International Statistical Review

International Statistical Review (2021) doi: 10.1111/insr.12471

Assessing Bayesian Semi-Parametric Log-Linear Models: An Application to Disclosure Risk Estimation

Cinzia Carota[1], Maurizio Filippone[2] and Silvia Polettini[3]

[1]*Dipartimento di Economia e Statistica "Cognetti de Martiis", Università di Torino, Lungo Dora Siena 100 A, Turin, 10153, Italy*
[2]*Department of Data Science, EURECOM Campus SophiaTech, 450 Route des Chappes, Biot, 06410, France*
[3]*Dipartimento di Scienze Sociali ed Economiche Sapienza Università di Roma P.le Aldo Moro 5, Rome, 00185, Italy*
E-mail: cinzia.carota@unito.it

Summary

We propose a method for identifying models with good predictive performance in the family of Bayesian log-linear mixed models with Dirichlet process random effects for count data. Their wide applicability makes the assessment of model performance crucial in many fields, including disclosure risk estimation, which is the focus of the present work.

Rather than assessing models on the whole contingency table, we target the specific objective of the analysis and propose a two-stage model selection procedure aimed at limiting a form of bias arising in the process of model selection. Our proposal combines two different criteria: at the first stage, a path in the model search space is identified through a strongly penalized log-likelihood; at the second, a small number of semi-parametric models is evaluated through a context-dependent score-based information criterion. Tested on a variety of contingency tables, our method proves to be able to identify models with good predictive performance in a few steps, even in the presence of large tables with many sampling and structural zeros. We carefully discuss the proposed method in the context of the literature on model assessment and contextualize the illustrative application in the recent debate on statistical disclosure limitation. Finally, we provide examples of further applications in different research areas.

Key words: Bayesian model selection; Dirichlet process random effects; Disclosure risk; Log-linear mixed models; Model's predictive performance; Selection-induced bias; Statistical disclosure limitation.

1 Introduction

Log-linear modelling provides a convenient way of investigating relationships among categorical variables in contingency tables. However, when the set of classifying variables is large or there are many categories, the induced table is not only large, but often sparse as well, with a huge set of alternative log-linear specifications. This poses severe issues in both model fitting and selection, such as unidentifiability of parameters, non-existence of maximum likelihood estimates, unreliability of degrees of freedom and indistinguishability of models with good

predictive performance. (Refer to e.g. Fienberg & Rinaldo, 2012; Piironen & Vehtari, 2017, respectively, and references therein). In Carota *et al*. (2015), we proposed a family of Bayesian log-linear models for disclosure risk estimation where the presence of nonparametric random effects allows to avoid the above-mentioned issues in model fitting. Capitalizing on the strengths of this family of models, here we reconsider disclosure control and develop an application-specific model selection method. Rather than assessing models on the whole contingency table, we target the specific objective of the analysis and develop a two-stage procedure aimed at limiting a form of bias arising in the process of model selection.

Under a Bayesian approach, models can be assessed and compared by evaluating their predictive accuracy on future datasets; the posterior predictive density can be used to assess the out-of-sample predictive performance of a model. Measures of predictive accuracy are often based on the log-predictive density, which represents a general summary of model fit. Therefore, an ideal measure of predictive accuracy could be the expected log-posterior predictive density for a future dataset. However this quantity has to be estimated by a suitable 'criterion' that can solely rely on the available data, thus measuring the within-sample predictive accuracy. Clearly, out-of-sample predictions will typically be less accurate than suggested by the within-sample predictive accuracy. This entails a form of bias that is referred to as within-sample error. At the same time, the variance of the criterion used to estimate the predictive accuracy is crucial in model comparison: indeed, in a large set of models to be compared, a criterion with large variance may lead to pick a model by chance rather than by merit. This is another form of bias that we refer to as selection-induced bias. Our model selection method is meant to address both issues, with special emphasis on the second (described in detail in Section 2.2). More precisely, we develop a two-stage procedure that combines the idea of assessing models via scores customized to the end user's utility (Underhill & Smith, 2016) and the great flexibility of the class of models in Carota *et al*. (2015), with the declared purpose of limiting the selection-induced bias. We also provide a careful discussion of this model selection method in light of previous and alternative methods (Sections 2.1,3 and 5).

Indeed, in certain estimation problems, rather than focusing on the overall model performance, it is sensible to favour the model that best estimates the quantity of interest; the criterion may therefore be tailored to the specific objective of the analysis. Disclosure risk estimation offers cogent evidence in favour of this approach. Specifically, we focus on disclosure control in microdata from social surveys released by national statistical institutes or other organisations (hereafter statistical agencies) for research purposes. This is a particularly challenging problem, recently debated in various fields, from computer science and mathematical statistics to social science and public health.

Microdata from social surveys typically include values of sensitive variables (such as income, health status, political orientation, …) and values of demographic variables. Of course, such data are disseminated without direct identifiers (name, surname, etc.), but some categorical variables might still be used as indirect identifiers, that is to say as *keys* for re-identification of respondents, because their values are also available from external sources in non-anonymized files. Examples of key variables include gender, ethnicity, marital status, place of residence, and so forth, also recorded in a number of administrative registers along with direct identifiers. When the released sample includes rare combinations of values of the key variables, the risk of disclosing respondents' identities is high if those combinations of values are also rare in the population. Actually, if variables are recorded without error, a potential intruder can easily discover the identity of data subjects by matching on keys, thus gaining access to their sensitive information. Breaches of confidentiality carry very serious legal consequences and undermine trust in the statistical agency, which becomes likely to get less, or less truthful, answers (i.e. lower quality data) from future respondents.

Disclosure risk estimation is a challenging inferential task since cross-classification of respondents according to the values of the key variables often results in a large contingency table with many sampling and structural zeros. For cells (combinations of values) with small sample frequencies, statistical agencies have to infer how large are the corresponding population frequencies. Traditionally, the literature focused on sample cells with frequency of 1, *sample uniques*, but sample doubles, triples, etc. can also be considered (refer to Section 5). Here, for simplicity, we follow the literature. Descriptions of the most common disclosure risk measures and rich lists of references can be found in Forster & Webb (2007) and Taylor *et al*. (2018).

Let F_k and f_k denote the population and sample frequencies in the kth cell, respectively, and let K be the total number of cells in the contingency tables spanned by the key variables. Two interesting measures of the global risk of re-identification, or disclosure risks of the sample to be released, are the number of sample uniques which are also population uniques,

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

and the number of correct guesses if each sample unique is matched with an individual randomly chosen from the corresponding population cell

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

where $I(A)$ denotes the indicator function of the event A. Usually, τ_1 and τ_2 are approximated by their expected values (e.g. Skinner & Shlomo, 2008) ignoring the source of variability due to the randomness of F_K, which is the reason why recently several authors have preferred to estimate Equations 1 and 2 (Carota *et al*., 2015; Manrique-Vallier & Reiter, 2012). As already mentioned, statistical agencies have a legal obligation to protect confidentiality by keeping the disclosure risk below certain thresholds. At the same time, their core mission is to release high-quality data effective for statistical purposes. Identifying the proper protection amounts to finding the proper balance between disclosure risk and data utility prior to any data release. This requires accurate and repeated estimates of suitable measures of the disclosure risk, which, in turn, demand for ready and safe identification of good models for risk estimation. This article tries attempts to provide a simple method to select useful models.

Before going into technical details, we recall that the debate on dissemination of microdata from social surveys (crucial for policy-relevant research as well as for academic research) has heated up after the US Census Bureau announced its plan to apply differential privacy for disclosure control in public use data products, including microdata derived from the 2020 Census and the American Community Survey (Abowd, 2018a, 2018b). According to differential privacy (a formal model of privacy protection emerged from the computer science literature described, for instance, in Rinott *et al*., 2018), exclusively data that have been perturbed and the corresponding perturbation mechanism can be released. This raised very strong reactions (refer to e.g. Ruggles *et al*., 2019, and related references) that in August 2019 induced the Census Bureau to announce that the earliest date for implementation of differential privacy for the American Community Survey will be 2025. In the context of the European Statistical System, the implementation of differential privacy in official statistics has been questioned (refer to e.g. Eurostat, 2017), and it is not clear if the discussion will be reopened (Eurostat, 2018). In the meanwhile, the General Data Protection Regulation (GDPR) entered into force in all EU member states. The GDPR strengthens the rights of data subjects and

obligations of data controllers; nevertheless, processing of personal data for statistical purposes is firmly guaranteed by the two legal principles of 'necessity' and 'proportionality' (Art. 89). In a nutshell, the differential privacy approach is elegant and aimed at an automated data dissemination, but substantial work on statistical inference from perturbed data is still ongoing (given that 'naive' inference is severely misleading) and the gap between legal and computer science definitions of privacy has to be bridged. Moreover, differential privacy seems to be unnecessarily protective and wasteful in terms of data quality when samples of microdata with small sampling fraction are released for research purposes (usually under special license agreements) (Shlomo *et al*., 2015, p. 307). These facts suggest to maintain the current approach to disclosure control, under which statistical agencies estimate the disclosure risk of a sample assuming the so-called matching on keys re-identification scenario.

The outline of the paper is as follows. Section 2 provides the necessary background. Section 3 presents our proposal and discusses its features. A detailed illustration is provided in Section 4. Finally, Section 5provides some concluding remarks, and discusses the generalizability of the method to a variety of inferential problems arising in completely different fields of research.

2 Background

The literature has highlighted the crucial role of the model in the process of risk estimation since the seminal paper by Bethlehem *et al*. (1990) (refer to e.g. Taylor *et al*., 2018, and references therein). Estimating the risk measures (1) and (2) amounts to predicting the unknown frequencies $(F_k - f_k)$ of the out-of-sample data exclusively for those cells where $f_k = 1$. In selecting a model for risk estimation, this will lead us to focus on measures of predictive accuracy restricted to such specific subset.

In this section, we recall the family of semi-parametric log-linear models proposed in Carota *et al*. (2015) and briefly discuss the issue of selecting a good log-linear model for disclosure risk estimation in a fully parametric family. Then, we discuss the selection-induced bias in the context of predictive methods for model selection.

2.1 Log-linear Models for Disclosure Risk Estimation and Previous Approaches to Model Selection

Carota *et al*. (2015) model population and sample frequencies, F_k and f_k, by independent Poisson distributions with rates λ_k and $\pi\lambda_k$, respectively, where π denotes the sampling fraction supposed to be known. The parameters $\lambda = (\lambda_1, \ldots, \lambda_k, \ldots, \lambda_K)$ are described by a log-linear model with mixed effects:

$$\lambda_k = e^{\mu_k},\ \mu_k = \mathbf{w}_k'\boldsymbol{\beta} + \phi_k,\ \boldsymbol{\beta} \sim N(\mathbf{0}, I\sigma^2),\ \phi_k | G \overset{i.i.d.}{\sim} G,\ G \sim \mathrm{DP}(m, G_0), \tag{3}$$

where $\mathbf{w}_k$ is a $q \times 1$ design vector depending on the values of the key variables in cell k, $\boldsymbol{\beta}$ is a $q \times 1$ vector of normally distributed fixed effects, and ϕ_k is a random effect accounting for cell specific deviations. The distribution of ϕ_k, denoted by G, is assumed to be unknown and a priori distributed according to a Dirichlet process, DP, with expectation G_0 and precision parameter m (Ferguson, 1973). Interesting choices of G_0 are a Normal or an Inverse Gaussian or a suitable transformation of a Gamma distribution (details in Section 4). In all such cases, the DP prior represents a relaxation of parametric distributional assumptions for the random effects whose implications on disclosure risk estimation have been well established in the literature (refer to Carlson, 2002; Elamir & Skinner, 2006; Skinner & Holmes, 1998, respectively). Consequently, the specification (3) implements a standard strategy in Bayesian nonparametric

(BNP) modelling (refer to e.g. Muller *et al*., 2015, chap. 7.1, and references therein). A DP prior is the simplest, and often computationally convenient, way to relax the choice of a parametric distribution G_0 for the random effects that (unlike the prior for the fixed effects) is rarely based on substantive prior information. A standard hyperprior for m is a gamma distribution and possible hyperparameters indexing G_0, in turn, can be easily modelled if additional flexibility is desired.

Indeed, such BNP modelling is often grounded on careful analyses of the impact of misspecification of G_0. For example, in the context of generalized linear mixed models used to analyse clustered or longitudinal data, McCulloch & Neuhaus (2011) elucidate situations in which such specification may and may not matter. In the case of disclosure risk estimation, instead, Elamir & Skinner (2006) show that any parametric distributional assumption for the random effects is irrelevant, as they find that, in order to obtain good risk estimates, the inclusion of random effects is not needed. For this reason, Skinner & Shlomo (2008) adopt a log-linear model without random effects and address the delicate issue of model selection (i.e. selection of suitable βs) from within this family. However, Carota *et al*. (2015) show that the inclusion of DP-distributed random effects in a log-linear model with parametric fixed effects results in greatly improved risk estimates and a drastic reduction in the number of fixed effects needed to achieve this goal. This radical change in perspective, consisting of a shift of focus from parametric fixed effects to nonparametric random effects, along with the results in Elamir & Skinner (2006) reveals a nonstandard rationale behind the otherwise standard BNP modelling strategy of Equation 3. Under this different perspective the parametric fixed effects are in a sense subordinate to the nonparametric random effects and not interesting for their own sake1 For this reason, Carota *et al*. (2015) refer to their approach as BNP log-linear modelling, though it is a semi-parametric approach.. Importantly, moreover, in small contingency tables, the reduction in the number of fixed effects needed to achieve good risk estimates is enough to make the vector $\boldsymbol{\beta}$ identifiable and to ensure existence of its maximum likelihood estimate, two properties that are not guaranteed in a fully parametric model. But two questions arise: what happens in larger tables? How is it possible to select a good model? In Section 3, we will address the issue of model selection in the presence of inferential targets (τ_1 and τ_2) that depend on a specific subset of values of the response f_k in any log-linear model, no matter how the priors of fixed and random effects are specified. All these facts will lead us to define an application-specific model selection method within the family (3), rather than apply techniques for variable selection based on a joint DP model for both fixed and random effects with a spike and slab base measure for βs (refer to Barcella *et al*., 2017, for a recent survey). Such approaches introduce a family of models larger than (3), namely the covariate-dependent Dirichlet process mixture models, and target the problem of cluster-specific covariate selection. As a consequence, observations are grouped on the basis of the potentially different effects of the covariates on the response variable. Highlighting the different, cluster specific, role of a large set of covariates, none of which in principle can be excluded from those larger models is out of the scope of our work. Instead, we are interested in selecting a preferably small, common to clusters, set of fixed effects βs that effectively supplements the nonparametric random effects in estimating the risk measures τ_1 and τ_2 within the family (3). This is the reason why we will pursue a forward search of a parsimonious model, imposing a strong penalty for overfitting.

We next recall some of the previous approaches to model selection for risk estimation. In the family of log-linear models without random effects, Skinner & Shlomo (2008) address model selection with the declared purpose of limiting the bias of risk estimates. Their proposal has a twofold justification. On the one hand, standard tools for assessing and selecting models, such as χ^2 goodness-of-fit tests and Akaike's information criterion, 'are not very successful in

deciding whether the disclosure risk measures will be well estimated' (Skinner & Shlomo, 2008, p. 991). On the other hand, in controlled settings, the bias of maximum likelihood estimators of the global risks $E(\tau_1) = \sum_{k=1}^{K} I(f_k = 1)\Pr(F_k = 1|f_k = 1)$ and $E(\tau_2) = \sum_{k=1}^{K} I(f_k = 1)E(1/F_k)$ evolves monotonically from overestimation to underestimation 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. Building on these findings, they develop a criterion ($\hat{B}$) that detects underfitting, being able to estimate the positive bias of risk estimators. Then, among models I, II, III,…, they select the least underfitting as the starting model and propose a stepwise forward model search aimed at minimizing the positive bias of the maximum likelihood estimators of $E(\tau_1)$ and $E(\tau_2)$. The authors anticipate that most likely a number of 'reasonable models' may exist, between which their criterion $\hat{B}$ is not able to discriminate (Skinner & Shlomo, 2008, pp. 993–994). Indeed, the real limitation of their approach is that it is highly exposed to the severe issues recalled at the beginning of Section 1 because reasonable models are very often too complex to be identifiable (details in Carota *et al*., 2015, p. 529). All these issues are instead avoided by Forster & Webb (2007), who focus exclusively on graphical decomposable log-linear models and account for model uncertainty by averaging inferences over that special sub-family of log-linear models.

2.2 Selection-Induced Bias

The literature on predictive methods for model assessment, selection and comparison (Vehtari & Ojanen, 2012; Gelman *et al*., 2014; Underhill & Smith, 2016; Piironen & Vehtari, 2017, and references therein) points out two critical issues in estimating model predictive accuracy. These points stem from the decomposition of the estimation error into bias and variance of any criterion and form the backdrop to our proposal:

(i) the need to correct for the bias arising from a double use of the data (*within-sample-error*), when model evaluation relies on predictions of the data used to fit it; and
(ii) the need to limit the variance in some way, since a criterion with a non-negligible variance has the potential for overfitting in the process of model selection by exploiting meaningless peculiarities of the sample over which it is evaluated. This form of overfitting, analogous to the more familiar one occurring in training the model, is termed *selection-induced bias* because it results in the undesirable optimistic bias in predictive performance evaluation that, in the presence of scarce data and large sets of models, often leads to select a model by chance rather than by merit.

It has long been known that any criterion suffers from selection-induced bias, and in the last decade the severity of such problem has been re-affirmed and quantified for a series of established and recent criteria, including cross validation, the Akaike Information Criterion (AIC), the Deviance Information Criterion (DIC), the Widely Applicable Information Criterion (WAIC), and many others (refer to e.g. Linhart & Zucchini, 1986; Miller, 1990; Chatfield, 1995; Zucchini, 2000; Vehtari & Ojanen, 2012; Gelman *et al*., 2014; Piironen & Vehtari, 2017). About (i) and (ii) Piironen & 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. However, the solution they provide to this problem (detailed in Piironen & Vehtari, 2017, p.775) cannot be easily implemented in many applications since it assumes that the true data generating model

can be conceptualized in some way (M-completed view), which is often an unrealistic assumption. This is the case, for instance, in the presence of large datasets with a complex dependence structure or when the search for good models is pragmatically, rather than theoretically motivated, as in disclosure risk estimation. So, we are left with traditional remedies against the selection-induced bias: restricting selection to a *small number of well-considered models* (this includes *regularization* and/or *early stopping*), or, alternatively, *model averaging*(Cawley & Talbot, 2010; Zucchini, 2000). In the next section we will argue that all of these remedies are, in various ways, implemented in our proposal, under the assumption that a good model is just a convenient proxy of the true model, neither included in the search space nor conceptualized in any way (M-open view). Once one accepts this assumption, '[…]the focus immediately shifts to identifying which aspects of the model performance are most important to the end user' (Underhill & Smith, 2016, p. 1006), paving the way for a context-dependent utility based approach to model selection. In our application, the context will be disclosure risk estimation, and we define a specific utility function for this particular target. However, this is not the only possible applied context; other possible applications are discussed in the final section.

3 New Model Selection Method

Our proposal is motivated by two ideas. First, standard tools for assessing and selecting models are not very successful in deciding whether the disclosure risk measures will be well estimated because of the peculiar structure of the problem at hand. In Equations 1 and 2 interest is restricted to a specific subset of the contingency table (the sample unique cells). It seems therefore inappropriate to assess whichever model by exploring its performance across all cells of the table. Second, we aim to leverage the potential of the family (3) to make model selection a feasible task, manageable with reasonably simple tools, given that, conversely, it is a daunting challenge in a fully parametric framework as in Skinner & Shlomo (2008) or Skinner & Holmes (1998) and Elamir & Skinner (2006). Driven by these motivations, we develop an application-specific method for selecting models within the family (3), which is also deliberately intended to limit the selection-induced bias, thus filling a gap in the literature. The proposed method relies on a suitable combination of two different criteria, and consists of two distinct yet complementary stages. One identifies a path of search, that is which semi-parametric models have to be evaluated and in which order. The other assesses the candidates and selects an optimal model through a measure of model predictive accuracy specifically tailored to the target, namely global disclosure risk estimation. The detail of the procedure is as follows.

(1) Building on findings in Carota *et al*. (2015), we take the semi-parametric independence model, shown to be a sort of 'default' model in that paper, as the starting model. We denote it by the shorthand NP + I, to emphasize both the nonparametric (NP) nature of the random effects $\boldsymbol{\phi}$ and the structure (I) of its parametric component, that is, the fixed effects $\boldsymbol{\beta} \sim N(\boldsymbol{0}, I\sigma^2)$. Hereafter σ^2 is assumed to be large, so as to express vague prior information. At this stage we focus on the parametric component of the model, and do a preliminary stepwise search in the space of graphical decomposable log-linear models without random effects. This prevents the unidentifiability issues recalled in Sections 1 and 2.1. Starting from the independence model (I), we repeatedly use a penalized log-likelihood with large penalty factor γ. We gradually move γ down on a grid selecting, at each step, the interaction terms that maximize

$$C_0(\gamma) = \sum_{k=1}^{K} \log(p(f_k|\hat{\boldsymbol{\beta}}_{ML})) - d \times \gamma, \tag{4}$$

where

$$p(f_k|\hat{\boldsymbol{\beta}}_{ML}) = \frac{\pi^{f_k}}{f_k!} e^{f_k \mathbf{w}_k' \hat{\boldsymbol{\beta}}_{ML}} e^{-\pi e^{\mathbf{w}_k' \hat{\boldsymbol{\beta}}_{ML}}},$$

$\hat{\boldsymbol{\beta}}_{ML}$ is the vector of maximum likelihood (ML) estimates of fixed effects, d is the difference between the number of parameters estimated under the current model and under the independence model, and γ controls the strength of the penalty. Being σ^2 large, at this stage $\hat{\boldsymbol{\beta}}_{ML}$ is used as an approximation to Bayesian estimates of fixed effects. Although $C_0(\gamma)$ bears some resemblance to the Akaike information criterion, it is indeed a 'sieve' criterion, devoted to drastically limit the number of interactions terms introduced in the model. As a starting value of γ, we select the penalty that allows to add the first interaction term to the independence model I; similarly, all subsequent steps are aimed at including only those interaction terms that satisfy the severe constraint of simplicity imposed through γ at each step. This results in a set of alternative, increasingly complex, parametric components (one for each step). Finally, we add DP random effects to each parametric component (set of fixed effects) identified at this stage of the procedure, thereby identifying a small number of candidate models in the family (3).

(2) Each semi-parametric candidate identified at stage (1) is then evaluated through the application-specific criterion C_1,

$$C_1 = \sum_{k=1}^{K} I(f_k = 1) \times \log\left(\int p(f_k|\lambda_k) p(\lambda_k|f_1, .., f_k) d\lambda_k \right), \tag{5}$$

where

$$p(f_k|\lambda_k) = \frac{1}{f_k!} (\pi\lambda_k)^{f_k} e^{-(\pi\lambda_k)}$$

and λ_k is defined as in Equation 3. This is the *log pointwise predictive density* (lppd, refer to Gelman *et al*., 2014, p. 1000) *restricted to the unique cells*, namely those crucial for estimating the global risks (1) and (2). C_1 is a Bayesian measure of the model's predictive accuracy, or performance. Rather than conditioning on a point estimate as at stage (1), it averages over the full posterior distribution and is computed using posterior simulations $\lambda^{(h)}$:

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

where H denotes the number of simulation draws necessary to fully capture the posterior of λ (see the Appendix A for implementation details). The higher C_1, the better the model.

Restricting model assessment to the sample uniques implies that we make the judgement on model performance *relative* to this subset of cells for each semi-parametric candidate in the path of search identified by means of C_0. As such, this does not necessarily imply a good fit on those cells, simply a relatively better fit. Following Underhill & Smith (2016), and their approach to utility based model selection, an alternative presentation of C_1 is provided next. In estimating the disclosure risk measures (1) and (2) good model performance over the sample uniques is the sole concern to statistical agencies. In this application, use of logarithmic scores restricted

to that subset of cells is therefore a very reasonable description of the end user's utility and model selection can proceed on the basis of the highest scoring model. In this sense, C_1 is a context-dependent score based Bayesian information criterion. However, unlike the context-dependent score-based Bayesian information criteria presented in Underhill & Smith (2016) and unlike standard criteria assessing the model's performance across the full joint distribution, C_1 does not include a correction for the within-sample-error. To prevent a negative impact on the selection-induced bias, we intentionally omit such estimated bias correction term, as it would introduce additional variability in the criterion. Refer to Zucchini (2000, p. 53) for general comments on the trade-off between (i) and (ii) and the advantages in terms of (ii) of avoiding a data-based correction for bias. This choice is particularly appropriate in our context where the within-sample-error is low. Indeed, here the number of sample uniques (hereafter denoted by U) is much smaller than the size of the data used to fit the model (the number of cells K, or the number of cells K minus the number of structural zeroes). Quoting Underhill and Smith (2016, p. 1027), the lack of such bias correction term can be interpreted as an extremization of the general benefit represented by the lower correction applicable when a criterion is 'based on the relevant marginal and conditional logarithmic scores of the variables of interest within a larger model'.

Let us now turn to $C_0(\gamma)$, the sieve criterion used at the first stage of the procedure to select suitable semi-parametric candidates. This intermediate, instrumental, criterion is based on a double use of the data (all K sample frequencies). Nevertheless, we do not correct it for the within-sample-error. Instead, we introduce in C_0 a heavy penalty for complexity, $d \times \gamma$, due to the large values of γ that we employ in the search. This choice indirectly implies a strongly non-uniform prior on models of the family (3), that is, a form of regularization by virtue of which limiting the selection-induced bias through a careful restriction of candidate models (refer to Section 2.2) reduces to identify the ones having the simplest parametric specification. With regard to the number of candidates to be considered, $C_0(\gamma)$ is not endowed with a stopping rule because it is used jointly with the criterion C_1. The semi-parametric models of increasing complexity selected by means of C_0 are assessed by C_1, and the search stops when C_1 begins to decline. A sound guarantee that a good model will be reached very quickly in the space of extremely simple candidates identified through C_0 is provided by a special feature of models belonging to the family (3). All of them are weighted averages of a huge collection of parametric models, which makes them extremely flexible, as illustrated next. In particular, we aim at clarifying why the complexity of good semi-parametric log-linear mixed models is scarcely sensitive to the size of the table under consideration, which is a peculiar and appealing feature of our model selection procedure.

That each candidate in the search space is an average model can be seen by writing the corresponding likelihood, $L(\boldsymbol{\beta}, m|f_1, \ldots, f_K, G_0)$ (refer to, e.g. Lo, 1984; Liu, 1996), as follows:

$$\sum_{c=1}^{K} \sum_{C:|C|=c} \left[\frac{\Gamma(m)\, m^c}{\Gamma(m+K)} \prod_{j=1}^{c} \Gamma(n_j) \right] \times \left[\prod_{j=1}^{c} \int \prod_{k \in \text{cluster } j} \frac{\pi^{f_k}}{f_k!} e^{f_k(\mathbf{w}_k'\boldsymbol{\beta} + \phi_j)} e^{-\pi e^{(\mathbf{w}_k'\boldsymbol{\beta} + \phi_j)}} dG_0(\phi_j) \right], \tag{6}$$

where each summand reads as the product of the two factors in square brackets. The first factor is the probability assigned to a given partition C of the K sample frequencies in c non-empty clusters by the multivariate Ewens distribution (Johnson *et al*., 2004, chap. 41): we denote it by $\Pr\{n_1, \ldots, n_c|m, C, c\}$ where n_j is the number of cells in the cluster j. The second factor is the likelihood corresponding to a log-linear model with the same fixed effects and a G_0-distributed random effect specific to each cluster j belonging to that partition, hereafter denoted by $L(\boldsymbol{\beta}, m|f_1, \ldots, f_K, G_0, C, c)$. Because the sum in (6) is over all possible partitions

C in c non-empty clusters (C: $|C|=c$) with $c = 1, .., K$, the total number of summands is equal to B_K, the Bell number. Hence, we can rewrite (6) as

$$L(\boldsymbol{\beta}, m|f_1, \ldots, f_K, G_0) = \sum^{B_K} \Pr\{n_1, \ldots, n_c|m, C, c\}\, L(\boldsymbol{\beta}, m|f_1, \ldots, f_K, G_0, C, c), \qquad (7)$$

showing that the likelihood corresponding to a given semi-parametric candidate is an average of B_K *parametric likelihoods* $L(\boldsymbol{\beta}, m|f_1, \ldots, f_K, G_0, C, c)$ according to specific weights on random partitions of $f_1, \ldots, f_K$. This implies various, useful consequences. Of special interest for model selection is that, as K increases, the stimulus received by the mechanism of model averaging in (7) is extraordinarily strong, because the number of terms summed in the likelihood increases as follows:

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

In practice, even with extremely large tables, this massive model averaging, implied by the presence of DP random effects, strongly limits the need for additional interaction terms to obtain 'reasonably good' models. More precisely, the number of models averaged in (7) grows so much with K to nearly compensate for the simultaneous worsening of risk overestimation due to the poorness of fixed effects, as explained by Skinner & Shlomo (2008). This is the reason why, when describing the proposed method, we claim that the criterion C_1 is used to evaluate a *small* set of semi-parametric candidates, and also why we can invariably start the search from the semi-parametric independence model, NP + I. In contrast, under the approach of Skinner & Shlomo (2008), the complexity of the optimal model increases remarkably with K (refer to p. 999, tables 5–7). There we can also observe an evolution of the starting model from the independence (I) to the all two-way interactions (II) model. Manrique-Vallier & Reiter (2012, supporting information) select the best log-linear model according to the criterion $\hat{B}$ of Skinner & Shlomo (2008): their results over samples of 5000 and 10 000 individuals confirm that the complexity of the starting as well as the selected model strongly depends on the table size.

The extreme flexibility of each model belonging to the family (3), originating from the fitting ability of B_K parametric log-linear mixed models and expressed by formula (7), is also the basic reason why the combination of two simple and quite rough criteria like C_0 and C_1 succeeds in selecting a good model. As a matter of fact, under the family (3), the challenging problem of selecting a model which leads to good estimates of the F_ks on cells where $f_k = 1$ (i.e. good predictions of $(F_k - 1)$ on such cells) reduces to the much easier selection of few interaction terms useful to enrich the parametric likelihoods $L(\boldsymbol{\beta}, m|f_1, \ldots, f_K, G_0, C, c)$ in (7) so as to enhance the above-mentioned correction for overestimation of the risks. This amounts to selecting a small, common to clusters, set of fixed effects βs that effectively supplement the DP random effects in inferring on a specific subset of the whole table. The sieve criterion $C_0(\gamma)$ allows us to identify few alternative sets of βs, thus strongly narrowing the search space. Roughly, it requires a good fit on all cells under a severe constraint of simplicity, sufficient to prevent any form of overfitting throughout the procedure. The goal of selecting the most appropriate interactions terms among the candidates identified by C_0 is then achieved through C_1, which focuses on unique cells. This makes the judgement on the performance of the semi-parametric candidates *relative* to that subset of cells, in the spirit of Underhill & Smith (2016). Both criteria are simply log-likelihoods adapted to achieve the above-mentioned goals, whose combined use within the family (3) implements an original mix of established remedies against the selection-induced bias. Traditionally, instead, these are often alternative remedies to each other. In Section 4,

we test the predictive ability of models selected by the pair (C_0, C_1) on different samples arising from very different settings. In all cases, an extremely small number of nonparametric candidates need to be evaluated to reach a 'reasonably good' model within the family (3). This makes the first-stage restriction to graphical decomposable models irrelevant in practice.

4 Illustration of the Proposed Method

We illustrate and test our model selection procedure in a range of contingency tables differing in size, reference population and spanning variables obtained from different sources, as detailed next. We consider two large datasets, taken here as reference pseudo-populations, from which we draw simple random samples with fraction $\pi = 0.05$. Because the pseudo-populations are known, we can compute the true risk measures and assess the performance of the proposed method. The first pseudo-population is the set of $N = 1\,150\,934$ individuals aged 21 and over in the 5% public use microdata sample of the US 2000 census for the state of California (IPUMS, Ruggles *et al*., 2017). The second consists of the $N = 794\,986$ individuals recorded in the 7% public use microdata sample of the Italian National Social Security Administration, 2004 (source: Work Histories Italian Panel, WHIP). For the California pseudo-population, we first reconsider the same table of 3 600 000 cells used in Manrique-Vallier & Reiter (2012), obtained by cross-classifying the ten key variables listed in Table 1 (left panel). 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. Because, 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, obtained from the WHIP pseudo-population by cross-classification of the eight key variables listed in the right panel of Table 1.

For each of the four contingency tables just described, we consider a set of semi-parametric models labelled SP_x, $x = a, b, ..,$ listed in Table 2 (first column). Such models have been selected by the criterion $C_0(\gamma)$, except for two models, SP_d and SP_e in the WHIP table, introduced in order to test the impact of a richer fixed effects specification on models' assessment.

For comparison, we also include the parametric counterparts of some of the selected semi-parametric models SP_x; these can be obtained from (3) as $m \to \infty$ and are labelled P_x. They are Bayesian log-linear models with the same fixed effects and a parametric (Gamma distributed, as detailed in the last column) random effect specific to each cell k, for $k = 1, .., K$. Conditionally on the random effects described in the second column of Table 2 (either nonparametric, NP, or parametric, P), the models we are considering only differ for the

Table 1. *Key variables under consideration (number of categories in parentheses) and their labels in the California (left) and WHIP—Work Histories Italian Panel—(right) data*

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)
INCR	Income (10)	FSIZE	Firm size (5)
DISAB	Disability (2)		
VETST	Veteran status (2)		

Table 2. *Log-linear models (in order of increasing complexity) for California and WHIP tables: model label, type of random effects, structure of fixed effects (parametric component), number of additional parameters compared to those included in the independence model, and prior on the random effects*

Label	*Random effects*	*Shorthand for fixed effects*	*Number of extra parameters*	*Prior for random effects*
California large				
SP_a	NP	I + SEX * VETST	1	DP(m, $Ga(a, b)$)
SP_b	NP	I + EMPST * INCR	18	DP(m, $Ga(a, b)$)
SP_c	NP	I+ SEX * VETST + EMPST * INCR	19	DP(m, $Ga(a, b)$)
P_a	P	I + SEX * VETST	1	$Ga(a, b)$
P_b	P	I + EMPST * INCR	18	$Ga(a, b)$
P_c	P	I+ SEX * VETST + EMPST * INCR	19	$Ga(a, b)$
California medium				
SP_a	NP	I + EMPST * SEX	3	DP(m, $Ga(a, b)$)
SP_b	NP	I + EMPST * INCR	18	DP(m, $Ga(a, b)$)
SP_c	NP	I + EMPST * SEX + EMPST * INCR	21	DP(m, $Ga(a, b)$)
NP + I	NP	I	—	DP(m, $Ga(a, b)$)
P_a	P	I + EMPST * SEX	3	$Ga(a, b)$
P_b	P	I + EMPST * INCR	18	$Ga(a, b)$
P_c	P	I + EMPST * SEX + EMPST * INCR	21	$Ga(a, b)$
P+I	P	I	—	$Ga(a, b)$
California small				
SP_a	NP	I + SEX * VETST	1	DP(m, $Ga(a, b)$)
SP_b	NP	I + MARST * RACE	20	DP(m, $Ga(a, b)$)
NP+I	NP	I	—	DP(m, $Ga(a, b)$)
P_a	P	I + SEX * VETST	1	$Ga(a, b)$)
P_b	P	I + MARST * RACE	20	$Ga(a, b)$)
P+I	P	I	—	$Ga(a, b)$
WHIP				
SP_a	NP	I + ESEC * WORKP	9	DP(m, $Ga(a, b)$)
SP_b	NP	I + ESEC * FSIZE	12	DP(m, $Ga(a, b)$)
SP_c	NP	I + ESEC * WORKP + ESEC * FSIZE	21	DP(m, $Ga(a, b)$)
SP_d	NP	I + ESEC * WORKP + ESEC * SEX + ESEC * WAGF + ESEC * FSIZE	27	DP(m, $Ga(a, b)$)
SP_e	NP	I + ESEC * WORKP + ESEC * SEX + ESEC * WAGF + AGE * WORKP	48	DP(m, $Ga(a, b)$)
NP + I	NP	I	—	DP(m, $Ga(a, b)$)
P+I	P	I	—	$Ga(a, b)$

For the last three tables, the starting model, NP + I, and its parametric counterpart, P + I, are also considered.

specification of the vector $\boldsymbol{\beta}$. For this reason, in the third column, we describe the structure of the parametric component. For instance, when the fixed effects comprise the main effects of all key variables and, in addition, the two-way interaction parameters between all levels of key variables SEX and WETST, we use the shorthand I + SEX * WETST.

Model complexity can be summarized by the number d of extra parameters implied by the interaction terms (e.g. SEX * WETST) added to the independence model; this is reported in the second last column of Table 2.

The last column shows the prior on the random effects. Throughout in this section, we reparametrize the random effects so that $\omega_k = e^{\phi_k}$ is drawn from a Dirichlet process with Gamma base measure, $Ga(a, b)$, where b denotes the rate parameter. Consequently, all P_x models in Table 2 are members of the family defined in Elamir & Skinner (2006). In practice, the latter is equivalent to the family of log-linear models without random effects of Skinner &

Shlomo (2008) because the corresponding risk estimates are nearly identical (Elamir & Skinner, 2006). Finally, we assume a standard prior for the precision parameter m, say $Ga(e, f)$. The hyperparameters are fixed so as to specify vague priors: we take $a = 1,\ b = 0.1; e = 1,\ f = 0.1$; and $\sigma^2 = 10$.

Values of C_1 for all models in Table 2 are presented in the fourth column of Table 3, with model rankings in parentheses (computational details are provided in the Appendix A). In columns 2 and 3, we also rank models according to the true estimation errors,

$$|\hat{\tau}_i - \tau_i|, \quad i = 1, 2,$$

that is the distance between the Bayesian estimate under the model, $\hat{\tau}_i$, and the true value of the risk τ_i computed by using the pseudo-population. Finally, to illustrate the impact that a data based correction for the within-sample-error (point (i) in Section 2.2) may have on the

Table 3. *Risk estimates under the models listed in Table 2, predictive measures and models' ranks (in brackets) based on the true estimation error* $|\hat{\tau}_i - \tau_i|,\ i = 1, 2$, *on* C_1 *and* WAIC_U

Model	$\hat{\tau}_1$	$\hat{\tau}_2$	C_1	WAIC_U
California large				
U = 11 421				
PU = 44 572	$\tau_1 = 2205$	$\tau_2 = 3949.7$		
SP_a	**2,245.2** (1)	4022.5 (1)	−21 914.5 (1)	−30 608.8 (5)
SP_b	2323.5 (3)	4090.5 (3)	−22 010.4 (2)	−30 676.1 (6)
SP_c	2272.2 (2)	4034.3 (2)	−22 032.9 (3)	−30 199.5 (4)
P_a	2706.6 (5)	4431.9 (5)	−29 745.4 (6)	−29 773.4 (3)
P_b	2727.1 (6)	4458.4 (6)	−29 594.9 (5)	−29 631.1 (2)
P_c	2652.9 (4)	4374.4 (4)	−29 299.7 (4)	−29 336.2 (1)
California medium				
U = 7669				
PU = 24 124	$\tau_1 = 1169$	$\tau_2 = 2314.6$		
SP_a	**1185.4** (1)	**2340.6** (1)	−13 624.5 (1)	−18 215.9 (4)
SP_b	1223.0 (3)	2371.8 (3)	−13 805.9 (3)	−18 133.6 (2)
SP_c	1225.0 (4)	2373.4 (4)	−13 812.5 (4)	−18 098.1 (1)
NP + I	**1189.2** (2)	**2345.3** (2)	−13 632.6 (2)	−18 195.9 (3)
P_a	1424.8 (8)	2523.6 (8)	−18 706.7 (8)	−18 734.6 (8)
P_b	1386.2 (5)	2487.3 (5)	−18 352.6 (5)	−18 387.9 (5)
P_c	1399.4 (6)	2500.9 (6)	−18 364.4 (6)	−18 399.1 (6)
P+I	1415.2 (7)	2511.7 (7)	−18 644.3 (7)	−18 670.4 (7)
California Small				
U = 3575				
PU = 10 355	$\tau_1 = 498$	$\tau_2 = 1023.4$		
SP_a	**479.8** (3)	**1003.2** (3)	−6212.9 (3)	−7886.3 (1)
SP_b	**483.8** (1)	**1011.3** (1)	−6183.8 (2)	−7993.0 (3)
NP + I	**480.3** (2)	**1008.6** (2)	−6175.9 (1)	−7982.6 (2)
P_a	581.6 (6)	1072.9 (5)	−8951.8 (4)	−8972.9 (4)
P_b	568.5 (4)	1065.0 (4)	−9023.4 (5)	−9055.2 (5)
P+I	579.9 (5)	1077.6 (6)	−9109.1 (6)	−9131.4 (6)
WHIP				
U = 7176				
PU = 17 630	$\tau_1 = 915$	$\tau_2 = 1948.1$		
SP_a	**917.9** (1)	**1981.2** (3)	−12 022.0 (2)	−16 107.4 (5)
SP_b	1003.1 (5)	2078.4 (5)	−12 261.6 (5)	−16 413.3 (7)
SP_c	**921.2** (3)	**1987.0** (4)	−12 128.4 (3)	−15 977.5 (3)
SP_d	**908.9** (2)	**1972.2** (2)	−12 134.7 (4)	−15 767.7 (2)
SP_e	**874.8** (4)	**1930.2** (1)	−12 010.1 (1)	−16 084.5 (4)
NP + I	1010.4 (6)	2083.4 (6)	−12 149.9 (5)	−16 195.7 (6)
P+I	1184.9 (7)	2289.9 (7)	−15 633.6 (7)	−15 650.3 (1)

For each contingency table, we report the number of sample uniques (U), the number of population uniques (PU) and the true values of τ_1 and τ_2.

variability of C_1, we rank models according to the Widely Applicable Information Criterion (WAIC, Watanabe, 2009) restricted to the sample uniques,

$$\mathrm{WAIC_U} = \sum_{k=1}^{K} I(f_k = 1) \times \left[\log\left(\int p(f_k|\lambda_k) p(\lambda_k|f_1, .., f_k) d\lambda_k \right) - var_{post}(\log p(f_k|\lambda)) \right].$$

$\mathrm{WAIC_U}$ is obtained by introducing the indicator function $I(f_k = 1)$ in the WAIC expression (in order to focus exclusively on the U sample uniques), and it is nothing but C_1 plus an estimated bias correction term analogous to p_{WAIC2} (refer to formula 11 Gelman *et al*., 2014, p. 1002), usually referred to as the effective number of parameters. The latter is computed considering the posterior variance of the log-predictive density for each data point f_k, that is, $V_{h=1}^{H} \log p(f_k|\lambda^h)$, where $V_{h=1}^{H}$ represents the sample variance $V_{h=1}^{H} a_h = \frac{1}{H-1} \sum_{h=1}^{H} (a_h - \overline{a})^2$.

Although the P_x models are not candidates (but just parametric counterparts of the semi-parametric candidates selected at the first stage of the procedure), they are included in the rankings presented in Table 3 to show that these largely underfitting models—leading to systematic overestimation of the global risks, as explained by Skinner & Shlomo (2008)—are preferred by the $\mathrm{WAIC_U}$ in two contingency tables (Large California and WHIP). This is the empirical evidence of substantial selection-induced bias and optimism in model performance evaluation due to the increase of the variance of the criterion because of an additional estimated term in C_1. Quoting Zucchini (2000), these are two examples of damage, rather than benefit, that may result from the attempt to correct for the within-sample-error. Such an attempt is particularly inappropriate in our problem: indeed C_1 is based solely on the U sample uniques and $\mathrm{U}/K \leq 0.001$ for all California tables, $\mathrm{U}/(K - \text{structuralzeroes}) \leq 0.0169$ for the WHIP table.

The good performance of C_1 in all contingency tables can be appreciated by benchmarking the corresponding models' rankings against the ones based on the true estimation error (second and third columns). First, the semi-parametric models, SP_x, are always preferred to their parametric counterparts, P_x, which will be ignored from now on. Second, the three criteria largely agree in ranking the SP_x models for all California tables. Importantly, moreover, in the rare cases where they disagree (second and third positions in the large California table; second and first positions in the small California table) the values of C_1 are so close to each other that we are actually warned about possible inversions of the corresponding positions in both rankings. Thirdly, in the awkward situation of disagreement between the two rankings based on the true estimation error of τ_1 and τ_2 (WHIP table), the ranking based on C_1 proves to be a very reasonable compromise between such two 'true' rankings. Put together, all these points show that the criterion C_1 is able to rank competing models in an approximately correct order, with a variability not dangerously inflated by an estimated bias correction term.

In addition to the previous points, in all contingency tables we observe that the two-stage procedure based on the pair (C_0, C_1) is able to identify a good model after very few steps in the search space. While this is not surprising in the small and medium California tables where NP + I confirms to be a good default model (in both tables there are no cogent reasons for additional interaction terms), this is an important result in the WHIP table where NP + I is just a starting model, inadequate for risk estimation, and in the large California table where it is even more inadequate (result not reported). In the latter table, 95% and 99% credible intervals under the model labelled SP_a include the true values of τ_1 and τ_2, respectively. As regards the WHIP table, we have four models under which 95% credible intervals include the true risks: SP_a, SP_c, SP_d and SP_e. Recall that SP_d and SP_e have been introduced in the evaluation (refer to Table 2) just to discuss the impact of enriching the fixed effects specification. Notably, they turn out to be

good models essentially because of the presence of a two-way interaction (ESEC * WORKP) already selected by C_0 and already included in SP_a and SP_c, two of the candidates identified at the first stage of our procedure. We stress that, among the latter models, C_1 prefers SP_a. Indeed, in a series of additional tests (not reported here for brevity), we observed that often a single two-way interaction is enough to enter the range of 'reasonably good' semi-parametric models. This leads us to conclude that, under the family (3), good risk estimates can be achieved by a very slight and easily identifiable adjustment of the parametric component in the starting model. This also implies that the restriction to decomposable graphical models imposed at stage (1) turns out to be a mere formality, irrelevant in practice.

5 Final Comments

We presented a new Bayesian method to select a good model for disclosure risk estimation in the family of log-linear mixed models with DP random effects.

In the literature, there is a lack of alternative methods for selecting log-linear models for disclosure risk estimation. The only available method (Skinner & Shlomo, 2008) has a number of severe issues essentially resulting from unidentifiability of fixed effects. More generally, outside the class of decomposable graphical models, such issues arise whenever the underlying contingency tables are sparse. These issues, the lack of genuine observed covariates and the peculiar nature of the problem at hand deterred us from adopting a formally more sophisticated approach like tackling model selection from within the nonparametric family of covariate-dependent Dirichlet process mixture models (refer to e.g.Barcella *et al*., 2017, and reference therein). Vice versa, all previous reasons solicited the proposal of a context-dependent method for selecting good models within the semi-parametric family of Carota *et al*. (2015). In such models, the DP random effects have a preeminent role in leading to good risk estimates, while the number of parametric fixed effects can be drastically limited. Therefore, we implemented a forward search for a parsimonious model within this family, under a pragmatic approach consonant with the problem at hand. We also recognize and capitalize on the nature of the problem at hand in two further important ways. The risk measures (1) and (2) are nonstandard estimands, because they are not entirely specified before observing the sample. Indeed, they are sums of functions of observable and unobservable variables (f_k and F_K, respectively), exclusively over cells where $f_k = 1$ (refer to Zhang, 2005). Thus, τ_1 and τ_2 clearly highlight a specific subset of values of the observed response f_k as the only subset of interest. In addition, estimating τ_1 and τ_2 amounts to predicting the unobserved quantities $F_k - f_k$ on that subset. As a consequence, on the one hand, we decided to focus on predictive measures of model's performance and work with a 'local' criterion, rather than assessing model fit on the whole contingency table. On the other hand, in a very natural way, we were induced to address the often neglected issue of selection-induced bias, while keeping down the within-sample-error.

In a nutshell, our proposal works as follows. Building on the great flexibility of the family of models under consideration, we defined two criteria, C_0 and C_1, to be jointly used in the forward search. C_0 identifies a small number of very simple candidate models (thereby also reducing the computational effort required in model selection) and the order in which they have to be evaluated. C_1 measures the predictive accuracy of the candidates on specific cells. According to Underhill & Smith (2016) and their utility-based approach to model selection, C_1 is a context-dependent scoring rule, and a good model is just a convenient proxy of the true model, neither included in the search space nor conceptualized in any way (M-open view). However, unlike Underhill & Smith (2016), we adapt our context-dependent scoring rule to face the selection-induced bias (ii), rather than the within sample error (i). Typically, (i) is addressed by including an estimated bias correction term in the criterion, with the drawback of exposing

it to (ii) because of an increase of its variance. Taking advantage of a natural reduction of the within-sample-error (the data are not used twice, because only a small subset of them—the U sample uniques—is re-used to evaluate C_1), we deliberately refrain from adding a data based correction for bias to C_1, as this would exacerbate the selection-induced bias, as shown in Section 4.

The applications investigated in this paper reveal that the proposed procedure is generally able to rank competing models in terms of their ability to produce good estimates of global disclosure risks in rather different settings and in a relatively simple and fast way.

Outside the application discussed in the paper, the same method can be directly adapted to different inferential problems involving datasets with a similar structure, but arising in completely different fields of research. Consider, for instance, the examples provided next. (a) In a given area, the observed number of animals in a species is extremely small in the presence of certain cultivation methods (such as specific combinations of type of seed, type and dose of fertilizer, and type and dose of herbicide). Ultimately, said cultivation methods are deemed dangerous for the survival of the species if the estimate of the number of animals living in the whole territory cultivated with those specific combinations falls below a given threshold. (b) In a clinical trial, a drug turns out to be effective only on subjects who have certain characteristics rare in the sample (e.g. genetic traits). The trial stops if said characteristics in the population of subjects affected by the disease are not frequent enough to make the production of the drug economically sustainable. (c) A sample survey aimed at studying innovation in industry shows that the top management of the most innovative companies, all other things being equal, includes subjects with degrees in Humanities, or of female gender, or of a different ethnicity than the one prevailing in the country where the company is based. This raises the question of whether uncommon cultural backgrounds or sensitivities are, together with the other things (values of other variables such as sector of activity, export-orientation, and so on), a stimulus for innovation. Testing this conjecture requires estimating the number of units in the population that exhibit certain specific characteristics (combinations of values) highlighted by the sampled units.

Indeed, all previous examples share with disclosure risk estimation the same structure: given certain combinations of values with small frequencies in the sample, we have to estimate how large the corresponding population frequencies are. This, in turn, implies that the observed sample contributes to define the estimand, which is not entirely specified before observation.

In all previous cases, an application-specific model selection criterion can be reasonably defined on the subset of cells indicated by the observed sample as ‘cells of interest’, and appropriately bent to face (ii) rather than (i). In order to achieve the latter goal, omitting an estimated bias correction term, as we have done in the illustrative application to disclosure risk estimation, is only apparently an extreme choice suitable only for extreme cases like the four contingency tables presented in Section 4 (where U is invariably negligible compared to K). Indeed, Zucchini (2000) discusses the advantages of such omission, in terms of (ii) versus (i), in a general criterion, that is, a criterion concerned with model’s performance across the full joint distribution (all K cells in our case). Clearly, for any application-specific criterion the cost of omitting an estimated bias correction term is lower (Underhill & Smith, 2016, p.1027). For the same reasons our method can be easily adapted to deal with different disclosure risk measures based not only on sample uniques, but, more generally, on the whole subset of cells with small sample frequencies (sample doubles, triples, and so on). All these facts and the great flexibility of our family of semi-parametric models, ensuring that a good model can be obtained by a slight and easily identifiable enrichment of the starting model, contribute to guaranteeing a wide generalizability and adaptability of the proposed model selection procedure to different circumstances.

Notes

[1]For this reason, Carota *et al*. (2015) refer to their approach as BNP log-linear modelling, though it is a semi-parametric approach.

References

Abowd, J.M. (2018a). Protecting the confidentiality of america's statistics: adopting modern disclosure avoidance methods at the census bureau. Census Blogs: Research Matters. Retrieved from Census Blogs: Research Matters. Available athttps://www.census.gov/newsroom/blogs/research-matters/2018/08/protecting_the_confi.html

Abowd, J.M. (2018b). The US Census Bureau adopts differential privacy. In KDD'18 proceedings of the 24th ACM SIGKDD international conference on knowledge discovery and data mining, pp. 2867–2867. London. https://doi.org/10.1145/3219819.3226070

Barcella, W., De Iorio, M. & Baio, G. (2017). A comparative review of variable selection techniques for covariate dependent Dirichlet process mixture models. *Canad. J. Statist.*, **45**, 254–273.

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

Carlson, M. (2002). Assessing microdata disclosure risk using the Poisson-inverse Gaussian distribution. *Stat. Transit.*, **5**, 901–925.

Carota, C., Filippone, M., Leombruni, R. & Polettini, S. (2015). Bayesian nonparametric disclosure risk estimation via mixed effects log-linear models. *Ann. Appl. Stat.*, **9**, 525–46.

Cawley, G.C. & Talbot, N.L.C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.*, **11**, 2079–2107.

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *J. Roy. Stat. Soc. A Sta.*, **158**, 419–46.

Duane, S., Kennedy, A.D., Pendleton, B.J. & Roweth, D. (1987). Hybrid Monte Carlo. *Phys. Lett. B*, **195**, 216–222.

Elamir, E.A.H. & Skinner, C.J. (2006). Record level measures of disclosure risk for survey microdata. *J. Off. Stat.*, **22**, 525–539.

Escobar, M.D. & West, M. (1994). Bayesian density estimation and inference using mixtures. *J. Am. Stat. Assoc.*, **90**, 577–588.

Eurostat (2017). Item 2.4—differential privacy. 9th meeting of the Expert Group on statistical disclosure control (EG SDC), Eurostat, 14 November 2017.

Eurostat (2018). Item 5—ideas for future projects. 10th meeting of the Expert Group on statistical disclosure control (EG SDC), Eurostat, 14 November 2018.

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

Fienberg, S.E. & Rinaldo, A. (2012). Maximum likelihood estimation in log-linear models. *Ann. Stat.*, **40**, 996–1023.

Forster, J.J. & Webb, E.L. (2007). Bayesian disclosure risk assessment: predicting small frequencies in contingency tables. *J. Roy. Statist. Soc. Ser. C.*, **56**, 551–570.

Gelman, A., Hwang, J. & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Stat. Comput.*, **6**, 997–1016.

Girolami, M. & Calderhead, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. Roy. Stat. Soc. B Met.*, **73**, 123–214.

Johnson, N.L., Kotz, S. & Balakrishnan, N. (2004). *Discrete multivariate distributions*. John Wiley and Sons: New York.

Linhart, H. & Zucchini, W. (1986). *Model Selection*. Wiley: New York.

Liu, J.S. (1996). Nonparametric hierarchical Bayes via sequential imputations. *Ann. Stat.*, **24**, 911–930.

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

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

McCulloch, C.E. & Neuhaus, J.M. (2011). Misspecifying the shape of a random effects distribution: why getting it wrong may not matter. *Stat. Sci.*, **26**, 388–402.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. & Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.*, **21**, 1087–1092.

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

Muller, P., Quintana, F.A., Jara, A. & Hanson, T. (2015). *Bayesian Nonparametric Data Analysis*. Springer: New York. https://doi.org/10.1007/978-3-319-18968-0

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.

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

Piironen, J. & Vehtari, A. (2017). Comparison of Bayesian predictive methods for model selection. *Stat. Comput.*, **27**, 711–735.
Rinott, Y., O'Keefe, C.M., Shlomo, N. & Skinner, C. (2018). Confidentiality and differential privacy in the dissemination of frequency tables. *Stat. Sci.*, **33**, 358–385.
Roberts, G.O. & Rosenthal, J.S. (2009). Examples of adaptive MCMC. *J. Comput. Graph. Stat.*, **18**, 349–367.
Ruggles, S., Fitch, C., Magnuson, D. & Schroeder, J. (2019). Differential privacy and census data: implications for social and economic research. *AEA Papers and Proceedings 2019*, **109**, 1–7.
Ruggles, S., Genadek, K., Goeken, R., Grover, J. & Sobek, M. (2017). Integrated public use microdata series: version 7.0 [dataset].
Shlomo, N., Antal, L. & Elliot, M. (2015). Measuring disclosure risk and data utility for flexible table generators. *J. Off. Stat.*, **31**, 305–324.
Skinner, C.J. & Holmes, D.J. (1998). Estimating the re-identification risk per record in microdata. *J. Off. Stat.*, **14**, 361–372.
Skinner, C. & Shlomo, N. (2008). Assessing identification risk in survey microdata using log-linear models. *J. Am. Stat. Assoc.*, **103**, 989–1001.
Taylor, L., Zhou, X.H. & Rise, P. (2018). A tutorial in assessing disclosure risk in microdata. *Stat. Med.*, **37**, 3693–3706.
Underhill, N.T. & Smith, J.Q. (2016). Context-Dependent Score Based Bayesian Information Criteria. *Bayesian Anal.*, **11**, 1005–1033.
Vehtari, A. & Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Stat. Surv.*, **6**, 142–228.
Watanabe, S. (2009). *Algebraic Geometry and Statistical Learning Theory.* Cambridge University Press: Cambridge.
Zhang, C.-H. (2005). Estimation of sums of random variables: examples and information bounds. *Ann. Statist.*, **33**, 2022–2041.
Zucchini, W. (2000). An introduction to model selection. *J. Math. Psychol.*, **44**, 41–61.

APPENDIX A

A1 Implementation of the Markov chain Monte Carlo 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 $\boldsymbol{\beta}|\text{rest}$, $\boldsymbol{\phi}|\text{rest}$, and $m|\text{rest}$. Samples from the posterior distribution over $\boldsymbol{\beta}$, $\boldsymbol{\phi}$, and m allows one to estimate per-cell risks through Monte Carlo averaging.

Sampling $\boldsymbol{\beta}$—The conditional distribution of $\boldsymbol{\beta}|\text{rest}$ is not of known form, given that the prior on $\boldsymbol{\beta}$ 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 & Rosenthal, 2009); these can include, for example, 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 & 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 $p(f_1, \ldots, f_K|\boldsymbol{\beta}, \text{rest})$. Defining $M$to be the metric tensor obtained as the Fisher Information of the model plus the negative Hessian of the prior, and ϵ 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 Mas follows $p(\boldsymbol{\beta}'|\boldsymbol{\beta}) = N(\boldsymbol{\beta}'|\boldsymbol{\mu}, \epsilon^2 M^{-1})$, with

$\boldsymbol{\mu} = \boldsymbol{\beta} + \frac{\epsilon^2}{2} M^{-1} \nabla_{\boldsymbol{\beta}} \log[p(f_1, \ldots, f_K | \boldsymbol{\beta}, \text{rest})]$. The complexity of the update is $\mathcal{O}(KD^3)$ where K is the number of cells and D is the size of $\boldsymbol{\beta}$; 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 $\boldsymbol{\beta}$ parameters.

Sampling $\boldsymbol{\phi}$—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 $\boldsymbol{\phi}$. We implemented the MCMC sampler proposed in the review paper of MCMC methods for DP models in Neal (2000) as Algorithm 3. 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 ϕ is given the *IG* distribution. This allows us to integrate out the values of ϕ analytically $\int p(f_k | \boldsymbol{\beta}, \phi) dG_0(\phi)$, where we expressed $p(f_k | \boldsymbol{\beta}, \phi)$ 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 $\mathcal{O}(K)$.

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 & West (1994).

[*Received April 2020; accepted August 2021*]