BAYESIAN NONPARAMETRIC DISCLOSURE RISK ESTIMATION VIA MIXED EFFECTS LOG-LINEAR MODELS

By Cinzia Carota*, Maurizio Filippone, Roberto Leombruni and Silvia Polettini*

The number of categorical observations that are unique in a sample and also unique, or rare, in the population is commonly used to measure the overall risk of disclosure in the sample data. Many authors have attempted to estimate risk by employing parametric models on cross classifications of the key variables, i.e. multi-way contingency tables of those categorical variables that permit the identification of individuals in the sample. In particular, parametric log-linear models or local smoothing polynomial models have been employed to capture the underlying probability structure of the contingency table. This paper proposes a nonparametric approach assuming a Poisson model with rates explained by a log-linear mixed model with Dirichlet process random effects. Risk estimates are obtained by carrying out a fully Bayesian treatment of the proposed model. The main finding is that parametric all two-way interactions log-linear models and semi-parametric log-linear models with main effects only produce roughly equivalent risk estimates. This fact is observed in applications to real data, and suggests that the latter can be adopted as "default" models, as they are able to produce reasonably good risk estimates and also to defuse potential shortcomings of traditional log-linear models.

1. Introduction. A major concern in releasing files of microdata arising from sample surveys is protecting the privacy of the subjects in the sample. The information contained in the files to be released consists of a set of identifying variables, usually categorical, along with some sensitive variables. The subset of identifying variables whose values in the population are also available to potential intruders from a source which is external to the data under consideration is referred to as the set of key variables. Using the key variables, an intruder with certain knowledge about a subject may identify him/her, thereby learning sensitive information about the subject, carried by the released data. Even if in certain cases synthetic or suitably altered data

*Work partially supported by the project PRIN 2008: New developments in sampling theory and practice, Project number 2008CEFF37, Sector: Economics and Statistics.

Keywords and phrases: Bayesian nonparametric models, confidentiality, disclosure risk, Dirichlet process, log-linear models, mixed effects models

can be released, in other cases sharing record-level data is part of the mission of many organizations, so that estimating disclosure risk measures is one of their obligations. For most socio-demographic survey data, the problem is usually tackled by considering a contingency table representing the cross-classification of individuals by the key variables and observing that if an individual belongs to a cell with small sample frequency an identification disclosure may occur when that combination of values of the key variables is also rare in the population. Often, attention is focused only on cell frequencies of 1 (sample uniques) and common disclosure risk measures are the number of sample uniques which are also population uniques and/or the expected number of correct guesses when each sample unique is matched with an individual randomly chosen from the corresponding population cell. Traditionally those measures are estimated by using parametric models. A common feature of many such models is the assumption of exchangeability of cells of the population contingency table, implying that all cells with the same sample frequency are assigned the same risk estimate. Skinner and Holmes (1998), Fienberg and Makov (1998), Elamir and Skinner (2006), Forster and Webb (2007) and Skinner and Shlomo (2008) introduce a log-linear model for the expected cell frequencies that overcomes this problem. Rinott and Shlomo (2006) and Rinott and Shlomo (2007a), instead, propose a local smoothing polynomial model based on the idea that one can learn about the risk in a given cell from neighbouring cells, if a suitable definition of closeness is possible (as it is, for instance, with ordinal key variables), without relying on a "neighbourhood" of cells determined by a log-linear model. Moreover, in those articles a variety of estimation strategies are suggested, including combinations of methods ranging from maximum likelihood estimates to fully Bayesian estimates, and also a method based on multiple imputation. Recently, Manrique-Vallier and Reiter (2012) have employed a Bayesian version of grade of membership (GoM) models to overcome potential shortcomings of log-linear models, which essentially result in: (i) bias of the risk estimates due to the sparsity of the contingency tables involved, and (ii) difficulties with model choice because of the huge number of competing log-linear models when high-order terms are included.

In this paper we introduce a Bayesian semi-parametric version of log-linear models, which specifically is a mixed effects log-linear model with a Dirichlet process (DP) prior (Ferguson, 1973) for modeling the random effects. On the one hand, suitable specifications of the base measure of the DP allow for useful extensions of many parametric models; here we generalize the one adopted in Skinner and Holmes (1998) and sketch the analog of one of those proposed in Elamir and Skinner (2006). On the other hand,

the assumption of DP random effects gives the modeling flexibility of accommodating any possible clustering of cells in the contingency table of the key variables, with cells in the same cluster receiving the same random effect. A practical consequence is that the huge number of patterns of dependence among cells automatically created by the proposed model may reduce the number of high-order terms required to achieve a satisfactory performance of risk estimators (see, e.g., Dorazio et al., 2008). In particular, in some applications to real data, one of which is presented in Section 4, we observed roughly equivalent risk estimates under semiparametric log-linear models with main effects only (independence models) and parametric all two-way interactions log-linear models. According to Manrique-Vallier and Reiter (2012, p. 1390), the latter "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". Therefore, our nonparametric independence models emerge as default models also able to substantially defuse the above mentioned shortcomings (i) and (ii). By including main effects only, the complexity of model choice is obviously reduced and the probability of observing sample marginal counts which are random zeros drastically falls, thereby reducing the bias of risk estimates. As it is well known, sparsity may prevent the estimation of high-order terms structurally required by traditional log-linear models: maximum likelihood estimates of parameters for cells corresponding to zero sample marginal counts do not exist, and treating the inestimable parameters as if they were zero implies that all other cells of the population contingency table are overestimated, with the consequence that risk measures are underestimated (Skinner and Shlomo, 2008). As in Manrique-Vallier and Reiter (2012), the estimation method we adopt is fully Bayesian and explicitly considers the randomness of population frequencies which thus represents an additional source of variability of risk estimators. In this respect our work is very different from previous works based on log-linear models, including the one by Rinott and Shlomo (2007b) on conditional variances and confidence intervals for disclosure risk measures. We also remark that in a disclosure limitation context, the sample to be released is unique and fixed and, unlike repeated sampling schemes, the Bayesian approach is quite appealing as it casts the risk estimation problem conditional on the observed data.

The outline of the paper is as follows. In Section 2 we define the disclosure risk measures we are interested in, describe the model adopted by Skinner and Holmes (1998) and introduce two possible nonparametric extensions of this model. In the first extension, we assume DP random effects

and, mimicking Skinner and Holmes (1998), we keep the fixed effects constant. Next, we relax the latter assumption and consider a model all effects of which are unknown. In both extended models the total mass parameter of the DP is also unknown and in Section 3 the MCMC methods used for inference are extensively discussed. Finally, in Section 4 we compare parametric and nonparametric models based on a random sample extracted from the population defined by the Italian National Social Security Administration (2004), benchmarking risk estimates against the true values of risks.

2. Semi-parametric Log-linear Models for disclosure risk estimation. In the contingency table of key variables we denote by f_k and F_k the sample and population frequencies in the k-th cell, respectively, and by K the total number of cells. Our goal is to estimate global risks of reidentification, or disclosure risks, defined as

$$\tau_1 = \sum_k^K I(f_k = 1, F_k = 1), \tag{1}$$

i.e. the number of sample uniques which are also population uniques, and

$$\tau_2 = \sum_k^K I(f_k = 1)\frac{1}{F_k}, \tag{2}$$

i.e. the expected number of correct guesses if each sample unique is matched with an individual randomly chosen from the corresponding population cell (see, e.g., Rinott and Shlomo, 2006). Usually these measures are approximated by $\tau_1^* = \sum_{k=1}^K I(f_k = 1)Pr\{F_k = 1|f_k = 1\}$ and $\tau_2^* = \sum_k^K I(f_k = 1)E(1/F_k|f_k = 1)$, i.e. $E(\tau_i|f_1, ..., f_K)$, $i = 1, 2$, under the assumption of cell independence, and estimated by using parametric models, which often are elaborations of the Poisson model.

Although many relevant models in the disclosure literature are parametric, we explore the possibility of dealing with this issue in a Bayesian nonparametric context, specifically extending the model and the estimation strategy introduced by Skinner and Holmes (1998). We briefly review their work and then report the proposed nonparametric extensions. Assuming that $F_k \sim Poisson(\lambda_k)$ and $f_k \sim Poisson(\pi\lambda_k)$ independently for $k = 1, \ldots, K$, Skinner and Holmes model the parameters λ_k through a log-linear model with mixed effects:

$$\lambda_k = e^{\mu_k}, \qquad \mu_k = \mathbf{w}_k'\boldsymbol{\beta} + \phi_k, \tag{3}$$

where $\mathbf{w}_k$ is a $q \times 1$ design vector depending on the values of the key variables in cell k, $\boldsymbol{\beta}$ is a $q \times 1$ parameter vector (typically main effects and low-order

interactions of the key variables), and ϕ_k is a random effect accounting for cell specific deviations. The sampling fraction π is supposed to be known. Finally, as far as the distribution of random effects goes, Skinner and Holmes assume that $\phi_k \sim \ iid\, \mathcal{N}(0, \sigma^2)$. This implies $\lambda_k \sim \ Lognormal(\mathbf{w}_k'\boldsymbol{\beta}, \sigma^2)$, independently for $k = 1, \ldots, K$.

In Skinner and Holmes (1998) the goal is to estimate τ_1^*, whose summands $Pr\{F_k = 1 | f_k = 1\}$, are given by $\tau_{1,k}^* = e^{-(1-\pi)\lambda_k}$. Their estimation strategy is as follows:

- preliminary estimates $(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$ of $\boldsymbol{\beta}$ and σ^2 are obtained from the sample frequencies f_k via iterative proportional fitting and by a conditional application of the moment method respectively;
- the pair $(\mathbf{w}_k'\boldsymbol{\beta}, \sigma^2)$ is substituted by $(\mathbf{w}_k'\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$ in the *Lognormal* prior;
- different estimates of the per-record risk of disclosure $\tau_{1,k}^*$ are derived:

(4) $$\hat{\tau}_{1,k}^* = \frac{\int e^{-\lambda_k} e^{-\frac{1}{2\hat{\sigma}^2}(\log \lambda_k - \mathbf{w}_k'\hat{\boldsymbol{\beta}})^2} d\lambda_k}{\int e^{-\pi\lambda_k} e^{-\frac{1}{2\hat{\sigma}^2}(\log \lambda_k - \mathbf{w}_k'\hat{\boldsymbol{\beta}})^2} d\lambda_k},$$

  obtained from the posterior of λ_k; $\hat{\tau}_{1,k}^* = e^{-(1-\pi)e^{\mathbf{w}_k'\hat{\boldsymbol{\beta}} + \frac{\hat{\sigma}^2}{2}}}$, obtained from the prior expected value of λ_k, and

(5) $$\hat{\tau}_{1,k}^* = e^{-(1-\pi)e^{\mathbf{w}_k'\hat{\boldsymbol{\beta}}}},$$

  obtained ignoring the randomness of λ_k (plug-in estimate).

Equation (4) is an empirical Bayes estimate of $\tau_{1,k}^*$; the second expression is a simplified empirical Bayes estimate of $\tau_{1,k}^*$. Equation (5) is recommended by Skinner and Holmes when the conditional moment method produces negative values of $\hat{\sigma}^2$, which is interpreted as evidence that the simpler model without random effects is more appropriate. In all three cases, this is a two-stage estimation procedure where, in the first stage, the association among cells is exploited to estimate the hyper-parameters of the *Lognormal* prior, while, in the second (and completely separate) stage, the estimates of $\tau_{1,k}^*$ are obtained cell by cell, independently.

More recently, Skinner and Shlomo (2008) resort to a log-linear model without random effects, so that $\tau_{1,k}^*$ is always estimated by equation (5), and similarly a plug-in estimate is used for the summands in τ_2^*, i.e. $\tau_{2,k}^* = E(1/F_k | f_k = 1) = \frac{1}{(1-\pi)\lambda_k}(1 - e^{-(1-\pi)\lambda_k})$. The inclusion of random effects in the log-linear model is also unnecessary according to Elamir and Skinner (2006), who, all other things being equal, take the priors of λ_ks to be independent Gamma distributions instead of Lognormal.

In this paper, we go back to the model defined in Skinner and Holmes (1998) and remove the assumption of normality of random effects. We model the distribution function G of the random effects as unknown and a priori distributed according to a DP $\mathcal{D}$ with base probability measure G_0 and total mass parameter m (Ferguson, 1973),

$$\phi_k|G \sim \ iid\ G, \qquad G \sim \mathcal{D}(m, G_0), \tag{6}$$

where $G_0 = N(\alpha, \sigma^2)$. Since $E(G) = G_0$ and m controls the variance of the process, in practice G_0 specifies one's "best guess" about an underlying model of the variation in ϕ, and m specifies the extent to which G_0 holds.

Two distinct generalizations of the Skinner and Holmes model are presented, based on different specifications of the model parameters. In the first extension, our prior on $\boldsymbol{\beta}$ degenerates at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_{\mathrm{ML}}$, where $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$ is the maximum likelihood estimate of the parameter vector. This extension is directly inspired by both the structure of the model and the estimation strategy in Skinner and Holmes (1998). Therefore, the corresponding risk estimates will be referred to as *nonparametric empirical Bayes* estimates of the risk and represent a generalization of (4). The hyper-parameters (α, σ^2) in the normal base measure G_0 are considered unknown; in the case when $\alpha = 0$ we simply have some variability around the normal model assumed by Skinner and Holmes. In the second extension, we add the uncertainty about $\boldsymbol{\beta}$. To overcome identifiability issues, following Li, Mueller and Lin (2011), we drop the overall effect β_0, referred to as the "intercept term", from $\boldsymbol{\beta}$ and attempt to infer α in G_0 instead.

The clustering induced by the DP prior on the random effects can be seen from a Polya-urn scheme representation of the joint distribution of realizations from a $\mathcal{D}(m, G_0)$ process. Blackwell and MacQueen (1973) provide this as the product of successive conditional distributions having the following form:

$$\phi_i|\phi_1, \ldots, \phi_{i-1}, M \sim \frac{m}{m+i-1} G_0(\phi_i) + \frac{1}{m+i-1} \sum_{k=1}^{i-1} \delta(\phi_k = \phi_i), \tag{7}$$

with $\delta(\cdot)$ denoting the Dirac delta function. The above representation shows that clusters in the K cells of the population contingency table are induced by the existence of a positive probability that a newly generated ϕ_i coincides with a previous one. Moreover, it shows that m, the mass or precision parameter of the DP, affects the number of clusters.

Therefore, under the previous assumptions, in the more general case the likelihood function turns out to be a sum of terms where all possible partitions (clusterings) C of the K cells in c nonempty clusters are considered

(see, e.g., Lo, 1984; Liu, 1996),

$$\sum_{c=1}^{K} \sum_{C:|C|=c} \frac{\Gamma(m)}{\Gamma(m+K)} m^c \prod_{j=1}^{c} \Gamma(n_j) \int p(\mathbf{f}_{(j)}|\boldsymbol{\beta}, \phi_j) dG_0(\phi_j), \tag{8}$$

where $\mathbf{f} = f_1, \ldots, f_K$ and n_j $(1 \leq n_j \leq K)$ denotes the number of cells in the j-th cluster,

$$\frac{\Gamma(m)}{\Gamma(m+K)} m^c \prod_{j=1}^{c} \Gamma(n_j) = Pr\{n_1, \ldots, n_c|C, c\}, \tag{9}$$

and finally

$$p(\mathbf{f}_{(j)}|\boldsymbol{\beta}, \phi_j) = \prod_{k \in \text{cluster } j} \frac{1}{f_k!} e^{\pi f_k(\mathbf{w}_k'\boldsymbol{\beta}+\phi_j)} e^{-e^{\pi(\mathbf{w}_k'\boldsymbol{\beta}+\phi_j)}}. \tag{10}$$

In the likelihood, starting from the latter formula, we can observe that the same random effect is assigned to all cells belonging to the same cluster, i.e. to $\mathbf{f}_{(j)}$, that $Pr\{n_1, \ldots, n_c|C, c\}$ is the multivariate Ewens distribution (MED) of K distinguishable objects, or cells $\{1, \ldots, K\}$ (see Takemura, 1999; Johnson, Kotz and Balakrishnan, 2004, chap. 41), and that the number of clusters in each partition ranges from 1 to K. We stress that the total number of summands in the likelihood is the Bell number, B_K, which is a combinatorial quantity assuming large values even for moderate K; just to fix ideas, when $K = 10$, $B_K = 115975$. The model by Skinner and Holmes (1998) corresponds to just one term (namely, $c = K$) in the likelihood and consequently, even for moderate values of K, our model implies a huge number of additional patterns of dependence among cells. In conclusion, the fixed effects included in the log-linear model imply specific patterns of dependence among cells (for instance, an independence model implies that inference on a given cell depends on all cells sharing a value of a key variable with it, since the sufficient statistics are given by the marginal counts); the addition of independent normal random effects allows for departures from the Poisson log-linear model such as overdispersion, but does not affect the way one can learn about a given cell from other cells. On the contrary, the addition of DP random effects implies that the model encompasses all other nonempty subsets of the K cells and that, for each given partition, a possible relation of dependence among cells in the same subset (i.e. whether or not one can learn from those cells about any fixed cell in such a subset) is explicitly evaluated. This suggests both the potential for the proposed model to improve the risk estimates and the computational complexity associated with

it. Similar advantages, with a similar effort, follow from suitable inclusion of DP random effects in the Elamir and Skinner (2006) model. Moreover, we conjecture that some advantages in terms of risk estimates might also be achieved by introducing DP random effects in the local smoothing polynomial model of Rinott and Shlomo (2006, 2007a), where for each fixed cell k, $k = 1, \ldots, K$, a "more local" neighbourhood, compared to the one implied by a log-linear model, is considered.

In order to describe the computational aspects, for each partition C in c clusters, we introduce a $K \times c$ allocation matrix A such that entries $a_{k,j} = 1$ when the random effect ϕ_k is from cluster j and zero otherwise. Then, setting $\phi_k = \eta_j$ when $\phi_k \in$ cluster j, we have $\boldsymbol{\phi} = A\boldsymbol{\eta}$, and the likelihood can be rewritten as

$$\sum_{c=1}^{K} \sum_{A \in \mathcal{A}_c} \frac{\Gamma(m)}{\Gamma(m+K)} m^c \prod_{j=1}^{c} \Gamma(n_j) \times \int \prod_{k=1}^{K} \frac{1}{f_k!} e^{\pi f_k (\mathbf{w}_k' \boldsymbol{\beta} + (A\boldsymbol{\eta})_k)} e^{-e^{\pi(\mathbf{w}_k' \boldsymbol{\beta} + (A\boldsymbol{\eta})_k)}} dG_0(\eta_1, \ldots, \eta_c), \tag{11}$$

where $\mathcal{A}_c$ is the set of all allocation matrices A and the parameters $\boldsymbol{\eta}$ are independent.

3. Inference. Rather than focusing on τ_1^* and τ_2^*, in this paper we directly estimate τ_1 and τ_2 in a fully Bayesian way. Note that τ_1 and τ_2 depend on the population and sample cell frequencies, and that $F_1, \ldots, F_K$ are unobservable random quantities (parameters), with $F_k|\lambda_k \sim Poisson(\lambda_k)$, $k = 1, \ldots, K$. To perform posterior inference, we consider values of λ_ks drawn from their joint posterior distribution and then values of $F_1, \ldots, F_K$ drawn from the corresponding Poisson distributions. In order to keep the notation uncluttered, let $\boldsymbol{\theta}$ denote the set of all parameters conditioning $\lambda_1, \ldots, \lambda_K$ for each of the models analyzed in this article. A Bayesian treatment of such models amounts to marginalizing out $\boldsymbol{\theta}$ from the terms in the re-identification risks τ_1 and τ_2. In a Monte Carlo sense, this can be achieved by

$$I\{F_k = 1 | f_k = 1, \mathbf{f}\} \simeq \frac{1}{H} \sum_{h=1}^{H} I\{F_k^{(h)} = 1 | f_k = 1, \boldsymbol{\theta}^{(h)}\} = \frac{1}{H} \sum_{h=1}^{H} \tau_{1,k}^{(h)} \tag{12}$$

$$(1/F_k | f_k = 1, \mathbf{f}) \simeq \frac{1}{H} \sum_{h=1}^{H} (1/F_k^{(h)} = 1 | f_k = 1, \boldsymbol{\theta}^{(h)}) = \frac{1}{H} \sum_{h=1}^{H} \tau_{2,k}^{(h)}, \tag{13}$$

where $\{\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(H)}\}$ denotes a set of H samples from the posterior distribution $p(\boldsymbol{\theta}|\mathbf{f})$. This approach takes into account the randomness of both groups of unobservable parameters (λ_ks and F_ks), with twofold consequences. First, since a posteriori the λ_ks are dependent on each other, we avoid the unrealistic assumption underlying the second stage of the estimation procedure of Skinner and Holmes (1998), where the cell risks are treated as if they were independent. Second, since the randomness of the F_ks is also explicitly considered, we obtain risk estimates whose variability depends on the variability of the F_ks as well as the variability of the λ_ks and the association between λ_ks. This means, for instance, that, in the more general case, our conditional (i.e., posterior) variance of τ_1, $Var(\tau_1|\mathbf{f}) = Var\left(\sum_k^K I(f_k = 1)I(F_k = 1|f_k = 1)\right)$, cannot be expressed in the form

$$\sum_k^K I(f_k = 1)Pr\{F_k = 1|f_k = 1\}(1 - Pr\{F_k = 1|f_k = 1\}) \tag{14}$$

as in Rinott and Shlomo (2007b) because of the covariances of the λ_ks. Moreover, we estimate these conditional variances (more precisely, the standard errors, s.e., provided in Section 4, Table 1) by using the posterior distributions of τ_i, $i = 1, 2$, instead of applying a plug-in method.

In order to obtain samples from the posterior distribution $p(\boldsymbol{\theta}|\mathbf{f})$, we propose to use Markov chain Monte Carlo (MCMC) techniques (Neal, 1993). In particular, we propose to use a Gibbs sampler where we sample one group of parameters at a time, namely $\boldsymbol{\beta}|$rest, $\boldsymbol{\phi}|$rest, $m|$rest, $(\alpha, \sigma^2)|$rest. The proposed Gibbs sampler steps are briefly discussed next.

Sampling $\boldsymbol{\beta}$ – Given the form of the Poisson likelihood, it is not possible to sample $\boldsymbol{\beta}$ using an exact Gibbs step, and so called Metropolis within Gibbs samplers need to be employed, whereby a proposal is accepted or rejected according to a Metropolis ratio (Roberts and Rosenthal, 2009). Recent work shows that it is possible to efficiently sample from the posterior distribution of parameters of linear models using so called *manifold MCMC* methods. Briefly, such samplers exploit the curvature of the logarithm of the likelihood $p(\mathbf{f}|\boldsymbol{\beta}, \text{rest})$ by constructing a proposal mechanism on the basis of the Fisher Information matrix (see Girolami and Calderhead, 2011, for further details). In this work we adopt Simplified Manifold Metropolis Adjusted Langevin Algorithm (SMMALA) to sample $\boldsymbol{\beta}$ as previously done in Filippone, Mira and Girolami (2011), which simulates a diffusion on the statistical manifold characterizing $p(\mathbf{f}|\boldsymbol{\beta}, \text{rest})$. Define M to be the metric tensor obtained as the Fisher Information of the model plus the negative Hessian of the prior, and ϵ to be a discretization parameter. SMMALA is essentially a Metropolis-Hastings sampler, with a position dependent proposal

akin to the Newton method in optimization, $p(\boldsymbol{\beta}'|\boldsymbol{\beta}) = \mathcal{N}(\boldsymbol{\beta}'|\boldsymbol{\mu}, \epsilon^2 M^{-1})$, with $\boldsymbol{\mu} = \boldsymbol{\beta} + \frac{\epsilon^2}{2} M^{-1} \nabla_{\boldsymbol{\beta}} \log[p(\mathbf{f}|\boldsymbol{\beta}, \text{rest})]$. Gradient and metric tensor can be computed in linear time in the number of cells K and in cubic time in the size of $\boldsymbol{\beta}$; therefore the method scales well to large data sets but it may be computationally intensive for highly parameterized models.

Sampling $\boldsymbol{\phi}$ – The representation of the random effects through the allocation matrix A makes it possible to apply simple schemes to obtain samples from the posterior of the random effects as extensively discussed in Neal (2000). In this work we adopted Algorithm 5 in Neal (2000), as it is easy to implement and as it achieves satisfactory performance in the given application.

Sampling m – In the literature, it has often been reported that inference in models involving DPs is heavily affected by the mass parameter m, and that setting it by means of Maximum Likelihood is bound to yield poor results (see, e.g., Liu, 1996). Rather than fixing this parameter, we propose to sample from its posterior distribution and to account for uncertainty about it when inferring τ_1 and τ_2. In order to do that, we log-transform m and sample $\psi_m = \log(m)$ instead, using a standard Metropolis-Hastings sampler. To avoid the Metropolis step, the approach of Escobar and West (1994) could also be employed.

Sampling α and σ^2 – Given that we chose a Gaussian base measure, by imposing a Gaussian prior on the mean α and an inverse gamma prior on the variance σ^2 of the base measure, we can exploit conjugacy and obtain the conditional distribution of α and σ^2 in closed form. This yields an exact Gibbs step to sample directly from $p(\alpha, \sigma^2|\text{rest})$.

4. Application to Italian National Social Security Administration data and discussion. To evaluate the performance of the proposed approach, we use data from the 7% microdata sample of the Italian National Social Security Administration, 2004. In the application the $N = 450,238$ individuals in the sample above whose workplace falls into 4 specific geographic areas are treated as the population. We draw a random sample with fraction $\pi = 0.1$, yielding $n = 45,023$, and consider five key variables (number of categories in parentheses), namely area (4), sex (2), age (11), ethnicity (5), and economic activity (9), giving a total of $K = 3,960$ cells. Such variables originate from a study on individual health and psychological well-being (W2H&back: From Work to Health and Back, Project funded by Regione Piemonte 2010). We also reconsider the same key variables except for age that is grouped in 6 bands, giving a smaller table with $K = 2,160$ cells.

In the application we examine several versions of the log-linear model with random effects (3), obtained by combining different specifications of its components. First of all, we consider either a parametric (P) or a nonparametric (NP) specification of the random effects. Under the parametric specification P, the random effects are modeled by a $\mathcal{N}(0, \sigma^2)$ distribution as in Skinner and Holmes (1998). Under the nonparametric specification NP, the random effects are assumed to follow a distribution drawn from a DP, whose base measure is $\mathcal{N}(\alpha, \sigma^2)$; the base measure may be assumed to have $\alpha = 0$, i.e. a zero mean (NP ZM), or more generally an unknown mean (NP UM). The previous options are then combined with different models for describing the fixed effects. In particular, we investigate several standard log-linear specifications, namely a model containing the intercept term only, referred to as the intercept model (O), the independence model (I) and the all two-way interactions model (II). We also consider two nonparametric models with no fixed effects (noF). In all cases, the presence or absence of the intercept β_0 in the linear combination defining the μ_k in (3) is denoted by "Yes" or "No", respectively. Finally, all models are estimated by the fully Bayesian method described in Section 3, except for two nonparametric models, where the prior of fixed effects is taken to be degenerate at $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$; here *Emp* denotes use of empirical Bayes estimates. The latter option is introduced to explore the performance of the estimation strategy adopted in Skinner and Holmes (1998) in the presence of DP random effects. We assume independent and reasonably vague Gaussian priors $\mathcal{N}(0, 10)$ on βs. In turn, the prior on α is taken to be $\mathcal{N}(0, 10)$ and the prior on σ^2 to be invGamma$(1, 1)$. Finally, we assume a Gamma$(1, 1)$ prior on m. Convergence of the chains in the MCMC sampling was checked using the Gelman and Rubin's potential scale reduction factor ($\hat{R}$; Gelman and Rubin, 1992), by running 10 parallel chains comprising $10,000$ iterations and assessing that chains had converged when $\hat{R} < 1.1$ for all the parameters. According to this criterion, all chains converged within a few thousands of iterations that were then discarded before evaluating the risk scores.

Table 1 reports true and estimated values of τ_1 and τ_2 (s.e. in parentheses) for eleven models formed by combining different modeling options as described above. Hereafter, these models will be denoted by labels denoting the selected modeling options. Inspection of Table 1 reveals a good performance of the all two-way interactions model among parametric models, which is in line with what reported in the literature. If, however, we enlarge the context to include nonparametric models, new and interesting findings are as follows:

1. the performance of nonparametric independence models, (NP & I)

TABLE 1

Estimated values of τ_1 and τ_2 by means of $\hat{\tau}_1$ and $\hat{\tau}_2$ for the two settings analysed ($K = 2,160$; $K = 3,960$). True values of the global risks are $\tau_1 = 18$ and $\tau_2 = 50.1$ in the small table, and $\tau_1 = 39$ and $\tau_2 = 94.4$ in the large table.

Model	Intercept	Fixed Effects	$K = 2,160$		$K = 3,960$	
			$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}_1$	$\hat{\tau}_2$
P	Yes	O	**0.0** (0.0)	**1.0** (0.0)	**0.0** (0.0)	**3.2** (0.0)
P	Yes	I	**20.5** (3.1)	**44.6** (2.1)	**32.1** (3.9)	**77.0** (2.8)
P	Yes	II	**21.5** (3.8)	**50.2** (3.2)	**32.5** (4.6)	**84.8** (3.7)
NP Emp UM	Yes	I	**19.6** (3.5)	**48.2** (2.9)	**32.5** (4.4)	**85.8** (3.8)
NP Emp UM	Yes	II	**17.5** (2.9)	**46.1** (2.0)	**26.0** (3.7)	**78.4** (2.6)
NP ZM	No	noF	**8.0** (3.7)	**36.6** (5.0)	**13.4** (5.4)	**68.9** (8.3)
NP ZM	No	I	**22.3** (3.7)	**52.3** (3.0)	**33.2** (4.9)	**87.9** (4.5)
NP ZM	No	II	**20.5** (3.6)	**49.5** (2.9)	**28.0** (3.5)	**80.1** (3.1)
NP UM	No	noF	**9.6** (4.2)	**42.7** (5.0)	**16.5** (5.7)	**76.4** (7.9)
NP UM	No	I	**22.1** (3.8)	**52.0** (3.2))	**32.1** (4.7)	**86.5** (4.2)
NP UM	No	II	**20.2** (3.6)	**48.9** (2.9)	**27.4** (4.0)	**79.1** (3.1)

in our notation, is comparable to that of the parametric all two-way interactions model, (P & II). This means that the DP prior is able to capture the essential features of heterogeneity without increasing the dimensionality of the problem.

2. The potential of the DP prior for capturing latent information not modeled by covariates can be noticed by comparing the results corresponding to the parametric log-linear model that only contains the overall mean, (P & O), and to the nonparametric models without fixed effects, (NP ZM & noF) and (NP UM & noF). The latter is the model used in Dorazio et al. (2008).
3. Because of the vague priors we adopted for $\boldsymbol{\beta}$ and σ^2, the fully Bayesian estimates of τ_1 obtained under the models (P & I) and (P & II) can be considered roughly equivalent to the Empirical Bayes estimates obtained from equation (4). The risk estimates obtained under the models (NP Emp UM & I) and (NP Emp UM & II) represent their nonparametric counterparts, i.e. a first minor generalization of the results in Skinner and Holmes (1998). Of course, using plug-in estimates of $\boldsymbol{\beta}$, a source of uncertainty is neglected; this explains why the corresponding standard errors are slightly smaller than the ones associated with the remaining nonparametric models.

All those findings can also be noticed by inspecting the plots in Figure 1 which presents the 2.5th, 5th, 50th, 95th and 97.5th percentiles of the posterior distribution of τ_i, i=1,2, under each of the models reported in Table 1.

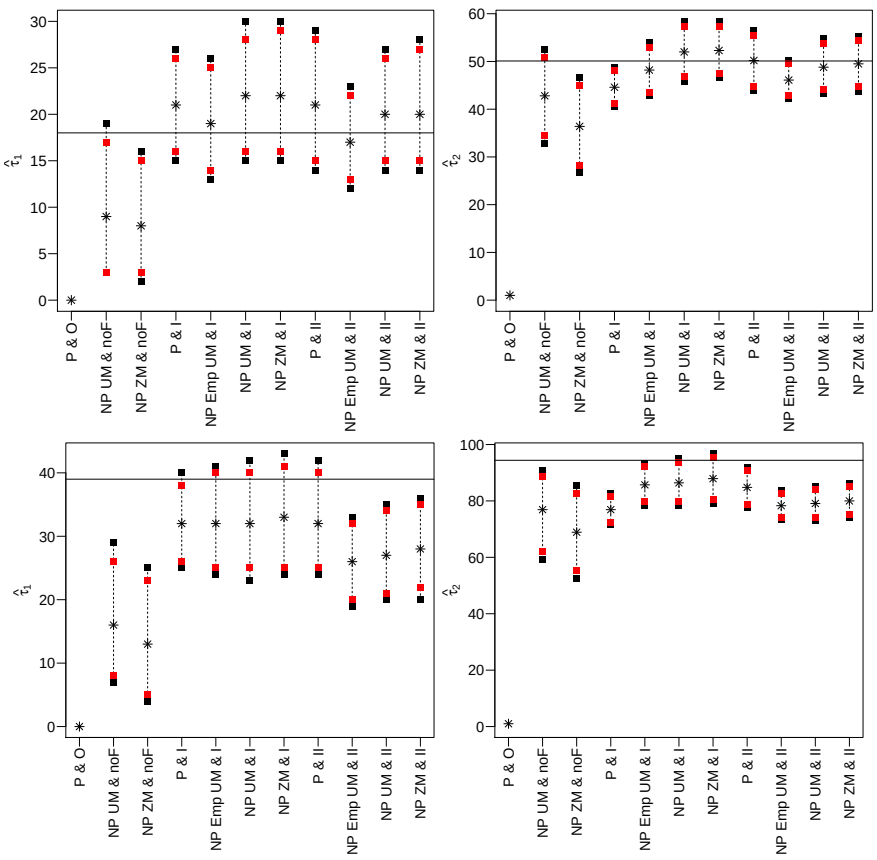

FIG 1. *Quantiles of the posterior distributions of τ_1^* (first column) and τ_2^* (second column) under all parametric and nonparametric models considered. The first (second) row refers to the table with $K = 2,160$ ($K = 3,960$) cells.*

Models appear in order of complexity of the log-linear specification; the solid horizontal lines represent the true risk values. About finding (2), for instance, considering the models without main effects and interaction terms (i.e. the first three models where the exchangeability assumption holds), Figure 1 clearly shows the impact on the posterior distribution of τ_i, $i = 1, 2$, arising from the assumption of DP random effects. Moreover, there is a clear indication that, as the complexity of the log-linear model increases, the variability corresponding to parametric models increases as well, while the variability corresponding to nonparametric models tends to decrease.

The latter fact is even more evident in Table 2 where, all other things being equal, the variability of the F_k is neglected, and we estimate τ_1^* and τ_2^* instead of τ_1 and τ_2. Similarly to what Manrique-Vallier and Reiter have ob-

TABLE 2

Estimated values of τ_1 and τ_2 by means of $\hat{\tau}_1^*$ and $\hat{\tau}_2^*$ for the two settings analysed ($K = 2,160$; $K = 3,960$). True values of the global risks are $\tau_1 = 18$ and $\tau_2 = 50.1$ in the large table and $\tau_1 = 39$ and $\tau_2 = 94.4$ in the small table.

Model	Intercept	Fixed Effects	$K = 2,160$		$K = 3,960$	
			$\hat{\tau}_1^*$	$\hat{\tau}_2^*$	$\hat{\tau}_1^*$	$\hat{\tau}_2^*$
P	Yes	O	**0.0** (0.0)	**1.0** (0.0)	**0.0** (0.0)	**3.2** (0.0)
P	Yes	I	**20.5** (0.7)	**44.6** (0.8)	**32.1** (1.0)	76.9 (1.4)
P	Yes	II	21.6 (2.3)	**50.2** (2.5)	**32.5** (2.4)	**84.8** (2.5)
NP Emp UM	Yes	I	19.5 (1.7)	**48.2** (2.0)	**32.5** (2.1)	**85.8** (2.6)
NP Emp UM	Yes	II	**17.5** (0.2)	**46.1** (0.3)	**26.0** (0.1)	**78.4** (0.2)
NP ZM	No	noF	**8.0** (2.9)	**36.6** (4.7)	**13.5** (4.5)	69.0 (7.9)
NP ZM	No	I	**22.3** (1.9)	**52.3** (2.2)	**33.2** (2.9)	**87.9** (3.6)
NP ZM	No	II	**20.5** (1.9)	**49.5** (2.1)	27.9 (1.1)	**80.1** (1.6)
NP UM	No	noF	**9.6** (3.1)	**42.7** (4.6)	**16.5** (4.5)	**76.4** (7.4)
NP UM	No	I	**22.1** (2.1)	**52.0** (2.4))	**32.1** (2.5)	**86.5** (3.2)
NP UM	No	II	**20.2** (1.9)	**48.9** (2.1)	**27.4** (1.4)	**79.1** (1.6)

served under their GoM models (Manrique-Vallier and Reiter, 2012, p.1389), in Table 2 we observe point estimates $\hat{\tau}_1^*$ and $\hat{\tau}_2^*$ nearly identical to the ones in Table 1, thereby confirming findings (1)-(3), with smaller standard errors since only the variability of λ is taken into account (the slight variations are underlined). Clearly, $\hat{\tau}_i^* = \sum_{k=1}^{K} \hat{\tau}_{i,k}^*$, $i = 1, 2$, where

$$\hat{\tau}_{1,k}^* = \frac{1}{H} \sum_{h=1}^{H} Pr\{F_k = 1 | f_k = 1, \boldsymbol{\theta}^{(h)}\}; \qquad \hat{\tau}_{1,k}^* = \frac{1}{H} \sum_{h=1}^{H} E\left(\frac{1}{F_k} | f_k = 1, \boldsymbol{\theta}^{(h)}\right).$$

By exploiting such a simplification, in the the rest of this Section we explore the behaviour of the per-cell risk estimates $\hat{\tau}_{1,k}^*$ and $\hat{\tau}_{2,k}^*$ to try to better understand our findings (1)-(3).

In Figures 2 and 3 we compare per-cell risk estimates $\hat{\tau}_{i,k}^*$ and true risks (red line) for $i = 1, 2$, respectively. We consider estimates from both the table with $K = 2,160$ cells (first row) and the table with $K = 3,960$ cells (second row) obtained under the parametric models (P & II) and (P & I) and under the three nonparametric independence models in Table 2. Cells containing sample uniques (189 in the first row and 356 in the second row) are arranged in increasing order of the per-cell true risk.

First of all we comment on the performance of nonparametric independence models compared to the other models considered. The per-cell risk estimates $\hat{\tau}_{i,k}^*$ corresponding to the (NP Emp UM & I) model and to the parametric independence model (P & I) perform similarly except for a noticeable reduction of the amplitude of oscillations in the presence of DP ran-

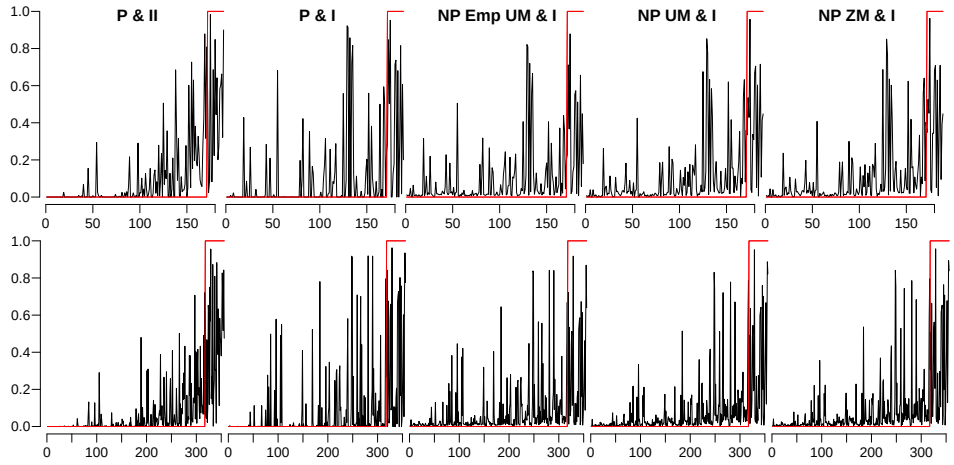

FIG 2. *Comparison of risk estimates* $\hat{\tau}^*_{1,k}$ *for sample unique cells under the all-two-way interactions parametric model (P & II) vs the corresponding estimates under the independence model (P & I) and three nonparametric models of type (NP & I). First (second) row represents estimated risks for the small (large) table with* $K = 2,160$ ($K = 3,960$). *Cells are arranged in decreasing order of population cell size; red line represents the true (0/1) risk.*

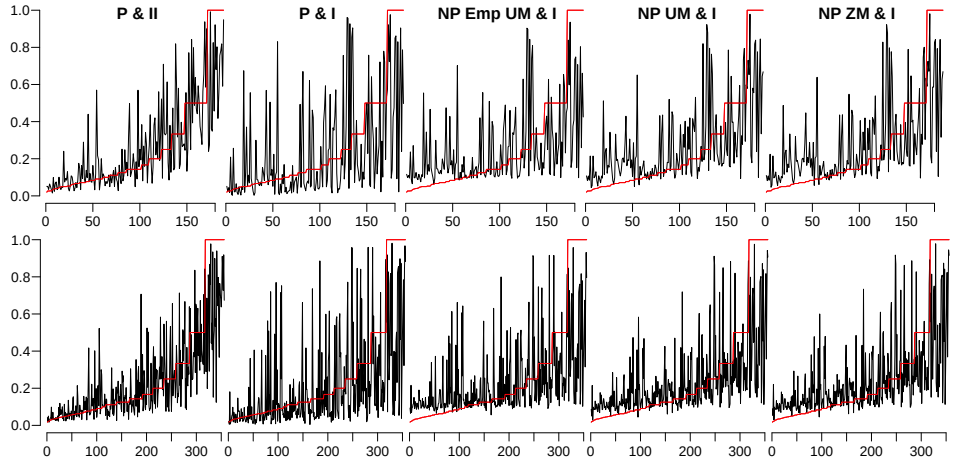

FIG 3. *Comparison of risk estimates* $\hat{\tau}^*_{2,k}$ *for sample unique cells under the all-two-way interactions parametric model (P & II) vs the corresponding estimates under independence model (P & I) and three nonparametric models of type (NP & I). First (second) row represents estimated risks for the small (large) table with* $K = 2,160$ ($K = 3,960$). *Cells are arranged in increasing order of true per-cell risk (red line).*

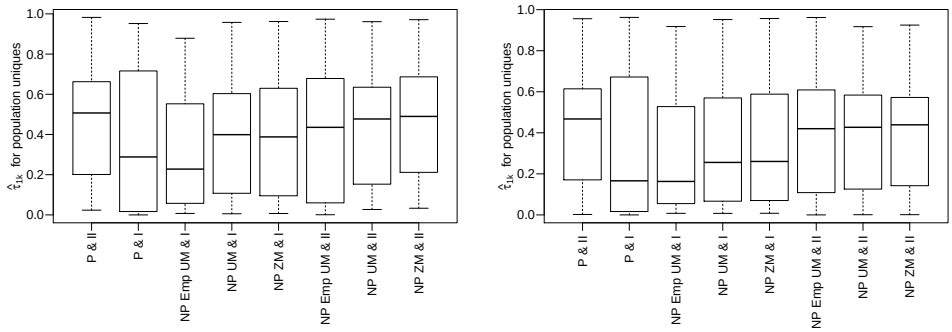

FIG 4. *Estimates of $\tau_{1,k}$ for population uniques only, under all models with covariates. Left (right) panel refers to the table with $K = 2,160$ ($K = 3,960$) cells.*

dom effects. In addition to smaller oscillations, the per-cell risk estimates $\hat{\tau}^*_{i,k}$ corresponding to the (NP UM & I) and (NP ZM & I) models — the fully Bayesian models where the DP random effects also affects estimation of fixed effects — exhibit an overall increasing trend as the population cell size decreases. This, compared to what observed under the all two-way interactions model (P & II), suggests that the DP random effects in some sense supplement the covariates in the log-linear model. Moreover, as to finding (1), inspection of Figure 3 reveals that the small differences between the global risks $\hat{\tau}^*_2$ found in Table 2 under the (P & II) and the (NP & I) models can be ascribed mainly to cells with very low risks, that are systematically over-estimated by the nonparametric independence models. See also Table 3 where the signed, absolute and squared errors for the estimation of $\tau_{2,k}$ are presented for both the small ($K = 2,160$) and large table ($K = 3,960$).

On the other hand, comparison of parametric and nonparametric independence models, (P & I) and (NP & I), in Figure 3 indicates a large improvement in per-cell risk estimates under the (NP & I) models, since in this case smaller oscillations are equivalent to smaller distances between risk estimates and true risks. See also Table 3. It is worth noting that such an improvement is achieved at the cost of one or two additional unknown parameters (m or m and β_0 respectively), while (P & II) requires hundreds of additional parameters with respect to (P & I), namely 357 and 257 interactions, gross of the unidentifiable ones, for the large and small table, respectively. Moreover, in Table 3 we can observe that, going from nonparametric independence models to nonparametric all two-way interactions models, the improvement of per-cell risk estimates for low risk cells tends to be greater than the improvement of per-cell risk estimates for high risk

TABLE 3

Signed, absolute and squared errors for the estimation of τ_{2k} in the small ($K = 2,160$) and large ($K = 3,960$) table using the models under analysis (above and below the horizontal line, respectively). The table also reports the same summaries restricted to cells having large ($F_k > 10$), and small ($F_k \leq 10$) frequency in the population, respectively.

Model	all cells			cells s. t. $F_k > 10$			cells s. t. $F_k \leq 10$		
	sign.	abs.	sq.	sign.	abs.	sq.	sign.	abs.	sq.
P & I	-5.6	52.9	27.9	9.4	10.9	5.4	-15.0	42.0	22.5
NP Emp UM & I	-1.9	47.3	22.3	10.8	10.8	4.2	-12.7	36.5	18.2
NP UM & I	1.8	46.9	21.7	10.6	10.6	4.1	-8.7	36.3	17.6
NP ZM & I	2.2	47.1	21.8	10.7	10.7	4.2	-8.5	36.4	17.7
P & II	0.1	45.4	21.8	7.7	8.5	3.2	-7.6	36.9	18.6
NP Emp UM & II	-4.0	43.5	21.1	6.8	7.7	2.8	-10.8	35.8	18.3
NP UM & II	-1.3	44.6	21.2	7.7	8.5	3.1	-9.0	36.2	18.1
NP ZM & II	-0.7	44.6	21.1	7.6	8.2	2.9	-8.2	36.4	18.2
P&I	-17.4	94.7	51.6	18.1	19.8	9.8	-35.5	75.0	41.8
NP Emp UM & I	-8.6	85.9	42.7	20.4	20.4	7.9	-29.0	65.5	34.8
NP UM & I	-7.9	80.5	38.8	18.1	18.1	6.8	-26.0	62.4	32.0
NP ZM & I	-6.5	81.4	39.3	18.7	18.7	7.1	-25.2	62.7	32.2
P & II	-9.6	84.7	41.8	14.3	15.3	5.2	-23.9	69.4	36.6
NP Emp UM & II	-16.0	81.8	41.2	12.8	13.9	4.7	-28.7	68.0	36.5
NP UM & II	-15.3	81.4	40.2	12.9	13.9	4.6	-28.1	67.5	35.6
NP ZM & II	-14.3	81.7	40.2	13.0	14.0	4.6	-27.3	67.6	35.6

cells. This fact may have a negative impact at level of global risk estimates inducing bias, as in the large table ($K = 3,960$); see also Figure 1. This fact can be even more clearly noticed in Figure 4, which presents boxplots for the estimated values of $\tau^*_{1,k}$ corresponding to population uniques only, under all models considered in Table 2. In the right-hand side of the Figure we can see that, when the true cell risks are 1, the distributions of cell risk estimates under the parametric and nonparametric all two-way interactions models are very similar. Therefore, the worse performance of global risk estimates we observed in Table 2 under the (NP & II) models than under the (P & II) model (see also Figure 1) is an unpleasant consequence of the greater improvement in cell risk estimates for cells where the true risk is zero achieved under the (NP & II) models than under the (P & II) model. In other words, under the (P & II) model, the over-estimation of cell risks whose true value is zero tends to balance the under-estimation of cell risks whose true values is one. A similar argument explains the good performance of the global risk estimates under the nonparametric independence models. In this respect, however, the nonparametric models we have considered are not equivalent to each other, the performance of the (NP Emp & I) model being less convincing. Further evidence of these facts is given in Figures 5

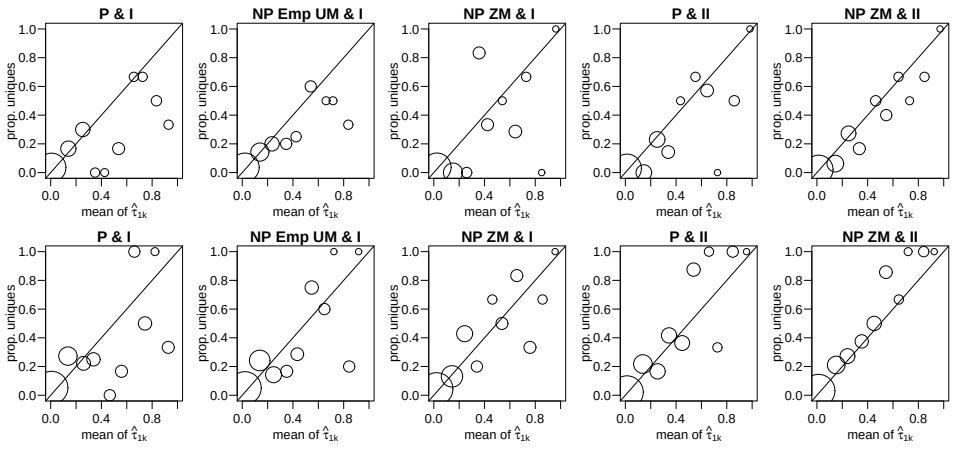

FIG 5. *Proportion of population uniques plotted against the average estimated risk* $\hat{\tau}^*_{1,k}$*, for cells categorized into 10 equal-width intervals according to the values of* $\hat{\tau}^*_{1,k}$*. The size of the plotting points depends on the number of cells in each interval. First (second) line refers to the table with* $K = 2,160$ *(*$K = 3,960$*) cells.*

and 6. In Figure 5 we consider a subset of the models presented in Table 2 and, as in Forster and Webb (2007), we plot the proportion of population uniques against the average value of $\hat{\tau}^*_{1,k}$, for cells categorized into 10 equal-width intervals according to the values of $\hat{\tau}^*_{1,k}$. Similarly, in Figure 6, as in Elamir and Skinner (2006) we plot the mean of $1/F_k$ against the mean of the estimated risk $\hat{\tau}^*_{2,k}$ after grouping cells into 10 intervals according to the values of $\hat{\tau}^*_{2,k}$.

Finally, Figure 7 and Figure 8 present the per-cell risk estimates $\hat{\tau}^*_{i,k}$ under parametric and nonparametric all two-ways interaction models, (P & II) and (NP & II), for $i = 1$ and $i = 2$, respectively. If we consider the deviances of the per-cell risk estimates $\hat{\tau}^*_{i,k}$ around their mean $\hat{\tau}^*_i/K$ for the four models in these figures, we obtain very similar values and, given that the $\hat{\tau}^*_{i,k}$ are in turn means within each cell, from

$$(s.e.(\hat{\tau}^*_i))^2 = \frac{1}{H}\sum_h^H \Big(\sum_k^K \tau^{*h}_{i,k} - \hat{\tau}^*_i\Big)^2 = A + B + C, \tag{15}$$

where

$$A = \sum_k^K \frac{1}{H}\sum_h^H (\tau^{*h}_{i,k})^2 - \sum_k^K (\hat{\tau}^*_{i,k})^2 \tag{16}$$

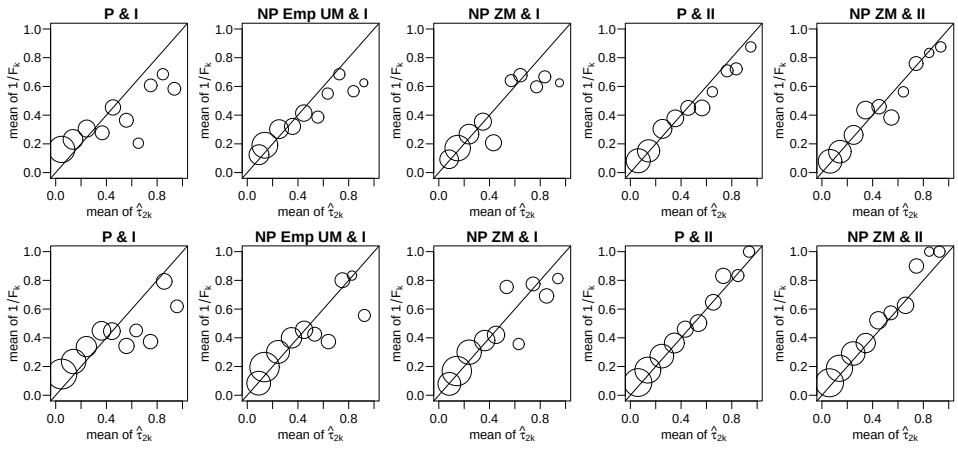

FIG 6. *Mean of* $1/F_k$ *against the mean of the estimated risk* $\hat{\tau}^*_{2,k}$*, for cells categorized into 10 equal-width intervals according to the values of* $\hat{\tau}^*_{1,k}$*. The size of the plotting points depends on the number of cells in each interval. First (second) line refers to the table with* $K = 2,160$ *(*$K = 3,960$*) cells.*

represents the sum of the variances within each cell,

$$B = \sum_{k}^{K} (\hat{\tau}^*_{i,k})^2 - K\Big(\frac{\hat{\tau}^*_i}{K}\Big)^2 \tag{17}$$

is the deviance between cells and

$$C = \sum_{k}^{K} \sum_{j \neq k}^{K} \frac{1}{H} \sum_{h}^{H} \tau^{*h}_{i,k} \tau^{*h}_{i,j} - K(K-1)\Big(\frac{\hat{\tau}^*_i}{K}\Big)^2 \tag{18}$$

is the sum of codeviances between cells, we can conclude that the smaller standard errors observed in Table 2 under the (NP & II) models (compared to the one under the (P & II) model) are essentially due to a reduction of the variances within cells and/or to codeviances between per-cell risks. On the other hand, the components A and C are essentially the only dominant factors of the s.e. under the nonparametric models without fixed effects ($B < 0.001$), and, for nonparametric models in general, the decrease of A and/or C as the complexity of the log-linear model increases prevails over the slight increase of B. Vice versa, going from the parametric independence model (P & I) to the all two-way interactions model (P & II), the component B slightly decreases and is overwhelmed by the increase of A and/or C. In this respect, consider that with parametric models the only way to increase the association between cells is by means the introduction of further covariates (interaction terms).

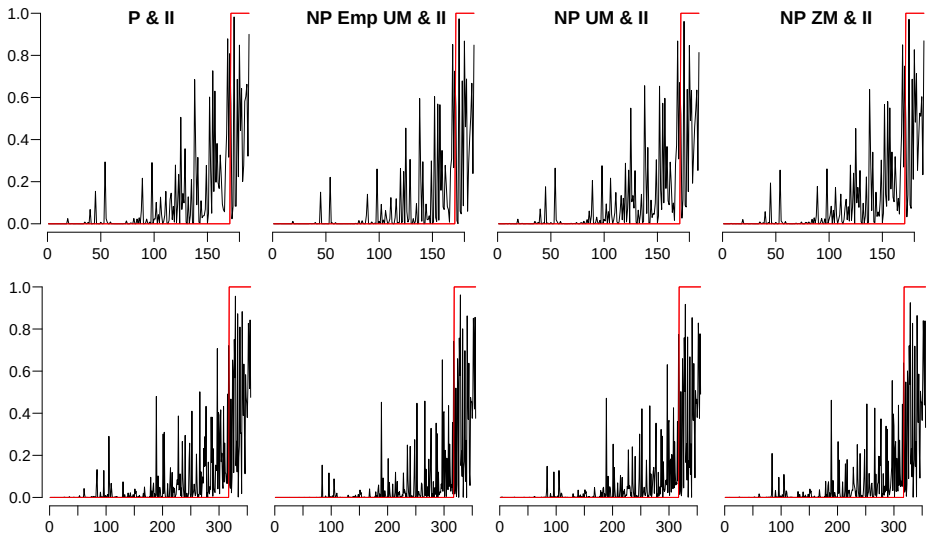

FIG 7. *Risk estimates* $\hat{\tau}^*_{1,k}$ *for sample unique cells under the all-two-way interactions parametric and nonparametric models (P & II, NP & II). First (second) row represents estimated risks for the small (large) table with* $K = 2,160$ ($K = 3,960$). *Cells are arranged in increasing order of true per-cell risk (red line).*

In conclusion, unlike the (P & II) model, for each cell the nonparametric independence models (NP & I) combine learning from two neighbourhoods of cells, one driven by the log-linear model, and one driven by the data and implied by the clustering of the random effects. In turn, this reduces the need for further covariates. In this respect, it is also worth saying that in our application the MCMC output for the nonparametric models shows that, as the complexity of the log-linear model increases, a posteriori the average number of clusters decreases (results not reported). This is a further indication that clusters play a supplementary role with respect to covariates in the log-linear model. Prospectively, moreover, we are induced to expect that the inclusion of DP random effects in a local smoothing polynomial model could improve the global risk estimates in a similar way, that is producing roughly similar risk estimates in the presence of a polynomial of lower degree than the one required without random effects or with normal random effects.

Even if the paper's primary interest is analyzing the impact of introducing DP random effects in the log-linear model (3) and model choice is beyond the scope of this paper, we observe that the (NP Emp UM & I), that represent the direct extension of the model presented in Skinner and Holmes (1998), produces good global risk estimates. Compared to this model, where two completely different estimation methods are combined as done in the largest

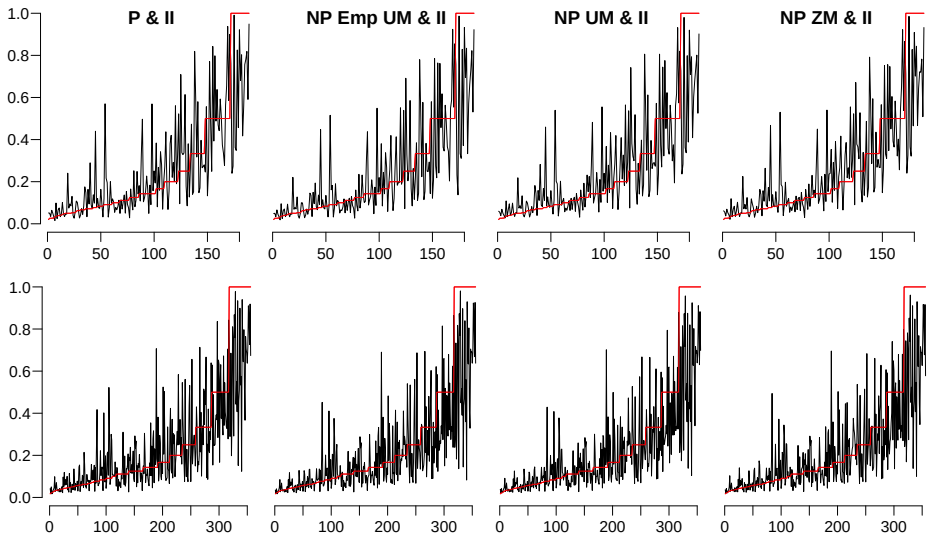

FIG 8. *Risk estimates* $\hat{\tau}^*_{2,k}$ *for sample unique cells under the all-two-way interactions parametric and nonparametric models (P & II, NP & II).First (second) row represents estimated risks for the small (large) table with* $K = 2,160$ ($K = 3,960$). *Cells are arranged in increasing order of true per-cell risk (red line).*

part of the literature, the fully Bayesian independence nonparametric models are also able to produce better estimates of the per-cell risks. Moreover for all our estimators a full account of all sources of uncertainty is produced.

References.

BLACKWELL, D. and MACQUEEN, J. B. (1973). Ferguson distributions via Pólya urn schemes. *The Annals of Statistics* **1** 353–355.

DORAZIO, R. M., MUKHERJEE, B., ZHANG, L., GHOSH, M., JELKS, H. L. and JORDAN, F. (2008). Modeling unobserved sources of heterogeneity in animal abundance using a Dirichlet process prior. *Biometrics* **64** 635–644.

ELAMIR, E. A. H. and SKINNER, C. J. (2006). Record Level Measures of Disclosure Risk for Survey Microdata. *Journal of Official Statistics* **22** 525–539.

ESCOBAR, M. D. and WEST, M. (1994). Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association* **90** 577–588.

FERGUSON, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics* **1** 209–230.

FIENBERG, S. E. and MAKOV, U. E. (1998). Confidentiality, Uniqueness, and Disclosure Limitation for Categorical Data. *Journal of Official Statistics* **14** 385–397.

FILIPPONE, M., MIRA, A. and GIROLAMI, M. (2011). Discussion of the paper: "Sampling schemes for generalized linear Dirichlet process random effects models" by M. Kyung, J. Gill, and G. Casella. *Statistical Methods & Applications* **20** 295-297.

FORSTER, J. J. and WEBB, E. L. (2007). Bayesian disclosure risk assessment: predicting small frequencies in contingency tables. *Journal of the Royal Statistical Society: Series C* **56** 551–570.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* **7** 457–472.

Girolami, M. and Calderhead, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B* **73** 123–214.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (2004). *Discrete Multivariate Distributions. Wiley series in probability and statistics*. John Wiley & Sons.

Li, Y., Mueller, P. and Lin, X. (2011). Center-adjusted inference for a nonparametric Bayesian random effect distribution. *Statistica Sinica* **21** 1201-1223.

Liu, J. S. (1996). Nonparametric Hierarchical Bayes via Sequential Imputations. *Annals of Statistics* **24** 911–930.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates. I. Density estimates. *Annals of Statistics* **12** 351–357.

Manrique-Vallier, D. and Reiter, J. P. (2012). Estimating Identification Disclosure Risk Using Mixed Membership Models. *Journal of the American Statistical Association* **107** 1385–1394.

Neal, R. M. (1993). Probabilistic Inference using Markov chain Monte Carlo Methods Technical Report No. CRG-TR-93-1, Dept. of Computer Science, University of Toronto.

Neal, R. M. (2000). Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics* **9** 249–265.

Rinott, Y. and Shlomo, N. (2006). A Generalized Negative Binomial Smoothing Model for Sample Disclosure Risk Estimation. In *Privacy in Statistical Databases*, (J. Domingo-Ferrer and L. Franconi, eds.). *Lecture Notes in Computer Science* **4302** 82-93. Springer Berlin / Heidelberg.

Rinott, Y. and Shlomo, N. (2007a). A Smoothing Model for Sample Disclosure Risk Estimation. In *Complex Datasets and Inverse Problems: Tomography, Networks and Beyond*, (R. Liu, W. Strawderman and C.-H. Zhang, eds.). *Lecture Notes–Monograph Series* **54** 161–171. Institute of Mathematical Statistics.

Rinott, Y. and Shlomo, N. (2007b). Variances and confidence intervals for sample disclosure risk measures. In *Proceedings of the 56th Session of the ISI*.

Roberts, G. O. and Rosenthal, J. S. (2009). Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics* **18** 349–367.

Skinner, C. J. and Holmes, D. J. (1998). Estimating the re-identification risk per record in microdata. *Journal of Official Statistics* **14** 361–372.

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

Takemura, A. (1999). Some superpopulation models for estimating the number of population uniques. In *Proceedings of the Conference on Statistical Data Protection* 45–58. Eurostat, Luxembourg.

Dipartimento di Economia e Statistica
Università di Torino
E-mail: cinzia.carota@unito.it
roberto.leombruni@unito.it

School of Computing Science
University of Glasgow
E-mail: maurizio.filippone@glasgow.ac.uk

Dipartimento di Metodi e modelli per
l'economia, il territorio e la finanza
Università di Roma "La Sapienza"
E-mail: silvia.polettini@uniroma1.it