The Annals of Applied Statistics
2015, Vol. 9, No. 1, 525–546
DOI: 10.1214/15-AOAS807

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

BY CINZIA CAROTA[*,1], MAURIZIO FILIPPONE[†], ROBERTO LEOMBRUNI[*,1] AND SILVIA POLETTINI[‡,2]

Università di Torino, University of Glasgow*[†] *and Università di Roma "La Sapienza"*[‡]

Statistical agencies and other institutions collect data under the promise to protect the confidentiality of respondents. When releasing microdata samples, the risk that records can be identified must be assessed. To this aim, a widely adopted approach is to isolate categorical variables key to the identification and analyze multi-way contingency tables of such variables. Common disclosure risk measures focus on sample unique cells in these tables and adopt parametric log-linear models as the standard statistical tools for the problem. Such models often have to deal with large and extremely sparse tables that pose a number of challenges to risk estimation. This paper proposes to overcome these problems by studying nonparametric alternatives based on Dirichlet process random effects. The main finding is that the inclusion of such random effects allows us to reduce considerably the number of fixed effects required to achieve reliable risk estimates. This is studied on applications to real data, suggesting, in particular, that our mixed models with main effects only produce roughly equivalent estimates compared to the all two-way interactions models, and are effective in defusing potential shortcomings of traditional log-linear models. This paper adopts a fully Bayesian approach that accounts for all sources of uncertainty, including that about the population frequencies, and supplies unconditional (posterior) variances and credible intervals.

1. Introduction. Statistical agencies and other institutions that release data arising from sample surveys are obliged to protect the confidentiality of respondent's identities and sensitive attributes. In socio-demographic surveys the observed variables are often categorical; some of these, called *key variables*, are identifying in that, being also available in external databases, allow potential intruders to disclose confidential information on records in the sample by matching on such keys. Assuming that there are no errors in the variables above, the problem of assessing disclosure risks associated with any proposed data release is often tackled by: (i) considering a contingency table representing the cross-classification of subjects by the key variables (often this is a very large and sparse table); (ii) observing

Received June 2013; revised October 2014.
[1]Supported in part from UniTo Project TO_Call3_2012_0119 "The Popart Network."
[2]Supported in part from Sapienza University by Grant C26A14PNSC.
Key words and phrases. Bayesian nonparametric models, confidentiality, disclosure risk, Dirichlet process, log-linear models, mixed effects models.

that a subject belonging to a cell with a sample frequency of 1 (sample unique) is at a relatively high risk of identification if there are few subjects in the population with that combination of the key variables.

Common disclosure risk measures are the number of sample uniques which are also population uniques and the expected number of correct guesses when each sample unique is matched with a subject randomly chosen from the corresponding population cell. Further measures can be found in Forster and Webb (2007) along with an extensive survey of the previous literature; in this paper we selectively review only those references that are closely related to the focus of our work.

Disclosure risk is traditionally estimated by parametric models; in this context, Skinner and Holmes (1998), Fienberg and Makov (1998), Carlson (2002), 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 the assumption of exchangeability of cells of the contingency table, implying constant risk estimates across cells having the same sample frequency. To learn about the risk in a given cell from neighboring cells without relying on the association structure implied by a log-linear model, Rinott and Shlomo (2006, 2007a) propose a local smoothing polynomial model, applicable to key variables for which a suitable definition of closeness is available. As far as estimation goes, the literature presents a whole variety of strategies, including combinations of methods ranging from maximum likelihood estimates to fully Bayesian estimates, and also a method based on multiple imputation.

Drawing from the above-mentioned literature, we propose a Bayesian semiparametric 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. As in Fienberg and Makov (1998), Forster and Webb (2007), and Manrique-Vallier and Reiter (2012, 2014), we adopt a fully Bayesian approach. Unlike repeated sampling schemes, the Bayesian framework is particularly appealing in a disclosure limitation context, where the sample to be released is unique and fixed. It also allows us to account for uncertainty about 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), as we provide unconditional variances and credible intervals for sample disclosure risk measures.

Emphasizing the random effects component of the model, we will refer to it as a *nonparametric* log-linear model, its *parametric* counterpart being a log-linear model with random effects modeled parametrically; fixed effects are always assigned a parametric prior, so no further distinctions are necessary.[3] Our nonparametric log-linear models are special cases of the family of hierarchical DPs [Teh

[3]The reason for such and related abuses of terminology is that often in the course of the paper we think of random effects conditionally on fixed effects and vice versa. This is also why we refer to *independence* in the sequel.

et al. (2006)] which also include some elements of the class of mixed membership models [which in turn include grade of membership models, Erosheva, Fienberg and Joutard (2007)] such as latent Dirichlet allocation models [Blei, Ng and Jordan (2003)].

The proposed nonparametric formulation has two major advantages. First, it may be interpreted as the nonparametric extension of some of the parametric models proposed in the literature (see Section 2). Second, and most importantly, in many applications to real data, two of which are presented in Section 4, we observed roughly equivalent global risk estimates under nonparametric log-linear models with main effects only (say, nonparametric *independence* models) compared to all two-way interactions log-linear models with and without random effects. Quoting Manrique-Vallier and Reiter [(2012), page 1390], the latter "have been found to produce reasonable results in many cases [Elamir and Skinner (2006), Fienberg and Makov (1998), Skinner and Shlomo (2008)], and so represent a default modeling position." Consequently, our main finding is that our nonparametric independence models can be used as default models, thereby avoiding the severe difficulties associated with complex log-linear model estimation in the presence of sparse tables [see, e.g., Fienberg and Rinaldo (2012)]. These difficulties arise from certain patterns of sampling zeroes which make the model nonidentifiable and result in nonexistent maximum likelihood estimators (MLE). This fact has long been known [Haberman (1974)], but recent research shows that nonexistent MLEs are likely to arise even in small tables, in the presence of positive margins and in frequently used models such as the all two-way interactions model. "Under a nonexistent MLE, the model is not identifiable, the asymptotic standard errors are not well defined and the number of degrees of freedom becomes meaningless" [Fienberg and Rinaldo (2012), page 997]. Moreover, common statistical packages are inadequate to cope with this problem, as detailed in Fienberg and Rinaldo (2007). The issue of nonexistence of MLE is also important in Bayesian analysis of contingency tables, but in our nonparametric models it is defused in two ways. First, the only fixed effects to be estimated are the main effects. This is a substantial simplification of the log-linear model significantly reducing the severity of the problem. Second, the vague prior we assign to fixed effects replaces the information content lacking in the data with the information contained in the prior about all cells. This obviates the need for ad hoc additions of small positive quantities to cells containing sampling zeroes [Fienberg and Rinaldo (2007), page 3437; Fienberg and Rinaldo (2012), page 1012], which is potentially severely misleading.

Recently, under the assumption that there are no structural zeroes in the contingency table, Manrique-Vallier and Reiter (2012) employ a Bayesian version of the grade of membership model for disclosure risk estimation, also discussing the model choice. This is a very challenging problem in complex log-linear models only addressed in Skinner and Shlomo (2008); another approach is Bayesian model averaging, pursued by Forster and Webb (2007) on decomposable graphical

models. In a subsequent paper, Manrique-Vallier and Reiter (2014) propose a truncated latent class model (LCM) for managing structural zeroes, thereby removing a traditional limitation of Bayesian latent structure models.

The paper is organized as follows: in Section 2 we define our model and interpret it in light of the existing literature; in Section 3 we describe in detail our estimation method. In Section 4 we compare parametric and nonparametric models based on a sample extracted from the population defined by the Italian National Social Security Administration (WHIP-Work Histories Italian Panel, Laboratorio Revelli, Centre for Employment Studies, http://www.laboratoriorevelli.it/whip), benchmarking risk estimates against the true values of global risks. The same comparison is also provided through a random sample from public use microdata from the state of California [IPUMS, Ruggles et al. (2010)]. In Section 5 we discuss comparisons between our nonparametric models and the LCMs of Manrique-Vallier and Reiter (2014), showing that both rely on the same basic assumptions, although implemented in different ways, which leads to different models with relative merits over each other. We also discuss some computational aspects, suggesting use of the Empirical Bayesian version of our model to reduce the computational burden for very large tables. Finally, in Section 6 we provide some final comments.

2. Log-linear models for disclosure risk estimation. Let f_k and F_k denote the sample and population frequencies in the kth cell, respectively, and let K be the total number of cells in the contingency table of the key variables. Our goal is to estimate global risks of re-identification, or disclosure risks, defined as

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

that is, the number of sample uniques which are also population uniques, and

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

that is, 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 their expectations $E(\tau_i | f_1, \ldots, f_K)$, $i = 1, 2$, namely, under the assumption of cell independence,

$$\tau_1^* = \sum_{k=1}^{K} I(f_k = 1)\Pr\{F_k = 1 | f_k = 1\} = \sum_{k=1}^{K} I(f_k = 1)\tau_{1,k}^*, \tag{3}$$

$$\tau_2^* = \sum_{k=1}^{K} I(f_k = 1)E(1/F_k | f_k = 1) = \sum_{k=1}^{K} I(f_k = 1)\tau_{2,k}^*, \tag{4}$$

and estimated by using parametric models, which are often elaborations of the Poisson model. Assuming

$$F_k \sim \text{Poisson}(\lambda_k) \quad \text{and} \quad f_k \sim \text{Poisson}(\pi\lambda_k) \tag{5}$$

independently for $k = 1, \ldots, K$, with π being the (known) sampling fraction, the terms in (3) and (4) can be expressed in closed form,

$$\tau^*_{1,k} = e^{-(1-\pi)\lambda_k}, \qquad \tau^*_{2,k} = \bigl(1 - e^{-(1-\pi)\lambda_k}\bigr)/\bigl((1-\pi)\lambda_k\bigr). \tag{6}$$

In a relevant part of the literature the Poisson assumption is integrated by log-linear modeling of cell means and, as mentioned in Section 1, the all two-way interactions model without random effects has been recognized as a useful default model by many authors [Elamir and Skinner (2006), Fienberg and Makov (1998), Skinner and Shlomo (2008)]; recent articles, however, show that inference in this model is not trivial with sparse tables. Even if the parameters of interest are the cell means λ_k, and the Iterative Proportional Fitting (IPF) is guaranteed to converge to the extended MLE by construction, the rate of convergence, with the noticeable exception of decomposable graphical models, can be very slow when the MLE is not defined. In conclusion, "The behavior of IPF when the MLE does not exist has not been carefully studied to date" [Fienberg and Rinaldo (2007), page 3438]. The previous facts, along with the nature of the problem, motivate our attempt to address it in a Bayesian nonparametric framework by introducing DP random effects. The assumption of a DP prior gives the modeling flexibility of accommodating any possible clustering of cells in the contingency table of the key variables, and implies that all possible clusters of cells are considered, 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 DP prior may reduce the number of high-order terms required in the log-linear model to achieve a satisfactory performance of risk estimators [see, e.g., Dorazio et al. (2008)]. Aiming at exploring this idea in real data applications, we build on work by Skinner and Holmes (1998) and related papers, such as Elamir and Skinner (2006) and Carlson (2002). Before describing our proposal, we briefly review the above references. Assuming (5), Skinner and Holmes model the parameters λ_k through a log-linear model with mixed effects:

$$\lambda_k = e^{\mu_k}, \qquad \mu_k = \mathbf{w}_k'\boldsymbol{\beta} + \phi_k, \tag{7}$$

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. Finally, ϕ_k i.i.d. $\sim \mathcal{N}(0, \sigma^2)$. Formula (7) can be re-expressed using multiplicative random effects as $\lambda_k = e^{\mathbf{w}_k'\boldsymbol{\beta}} e^{\phi_k} = \xi_k\omega_k$, hence $\lambda_k|(\boldsymbol{\beta}, \sigma^2) \sim \text{Lognormal}(\mathbf{w}_k'\boldsymbol{\beta}, \sigma^2)$, independently for $k = 1, \ldots, K$.

Skinner and Holmes (1998) estimate τ_1^* of formula (3) by a two-stage procedure: in the first stage, the association among cells is exploited to estimate the

hyperparameters $\boldsymbol{\beta}$ and σ^2 of the Lognormal prior; in the second (and completely separate) stage, estimates of $\tau^*_{1,k}$ are obtained cell by cell, independently. When the preliminary estimate of σ^2 is positive, this procedure leads to empirical Bayes estimates of the $\tau^*_{1,k}$'s in (6), otherwise the random effects ϕ_k's are removed, and plug-in estimates of the $\tau^*_{1,k}$'s are derived by using ML estimates $\hat{\xi}_k = e^{\mathbf{w}'_k\hat{\boldsymbol{\beta}}}$. In the same framework, but focusing on estimation of $\tau^*_{2,k}$ in (6), Elamir and Skinner (2006) assume independent Gamma priors in place of Lognormals on λ_k's, and find that the addition of parametric random effects does not improve risk estimates; as a consequence, they suggest to adopt plug-in estimates. Conjugate Gamma priors guarantee computational advantages, as do the Inverse Gaussian distributions (IG) described in Carlson (2002).

Our proposal is as follows: we keep the mixed effects log-linear structure (7), but remove the assumption of normality. 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)],

$$\phi_k|G \sim \text{i.i.d. } G, \qquad G \sim \mathcal{D}(m, G_0). \tag{8}$$

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. Within the class of models just defined, we consider three specifications of G_0 that lead to three different direct generalizations of the existing literature, namely, Skinner and Holmes (1998), when $G_0 = N(\alpha, \sigma^2)$; Carlson (2002), when $G_0 = \text{IG}(\alpha, \sigma^2)$; Elamir and Skinner (2006), when $G_0 = \text{LG}(a, b)$, where LG denotes the distribution of a log transformation of a Gamma(a, b) variate ω, with $f(\omega; a, b) = b^a/\Gamma(a)\omega^{a-1}e^{-b\omega}$. The hyperparameters in the base measure G_0 can be fixed, which is how we proceed, or be given a prior distribution. While in the corresponding parametric approaches a fixed distribution $G = G_0$ is selected and its hyperparameters are estimated, we take the opposite perspective, that is, we assume a random G while holding the hyperparameters of its mean distribution G_0 fixed, and chosen so as to obtain a vague specification.

The estimation of risk measures under the proposed model is discussed in Section 3. Here we analyze the implications of our nonparametric specification of random effects and the advantages over the parametric counterparts of our model. 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 $\mathcal{D}(m, G_0)$. Blackwell and MacQueen (1973) provide this as the product of successive conditional distributions:

$$\phi_i|\phi_1, \ldots, \phi_{i-1}, \qquad M \sim \frac{m}{m+i-1}G_0(\phi_i) + \frac{1}{m+i-1}\sum_{k=1}^{i-1}\delta(\phi_k = \phi_i), \tag{9}$$

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. It also shows that m, the mass or precision parameter of the DP, affects the expected number of clusters.

Therefore, under the previous assumptions, the likelihood function turns out to be a sum of terms where all possible partitions (clusterings) C of the K cells into c nonempty clusters are considered [see, e.g., Liu (1996), Lo (1984)],

$$\sum_{c=1}^{K} \sum_{C:|C|=c} \frac{\Gamma(m)}{\Gamma(m+K)} m^c \prod_{j=1}^{c} \Gamma(n_j) \int p(\mathbf{f}_{(j)}|\boldsymbol{\beta}, \phi_j)\, dG_0(\phi_j), \tag{10}$$

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

$$\frac{\Gamma(m)}{\Gamma(m+K)} m^c \prod_{j=1}^{c} \Gamma(n_j) = \Pr\{n_1, \ldots, n_c|C, c\}, \tag{11}$$

and finally

$$p(\mathbf{f}_{(j)}|\boldsymbol{\beta}, \phi_j) = \prod_{k \in \text{cluster}\, j} \frac{1}{f_k!} e^{\pi f_k(\mathbf{w}_k'\boldsymbol{\beta}+\phi_j)} e^{-e^{\pi(\mathbf{w}_k'\boldsymbol{\beta}+\phi_j)}}. \tag{12}$$

In the likelihood, starting from the latter formula, we notice that the same random effect is assigned to all cells belonging to the same cluster, that is, 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 (1997), Chapter 41], and that the number of clusters in each partition ranges from 1 to K. We stress that the total number of terms in the likelihood (10) is the Bell number, B_K, which is a combinatorial quantity assuming large values even for moderate K; just to fix ideas, when $K = 10$, $B_K = 115{,}975$. The parametric counterparts of our nonparametric random effects models correspond 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.

The above considerations show that the intrinsic characteristics of DP random effects set them apart from parametric random effects for their potential to improve upon the fixed effects component of the log-linear model. Indeed, 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 parametric random effects, $N(\alpha, \sigma^2)$, $\text{IG}(\alpha, \sigma^2)$ or $\text{LG}(a, b)$, allows for departures from the Poisson log-linear model such as overdispersion, but does not significantly affect

the way one can learn about a given cell from other cells. In contrast, the inclusion of DP random effects implies that, in addition to the above-mentioned fixed effects patterns, the model encompasses all other nonempty subsets of the K cells. For each given partition, a possible relation of dependence among cells in the same subset is explicitly evaluated. In other words, to learn about a given cell, additional information is borrowed from cells belonging to the same subset, for each subset to which the cell can be assigned in the context of all possible partitions in nonempty subsets of the K cells. This suggests both the potential for the proposed model to improve the risk estimates and the associated computational complexity. Furthermore, the results under our nonparametric models can be interpreted as averages over mixed effects log-linear models with different clusterizations of parametric random effects.

3. Inference. In this section we describe how to estimate not only τ_1^* and τ_2^* in (3) and (4), as most of the literature based on log-linear models does, but also τ_1 and τ_2 and their terms τ_{1k} and τ_{2k} in (1) and (2), in a fully Bayesian way. This approach is inspired by Manrique-Vallier and Reiter (2012, 2014); see also Fienberg and Makov (1998). In order to keep the notation uncluttered, let $\boldsymbol{\theta}$ denote the set of all model parameters conditioning $\lambda_1, \ldots, \lambda_K$ for each of the models analyzed in this article. The posterior distribution over $\boldsymbol{\theta}$ is not available in closed form for any of the models considered here. We employ Markov Chain Monte Carlo (MCMC) techniques [Neal (1993)] to obtain samples from $p(\boldsymbol{\theta}|f_1, \ldots, f_K)$; 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 and $m|$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 log-likelihood $\log[p(f_1, \ldots, f_K|\boldsymbol{\beta}, \text{rest})]$ 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 a 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 $p(f_1, \ldots, f_K|\boldsymbol{\beta}, \text{rest})$. 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, $p(\boldsymbol{\beta}'|\boldsymbol{\beta}) = N(\boldsymbol{\beta}'|\boldsymbol{\mu}, \varepsilon^2 M^{-1})$, with $\boldsymbol{\mu} = \boldsymbol{\beta} + \frac{\varepsilon^2}{2} M^{-1} \nabla_{\boldsymbol{\beta}} \log[p(f_1, \ldots, f_K|\boldsymbol{\beta}, \text{rest})]$. 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 $\boldsymbol{\phi}$. An extensive treatment of MCMC for DP models can be found in Neal (2000), where we refer the reader for full details. Drawing samples from the posterior distribution over the random effects entails allocating cells to an unknown number of clusters and drawing a value for the random effect for each cluster. The way in which these steps are carried out depends on whether it is possible to exploit conjugacy of the base measure, that is, whether the integral $\int p(f_k|\boldsymbol{\beta}, \phi)\, dG_0(\phi)$ can be evaluated analytically.[4]

In the applications presented in Section 4, we choose a LG distribution for G_0 so that $\omega = e^{\phi}$ is given a Gamma base measure. In this case we can exploit conjugacy with the Poisson likelihood; a similar argument applies when ϕ is given the IG distribution, for which the integral above is analytically tractable. When conjugacy holds, a simple and efficient algorithm can be constructed to draw samples from the full conditional distribution over the random effects, which is referred to as Algorithm 3 in Neal (2000). First, the allocation of cells to clusters is updated for one cell at a time, integrating out analytically the dependency from the actual value that the random effects can take, and allowing the total number of clusters to vary across iterations. Second, the value of the random effect pertaining to each cluster can be drawn directly from a known distribution [which is a Gamma in the extension of Elamir and Skinner (2006)], again due to the fact that the likelihood and the DP base measure form a conjugate pair. The sampling of $\boldsymbol{\phi}$ has a computational cost that scales linearly with the number of cells.

Instead, when we extend the model proposed by Skinner and Holmes (1998), the normal distribution does not enjoy the above-mentioned conjugacy property; for this reason, sampling schemes for nonconjugate base measures described, for example, in Neal (2000), must be employed, and these usually lead to less efficient MCMC sampling schemes.

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 draw samples from its posterior distribution and to account for uncertainty about it when inferring risk measures. By selecting a Gamma prior over m, it is possible to employ the approach of Escobar and West (1995) to draw samples from the posterior distribution over $m|\text{rest}$ directly.

MCMC estimates. Once H samples from the posterior distribution over $\boldsymbol{\theta}$ are available, it is possible to obtain Monte Carlo estimates of per-cell risks by refer-

[4]Note that here $p(f_k|\boldsymbol{\beta}, \phi)$ represents the likelihood based on a single datum, that is, one of the terms in the product (12).

ring to (6):

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

$$\hat{\tau}^*_{2,k} = \frac{1}{H}\sum_{h=1}^{H} E\left(\frac{1}{F_k}\Big| f_k = 1, \boldsymbol{\theta}^{(h)}\right),$$

which in turn lead to global risk estimates $\hat{\tau}^*_i = \sum_{k=1}^{K} \hat{\tau}^*_{i,k}$, $i = 1, 2$.

Fully Bayesian estimates of $\tau_i, i = 1, 2$, instead require taking into account a further source of variability induced by the randomness of the unobserved $F_1, \ldots, F_K$. In particular, observing that the terms $\tau_{i,k}$ in τ_i, are $\tau_{i,k} = \tau_{i,k}(f_k, \lambda_k, F_k)$ $(i = 1, 2)$ where $F_1, \ldots, F_K$ are unknown random quantities, with $F_k|\lambda_k \sim \text{Poisson}(\lambda_k)$, $k = 1, \ldots, K$, we consider values of λ_k's drawn from their joint posterior distribution and then values of $F_1, \ldots, F_K$ drawn from the corresponding Poisson distributions. This allows us to derive a sample of $\tau_{i,k}$, $i = 1, 2$, from which it is possible to characterize the posterior distribution of global and cell-specific risk values by standard Monte Carlo techniques. Accounting for randomness of both groups of unobserved parameters (λ_k's and F_k's) has two important implications. First, since a posteriori the λ_k's 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 uncertainty on the F_k's is also explicitly considered, we obtain risk estimates whose variability depends on the variability of the F_k's as well as the variability of the λ_k's and the association between λ_k's. This means, for instance, that our posterior variance of τ_1, $\text{Var}(\tau_1|f_1, \ldots, f_K) = \text{Var}(\sum_k^K I(f_k = 1) I(F_k = 1|f_k = 1)|f_1, \ldots, f_K)$, cannot be expressed in the form [Rinott and Shlomo (2007b)]

$$\sum_{k}^{K} I(f_k = 1) \Pr\{F_k = 1 | f_k = 1\}\big(1 - \Pr\{F_k = 1 | f_k = 1\}\big) \tag{13}$$

because of the covariances of the λ_k's. Moreover, our variances, and the corresponding standard deviations (s.d.), provided in Table 1, are derived from the posterior distributions of τ_i, $i = 1, 2$, rather than by plug-in.

As mentioned in Section 1, the issue of nonexistence of MLE (due to data being not fully informative about model parameters) is also important in Bayesian analysis of log-linear models. The vague prior we specify in Section 4 for the fixed effects replaces the information content lacking in the data with the information contained in the prior. This prior is especially useful to estimate the all two-way interactions model that we consider for comparison, as it makes the posterior information matrix of $\boldsymbol{\beta}$ not rank deficient. This is the way we can avoid ad hoc additions of small positive quantities to cells containing sampling zeroes.

TABLE 1
Estimated values of τ_1 and τ_2 by means of $\hat{\tau}_1$ and $\hat{\tau}_2$ (top panel) and $\hat{\tau}_1^*$ and $\hat{\tau}_2^*$ (bottom panel) for the California and WHIP tables. Posterior standard deviations in parentheses

	California		WHIP	
Model	$\tau_1 = 211$	$\tau_2 = 499.8$	$\tau_1 = 915$	$\tau_2 = 1948.1$
	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}_1$	$\hat{\tau}_2$
(P+O)	0.0 (0.1)	170.5 (1.4)	1180.7 (33.2)	3322.2 (24.8)
(P+I)	255.4 (10.4)	518.8 (7.6)	1184.9 (23.7)	2289.9 (17.1)
(P+II)	253.9 (11.1)	537.5 (8.6)	958.4 (22.4)	1996.2 (17.5)
(NP+O)	700.0 (232.1)	910.0 (198.8)	2397.8 (459.5)	3042.6 (405.1)
(NP+I)	217.0 (12.2)	503.7 (10.9)	1010.4 (29.8)	2083.4 (28.3)
(NP+I) Emp	241.8 (12.3)	528.8 (10.8)	970.2 (32.7)	2046.0 (32.4)
	$\hat{\tau}_1^*$	$\hat{\tau}_2^*$	$\hat{\tau}_1^*$	$\hat{\tau}_2^*$
(P+O)	0.0 (0.0)	170.6 (0.7)	1180.6 (10.8)	3322.1 (10.8)
(P+I)	255.3 (3.3)	518.8 (4.0)	1184.9 (9.4)	2290.0 (9.7)
(P+II)	254.0 (4.7)	537.6 (5.5)	958.5 (10.7)	1996.3 (11.9)
(NP+O)	700.1 (231.8)	910.1 (198.7)	2397.8 (458.8)	3042.6 (404.9)
(NP+I)	217.0 (7.5)	503.7 (8.7)	1010.3 (21.9)	2083.4 (24.9)
(NP+I) Emp	241.7 (7.4)	528.7 (8.5)	970.2 (26.0)	2046.0 (29.6)
II	250.4 (–)	536.7 (–)	946.8 (–)	1992.4 (–)

4. Applications. To evaluate the performance of the proposed approach in practical settings, we apply our nonparametric risk estimators to two tables with different sizes and degrees of sparsity. We consider data from the 5% Public Use Microdata Sample of the U.S. 2000 Census for the state of California [IPUMS, Ruggles et al. (2010)], treating the set of individuals aged 21 and older as the population. We also use data from the 7% microdata sample of the Italian National Social Security Administration (WHIP-Work Histories Italian Panel, Laboratorio Revelli, Centre for Employment Studies, http://www.laboratoriorevelli.it/whip), treated here as the population. In both cases we draw random samples with fraction $\pi = 0.05$. The key variables considered for the WHIP data are sex (2), age (12), area of origin (11), region of work (20), economic sector (4), wages guarantee fund (2), working position (4) and firm size (5), leading to a table of 844,800 cells, of which 5017 (0.59%) are nonempty. The California table comprises the following key variables: number of children (10), age (10), sex (2), marital status (6), race (5), employment status (3) and education (5), for a total of 90,000 cells, of which 4707 (5.2%) are nonempty. These variables are a subset of those specified in Manrique-Vallier and Reiter (2012) that we follow for categorization of the key variables and selection of the reference population; the latter excludes the presence of impossible, or otherwise predetermined, combinations, that is, structural zeroes. The expected cell probabilities (λ_k) in cells containing structural zeroes are assigned a degenerate prior; loosely speaking, this has to be interpreted as a

"conventional" way to state that all such cells have to be ignored in the fitting of the model so that they cannot bias estimates in the remaining "nonstructural zero" cells.

In the applications we focus on one of the nonparametric models presented in Section 2, namely, the extension of the model proposed by Elamir and Skinner (2006). We examine several choices of the log-linear component describing the fixed effects; in particular, we investigate a model with no fixed effects, referred to as the overall mean model (O), the main effects or independence model (I) and the all two-way interactions model (II). For comparison we fit both the parametric (P) and nonparametric (NP) random effects versions of the above-mentioned models. For simplicity, hereafter, the above models will be identified by labels denoting the selected modeling options, so, for instance, (NP+I) is the nonparametric model with main effects only, and (P+II) and (II) are the all two-ways interactions models with and without parametric random effects, respectively.

Under the parametric specification P, the random effects ϕ are modeled by a LG(a, b) prior, whereas under the nonparametric specification NP, the random effects are assumed to follow a distribution drawn from a DP whose base measure is LG(a, b). In both cases, the hyperparameters (a, b) are fixed so that this is a vague prior: $a = 1$, $b = 0.1$ (b is the rate parameter). Since we drop from $\boldsymbol{\beta}$ the overall effect β_0 to overcome identifiability issues, β_0 is incorporated into the mean of the random effects. Therefore, the assumption of Elamir and Skinner (2006), who take the mean of the Gamma distribution of the multiplicative random effects ω to be 1, is compatible with ours: by fixing $a \neq b$, that is, a prior mean that differs from 1, we simply allow for an overall effect. For the components of $\boldsymbol{\beta}$ we assume independent and reasonably vague Gaussian priors $N(0, 10)$. Finally, we take a Gamma$(1, 0.1)$ prior on m. All models are estimated by the fully Bayesian method[5] described in Section 3, with the exception of one nonparametric independence model where the prior on the fixed effects is taken to be degenerate at the MLE of $\boldsymbol{\xi}$, $\hat{\boldsymbol{\xi}}_{\mathrm{ML}}$. We label the corresponding approach by (NP+I) Emp to indicate that we rely on empirical Bayes estimation in the presence of DP random effects. Note that the California table is free of structural zeroes, so that the log-linear model with main effects only is in fact an independence, that is, decomposable, model, and $\hat{\boldsymbol{\xi}}_{\mathrm{ML}}$ exists since all observed unidimensional margins are positive. This is not the case for the large WHIP table where the main effects model represents a quasi-independence model. Here we simply use the $\hat{\boldsymbol{\xi}}$ obtained by IPF (for which the R routine converged within 15 iterations with a tolerance of 10^{-8}), assuming it is the extended MLE.

In the implementation of the MCMC sampling, convergence of the chains was checked using Gelman and Rubin's potential scale reduction factor [$\hat{R}$; Gelman and Rubin (1992)] by running 10 parallel chains and assessing that chains had

[5]Suitably modified when estimating the models (P+I) and (P+II).

converged when $\hat{R} < 1.1$ for all the parameters. According to this criterion, all chains converged within five thousands iterations that were then discarded before running the chains for a further 10,000 iterations that were used to evaluate the risk scores.

We note here that, for the (P+II) model, the $K \times q$ design matrix associated with the log-linear model component is very large ($q > 10^3$ and $K \sim 10^6$), which caused some difficulties when running the adopted sampling scheme. Indeed, each update of $\boldsymbol{\beta}$ requires evaluating and factorizing a $q \times q$ matrix, leading to running times that are beyond usability (weeks). This is the main reason why we considered a subset of the variables in the California table analyzed in Manrique-Vallier and Reiter (2012). For the parametric models that are introduced for comparison we therefore tested an alternative where we approximated the posterior distribution over $\boldsymbol{\beta}$ by a Gaussian. In particular, we carried out a Laplace Approximation, where the approximating Gaussian has a mean equal to the mode of the posterior distribution and the inverse covariance is equal to the negative Hessian of the logarithm of the posterior density at the mode [Tierney and Kadane (1986)]. Computationally, this procedure has the following advantages. First, the mode-finding procedure can be implemented in a way that it does not require factorization or storage of large matrices, for example, by feeding log-posterior and its gradient to standard optimization routines. Second, once the mode is located, drawing samples from the approximate posterior over $\boldsymbol{\beta}$ requires that the $q \times q$ covariance matrix is computed and factorized only once. Interestingly, in cases where we could run the sampling from the posterior over $\boldsymbol{\beta}$, we noticed that the risks obtained by the approximate method were strikingly close to one another. For this reason, the results that we report for the (P+I) and the (P+II) models refer to the approximate method.

Table 1 reports true and estimated values of τ_1 and τ_2 (s.d. in parentheses) for six models formed by combining different modeling options as described above. In addition, risks obtained under the default log-linear model (II) without random effects and fitted by the IPF are included for reference. First of all, the very small difference in the results under the (II) and (P+II) models confirms the findings in Elamir and Skinner (2006). Moreover, similar to what Manrique-Vallier and Reiter [(2012), page 1389] have observed under their GoM models, point estimates $\hat{\tau}_1^*$ and $\hat{\tau}_2^*$ are nearly identical to $\hat{\tau}_1$ and $\hat{\tau}_2$ with smaller posterior standard deviations, since the former do not take into account the variability of F_k's. The 2.5th, 5th, 50th, 95th and 97.5th percentiles of the posterior distribution of τ_i, $i = 1, 2$, under a subset of the models reported in Table 1, are presented in Figure 1 where models appear in order of complexity of the log-linear specification and the solid vertical lines represent the true risk values.

Inspection of Table 1, and related Figure 1, confirms that the parametric all two-way interactions model (P+II) outperforms the (P+I) model in the large table, which is in line with what was reported in the literature. If, however, we include nonparametric models in the analysis, new and interesting findings are as follows:

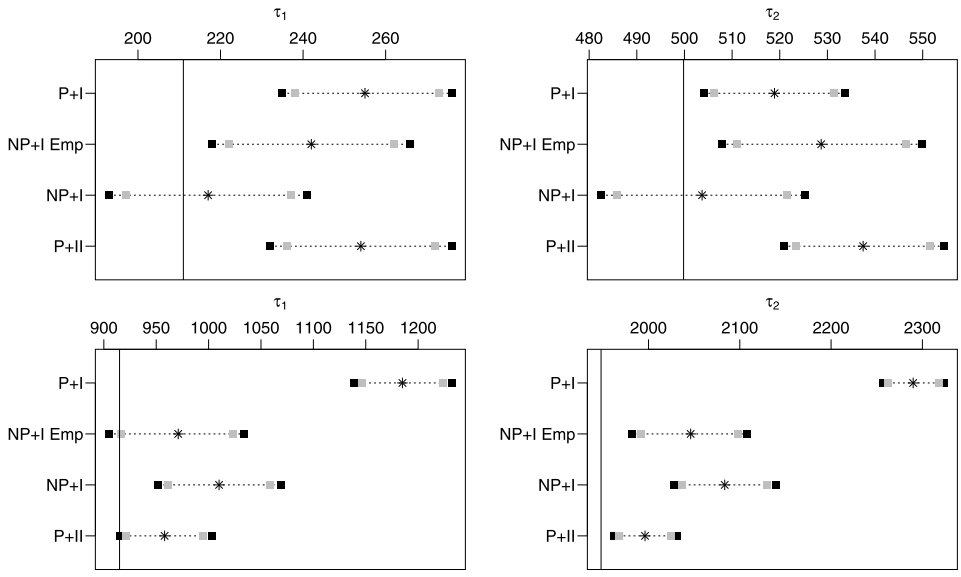

FIG. 1. *Quantiles of the posterior distributions of* τ_1 (*first column*) *and* τ_2 (*second column*) *under a subset of parametric and nonparametric models considered. First* (*second*) *row refers to the California* (*WHIP*) *table. Gray squares*: 5*th*, 95*th percentiles*; *black squares*: 2.5*th*, 97.5*th percentiles*; *stars*: *median of the posterior distributions. Vertical segments represent the true risks.*

1. The potential of the DP prior for capturing association not modeled by the fixed effects can be noticed by comparing the results under the two models that, conditionally on the random effects, rely on the exchangeability assumption, namely, the parametric no fixed effects log-linear model (P+O) and its nonparametric counterpart (NP+O). The latter is the model used in Dorazio et al. (2008).

2. When risks are estimated by nonparametric models, the tendency of risk estimates to decrease as the complexity of the model increases, shown in Skinner and Shlomo [(2008), Table 1, going, in particular, from I to II], can be observed in both California and WHIP tables at a lower level, that is, going from the (NP+O) model to the (NP+I) and (NP+I) Emp models.

3. The performance of the nonparametric independence model, (NP+I) Emp, is roughly 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 the need for additional terms (interactions) in the vector of fixed effects. Considering, moreover, the good performance of the (NP+I) model in the California table, we are induced to conclude that, in the presence of DP random effects, the number of fixed effects required to obtain reasonable global risk estimates is lower than in the parametric case and less sensitive to the size of the table K. This is in line with finding 2.

4. Although we do not specifically address the challenging problem of model choice, our approach may contribute to lessen its scale and complexity. Indeed, the

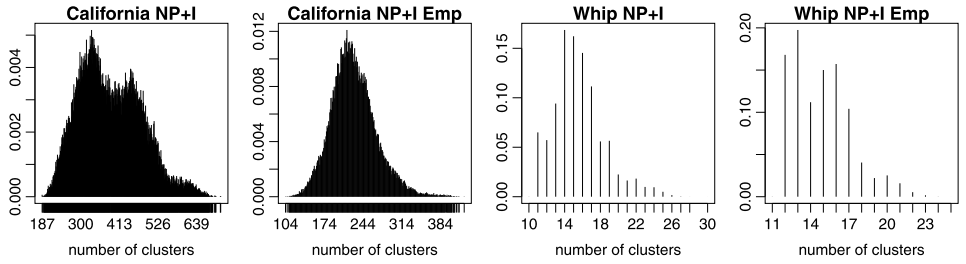

FIG. 2. *Posterior distribution of the number of clusters for the California and WHIP tables when using the (NP+I) and (NP+I) Emp models.*

(NP+I) Emp model can be taken as the initial model in a forward model selection procedure. The significant reduction of the space of adjacent models that need to be examined at each step would mitigate the difficulties associated with model choice. This point will be explored in future work.

By comparing parametric and nonparametric independence models, (P+I), (NP+I) and (NP+I) Emp, Figure 1 allows us to see how strongly DP random effects integrate into a log-linear model with main effects only and contribute to improving global risk estimates even for the large WHIP table for which the fit of the parametric independence model is particularly poor.

To appreciate the role played by the clustering mechanism induced by the DP, Figure 2 provides a representation of the posterior distribution of the number of clusters under the proposed (NP+I) and (NP+I) Emp models. There is a striking difference between the distribution of the number of clusters for the California and WHIP tables. The fact that in the California table the number of clusters is large seems to reflect the ability of the (NP+I) model to perform extremely well in the estimation of risk. In the case of the WHIP table, the introduction of the DP distributed random effects, although significantly improving on the estimation of risk with respect to the (P+I) model, does not completely account for the lack of fit.

For the California table we also explored the frequentist properties of our approach through a simulation study comprising 100 samples, where we evaluated the frequentist coverage of the credible intervals based on the 2.5th and 97.5th percentiles of the posterior distribution of $\tau_i, i = 1, 2$. We observed that, under the (NP+I) model, all of them include the true value of τ_1 and 76 include the true value of τ_2.

In the rest of this section we explore the behavior of per-cell risk estimates by using, for simplicity, $\hat{\tau}^*_{1,k}$ and $\hat{\tau}^*_{2,k}$. In Figure 3, for a subset of the models presented in Table 1, and proceeding as in Figure 4 of 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 $\hat{\tau}^*_{1,k}$. Visual assessment of the relative proximity to the diagonal gives an idea of how

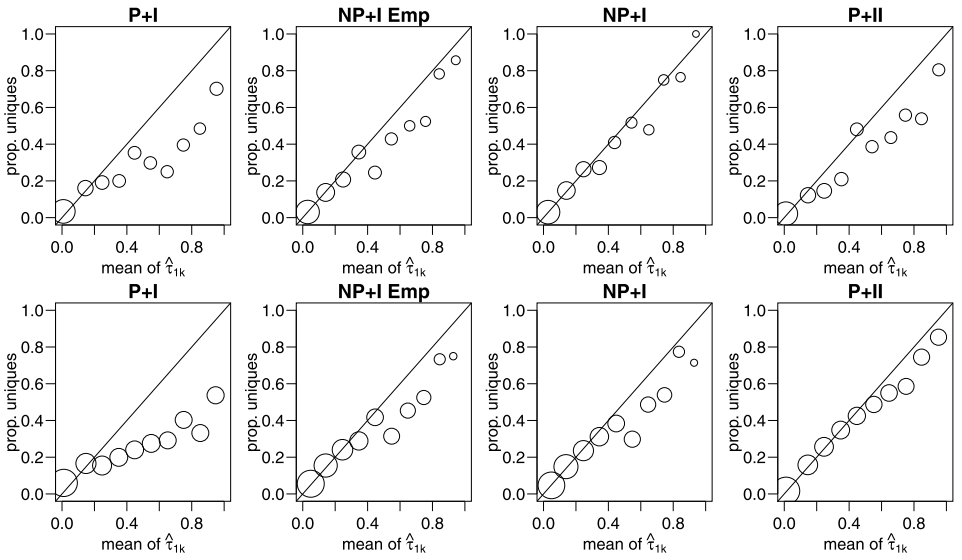

FIG. 3. *Proportion of population uniques plotted against the average estimated risk* $\hat{\tau}^*_{1,k}$, *for cells categorized into* 10 *equal-width intervals according to the values of* $\hat{\tau}^*_{1,k}$. *The size of the plotting points depends on the number of cells in each interval. First line*: *California table*; *second line*: *WHIP table.*

accurately each model can predict population unique cells. Similarly, in Figure 4, 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 $\hat{\tau}^*_{2,k}$.

In Figure 5 we compare per-cell risk estimates $\hat{\tau}^*_{i,k}$ and true risks (bold lines) for $i = 1, 2$, respectively. We consider estimates from the California table for which the (NP+I) model outperforms the parametric model (P+II) and the parametric independence model (P+I). Cells containing sample uniques are arranged in increasing order of the true per-cell risk; in turn, for each level of the true per-cell risk, estimates are arranged in decreasing order of population cell size and increasing order of estimated risk. This allows us to observe overestimates and underestimates in all cells under the two models under examination. By drawing cutoff points (not included) in the first two plots of the figure, we can also visualize the corresponding false positive and false negative cells. We can conclude that the (NP+I) model improves risk estimates $\hat{\tau}_{1,k}$ in cells with intermediate population frequencies, while in cells with extreme (very large or 1) population frequencies, the (P+II) model tends to produce better results at the cell level; however, this is not sufficient for the (P+II) model to outperform the (NP+I) model in the estimation of the global risk. This fact is even more apparent when inspecting the last two plots in Figure 5. The results just analyzed indicate that, compared to the all two-way parametric random effects log-linear model, the proposed approach does

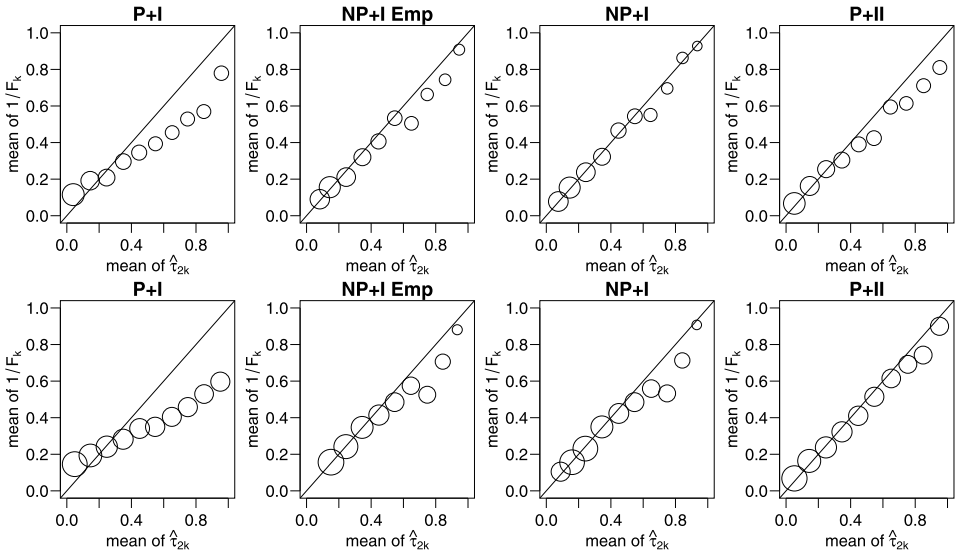

FIG. 4. *Mean of* $1/F_k$ *against the mean of the estimated risk* $\hat{\tau}^*_{2,k}$, *for cells categorized into* 10 *equal-width intervals according to the values of* $\hat{\tau}^*_{2,k}$. *The size of the plotting points depends on the number of cells in each interval. First line*: *California table*; *second line*: *WHIP table.*

not produce uniformly better per-cell risk estimates. While in this paper we have mainly focused on measures of global risk, the specific problem of per-cell risk estimation could be tackled in a different way, that we plan to explore in future work.

5. Computational aspects and comparison with other approaches. In this section we discuss computational costs and applicability to large tables of our proposal, in comparison with other approaches in the recent literature related to our problem.

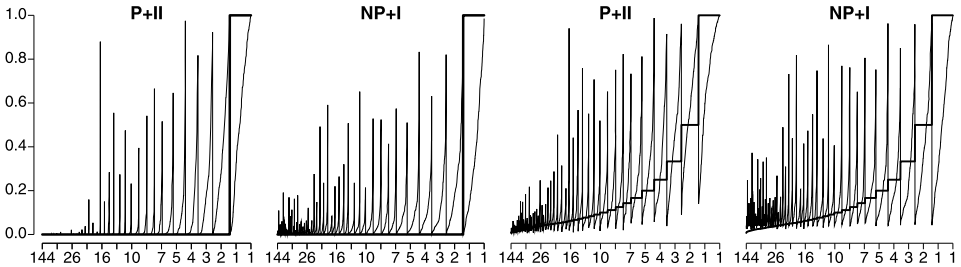

FIG. 5. *Comparison of risk estimates* $\hat{\tau}^*_{1,k}$ (*first two plots*) *and* $\hat{\tau}^*_{2,k}$ (*last two plots*) *for sample unique cells under the* (*P*+*II*) *and* (*NP*+*I*) *models for the California table. Bold lines represent the true risks.*

As already discussed, in large and sparse tables maximum likelihood estimation of standard log-linear models—in particular the all two way interaction model (II) in our application—and model search become highly challenging, as the parameter space quickly explodes and a number of parameters may result in being unidentifiable due to sparsity. Assigning a prior to these parameters and carrying out Maximum-A-Posteriori (MAP) estimation instead of Maximum Likelihood, allowed us to somewhat work around this problem. However, locating the mode of the posterior distribution of the parameters requires employing iterative search algorithms that are computationally intensive and potentially slow to converge.

Vice versa, the computational performance of our nonparametric independence models depends on the interplay of two different elements, namely, (i) estimation of the parametric fixed effects $\boldsymbol{\beta}$; and (ii) estimation of the nonparametric random effects. As to (i), it is the number of main effects that determines the computational scale of the problem, that, albeit cubic in the size of $\boldsymbol{\beta}$, remains much smaller than the table size. Nonetheless, when the size of $\boldsymbol{\beta}$ is very large, storing of information matrices might be challenging; in that case we suggest use of the Empirical Bayesian version of the nonparametric independence model. This approach, akin to the estimation strategy of Skinner and Holmes (1998), is an appealing alternative, since it relies on IPF that converges in at most two steps in decomposable models.

(ii) is related to the allocation of random effects; the proposed nonparametric methods scale linearly with the number of cells, which makes our proposal suitable for applications to large tables. Although it is not possible to provide any guarantees on convergence speed of the MCMC approach to the posterior distribution over the parameters, in all tables that we studied in this work, we found that convergence of the chains was reached after a few thousand iterations.

By using a log-linear representation of the latent class model (LCM), our (NP+I) model and the LCM recently applied by Manrique-Vallier and Reiter (2014) can be shown to rely on the same basic assumptions, that is, independence of the key variables conditionally on an unobserved variable S and a prior for such unobserved variable S somewhat related to the Dirichlet process [see also Si and Reiter (2013)]. The latter assumption, however, is applied at the level of individuals (through an individual latent class Z whose prior is a finite stick breaking process) in the LCM, while it is applied at the level of cells (via the cell-specific DP random effect ϕ) in the (NP+I) model. This implies different allocations to clusters and different sampling schemes in the two cases. Practical consequences are that as the sample size increases, the (NP+I) model does not require any additional computational costs, while it scales as discussed above with the number of cells. Vice versa, the LCM scales easily with the number of cells, as emphasized in Manrique-Vallier and Reiter (2014), but has to sustain a nonnegligible computational cost as the sample size increases. This may be an advantage of our method, as in the practice of Statistical Institutes the sample sizes are often much larger than those considered in the literature based on LCMs. Note that while in

our applications the sampling fraction is higher than what could commonly be used in practice, the absolute size ($n = 57{,}547$; $n = 40{,}122$ for the California and WHIP data, resp.) is the same order of magnitude of many surveys on individuals conducted, for instance, by the Italian National Statistical Institute. A second practical issue relates to structural zeroes, which, at the level of cells, are very simply managed in our nonparametric approach (by a degenerate prior on those cells), while they require a specific technique in the LCM, that is, the one introduced by Manrique-Vallier and Reiter (2014). This also means that our approach has the same advantages mentioned by Manrique-Vallier and Reiter (2014), such as applicability to variables with skip patterns or when certain combinations have been effectively eliminated from the sample by design.

In conclusion, (NP+I) and LC models are built on the same basic ingredients though implemented in different ways, thereby producing the different advantages and disadvantages—in terms of scalability, structural zeroes and applicability—just summarized.

The actual computational times associated with our proposal clearly depend on the size of the table to be analyzed. Recent developments on Bayesian LCMs show that they can deal with extremely large tables, in the order of 10^{40} as illustrated by Si and Reiter (2013) for a multiple imputation problem. Being able to treat extremely large tables in short computational times is undoubtedly important. Although "Big Data" issues are likely to have an impact in the context of disclosure risk estimation (in terms of disclosure scenario and type and number of key variables), we deem that tables of the above size may be less common than in other related fields. Indeed, when the number of cells is much higher than the population size, the average population cell size N/K, whatever the sample, is very low. Under such circumstances Statistical Institutes may judge releasing information on the key variables at that level of detail too risky and may prefer to recode/merge the key variables and/or decrease their detail before proceeding to assess the risk formally through a suitable statistical model.

6. Final comments. In this article we investigated the role of random effects in log-linear models for disclosure risk estimation. We show in theory and through real data applications that modeling the random effects nonparametrically does improve upon the log-linear model, because it allows to simplify to a large extent the structure of fixed effects required to achieve good risk estimates. Therefore, the utility of our nonparametric approach increases with the size and the degree of sparsity of the table, since problems with nonestimable parameters in fixed effects log-linear models increase disproportionately with the number of terms included. Quoting Fienberg and Rinaldo (2007), "the number of possible patterns of zero counts invalidating the MLE exhibits an exploding behavior as the number of classifying variables or categories grows."

Unlike parametric random effects models, for each cell our nonparametric models combine learning from two types of neighborhoods, one driven by the fixed ef-

fects, and the other driven by the data and implied by the clustering of the random effects.

Interestingly, in the applications the empirical Bayesian version of our (NP+I) model emerges as the nonparametric equivalent of the parametric model (P+II), indicated in the literature as the default approach in risk estimation. This evidence is found in tables with rather different structures and dimensions. Moreover, in the analysis of the California data set the (NP+I) model greatly improves the performance of the parametric model in terms of global risk estimation.

The striking impact of the inclusion of DP random effects in the (P+I) model indicates that enlarging the simple (NP+I) Emp model by adding a few interaction terms can be expected to produce satisfactory results. Even if we do not address the issue of model selection, the previous remark opens the door to a model search approach that takes our (NP+I) Emp model as the starting point, thus lessening the scale and complexity of the problem, since the space of adjacent models to be examined is significantly reduced.

We emphasize that the previous ones are general results, that is, a reduction in the number of fixed effects in the presence of DP random effects—with the mentioned benefits in terms of estimability in sparse tables and simplification of model search—can be expected in different applications of log-linear models, not only in disclosure risk estimation.

Having adopted a fully Bayesian approach allowed us to account for all sources of uncertainty (about λ_k's, F_k's) in the estimation of risk. In (P+I) and (P+II) models, this is an advantage compared to the empirical Bayes procedure by Elamir and Skinner [(2006), Section 3.3], even though we can expect numerical agreement between their estimates because of the vague priors adopted for the fixed effects. As to our (NP+I) Emp model, which replicates in a nonparametric context the estimation strategy of Skinner and Holmes (1998), although it neglects the variability of the fixed effects, it incorporates other sources of uncertainty, such as the population frequencies. Although our approach generalizes existing models mentioned in Section 2, there are important differences from the previous literature, including Rinott and Shlomo (2007b), as our risk estimates are endowed with unconditional (posterior) variances and we can also produce credible intervals, that is, posterior probability intervals.

As regards the assumptions underlying our Bayesian models, all of them are explicit and more flexible than the ones underlying a log-linear model without random effects. Indeed, we have selected vague priors and modeled the random effects nonparametrically, which is a further relaxation of the hypotheses.

While in this paper we have mainly focused on measures of global risk, the applications indicate that, compared to the all two-way parametric random effects log-linear model, the proposed approach does not produce uniformly better per-cell risk estimates even when the global risk estimates under the (NP+I) model outperform those obtained under the (P+II) model. The specific problem of per-cell risk estimation could be tackled in a different way, that we plan to explore in future work.

Acknowledgments. The authors wish to thank the Associate Editor and the anonymous reviewers for their valuable and constructive comments.

REFERENCES

BLACKWELL, D. and MACQUEEN, J. B. (1973). Ferguson distributions via Pólya urn schemes. *Ann. Statist.* **1** 353–355. MR0362614

BLEI, D. M., NG, A. Y. and JORDAN, M. I. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3** 993–1022.

CARLSON, M. (2002). Assessing microdata disclosure risk using the Poisson-inverse Gaussian distribution. *Statistics in Transition* **5** 901–925.

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, 670–671. MR2432438

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.

EROSHEVA, E. A., FIENBERG, S. E. and JOUTARD, C. (2007). Describing disability through individual-level mixture models for multivariate binary data. *Ann. Appl. Stat.* **1** 502–537. MR2415745

ESCOBAR, M. D. and WEST, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90** 577–588. MR1340510

FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230. MR0350949

FIENBERG, S. E. and MAKOV, U. E. (1998). Confidentiality, uniqueness, and disclosure limitation for categorical data. *Journal of Official Statistics* **14** 385–397.

FIENBERG, S. E. and RINALDO, A. (2007). Three centuries of categorical data analysis: log-linear models and maximum likelihood estimation. *J. Statist. Plann. Inference* **137** 3430–3445. MR2363267

FIENBERG, S. E. and RINALDO, A. (2012). Maximum likelihood estimation in log-linear models. *Ann. Statist.* **40** 996–1023. MR2985941

FILIPPONE, M., MIRA, A. and GIROLAMI, M. (2011). Discussion of: "Sampling schemes for generalized linear Dirichlet process random effects models", by M. Kyung, J. Gill, and G. Casella [MR2859768]. *Stat. Methods Appl.* **20** 295–297. MR2859770

FORSTER, J. J. and WEBB, E. L. (2007). Bayesian disclosure risk assessment: Predicting small frequencies in contingency tables. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **56** 551–570. MR2405419

GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–472.

GIROLAMI, M. and CALDERHEAD, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 123–214. MR2814492

HABERMAN, S. J. (1974). *The Analysis of Frequency Data*. Univ. Chicago Press, Chicago, IL. MR0408098

JOHNSON, N. L., KOTZ, S. and BALAKRISHNAN, N. (1997). *Discrete Multivariate Distributions*. Wiley, New York.

LIU, J. S. (1996). Nonparametric hierarchical Bayes via sequential imputations. *Ann. Statist.* **24** 911–930. MR1401830

LO, A. Y. (1984). On a class of Bayesian nonparametric estimates. I. Density estimates. *Ann. Statist.* **12** 351–357. MR0733519

MANRIQUE-VALLIER, D. and REITER, J. P. (2012). Estimating identification disclosure risk using mixed membership models. *J. Amer. Statist. Assoc.* **107** 1385–1394. MR3036402

MANRIQUE-VALLIER, D. and REITER, J. P. (2014). Bayesian estimation of discrete multivariate latent structure models with structural zeros. *J. Comput. Graph. Statist.* **23** 1061–1079. MR3270711

NEAL, R. M. (1993). Probabilistic inference using Markov chain Monte Carlo methods. Technical Report No.CRG-TR-93-1, Dept. of Computer Science, Univ. Toronto.

NEAL, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* **9** 249–265. MR1823804

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.

RINOTT, Y. and SHLOMO, N. (2007a). A smoothing model for sample disclosure risk estimation. In *Complex Datasets and Inverse Problems* (R. Liu, W. Strawderman and C.-H. Zhang, eds.). *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **54** 161–171. IMS, Beachwood, OH. MR2459186

RINOTT, Y. and SHLOMO, N. (2007b). Variances and confidence intervals for sample disclosure risk measures. In *Bulletin of the International Statistical Institute*: *Proceedings of the* 56*th Session of the International Statistical Institute*, *ISI'*07, *Lisbon*. *August* 22–29 1090–1096.

ROBERTS, G. O. and ROSENTHAL, J. S. (2009). Examples of adaptive MCMC. *J. Comput. Graph. Statist.* **18** 349–367. MR2749836

RUGGLES, S., ALEXANDER, J. T., GENADEK, K., GOEKEN, R., SCHROEDER, M. B. and SOBEK, M. (2010). Integrated public use microdata series: Version 5.0 [Machine-readable database]. University of Minnesota, Minneapolis. Available at https://usa.ipums.org/usa/.

SI, Y. and REITER, J. P. (2013). Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *Journal of Educational and Behavioral Statistics* **38** 499–521.

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. *J. Amer. Statist. Assoc.* **103** 989–1001. MR2462887

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.

TEH, Y. W., JORDAN, M. I., BEAL, M. J. and BLEI, D. M. (2006). Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.* **101** 1566–1581. MR2279480

TIERNEY, L. and KADANE, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.* **81** 82–86. MR0830567

C. CAROTA
R. LEOMBRUNI
DIPARTIMENTO DI ECONOMIA E STATISTICA
UNIVERSITÀ DI TORINO
LUNGO DORA SIENA 100A
10153 TORINO
ITALY
E-MAIL: cinzia.carota@unito.it
roberto.leombruni@unito.it

M. FILIPPONE
SCHOOL OF COMPUTING SCIENCE
UNIVERSITY OF GLASGOW
18 LILYBANK GARDENS
G12 8QQ, GLASGOW
SCOTLAND
E-MAIL: maurizio.filippone@glasgow.ac.uk

S. POLETTINI
DIPARTIMENTO DI METODI E MODELLI PER
L'ECONOMIA, IL TERRITORIO E LA FINANZA
SAPIENZA UNIVERSITA' DI ROMA
VIA DEL CASTRO LAURENZIANO 9
00161 ROMA
ITALY
E-MAIL: silvia.polettini@uniroma1.it