Assessing Bayesian Nonparametric Log-Linear Models: an application to Disclosure Risk estimation

Cinzia Carota¹, Maurizio Filippone² and Silvia Polettini³

¹Dipartimento di Economia e Statistica
Università di Torino
e-mail: cinzia.carota@unito.it

²Department of Data Science
EURECOM, France
e-mail: maurizio.filippone@eurecom.fr

³Dipartimento di Scienze Sociali ed Economiche
Università di Roma “La Sapienza”
e-mail: silvia.polettini@uniroma1.it

Abstract: We present a method for identification of models with good predictive performances in the family of Bayesian log-linear mixed models with Dirichlet process random effects. Such a problem arises in many different applications; here we consider it in the context of disclosure risk estimation, an increasingly relevant issue raised by the increasing demand for data collected under a pledge of confidentiality. Two different criteria are proposed and jointly used via a two-stage selection procedure, in a M-open view. The first stage is devoted to identifying a path of search; then, at the second, a small number of nonparametric models is evaluated through an application-specific score based Bayesian information criterion. We test our method on a variety of contingency tables based on microdata samples from the US Census Bureau and the Italian National Security Administration, treated here as populations, and carefully discuss its features. This leads us to a journey around different forms and sources of bias along which we show that (i) while based on the so called “score+search” paradigm, our method is by construction well protected from the selection-induced bias, and (ii) models with good performances are invariably characterized by an extraordinarily simple structure of fixed effects. The complexity of model selection - a very challenging and difficult task in a strictly parametric context with large and sparse tables - is therefore significantly defused by our approach. An attractive collateral result of our analysis are fruitful new ideas about modeling in small area estimation problems, where interest is in total counts over cells with a small number of observations.

Keywords and phrases: Bayesian model selection, Disclosure risk, Dirichlet process random effects, Log-linear mixed models, Model’s predictive performance, Selection-induced bias, Small area estimation.

1. Introduction

Log-linear modeling provides a convenient way of investigating relationships between categorical variables in contingency tables. However, when the set
classifying variables is large or there are many categories, the induced table not only is large, but often also sparse, with a huge set of alternative log-linear specifications. This poses challenging theoretical issues in both model fitting and selection (see e.g. Fienberg and Rinaldo, 2012; Piironen and Vehtari, 2017, respectively, and references therein). In a recent paper (Carota et al., 2015), we proposed a class of Bayesian log-linear models with nonparametric random effects useful to overcome the above mentioned problems in disclosure risk estimation, an increasingly relevant issue jointly raised by the increasing demand for data collected under a pledge of confidentiality and refinement of record linkage techniques. We also suggested that likely under this approach model selection may be pursued within a narrower search space. Indeed, here we argue that under this class of models there is room for a novel approach to model selection: we propose a method aimed at limiting the bias arising in the phase of model selection, rather than in model fitting as prescribed by the method proposed in Skinner and Shlomo (2008). Our main contribution, presented in Section 4, is a two-stage model selection procedure tailored to estimating global measures of disclosure risk, in a M-open view, i.e. avoiding the unrealistic assumption that the true data generating model is included in the model space. The first stage is devoted to identify a small number of nonparametric models that are then evaluated through a new measure of predictive performance, essentially an application-specific score based Bayesian information criterion. While based on the so called “score+search” paradigm, the proposed selection method is by construction well protected from the selection-induced bias, a serious and ubiquitous problem, often neglected in the literature, but recently emphasized especially in machine learning journals (see, e.g., Reunanen, 2003; Varma and Simon, 2006; Cawley and Talbot, 2007, 2010). See also Piironen and Vehtari (2017). We discuss this and other forms of bias and the distinctive features of our selection method both in the context of the recent literature on predictive methods for model assessment, selection and comparison (Vehtari and Ojanen, 2012; Gelman et al., 2014; Underhill and Smith, 2016; Piironen and Vehtari, 2017) and in relation to the specialized literature on model selection for disclosure risk estimation (Forster and Webb, 2007; Skinner and Shlomo, 2008; Manrique-Vallier and Reiter, 2012, 2014). An attractive collateral result of our analysis are ideas fruitfully applicable in different fields. Focusing on the relationship between good global risk estimates and good per-cell risk estimates, we show that overfitting nonparametric log-linear models for the purpose of global risk estimation can be advantageously exploited in specific applications where the interest is in total counts over cells with small sample frequencies, as, for instance, small area estimation. This is discussed in Section 5.

The remainder of the paper is structured as follows. Section 2 introduces the problem of data confidentiality and reviews current approaches to model selection for disclosure risk estimation. Section 3 provides materials and motivations for the model selection procedure presented in Section 4 and suggests the application sketched in Section 5. There, based on a variety of explorations of different models over different real data tables, first, we show that nonparametric random effects are a powerful adaptive remedy against the positive bias
(that is, overestimation of risks) found by Skinner and Shlomo (2008) under insufficiently rich, under-fitting log-linear parametrizations; successively, we illustrate in terms of per-cell risk estimates the nature of such corrections for overestimation of global risks, thereby finding further interesting applications of the class of models under consideration.

2. Data confidentiality and model selection for disclosure risk estimation

In meeting the request to always increase the information content and the detail of the statistical outputs, Statistical Offices must comply with the legal obligation to protect confidentiality of respondents, targeting at maximizing the analytical content of the released data without disclosing confidential information attached to specific individuals or entities. We consider the release of social survey microdata; in this case certain publicly available categorical characteristics of the sampled records (e.g. age, gender, education, geography, household type, ...) can be used as key variables to match sampled records with units in the population and disclosure risk can be defined on cells of multi-way contingency tables of such variables. Following the previous literature (see, e.g., Bethlehem et al., 1990; Chen and Keller-McNulty, 1998; Skinner and Elliot, 2002; Skinner and Holmes, 1998), we define the risk of re-identification in terms of cells containing one or a small number of individuals, that is “unique” and “small” cells, of the sample and population contingency tables, respectively. The increasing availability of information from external sources and the request to maximize the detail of the released variables often imply the presence of a large number of key variables, some with many categories, with the consequence that the number of cells in the associated contingency table is much larger than the sample size and risk assessment in actual data releases requires to handle extremely large and sparse tables. Data are often protected by reducing the detail (global recoding) or the number (global suppression) of key variables, which lowers the disclosure risk, but also deteriorates the analytical validity of the data. Identifying the proper protection amounts to finding the proper balance between disclosure risk and data utility. This requires accurate and repeated estimates of disclosure risk prior to any proposed data release, which, in turn, demand for ready and safe identification of good models for risk estimation.

In the literature, risk measures are estimated by introducing suitable models on the contingency tables defined by the key variables. Let \( F_k \) and \( f_k \) be the frequencies in the population and sample contingency tables defined by the key variables, and let \( K \) be the total number of cells. Two widely accepted measures of the global risk of re-identification, or disclosure risks, are the number of sample uniques which are also population uniques,

\[
\tau_1 = \sum_{k=1}^{K} I(f_k = 1, F_k = 1) = \sum_{k: f_k = 1} I(F_k = 1 | f_k = 1) := \sum_{k: f_k = 1} \tau_{1k}, \quad (1)
\]
and 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),

\[
\tau_2 = \sum_{k=1}^{K} I(f_k = 1) \frac{1}{F_k} = \sum_{k:k_f = 1} \frac{1}{F_k} := \sum_{k:k_f = 1} \tau_{2k}.
\]

(2)

Often (1) and (2) are approximated by \( \tau_1^* = \sum_{k=1}^{K} I(f_k = 1) Pr\{F_k = 1|f_k = 1\} := \sum_{k:k_f = 1} \tau_{1k}^* \) and \( \tau_2^* = \sum_{k=1}^{K} I(f_k = 1) E(1/F_k|f_k = 1) := \sum_{k:k_f = 1} \tau_{2k}^* \), i.e., \( E(\tau_i|f_1, ..., f_K) \), \( i = 1, 2 \), under the assumption of cell independence.

Often (1) and (2) are approximated by \( \tau_1^* = \sum_{k=1}^{K} I(f_k = 1) Pr\{F_k = 1|f_k = 1\} := \sum_{k:k_f = 1} \tau_{1k}^* \) and \( \tau_2^* = \sum_{k=1}^{K} I(f_k = 1) E(1/F_k|f_k = 1) := \sum_{k:k_f = 1} \tau_{2k}^* \), i.e., \( E(\tau_i|f_1, ..., f_K) \), \( i = 1, 2 \), under the assumption of cell independence.

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, thereby overcoming the assumption of exchangeability of cells (see e.g. Bethlehem et al., 1990) implying the unrealistic consequence that risk estimates are constant across cells having the same sample frequencies, but different combinations of key variables. Per-cell estimates may be used to highlight high risk combinations or aggregated to produce an overall, global, risk measure. However, as mentioned in Section 1, in disclosure risk estimation log-linear models have almost invariably to deal with extremely sparse tables which pose a number of challenges described, e.g., in Fienberg and Rinaldo (2007) and Fienberg and Rinaldo (2012). Manrique-Vallier and Reiter (2012) and Manrique-Vallier and Reiter (2014) adopt Bayesian latent structure models that does not suffer from the potential shortcomings of log-linear models. Carota et al. (2015) suggest instead to overcome them by adopting log-linear models with a standard estimable structure for the parametric fixed effects, specifically the independence structure, whose lack of fit is compensated for by nonparametric random effects described by a Dirichlet Process (DP).

Model specification for risk estimation is a somehow neglected problem. Here we briefly review the current approaches to model selection, specifically devoted to estimate global measures of disclosure risk. Skinner and Shlomo (2008), recognizing the peculiarity of the inferential problem under consideration, suggest a model search algorithm based on a predictive criterion that we describe next. Instead, Forster and Webb (2007), restrict their attention to the special sub-family of decomposable graphical log-linear models (thereby avoiding problems in model fitting), and account for model uncertainty by averaging inferences over that very specific sub-family. Manrique-Vallier and Reiter (2012) and Manrique-Vallier and Reiter (2014) reconsider the problem, though under a different model specification, as will be discussed in the sequel.

Skinner and Shlomo (2008) model population and sample frequencies by independent Poisson distributions with rates \( \lambda_k \) and \( \pi \lambda_k \), respectively, and describe the parameters \( \lambda = (\lambda_1, \ldots, \lambda_k, \ldots, \lambda_K) \) by a log-linear model,

\[
\lambda_k = e^{\mu_k}, \quad \mu_k = W_k^T \beta,
\]

(3)

where \( \pi \) is the sampling fraction supposed to be known, \( W_k \) is a \( q \times 1 \) design
vector depending on the values of the key variables in cell $k$ and $\beta$ is a $q \times 1$ vector of fixed effects.

Clearly, over-parametrized log-linear models tend to overfit the data for the trivial reason that they tend to the saturated model. Specifically, an exceedingly complex model induces both bias in estimators of the Poisson parameters $\lambda$ and an inflation of their variances. Skinner and Shlomo (2008) rely on maximum likelihood (ML) estimates of the log-linear model parameters $\beta$ (and therefore of $\lambda$) that they plug into the expressions for the global disclosure risks $\hat{\tau}^*_i$, $i = 1, 2$, under the Poisson model. Interestingly, Skinner and Shlomo (2008) notice that the behaviour of such risk estimates evolves regularly with the complexity of the assumed log-linear specification. In particular, the bias evolves monotonically from underestimation to overestimation of the global disclosure risks $\tau_i$, $i = 1, 2$, when going from the independence model (I), to the all two-way interactions model (II), to the all three-way interactions model (III), and so on. Among such models, they select the least underfitting as the starting model, and propose a stepwise forward model search aimed at minimizing the (positive) bias of the corresponding risk estimator. The optimal model is the one that achieves the “best” compromise between over- and underestimation according to a minimum error criterion, denoted by $\hat{B}$. The authors anticipate that most likely a number of “reasonable models” may exist, between which the criterion is not able to discriminate, and also suggest to use the differences between the estimates for each of these models as a diagnostic to check the sensitivity of the measures to the model specification (see Skinner and Shlomo, 2008, p.994).

The latter, which in general is very high (see also Manrique-Vallier and Reiter, 2012; Fienberg and Makov, 1998), is found to be small across the reasonably good models, implying a form of robustness. Manrique-Vallier and Reiter (2012) stress that, when dealing with large tables, the models to be compared under the above approach amount to several hundreds, and exploration of the whole space becomes unfeasible. They avoid such drawback by abandoning the log-linear formulation, relying instead on a Bayesian version of the Grade of Membership (GoM) model. This is a latent class model characterized by a pre-specified, small, number of classes (extreme profiles), with individuals being allowed to belong to more than one class simultaneously (mixed membership model). Model complexity is driven by the number of extreme profiles $E$ ($K$ in their notation), and it turns out that risk estimates are extremely sensitive to the value of $E$. However, as $E$ increases, the estimates exhibit a typical monotonically decreasing pattern which, past a given threshold for the number of latent classes, tends to become stable. This leads Manrique-Vallier and Reiter (2012) to propose the following empirical model selection strategy: starting from a given, small, value for $E$, progressively increase it by a multiple of 3 or 5, depending on the sample size, until there is evidence of stabilization. The computational cost of such model selection procedure is avoided in Manrique-Vallier and Reiter (2014) by speci-

\footnote{We argued that $\hat{B}$ may be viewed as an estimator of the bias of $\hat{\tau}^*_i$, $i = 1, 2$, in the presence of underfitting, when the bias may be expected to be positive. The properties of $\hat{B}$ in the case of overfitting are more difficult to assess.; Skinner and Shlomo (2008), p.993}
fying a large number of classes (50) for the truncated latent class model they adopt to allow for structural zeroes. Similarly, Si and Reiter (2013), following Dunson and Xing (2009), use a mixture of independent multinomial distributions, with the mixture distribution being modelled by a DP prior. The number of classes is not fixed a priori, yet for computational convenience the authors resort to a truncated representation of the Dirichlet process, thereby fixing the number of classes to a large value, analogously to Manrique-Vallier and Reiter (2014).

In this article we focus on the class of Bayesian log-linear mixed models with DP random effects proposed in Carota et al. (2015) and seek an “optimal” model within this class. Under the same assumptions and notation as in (3), we model the parameters \( \lambda_k \) through a log-linear model with mixed effects:

\[
\lambda_k = e^{\mu_k}, \quad \mu_k = w_k^\prime \beta + \phi_k, \quad \phi_k|G \overset{\text{i.i.d.}}{\sim} G, \quad G \sim \mathcal{D}(m, G_0),
\]

where, all other symbols being as in (3), \( \phi_k \) is a random effect accounting for cell specific deviations, whose distribution function, denoted by \( G \), is assumed to be unknown and a priori distributed according to a DP, \( \mathcal{D} \), with base probability measure \( G_0 \) and total mass parameter \( m \) (Ferguson, 1973). \( G_0 \) is the mean of the Dirichlet process, while \( m \) controls the variance of the process. Suitable parametric priors are also assigned to \( m \) and to the fixed effects \( \beta \).

With a slight abuse of terminology, in order to stress the nonparametric nature of random effects, as in Carota et al. (2015) we refer to this approach as Bayesian nonparametric (NP) log-linear modelling, though recognizing that it is in fact a semi-parametric approach under which nonparametric, DP distributed, random effects are added to the log-linear formulation (3) for the Poisson means. In describing our models, we will focus on their parametric and nonparametric components separately; for instance the nonparametric independence model, that we take as a “default” model, is denoted by NP+I, to emphasize the probabilistic nature of random effects (NP) and the structure of its parametric component (I), i.e. of fixed effects, respectively. The parametric counterparts of models (4), that is, models where parametric random effects are added to \( w_k^\prime \beta \) in (3), are analysed in Skinner and Holmes (1998) and Elamir and Skinner (2006), who observe a practical equivalence, in terms of risk estimation, with log-linear models without random effects. In the above case we will speak of parametric log-linear models (or simply parametric models, labelled P) and parametric risk estimates to emphasize the nature of the random effects.

In the next section we reconsider the association found by Skinner and Shlomo (2008) between underfitting and over-estimation, while the relation between overfitting and under-estimation will be discussed in Section 5. According to the previous authors, such positive/negative bias is due to structural zeroes and sampling zeroes, respectively (for details see Skinner and Shlomo, 2008, pp.991-992). Although our view about the source of over-estimation is different - structural zeroes are removed from our analysis since the beginning (see Carota et al., 2015, p.535)-, a careful analysis of both such associations conducted from within the enlarged family of log-linear models (4) provides fruitful
ideas about two very challenging issues: model selection, and small area estimation, respectively.

3. Underfitting and over-estimation: a Bayesian nonparametric correction

In this Section, we illustrate how the performance of simple models of type (3) is improved by the addition of Dirichlet process random effects. A series of preliminary explorations reveals that such nonparametric log-linear models attain extraordinarily appropriate corrections of the positive bias of risk estimates corresponding to the models of type (3) from which they are built. Moreover, good nonparametric models are invariably found in very close neighbourhoods of the nonparametric independence one, NP+I. Here we present a selection of these explorations. They serve not only to highlight the ability of DP random effects to correct for the overestimation of global risks typical of oversimplified/underfitting log-linear specifications indicated by Skinner and Shlomo (2008), but also to illustrate the nature of such corrections in terms of per-cell risk estimates. Moreover, these explorations introduce and motivate the model selection procedure proposed in the next Section. There we will also provide a theoretical justification for the model performances observed in this Section. Models are tested on a range of contingency tables differing in size, reference population and spanning variables obtained from different sources as detailed in the supplementary material A.1. First, we consider three tables of decreasing dimension (K=3,600,000; 900,000 and 360,000 cells, respectively) built from the 5% public use microdata sample of the U.S. 2000 census for the state of California (IPUMS Ruggles et al., 2017). To allow for structural zeroes, we also consider a table of 844,800 cells, half of which structurally empty, built from the set of N = 794,986 individuals recorded in the 7% public use microdata sample of the Italian National Social Security Administration, 2004 (source: Work Histories Italian Panel, WHIP). Such microdata samples are treated here as populations: we take simple random samples from those populations and benchmark estimated risks under different models against their “true” values. For each of the four tables above, in this Section we explore a set of nonparametric models built as follows. We first specify a set of models of type (3) selected by maximizing the log-likelihood under a severe penalty for complexity. Of course, these models are not optimal for disclosure risk estimation, but, conditionally on the severe constraint of simplicity imposed through the penalty term, they are optimal for estimation of the cell parameters $\lambda$. To each of such models we associate a nonparametric model of type (4) by adding DP random effects. A formal description of the criterion used at this preliminary stage is provided in Section 4, where it is denoted by $C_0(\gamma)$. For each of our test tables, Table 1 (first three rows) lists the nonparametric models selected for exploration. Models are reported in order of incremental complexity w.r.t. our default model, NP+I; complexity is described through: (a) the number of two-way interactions (interaction terms) included in the model, and (b) the number of associated
interaction parameters (further details can be found in the supplementary material A.1). As indicated by the entries of column (a) in Table 1, the models selected through $C_0(\gamma)$ are close neighborhoods of the nonparametric independence one. To illustrate the effect of a richer fixed effects specification, for the Whip table we also include in the exploration two models encompassing four two-way interaction terms, corresponding to a much lower penalty for the log-likelihood. All parameters of the models considered in this section are assigned vague priors, as detailed in the supplementary material A.1. For computational details regarding the applications presented here and throughout the article see instead the supplementary material B.

Table 1

List of nonparametric models explored in California and Whip tables, and their complexity. Complexity is evaluated focusing on the parametric component and measured through: (a) the number of two-way interactions included in the model, and (b) the number of associated interaction parameters. Since our reference model is the nonparametric independence one (NP+I), we restrict attention to the terms that are added to the main effects of the spanning variables in the fixed effects specification.

| Label | California | Whip |
|-------|------------|------|
|       | Large: K=3600000. | Medium: K=900000. | Small: K=36000 |
|       | (a) | (b) | (a) | (b) | (a) | (b) |
| NP_a  | 1 | 1 | 1 | 3 | 1 | 9 |
| NP_b  | 1 | 18 | 1 | 18 | 1 | 20 |
| NP_c  | 2 | 19 | 2 | 21 | 1 | 12 |
| NP_d  | 4 | 27 | 4 | 48 | 2 | 21 |

Figures 1 and 2 present posterior medians and posterior credible intervals (95% and 99%) for the global risks $\tau_1$ and $\tau_2$ under all nonparametric models presented in Table 1, for California and WHIP tables, respectively. Figure 1 includes the parametric counterparts of the above models to assess their relative performance in terms of risk estimates. Clearly, in the presence of DP random effects, good risk estimates stem from a few, simple fixed effects, which are otherwise insufficient to produce adequate inferences in the presence of parametric random effects, or, equivalently (Elamir and Skinner, 2006), in the absence of random effects under a formulation of type (3). Figure 2 confirms that simple nonparametric models are able to produce good risk estimates and suggests the result of selecting a richer fixed effects specification. Actually, in all of our preliminary explorations we observed that the mere inclusion of one specific interaction term may mark the difference between good models and inadequate models. Indeed, the good performance of the four models $NP_c$, $NP_d$, $NP_e$ and $NP_a$ is due to the presence of a specific interaction term, as we will see in Section 4. Figure 2 also presents the risk estimates obtained under the parametric independence model (P+I), as its performance will be further analysed in the sequel (see Figure 7 in the supplementary material A.2).

Analysing per-cell risks, we next clarify how the over-estimation systematically associated with the “too simple” parametric models reported in Figure 1 is corrected for by the DP random effects at the cell level. When comparing
Fig 1. California data: true values (horizontal solid line) and quantiles (0.005, 0.025, 0.50, 0.975, 0.995) of the posterior distributions of $\tau_1$ (first column) and $\tau_2$ (second column) under the set of nonparametric models (NP) reported in Table 1, and under their parametric counterparts (P). First row: Large table; second row: Medium table; third row: Small table. For the latter two tables nonparametric and parametric independence models (NP+I and P+I, respectively) are also reported.
parametric vs. nonparametric estimates of per cell risks $\tau_{1,k}$ and $\tau_{2,k}$, in all of our preliminary tests as well as under the models analysed in this paper we always noticed a typical sigmoidal shape shown e.g. in the center and right-hand columns of Figure 3, where estimates of $\tau_{1,k}$ are reported for the three models explored in the Large California table. (See also Fig. 7 in the supplementary material A.2, where a similar analysis is conducted focusing on the independence models considered in the remaining three tables). This shape reveals that the presence of DP random effects invariably increases per-cell risk estimates that are small under the parametric model, but, even more, decreases per-cell risk estimates that are large under the parametric model, which is the feature that invariably results in a significantly improved balance, i.e. reduced bias, at global level.

The regularity and persistence of such adjustment is impressive if one considers the diversity of the four contingency tables we are dealing with. Indeed risks are estimated over tables of different sizes, with and without structural zeroes, with very different proportions of population uniques, doubles and so on. Notice that under the approach of Skinner and Shlomo (2008) the complexity of the
optimal model increases remarkably with $K$, as shown in Tables 5-7 on p. 999 of their article where we can also observe an evolution of the starting model from the independence (I) to the all two way interactions (II) model. Manrique-Vallier and Reiter (2012, online supplement) select the best log-linear model according to the criterion of Skinner and Shlomo (2008) for the California Large table. Their results over samples of 5000 and 10000 individuals confirm that the complexity of the starting as well as the selected model strongly depends on the table’s size. In contrast, under our approach, the increase of complexity needed to obtain very good fitting nonparametric models is extraordinarily limited. Moreover,
the NP+I model invariably emerges as the starting model (see Figures 1 and 2, and compare with Carota et al., 2015, where such model is presented as a sort of default model). All these findings not only induce us to conclude that the DP random effects are a formidable adaptive correction for the over-estimation found by Skinner and Shlomo (2008) under “too simple”, underfitting, log-linear models of type (3), but also invite us to take advantage of this by proposing a new method for model selection.

So far, we discussed a notion of positive bias due to underfitting and arising in the model estimation phase; in Section 5 we will elaborate on the negative bias due to overfitting. In the next section, instead, we present our model selection procedure, which is derived with special attention to the different sources of bias arising in the model selection phase briefly reviewed in the initial paragraphs. We stress the difference with the criterion \( \hat{B} \) of Skinner and Shlomo (2008), that, as recalled in Section 2, balances positive and negative bias arising in model fitting.

4. New model selection method

Vehtari and Ojanen (2012) give a comprehensive survey of established and recent Bayesian predictive methods for model assessment, selection and comparison. Here we selectively review only those references useful to illustrate and comparatively discuss our proposal. The literature (see also Gelman et al., 2014; Underhill and Smith, 2016; Piironen and Vehtari, 2017) clearly indicates that the challenge in estimating predictive model accuracy is twofold:

(i) to correct for the bias inherent in evaluating a model’s predictions of the data that were used to fit it (within-sample-error), and

(ii) to address in some way the selection-induced bias, i.e. the undesirable optimistic bias in predictive performance evaluation that, in large sets of models, often leads to select a model by chance rather than by merit.

As to (i), several proposals of bias correction are available in the literature (discussed, for instance, in Vehtari and Ojanen, 2012; Gelman et al., 2014; Underhill and Smith, 2016, and related articles), but they often incur in the second type of bias. As to (ii), it has long been known that any criterion suffers from selection bias (see e.g. Linhart and Zucchini, 1986; Miller, 1990; Chatfield, 1995; Zucchini, 2000; Vehtari and Ojanen, 2012; Gelman et al., 2014; Piironen and Vehtari, 2017, and references therein), and that it stems from a form of overfitting in model selection, analogous to the more familiar one occurring in training the model. In the last decade, however, the severity of this problem has been re-affirmed and quantified for a series of established and recent criteria, especially in the machine learning literature (e.g. Reunanen, 2003; Varma and Simon, 2006; Cawley and Talbot, 2007, 2010) where reliable model performance evaluation not only is crucial in many practical applications, but is required for fair comparison of machine learning algorithms. A criterion with non-negligible variance has the potential for overfitting in the optimization phase.
by exploiting meaningless peculiarities of the sample over which it is evaluated. Cross validation, widely used criteria adjusting for within-sample-error such as AIC, DIC and WAIC which are approximations to different versions of cross validation, and many other criteria are all severely prone to selection bias (see, e.g., Zucchini, 2000; Cawley and Talbot, 2010; Piironen and Vehtari, 2017). About (i) and (ii) and the underlying decomposition of the estimation error into bias and variance, Piironen and Vehtari (2017), p. 718, wrote: “[...] the unbiasedness is intrinsically unimportant for a model selection criterion” and “it is more important to be able to rank competing models in an approximately correct order with a low variability.” They also comment that, nonetheless, most literature focuses on unbiased estimates of the model predictive accuracy and provides little guidance on how to reduce the selection induced bias. In our application, however, their solution, namely the projection approach in a so-called M-completed view (for details see Piironen and Vehtari, 2017, section 2.4), cannot be easily implemented, so we are left with traditional remedies against the selection bias: restricting selection to a small number of well-considered models (this includes regularization and/or early stopping), or, alternatively, model averaging.

In what follows we propose a new model selection procedure and argue that all of these remedies are in some way applied under our approach, under the assumption that the true model is not necessarily included in the explored model space and that a good model is just a useful approximation to the true model. As argued in Underhill and Smith (2016, p.1006), once one accepts this assumption, the focus immediately shifts to identifying which aspects of the model performance are most important to the end user, thereby relegating all the others to an intermediate, instrumental role.

Our new two-stage model selection procedure is based on the so called “score + search” paradigm. The first stage is devoted to identify a path of search, i.e. which nonparametric models and in which order have to be evaluated; at the second stage, an optimal model is selected through a new measure of model predictive accuracy specifically tailored to disclosure risk estimation. The detail of the procedure is as follows. Building on findings in Section 2, we start from the nonparametric independence model, NP+I. At the first stage we focus exclusively on its parametric component, i.e. we restrict the search to fixed effects models of type (3), and do a preliminary stepwise search. Starting from a large value of a penalty factor $\gamma$, we gradually move it down and select, at each step, the interaction term which maximizes a penalized log-likelihood, referred to as the criterion $C_0(\gamma)$,

$$C_0(\gamma) = \sum_{k=1}^{K} \log(p(f_k|\hat{\beta}_{ML})) - d \times \gamma,$$

where $p(f_k|\hat{\beta}_{ML}) = \frac{w_k^{f_k}e^{f_k\hat{\beta}_{ML}}}{\int f_k w_k e^{f_k\hat{\beta}_{ML}}}$, $d$ is the difference between the number of parameters estimated under the current model and under the independence model I, and $\gamma$ controls the strength of the penalty. In principle this
search is restricted to decomposable models (to have a guarantee of existence of the ML estimates $\hat{\beta}_{ML}$, reliability of the degrees of freedom, and so on), but we will see that such a restriction is ineffective in practice, since we can stop the search after a few steps.\footnote{At this stage we use $\hat{\beta}_{ML}$ as an approximation to Bayesian estimates of the fixed effects $\beta$ since they are assigned a vague prior (details in the supplementary materials A.1).}

At the second stage of the procedure, a small set of candidate nonparametric models - obtained by adding DP random effects to the parametric component identified at the first stage - is evaluated through the application-specific criterion $C_1$,

$$C_1 = \sum_{k=1}^{K} I(f_k = 1) \times \log \left( \int p(f_k | \lambda_k) p(\lambda_k | f_1, \ldots, f_k) d\lambda_k \right), \quad (6)$$

where $p(f_k | \lambda_k) = \frac{1}{\pi_k!} (\pi \lambda_k)^{f_k} e^{-\pi \lambda_k}$ and $\lambda_k$ is defined as in (4). This is the log pointwise predictive density (lppd, see Gelman et al., 2014) restricted to the unique cells, namely those crucial for estimating the global risks (1) and (2). $C_1$ measures the model predictive accuracy, or performance, and is computed using posterior simulations $\lambda^{(h)}$, $h = 1, \ldots, H$:

$$\sum_{k=1}^{K} I(f_k = 1) \times \log \left( \frac{1}{H} \sum_{h=1}^{H} p(f_k | \lambda^{(h)}_k) \right)$$

(see the supplementary material B for implementation details). Following Underhill and Smith (2016), and their description of different approaches to utility based model selection, $C_1$ can also be presented as a score based Bayesian information criterion whose particular scoring rule avoids that the good performance of a model on the subset of cells of interest is disguised by poor performance on the other cells, as it might be the case for criteria concerned with the model’s performance across the full joint distribution. This is particularly appropriate in the case of disclosure risk estimation, as sample uniques usually are a very small subset of the total number of cells.

Of course, models with high values of $C_1$ are preferred. Being high predictive accuracy equivalent to low predictive error, in this respect $C_1$ and the criterion $\hat{B}$ by Skinner and Shlomo (2008) are not different. Values of $C_1$ for all models explored in Figures 1 and 2 (see also Table 6 in the supplementary material A.1) are presented in Table 2 (computational details are provided in the supplementary material B). In parentheses we show different models’ rankings: based on $C_1$; on the true estimation errors, that is, the distance between the true value of $\tau_i$ and its Bayesian point estimate under the model; and based on the widely applicable information criterion (Watanabe, 2009) restricted to the sample uniques, say WAIC_U. Although parametric models are not candidate models, but just parametric counterparts of the candidate nonparametric models selected at the first stage of the procedure, they are included in the rankings to show that...
in some contingency tables (Large California and WHIP) the WAICU selects a largely sub-optimal parametric model, despite the small number of models under evaluation. This is the empirical evidence of substantial selection bias and optimism in the performance evaluation due to the increase of the variance of the criterion implied by the presence of a further estimated term. Instead, C1 provides a reasonably good ranking of models in all contingency tables, and when (see, for instance, second and third positions in the Large California table and second and first positions in the Small California table) the rankings based on the true estimation error do not agree with those based on C1, yet the relative positions defined by C1 do not differ substantially from the “true” ones. Importantly, moreover, the corresponding values of C1 are so close to each other that we are actually warned about possible inversions of positions in the

---

### Table 2

Risk estimates under the models explored in Figures 1 and 2, predictive measures and models’ ranks (in brackets) based on the true estimation error, on C1 and WAICU. For each real data table we report the number of sample uniques (U) and the true values of τ1 and τ2.

| Model         | C1         | τ1       | τ2       | WAICU     |
|---------------|------------|----------|----------|-----------|
| California Large (U=11421) |            |          |          |           |
| NP_a          | 2245.2     | 2205.3   | 3849.7   |           |
| NP_b          | 2323.5     | 2090.5   | 3950.6   |           |
| NP_c          | 2272.2     | 2034.3   | 3909.4   |           |
| P_a           | 2706.6     | 4431.9   | 2974.5   |           |
| P_b           | 2727.1     | 4458.4   | 2959.4   |           |
| P_c           | 2652.9     | 4374.4   | 2929.7   |           |
| California Medium (U=7609) |            |          |          |           |
| NP_a          | 1185.4     | 2340.6   | 4981.5   |           |
| NP_b          | 1223.9     | 2371.8   | 4821.3   |           |
| NP_c          | 1220.9     | 2374.4   | 4823.9   |           |
| P_a           | 1189.2     | 2345.3   | 4823.9   |           |
| P_b           | 1424.8     | 2523.6   | 4823.9   |           |
| P_c           | 1386.2     | 2487.3   | 4823.9   |           |
| P+I           | 1399.4     | 2500.9   | 4823.9   |           |
| California Small (U=3575) |            |          |          |           |
| NP_a          | 479.8      | 1003.2   | 7886.3   |           |
| NP_b          | 483.8      | 1011.3   | 7993.0   |           |
| NP_c          | 480.3      | 1008.6   | 7982.6   |           |
| P_a           | 581.6      | 1072.9   | 8951.8   |           |
| P_b           | 568.5      | 1065.0   | 9023.4   |           |
| P+I           | 579.9      | 1077.6   | 9109.1   |           |
| WHIP (U=7176) |            |          |          |           |
| NP_a          | 917.9      | 1981.2   | 1610.7   |           |
| NP_b          | 1003.1     | 2078.4   | 1641.3   |           |
| NP_c          | 921.2      | 1987.0   | 1597.7   |           |
| P_a           | 908.9      | 1972.2   | 1576.7   |           |
| P_b           | 874.8      | 1930.2   | 1608.4   |           |
| P+I           | 1010.4     | 2083.4   | 1619.5   |           |

---

To be more precise, WAICU is obtained by adding to C1 a data based bias correction which, analogously to $\rho_{WAIC2}$ in Gelman et al. (2014), uses the posterior variance of individual terms in the log-predictive density summed over the U sample uniques.
ranking. In the WHIP table it is also worth noticing that the NP$_e$ and NP$_d$ models, featuring four interaction terms, turn out to be good models essentially because of the presence of an interaction term (ESEC*WORKP, as can be seen in the supplementary material A.1, Table 6) already selected at the first stage of the procedure by means of $C_0(\gamma)$ and included in two models, NP$_a$ and NP$_c$, among which $C_1$ selects the one including this two-way interaction only. It is this term that marks the difference between good and inadequate models; in fact, in all our preliminary explorations we observed that adding to the independence model a single interaction term (selected by using $C_0(\gamma)$) is enough to enter the range of reasonably good nonparametric models. A theoretical justification to this is provided later, within the discussion of our model selection method. In the remainder of the section we will further comment on the pair ($C_0$, $C_1$) and the results they produce with special attention to both the challenging issues (i) and (ii) recalled at the beginning of the Section. Furthermore, we will discuss the proposed procedure in light of the specialized literature on model selection for disclosure risk estimation reviewed in Section 2. We will see that the problem of model choice, which is a very difficult task in a strictly parametric context, is significantly defused by our approach.

As regards (i), as already stressed, the criterion $C_1$ is based solely on the sample uniques, i.e. a very small subset of the $K$ cells used to fit the model. For illustration, see Table 2 where $U/K \leq 0.001$ for all California tables and $U/(K - \# \text{ structural zeroes}) \leq 0.0169$ for the WHIP table. This fact results in a negligible within-sample-error leading us to omit a bias correction term. Importantly, this means that no additional variability is introduced in the criterion by the bias adjustment. A comment of the same nature can be found in Zucchini (2000, p.53) about the simple criterion $C_{K-L}^{*}$ compared to its bias corrected version $C_{K-L}$. With reference to Underhill and Smith (2016, p.1027), whose BPSIC is a Bayesian, very adaptable and refined evolution of the criterion $C_{K-L}$ (see also Linhart and Zucchini, 1986), our omission can be interpreted as an extremization of the general benefit represented by the lower bias correction term applicable when a criterion is “based on the relevant marginal and conditional logarithmic scores of the variables of interest within a larger model”.

As regards (ii), we first comment on $C_0(\gamma)$, the criterion used to select suitable candidate nonparametric models at the first stage of the procedure. While based on a double use of the $K$ sample frequencies, in place of a bias correction term limiting the within-sample-error, (5) includes a relevant “economic” component, $d \times \gamma$, due to the large values of $\gamma$ we employ in the search. This is a request of simplicity, independent of the sample, directly suggested by the great ability of the DP random effects to correct for the over-estimation associated with under-fitting parametric log-linear models observed in Section 2. As a matter of fact, this structure of $C_0(\gamma)$ limits two very different types of bias at the same time: the selection bias (avoiding the additional variability introduced by the bias adjustment) on the one hand, and the unbalance of over-estimates and under-estimates of per cell risks under over-parametrized nonparametric models on the other, as we will see in Section 5. Of course, such a structure of $C_0(\gamma)$ indirectly implies a strongly non uniform prior on the models under consideration.
(in principle all decomposable log-linear models). Finally, $C_0(\gamma)$ is not endowed with a stopping rule since it is used jointly with the application-specific criterion $C_1$. The search stops when, for nonparametric models of increasing complexity, $C_1$ ceases to improve and begins to decline. For instance, re-using the expression of Skinner and Shlomo (2008), in all contingency tables presented in Section 2 two-way interactions identified through $C_0(\gamma)$ are enough to enter the range of “reasonably good” nonparametric models according to $C_1$. This also means that, in different applications, $C_0(\gamma)$ can be employed to identify suitable nonparametric log-linear models jointly with different application-specific criteria. Comparatively, therefore, it emerges as a general purpose criterion. Turning now to the second stage criterion $C_1$, we point out that each of the nonparametric models evaluated through (6), or, possibly, different application-specific criteria, is an average model. This can be seen in the likelihood $L(\beta, m|f_1, \ldots, f_K)$, which is a sum of $B_K$ terms, where $B_K$ is the Bell number, resulting from all possible partitions (clusterings) $C$ of the $K$ sample frequencies $f_1, \ldots, f_K$ in $c$ nonempty clusters ($1 \leq c \leq K$) (see, e.g., Lo, 1984; Liu, 1996). Denoting by $n_j$ the number of cells included in the $j$-th cluster ($1 \leq n_j \leq K$), we can interpret each term in this sum\(^4\) as the product of two factors:

$$
\prod_{j=1}^{c} \int \prod_{k \in \text{cluster } j} \frac{\pi f_k}{f_k!} f_k(w_k^0, \beta + \phi_j) e^{-\pi e(w_k^0, \beta + \phi_j)} dG_0(\phi_j),
$$

i.e. the likelihood corresponding to a parametric log-linear mixed model with the same (very few) fixed effects and a $G_0$ distributed random effect specific to each cluster $j$ belonging to a fixed partition $C$ of $f_1, \ldots, f_K$ into $c$ clusters, times

$$
\left[\Gamma(m + K)\right]^{-1} \Gamma(m) \prod_{j} \Gamma(n_j) \times \cdots \times \Gamma(n_c),
$$

i.e. the probability $Pr\{n_1, \ldots, n_c|C, c\}$ assigned to such partition by the multivariate Ewens distribution (Takemura, 1999; Johnson et al., 2004, chap. 41).

In other words, $L(\beta, m|f)$ is an average of $B_K$ parametric likelihoods (7) according to specific weights on random partitions of the $K$ sample frequencies. This implies various, useful consequences. Of special interest here is that, as $K$ increases, the stimulus received by the mechanism of model averaging just described is extraordinarily strong, since the number of terms summed in the likelihood increases as follows

$$
B_{K+1} = \sum_{s=0}^{K} \binom{K}{s} B_K.
$$

This massive model averaging action implied by the presence of DP random effects strongly limits the need for additional interaction terms to obtain good

\(^4\)In our case (see Carota et al., 2015) it can be written as follows:

$$
\sum_{c=1}^{K} \sum_{C|C_{1}=c} \frac{\Gamma(m) \prod_{j} \Gamma(n_j)}{\Gamma(m + K)} \prod_{j=1}^{c} \int \prod_{k \in \text{cluster } j} \frac{\pi f_k}{f_k!} f_k(w_k^0, \beta + \phi_j) e^{-\pi e(w_k^0, \beta + \phi_j)} dG_0(\phi_j),
$$


models (see Table 2), and explains the permanence of the NP+I model as the starting model. More explicitly, each given candidate nonparametric model is extraordinarily boosted by the increase in $K$, since, for each partition of the $K$ sample frequencies, a possible relation of dependence among observations in the same cluster is explicitly evaluated and exploited for inference. As opposite, as $K$ increases, the parametric counterpart of each candidate nonparametric model becomes more and more under-fitting. This explains the increasing gap between performances of parametric and nonparametric models observed in Figure 1 for California tables of increasing sizes (roughly, going from the Small to the Large table, the estimation error under parametric models increases tenfold.)

In conclusion, the strongly adaptive reinforcement of each candidate nonparametric model that occurs under our approach to model selection produces both a data-driven significant restriction of the search space and a reduction of the sensitivity of risk estimates to the specification of the model, that is a form of robustness. Such two facts concur to produce a substantial simplification of the model selection task.

In comparison with Forster and Webb (2007), in our model selection procedure model averaging and restriction of the set of possible specifications for fixed effects to decomposable structures are applied in reverse order. Further distinctive points of difference are that in our case: 1) model averaging stems automatically from having extended the model class to the family of log-linear mixed models with DP random effects; 2) for each candidate nonparametric model, averaging is performed over $B_K$ parametric models according to the weights just discussed, rather than over the inferences corresponding to the whole class of graphical log-linear models according to weights given by a posterior distribution on the whole model space; 3) the restriction of the possible alternative specifications of fixed effects to the decomposable structures is ineffective in practice, given the extreme closeness of reasonably good nonparametric log-linear models to the starting model, NP+I. This is indicated in all our examples by the values of $C_1$, which induce to stop the search early, and is confirmed by benchmarking our estimates to the true values of risks. In comparison with Skinner and Shlomo (2008), our model selection procedure not only avoids possible problems with nonexistence of ML estimates of their candidates of type (3), but also limits the selection-induced bias due to the large number of models under evaluation. Finally, a point by point comparison with Manrique-Vallier and Reiter (2012) is less agile because of the very different modeling framework they introduce to describe contingency tables. Overall, however, our approach to model selection for disclosure risk estimation relies on a pair of simple, easily applicable, criteria and such a strong reduction of the search space that, in comparison with standard practice, it probably is a way forward.

5. Overfitting and under-estimation: ideas for small area estimation

In this Section we contrast more and less parsimonious nonparametric models for risk estimation, and reconsider the bias that arises in the model fitting
Table 3

| Model    | $\hat{\tau}_1$ | $\hat{\tau}_2$ | $\hat{\tau}^*_1$ | $\hat{\tau}^*_2$ |
|----------|-----------------|-----------------|-------------------|-------------------|
| NP+I     | 32.1 (2.5)      | 86.5 (3.2)      | 32.1 (4.7)        | 86.5 (4.2)        |
| NP+II    | 27.4 (1.4)      | 79.1 (1.6)      | 27.4 (4.0)        | 79.1 (3.1)        |
| P+I      | 32.1 (1.0)      | 76.9 (1.4)      | 32.1 (3.9)        | 77.0 (2.8)        |
| P+II     | 32.5 (2.4)      | 84.8 (2.5)      | 32.5 (4.6)        | 84.8 (3.7)        |

phase, exploring models of type (4) in a different setting, namely in a small and dense contingency table. The reason of this choice is twofold: from a practical perspective, this exploration is a test whose results are useful in applications where the size and sparsity of tables are not as extreme as in disclosure risk estimation; from a technical perspective, such choice guarantees against the severe issues arising when fitting complex log-linear models in the presence of sparse tables (e.g. Fienberg and Rinaldo, 2012).

For illustration, we reconsider the WHIP data (the 7% microdata sample from the Italian National Social Security Administration 2004). As described in the supplementary material A.3, we now obtain a new contingency table of size $K = 3,960$, referred to as the Small WHIP table, based on five spanning (key) variables. With this data at hand, from within our enlarged class of models (4), we reconsider the association between under-estimation of risks and specification of overfitting models discussed in Skinner and Shlomo (2008). So far, the performance of the family of nonparametric log-linear mixed models with DP random effects has been evaluated with the purpose of estimating an overall measure of disclosure risk. Here, given that global risks are estimated by summing the cell-specific quantities $\hat{\tau}_{i,k}$, $i = 1, 2$, we perform a close analysis of the latter in order to draw useful directions for future research. In particular, to highlight that nonparametric models with a small or a large number of parameters may complement each other when different viewpoints are taken, we now compare two “reference” models, namely the nonparametric independence and all-two-way interactions models (NP+I and NP+II, representative of more and less parsimonious models) with respect to both of the above recalled estimation goals, that is, global and cell-specific risk estimation. We show that a model with a large number of parameters having a poor performance in global disclosure risk estimation, i.e. an “over-parametrised”/overfitting model for this purpose, may perform better than a less parametrised model at estimating the per cell risks over certain subsets of sample unique cells, specifically, the low risk cells. For illustration, it will also be useful to consider the fully parametric counterparts of these models (P+I and P+II, respectively).

Table 3 reports true and estimated values of the global risks $\tau_1$ and $\tau_2$ (standard errors, s.e., in parentheses). Plots in Figure 4 present the 2.5th, 5th, 50th, 95th and 97.5th percentiles of the posterior distribution of $\tau_i$, $i=1,2$, under the same models. Details about priors on model’s parameters and computations are in the supplementary materials A.3 and B, respectively. In Figure 4 we notice the expectedly good performance of the default nonparametric model NP+I,
roughly comparable to that of the parametric model, P+II. At the same time, we observe that for nonparametric models the bias in estimating global risks increases with the number of fixed effects (see also Table 3). Moreover, Table 3 provides a clear indication that, as the model complexity increases, the variability corresponding to parametric models increases as well, while the variability corresponding to nonparametric models tends to decrease. Since all of these phe-

![Figure 4](image_url)

Fig 4. True values (horizontal solid line) and quantiles (0.005, 0.025, 0.50, 0.975, 0.995) of the posterior distributions of $\tau_1$ (left) and $\tau_2$ (right) under all parametric and nonparametric models considered for the Small WHIP table.

nomena are clearly noticeable by considering $\hat{\tau}_i$ or $\hat{\tau}_i^*$, $i = 1, 2$, to quickly get at the root of the behaviour of our estimators, in the rest of the Section we focus on per-cell risk estimates $\hat{\tau}_{i,k}$, $i = 1, 2$. This means that the variability of the $F_k$s is neglected, but provides us with an effective simplification. For instance, Table 3 clearly shows that the NP+II model is an overparametrized model when estimation of the global risks (1) and (2) is sought; but let us now consider the estimation of per cell risks. Analyses of the performance of estimators of $\tau_{2,k}$, reported in Table 4, show that, going from the NP+I to the NP+II model, the improvement of per-cell risk estimates for low risk cells (large $F_k$s) tends to be greater than the improvement of per-cell risk estimates for high risk cells (small $F_k$s), a fact that may explain the increasing negative bias observed at global level in Table 3. An analogous suggestion comes from Figure 5, concerning estimators of $\tau_{1,k}$. Figure 5 presents the boxplots of estimates of $\tau_{1,k}$ restricted to cells which are population uniques ($F_k = 1$). It shows that the two-way interactions models, NP+II and P+II, are by far superior to the nonparametric independence model, NP+I, on those cells. Therefore, from this Figure we can conclude that the worse performance at global level of the NP+II model observed in Table 3, not only in comparison with the NP+I model but also with the P+II model, is an unpleasant consequence of the the greater improvement achieved by the nonparametric all-two-way interaction model NP+II on cells where the true risk $\tau_{1,k}$ is zero ($F_k > 1$). Further evidence about this fact is given in Figure 6. In other words, these analyses reveal a trade-off between
Table 4
Signed, absolute and squared errors for the estimation of $\tau_{2k}^*$ under all models considered in the Small WHIP table: for all cells (left panel), restricted to cells having large frequency in the population ($F_k > 3$ and $F_k > 10$; central panel), and restricted to cells having small frequency in the population ($F_k \leq 3$ and $F_k \leq 10$; right panel).

| Model | all cells | cells s. t. $F_k > 3$ | cells s. t. $F_k \leq 3$ |
|-------|-----------|----------------------|--------------------------|
|       | sign. abs. sq. | sign. abs. sq. | sign. abs. sq. |
| NP+I  | -7.9 80.5 38.7 | 31.4 38.7 13.3 | -39.3 41.8 25.4 |
| NP+II | -15.3 81.4 40.2 | 24.8 37.6 12.6 | -40.1 43.8 27.6 |
| P+I   | -17.4 94.7 51.6 | 27.8 46.3 19.2 | -45.2 48.4 32.4 |
| P+II  | -9.6 84.7 41.8 | 29.6 41.6 14.9 | -39.2 43.1 26.9 |

In conclusion, through a long journey around different forms and sources of bias, we gained a new perspective on model selection for disclosure risk estimation yielding an effective, simple method also able to face the challenging issue, rarely treated in the literature, of the selection-induced bias. In addition, we acquired new solid suggestions about modelling in small area estimation problems.
Fig 5. Boxplots of the estimated values of $\tau^*_1$ for population uniques only, i.e. cells where $\tau^*_1 = 1$, under the nonparametric independence and all two-way interactions models, NP+I and NP+II, respectively, and their parametric counterparts, P+I and P+II; Small WHIP table.

References

Bethlehem, J., Keller, W., and Pannekoek, J. (1990). “Disclosure control of microdata.” *Journal of the American Statistical Association*, 85: 38–45.

Carlson, M. (2002). “Assessing Microdata Disclosure Risk Using the Poisson-Inverse Gaussian Distribution.” *Statistics in Transition*, 5: 901–925.

Carota, C., Filippone, M., Leonbruni, R., and Polettini, S. (2015). “Bayesian nonparametric disclosure risk estimation via mixed effects log-linear models.” *Annals of Applied Statistics*, 9: 525–46.

Cawley, G. C. and Talbot, N. L. C. (2007). “Preventing over-fitting during model selection via Bayesian regularization of the hyper-parameters.” *Journal of Machine Learning Research*, 8: 841–861.

— (2010). “On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation.” *Journal of Machine Learning Research*, 11: 2079–2107.

Chatfield, C. (1995). “Model uncertainty, data mining and statistical inference.” *Journal of the Royal Statistical Society, Series A*, 158: 419–46.

Chen, G. and Keller-McNulty, S. (1998). “Estimation of Identification Disclosure Risk in Microdata.” *Journal of Official Statistics*, 14: 79–95.

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). “Hybrid Monte Carlo.” *Physics Letters B*, 195(2): 216–222.
Fig 6. Risk estimates $\hat{\tau}_{1,k}$ (top row) and $\hat{\tau}_{2,k}$ (bottom row) for sample unique cells under the all-two-way interactions parametric (left) and nonparametric (right) models in the Small WHIP table. Cells are arranged in increasing order of true per-cell risk (red line).
Dunson, D. B. and Xing, C. (2009). “Nonparametric Bayes Modeling of Multivariate Categorical Data.” *Journal of the American Statistical Association*, 104(487): 1042–1051.

Elamir, E. A. H. and Skinner, C. J. (2006). “Record Level Measures of Disclosure Risk for Survey Microdata.” *Journal of Official Statistics*, 22(3): 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(2): 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.

Fienberg, S. E. and Rinaldo, A. (2007). “Three centuries of categorical data analysis: Log-linear models and maximum likelihood estimation.” *Journal of Statistical Planning and Inference*, 137(11): 3430–3445. Special Issue: In Celebration of the Centennial of The Birth of Samarendra Nath Roy (1906-1964).

Fienberg, S. E. (2012). “Maximum likelihood estimation in log-linear models.” *The Annals of Statistics*, 40(2): 996–1023.

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(5): 551–570.

Gelman, A., Hwang, J., and Vehtari, A. (2014). “Understanding predictive information criteria for Bayesian models.” *Statistics and computing*, 6: 997–1016.

Girolami, M. and Calderhead, B. (2011). “Riemann manifold Langevin and Hamiltonian Monte Carlo methods.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2): 123–214.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (2004). *Discrete Multivariate Distributions. Wiley Series in Probability and Statistics*. New York: John Wiley and Sons.

Linhart, H. and Zucchini, W. (1986). *Model selection*. New York: Wiley.

Liu, J. S. (1996). “Nonparametric Hierarchical Bayes via Sequential Imputations.” *Annals of Statistics*, 24(3): 911–930.

Lo, A. Y. (1984). “On a class of Bayesian nonparametric estimates. I. Density estimates.” *Annals of Statistics*, 12(1): 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.

— (2014). “Bayesian Estimation of Discrete Multivariate Latent Structure Mod-
els with Structural Zeros.” *Journal of Computational and Graphical Statistics*, 23: 1061–1079.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). “Equation of state calculations by fast computing machines.” *The Journal of Chemical Physics*, 21(6): 1087–1092.

URL http://dx.doi.org/10.1063/1.1699114

Miller, A. (1990). *Subset selection in regression*. London: Chapman and Hall.

Neal, R. M. (1993). “Probabilistic Inference using Markov chain Monte Carlo Methods.” Technical Report CRG-TR-93-1, Dept. of Computer Science, University of Toronto.

— (2000). “Markov Chain Sampling Methods for Dirichlet Process Mixture Models.” *Journal of Computational and Graphical Statistics*, 9(2): 249–265.

URL http://www-clmc.usc.edu/~{}cs599_ct/neal-TR1998.pdf

Piironen, J. and Vehtari, A. (2017). “Comparison of Bayesian predictive methods for model selection.” *Statistics and Computing*, 27: 711–735.

Reunanen, J. (2003). “Overfitting in making comparisons between variable selection methods.” *Journal of Machine Learning Research*, 3: 1371–1382.

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

Roberts, G. O. and Rosenthal, J. S. (2009). “Examples of adaptive MCMC.” *Journal of Computational and Graphical Statistics*, 18(2): 349–367.

URL http://dx.doi.org/10.1198/jcgs.2009.06134

Ruggles, S., Genadek, K., Goeken, R., Grover, J., and Sobek, M. (2017). “Integrated Public Use Microdata Series: Version 7.0 [dataset].”

URL https://doi.org/10.18128/D010.V7.0

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(5): 499–521.

Skinner, C. and Elliot, M. (2002). “A measure of disclosure risk for microdata.” *Journal of the Royal Statistical Society, Series B*, 64: 855–867.

Skinner, C. and Shlomo, N. (2008). “Assessing Identification Risk in Survey Microdata Using Log-Linear Models.” *Journal of the American Statistical Association*, 103(483): 989–1001.

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

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

Underhill, N. T. and Smith, J. Q. (2016). “Context-Dependent Score Based Bayesian Information Criteria.” *Bayesian Analysis*, 11: 1005–1033.

Varma, S. and Simon, R. (2006). “Bias in error estimation when using cross-validation for model selection.” *BMC Bioinformatics*, 7: 91.

Vehtari, A. and Ojanen, J. (2012). “A survey of Bayesian predictive methods for model assessment, selection and comparison.” *Statistics Surveys*, 6: 142–228.
Watanabe, S. (2009). *Algebraic Geometry and Statistical Learning Theory*. Cambridge: Cambridge University Press.

Zucchini, W. (2000). “An Introduction to Model Selection.” *Journal of Mathematical Psychology*, 44: 41–61.
SUPPLEMENTARY MATERIAL

A.1. Illustrative tables and models used in Sections 3 and 4

In this material we provide details about the data and models that were used in the illustrative examples summarized in Sections 3 and 4 of the article.

First, we reconsider the same table of 3,600,000 cells used in Manrique-Vallier and Reiter (2012). A description of the ten key variables generating such “large” California table is provided in Table 5 (left panel). It is built from the 5% public use microdata sample of the U.S. 2000 census for the state of California (IPUMS Ruggles et al., 2017); the set of $N = 1,150,934$ individuals aged 21 and over is taken as the reference population. In addition, two new contingency tables are obtained by global suppression of variables DISAB and VETST, yielding a “medium” table of 900,000 cells, and global suppression of variable INCR, yielding a “small” table of 360,000 cells. Clearly, global suppression is not applied here as a protection strategy, but rather as a way to create contingency tables with different characteristics.

Since by design the previous tables do not include structural zeroes, we also consider a contingency table of 844,800 cells, half of which are structurally empty. The data come from the set of $N = 794,986$ individuals recorded in the 7% public use microdata sample of the Italian National Social Security Administration, 2004 (source: Work Histories Italian Panel, WHIP), treated here as the population; these records are cross-classified according to the eight key variables listed in Table 5 (right panel). In all cases we draw random samples with fraction $\pi = 0.05$.

Table 5

| Large California | WHIP |
|------------------|------|
| **Label** | **Variable** | **Label** | **Variable** |
| CHIL | number of children (10) | AORIG | area of origin (11) |
| AGE | age (10) | AGE | age (12) |
| SEX | sex (2) | SEX | sex (2) |
| MARST | marital status (6) | RWORK | region of work (20) |
| RACE | race (5) | ESEC | economic sector (4) |
| EDU | education (5) | WAGF | wage guaranteed fund (2) |
| EMPST | employment status (3) | WORKP | working position (4) |
| INCN | income (10) | FSIZE | firm size (5) |
| DISAB | disability (2) | | |
| VETST | veteran status (2) | | |

For each of the four contingency tables just described, we consider the models listed in Table 6. The nonparametric (NP) models are selected as illustrated at the beginning of Section 3 and formally described in Section 4. Often, for comparison, we also consider the parametric counterparts of such models, labelled P. The latter are Bayesian log-linear models with the same fixed effects.
Table 6
Log-linear models for California and WHIP tables: model label, structure of fixed effects (parametric component), prior on the random effects, and number of additional parameters compared to those included in the independence model.

| Label   | Shorthand for fixed effects | Prior for random effects | n. of extra parameters |
|---------|-----------------------------|--------------------------|------------------------|
| California Large | | | |
| NP_a   | I + SEX*VETST                | \(D(m, Ga(a, b))\) | 1                     |
| NP_b   | I + EMPST*INCR              | \(D(m, Ga(a, b))\) | 18                    |
| NP_c   | I+ SEX*VETST + EMPST*INCR   | \(D(m, Ga(a, b))\) | 19                    |
| P_a    | I + SEX*VETST                | \(Ga(a, b)\) | 1                     |
| P_b    | I + EMPST*INCR              | \(Ga(a, b)\) | 18                    |
| P_c    | I+ SEX*VETST + EMPST*INCR   | \(Ga(a, b)\) | 19                    |
| California Medium | | | |
| NP_a   | I + EMPST*SEX                | \(D(m, Ga(a, b))\) | 3                     |
| NP_b   | I + EMPST*INCR              | \(D(m, Ga(a, b))\) | 18                    |
| NP_c   | I + EMPST*SEX + EMPST*INCR  | \(D(m, Ga(a, b))\) | 21                    |
| NP+I   | I                           | \(D(m, Ga(a, b))\) | -                     |
| P_a    | I + EMPST*SEX                | \(Ga(a, b)\) | 3                     |
| P_b    | I + EMPST*INCR              | \(Ga(a, b)\) | 18                    |
| P_c    | I + EMPST*SEX + EMPST*INCR  | \(Ga(a, b)\) | 21                    |
| P+I    | I                           | \(Ga(a, b)\) | -                     |
| California Small | | | |
| NP_a   | I + SEX*VETST                | \(D(m, Ga(a, b))\) | 1                     |
| NP_b   | I + MARST*RACE              | \(D(m, Ga(a, b))\) | 20                    |
| NP+I   | I                           | \(D(m, Ga(a, b))\) | -                     |
| P_a    | I + SEX*VETST                | \(Ga(a, b)\) | 1                     |
| P_b    | I + MARST*RACE              | \(Ga(a, b)\) | 20                    |
| P+I    | I                           | \(Ga(a, b)\) | -                     |
| WHIP   | | | |
| NP_a   | I + ESEC*WORKP              | \(D(m, Ga(a, b))\) | 9                     |
| NP_b   | I + ESEC*FSIZE              | \(D(m, Ga(a, b))\) | 12                    |
| NP_c   | I + ESEC*WORKP + ESEC*FSIZE | \(D(m, Ga(a, b))\) | 21                    |
| NP_d   | I + ESEC*WORKP + ESEC*SEX   | \(D(m, Ga(a, b))\) | 27                    |
|       |                             | ESEC*WAGF + ESEC*FSIZE  |                       |
| NP_e   | I + ESEC*WORKP + ESEC*SEX   | \(D(m, Ga(a, b))\) | 48                    |
|       |                             | ESEC*WAGF + AGE*WORKP   |                       |
| NP+I   | I                           | \(D(m, Ga(a, b))\) | -                     |
| P+I    | I                           | \(Ga(a, b)\) | -                     |

and parametric (Gamma distributed, as it is explained later) random effects: they represent the special case of NP models for an indefinitely large \(m\). As recalled in Section 2, in practice such parametric counterparts are equivalent to log-linear models without random effects \(3\), since the corresponding risk estimates are nearly identical (Elamir and Skinner, 2006).

Using the fact that, conditionally on the random effects (\(NP\) or \(P\)), our models differ only for the specification of the vector \(\beta\), we further distinguish them by referring to their parametric component. For instance, as we use the shorthand notation \(NP+I\) to denote the nonparametric model whose fixed effects include the main effects of all key variables (nonparametric independence model); likewise, we denote by \(NP+I+A*B\) the nonparametric model whose fixed effects additionally include the two-way interaction parameters between all levels of key variables A and B. Therefore, a comprehensive description of all
models explored in Sections 3 and 4 can be read in columns of Table 6: model label (dictated by the nature of random effects), shorthand for fixed effects included in the model, prior on random effects and number, $d$, of extra parameters, implied by the interaction terms (e.g., A*B) added to the independence model.

In all of these applications we reparametrize the random effects so that $\omega_k = e^{\phi_k}$ is drawn from a Dirichlet process with base measure $G_\alpha(a, b)$ (where $b$ is the rate parameter), thereby extending the model defined in Elamir and Skinner (2006), and assume a Gamma-distributed precision $m$. The hyperparameters are fixed so as to specify vague priors: for the random effects we take $a = 1, b = 0.1$; as regards the fixed effects $\beta$, we consider a vague Gaussian prior, $N(0, I\sigma^2)$; finally, we take $m \sim \text{Gamma}(1, 0.1)$.

A.2. Analysis of per cell risk estimates under nonparametric vs. parametric independence models

Figure 7 presented in this material compares the estimates of $\tau_{1,k}$ obtained under the nonparametric independence model NP+I and under its parametric counterpart P+I, in three different tables: California Medium, California Small and Whip. We can observe different sigmoidal shapes: two of them (first two rows) describe corrections of parametric per-cell risk estimates sufficient to obtain good nonparametric estimates of global risks (see Figure 1); the sigmoidal shape in the third row, though much more pronounced, does not achieve the same result (see Figure 2).

A.3. Illustrative table and models used in Section 5

In Section 5 we perform a detailed analysis of per cell risk estimates, comparing the estimation bias associated to more and less parsimonious nonparametric models. We refer to a small and dense contingency table, called the Small WHIP table, that, once again, is obtained from the 7% microdata sample of the Italian National Social Security Administration, 2004. This time we treat the $N = 450,238$ individuals whose workplace falls into 4 specific geographic areas as the reference population, from which we draw a random sample with fraction $\pi = 0.1$. The table has the following spanning (key) variables (number of levels in parentheses): geographic area (4), sex (2), age (11), ethnicity (5), and economic activity (9), returning a total of $K = 3,960$ cells.

We consider as “reference” models (representative of more and less parsimonious models) the nonparametric independence and the all-two-way interactions models, NP+I and NP+II, respectively, also estimating their parametric counterparts, P+I and P+II, for comparison. In this application the base measure of the DP prior for random effects is $G_\alpha = \mathcal{N}(\alpha, \sigma^2)$, which extends the model in Skinner and Holmes (1998); we assume $\alpha \sim \mathcal{N}(0, 10)$, $\sigma^2 \sim \text{invGamma}(1, 1)$ and $m \sim \text{Gamma}(1, 1)$. Finally, we take a reasonably vague Gaussian prior $\mathcal{N}(0, 10I)$ on $\beta$. 
A.4. Some remarks on standard errors of global risk estimators

Here we try to explain why the standard error decreases when going from the NP+I model to the NP+II model, and viceversa increases when going from the P+I to the P+II model. If we consider the deviances of the per-cell risk estimates $\hat{\tau}_{i,k}$ around their mean $\hat{\tau}_i^*/K$ (hereafter denoted by $D_b$) for the parametric and nonparametric all-two-way interactions models, we obtain very similar values and, given that the $\hat{\tau}_{i,k}$ are in turn means within each cell, from

$$(s.e.(\hat{\tau}_i^*))^2 = \frac{1}{H} \sum_h \left( \sum_k \tau_{i,k}^{*(h)} - \hat{\tau}_i^* \right)^2 = V_w + D_b + C_b,$$  \hspace{1cm} (8)$$

where

$$V_w = \sum_k \frac{1}{H} \sum_h (\tau_{i,k}^{*(h)})^2 - \sum_k (\hat{\tau}_{i,k}^*)^2$$  \hspace{1cm} (9)$$

is the sum of the variances within each cell,

$$D_b = \sum_k (\hat{\tau}_{i,k}^*)^2 - K \left( \frac{\hat{\tau}_i^*}{K} \right)^2$$  \hspace{1cm} (10)$$

is the deviance between cells and

$$C_b = \sum_k \sum_{j \neq k} \frac{1}{H} \sum_h \tau_{i,k}^{*(h)} \tau_{i,j}^{*(h)} - K(K-1) \left( \frac{\hat{\tau}_i^*}{K} \right)^2$$  \hspace{1cm} (11)$$

is the sum of codeviances between cells, we can conclude that the smaller standard error observed in Table 3 under the NP+II model compared to the one under the P+II model is due to smaller values of the variances within cells and/or of codeviances between per-cell risks. If, in addition, we consider a nonparametric model without fixed effects -under which $\hat{\tau}_1^* = 16.5(4.5)$ and $\hat{\tau}_2^* = 76.4(7.4)$-, where the components $V_w$ and $C_b$ are essentially the only relevant components of the s.e. being $D_b < 0.001$, we can affirm that, as the complexity of the nonparametric log-linear model increases, the decrease of $V_w$ and/or $C_b$ prevails over the slight increase of $D_b$. Vice versa, going from the parametric independence model P+I to the all two-way interactions model P+II, the component $D_b$ slightly decreases and is overwhelmed by the increase of $V_w$ and/or $C_b$. In this respect, consider that under parametric models the only way to increase the association between cells is by means of the introduction of further interaction terms.

B. Implementation of the MCMC approach

The Markov Chain Monte Carlo (MCMC) sampler employed here is a Gibbs sampler, where groups of parameters are sampled one after the other. In particular, the sequence of MCMC steps amounts in drawing samples from the...
conditionals $\beta|\text{rest}$, $\phi|\text{rest}$, and $m|\text{rest}$. Samples from the posterior distribution over $\beta$, $\phi$, and $m$ allows one to estimate per-cell risks through Monte Carlo averaging.

**Sampling $\beta$** – The conditional distribution of $\beta|\text{rest}$ is not of known form, given that the prior on $\beta$ is Gaussian and the likelihood is Poisson. Therefore, we employ Metropolis-within-Gibbs samplers, where a proposal is accepted or rejected according to a Metropolis ratio (Roberts and Rosenthal, 2009); these can include, e.g., Metropolis-Hastings (Metropolis et al., 1953) or Hybrid Monte Carlo (Duane et al., 1987; Neal, 1993), but in this work we employ the so-called Simplified Manifold Adjusted Langevin Algorithm (SMMALA) (Girolami and Calderhead, 2011). SMMALA is one instance of manifold MCMC methods, which are characterized by the fact that they exploit the curvature of the log-likelihood, allowing for efficient moves in the parameter space. SMMALA has been shown to be effective for problems similar to the ones considered here, where the posterior is unimodal and is not characterized by strong skewness. SMMALA approximates the diffusion on the statistical manifold characterizing $p(f_1, \ldots, f_K|\beta, \text{rest})$. Defining $M$ to be the metric tensor obtained as the Fisher Information of the model plus the negative Hessian of the prior, and $\epsilon$ to be a discretization parameter, SMMALA can be thought of as a Metropolis-Hastings sampler with a position-dependent proposal. The curvature of the log-likelihood determines the step-size of the proposal through the metric tensor $M$ as follows $p(\beta'|\beta) = N(\beta'|\mu, \epsilon^2 M^{-1})$, with $\mu = \beta + \frac{\epsilon^2}{2} M^{-1} \nabla_{\beta} \log[p(f_1, \ldots, f_K|\beta, \text{rest})]$. The complexity of the update is $\mathcal{O}(KD^3)$ where $K$ is the number of cells and $D$ is the size of $\beta$; the linearity in $K$ makes it well suited in applications where the number of cells is large, while the cubic scaling in $D$ makes it suitable for models with a small number of $\beta$ parameters.

**Sampling $\phi$** – In Sections 3 and 4 of this work, we exploit the conjugacy between the base Gamma measure and the Poisson likelihood to derive an efficient sampler for $\phi$. We implemented the MCMC sampler “Algorithm 3” proposed in the review paper of MCMC methods for DP models in Neal (2000). In a nutshell, we choose a distribution for $G_0$ such that $\omega = e^\phi$ is given a Gamma base measure, for which we can exploit conjugacy with the Poisson likelihood. A similar argument holds when $\phi$ is given the IG distribution. This allows us to integrate out the values of $\phi$ analytically $\int p(f_k|\beta, \phi) dG_0(\phi)$, where we expressed $p(f_k|\beta, \phi)$ as the likelihood for a single point. As a result, it is possible to derive a sampler that allocates cells to an unknown number of clusters and to draw directly a value for the random effect for each cluster. The complexity of the update is $\mathcal{O}(K)$.

In Section 5, where the base measure is not conjugate with the Poisson likelihood, we adopt the “Algorithm 5” in Neal (2000), as it easy to implement and as it achieves satisfactory performance in the given application.

**Sampling $m$** – We choose a Gamma prior for the $m$ parameter. With this choice, it is possible to draw samples from the posterior distribution over $m|\text{rest}$ directly following Escobar and West (1994).
Fig 7. California Medium, California Small and WHIP tables (first, second and third row, respectively). For each of the three tables, the figures report estimates of per cell risks, $\hat{\tau}_{1,k}$, under the nonparametric independence model (horizontal axis) and under its parametric counterpart, in the presence of a vague Gaussian prior on $\beta$ (column b), or a degenerate prior putting mass at the ML estimate of $\beta$ (column c). The latter estimates are referred to as nonparametric empirical Bayesian estimates, and the corresponding model is denoted by the additional subscript “Emp”. For the sake of completeness, column (a) reports both the nonparametric and nonparametric empirical Bayesian per cell risk estimates of $\tau_{1,k}$. 