate of τ_1. However, given that file-level disclosure risk measures provide a simple 1-dimensional summary of the disclosure risk of a specific dataset and allow straightforward and easy to visualise comparisons among different methodologies, we will focus our attention only on them in the rest of the paper.

4 Extension to Structural Zeros

Structural zeros are combinations of key variables that lead to impossible values, like a five-year-old veteran or a pregnant male. In real datasets, structural zeros might account for an extremely large proportion of the possible cells $|\mathcal{C}|$,

often above 90-95%. If a Bayesian model does not take into account the presence of structural zeros, its posterior estimates can deteriorate dramatically, as shown in an example in Section 5. This is because, if the prior distribution assigns positive probability to every possible cell in $\mathcal{C}$, the posterior distribution will also assign some mass to every cell. Even if the posterior mass assigned to each structural zero cell is very low, if the number of these cells is very large, their overall posterior mass will be far from being negligible.

Structural zeros should not be confused with sparsity for tabular data, as considered for example in Snoke et al. (2025). Sparsity occurs in tabular data because, when the number of key variables J is large, the vector of cross-classified cells will have a very large dimension $|\mathcal{C}|$. If the sample size is not too large, most of the cells in the vector of sample frequencies $(f_1, \ldots, f_{|\mathcal{C}|})$ will be equal to zero. Instead, structural zeros are cells that are frequency zero in the sample not due to the small sample size, but to the specific nature of their categorical variable values. They correspond to combinations of key variables that are deemed impossible to observe also in the population. In the illustration of Section 5, the real data example includes both sparsity, due to the small sample sizes, and the presence of structural zeros.

Following the general algorithm of Manrique-Vallier and Reiter (2014), in this Section, we describe an MCMC algorithm to perform posterior inference on the model parameters and disclosure risk measure τ_1, in the presence of structural zeros. The main idea of the algorithm is to consider the observed sample, $X_{1:n}$ of size n, as a truncated version of larger sample $X_{1:n+n_0}$ of size $n + n_0$ sampled from the model of section 3.1, and in which n of these observations have fallen into admissible cells, while the other n_0 have taken values in structural zeros cells. Then, the algorithm is a data augmentation scheme in which in steps 7-9, we sample the latent variables (n_0 truncated observations in the structural zeros, denoted $Z_{(n+1):(n+n_0)}$ $X_{(n+1):(n+n_0)}$) given the observed variables and common parameters, and in steps 3-5 we sample the common parameters given both the observed and latent variables.

Structural zeros can be defined in terms of marginal conditions. These are conditions that fix 2 or more key variables to some specific values. For example $\mu = \{*, 1, *, *, 2, *\}$ is the marginal condition on a dataset with 6 key variables and includes all cells taking value 1 in the second variable and value 2 in the fifth one, and the placeholder symbol $*$ means that that variable is unrestricted. Conditions that fix more than one category in a specific variable can be written separately as unions of multiple marginal conditions. Moreover, a set of overlapping marginal conditions can always be rewritten as a (possibly larger) set of *disjoint* marginal conditions. For example, let us suppose to have 3 binary key variables, the two overlapping conditions $\tilde{\mu}_1 = \{*, 1, 2\}$ and $\tilde{\mu}_2 = \{1, 1, *\}$ (cell $\{1, 1, 2\}$ belongs to both conditions) can be rewritten as disjoint conditions $\mu_1 = \{*, 1, 2\}$ and $\mu_2 = \{1, 1, 1\}$, i.e. $\tilde{\mu}_1 \cup \tilde{\mu}_2 = \mu_1 \cup \mu_2$.

Section 4.2 of Manrique-Vallier and Reiter (2014) presents a simple algorithm to transform a set of *overlapping* marginal conditions into a set of *disjoint* ones. This algorithm is run as a pre-processing step before implementing the MCMC. Therefore, we can assume to have a set of C *disjoint* marginal conditions, denoted $S_d = \{\mu_1, \ldots, \mu_C\}$, specifying sets of impossible cells, and $S = \cup_{c=1}^{C} \mu_c$ the subset of sample space $\mathcal{C}$ corresponding to structural zeros. In the MCMC sampler, for each marginal constraint μ_c, steps 7-9 simulate the truncated observations from $X_{(n+1):(n+n_0)}$ that fall into the cells specified by μ_c, and their corresponding mixture classification variables $Z_{(n+1):(n+n_0)}$. Specifically, step 7 computes the probability p_c of all cells in μ_c, step 8 samples the number n_c of truncated observations from $X_{(n+1):(n+n_0)}$ in μ_c, and finally step 9 samples their mixture classification Z_i given the event $X_i \in \mu_c$.

4.1 MCMC Algorithm including structural zeros

The MCMC sampler of Section 3.2.1 can be extended following (Manrique-Vallier and Reiter 2014) to account for the presence of structural zeros. Specifically, we repeat the following steps.

1. Sample $Z_{i,j}$: for $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, J\}$, sample $Z_{i,j}$ from (3).
2. Sample m_{ik}: for $i = \{1, \ldots, n + n_0\}$ and $k = \{1, \ldots, K_n\}$, sample m_{ik} as in step 2 of the sampler in Section 3.2.
3. Draw $\left(g_{0,k}\right)$: Compute $m_{\cdot k} = \sum_{i=1}^{n+n_0} m_{ik}$ for $k \in \{1, \ldots, K_n\}$, and sample $\left(g_{0,k}\right)$ from (4).
4. Draw $\left(g_{i,k}\right)$: for $i \in \{1, \ldots, n\}$ sample $\left(g_{i,k}\right)$ from (5).
5. Draw $\theta^{(k)}$: For $k \in \{1, \ldots, K_n\}$ and $j \in \{1, \ldots, J\}$ sample $\theta^{(k)}_{j,\cdot}$ according to,

$$\text{Dirichlet}\left(1 + \sum_{i=1}^{n+n_0} \mathbb{I}\left(Z_{ij} = k, X_{ij} = 1\right), ..., 1 + \sum_{i=1}^{n+n_0} \mathbb{I}\left(Z_{ij} = k, X_{ij} = n_j\right)\right)$$

6. Update α_0, α_i, for $i \in \{1, \ldots, n\}$, as in step 6 of the sampler in Section 3.2.
7. Compute $(p_1, \ldots, p_C)$: for $c \in \{1, \ldots, C\}$, compute with Monte Carlo

$$p_c := \mathbb{P}\left(X_i \in \mu_c | G_0, \alpha\right) = \int \mathbb{P}\left(X_i \in \mu_c | G_i, \alpha\right) \mathbb{P}\left(G_i | G_0, \alpha\right) dP\left(G_i\right)$$

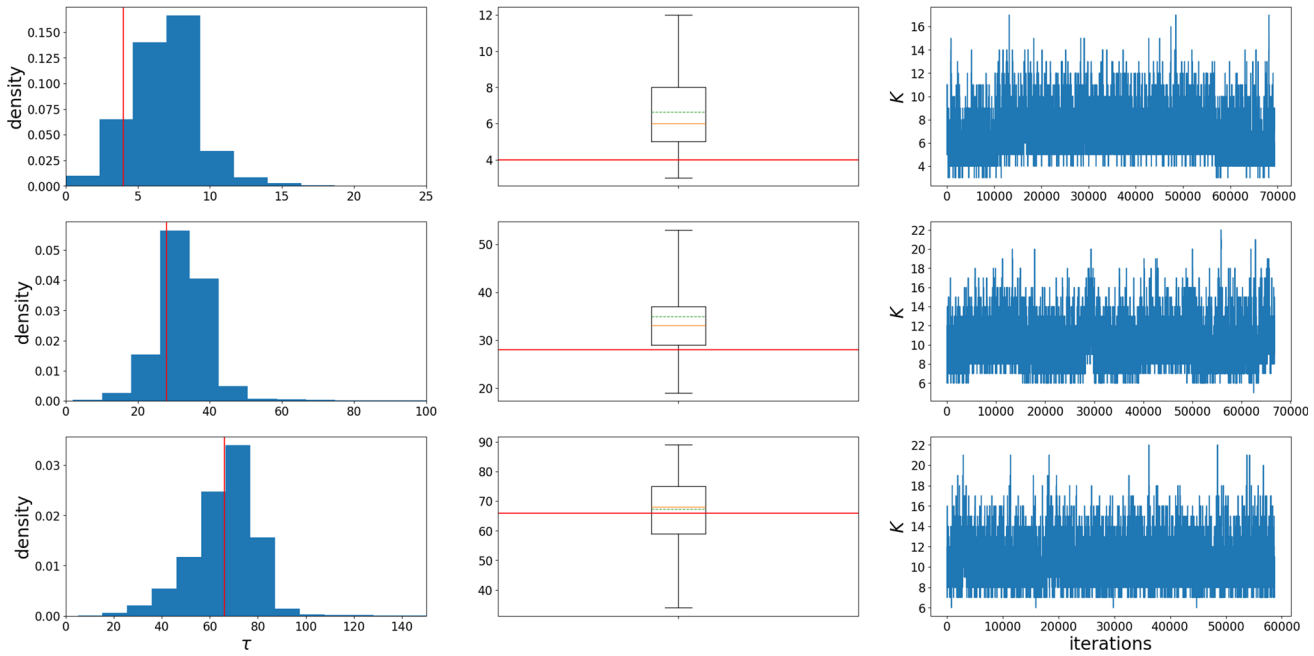

Fig. 1 HDP on synthetic data with τ_1 estimated via sampling. In red the true τ_1. For the box plots: in orange the median, in dashed green the mean, and the whiskers show 95% credible intervals

$$= \int \prod_{j\in\{1,\dots,J\}:\mu_{c,j}\neq *} \left(\sum_{k=1}^{K_n} g_{i,k}\theta^{(k)}_{j,\mu_{c,j}} + g_{i,0}\frac{1}{n_{j\cdot\cdot}} \right) \mathrm{Dir}\left((g_{i,k})|(g_{0,k})\right) dg_{i,k}$$

8. Draw $(n_1, \dots, n_C)$: sample a vector

$$(n_1, \dots, n_C) \sim \mathrm{NM}\,(n, p_1, \dots, p_C)\,,$$

where NM denotes a Negative Multinomial distribution, with mass function,

$$p\,(n_1, \dots, n_C|p_1, \dots, p_C) = \frac{\Gamma\,(n+n_0)}{\Gamma\,(n)\prod_{c=1}^{C} n_c!}\left(1-\sum_{c=1}^{C} p_c\right)^{n}\prod_{c=1}^{C} p_c^{n_c}$$

where $n_0 := \sum_{i=1}^{C} n_i$.

9. Sample $Z_{(n+1):(n+n_0)}$, $X_{(n+1):(n+n_0)}$: for $c \in \{1, \dots, C\}$ and for $i \in \{n+\sum_{l=1}^{c-1} n_l + 1, n+\sum_{l=1}^{c-1} n_l + 2, \dots, n+\sum_{l=1}^{c-1} n_l + n_c\}$ (with the proviso that $\sum_{l=1}^{0} n_l = 0$):
   - Draw $\left(g_{i,k}\right)$ from (5). Then, for $j \in \{1, \dots, J\}$:
     - If $\mu_{c,j} \neq *$: Set $X_{i,j} = \mu_{c,j}$ and sample $Z_{i,j}$ from (3).
     - If $\mu_{c,j} = *$: Sample $\mathbb{P}\left(Z_{i,j} = k|\left(g_{i,k}\right)\right) = g_{i,k}$, and sample $X_{i,j}|Z_{i,j}, \theta \sim \theta^{(Z_{i,j})}_{j,\cdot}$.

When the structural zeros account for the majority of cells, the probabilities and counts from steps 7 and 8 can become very large. This implies that, at step 9, many variables have to be simulated, and this slows down the algorithm significantly. In the Online Supplementary Material, we describe an approximation of step 9 that reduces the computational cost dramatically and produces similar estimates of τ_1. Remark also that even the estimate of τ_1 should account for the presence of structural zeros, and, as already mentioned, we describe this modification in the Online smentary Material.

5 Experiments

This section is composed of two parts. In Subsection 5.1, we compare the parametric and non-parametric versions of the mixed membership model on synthetic data. In Subsection 5.2, we test the performance of the non-parametric model on a real dataset in two scenarios, with and without modelling the structural zeros.

The code to reproduce the experiments is open source and available at https://github.com/LorenzoRimella/BNP_DR. All the experiments were run on a 32 GB Tesla V100 GPU available on "The High-End Computing" (HEC) facility at Lancaster University. In the same Github folder, the implementation of the mixed membership model of Manrique-Vallier and Reiter (2012) is also available to reproduce the comparison. We did not compare with the original code of Manrique-Vallier and Reiter (2012) and Manrique-

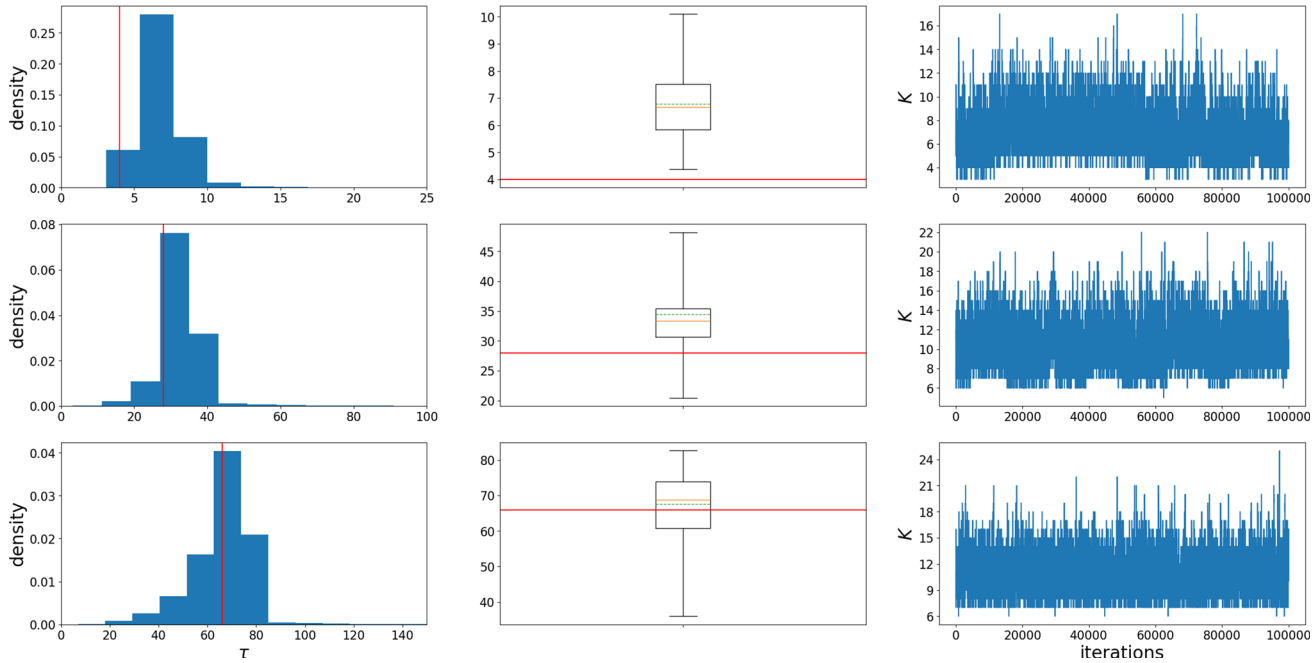

Fig. 2 HDP on synthetic data with τ_1 estimated via Monte Carlo. In red the true τ_1. For the box plots: in orange the median, in dashed green the mean, and the whiskers show 95% credible intervals

Vallier and Reiter (2014) as we did not find open-source references in the manuscripts.

In the Online Supplementary Material, we have also included some notes on how we implement the algorithm to exploit parallel computing.

5.1 Synthetic data

We generated synthetic data of size $N = 712174$ from the mixed membership models of Manrique-Vallier and Reiter (2012) with $K = 15$, for $J = 10$ categorical variables, and different numbers of categories n_j per variable, ranging from 2 to 11. We draw three samples of sizes $n = 1000, 5000, 10000$, and run three MCMC samplers: i) the non-parametric HDP model with τ_1 estimated via population sampling; ii) the non-parametric HDP model with τ_1 estimated via Monte Carlo sampling; iii) the algorithm from Manrique-Vallier and Reiter (2012) for different values of K, for 400k iterations, out of which 300k discarded as a burn-in. As a further comparison, we also report the results obtained by an 'oracle' version of log-linear models. This oracle model, which requires knowing the true value of τ_1, is obtained by implementing many different log-linear specifications proposed in Skinner and Shlomo (2008); Snoke et al. (2025), including pairwise interaction and LASSO regularization, and selecting that with best estimates of τ_1. Alternatively, a model selection procedure, such as that proposed in Skinner and Shlomo (2008), could be used to select a specific log-linear competitor. However, we focus on the 'oracle' version instead, which produces the best competitor among all log-linear model specifications.

Figures 1-2 display the histogram estimates (first column) and box plots (second column) of τ_1 and the trace plots of the number of mixture components K_n (third column) for the HDP model, using population and Monte Carlo sampling, respectively. Different rows in the figures correspond to different sample sizes. The non-parametric model performs well on synthetic data and is capable of recovering the true value of τ_1, which is 4, 28, 66 respectively and falls within 95% credible intervals for all three sample sizes. Moreover, the algorithms with population sampling and Monte Carlo sampling seem to have comparable performance, with the Monte Carlo approximation narrowing the credible intervals. In view of the computational time gain, the Monte Carlo sampling approach seems preferable, and we have focused on that in the real-data example. Also, in Figure 3, the results of the parametric model are shown, from which it seems the model might slightly underestimate τ_1. The overall posterior mean point estimates and standard deviations of τ_1 are summarized in Table 1 together with the computational times to run the algorithms.

As already mentioned, for completeness, we have also included some comparison with a log-linear Poisson model as in Skinner and Shlomo (2008). We find that the log-linear models are several orders of magnitude faster but significantly overestimate τ_1. We also tried to include all the pairwise interaction terms and a LASSO penalization as suggested in Snoke et al. (2025), but we did not see substantial

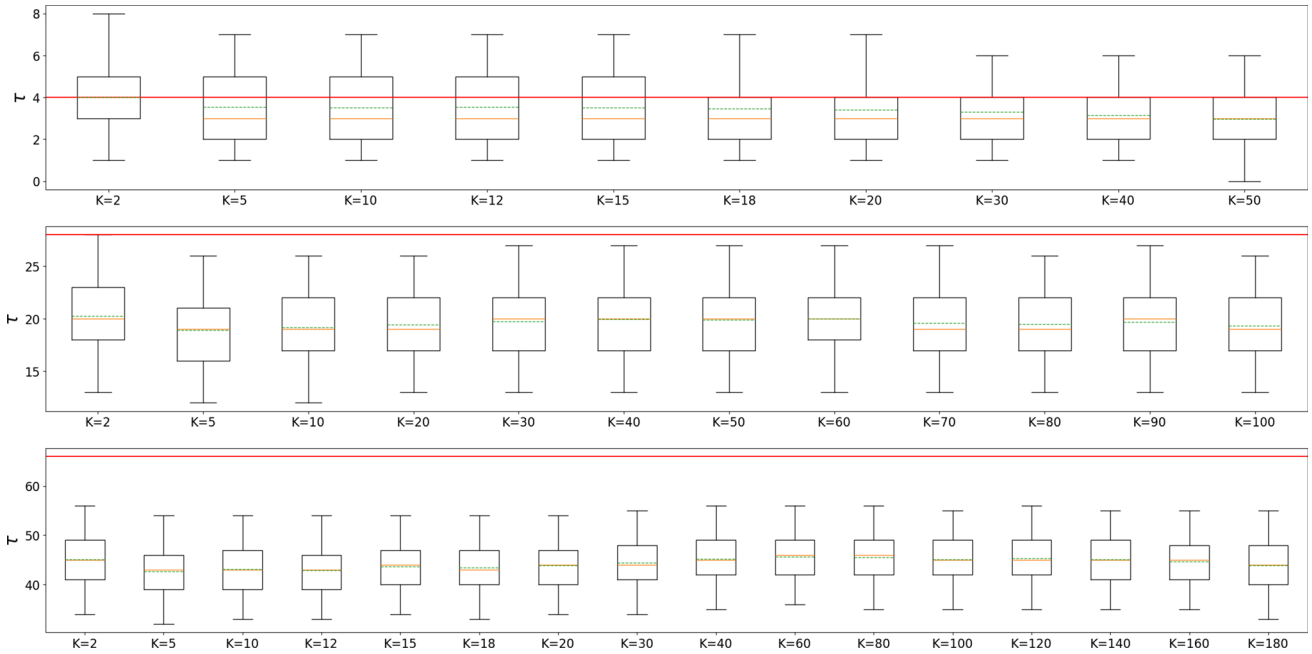

Fig. 3 Estimates of τ_1 with the parametric model for different values of K. In red the true τ_1. For the box plots: in orange the median, in dashed green the mean, and the whiskers show 95% credible intervals

improvements, see the Online Supplementary Material. In all Tables, we only report an 'oracle' version of log-linear models, i.e. the log-linear specification that results in the best estimates possible, which might vary from sample to sample and requires knowing τ_1. The underperformance of the log-linear models in this simulation setting might be due to the fact that the synthetic data are simulated from a mixed membership model, making the log-linear model a misspecified choice.

5.2 Real data

We test the performance of the HDP model on a real dataset from the 5% public use microdata sample (PUMS) of the American Community Survey, which is fielded by the U.S. Census Bureau, for the state of New York (Ruggles et al. 2024). In this illustrative real-data example, we consider the entire dataset as the "population", and draw a sample at random from it, which is then used to estimate the actual value of τ_1. The dataset contains information about the following 10 categorical variables, observed for a population of 953076 individuals: ownership of dwelling (OWNERSHIP: 3 levels), mortgage status (MORTGAGE: 4 levels), age bands (AGE: 9 levels), sex (SEX: 2 levels), marital status (MARST: 6 levels), race identification band (RACESING: 5 levels), education level (EDUC: 11 levels), employment status (EMPSTAT: 4 levels), work disability (DISABWRK: 3 levels), and veteran status (VETSTAT: 3 levels).

This data results in a contingency table of 2566080 cells in total, many of which can be considered as structural zeros. For example, from Table 2, which cross-classifies age and education, we can see that there are some obvious structural zeros. These are due to the impossibility of some values of the categorical variables to coexist, e.g. age below 14 (level 1) with the highest level of education (level 11). Following (Manrique-Vallier and Reiter 2014), we recover 60 overlapping marginal conditions, resulting in 557 disjoint marginal conditions and representing 2317030 cells of our contingency table (approx. 90%). As pointed out in Section 4, this real-data example includes both sparsity and structural zeros. Indeed, the total number of cells $|\mathcal{C}|$ is approximately 2.5 million, of which approximately $250k$ cells are considered to be non-structural zero cells, while the remaining 2.25 million are structural zeros. With sample sizes of just $1k$-$10k$ observations, the observed frequency vectors restricted to non-structural zeros cell are still extremely 'sparse' and composed mainly of cells with zero observed frequency (at most 1-2% of $250k$ non-structural zero cells have non-zero observed frequency).

As an initial demonstration of the performance of the non-parametric model, we first pre-processed the data to remove the majority of structural zeros and then implemented the algorithm from Section 3.2. Specifically, we have removed all the individuals that were younger than 18, and we dropped the categorical variables OWNERSHP and MORTGAGE. This results in a dataset with 712174 individuals and a significantly smaller contingency table of 39600 cells, with many

Table 1 τ_1 estimates for synthetic data using: the HDP model under sampling and Monte Carlo, the parametric model with the best choice of K (the closest posterior mean to the true τ_1), and a log-linear model. The log-linear model is chosen according to an oracle

Algorithm type	n	True τ_1	Est. τ_1	Comp. time (hours, max = 12h)
HDP sampling	1000	4	6.66+/-2.42	12.01
HDP sampling	5000	28	34.92+/-23.63	12.01
HDP sampling	10000	66	67.33+/-25.42	12.0
HDP Monte Carlo	1000	4	6.79+/-1.55	1.49
HDP Monte Carlo	5000	28	34.48+/-20.86	3.13
HDP Monte Carlo	10000	66	67.64+/-24.78	5.35
Parametric K=2	1000	4	4.01+/-1.68	5.39
Parametric K=2	5000	28	20.26+/-3.67	5.95
Parametric K=60	10000	66	45.63+/-5.25	6.89
log-linear Poisson	1000	4	10.97	0.0024
log-linear Poisson	5000	28	55.28	0.0024
log-linear Poisson	10000	66	110.76	0.0024

Table 2 Cross table between AGE and EDUC showing the presence of structural zeros

EDUC	0	1	2	3	4	5	6	7	8	9	10
AGE											
0	69386	77787	51360	2180	298	50	1	0	0	0	0
1	36	12	5016	6070	1867	176	115	0	0	0	0
2	44	6	933	4159	5774	1807	330	33	7	0	0
3	51	17	448	963	4157	5565	2178	73	3	0	0
4	754	268	2085	2011	3189	6908	34299	19418	4507	8339	762
5	1673	507	4022	2857	3639	3966	50900	18474	13830	29715	15523
6	2846	944	6466	3943	4822	5220	84127	27168	21237	35888	27232
7	3444	1764	10976	4872	5860	5098	73838	17735	9791	21102	23182
8	2402	1306	11635	3765	4428	3490	38307	6884	1973	6629	6154

fewer zero cells. For example, now the dataset contains only rows from categories 4 to 8 of Table 2, hence reducing substantially the number of empty cells.

After drawing three samples of sizes 1000, 5000, 10000, the MCMC of Section 3.2 was run for 300k iterations, with the first 200k iterations discarded as a burn-in. Figure 4 displays the posterior histogram and box plot of τ_1, together with the trace plots of the number of mixture components K_n. Moreover, the first three rows of Table 3 summarize point estimates and credible intervals of τ_1. The true value of τ_1 is mostly within the 95% credible interval. Note that the slight deterioration in performance compared to the synthetic data example might also be due to the presence of some additional structural zeros that have not been completely removed with the pre-processing step.

As a comparison, we consider a few log-linear model specifications: the simple log-linear model (Skinner and Shlomo 2008), the log-linear model with pairwise interaction (Skinner and Shlomo 2008), and the log-linear model with pairwise interaction and LASSO penalization (Snoke et al. 2025). We did not perform model selection for the LASSO penalization coefficient, as in Snoke et al. (2025), but rather considered an 'oracle' selecting its best value, which was $\lambda = 1000$, in terms of distance from the true value of τ_1, see Table 3. The 'oracle' version of the method from Snoke et al. (2025) performs very well, suggesting its use in the absence of structural zeros, if the proposed model selection step of λ was producing values comparable with the oracle ones (which requires knowledge of τ_1). More details on the results are available in the Online Supplementary Material.

We now consider the raw data, in which structural zeros have not been removed. In order to show how much they can deteriorate the performance of the algorithm, we first run the algorithm of Section 3.2, which does not model structural zeros. The results are displayed in Figure 5 and in Table 3. Even with 300k iterations, the Markov chain fails to properly converge and estimates τ_1 to be very far from the true values. Indeed, the algorithm significantly overestimates τ, as sample unique values now have a very low posterior probability of being sampled again. This is because most of the posterior probability mass is now assigned to the structural zeros cells, and much larger sample sizes are probably needed to wash out the effect of the prior and shrink this probability to zero.

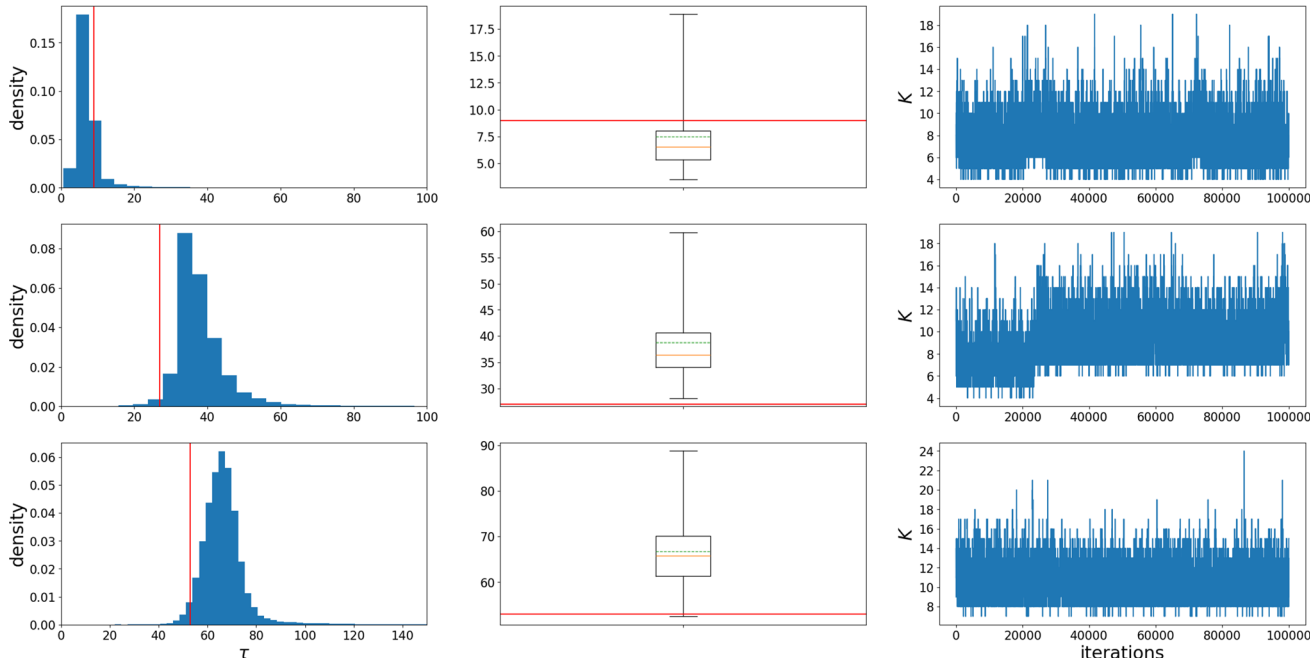

Fig. 4 Estimates of τ_1 with HDP, algorithm from Section 3.2, on real data, after pre-processing to remove structural zeros. In red the true τ_1. For the box plots: in orange the median, in dashed green the mean, and the whiskers show 95% credible intervals

Table 3 τ_1 estimates from the real data with and without structural zeros (SZ). The HDP has been launched with and without the adjustment for structural zeros. The log-linear model is chosen according to an oracle. The HDP in the presence of structural zeros has two modes when $n = 10000$, which are both reported.

Algorithm type	n	True τ	Est. τ	Structural zeros	Comp. time (hours)
HDP Monte Carlo	1000	9	7.49+/-5.32	False	2.01
HDP Monte Carlo	5000	27	38.8+/-12.0	False	4.52
HDP Monte Carlo	10000	53	66.72+/-10.85	False	7.20
log-linear Poisson	1000	9	8.87	False	0.0028
log-linear Poisson	5000	27	24.38	False	0.0029
log-linear Poisson	10000	53	64.08	False	0.0029
HDP Monte Carlo	1000	11	57.98+/-13.25	True	2.04
HDP Monte Carlo	5000	55	186.06+/-46.67	True	4.48
HDP Monte Carlo	10000	88	288.7+/-47.15	True	7.17
log-linear Poisson	1000	11	157.80	True	0.21
log-linear Poisson	5000	55	496.55	True	0.22
log-linear Poisson	10000	88	769.87	True	0.22
SZ HDP Monte Carlo	1000	11	10.81+/-1.66	True	7.61
SZ HDP Monte Carlo	5000	55	56.16+/-3.78	True	9.08
SZ HDP Monte Carlo	10000	88	189.46+/-5.76	True	12.58
			101.98+/-5.27		

We then consider the log-linear models with 'oracle' for selection of the model specification and regularization parameter. The number of possible cells in this experiment is 2566080, compared to just 39600 after preprocessing. This fact significantly impacts the memory cost of the algorithm and its running time. The running time is still considerably cheaper compared to mixed membership models, even though it seems to scale worse with the total number of cells. Looking at the accuracy of the estimates, the estimated values of τ_1 are considerably far from the true values, with estimates that are even 10 times higher than τ_1, making the log-linear models not suitable in the presence of structural zeros. We refer to Table 3 for the best 'oracle' log-linear model, and to the Online Supplementary Material for the other sub-optimal specifications of log-linear models.

Finally, we run the algorithm of Subsection 4.1, which accounts for the presence of structural zeros in the data. Figure 6 and the last three rows of Table 3 show the results of

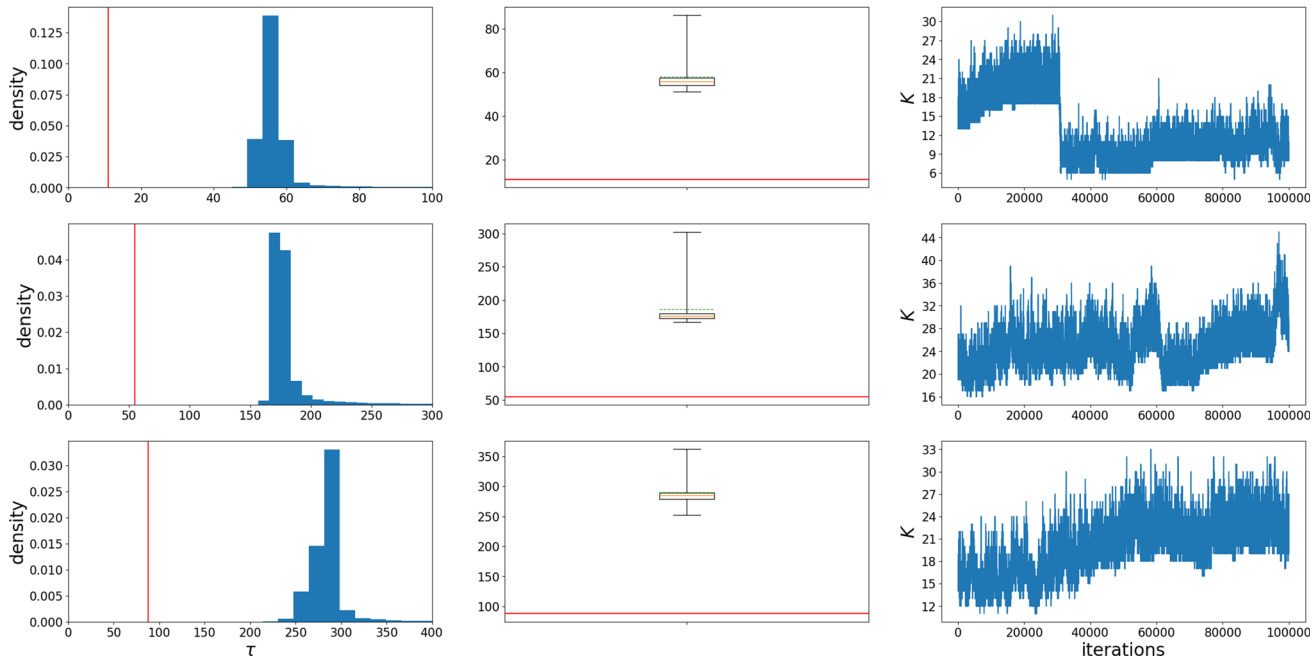

Fig. 5 Estimates of τ_1 with HDP, algorithm from Section 3.2, on real data, without pre-processing to remove structural zeros. In red the true τ_1. For the box plots: in orange the median, in dashed green the mean, and the whiskers show 95% credible intervals

100K iterations, obtained after the burn-in period of 100k, for three samples of sizes 1000, 5000, and 10000. The first two rows of Figure 6 display the histogram estimators and box plots of τ_1 and the trace plots of K_n, for the samples 1000 and 5000. From the plots, we can see that, for these samples, the MCMC has converged to stationarity and the estimates τ_1 are good, with the true value being within the 95% posterior credible intervals. However, we should warn that for some samples, the posterior distribution can become multimodal. This is the case for the chosen sample of size 10000, third row in Figure 6. It is indeed evident that the posterior is bimodal. Due to the initialization chain, the chain spends many iterations in the sub-optimal mode before jumping to the best mode. Estimates obtained by discarding more iterations of burn-in, or running for Markov chain longer, produce reasonable estimates of τ_1. Alternatively, a clustering algorithm can be applied to separate MCMC iterations from the two modes as a post-processing step.

6 Discussion

In this work, we have proposed a Bayesian non-parametric approach, based on hierarchical modelling, that generalizes parametric mixed membership models to estimate measures of disclosure risk. The proposed approach does not have any tuning parameters and performs well in the experiments, even with samples as small as 1% of the entire population. Also, the methodology can be extended to account for the presence of many structural zeros in the data through a data augmentation scheme. Moreover, fast Monte Carlo approximation schemes have been suggested, which can reduce the computational cost of running the algorithms dramatically, hence making the approach applicable also in the presence of large population sizes N.

It is important to comment on the trade-off between our method and log-linear models. Log-linear models are computationally fast and perform well without structural zeros. However, they can perform poorly in the presence of structural zeros, as shown in Section 5.2. Moreover, they do not scale well with the number of categorical variables and the overall number of cells. Indeed, their computational cost, see Table 3, becomes 100 times slower when including structural zeros. The HDP method is computationally more demanding, but also performs well in the presence of structural zeros. Moreover, the computational cost seems to scale well with the overall number of cells, with an increase by a factor of only two/three times when including structural zeros, see Table 3. These considerations suggest that log-linear models might be preferable in the presence of a low computational budget, few categorical variables, and no structural zeros, while the HDP method is preferable in the presence of structural zeros and/or many categorical variables.

In terms of improvements, we have shown in the experiment section how the posterior distribution of the augmented model, accounting for structural zeros, can become multimodal, depending on the observed sample. If the MCMC algorithm is poorly initialized and not run long enough, it can get stuck in a sub-optimal mode, hence producing misleading

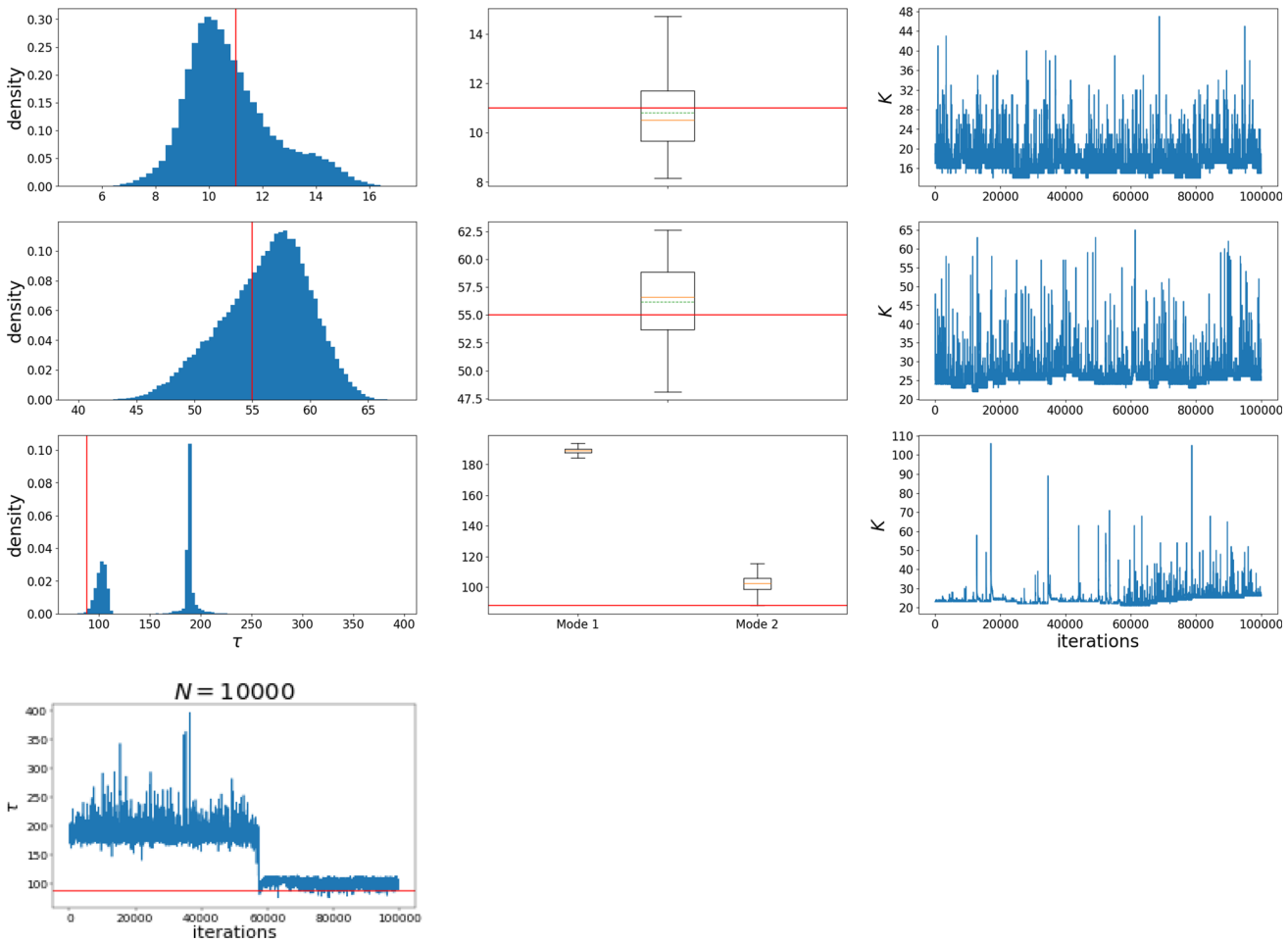

Fig. 6 Estimates of τ_1 with HDP, algorithm from Section 4.1, on real data, without pre-processing to remove structural zeros. The last row reports the trace plot for $N = 10000$. In red the true τ_1. For the box plots: in orange the median, in dashed green the mean, and the whiskers show 95% credible intervals

estimates of τ_1. Therefore, we recommend the practitioner to start the algorithm from different initial values of the parameters, in order to detect whether any multimodality is present. If so, the MCMC should be run for many iterations to properly explore the parameter space and obtain accurate estimates. Alternatively, a clustering algorithm could be applied as a post-processing step to separate MCMC iterations obtained from different modes. As a direction for improvement, it would be useful to define an automatic approach to detect multimodality in the posterior distribution of mixed membership models and, if so, a way of properly handling it, like for example by restarting the algorithm from different initial values, choosing a good way to initialize the algorithm using the observed sample, or trying to improve step sizes and movement directions of the chain, to facilitate jumps from sub-optimal to optimal modes.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11222-025-10693-9.

Author Contributions M.B. and L.R. equal contributions to the paper

Data Availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

Bethlehem, J.G., Keller, W.J., Pannekoek, J.: Disclosure control of microdata. J. Am. Stat. Assoc. **85**(409), 38–45 (1990)

Carota, C., Filippone, M., Leombruni, R., Polettini, S.: Bayesian nonparametric disclosure risk estimation via mixed effects log-linear models. Ann. Appl. Stat. **9**, 525–546 (2015)

Carota, C., Filippone, M., Polettini, S.: Assessing bayesian semiparametric log-linear models: an application to discolure risk estimation. Int. Stat. Rev. **90**(1), 165–183 (2022)

Duncan, G.T., Lambert, D.: Disclosure-limited data dissemination. J. Am. Stat. Assoc. **81**(393), 10–18 (1986)

Duncan, G., Lambert, D.: The risk of disclosure for microdata. J. Bus. Econ. Stat. **7**(2), 207–217 (1989)

Ferguson, T.S.: A bayesian analysis of some nonparametric problems. Ann. Stat. **1**(2), 209–230 (1973)

Lambert, D.: Measures of disclosure risk and harm. J. Off. Stat. **9**, 313–313 (1993)

Manrique-Vallier, D., Hu, J.: Bayesian non-parametric generation of fully synthetic multivariate categorical data in the presence of structural zeros. J. R. Stat. Soc. Ser. A Stat. Soc. **181**(3), 635–647 (2018). https://doi.org/10.1111/rssa.12352. (https://academic.oup.com/jrsssa/article-pdf/181/3/635/49449396/jrsssa_181_3_635.pdf)

Manrique-Vallier, D., Reiter, J.P.: Estimating identification disclosure risk using mixed membership models. J. Am. Stat. Assoc. **107**(500), 1385–1394 (2012)

Manrique-Vallier, D., Reiter, J.P.: Bayesian estimation of discrete multivariate latent structure models with structural zeros. J. Comput. Graph. Stat. **23**(4), 1061–1079 (2014)

Matthews, G.J., Harel, O., et al.: Data confidentiality: a review of methods for statistical disclosure limitation and methods for assessing privacy. Stat. Surveys **5**, 1–29 (2011)

Reiter, J.: Estimating risks of identification disclosure in microdata. J. Am. Stat. Assoc. **100**, 1103–1112 (2005)

Rinott, Y., Shlomo, N.: A Generalized Negative Binomial Smoothing Model for Sample Disclosure Risk Estimation. In Privacy in Statistical Databases. Lecture Notes in Computer Science, Springer, Heidelberg (2006)

Ruggles, S., Flood, S., Sobek, M., Backman, D., Chen, A., Cooper, G., Richards, S., Rodgers, R., Schouweiler, M.: IPUMS USA: Version 15.0 [dataset]. IPUMS. Minneapolis, MN. (2024). https://doi.org/10.18128/D010.V15.0

Sethuraman, J.: A constructive definition of dirichlet priors. Stat. Sin. **4**, 639–650 (1994)

Shlomo, N., Skinner, C., et al.: Assessing the protection provided by misclassification-based disclosure limitation methods for survey microdata. Ann. Appl. Stat. **4**(3), 1291–1310 (2010)

Skinner, C.J., Elliot, M.: A measure of disclosure risk for microdata. J. R. Stat. Soc.: series B (statistical methodology) **64**(4), 855–867 (2002)

Skinner, C., Shlomo, N.: Assessing identification risk in survey microdata using log-linear models. J. Am. Stat. Assoc. **103**(483), 989–1001 (2008)

Skinner, C., Shlomo, N.: Assessing identification risk in survey microdata using log-linear models. J. Am. Stat. Assoc. **103**(483), 989–1001 (2008)

Skinner, C., Marsh, C., Openshaw, S., Wymer, C.: Disclosure control for census microdata. J. Off. Stat. **10**, 31–31 (1994)

Snoke, J., Meijer, E., Phillips, D., Wilkens, J., Lee, J.: Synthesizing surveys with multiple units of observation: An application to the longitudinal aging study in india. Journal of Survey Statistics and Methodology, 047 (2025)

Sweeney, L.: Computational disclosure control: Theory and practice. PhD dissertation, Massachusetts Institute of Technology (2001)

Teh, Y.W., Jordan, M.I.: 5. Bayesian Nonparametrics., pp. 158–207. Cambridge University Press, Cambridge (2010)

Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical dirichlet processes. J. Am. Stat. Assoc. **101**, 1566–1581 (2006)

Willenborg, L., De Waal, T.: Elements of Statistical Disclosure Control vol. 155. Springer, New York (2012). https://doi.org/10.1007/978-1-4613-0121-9