Disclosure risk estimation via nonparametric log-linear models

Cinzia Carota, Maurizio Filippone, Roberto Leombruni, Silvia Polettini

Abstract A major concern in releasing microdata sets is protecting the privacy of individuals in the sample. Consider a data set in the form of a high-dimensional contingency table. If an individual belongs to a cell with small frequency, an intruder with certain knowledge about the individual may identify him and learn sensitive information about him in the data. To estimate the risk of such breach of confidentiality we introduce several nonparametric models which represent progressive extensions of the one adopted by Skinner and Holmes (1998). The latter is a Poisson model with rates modeled through a mixed effects log-linear model with normal random effects. In the first extension, we assume Dirichlet process random effects and, mimicking Skinner and Holmes (1998), we keep the fixed effects constant. Next, we relax the latter assumption and consider a model all effects of which are unknown. In both extended models the total mass parameter of the Dirichlet process is also unknown. The MCMC methods used for inference are extensively discussed. An application to real data concludes the article.

Key words: bayesian log-linear models, confidentiality, disclosure risk, Dirichlet process, mixed effects models, nonparametric models

Cinzia Carota
Department of Economics Cognetti De Martiis, Univeristà di Torino, via Po, 53, 10123 Torino
e-mail: cinzia.carota@unito.it

Maurizio Filippone
School of Computing Science, University of Glasgow, Sir Alwyn Williams Building, 18 Lilybank Gardens, Glasgow, UK e-mail: Maurizio.Filippone@glasgow.ac.uk

Roberto Leombruni
Department of Economics Cognetti De Martiis, Univeristà di Torino, via Po, 53, 10123 Torino
e-mail: robeto.leombruni@unito.it

Silvia Polettini
Dipartimento di Scienze e Biotecnologie Medico-chirurgiche, Sapienza Univeristà di Roma, Corso della Repubblica, 79 Latina e-mail: silvia.polettini@uniroma1.it

1 Introduction

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

Consider the contingency table representing the cross-classification of individuals by the key variables: if an individual belongs to a cell with small sample frequency, a disclosure may occur. Often, the risk of such a breach of confidentiality is measured by considering only cell frequencies of 1 (sample uniques) in the table of key variables. Note, however, that such low frequencies may arise merely because of sampling and therefore it is not sufficient for a record to be at risk to be a sample unique.

In this article we denote by f_k and F_k respectively the sample and population frequencies in the k-th cell of the contingency table of key variables, and by K the total number of cells. Our goal is to estimate global risks of re-identification, or disclosure risks, defined as [15]

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

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

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

which is the expected number of correct guesses if each sample unique is matched with an individual randomly chosen from the corresponding population cell. Usually, those risks are estimated using parametric models, often based on the Poisson distribution. The work of Bethlehem *et al.* [1] represents the first approach to defining a statistical model for samples where the identifying variables form a contingency table. The model is a hierarchical Poisson-Gamma superpopulation model where $F_k \sim Poisson(\lambda_k)$ and $f_k | F_k \sim binomial(F_k, \pi)$ with known constant sampling probability π. The model was used to deduce (1) and can be seen as an approximation to the Dirichlet-multinomial model analysed by Takemura [16].

A common feature of the models just described is the assumption of exchangeability of cells of the population contingency table, implying that all cells with the same sample frequency are assigned the same risk estimate. Skinner and Holmes [15] and Forster and Webb [4] whithin a Bayesian formulation and Skinner and Shlomo [14] in a frequentist setting introduce a log-linear model for the expected cell frequencies that overcomes this problem. Rinott and Shlomo [12], instead, propose a generalized negative binomial model with local neighbourhood smoothing.

2 Nonparametric Log-linear Models for disclosure risk estimation

As shown in the previous section, many relevant models in the disclosure literature are parametric. In this work, we explore the possibility of dealing with this issue in a Bayesian nonparametric context, extending the model and the estimation method introduced by Skinner and Holmes [15]. We briefly review the work in [15] and then report the proposed nonparametric extensions. Assuming that $F_k \sim Poisson(\lambda_k)$ and $f_k \sim Poisson(\pi\lambda_k)$ independently for $k = 1,..,K$, [15] model the parameters λ_k through a log-linear model with mixed effects:

$$\lambda_k = exp(\mu_k), \quad \mu_k = w_k'\beta + \phi_k, \tag{2}$$

where w_k is a $q \times 1$ design vector depending on the values of the key variables in cell k, β is a $q \times 1$ parameter vector (typically main effects and low-order interactions of the key variables), and ϕ_k is a random effect accounting for cell specific deviations. The sampling fraction π is supposed to be known. Finally, [15] assume that $\phi_k \sim iid\ \mathcal{N}(0, \sigma^2)$. This implies $\lambda_k \sim Lognormal(w_k'\beta, \sigma^2)$, independently for $k = 1,..,K$.

The goal of [15] is to estimate τ_1, whose summands $Pr\{F_k = 1 | f_k = 1\}$, hereafter denoted by $\tau_{1,k}$, are given by $e^{-(1-\pi)\lambda_k}$. Their estimation strategy is as follows:

- preliminary estimates $(\hat{\beta}, \hat{\sigma}^2)$ of β and σ^2 are obtained from the sample frequencies f_k via iterative proportional fitting and by a conditional application of the moment method respectively. Sometimes, however, the value of $\hat{\sigma}^2$ turns out to be negative and, in this case, the authors suggest to use a log-linear model without random effects;
- the pair $(w_k'\beta, \sigma^2)$ is substituted by $(w_k'\hat{\beta}, \hat{\sigma}^2)$ in the *Lognormal* prior;
- different estimates of the per record risk of disclosure $\tau_{1,k}$ are derived:

$$\hat{\tau}_{1,k} = \frac{\int e^{-\lambda_k} e^{-\frac{1}{2\hat{\sigma}^2}(log\lambda_k - w_k'\hat{\beta})^2} d\lambda_k}{\int e^{-\pi\lambda_k} e^{-\frac{1}{2\hat{\sigma}^2}(log\lambda_k - w_k'\hat{\beta})^2} d\lambda_k}, \tag{3}$$

obtained from the posterior of λ_k;

$$\hat{\tau}_{1,k} = e^{-(1-\pi)e^{w_k'\hat{\beta} - \frac{\hat{\sigma}^2}{2}}}, \tag{4}$$

obtained from the prior expected value of λ_k;

$$\hat{\tau}_{1,k} = e^{-(1-\pi)e^{w_k'\hat{\beta}}} \tag{5}$$

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

Equation (3) is an empirical Bayes estimate of $\tau_{1,k}$, equation (4) is a 'simpler' empirical Bayes estimate of $\tau_{1,k}$, and, finally, equation (5) has to be used when the conditional moment method produces negative values of $\hat{\sigma}^2$. In short, this is a two-stage estimation procedure where, in the first stage, the association among cells is exploited to estimate the hyper-parameters of the *Lognormal* prior, while, in the second (and completely separate) stage, the estimates of $\tau_{1,k}$ are obtained cell by cell, independently.

More recently, Skinner and Shlomo [14] resort to a log-linear model without random effects, so that $\tau_{1,k}$ is always estimated by equation (5), and a similar estimate is used for the terms in τ_2, i.e. $E(1/F_k|f_k = 1) = \frac{1}{(1-\pi)\lambda_k}(-e^{(1-\pi)\lambda_k})$.

In our paper, we go back to the model in Skinner and Holmes [15] and, all other things being equal, we assume that the distribution of the random effects, say G, is unknown and a priori distributed according to a Dirichlet process $\mathscr{D}$ with base probability measure G_0 and total mass parameter m [3],

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

Since m controls the variance of the process, in practice G_0 specifies one's 'best guess' about an underlying model of the variation in ϕ, and m specifies the extent to which G_0 holds. We consider two different extensions of the Skinner and Holmes model by introducing different specifications for G_0 and the prior distribution of β. In the first extension, we fix $\beta = \hat{\beta}_{ML}$, where $\hat{\beta}_{ML}$ is the maximum likelihood estimate of the parameter vector, and assume $G_0 = \mathscr{N}(0,\sigma^2)$ with unknown variance σ^2. This extension is directly inspired by both the structure of the model and the estimation strategy in [15]. Therefore, the corresponding risk estimates will be referred to as *nonparametric empirical Bayes* estimates of the risk and represent a generalization of (3). In the second extension we assume that the fixed effects are unknown with a normal prior distribution. To overcome identifiability problems, we follow Li *et al.* [7] and partition the vector β to separate the intercept term, β_0, from main effects and interaction terms, referred to as $\beta_{covariates}$, $\beta = (\beta_0, \beta_{covariates})'$. We assume a reasonably vague Gaussian prior on $\beta_{covariates}$ ($\mathscr{N}(0,10)$) and set $G_0 = \mathscr{N}(\beta_0, \sigma^2)$. In turn, the prior on β_0 is taken to be $\mathscr{N}(0,10)$ and the prior on σ^2 to be invGamma$(1,1)$. Finally, we assume a Gamma$(1,1)$ prior on m.

Under these assumptions, in the more general case, the likelihood turns out to be (see Lo [9] and Liu [8]),

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

where C is a partition of cells $\{1,..,K\}$ in c groups (or clusters), n_j is the number of observations in the j-th cluster, $1 \le n_j \le K$, and finally

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

In the likelihood we observe that the same random effect is assigned to all cells belonging to the same cluster, and that the number of such partitions of the K cells is unknown.

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

$$L(\beta|\underline{f}) = \sum_{c=1}^{K} \sum_{A \in \mathscr{A}_c} \frac{\Gamma(m)}{\Gamma(m+K)} m^c \times$$

$$\prod_{j=1}^{c} \Gamma(n_j) \int \prod_{k=1}^{K} \frac{1}{f_k!} e^{\pi f_k (w_k' \beta + (A\eta)_k)} e^{-e^{\pi (w_k' \beta + (A\eta)_k)}} dG_0(\eta_1, .., \eta_c),$$

where $\mathscr{A}_c$ is the set of all allocation matrices A.

In this work, we adopt a Bayesian treatment to evaluate the re-identification risk. In order to keep the notation uncluttered, let θ denote the set of all parameters in a given model. A Bayesian treatment of the models above described allows to evaluate the terms in the re-identification risks τ_1 and τ_2 as:

$$Pr\{F_k = 1 | f_k = 1, \mathbf{f}\} = \int Pr\{F_k = 1 | f_k = 1, \theta\} p(\theta|\mathbf{f}) d\theta \tag{7}$$

$$E(1/F_k | f_k = 1, \mathbf{f}) = \int E(1/F_k | f_k = 1, \theta) p(\theta|\mathbf{f}) d\theta. \tag{8}$$

Those expressions show how it is possible to evaluate τ_1 and τ_2 by integrating out all model parameters, and the importance of the role played by the posterior distribution $p(\theta|\mathbf{f})$. Given a prior on the parameters $p(\theta)$, we can readily employ Bayes' theorem and obtain the posterior distribution as $p(\theta|\mathbf{f}) = Z^{-1} p(\mathbf{f}|\theta) p(\theta)$, where Z is a proper normalizing constant to ensure that the right hand side is a probability density in the parameter space.

Obtaining the posterior $p(\theta|\mathbf{f})$ analytically is intractable, so we propose to evaluate eq. 7 and eq. 8 by means of Monte Carlo integration [10]. Denote by $\{\theta^{(1)}, \ldots, \theta^{(h)}\}$ a set of h samples from the posterior distribution $p(\theta|\mathbf{f})$. The Monte Carlo estimates of 7 and 8 are simply $\frac{1}{h} \sum_{i=1}^{h} Pr\{F_k = 1 | f_k = 1, \theta^{(h)}\}$, and $\frac{1}{h} \sum_{i=1}^{h} E(1/F_k | f_k = 1, \theta^{(h)})$.

In order to obtain samples from the posterior distribution $p(\theta|\mathbf{f})$, we propose to use Markov chain Monte Carlo (MCMC) techniques [10]. In particular, we propose to use a Gibbs sampler where we sample one group of parameters at a time, namely $\beta|$rest, $\phi|$rest, $m|$rest, $(\beta_0, \sigma^2)|$rest. Convergence of the chains was checked using the Gelman and Rubin's potential scale reduction factor ($\hat{R}$; [5]), by running 10 parallel chains comprising $10,000$ iterations and assessing that chains had converged when $\hat{R} < 1.1$ for all the parameters. According to this criterion, all chains converged within a few thousands of iterations that were then discarded before evaluating the risk scores. The proposed Gibbs sampler steps are briefly discussed next.

Sampling β – Given the form of the Poisson likelihood, it is not possible to sample β using an exact Gibbs step, and so called Metropolis within Gibbs samplers need to be employed [13]. Recent work shows that it is possible to efficiently sample from the posterior distribution of parameters of linear models using the so called *manifold MCMC* methods [6]. Briefly, such samplers exploit the curvature of the log-joint density by constructing a proposal mechanism on the basis of the Fisher Information of the model (see [6] for further details). In this work we adopt Simplified Manifold Metropolis Adjusted Langevin Algorithm (S-MMALA) to sample β, which simulates a diffusion on the statistical manifold. Define M to be the metric tensor obtained as the Fisher Information of the model plus the negative Hessian of the prior, and ε to be a discretization parameter. SM-MALA is essentially a Metropolis-Hastings sampler, with a position dependent proposal akin to the Newton method in optimisation $p(\beta'|\beta) = \mathscr{N}(\beta'|\mu, \varepsilon^2 M^{-1})$, with $\mu = \beta + \frac{\varepsilon^2}{2} M^{-1} \nabla_\beta \log[p(\mathbf{f}|\beta, \mathrm{rest})]$. Gradient and metric tensor can be computed in linear time in the number of cells K and therefore the method scales well to large data sets.

Sampling ϕ – The representation of the random effects through the allocation matrix A allows to apply simple sampling schemes as in [11] to obtain samples from the posterior of the random effects. In this work we adopted Algorithm 5 in [11], as it is easy to implement and as it achieves satisfactory performance in the given application.

Sampling m – In the literature, it has been often reported that inference in models involving Dirichlet Processes is heavily affected by the mass parameter m, and that setting it by means of Maximum Likelihood is bound to yield poor results [8]. Rather than fixing this parameter, we propose to sample from its posterior distribution and to account for uncertainty about it in the evaluation of τ_1 and τ_2. In order to do that, we log-transform m and sample $\psi_m = \log(m)$ instead, using a standard Metropolis-Hastings sampler.

Sampling β_0 and σ^2 – Given that we chose a Gaussian base measure, by imposing a Gaussian prior on the mean β_0 and an inverse gamma prior on the variance σ^2 of the base measure, we can exploit conjugacy in the sampling and obtain the conditional distribution of β_0 and σ^2 in closed form.

3 Results and Discussion

In this section, we compare the proposed models by applying them to a random sample drawn from $N = 450,238$ individuals (source: Work Histories Italian Panel, a linked employer-employee longitudinal database built upon a 1% sample of the National Social Security Administration archives and treated here as the population). From this population, we consider a sampling fraction $\pi = 0.1$ yielding $n = 45,023$, and five key variables (number of categories in parentheses): area (4), sex (2), age (11), ethnicity (5), and economic activity (9), giving $K = 3,960$. Next, we reconsider the same key variables except for age that is grouped in 6 bands, giving $K = 2,160$. Table 1 presents true and estimated values of τ_1 and τ_2 for these two

Table 1 Comparison of estimated values of τ_1 and τ_2 for the two settings analysed ($K = 2,160$; $K = 3,960$). True values of the global risks are $\tau_1 = 18$ and $\tau_2 = 50.1$ in the data set with $K = 2,160$, and $\tau_1 = 39$ and $\tau_2 = 94.4$ in the data set with $K = 3,960$.

			$K = 2,160$		$K = 3,960$	
Model	Intercept	Log-linear model	τ_1	τ_2	τ_1	τ_2
P	Yes	O	0.0	1.0	0.0	3.2
P	Yes	I	20.4	44.5	32.0	76.9
P	Yes	II	21.6	50.2	32.5	84.8
NP Emp NZM	Yes	I	19.6	48.2	32.6	85.8
NP Emp NZM	Yes	II	17.5	46.1	26.0	78.4
NP ZM	No	-	8.0	36.6	13.5	69.1
NP ZM	No	I	22.2	52.2	33.4	88.0
NP ZM	No	II	20.5	49.5	26.6	78.5
NP NZM	No	-	9.6	42.7	16.6	76.6
NP NZM	No	I	21.8	51.6	32.0	86.3
NP NZM	No	II	20.2	48.9	26.5	78.3

settings, for three log-linear models (O='intercept model', I='independence model', II= 'all two-way interactions model'), and four different assumptions on mixed effects (P='normal mixed effects', NP Emp NZM = 'Empirical Bayes estimates of fixed effects and Dirichlet process with non zero mean random effects', NP NZM = 'Dirichlet process with non zero mean random effects', NP ZM = 'Dirichlet process with zero mean random effects'). In the second column of Table 1 we denote by 'Yes' the presence of an intercept in the linear combination defining the λ_k, and by 'No' its absence.

Table 1 shows a good performance of the all two-way interactions model among parametric models (P) which is in line with what reported in the literature (see [15] and [14]) . Although a proper comparison of the results reported in Table 1 would require an assessment of the posterior variability, new and interesting findings are:
1) the performance of nonparametric log-linear independence models, say (NP+I), is comparable to that of the parametric log-linear all two-way interactions model, say (P+II). This means that the Dirichlet process prior is able to capture the essential features of heterogeneity without increasing the dimensionality of the problem.
2) the potential of the Dirichlet process prior for capturing latent infomation not modeled by covariates can be appreciated by comparing the parametric log-linear model that only contains the overall mean, say (P+O) and the nonparametric models NP ZM and (NP NZM+O). The latter is the model used in Dorazio *et al.* (2008) [2]; it is more flexible than NP ZM, since the intercept term β_0 migrates to the role of unknown mean of the base measure of the Dirichlet process.
3) For log-linear models I and II, the comparison between nonparametric empirical Bayes estimates of τ_1 and the estimates obtained from equation (3) is equivalent to the comparison between nonparametric empirical Bayes estimates of τ_1 and fully Bayesian parametric estimates reported in the second and third rows of Table 1, since the latter estimates are obtained from vague priors.
It is worth noting that, in nonparametric models (NP), as the complexity of the log-

linear model increases, a posteriori the average number of clusters decreases (results not reported). Those considerations suggest that it is worth studying whether the proposed nonparametric log-linear independence models (NP+I) allows to obtain accurate risk estimation in very sparse tables with huge numbers of cells compared to an all two-way interaction model (P+II). Since the number of unknown parameters involved in (NP+I) and (P+II) models is significantly different, and the number of corresponding zero marginal counts is also very different, a more detailed analysis of the results obtained under the models (NP+I) and (P+II) seems to be a promising avenue of future research.

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

References

1. Bethlehem, J., Keller, W., Pannekoek, J.: Disclosure control of microdata. Journal of the American Statistical Association **85**, 38–45 (1990)
2. Dorazio, R., Mukherjee, B., Zhang, L., Ghosh, M., Jelks, H., Jordan, F.: Modeling unobserved sources of heterogeneity in animal abundance using a dirichlet process prior. Biometrics **64**, 635–644 (2008)
3. Ferguson, T.S.: A Bayesian analysis of some nonparametric problems. The Annals of Statistics **1**(2), 209–230 (1973)
4. Forster, J.J., Webb, E.L.: Bayesian disclosure risk assessment: predicting small frequencies in contingency tables. Journal of the Royal Statistical Society: Series C **56**(5), 551–570 (2007)
5. Gelman, A., Rubin, D.B.: Inference from iterative simulation using multiple sequences. Statistical Science **7**(4), 457–472 (1992)
6. Girolami, M., Calderhead, B.: Riemann manifold Langevin and Hamiltonian Monte Carlo methods. Journal of the Royal Statistical Society: Series B **73**(2), 123–214 (2011)
7. Li, Y., Mueller, P., Lin, X.: Center-adjusted inference for a nonparametric Bayesian random effect distribution. Statistica Sinica **21**(3), 1201–1223 (2011)
8. Liu, J.S.: Nonparametric hierarchical Bayes via sequential imputations. Annals of Statistics **24**(3), 911–930 (1996)
9. Lo, A.Y.: On a class of Bayesian nonparametric estimates. I. Density estimates. Annals of Statistics **12**(1), 351–357 (1984)
10. Neal, R.M.: Probabilistic inference using Markov chain Monte Carlo methods. Tech. Rep. CRG-TR-93-1, Dept. of Computer Science, University of Toronto (1993)
11. Neal, R.M.: Markov Chain Sampling Methods for Dirichlet Process Mixture Models. Journal of Computational and Graphical Statistics **9**(2), 249–265 (2000)
12. Rinott, Y., Shlomo, N.: A generalized negative binomial smoothing model for sample disclosure risk estimation. In: J. Domingo-Ferrer, L. Franconi (eds.) Privacy in Statistical Databases, *Lecture Notes in Computer Science*, vol. 4302, pp. 82–93. Springer Berlin / Heidelberg (2006)
13. Roberts, G.O., Rosenthal, J.S.: Examples of adaptive MCMC. Journal of Computational and Graphical Statistics **18**(2), 349–367 (2009)
14. Skinner, C., Shlomo, N.: Assessing identification risk in survey microdata using log-linear models. Journal of the American Statistical Association **103**(483), 989–1001 (2008)
15. Skinner, C.J., Holmes, D.J.: Estimating the re-identification risk per record in microdata. Journal of Official Statistics **14**, 361–372 (1998)
16. Takemura, A.: Some superpopulation models for estimating the number of population uniques. In: Proceedings of the Conference on Statistical Data Protection, pp. 45–58. Lisbon (1999). Eurostat, Luxembourg