Issues on Bayesian nonparametric measures of disclosure risk

Questioni su misure Bayesiane nonparametriche di rischio di ”disclosure”

Federico Camerlenghi, Cinzia Carota and Stefano Favaro

Abstract Consider a microdata sample from a finite population, such that each record contains two disjoint types of information: identifying and sensitive information. Any decision about releasing data is supported by the estimation of measures of disclosure risk, which are functionals of the number of records with a unique combination of values of identifying variables. The work of [7] first explored the use of exchangeable random partitions to estimate a common measure of disclosure risk: the number of unique sample records that are also unique population records. In this paper we revisit the work of [7] from a Bayesian nonparametric perspective, and we discuss new potential research directions in the fields.

Abstract *Si consideri un campione di microdati da una popolazione finita, dove ogni record contiene due informazioni disgiunte: informazioni identificative e sensibili. Ogni decisione sul rilascio dei dati è supportata dalla stima di misure del rischio di ”disclosure”, che sono funzione del numero di record con una combinazione unica di valori delle variabili identificative. Il lavoro di [7] ha studiato per la prima volta l’uso delle partizioni aleatorie scambiabili per la stima di un comune rischio di ”disclosure”: il numero di record campionari unici che sono anche record di popolazione unici. In questo articolo rivisitiamo il lavoro di [7] in ottica Bayesiana nonparametrica, e discutiamo nuove direzioni di ricerca.*

Key words: Bayesian nonparametrics; Dirichlet process prior; disclosure risk; empirical Bayes; exchangeable random partitions; identifying and sensitive information.

Federico Camerlenghi
University Milano - Bicocca, Piazza dell’Ateneo Nuovo 1, 20126 Milano, Italy. e-mail: federico.camerlenghi@unimib.it

Cinzia Carota
University of Torino, Lungo Dora Siena 100 A, 10153 Torino, Italy. e-mail: cinzia.carota@unito.it

Stefano Favaro
University of Torino, Corso Unione Sovietica 218bis, 10134 Torino, Italy. e-mail: stefano.favaro@unito.it

1 Introduction

Consider a microdata sample $(X_1,\ldots,X_n)$ from a finite population of size $N > n$, such that each sample record X_i contains two disjoint types of information: identifying information and sensitive information. Identifying information consists of the values of categorical variables, which might be matchable to known units of the population. See, e.g., [1] and [9] for a comprehensive account on these types of information. A risk of disclosure arises from the possibility that an intruder might succeed in identifying a microdata unit through such a matching and hence be able to disclose sensitive information of the unit. To prevent disclosure, any decision about releasing data is supported by the estimation of measures of disclosure risk, which are suitable functionals of the number of records with a unique combination of values of identifying variables. Indeed, assuming no errors in the matching process or data sources, for unique records the match is guaranteed to be correct. When sample records are cross-classified according to the identifying variables, the microdata sample is partitioned in $K_n \leq n$ non-empty cells, labelled by $\{X_1^*,\ldots X_{K_n}\}$, $M_{i,n}$ of which have frequency i, for $i = 1,\ldots,n$. Two common measures of disclosure risk are: i) the number ν_1 of unique population records; ii) the number τ_1 of unique sample records that are also unique population records.

The work of [7] first explored the use of exchangeable random partition to estimate τ_1. His ideas can be described in terms of an urn scheme, where records belonging to the cell X_i^* are depicted as balls of the same color. Using Samuels' terminology, let $(X_i)_{i\geq 1}$ be a superpopulation of colored balls belonging to a (ideally) infinite number of colors $(X_i^*)_{i\geq 1}$ with unknown composition $P = (p_i)_{i\geq 1}$, i.e. p_i is the probability of drawing a ball of color X_i^*, with $\sum_{i\geq 1} p_i = 1$ almost surely. Then, consider a population $(X_1,\ldots,X_N)$ which a random sample from P such that: i) $(X_1,\ldots,X_n)$ is an initial observable random sample from P; ii) $(X_{n+1},\ldots,X_{N-n})$ is an additional unobservable random sample from P. In particular, $(X_1,\ldots,X_N)$ takes on the interpretation of the population records, of which the subsample $(X_1,\ldots,X_n)$ are the observable sample records. A natural way to make inference on the composition of the population $(X_1,\ldots,X_N)$ from $(X_1,\ldots,X_n)$ is to immagine sampling the remainder of the population $(X_{n+1},\ldots,X_{N-n})$ from the posterior distribution of P given $(X_1,\ldots,X_n)$. The problem of estimating τ_1 can thus be stated as follows: given a urn whose initial composition is $(X_1,\ldots,X_n)$, and given $(N-n)$ draws from the urn, how many colors which are unique in the initial state will remain unique at the final state?

[7] addressed this problem under an urn scheme introduced by [5] in (mathematical) population genetics. Specifically, consider an urn that initially contains only a black ball with mass $\theta > 0$, and apply iteratively the following sampling scheme: i) if we pick the black ball then it is returned with a ball of a new color with mass 1; ii) if we pick a non-black ball then it is returned with a ball of the same color with mass 1. Let $M_{i,n}$ denote the number of colors with frequency $i > 1$ after n draws, namely the number of unique sample records. By relying on the sole sampling scheme, [7] showed that the expected τ_1 is

$$m_{1,n}\frac{n+\theta-1}{N+\theta-1}, \tag{1}$$

with $m_{1,n}$ being the observed number of unique cells. [7] then specified θ via maximum likelihood, that is by choosing the value $\hat{\theta}$ that maximizes the likelihood function for sample records. From [5] this corresponds to solve, with respect to θ, the equation $k_n = \sum_{1\leq j\leq n-1}\theta/(\theta+j)$, with k_n being the observed number of distinct cells. Experiments in [7] show that the estimator (1) leads to a systematic underestimation of τ_1, which worsen as N and n become large.

In this paper we revisit and discuss the work of [7] from a Bayesian nonparametric perspective. We derive Bayesian nonparametric estimators of ν_1 and τ_1 under a Dirichlet process prior ([4]) for the unknown composition P of the superpopulation. These estimators have simple closed-form expressions, and they are obtained by a direct application of results by [3] on conditional formule for Gibbs-type exchangeable random partitions. Not surprisingly, our estimator of τ_1 coincides with Samuels' estimator (1), showing that (1) is a Bayesian nonparametric estimator of τ_1, with respect to a squared loss function, under a Dirichlet process prior for P. This provides with a Bayesian derivation of Samuels' estimator, and it paves the way to discuss the following issues: i) the problem of uncertainty quantification for Samuels' estimator; ii) the interpretation of (1) as a Bayesian smoothed version of the nonparametric estimator $m_{1,n}n/N$, which is the naive estimator of τ_1 when $m_{1,n}$ is used as an estimator of ν_1 ([1]); iii) the problem of estimating ν_1 and τ_1 under generalized Dirichlet priors, e.g., the two parameter Poisson-Dirichlet prior ([6]) and, more generally, any prior in the class of Poisson-Kingman models ([6]); iv) the problem of extending the approach of [7] to deal with the presence of structurally empty cells (structural zeros).

2 Bayesian nonparametric measures of disclosure risk

For any $\theta > 0$, let $(v_i)_{i\geq 1}$ be a collection of independent random variables with v_i distributed according to a Beta distribution with parameter $(1,\theta)$, and let $(p_i)_{i\geq 1}$ be a sequence of random variables defined as follows: $p_1 = v_1$ and $p_i = v_i\prod_{1\leq j\leq i-1}(1-v_j)$, for any $i\geq 2$. Furthermore, let $(X_i^*)_{i\geq 1}$ be a collection of random variables independent of $(v_i)_{i\geq 1}$ and independent and identically distributed according to a nonatomic probability measure α_0. The Dirichlet process prior with parameter θ and base distribution α_0 is defined as the law of the random probability measure $P_{\theta,\alpha_0} = \sum_{i\geq 1} p_i\delta_{X_i^*}$. We assume a Dirichlet process prior on the unknown composition P of the superpopulation. Then, the sample record $(X_1,\ldots,X_n)$ is a random samples from P_{θ,α_0}, i.e.,

$$\begin{aligned} X_1,\ldots,X_n\,|\,P_\theta &\overset{iid}{\sim} P_{\theta,\alpha_0} \\ P_{\theta,\alpha_0} &\sim \mathscr{D} \end{aligned} \tag{2}$$

with $\mathscr{D}$ denoting the law of P_{θ,α_0}. Due to the (almost sure) discreteness of $\mathscr{D}$, we expect ties in a sample $(X_1,\dots,X_n)$ from P_{θ,α_0}. Specifically, $(X_1,\dots,X_n)$ features $K_n = k_n \leq n$ distinct types, labelled by $\{X_1^*,\dots,X_{K_n}^*\}$, with frequencies $\mathbf{N}_n = (N_{1,n},\dots,N_{K_n,n}) = (n_{1,n},\dots,n_{k_n,n})$ such that $N_{i,n} \geq 1$ and $\sum_{1\leq i\leq K_n} N_{i,n} = n$. The distribution of the random variable $(K_n,N_{1,n},\dots,N_{K_n,n})$ models the composition of the (observable) sample records. We refer to [6] for a comprehensive account on the distribution of $(K_n,N_{1,n},\dots,N_{K_n,n})$, which is known as the Ewens sampling formula.

We now present a description of the composition of the additional (unobservable) random sample $(X_{n+1},\dots,X_{N-n})$. As recalled in the introduction, this is assumed to be a random sample from the posterior distribution of P_θ given $(X_1,\dots,X_n)$. In particular, due to the conjugacy of the Dirichlet process prior ([4]), the law of P_θ given $(X_1,\dots,X_n)$ is the law of the Dirichlet process $P_{\theta+n,\alpha_0+e_n}$, with $e_n = n^{-1}\sum_{1\leq i\leq n}\delta_{X_i}$. Let $\{Y_1^*,\dots,Y_{J_{N-n}}^*\}$ be the labels of the $J_{N-n} \leq N-n$ distinct types in the additional sample$(X_{n+1},\dots,X_{N-n})$ that do not coincide with any of the X_i^*. Moreover, let $0 \leq V_{N-n} \leq n$ be the number of X_{n+i}'s that do not coincide with any of the X_i^*'s, and set

i) $\mathbf{S}_{N-n} = (S_{1,N-n},\dots,S_{K_n,N-n})$, where $S_{j,N-n}$ is the number of X_{n+i}'s that coincide with the label X_j^*, for any $j = 1,\dots,K_n$, such that $S_{j,n} \geq 0$ and $\sum_{1\leq j\leq K_n} S_{j,n} = n - V_n$;
ii) $\mathbf{R}_{N-n} = (R_{1,N-n},\dots,R_{J_{N-n},N-n})$, where $R_{j,N-n}$ be the number of X_{n+i}'s that coincide with the label Y_j^*, for any $j = 1,\dots,J_{N-n}$, such that $R_{j,n} \geq 1$ and $\sum_{1\leq j\leq J_n} R_{j,n} = V_n$.

The conditional distribution of the random variable $(\mathbf{S}_{N-n},V_{N-n},J_{N-n},\mathbf{R}_{N-n})$ given $(K_n,N_{1,n},\dots,N_{K_n,n})$ models the composition of the additional unobservable sample. We refer to [3] for additional details on $(\mathbf{S}_{N-n},V_{N-n},J_{N-n},\mathbf{R}_{N-n})$, as well as for its numerous distributional properties, i.e. conditional Ewens sampling formula.

Under the Bayesian nonparametric model (2), we consider the problem of estimating ν_1 and τ_1. First, we give a formal definition of ν_1 and τ_1 in terms of the random variables K_n, $N_{i,n}$, J_{N-n}, $S_{i,N-n}$ and $R_{i,N-n}$ introduced above. In particular, we can write

$$\nu_1 = \sum_{i=1}^{K_n} \mathbb{1}_{\{N_{i,n}+S_{i,N-n}=1\}} + \sum_{i=1}^{J_n} \mathbb{1}_{\{R_{i,N-n}=1\}}$$

and

$$\tau_1 = \sum_{i=1}^{K_n} \mathbb{1}_{\{N_{i,n}+S_{i,N-n}=1\}}.$$

We consider the problem of deriving a Bayesian nonparametric estimator of ν_1 and τ_1 with respect a squared loss function. That is, we want to compute the conditional expectation of ν_1 given the sample records $(X_1,\dots,X_n)$, and the conditional expectation of τ_1 given the sample records $(X_1,\dots,X_n)$. These results follows by a direct application of Equation 23 and Equation 24 in [3]. It is worth pointing out that results in [3] can be applied also to derive high-order moments of the posterior distribution. Furthermore, these moments may lead, with additional effort, to the posterior distribution of ν_1 given $(X_1,\ldots,X_n)$. Here, for the sake of simplicity, we focus on the expected value of the posterior distribution.

Proposition 1. Let $(X_1,\ldots,X_n)$ be sample records consisting of $K_n = k_n$ non-empty cells with corresponding frequencies $\mathbf{N}_n = (n_{1,n},\ldots,n_{k_n,n})$ and such that $m_{1,n} = \sum_{1\le i\le k_n} \mathbb{1}_{\{n_{i,n}=1\}}$. Under the Bayesian nonparametric model model (2), one has

$$\mathbb{E}\left[\sum_{i=1}^{K_n} \mathbb{1}_{\{N_{i,n}+S_{i,N-n}=1\}}\,|\,X_1,\ldots,X_n\right] = \frac{\theta+n-1}{\theta+N-1}m_{1,n}$$

and

$$\mathbb{E}\left[\sum_{i=1}^{J_n} \mathbb{1}_{\{R_{i,N-n}=1\}}\,|\,X_1,\ldots,X_n\right] = \frac{\theta}{\theta+N-1}(N-n).$$

Proposition 1 provides a simple closed-form expression for a Bayesian nonparametric estimator, with respect to a squared loss function, of ν_1 and τ_1. In particular, we have

$$\hat{\nu}_1 := \mathbb{E}[\nu_1\,|\,X_1,\ldots,X_n] = \frac{\theta+n-1}{\theta+N-1}m_{1,n} + \frac{\theta}{\theta+N-1}(N-n) \qquad (3)$$

and

$$\hat{\tau}_1 := \mathbb{E}[\tau_1\,|\,X_1,\ldots,X_n] = \frac{\theta+n-1}{\theta+N-1}m_{1,n}. \qquad (4)$$

Observe that the number $M_{1,n}$ of unique sample records is sufficient to estimate the measures τ_1 and ν_1. That is $M_{1,n}$ is the sole information contained in the random sample $(X_1,\ldots,X_n)$ that is required to estimate τ_1 and ν_1. To the best of our knowledge, the estimator $\hat{\nu}_1$ in Equation (3) is the first example of a Bayesian nonparametric estimator of ν_1. With regards to Equation (4), the Bayesian nonparametric estimator $\hat{\tau}_1$ coincides with Samuels' estimator (1). Then, our result provides with a proper Bayesian derivation of Samuels' estimator (1) as posterior expectation under a Dirichlet process prior. While this is an interesting result, it is not surprising due to the well-known interplay between Hoppe's sampling scheme and the Dirichlet prior ([5]).

3 Discussion

We revisited the work of [7] from a Bayesian nonparametric perspective. In particular, we introduced a Bayesian nonparametric estimator of ν_1, and we showed that Samuels' estimator (1) is a Bayesian nonparametric estimator of τ_1, with respect to a squared loss function, under a Dirichlet process prior for P. Then, the estimator $\hat{\tau}_1$ may be viewed as a Bayesian smoothed version of the naive estimator $m_{1,n}n/N$ discussed by [1] and [9]. In particular, as N and n become large, the effect of the smoothing prior parameter θ vanishes, and $\hat{\tau}_1$ becomes approximately the naive estimator. This motivates the underestimation phenomenon discussed in [7]. Our Bayesian derivation of Samuels' estimator (1) introduces a natural way to deal with the problem of uncertainty quantification, which was missing in the original work of [7]. With some effort results in Section 3.1. of [3] can be applied to derive the posterior distribution of ν_1 given the sample records $(X_1,\dots,X_n)$, and the posterior distribution of τ_1 given the sample records $(X_1,\dots,X_n)$. This posterior distributions then provide with natural tools for quantifying uncertainty for the estimators $\hat{\nu}_1$ and $\hat{\tau}_1$.

Given our derivation of Samuels' estimator (1), one may consider the use of different prior distributions for the unknown composition P of the superpopulation. For instance, one may consider the use of the celebrated two-parameter Poisson-Dirichlet prior ([6]). Under this prior assumption one may still relies on results in [3] to derive the posterior distribution of ν_1 given the sample records $(X_1,\dots,X_n)$, and the posterior distribution of τ_1 given the sample records $(X_1,\dots,X_n)$. In general, results [3] provide useful tools to deal with Bayesian nonparametric estimation of ν_1 and τ_1 under sufficiently flexible prior assumptions. It remains an open problem to adapt our Bayesian nonparametric approach to deal with structural zeros. In that regard a concrete direction of research would consist in making use of a Dirichlet process prior with a spike and slab base measure (see, e.g., [8] and [2]). In other terms, the nonatomic base distribution α_0 is replaced by a base distribution of the form $\alpha_0(\zeta) = \zeta\delta_0 + (1-\zeta)\alpha_0$, with $\zeta \in [0,1]$ and α_0 being a nonatomic distribution. The parameter ζ in then used to include the information on structural zeros, taking on the interpretation of the proportion of structural zeros in the population or records.

References

[1] Bethlehem, J.G., Keller, W.J. and Pannekoek, J. (1990). Disclosure control of microdata. *J. Amer. Stat. Ass.*, **85** 38–45.

[2] Canale, A., Lijoi, A., Nipoti, B. and Prünster, I. (2017). On the Pitman-Yor process with spike and slab base measure. *Biometrika*, **104** 681–697.

[3] Favaro, S., Lijoi, A. and Prünster, I. (2013). Conditional formulae for Gibb-type exchangeable random partitions. *Ann. Appl. Probab.*, **23**, 1721–1754.

[4] Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209–230.

[5] Hoppe, F.H. (1984). Pólya-like urns and the Ewens sampling formula. *J. Math. Biol.*, **20**, 91–94.

[6] Pitman, J. (2006). *Combinatorial Stochastic Processes.* Ecole d'Eté de Probabilités de Saint-Flour XXXII. Lecture notes in mathematics, Springer - New York.

[7] Samuels, S.M. (1998). A Bayesian, species-sampling-inspired approach to the uniques problem in microdata disclosure risk assessment. *J. Off. Statist.* **14**, 373–383.

[8] Scarpa, B. and Dunson, D. (2009). Bayesian hierarchical functional data analysis via contaminated informative priors.. *Biometrics*, **65** 772–780.

[9] Skinner, C.J., Marsh, C., Openshaw, S. and Wymer, C. (1994). Disclosure control for census microdata. *J. Off. Statist.*, **10**, 31–51.