A Generalized Negative Binomial Smoothing Model for Sample Disclosure Risk Estimation

Yosef Rinott[1] and Natalie Shlomo[2,⋆]

[1] The Hebrew University of Jerusalem, Israel
[2] The Hebrew University of Jerusalem, Israel and Southampton Statistical Sciences Research Institute, University of Southampton, UK

Abstract. We deal with the issue of risk estimation in a sample frequency table to be released by an agency. Risk arises from non-empty sample cells which represent small population cells and from population uniques in particular. Therefore risk estimation requires assessing which of the relevant population cells are indeed small. Various methods have been proposed for this task, and we present a new method in which estimation of a population cell frequency is based on smoothing using a local neighborhood of this cell, that is, cells having similar or close values in all attributes.

The statistical model we use is a *generalized Negative Binomial* model which subsumes the Poisson and Negative Binomial models. We provide some preliminary results and experiments with this method.

Comparisons of the new approach are made to a method based on *Poisson regression log-linear hierarchical model*, in which inference on a given cell is based on classical models of contingency tables. Such models connect each cell to a 'neighborhood' of cells with one or several common attributes, but some other attributes may differ significantly. We also compare to the *Argus* Negative Binomial method in which inference on a given cell is based only on sampling weights, without learning from any type of 'neighborhood' of the given cell and without making use of the structure of the table.

1 Introduction

Let $\mathbf{f} = \{f_k\}$ denote an m-way frequency table, which is a sample from a population table $\mathbf{F} = \{F_k\}$, where $k = (k_1, ..., k_m)$ indicates a cell, and f_k and F_k denote the frequency in cell k in the sample and population, respectively, and the number of cells is denoted by K. Formally, the sample and population sizes in our models are random and their expectations are denoted by n and N respectively. We formally assume that n and N are known, but in practice they are usually replaced by their natural estimators: the actual sample and population sizes, assumed to be known, and without further comment.

[⋆] Both authors' research was partially supported by the Israel Science Foundation grant No. 473/04.

J. Domingo-Ferrer and L. Franconi (Eds.): PSD 2006, LNCS 4302, pp. 82–93, 2006.

The m attributes in the table are considered to be *key variables*, that is, variables which are to some extent accessible to the public or to potential intruders. Disclosure risk arises from cells in which both f_k and F_k are positive and small, and in particular when $f_k = F_k = 1$ (sample and population *uniques*). An intruder who locates a sample unique in cell k, say, and is aware of the fact that in the population the combination of values $k = (k_1, ..., k_m)$ is unique ($F_k = 1$) or rare (F_k small) but matches an individual of interest, can identify this individual on the basis of these m attributes. If the sample contains information on the values of other attributes, then these can now be inferred for the individual in question, and his privacy is violated.

Individual risk measures will be briefly discussed in Section 2 and we start with *global risk measures* which quantify an aspect of the total risk in the file by aggregating risk over the individual cells. For simplicity we shall focus here only on two global measure, which are based on sample uniques:

$$\tau_1 = \sum_k \mathbf{I}(f_k = 1, F_k = 1)\,, \qquad \tau_2 = \sum_k \mathbf{I}(f_k = 1)\frac{1}{F_k}\,,$$

where $\mathbf{I}$ denotes the indicator function. Note that τ_1 counts the number of *sample uniques* which are also *population uniques*, and τ_2 is the expected number of correct guesses if each sample unique is matched to a randomly chosen individual from the same population cell. These measures are somewhat arbitrary, and one could consider measures which reflect matching of individuals that are not sample uniques, possibly with some restrictions on cell sizes. Also, it may make sense to normalize these measures by some measure of the total size of the table, by the number of sample uniques, or by some measure of the information value of the data.

Various individual and global risk measures have been proposed in the literature, see e.g. Franconi, and Polettini (2004) and references therein, Skinner and Holmes (1998), Elamir and Skinner (2006), Rinott (2003).

In Sections 2 and 3 we propose and explain a new method of estimation of quantities like τ_1 and τ_2, using a *generalized Negative Binomial model*, and *local smoothing* of frequency tables, Simonoff (1998). The method is based on the idea that one can learn about a given population cell from neighboring cells, if a suitable definition of closeness or neighbors is possible, by standard smoothing techniques, without relying on complex dependence structure modeling. This method differs from that of Elamir and Skinner (2006), in which one uses classical hierarchical log-linear models, which means inferring on a given cell by using cells that could be very different in many attribute values. For example, in the independence model, inference on a cell uses all cells which have at least one common attribute with the given cell, but all others may be very different. Thus neighborhoods formed by the classical log-linear model theory seem to be too large for our purposes. This point is explained in detail in Rinott and Shlomo (2005). On the other hand, the Argus approach, see, e.g., Franconi, and Polettini (2004), uses no neighborhoods at all and ignores the table structure. We consider the smoothing approach simple conceptually but not necessarily in terms of the computations required.

In this paper it is assumed that $\mathbf{f}$ is known, and $\mathbf{F}$ is an unknown parameter (on which there may be some partial information) and the quantities τ_1 and τ_2 should be estimated. Note that they are not proper parameters, since they involve both the sample $\mathbf{f}$ and the parameter $\mathbf{F}$.

The methods discussed in this paper consist of modeling the conditional distribution of $\mathbf{F}|\mathbf{f}$, estimating parameters in this distribution and then using estimates of the form

$$\hat{\tau}_1 = \sum_k \mathbf{I}(f_k = 1)\hat{P}(F_k = 1|f_k = 1), \quad \hat{\tau}_2 = \sum_k \mathbf{I}(f_k = 1)\hat{E}[\frac{1}{F_k}|f_k = 1], \tag{1}$$

where $\hat{P}$ and $\hat{E}$ denote estimates of the relevant conditional probability and expectation. For a general theory of estimates of this type see Zhang (2005) and references therein. Some direct variance estimates appear in Rinott (2003).

2 The Model

For completeness we briefly introduce the Poisson and Negative Binomial models. More details can be found, for example, in Bethlehem et al. (1990), Cameron and Trivedi (1998), Rinott (2003).

We assume $F_k \sim$ Poisson$(N\gamma_k)$, independently, with $\sum \gamma_k = 1$. Binomial (or Poisson) sampling from F_k means that $f_k|F_k \sim Bin(F_k, \pi_k)$, π_k being the (known) sampling fraction in cell k. These are common assumptions in the frequency table literature, where it is convenient for log-linear modeling to assume that all π_k's are equal, an assumption not made here. However, we assume that the inclusion probabilities π_k are fixed within cells. In certain cases such an assumption may not hold, and more complex models may be required.

By standard calculations we then have

$$f_k \sim \text{ Poisson}(N\gamma_k\pi_k) \text{ and } F_k \,|\, f_k \sim f_k + \text{ Poisson}(N\gamma_k(1-\pi_k))\,, \tag{2}$$

leading to the Poisson model (see references below).

We now add the Bayesian assumption $\gamma_k \sim$ Gamma(α, β) independently. (Later we assume a common value for α and β in some neighborhoods of cells, rather than the whole table.)

Then

$$f_k \sim NB(\alpha, p_k = \frac{1}{1 + N\pi_k\beta}), \tag{3}$$

the *generalized Negative Binomial distribution*, defined for any $\alpha > 0$ by

$$X \sim NB(\alpha, p) \text{ if } P(X = x) = \frac{\Gamma(x+\alpha)}{\Gamma(x+1)\Gamma(\alpha)}(1-p)^x p^\alpha, \; x = 0, 1, 2, \ldots, \tag{4}$$

which for a natural α counts the number of *failures* until α successes occur in independent Bernoulli trials with probability of success p. For this distribution we have $\mu = EX = \alpha(1-p)/p$, $\text{Var}X = \alpha(1-p)/p^2 = \mu + \mu^2/\alpha$, and the probability generating function is $g(t) = Et^X = p^\alpha/[1-(1-p)t]^\alpha$, see Cameron and Trivedi (1998, p 375).

With the above parametrization $\mu_k \equiv Ef_k = N\pi_k\alpha\beta$, and for $b > 0$

$$E[1/(b+X)] = \int_0^1 t^{b-1}g(t)dt. \tag{5}$$

Further calculations yield

$$F_k \mid f_k \sim f_k + NB(\,\alpha + f_k,\ \rho_k = \frac{N\pi_k\beta + 1}{N\beta + 1}\,), \tag{6}$$

and clearly $F_k \geq f_k$.

This is the generalized Negative Binomial model used in this paper.

As $\alpha \to 0$ and $\beta \to \infty$ we obtain $F_k \mid f_k \sim f_k + NB(f_k, \pi_k)$, which is exactly the Negative Binomial assumption used in the *Argus* method. See Franconi and Polettini (2004) and references therein for details. If $\alpha \to \infty$ and $\alpha\beta \to$ constant, the Poisson model used in this context by Skinner and Holmes (1998) and Elamir and Skinner (2006) is obtained. Therefore the generalized Negative Binomial subsumes both models.

Using (5), (6) and setting $\rho_k = (N\pi_k\beta + 1)/(N\beta + 1)$, it is easy to compute *individual risk measures* for cell k, defined by

$$P(F_k = 1|f_k = 1) = \rho_k^{1+\alpha}, \qquad E[\frac{1}{F_k}|f_k = 1] = \frac{\rho_k(1-\rho_k^\alpha)}{\alpha(1-\rho_k)}. \tag{7}$$

3 Smoothing Polynomials and Local Neighborhoods

Our goal in this section is to estimate the parameters of the model so that we can estimate the quantities in (7). The global risk measures will then be estimated as indicated in (1).

The estimation question here is essentially the following: given, say, a sample unique, how likely is it to be also a population unique, or arise from a small population cell. If a sample unique is found in a part of the sample table where neighboring cells (by some reasonable metric, to be discussed later) are small or empty, then it seems reasonable to believe that it is more likely to have arisen from a small population cell. This motivates our attempt to study local neighborhoods, and compare the results to those obtained by using model-driven neighborhood arising in hierarchical log-linear models, where it seems that the neighborhoods may be too large, and the Argus method which uses no neighborhoods.

Consider frequency tables in which some of the attributes are ordinal, and define closeness between categories of an attribute in terms of the order, or more generally, suppose that for a certain attribute one can say that some values of the attribute are closer to a given value than others. For example, Age and number of Years of Education are ordinal attributes, and naturally the age of 16 is closer to 15 than to 20, say, while Occupation is not ordinal, but one can try to define reasonable notions of closeness between different occupations. The attribute values of variables which are purely categorical will be kept fixed within a neighborhood, and ordinal variables will vary within a range that defines the neighborhood.

Classical log-linear models do not take such closeness into account, and therefore, when such models are used for individual cell parameter estimation, the estimates involve data in cells which may be rather remote from the estimated cell. On the other hand, as mentioned above, the *Argus* method bases its estimation only on the sampling weights in the estimated population cell. There is no learning from other cells, the structure of the table plays no role, and each cell's parameter is estimated separately.

Our approach consists of using local neighborhood smoothing which will be described in (10) below, along with the generalized Negative Binomial model of (3)-(6). We thus assume that $f_k \sim NB(\alpha, p_k = \frac{1}{1+N\pi_k\beta})$, and therefore $\mu_k \equiv Ef_k = \alpha(1-p_k)/p_k = N\pi_k\alpha\beta$, see (3) and the subsequent relations.

We describe the proposed estimation method for μ and α. These estimates will be transformed to estimates of the parameters appearing in the individual risk measures (7), which in turn lead to estimates of the global risk measures using (1).

For each fixed cell k we define a neighborhood of cells $M = M_k$ (where $k \in M$) and estimate the values of μ_k and α_k using neighboring cells $k' \in M_k$ and the assumption

$$f_{k'} \sim NB(\alpha_k, p_{k'} = \frac{1}{1+N\pi_{k'}\beta_k}), \tag{8}$$

where α_k and β_k are fixed in the neighborhood and do not depend on k', while $p_{k'}$ actually depends also on k. Since we now fix k we suppress it as an index in α, β or $p_{k'}$, and write $Ef_{k'} = \mu_{k'} = \alpha(1-p_{k'})/p_{k'}$. For the fixed k, set $\boldsymbol{\mu} = \{\mu_{k'} : k' \in M\}$, so the index k is suppressed also in μ. We consider the likelihood of the observations $\{f_{k'} : k' \in M\}$ in a neighborhood $M = M_k$ of k based on (8), and using different parameterizations which include μ and $a = 1/\alpha$

$$\begin{aligned} L(a,\boldsymbol{\mu}) \equiv L(a,\boldsymbol{\mu};\{f_{k'} : k' \in M\}) &= \prod_{k'\in M} \frac{\Gamma(x+\alpha)}{\Gamma(x+1)\Gamma(\alpha)}(1-p_{k'})^{f_{k'}} p_{k'}^{\alpha} \\ &= \prod_{k'\in M} \frac{\Gamma(x+\alpha)}{\Gamma(x+1)\Gamma(\alpha)}[1-\alpha/(\mu_{k'}+\alpha)]^{f_{k'}}[\alpha/(\mu_{k'}+\alpha)]^{\alpha}. \end{aligned} \tag{9}$$

We emphasize again that although in the above formulas only dependence on k' is shown, it should be noted that α, β and μ depend on k, and therefore $p_{k'}$ and $\mu_{k'}$ depend both on k and k'.

For each k we will estimate $\alpha = \alpha_k$ and $\mu_{k'}$ for $k' \in M = M_k$ using the likelihood (9) and a smoothing model described next, and then use the estimates of α_k and μ_k (not using the $\mu_{k'}$ estimates for $k' \neq k$) for further risk estimates, as discussed below.

Following Simonoff (1998), see also references therein, we use a local smoothing polynomial model.

For convenience of notation we now assume $m = 2$ (a two-way table); the extension to any m is straightforward. For each fixed $k = (k_1, k_2)$ separately, we write the log-linear model below for $\mu_{k'}$ in terms of the parameters $\boldsymbol{\theta}=(\theta_0, \theta_1, \vartheta_1, \ldots, \theta_t, \vartheta_t)$, with $k' = (k'_1, k'_2)$ varying in the neighborhood $M = M_k$ of k :

$$\log \mu_{k'}(\boldsymbol{\theta}) = \theta_0 + \theta_1(k'_1 - k_1) + \vartheta_1(k'_2 - k_2) + \ldots + \theta_t(k'_1 - k_1)^t + \vartheta_t(k'_2 - k_2)^t, \tag{10}$$

for some natural number t. One can hope that such a polynomial, with a suitable t, provides a reasonable approximation to $\log \mu_{k'}$ if $k' = (k'_1, k'_2)$ is in a small neighborhood of $k = (k_1, k_2)$. Substituting (10) into the likelihood function (9) using the relations between parameterizations as described above we obtain the likelihood function $L(a, \boldsymbol{\theta})$.

Our next goal is to maximize it as a function of $a = 1/\alpha$ and $\boldsymbol{\theta}$. This maximization takes place in principle for each cell k (although it may suffice for our purposes to carry it out for sample uniques only, that is, for cells such that $f_k = 1$). A source of difficulty here is that $\log L(a, \boldsymbol{\theta})$ is concave in $\boldsymbol{\theta}$, but not jointly in $(a, \boldsymbol{\theta})$, and therefore local maxima may occur, Hessians are not necessarily positive definite, and standard algorithms may not converge to the real MLE. This difficulty does not arise in the Poisson case of log-linear models of this type, where the log-likelihood is concave, see Rinott and Shlomo (2006), for a detailed discussion of the Poisson model. There are several options for maximization. SAS uses a *Newton-Raphson Ridge Optimization* (NRRIDG) which adds a multiple of the identity matrix to the Hessian when the latter is not positive definite, and also the *Fisher Scoring Algorithm* which replaces the Hessian by its expectation which is the information matrix, (using the parameter estimates of the current iteration), thus making it positive definite. We used our own program of the latter algorithm.

The components of the gradient of the log-likelihood function are obtained by differentiation and some manipulations as in Cameron and Trivedi (1998 p. 71), taking the form:

$$\frac{\partial \log L(a, \boldsymbol{\theta})}{\partial a} = \sum_{k' \in M} \left\{ \frac{1}{a^2} \left(\log(1 + a\mu_{k'}) - \sum_{v=0}^{f_{k'}-1} \frac{1}{v + a^{-1}} \right) + \frac{f_{k'} - \mu_{k'}}{a(1 + a\mu_{k'})} \right\}$$

$$\frac{\partial \log L(a, \boldsymbol{\theta})}{\partial \theta_\ell} = \sum_{k' \in M} \frac{f_{k'} - \mu_{k'}}{(1 + a\mu_{k'})} (k'_1 - k_1)^\ell, \quad \ell = 0, \ldots, t$$

$$\frac{\partial \log L(a, \boldsymbol{\theta})}{\partial \vartheta_\ell} = \sum_{k' \in M} \frac{f_{k'} - \mu_{k'}}{(1 + a\mu_{k'})} (k'_2 - k_2)^\ell, \quad \ell = 1, \ldots, t.$$

Note that in the solution to the related normal equations, the resulting vector $(a, \boldsymbol{\theta})$ depends on k.

The Hessian is calculated as follows:

$$\frac{\partial^2 \log L(a,\boldsymbol{\theta})}{\partial a^2} = \sum_{k'\in M} \left\{ \frac{-2}{a^3}\left(\log(1+a\mu_{k'}) - \sum_{\upsilon=0}^{f_{k'}-1}\frac{1}{\upsilon+a^{-1}}\right) + \frac{1}{a^2}\left(\frac{\mu_{k'}}{1+a\mu_{k'}} - \sum_{\upsilon=0}^{f_{k'}-1}\frac{1}{(a\upsilon+1)^2}\right) - \frac{(f_{k'}-\mu_{k'})(1+2a\mu_{k'})}{a^2(1+a\mu_{k'})^2}\right\},$$

$$\frac{\partial^2 \log L(a,\boldsymbol{\theta})}{\partial\theta_\ell \partial a} = -\sum_{k'\in M}\frac{(f_{k'}-\mu_{k'})\mu_{k'}}{(1+a\mu_{k'})^2}(k'_1-k_1)^\ell,\ \ \ell=0,\ldots,t,$$

and

$$\frac{\partial^2 \log L(a,\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j} = -\sum_{k'\in M}\frac{(1+af_{k'})\mu_{k'}}{(1+a\mu_{k'})^2}(k'_1-k_1)^{i+j}\ \ i,j=0,\ldots,t.$$

$$\frac{\partial^2 \log L(a,\boldsymbol{\theta})}{\partial\theta_i \partial\vartheta_j} = -\sum_{k'\in M}\frac{(1+af_{k'})\mu_{k'}}{(1+a\mu_{k'})^2}(k'_1-k_1)^{i}(k'_2-k_2)^{j},\ \ i=0,\ldots,t,\ j=1,\ldots,t.$$

$$\frac{\partial^2 \log L(a,\boldsymbol{\theta})}{\partial\vartheta_i \partial\vartheta_j} = -\sum_{k'\in M}\frac{(1+af_{k'})\mu_{k'}}{(1+a\mu_{k'})^2}(k'_2-k_2)^{i+j},\ \ i,j=1,\ldots,t.$$

With $\arg\max L(a,\boldsymbol{\theta}) = (\hat{a},\hat{\boldsymbol{\theta}})$, and $\hat{\theta}_0$ denoting the first component of $\hat{\boldsymbol{\theta}}$, we finally obtain our estimate of $\mu_k = \mu_{(k_1,k_2)}$ in the form

$$\hat{\mu}_k \equiv \mu_k(\hat{\boldsymbol{\theta}}) = \exp(\hat{\theta}_0), \tag{11}$$

where the second equality is explained by taking $k' = k = (k_1,k_2)$ in (10).

To summarize, we obtain the estimates $\hat{a}_k$, $\hat{\boldsymbol{\theta}}$ both depending on k by a separate maximization for each k as explained above, leading to the estimates $\hat{a}_k$, and $\hat{\mu}_k$ of (11). For the risk measure discussed in this paper, it suffices to compute these estimates for cells k which are sample uniques, that is, $f_k = 1$

Having estimated $\hat{a}_k, \hat{\mu}_k$ for each cell k separately on the basis of a neighborhood M_k, we use them to estimate the quantities ρ_k and $\alpha = \alpha_k$ which are obtained by tracing back the reparameterizations. Using the relations $\rho_k = \frac{N\pi_k\beta+1}{N\beta+1}$, and $\mu_k = N\pi_k\alpha_k\beta_k$ we readily obtain

$$\rho_k = \frac{\mu_k + \alpha_k}{\mu_k/\pi_k + \alpha_k}, \qquad \alpha_k = 1/a_k.$$

We plug our estimates $\hat{a}_k, \hat{\mu}_k$ in the latter formula, and then plug the resulting estimates of α_k and ρ_k into (7), to obtain the individual risk estimates. The global risk measures are estimated as indicated in (1).

4 Experiments with Neighborhoods

We present a few experiments. Our results are preliminary as already mentioned and more work is needed on the approach itself and on classifying types of data for which it might work.

For the computations we used our versions of the *Argus* and log-linear models methods, programmed on the SAS system. The weights w_i for the *Argus* method in all our examples were computed by post-stratification on Sex by Age by Geographical location (the latter is not one of the attributes in any of the tables, but it was used for post-stratification). These variables are commonly used for post-stratification, other strata may give different, and perhaps better results.

In the experiments below we compare results of our NB smoothing method with the Argus estimates and with Poisson hierarchical log-linear models (Elamir and Skinner 2006), with two log-linear models: one of independence, the other including all two-way interactions.

We defined neighborhoods M of k by varying around k coordinates corresponding to attributes that are ordinal, allowing in each coordinate a fixed maximal distance, which is equivalent to using a ball in the sup-norm, or intersection of sup-norm and ℓ_1 balls (see below). In principle we would use close values in non-ordinal attributes when possible (e.g., in Occupation). Attributes in which closeness of values cannot be defined, such as Sex remain constant in the whole neighborhood and therefore in our experiments neighborhoods always consist of individuals of the same Sex.

In all experiments we took a real population data file of size N given in the form of a contingency table with K cells, and from it we took a simple random sample of size n, so that always $\pi_k = n/N$. Our approach and formulas have the advantage of allowing for variable π_k's, but taking them all equal enables us to compare to the log-linear models method, where equal π_k's are required. Since the population and the sample are known to us, we can compute the *true values* of τ_1 and τ_2 and their estimates by the different methods, and compare.

Example 1. Population : an extract from the 1995 Israeli Census. $N = 37,586$, $n = 3,759$, $K = 11,648$. Attributes (with number of levels in parentheses): Sex(2) * Age Groups (32) * Income Groups(14) * Years of Study (13).

In this small experiment we tried our proposed smoothing polynomial model of (10) for $t = 2$. We considered one type of neighborhood here, constructed by fixing Sex and varying each of the other attribute value in k by at most c values up or down, that is, the neighborhood of each cell k (with a fixed Sex value) is of the type

$$M = \{k' : k'_1 = k_1, \max_{2 \le i \le m} |k'_i - k_i| \le c\}. \tag{12}$$

With $m = 4$ and one variables fixed we vary three variables, each over a range of five values for $c = 2$, , so we have $|M| = 5^3 = 125$, and taking $c = 3$ we have $|M| = 7^3 = 343$.

For cells near the boundaries some of the cells in their neighborhoods do not exist; here we set non-existing cells' frequencies to be zero, but other possibilities can be considered.

The table below presents the true τ values and their estimates by the methods described above.

Model	τ_1	τ_2
True Values	187	452.0
Argus	137.2	346.4
Log Linear Model: Independence	217.3	518.0
Log Linear Model: 2-Way Interactions	167.2	432.8
NB Smoothing $t = 2$ $\|M\| = 125$	181.9	461.3
NB Smoothing $t = 2$ $\|M\| = 343$	179.6	449.8

Example 2. Population : an extract from the 1995 Israeli Census. $N = 746,949$, $n = 14,939$, $K = 337,920$. Attributes: Sex (2) * Age Groups (16) * Years of Study (10) * Number of Years in Israel (11) * Income Groups (12) * Number of Persons in Household (8). Note that this is a very sparse table.

We applied the smoothing polynomial of (10) for $t = 2$ and neighborhoods obtained by varying all attributes except for Sex which was fixed. Neighborhoods are of the type

$$M = \{k' : k'_1 = k_1, \max_{2 \le i \le m} |k'_i - k_i| \le c, \sum_i |k'_i - k_i| \le d\}, \tag{13}$$

with $c = 2$; $d = 4$ and 6, and $|M| = 581$ and $1,893$, respectively. The results are given in the table below.

Model	τ_1	τ_2
True Values	430	1,125.8
Argus	114.5	456.0
Log Linear Model: Independence	773.8	1,774.1
Log Linear Model: 2-Way Interactions	470.0	1,178.1
NB Smoothing $t = 2$ $\|M\| = 581$	300.7	999.4
NB Smoothing $t = 2$ $\|M\| = 1,893$	461.9	1,179.6

Example 3. Population : an extract from the 1995 Israeli Census. $N = 746,949$, $n = 7,470$, $K = 42,240$. Attributes: Sex (2) * Age Groups (16) * Years of Study (10) * Number of Years in Israel (11) * Income Groups (12).

We applied the smoothing polynomial of (10) for $t = 2$ and neighborhoods obtained by varying all attributes except for Sex which was fixed. Neighborhoods are as in (13) with $c = 2$ and $d = 4$ and $|M| = 257$; $c = 2$, $d = 6$, and $|M| = 545$, and $c = 2$, $d = 8$ and $|M| = 625$. Smaller neighborhoods did not yield good estimates. The results are given in the table below.

Model	τ_1	τ_2
True Values	42	171.2
Argus	20.7	95.4
Log Linear Model: Independence	28.8	191.5
Log Linear Model: 2-Way Interactions	35.8	164.1
NB Smoothing $t = 2$ $\|M\| = 257$	24.7	147.5
NB Smoothing $t = 2$ $\|M\| = 545$	39.3	174.8
NB Smoothing $t = 2$ $\|M\| = 625$	45.8	184.4

Discussion of examples. The log-linear model method was tested in Skinner and Shlomo (2005, 2006) and references therein, and based on model selection techniques and goodness of fit criteria, yields good estimates for disclosure risk measures for the types of experiments done here. Di Consiglio et al. (2003) presented experiments for individual risk assessment with Argus, which seems to perform less well than the log-linear method in many of our experiments with global risk measures. Our new method still requires fine-tuning. At present the results seem comparable or somewhat better than the Poisson hierarchical log-linear method. In Rinott and Shlomo (2006) we performed experiments of this kind on a smoothing method based on the Poisson rather than the Negative Binomial distribution. So far the present Negative Binomial model improves all the results, and seems potentially promising.

Naturally, more variables and sparse data sets with a large number of cells are typical and need to be tested. Such files will cause difficulties to any method, and this is where the different methods should be compared. In sparse multi-way tables, model selection will be crucial but difficult for the log-linear method, and perhaps simpler for the smoothing approach.

Our proposed method is at a preliminary stage and requires more work. Particular directions are the following:

1. Adjust the parameter estimates to fit known population marginals obtained from prior knowledge and sampling weights, and vary the sampling fractions π_k. In all our experiments so far we used constant π_k's but unlike methods based on log-linear models, the formulas given here allow for variables π_k's, and we intend to try variable π_k's obtained by sampling design or post-stratification.
2. Use goodness of fit measures and information on population marginals and sampling weights to select the type and size of the neighborhoods, and the degree of the smoothing polynomial in (10).

The examples show a typical monotonicity phenomenon discussed also in the papers of Rinott and Shlomo (2005, 2006): the risk measure estimates decrease

as a function of the size of the log-linear model (that is, with one exception of τ_1 in Example 3, the two-way models always yield lower estimates than the independence model). In the present smoothing approach the risk estimates always decrease with the size of the neighborhood. These two facts can be explained in the same way: the better the fit to the sample data, the smaller the risk estimates. A larger log-linear model or a smaller smoothing neighborhood correspond to a better fit and therefore yield smaller risk estimates. In the presence of such monotonicity, a study of suitable goodness of fit measures to choose the right model is critical.
3. We intend to test this method also for individual risk measure estimates, which are important in themselves, and may also shed more light on efficient neighborhood and model selection. Our preliminary experiments suggest that the smoothing approach performs relatively well in estimating individual risk.

References

1. Benedetti, R., Franconi, L. and Piersimoni, F.: Per-record risk of disclosure in dependent data. *Proceedings of the Conference on Statistical Data Protection, Lisbon March 1998.* (1999) European Communities, Luxembourg.
2. Bethlehem, J., Keller, W., and Pannekoek, J.: Disclosure Control of Microdata. *J. Amer. Statist. soc.* **8**5 (1990) 38–45.
3. Cameron, A. C., and Trivedi, P. K.: *Regression analysis of count data*, Econometric Society Monographs **3**0 (1998) Cambridge University Press.
4. Di Consiglio, L., Franconi, L., and Seri, G.: Assessing individual risk of disclosure: an experiment. *Proceedings of the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality* (2003) Luxemburg 286–298.
5. Elamir, E. and Skinner, C.: Record-level measures of disclosure risk for survey microdata, *J. Official Statist* **2**2 (2006) to appear.
6. Franconi, L. and Polettini, S.: Individual risk estimation in mu-argus: a review. In J. Domingo-Ferrer and V. Torra, editors, *Privacy in Statistical Databases*, volume **3**050 of LNCS, Springer Berlin Heidelberg (2004) 262-272.
7. Polettini, S. and Seri, G.: Guidelines for the protection of social micro-data using individual risk methodology - Application within mu-argus version 3.2. *CASC Project Deliverable No. 1.2-D3* (2003). http://neon.vb.cbs.nl/casc/
8. Rinott, Y.: On models for statistical disclosure risk estimation. *Proceedings of the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*, Luxemburg (2003) 275-285.
9. Rinott, Y. and Shlomo, N.: A neighborhood regression model for sample disclosure risk estimation. *Proceedings of the Joint UNECE/Eurostat work session on statistical data confidentiality* Geneva, Switzerland (2005) 79-87.
10. Rinott, Y. and Shlomo, N.: A smoothing model for sample disclosure risk estimation. (2006) Submitted.
11. Simonoff S. J.: Three sides of smoothing: categorical Data smoothing, nonparametric regression, and density estimation, *International Statistical Review*, **66** (1998) 137-156.
12. Skinner, C. and Holmes, D.: Estimating the Re-identification Risk Per Record in Microdata, *J. Official Statist.* **14** (1998) 361-372.

13. Skinner, C. and Shlomo, N.: Assessing disclosure risk in microdata using record-level measures. In *Proceedings of the Joint UNECE/Eurostat work session on statistical data confidentiality* Geneva, Switzerland (2005) 69-78.
14. Skinner, C. and Shlomo, N.: Assessing identification risk in survey microdata using log-linear models. (2006) Submitted.
15. Willenborg, L. and de Waal T.: *Elements of Statistical Disclosure Control*, Lecture Notes in Statistics, **155** (2001) Springer, New York.
16. Zhang C.-H.: Estimation of sums of random variables: examples and information bounds. *Ann. Statist.* **33** (2005) 2022-2041.