Evaluation of per-record identification risk by additive modeling of interaction for contingency table cell probabilities

Akimichi TAKEMURA
Department of Mathematical Informatics
Graduate School of Information Science and Technology
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, JAPAN
takemura@stat.t.u-tokyo.ac.jp

Abstract

We propose to fit a Lancaster-type additive model of interaction terms for cell probabilities of contingency tables to evaluate the conditional probability of population uniqueness of sample unique records in microdata sets. Moment estimation of the Lancaster-type additive model is straightforward and the proposed estimation procedure is intuitively appealing from the viewpoint of disclosure risk assessment. In order to increase flexibility of the procedure, we also consider Ridge type shrinkage of the Lancaster-type additive model towards the independence model. For illustration we apply the proposed procedure to a test data set based on 1990 U.S. Census PUMS data.

1. Introduction

In evaluating the disclosure risk of a given microdata set, the number (or the proportion) of the population uniques among the sample unique records with respect to a given set of key variables is an important overall measure of the disclosure risk. For estimating the number of population uniques, various models have been proposed, including Poisson-Gamma model (Bethlehem et al. (1990)), Ewens sampling formula (e.g. Hoshino and Takemura (1998)), and more recently, Pitman model (Hoshino (2001)). These models treat the sample unique records exchangeably. Therefore under these models estimated conditional probability of population uniqueness is common for every sample unique record and is equal to the estimated proportion of population uniques among the sample unique records. However it is clear that some sample unique records are more likely to be population uniques than other records, depending on the intuitive "rareness" of the sample unique records. If a sample unique has very rare combination of observed characteristics, it is likely to be a population unique.

One way of evaluating the per-record identification risk is modeling of cell probabilities of the contingency table corresponding to a microdata set, where all the key variables of the microdata set are categorized and the joint frequencies of the key variables are counted.

Then the per-record identification risk can be evaluated in terms of the estimated conditional probability of population uniqueness of sample uniques cells. This approach was investigated in Skinner and Holmes (1998) and Fienberg and Makov (1998). They used the standard log-linear model of cell probabilities of contingency tables. In this paper we consider fitting the Lancaster-type additive model of interaction terms, because of its simple computation and interpretation.

In actual evaluation of disclosure risk, we often have to consider 10 or more possible key variables. Then the contingency table is m-way, where m is greater than equal to 10. Suppose that each variable has 10 categories, then the total number of the cells in the contingency table is 10^m. This shows that we have to deal with contingency table of very large size in disclosure risk assessment. In the application in Section 4 the total number of the cells is about 1/4 billion. The maximum likelihood estimation of log-linear model becomes computationally very expensive for large tables and simplicity of additive modeling of cell probabilities is attractive from computational viewpoint. On the other hand a difficulty with additive modeling is that estimated cell probabilities are not necessarily non-negative. Relative merits of log-linear model and additive model are discussed in Section 5.

Throughout this paper we assume the following simple superpopulation model: the cell probabilities of the contingency table are unknown but fixed and each of the N individuals of the population falls into a cell by an independent multivariate Bernoulli trial. Concerning the sampling we assume simple random sampling of n individuals without replacement. In this setting, the unobserved $N-n$ individuals are distributed independently of the observed n individuals and the evaluation of conditional probability is simply derived from the multinomial probability of the unobserved individuals.

In Section 2 we introduce the Lancaster-type additive model of interaction terms and its moment estimation. We also introduce Ridge type shrinkage of the Lancaster-type additive model towards the independence model. In Section 3 we discuss the estimation of the number of population uniques among the sample uniques based on the estimated Lancaster-type additive model. We also propose some procedures of checking the fit of the estimated model. In Section 4 we apply the proposed model to a test data set based on 1990 U.S. Census PUMS data. Finally in Section 5 we discuss relative merits of the log-linear model and the additive model for disclosure risk assessment.

2. Lancaster-type additive model and its shrinkage to independence

Here we describe the Lancaster-type additive model. For simplicity of notation we describe the model for 3-way contingency tables, although in actual applications we need to use m-way tables, where m (the number of key variables) is often around 10. Extension of the model to higher order tables is trivial except for notational complication.

Let p_{ijk} denote the cell probability of an $I \times J \times K$ contingency table. Denote the one-

dimensional marginal probabilities by $p_{i\cdot\cdot}, p_{\cdot j\cdot}, p_{\cdot\cdot k}$ and the two-dimensional marginal probabilities by $p_{ij\cdot}, p_{i\cdot k}, p_{\cdot jk}$. The Lancaster-type additive model without three-variable interactions in Lancaster's sense (Darroch (1974), Lancaster (1971), Zentgraf (1975), Chapter 12 of Lancaster (1969)) is defined by

$$p_{ijk} = p_{i\cdot\cdot}p_{\cdot j\cdot}p_{\cdot\cdot k}\Big\{1 + \Big(\frac{p_{ij\cdot}}{p_{i\cdot\cdot}p_{\cdot j\cdot}} - 1\Big) + \Big(\frac{p_{i\cdot k}}{p_{i\cdot\cdot}p_{\cdot\cdot k}} - 1\Big) + \Big(\frac{p_{\cdot jk}}{p_{\cdot j\cdot}p_{\cdot\cdot k}} - 1\Big)\Big\}. \tag{1}$$

Note that the independence model $p_{ijk} = p_{i\cdot\cdot}p_{\cdot j\cdot}p_{\cdot\cdot k}$ is modified by two-variable interaction terms of the form $p_{ij\cdot}/(p_{i\cdot\cdot}p_{\cdot j\cdot}) - 1$. If we sum this term with respect to i, j, k, we have

$$\sum_{i,j,k} p_{i\cdot\cdot}p_{\cdot j\cdot}p_{\cdot\cdot k}\Big(\frac{p_{ij\cdot}}{p_{i\cdot\cdot}p_{\cdot j\cdot}} - 1\Big) = \sum_{i,j,k} p_{ij\cdot}p_{\cdot\cdot k} - \sum_{i,j,k} p_{i\cdot\cdot}p_{\cdot j\cdot}p_{\cdot\cdot k} = 1 - 1 = 0.$$

Therefore the modification terms sum to 0 and p_{ijk} of (1) sum to 1. As shown in Darroch (1974), (1) is equivalent to the following hypothesis of no three-variable interaction in the sense of Lancaster:

$$H: \ \frac{p_{ijk}}{p_{i\cdot\cdot}p_{\cdot j\cdot}p_{\cdot\cdot k}} = (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk}.$$

See Darroch and Speed (1983), O'Neill (1982) on the relation between the Lancaster-type additive model and the standard log-linear model.

The moment estimation of (1) is straightforward. Let n denote the sample size and let $n_{i\cdot\cdot}, n_{ij\cdot}, \ldots,$ denote the sample marginal frequencies. Then the cell probability is estimated as

$$\begin{aligned}
\hat{p}_{ijk} &= \hat{p}_{i\cdot\cdot}\hat{p}_{\cdot j\cdot}\hat{p}_{\cdot\cdot k}\Big\{1 + \Big(\frac{\hat{p}_{ij\cdot}}{\hat{p}_{i\cdot\cdot}\hat{p}_{\cdot j\cdot}} - 1\Big) + \Big(\frac{\hat{p}_{i\cdot k}}{\hat{p}_{i\cdot\cdot}\hat{p}_{\cdot\cdot k}} - 1\Big) + \Big(\frac{\hat{p}_{\cdot jk}}{\hat{p}_{\cdot j\cdot}\hat{p}_{\cdot\cdot k}} - 1\Big)\Big\} \\
&= \frac{n_{i\cdot\cdot}}{n}\frac{n_{\cdot j\cdot}}{n}\frac{n_{\cdot\cdot k}}{n}\Big\{1 + \Big(\frac{nn_{ij\cdot}}{n_{i\cdot\cdot}n_{\cdot j\cdot}} - 1\Big) + \Big(\frac{nn_{i\cdot k}}{n_{i\cdot\cdot}n_{\cdot\cdot k}} - 1\Big) + \Big(\frac{nn_{\cdot jk}}{n_{\cdot j\cdot}n_{\cdot\cdot k}} - 1\Big)\Big\}\ .
\end{aligned} \tag{2}$$

Note that the moment estimate in (2) can be negative. If $\hat{p}_{ijk} < 0$ for some cell (i, j, k), we should first replace it by 0. Actually in the analysis of a test data set in Section 4 we have a large number of negative estimated cell probabilities. $\hat{p}_{ijk}$ of (2) always sum to 1. Therefore by replacing negative estimates by 0, the sum exceeds 1. We define renormalized non-negative moment estimate by

$$\frac{1}{c} \times \max(\hat{p}_{ijk}, 0), \qquad c = \sum_{\hat{p}_{i'j'k'} \geq 0} \hat{p}_{i'j'k'} = 1 - \sum_{\hat{p}_{i'j'k'} < 0} \hat{p}_{i'j'k'}. \tag{3}$$

When renormalized, the simplicity of the moment estimation is somewhat lost. We might consider maximum likelihood estimation of the additive model (1). However as discussed in Section 2 of Darroch and Speed (1983) the maximum likelihood estimation of the additive model seems to be difficult. Also it should be noted that there is some computational difficulty of evaluating the renormalizing constant c in (3) for large contingency tables. We discuss this point at the end of this section.

We have presented Lancaster-type additive model for 3-way contingency table. Generalization to m-way table is straightforward. The model can also be extended to include some higher order interactions.

Although the model (1) contains only up to the two-variable interactions, the number of estimated interaction terms may be large. Let $I_l,\ l = 1, \ldots, m$, denote the number of categories of the l-th variable. Then the total number of estimated parameters is roughly equal to

$$I_1 I_2 + I_1 I_3 + \cdots + I_{m-1} I_m.$$

For example, when $m = 10$ and $I_l = 10,\ l = 1, \ldots, m$

$$I_1 I_2 + I_1 I_3 + \cdots + I_{m-1} I_m = 4500.$$

Therefore there is a question of stability of the estimated cell probabilities. For this reason, we introduce a shrinkage factor $0 \leq \lambda \leq 1$ and consider Ridge type shrinkage of $\hat{p}_{ijk}$ of (2) towards the independence model:

$$\hat{p}_{ijk}(\lambda) = \frac{n_{i\cdot\cdot}}{n}\frac{n_{\cdot j\cdot}}{n}\frac{n_{\cdot\cdot k}}{n}\Big\{1 + \lambda\Big(\frac{nn_{ij\cdot}}{n_{i\cdot\cdot}n_{\cdot j\cdot}} - 1\Big) + \lambda\Big(\frac{nn_{i\cdot k}}{n_{i\cdot\cdot}n_{\cdot\cdot k}} - 1\Big) + \lambda\Big(\frac{nn_{\cdot jk}}{n_{\cdot j\cdot}n_{\cdot\cdot k}} - 1\Big)\Big\}. \tag{4}$$

The case $\lambda = 0$ is the independence model and the case $\lambda = 1$ is the full additive model (2). Note that the number of the parameters in the independence model is $I_1 + \cdots + I_m - m$. This is much smaller and we can expect stability in estimation by shrinking towards $\lambda = 0$.

For $\lambda > 0$ (4) can be negative. As above we consider replacing negative estimates by 0 and renormalizing them:

$$\frac{1}{c(\lambda)} \times \max(\hat{p}_{ijk}(\lambda), 0), \quad c(\lambda) = \sum_{\hat{p}_{i'j'k'}(\lambda) \geq 0} \hat{p}_{i'j'k'}(\lambda) = 1 - \sum_{\hat{p}_{i'j'k'}(\lambda) < 0} \hat{p}_{i'j'k'}(\lambda). \tag{5}$$

In the example of Section 4 we will see that $c(\lambda)$ can be substantially larger than 1. This indicates lack of fit of the additive model.

In the example in Section 4 we vary $\lambda = 0.0, 0.2, 0.4, 0.6, 0.8, 1.0$ and check the stability of the estimated models. It may be more appropriate to choose the value of λ from the data itself. For example we might consider the set of random variable $\{nn_{ij\cdot} - n_{i\cdot\cdot}n_{\cdot j\cdot}\}$ approximately normally distributed and use Stein type estimator. This seems somewhat too complicated for the present application and here we only consider fixed number of values of λ.

We have discussed how to deal with negative estimates of cell probabilities. This is closely related to the problem of structural zeros. The estimator $\hat{p}_{ijk}$ of (2) has the simple form because of the assumption of no structural zeros in the $I \times J \times K$ contingency table. In actual contingency tables corresponding to microdata sets from official statistics, there are usually many structural zeros. The Lancaster-type additive model loses its simplicity when the structural zeros are incorporated into the model. In this sense, we regard the model as convenient approximate model for quick evaluation of disclosure risk.

Finally we discuss the evaluation of the renormalizing constants c in (3) and $c(\lambda)$ in (5). As discussed above we have in mind contingency tables with large number of cells and actual summation in (3) and (5) might be time consuming. In this case we can use Monte Carlo simulation for approximate evaluation of c or $c(\lambda)$. Approximate value of c or $c(\lambda)$ is sufficient for the evaluation of the conditional probability of population uniqueness. Write (4) as

$$\hat{p}_{ijk}(\lambda) = p^0_{ijk} \times w_{ijk}(\lambda), \tag{6}$$

where

$$p^0_{ijk} = \hat{p}_{i..}\hat{p}_{.j.}\hat{p}_{..k} \tag{7}$$

denotes the estimate under the independence model and

$$w_{ijk}(\lambda) = 1 + \lambda\Big(\frac{nn_{ij.}}{n_{i..}n_{.j.}} - 1\Big) + \lambda\Big(\frac{nn_{i.k}}{n_{i..}n_{..k}} - 1\Big) + \lambda\Big(\frac{nn_{.jk}}{n_{.j.}n_{..k}} - 1\Big).$$

It is simple to resample from the independence model $\{p^0_{ijk}\}$. Let X denote the $n \times m$ data matrix of observations on m key variables of n individuals. The easiest way to resample from the independence model is to randomly choose one observation from each column of X, independently from column to column. Let $E_{p^0}(\cdot)$ denote the expected value under the estimated independence model. Then

$$1 - c(\lambda) = \sum_{\hat{p}_{ijk}(\lambda)<0} \hat{p}_{ijk}(\lambda) = E_{p^0}\Big[w_{ijk}(\lambda) I\{w_{ijk}(\lambda) < 0\}\Big], \tag{8}$$

where $I\{w_{ijk}(\lambda) < 0\}$ denotes the indicator function of the event $w_{ijk}(\lambda) < 0$. Therefore $1 - c(\lambda)$ can be estimated by the resampling average of $w_{ijk}(\lambda) I\{w_{ijk}(\lambda) < 0\}$ under the estimated independence model. Replication size of 200,000 seems to be sufficient for our purposes.

3. Estimation of the number of population uniques and diagnostics of the model

Once the cell probability is estimated, the per-record disclosure risk for sample unique record (or cell) is given as follows. As discussed in Section 1 we assume that N individuals of the population fall into the cells of the contingency table according to the multinomial scheme and the sampling of n individuals is by simple random sampling without replacement. Let n_{ijk} and N_{ijk} denote the sample and the population cell frequencies. Then given that a cell is a sample unique ($n_{ijk} = 1$), the conditional probability of the cell being a population unique ($N_{ijk} = 1$) is written as

$$P(N_{ijk} = 1 \mid n_{ijk} = 1) = (1 - p_{ijk})^{N-n}. \tag{9}$$

If we replace p_{ijk} by its estimate $\hat{p}_{ijk}$, we obtain an estimated value of the conditional probability. The number of population uniques in the microdata set is now estimated by

$$\sum_{(i,j,k)\,:\,n_{ijk}=1} \hat{P}(N_{ijk} = 1 \mid n_{ijk} = 1) = \sum_{(i,j,k)\,:\,n_{ijk}=1} (1 - \hat{p}_{ijk})^{N-n}. \tag{10}$$

Furthermore a simple approximate confidence interval of the number of population uniques can be obtained based on Poisson distribution with mean given by (10).

Note that (9) is a decreasing function of p_{ijk} and this reflects an intuitively obvious fact that the disclosure risk of a sample unique is high if the estimated probability of the cell is small. $\hat{p}_{ijk}$ of (2) is small if univariate relative frequencies are small and the terms like $(nn_{ij\cdot})/(n_{i\cdot\cdot}n_{\cdot j\cdot})$ are small. This term is the ratio of the actual two-dimensional frequency n_{ij} to the estimated frequency $n_{i\cdot\cdot}n_{\cdot j\cdot}/n$ under the independence model. Therefore $\hat{p}_{ijk}$ of (2) combines univariate rareness and bivariate rareness, which are routinely considered in disclosure control practices. This argument shows that $\hat{p}_{ijk}$ of (2) is a reasonable measure of the identification risk of a sample unique cell (i, j, k).

We now discuss how to assess the goodness of fit of the Lancaster-type additive model. As stated in Section 1 the standard superpopulation models treat the sample uniques exchangeably. More precisely under these models we only consider "size indices" (Sibuya (1993), Sibuya and Yamato (1995)) or "frequency of frequencies" (Good (1965))

$$s_0, s_1, s_2, \ldots,$$

where s_i is the number of the cells of frequency i in the sample. For the purpose of checking the goodness of fit of these models, we can evaluate expected values of size indices under estimated models and compare the expected values with actual size indices. For the Lancaster-type additive model expected values of size indices can be evaluated by resampling.

Since the random variables are discrete it is conceptually very simple to resample from the estimated model. However in the case of very large contingency table, it seems better to avoid summing large number of small estimated probabilities in resampling. Here we consider using Metropolis-Hastings type algorithm of Markov Chain Monte Carlo method (Hastings (1970)) using the independence model $\{p^0_{ijk}\}$ of (7) as the Markov transition kernel. Let (i, j, k) be the current cell. We choose a candidate of the next cell (i', j', k') according to the independence model $\{p^0_{i'j'k'}\}$. If $w_{i'j'k'}(\lambda) < 0$ we simply ignore this cell (not counting as a step of the Markov chain) and choose another candidate (i', j', k'). If $w_{i'j'k'}(\lambda) \geq 0$ we move to (i', j', k') with probability

$$p = \min\left\{\frac{w_{i'j'k'}(\lambda)}{w_{ijk}(\lambda)}, 1\right\}, \tag{11}$$

and stays at (i, j, k) with probability $1 - p$. Even if we stay at (i, j, k) we count this as one step of the Markov chain. From the general results on Markov Chain Monte Carlo method we see that the estimated renormalized model $\{(1/c(\lambda)) \times \max(\hat{p}_{ijk}(\lambda), 0)\}$ forms the unique stationary distribution of the Markov chain and by iterating the chain long enough we can resample from the estimated model. It should be noted that we should use intermittent observations from the chain and avoid using consecutive observations. This is because when the Markov chain stays at the same cell consecutive observations at the same cell has an obvious downward bias for the number of the sample unique cells.

4. Application to a test data set from U.S. Census PUMS data

We applied the Lancaster-type additive model to a test data set subsampled from 1990 U.S. Census of Population and Housing Public Use Microdata Samples. We subsampled $n = 9809$ individuals from the state of Washington and chose $m = 10$ variables for experimental purpose:

1. Relationship (14 categories), 2. Sex (2), 3. Age (91), 4. Marital status (5),
5. Place of birth (14), 6. Spouse present/absent (7), 7. Own child (2),
8. Age of own child (5), 9. Related child (2), 10. Detailed relationship (10).

The population size is $N = 4,867,000$. The dataset can be viewed as a contingency table of the type

$$14 \times 2 \times 91 \times 5 \times 14 \times 7 \times 2 \times 5 \times 2 \times 10$$

with 249,704,000 cells. We see that the contingency table is very sparse with only $n = 9809$ counts among 249,704,000 cells Fitting log-linear model to contingency table of this size is computationally fairly difficult. We took these $m = 10$ variables from a PUMS data set without further global recoding. For example we used the age itself with 91 categories. This is somewhat unrealistic for evaluation of disclosure risk. On the other hand there are other possible key variables in the original PUMS data set. We also intended to check how the proposed model works for large contingency tables.

Although the (formal) total number of cells 249,704,000 is very large, the effective total number should be much smaller because of structural zeros. For example there is no age of own child if there is no own child. In this case the age of own child is coded as N/A in the data set. Also there is an obvious relation between age and marital status. The existence of large number of structural zeros seems to adversely affect the fit of the model as discussed below.

For reference we show first few lines of 9809×10 data matrix.

```
00,0,17,4,10,6,0,0,0,0
00,0,17,4,52,6,0,0,0,0
00,0,18,0,23,1,0,0,0,0
00,0,18,0,24,1,0,0,0,0
00,0,18,0,51,1,0,0,0,0
```

The frequencies of the cell sizes (size indices, frequency of frequencies) $s_1, s_2, \ldots$, of this data set is given as follows.

Cell size	1	2	3	4	5	6	7	8	9	10	11 $\leq$
Frequency	2249	521	275	132	104	60	59	34	46	19	124

We fitted the Lancaster-type additive model to this data set with Ridge type shrinkage with $\lambda = 0.0, 0.2, 0.4, 0.6, 0.8, 1.0$. The estimated number of population uniques among the 2249 sample uniques based on (10) is given in Table 1.

Table 1: Estimated number of the population uniques

λ	0.0	0.2	0.4	0.6	0.8	1.0	Ewens	Pitman
no renormalization	885.02	480.27	372.82	315.93	278.88	252.13	5.88	213.96
renormalization	885.02	482.94	403.46	376.21	362.50	354.68		
sum of negative prob.	0	–0.013	–0.193	–0.465	–0.758	–1.065		

In Table 1 the numbers below "Ewens" and "Pitman" show the estimated numbers of the population uniques under these models. Derivation of these numbers is discussed below. "No renormalization" means using (4) ignoring negative estimated cell probabilities and "Renormalization" means that the renormalization in (5) was applied. Sum of negative probabilities stands for $1 - c(\lambda)$ in (8), where the summation is evaluated by Monte Carlo simulation with replication size 200,000.

The sum of negative estimated cell probabilities is substantial for $\lambda \geq 0.4$. It is -0.465 for $\lambda = 0.6$, which indicates that negative estimated cell probabilities are very frequent. This indicates that the fit of our additive model is not very good for this data set. However it should be mentioned that for the 3623 nonempty cells in the data set, all estimated probabilities were positive even for $\lambda = 1$. This implies that all the 3623 nonempty cells tend to have large marginal bivariate frequencies. This seems to correspond to positive correlation among the variables. It should also be noted that renormalization does not greatly affect the estimated number of population uniques.

The estimated number of population uniques decreases as λ increases from 0 to 1. A possible explanation on this decrease is as follows. Under the independence model $\lambda = 0$, estimated cell probability is positive for every cell, whenever the univariate marginal frequencies are all positive. This implies that the probability mass is spread all over the contingency table. However as discussed above there exist correlations among variables and in addition there are many structural zeros. Therefore under the independence model the probabilities of nonempty cells tend to be underestimated and the probabilities of empty cells tend to be overestimated. Since sample unique cells are nonempty, the conditional probabilities of population uniqueness $(1 - \hat{p}_{i_1 \cdots i_m})^{N-n}$ for the sample unique cells $(n_{i_1 \cdots i_m} = 1)$ are overestimated leading to an overestimated number of population uniques. From this viewpoint the decrease of the estimated number of population uniques as λ increases from 0 to 1 suggests that the fit of the model is improving by incorporating two-variable interactions. Obviously this argument is not totally persuasive in view of lack of fit indicated by the sum of negative estimated cell probabilities.

We briefly discuss estimation of the number of population uniques under the Ewens model and the Pitman model in Table 1. A simple moment estimate under the Ewens model is given by (Hoshino and Takemura (1998))

$$\frac{s_1 n(n-1)}{n(N-1)-s_1(N-n)} \doteq s_1 \frac{n}{n-s_1}\frac{n}{N} = 5.88.$$

For the Pitman model let $u = 3623$ denote the number of nonempty cells in the sample and let $\hat{\alpha} = s_1/u = 2249/3623 = 0.621$. The following simple moment estimate under the Pitman model is proposed by Hoshino (2001):

$$s_1\Big(\frac{n}{N}\Big)^{1-\hat{\alpha}} = 213.96.$$

In our experiences the Ewens model and its related models including Poisson-Gamma model tend to underestimate the number of the population uniques and the Pitman model behaves in an opposite manner. For this example our estimate under the Lancaster-type additive model is even larger than that of the Pitman model.

We now resample from the estimated Lancaster-type additive model for the purpose of checking the goodness of fit of the model. We use the MCMC method described at the end of Section 3. In each run of the Markov chain, we discarded the first 5000 steps and then we recorded $n = 9809$ observations, which are 20 steps apart from each other. From the observations we obtained sample size indices. For each value of λ we performed 100 runs of the Markov chain and took the average of sample size indices. The result is given in Table 2.

Table 2: Expected size indices under the estimated model

Cell size	1	2	3	4	5	6	7	8	9	10	11 $\leq$
$\lambda = 0.0$	9060	323.8	28.6	3.28	0.4	0.04	0	0	0	0	0
$\lambda = 0.2$	8769.2	412.9	52.1	10.3	2.43	0.53	0.08	0.02	0.01	0	0
$\lambda = 0.4$	8435.8	502.6	82.8	19.6	5.29	1.52	0.55	0.18	0.06	0	0
$\lambda = 0.6$	8283.8	538.4	96.9	24.1	7.57	2.51	0.72	0.25	0.11	0.01	0.45
$\lambda = 0.8$	8189.4	559.1	104.6	28.2	8.67	3.0	1.24	0.3	0.18	0.06	0.41
$\lambda = 1.0$	8132.6	574.4	109.0	29.7	9.67	3.29	1.12	0.5	0.18	0.03	0.22
Actual	2249	521	275	132	104	60	59	34	46	19	124

As seen from Table 2 the expected number of sample uniques is very much higher than the actual value of 2249. Overall the distribution of the size indices are shifted downward compared to the observer size indices. This is another indication that the fit of the proposed model is not good for this data set. Again the reason for this lack of fit seems to be excessive spread of the probability mass over the whole contingency table leading to many cells with very small probability. This tendency is expected for the case of the

independence model $\lambda = 0$ as discussed above. In Table 2 the expected number of the sample uniques decrease as λ increases from 0 to 1. This again suggests that the fit of the model is improving by incorporating two-variable interactions. Although there is no direct relationship between the expected number of sample uniques and the estimated number of population uniques under the Lancaster-type additive model, Table 2 suggests that the number of population uniques is overestimated by Lancaster-type additive model in Table 1.

5. Comparison to the log-linear model and some discussion

In this study we have used the Lancaster-type additive model of the interaction terms. The standard model for analyzing contingency tables is the log-linear model (e.g. Bishop, Fienberg and Holland (1975)). With the log-linear model estimates are always nonnegative. Furthermore the problem of structural zeros can be appropriately handled in the framework of the maximum likelihood estimation (Chapter 5 of Bishop, Fienberg and Holland (1975)). However for large contingency tables there is a computational difficulty in fitting the log-linear model.

Consider $I \times J \times K$ 3-way contingency table. The log-linear model without three-variable interaction is defined by

$$\log p_{ijk} = (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk}. \tag{12}$$

For given $\{(\alpha\beta)_{ij}, (\alpha\gamma)_{ik}, (\beta\gamma)_{jk}\}$ the normalizing constant c is defined by

$$1 = \frac{1}{c} \times \sum_{i,j,k} \exp((\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{ik}). \tag{13}$$

It seems that for this log-linear model the expression for c can not be simplified and we need to perform the actual summation for the exact evaluation of c. In the framework of exponential family, c corresponds to the moment generating function and its partial derivatives with respect to the natural parameters $\{(\alpha\beta)_{ij}, (\alpha\gamma)_{ik}, (\beta\gamma)_{jk}\}$ are needed in many algorithms for the maximum likelihood estimation. Avoiding the actual summation by simulation as in (8) does not seem to be appropriate for the iterative steps of maximum likelihood estimation. The difficulty in evaluating the normalizing constant in exponential family is well known in other contexts, e.g., Gibbs distribution in spatial statistics or Boltzman machine in neural networks (e.g. Titterington and Anderson (1994)). The normalizing constant c in (13) is called the "partition function" in these areas and its evaluation is an important topic.

In terms of the expectation parameter, instead of natural parameter, the iterative proportional fitting (Section 3.5 of Bishop, Fienberg and Holland (1975), Chapter 4 of Lauritzen (1996)) is the standard procedure for maximum likelihood estimation. In the

iterative proportional fitting we do not need to compute the normalizing constant. However we need to keep the expected cell sizes of all the cells in computer memory at each step of the iteration. This is not feasible for large contingency tables.

Compared to the log-linear model, the simplicity of moment estimation of the additive model is attractive. We could regard the additive model as a quick alternative for the log-linear model. It is important to further investigate the feasibility of fitting the log-linear model utilizing the ever increasing computing power.

We might interpret $\hat{p}_{ijk}(\lambda)$ only as a relative measure of the per-record disclosure risk. Smaller the value of $\hat{p}_{ijk}(\lambda)$, riskier the sample unique. The total number of the population uniques can be separately estimated by standard models.

Concerning the test data set in Section 4, the log-likelihood model including all the two-variable interactions might not improve the fit over the additive model, because our analysis suggests that two-variable interactions do not fully capture the dependence structure of the variables and the structural zeros.

It is theoretically most desirable to specify all the structural zeros based on the definition of the variables. However this is often too cumbersome. One practical approach would be to regard all sampling zeros in bivariate marginal tables as structural zeros. We might simply let the estimated cell probability be zero, whenever the cell belongs to an empty marginal cell in bivariate marginal tables.

Acknowledgment. I am indebted to Shidou Sai for the preparation of the data set treated in Section 4 and to Satoshi Kuriki for relevant references on additive model.

REFERENCES

[1] Bethlehem, J.G., Keller, W.J. and Pannekoek, J. (1990). Disclosure control of microdata. *Journal of the American Statistical Association*, **85**, 38–45.

[2] Bishop, Y.M.M., Fienberg, S.E. and Holland P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*, The MIT Press, Cambridge.

[3] Darroch, J. N. (1974). Multiplicative and additive interaction in contingency tables. *Biometrika*, **61**, 207–214.

[4] Darroch, J.N. and Speed, T.P. (1983). Additive and multiplicative models and interactions. *Ann. Statist.*, **11**, 724–738.

[5] Fienberg, S.E. and Makov, U.E. (1998). Confidentiality, uniqueness, and disclosure limitation for categorical data. *Journal of Official Statistics*, **14**, 385–397.

[6] Good, I.J. (1965). *The Estimation of Probabilities: An Essay on Modern Bayesian Methods.* MIT Press, Cambridge, Massachusetts.

[7] Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

[8] Hoshino, N. (2001). Applying Pitman's sampling formula to microdata disclosure risk assessment. To appear in *Journal of Official Statistics*.

[9] Hoshino, N. and Takemura, A. (1998). Relationship between logarithmic series model and other superpopulation model useful for microdata disclosure risk assessment. *Journal of the Japan Statistical Society*, **28**, 125–134.

[10] Lancaster, H.O. (1969). *The Chi-Squared Distribution.* Wiley, London.

[11] Lancaster, H.O. (1971). The multiplicative definition of interaction. *Austral. J. Statist.*, **13**, 36–44.

[12] Lauritzen, S.L. (1996). *Graphical Models*. Clarendon Press, Oxford.

[13] O'Neill, M.E. (1982). A comparison of the additive and multiplicative definitions of second-order interaction in $2 \times 2 \times 2$ contingency tables. *J. Statist. Comput. Simul.*, **15**, 33–55.

[14] Sibuya, M. (1993). A random clustering process. *Ann. Inst. Stat. Math.*, **45**, 459–465.

[15] Sibuya, M. and Yamato, H. (1995). Characterization of some random partitions. *Japan J. Indust. Appl. Math.*, **12**, 237–263.

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

[17] Titterington, D.M. and Anderson, N.H. (1994). Boltzmann machines. in *Probability, Statistics and Optimisation.* F. P. Kelly ed., Wiley, Chichester.

[18] Zentgraf, R. (1975). A note on Lancaster's definition of higher-order interactions. *Biometrika*, **62**, 375-378.