STATISTICA, *anno LXIV, n. 4, 2004*

A TWO-STEP SMOOTHING PROCEDURE FOR THE ANALYSIS OF SPARSE CONTINGENCY TABLES WITH ORDERED CATEGORIES

Riccardo Borgoni

1. INTRODUCTION

The main interest of many statistical analyses of economical, demographic and social phenomena consists of investigating the multivariate structure of the relationships behind the data. Quite often the collected information is coded in terms of a categorical variable either because the nature of the considered characteristic is intrinsically discrete or nominal or because the variable results from the categorization into groups of an inherently continuous one. In the latter case the resulting variable will take values on an ordinal scale like classes of years or income. Even in the former case however, the variable can be naturally ordered. For instance the political interest of a person can be classified as moderate, medium or active, the level of attained education can be classified as low, medium and high and so on. A number of different methods have been proposed for the analysis of categorical data. Systematic reviews on categorical data analysis are given, among others, by Bishop *et al.* (1975) and Agresti (2002); methods for the analysis of ordinal categorical data are extensively addressed by Agresti (1984).

In many situations it may happen that the number of cells may be close to, or even greater than, the number of the available observations resulting in very small or even zero cell counts. In this case a contingency table is usually referred to as a sparse table. In such a situation the usual statistical procedures may lose the optimal properties they have for large samples.

In a hypothesis test framework the asymptotic inference for tests of goodness of fit and of multidimensional association is often unreliable for sparse data (Agresti and Yang, 1987, Contini and Lovison, 1993, Haberman, 1977). Therefore many authors propose to base the inference on an exact conditional distribution. The inference is based on the distribution of the sufficient statistics for the parameters of interest given the sufficient statistics for the nuisance parameters which do not depend on these parameters. A review of exact inference methods for contingency tables is given by Agresti (1992). The case of contingency tables with ordered categories in particular is considered in Agresti *et al.* (1992). Forster *et al.* (1996, 2003) propose a MCMC approach to sample the exact conditional distribution in the case of high-dimensional multivariate distributions. Simonoff (1986) and Van Davier (1997) investigate the possibility of using jacknife and bootstrap tests in a sparse context.

On the estimation side the inference faces relevant problems as well. It is well known, for instance, that the Maximum Likelihood Estimator of the probability mass function under a multinomial sampling is the empirical distribution, that is the ratio n_i/n where n_i is the number of occurrences in the generic cell i and n the sample size. The Law of Large Numbers states that this is a consistent estimator of the true probability, even in a strong sense, as long as both n_i and n tend to infinity. This framework however does not fit the evidence as we are in the presence of small frequencies. Bishop *et al.* (1975) introduced the idea of sparse asymptotic to give a more realistic asymptotic framework for situations like this. The sparse asymptotic properties of an estimator are investigated assuming that the number of cells, say K, goes to infinity with n. Hall and Titterington (1987) derived the optimal convergence rate showing that MLE does not achieve this rate under some hypotheses on the way in which K goes to infinity as a function of n.

In order to get around the problem of sparseness in the case of ordinal categorical data, some authors suggested the use of non parametric estimators. Simonoff (1983) proposes a penalised likelihood estimator for sparse categorical data. Hall and Titterington (1987) introduce a kernel estimator for multinomial count showing its optimality in a sparse asymptotic framework. Aerts *et al.* (1997) use a local polynomial approach for estimating the probabilities of a sparse table. Simonoff (1996) reviews many of the non-parametric methods for sparse contingency tables. In a Bayesian framework Giudici (1998) proposes a graphical model approach for smoothing a sparse contingency and Tarantola and Dalla Portas (2001) suggest reducing the table size to cope with sparseness by merging adjacent cells in such a way that the underling conditional independence structure is preserved.

This paper focuses on the analysis of contingency tables with ordered categories in a sparse context. In particular, being aware of the advantages provided by smoothing procedures in estimating the cell probabilities under sparseness condition, the capability of a two-step smoothing technique in addressing the multivariate structure of multidimensional data is investigated. Information measures are used to characterise this structure and quantify the discrepancy between models.

Section 2 reviews some main ideas of the information theory approach to the analysis of contingency tables and some information measures. In section 3 a two-step algorithm for estimating the mass probability function is introduced whilst section 4 contains some simulation results on its performance.

2. INFORMATION-BASED MODELS

For a long time information theory has achieved an important role in the analysis of contingency tables. A cornerstone in the information theory approach to statistical inference is the book of Kullback (1968). Ghokale and Kullback (1978) discuss in detail the case of contingency table analysis whilst Krippendorff (1986) embeds it in a structural modelling framework for qualitative data. Borgoni *et al.* (1998) provide an application of this approach to the analysis of longitudinal categorical data.

Measures of information, like entropy or cross entropy, are mostly used in statistics to quantify the variability of categorical and nominal variables. Non-parametric information measures (Papaioannou, 1985) have been often used to quantify the amount of information in the data explained or ignored by a model allowing comparisons among alternative models.

The measure we consider hereafter is well known, namely the Kullback and Leibler divergence function (KL) together with some indices derived from it.

Assuming that two probability measures μ and π on a given probability space $(\chi,\Im)$ are absolutely continuous with respect to one another, the Kullback and Leibler information measure is defined as:

$$I(\mu,\pi)=\int_{\chi}\log\frac{f(x)}{g(x)}d\mu(x)$$

where $f(x)$ and $g(x)$ are the Radom-Nikodym derivatives of μ and π with respect to an absolutely continuous measure. A number of example supporting the term "*information*" for the above function are given by Kullback (1968). They relate for instance to the possibility of interpreting it, or some suitable specification of it, as the mean information provided by one observation x to discriminate about a set of mutually exclusive and exhaustive hypotheses or as a measure of the relation between the transmitted and the received signal through a transmission channel.

In the case of two mass probability functions on the same finite support of cardinality r, $\boldsymbol{\mu}=(\mu_1,\cdots\mu_r)'$ and $\boldsymbol{\pi}=(\pi_1,\cdots\pi_r)'$ where μ_i and π_i i=1,...,r, are the probabilities of the i-th category, the KL divergence specifies in

$$I(\boldsymbol{\mu},\boldsymbol{\pi})=\sum_{i=1}^{r}\mu_i\log(\mu_i/\pi_i),\quad \text{assuming by definition}\quad 0\log(0/0)=0.$$

Although not being a real metric in a topological sense (Csiszàr, 1995) the KL divergence has proven to own relevant geometric properties which make it suitable for comparing probability functions and hence for comparing alternative models.

The models considered in this paper are hierarchical models. Each model is defined in terms of a set of s components K_e e=1,...,s. A component represents a subset of the variables included in the model and a hierarchical model is such that it always includes lower order components contained in a higher order component of the model. A component which is a singleton is called a main effect otherwise it is called an interaction. An interaction represents what is unique to a set of variables and not reducible to any of its subsets. In terms of the notation used in the paper, the components are separated from each other by ":", for example given a random vector (X_1, X_2, X_3) a main effect model is denoted as X_1:X_2:X_3. The way in which such components interact with each other depends on the shared variables. The notation introduced so far is equivalent to specifying a log-linear model (Agresti, 2002) for the considered contingency table. As in the case of log linear models, each component identifies a marginal distribution. The joint probability distribution on the whole space conforms to each marginal distribution and the shared variables identify relationships of conditional independence. Components which do not share any variable represent therefore a relationship of marginal independence. The model consisting of all possible interactions is called *saturated,* and it will be denoted by m_0 in what follows, whilst the independent model, m_{ind}, excludes all the interactions between the considered variables.

A model m_j is called a *descendent* of (or *nested* in) another model m_i if each interaction of the first is included in the second. The hierarchical relation between the two models is denoted as $m_i \rightarrow m_j$. Two models are incompatible if they are not nested. It can be proved that "→" is a relation of partial order on the set $M=\{m_i, i=1, ...\}$ of the models covering a given space and therefore the pair $(M,\rightarrow)$ identifies a lattice. The possibility of identifying such an order provides us with the opportunity to define paths on the lattice and therefore optimal searching algorithms may be implemented.

Assuming π^i and π^0 be the probability r-vectors associated with a model m_i and the saturated model respectively, $I(m_0 \rightarrow m_i)$ is meant to represent the KL divergence between the model m_i and the saturated one and it can be naturally seen as a measure of goodness of fit for m_i once it is calculated on the sample estimates.

In order to compare any pair of descendent models m_i and m_j such that $m_i \rightarrow m_j$, the KL divergence can be extended as

$$I(m_i \rightarrow m_j) = \sum_{h=1}^{r} \pi_h^0 \log(\pi_h^i / \pi_h^j) \quad \text{with} \quad 0\log(0/0) = 0\,.$$

The following additive properties hold:

$$I(m_0 \rightarrow m_{ind}) = I(m_0 \rightarrow m_i) + I(m_i \rightarrow m_{ind}) \tag{1}$$

$$I(m_0 \rightarrow m_{ind}) = I(m_0 \rightarrow m_i) + I(m_i \rightarrow m_j) + I(m_j \rightarrow m_{ind}) \tag{2}$$

In particular the first equation splits the maximal distance in the lattice of the considered models in two components: $I(m_i \rightarrow m_{ind})$ is the amount of divergence explained by a model m_i and $I(m_0 \rightarrow m_i)$ is what the model ignores. Given (1) the following indexes:

$$I(m_0 \to m_i)/I(m_0 \to m_{ind}) \tag{3}$$

$$I(m_i \to m_{ind})/I(m_0 \to m_{ind}) \tag{4}$$

represent the proportion of the ignored and explained information. They take value 1 when all the information is ignored by m_i and when m_i explains all the information present in the data respectively.

3. A TWO-STEP PROCEDURE FOR SMOOTHING SPARSE TABLES

In the contingency tables analysis, the problem of the estimation of the multivariate distribution associated to a given model of association is usually addressed via maximum likelihood. However a number of different approaches are proposed in the literature. A well known method is the Maximum Entropy Principle (MPE). According to the MEP, given the set of constraints required by a model, the estimate of the probability law is obtained by maximising a suitable entropy function. A measure often proposed in the literature is the Kullback and Leibler divergence function. The resulting estimators are often referred as Minimum Discriminant Information (MDI) and this approach can be included in the wider class of the Minimum Distance Estimators.

The constraints[1] on the probability distribution (sometime called *information constraints*) are expressed in terms of expected values of appropriate transformations of a set of random variables (Soofi, 1994) and, in the case of contingency tables they can be stated as linear functions of the cell probabilities.

The optimality of the resulting estimators is investigated, among others, by Darroch and Ratcliff (1972). The authors also provide an iterative algorithm called Generalised Iterative Scaling (GIS) for solving the optimum problems of the MDI estimation. Most of the investigated properties, however, are large sample properties and therefore not suitable for a sparse context.

In this section a two-step procedure is introduced. In the first step the probability distribution is esteemed via a non-parametric technique. In particular a kernel-type smoother is used. In the second step the output of the first stage is used in a MDI paradigm, that is to say that the smoothed table is taken as the input for the GIS algorithm which produces the final smoothing according to an assumed model.

Being aware of the improvement in the estimates due to the smoothing step, the aim is to assess whether or not this improvement mirrors in more accurate estimates of the interaction structure of the involved variables and hence in a more powerful tool to detect the underlying multivariate structure of the data.

[1] According to Ghokale and Kullaback (1978) the constraints are called internal if the linear functions are defined in terms of a set of marginal distributions of the observed table, and that are the ones considered in the present paper, and external if the constraints are not defined in terms of the data.

It can be observed that for non-sparse tables the smoothed estimates are proved to be very close to the relative frequency therefore it does not matter whether or not they are used.

In the remaining part of this section the two steps of the estimation procedure are described in detail.

The First step

The first step of the procedure gives a smoothed version of the contingency table. In particular a kernel smoother is used. Kernel estimators for discrete probability distributions adapt kernel estimators for densities to the discrete case (Simonoff, 1996). Assuming that the probabilities associated to adjacent cells are similar, the idea of the estimator is borrowing strength from neighbouring cells in order to improve the estimates of less frequented categories.

Assuming $\boldsymbol{S}=(X_1, \quad \cdots \quad X_n)'$ being a sample drawn from a probability mass vector $\mathbf{p}$ of r components, the kernel estimator of $p_i=p(i)$ is defined as (Hall and Titterington, 1987)

$$\tilde{p}(i)=\tilde{p}(i|h,\boldsymbol{S})=\frac{h}{n}\sum_{j=1}^{n}W_h\{h(i-X_j)\}=\frac{h}{n}\sum_{l=1}^{r}n_l W_h\{h(i-l)\},$$

where h is the smoothing parameter, $W_h(x)$ is a kernel function and n_l the frequency of cell l, l=1,...,r. Under some regularity conditions this estimator has good properties both under standard and sparse asymptotic conditions (Bowman, Hall and Titterington, 1984).

Assuming $\overline{\mathbf{p}}$ being an estimator of the probability mass function $\mathbf{p}$ and defining the Mean Summed Squared Error (MSSE) of $\overline{\mathbf{p}}$ as $E\left\{\sum_{i=1}^{r}(\overline{p}_i-p_i)^2\right\}$, Hall and Titterington (1987) derived the optimal convergence rate in terms of the MSSE under sparse asymptotic conditions for any estimator $\overline{\mathbf{p}}$ of $\mathbf{p}$. Specifically assuming that the vector $\mathbf{p}$ is generated by an underlying density function $f(x)$ with s-bounded continuous derivatives on a compact support through the relations $p_i=\int_{(i-1)/r}^{i/r}f(u)du$, they proved the optimal rate to be $O(n^{-2s/(2s+1)}\delta)$ if $n^{-1/(2s+1)}\delta\rightarrow 0$ as $n\rightarrow\infty$, $\delta=\delta_n$ being a sequence such that $\delta\rightarrow 0$ as $n\rightarrow\infty$. Moreover they showed that the kernel estimator achieves this rate.

It has been observed that kernel convolution smoothers have difficulties at and near the edges if the estimation is attempted over a region with known boundaries and are particularly biased in this part of the support even when the estimation of the density of an absolutely continuous random variable is of concern (Jones, 1993). To face this sort of bias, kernel estimators are proposed to be corrected on the boundaries. For categorical data Dong and Simonoff (1994) proposed a boundary correction which consists in replacing the kernel function on the cells near the boundary with another suitable kernel (i.e. a function which satisfies the so called second order conditions on this part of the range). This boundary corrected version is used in this paper.

The data sparseness problem occurs more heavily in a multivariate framework. The kernel estimators can be generalised to the multidimensional context in rather a straightforward way (Grund, 1993). Being **i** a *d*-vector of indexes identifying a cell of a *d*-way table, the kernel estimator of the probability of the cell is defined as

$$\tilde{p}(\mathbf{i}) = \tilde{p}(\mathbf{i}|\mathbf{h},\boldsymbol{S}) = \sum_{\mathbf{l}} \hat{p}_{\mathbf{l}} W_{\mathbf{i}}(\mathbf{l},\mathbf{h}),$$

where $\mathbf{h} \in [0,1]^{\mathrm{d}}$ is the vector of smoothing parameters and *d* is the dimension of the table. As in the univariate case, the kernel function $W_{\mathbf{i}}(\mathbf{l},\mathbf{h})$ weights the probability of each multi-cell **l** in a multivariate neighbourhood of the current smoothed multi-cell **i** where the smoothing window is defined in terms of the parameters **h**. The usual way to define the multidimensional kernel function consists in using a product of univariate kernels where each of them is obtained from a density with a fairly regular compact support. The previous formula does not necessarily imply that the smoothing parameter is the same for each dimension and each component of the vector **h** may take a different value. Dong and Simonoff (1995) generalised the boundary-corrected estimator to the *d*-dimensional case. Although this estimator achieves a good performance in terms of asymptotic properties, it has the drawback of allowing negative estimates of the cell probabilities which is a particularly unattractive feature since a probability less than zero is clearly meaningless. In order to guarantee a positive estimate, Dong and Simonoff (1995) introduced a suitable further correction. Such a correction is based on a geometric combination of kernel estimators defined in terms of a different width. They found the resulting estimator consistent in terms of Summed Squared Error with a rate of convergence $O_p(r^{-1}n^{-8/(d+8)})$ for all *d*. Hereafter we refer to a simple version proposed by Dong and Simonoff which takes the form:

$$p^{\circ}(\boldsymbol{i}\,|\,\boldsymbol{h}) = \tilde{p}(\boldsymbol{i}\,|\,\boldsymbol{h})^{4/3}\,\tilde{p}(\boldsymbol{i}\,|\,2\boldsymbol{h})^{-1/3}.$$

The Second Step

As mentioned above the MDI estimator of a probability mass function can be obtained by minimising a divergence function under a set of linear constraints. In particular the divergence considered here is the KL information function.

Assuming $\boldsymbol{\mu}$ a given probability mass function on a support of cardinality *r* the problem can be formalised as

$$\underset{\boldsymbol{\pi}}{\arg\min}\, I(\boldsymbol{\mu},\boldsymbol{\pi}) \quad \text{given} \quad \sum_{i=1}^{r} b_{s,i}\mu_i = c_s \qquad s = 1,\cdots,v$$

where *v* is the number of constraints.

The fixed probability vector can be chosen on a priori ground. If the uniform distribution is chosen the MDI problem is equivalent to a Maximum Entropy Estimator where the entropy function is the Shannon's entropy.

The solution of the optimum problem is obtained by GIS, which is an iterative algorithm which adjusts, in each iteration, the probabilities estimated in the previous step until a tolerance level is reached (Darroch and Ratcliff, 1972). For acyclic models (models in which any component does not influence itself directly or indirectly and for which the associated probability distribution can be computed algebraically) the algorithm converges since the first iteration. The obtained estimates are maximum likelihood estimates.

In order to assess the goodness of fit of the estimated models, Ghokale and Kullback (1978) recommended using the statistic $G^2 = b \times n \times I(\mathbf{p}^* \to \hat{\mathbf{p}})$, where b is a constant depending on the base of the logarithm used in computing the divergence, $\mathbf{p}^*$ is the MDI estimator under the considered model and $\hat{\mathbf{p}}$ is the observed empirical distribution (the estimate under the saturated model). G^2 tends to a chi-square distribution under standard asymptotic conditions with a number of degrees of freedom equal to the difference between the degrees of freedom associated to each of the two considered nested models, at their turn determined on the basis of the number of constraints imposed by each component[2].

4. A MONTE CARLO SIMULATION STUDY

In order to assess the performance of the proposed estimator, a Monte Carlo experiment was performed. This section describes the details of the simulation study. In particular the simulation design is described in section 4.1. Details on the data generation and on the implemented procedures are reported in section 4.2 and 4.3 respectively. Finally, the main results are discussed in section 4.4.

4.1 *The simulation design*

Each Monte Carlo experiment is structured as follows. A trivariate space of ordinal categorical data is considered and a sparse table with a known association structure is generated (the data generation procedure is described in detail in section 4.2).

The simulation study aims to be explorative in nature. The goal is to assess whether or not the two-step procedure manages to increase the chance of finding out the real structure which is behind the data. At the same time the simulation aims to identify what sort of error is more likely to occur, i.e. whether the algorithm tends to identify a structure which is more or less complex than the one

[2] For the external constraints problem Ghokale and Kullaback (1978) suggested a statistic which reverses the roles of $\mathbf{p}^*$ and $\hat{\mathbf{p}}$. Both these functionals belong to the Power Divergence Statistics (Cressie and Read, 1988) and have a similar asymptotic behaviour. It may also be noticed that for an external constraint problem the MDI estimator is not equal to the maximum likelihood estimator but it has the same asymptotic properties.

which was actually used to generate the data. Ideally this would require applying the two-step procedure to all of the models pertinent to the considered space, and compare the outcomes with the data looking for the best fit. In order to speed up the simulation an automatic top-down search algorithm was implemented which reduces the number of comparisons. This, by the way, mirrors what actually happens in many practical applications when, starting from the observed data, a model, which may be considered optimal in some respects, is sought by the analyst by using an automatic search procedure.

More specifically the table is first smoothed using the kernel-type smoother. Then a backwards search procedure searching for an optimal model is implemented for each of the visited models. It takes the smoothed table as an input. The GIS algorithm is then applied in order to estimate the probability distribution of each visited model (this stage is described in more details in section 4.3). Furthermore for each generated sparse table, a backwards model selection procedure based on standard (asymptotic) maximum likelihood inference is also implemented. The resulting models from the two procedures are stored in a file and the whole process is repeated a number of times.

4.2 *Data generation*

If the multivariate joint law of a random vector can be specified then the random vector can be generated (Johnson, 1987). In many cases however it is difficult to specify such a joint distribution although it is usually possible to specify the marginal distributions and some measures of dependence among the single random variables. In what follows the data generation is worked out by supposing that the levels of each classification factor are realisations of a multivariate Poisson variable with a given vector of means. It is usually assumed that a random vector is multivariate Poisson distributed if all its univariate marginal distributions are Poisson and if each Poisson component correlates to the other according to a given correlation matrix. A method to generate a multivariate Poisson vector based on the self-decomposability property of the Poisson distribution (Steutel and Van Harn, 1979) has been proposed by Sim (1993).

The algorithm used for generating the data in what follows is an efficient way to generate very large contingency tables. The algorithm takes the vector of marginal means and the correlation matrix as an input and gives a vector of random variables whose marginal distributions are Poisson with the given parameters as an output. Specifically let $X = (X_1,...,X_P)'$ be a vector of Poisson variables with vector of means $\boldsymbol{\mu}$ and correlation matrix $\boldsymbol{\Sigma}$ and $Y = (Y_1,..., Y_M)'$ be a vector of independent Poisson variables with vector of parameters $\boldsymbol{\lambda}$. The algorithm computes the vector $\boldsymbol{\lambda}$ satisfying the conditions

$$\mathrm{E}(\mathbf{X}) = \mathrm{E}(\mathbf{TY}) = \mathbf{T}\boldsymbol{\lambda} = \boldsymbol{\mu} \text{ and } \mathrm{Cov}(\mathbf{X}) = \mathrm{Cov}(\mathbf{TY}) = \mathbf{T} \times \mathrm{diag}(\boldsymbol{\lambda}) \times \mathbf{T}' = \boldsymbol{\Sigma}$$

where T is a $P \times M$ incidence matrix (i.e. a matrix of 0's and 1's). The problem of determining T can be led back to a linear programming problem of the form

$$\mathbf{H}\boldsymbol{\lambda} = \mathbf{c} \quad \text{under constraint} \quad \boldsymbol{\lambda} > \mathbf{0}$$

where **H** is a $[P\times(P+1)/2]\times M$ matrix depending on **T** and **Σ** and solved via the simplex algorithm. The generation of multivariate Poisson variables can then be obtained by the transformation **TY** without knowing the probability law of Xs but only its first two moments and the probability law of the Ys components.

Here the correlation structure is seen as a proxy of the interaction structure. For instance the model $X_1X_2{:}X_3$, where X_1 and X_2 are the only two variables which interact each other, is approximated by a correlation matrix corresponding to a marginal correlation between X_1 and X_2 and a null correlation between any other pair of variables i.e. a correlation matrix which takes the form

$$\begin{bmatrix} 1 & \rho & 0 \\ \rho & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

where ρ is a positive value less than 1.

After generating individual records they are cross-tabulated in order to produce the contingency table to be passed as the argument to the estimation procedure.

4.3 *The searching procedure*

In the first step of the estimation procedure a boundary corrected kernel estimator is computed using a product kernel of Epanechnikov's univariate functions. The estimator is further corrected through a geometrical combination as described in section 2.

A backwards searching is then applied to the smoothed table. The algorithm used is similar to the one originally proposed by Lin (1982). It consists of two parts.

In each of them the smoothed table is used as an input for the GIS algorithm which gives the final MDI estimate according to two alternative models of association.

Specifically the first part of the Lin's algorithm selects a uniform model[3]. The second part of the algorithm consists in a backwards elimination procedure which deletes those effects not contributing to the fit.

The procedure stops when a model is found which differs from the previous one for an amount of information bigger than a given threshold. Specifically the stop rule is based on the index of residual information (3) introduced in section 2 and stops when such an index gets greater than a given value. The used thresholds for the residual information are 25% to select the uniform model and 15% to evaluate models between two adjacent uniform models[4].

[3] A *uniform model* is defined as a model which includes all interactions of a given order and none interaction of higher order.

[4] The chosen thresholds are suggested in Borgoni (1999) where an extensive simulation analysis on a grid of alternative values was performed.

4.4 *Results*

A first set of simulations based on 500 iterations is run in a very sparse context. A sample of 13.000 records is generated according to two different models: $X_1X_2:X_2X_3$ and $X_1X_2:X_3$. Different values of the correlation coefficients, 0.2, 0.5 and 0.8 are considered in order to investigate the effect of a different degree of interaction among variables.

The vector of the marginal means is (4.5, 4.5, 4.5)'. This produces an average number of categories for each dimension around 16 and an average size of cell frequency of 3.3 (see the last four columns of table 1).

Before looking at the Monte Carlo results in details it could be noted that the degree of association amongst the variables affects, given the sample size, both the number of empty cells and the cell frequency. This is easy to be seen in a rectangular table as the frequencies tend to assume specific patterns, for instance to become diagonal in the case of a positive association, when the interaction between two ordered classification factors gets higher. For a dimension larger than two this cannot be visualised anymore. In order to understand the effect on the structure of the sparseness in a multivariate table due to a gradually higher correlation amongst the marginal variables, a small set of simulations using the algorithm for generating multivariate Poisson data proposed above were run for four different values of the correlation coefficient, ρ=0.2, 0.4, 0.6 and 0.8 (100 tables were generated in each of them). The average rate of empty cells as well as the minimum and the maximum rate in each of the four sets of simulations were computed and plotted versus the value of the correlation coefficient in Figure 1. Also the average of the highest frequency obtained in each of the onehundred generated table is reported there. Clearly the distribution gets more and more concentrated in a fewer number of cells as ρ increases.

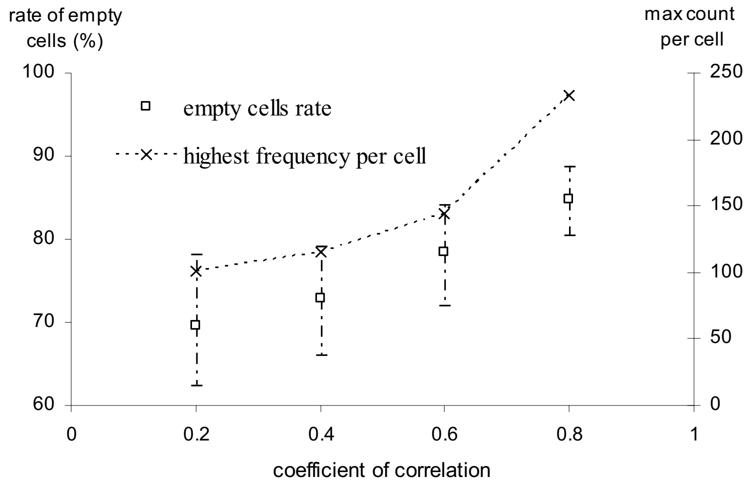

Figure 1 – Average rate of empty cell (%) (horizontal bars represent the maximum and the minimum rate) and average highest cell frequency across 100 simulated tables versus the correlation coefficient.

This exercise underlines that it is worthwhile to check also the role that a different strength of the relation amongst the variables could play as this affects the number of zeros and the structure of the sparseness in a table. This is the reason why a grid of different values of ρ is considered in the Monte Carlo study the results of which are reported in Table 1.

In both the two sets of simulations, i.e. the data are simulated by model $X_1X_2{:}X_2X_3$ and $X_1X_2{:}X_3$, the two-step algorithm proposed here looks to perform quite well always picking the model which actually generates the data out.

On the other hand the search procedure based on the asymptotic distributional properties of the likelihood ratio test tends to introduce spurious effects. In particular this is the truer the higher the correlation amongst the variables. In the case of ρ=0.5, for instance, this procedure always selects the model $X_1X_2{:}X_1X_3{:}X_2X_3$ whilst the actual model behind the data was $X_1X_2{:}X_2X_3$ and basically the same happens when ρ=0.8. Surprisingly the standard procedure is still very conservative (25% percent of the times the uniform model of order two is picked out) also when the association amongst the variables is indeed low (ρ = 0.2).

The bottom part of table 1 presents the results when the data are generated by model $X_1X_2{:}X_3$. As one might expect the performance of the chi-square based procedure improves as the structure of the generating model becomes simpler (i.e. less parameters are involved). This is because the dimension of the appropriate sufficient statistics of the model (that is the marginal distributions of the table) is smaller and therefore those statistics are less likely to be sparse even the whole table may be very sparse. In this case, in fact, it looks like the strength of the relation between the pair of correlated variables affects the output of the search much less than in the previous case. Although it may be observed that still in a number of cases ranging from 7.8% (ρ = 0.8) to 10% (ρ=0.2), the standard asymptotic inference suggests to keep spurious relationships.

A second set of simulations has been run under a less extreme sparseness condition. In this case using the same parameters for the marginal Poisson distributions involved in the data generation mentioned above and the same grid of correlation values, 500 samples of 17000 units were simulated from the model $X_1X_2{:}X_2X_3$ and the procedures described in the previous section[5] applied. Results are reported in table 2.

The findings are analogous to the ones obtained for the simulation run in the previous more extreme sparse case. Also in this case a good performance of the

[5] The whole simulation procedure was implemented by a Fortran code. In the case of the two-step approach the total amount of time for running a slot of 500 iterations ranged from 246 minutes to 300 minutes across the performed simulations on a 1.70GHz Celeron Processor with 260Kb of RAM on a MS Windows 2000 platform. Most of the time was due to the smoothing step. The data generation took a negligible amount of time and there are basically no differences in the computational time due to the size of the generated samples. It could be noted that only a few minutes were necessary to implement the search procedure in the case of no smoothing. In other words when the analysis of a very large table is of interest the amount of time necessary to smooth it has to be taken into account as a possible drawback of this approach.

proposed procedure was found whilst the one based upon the likelihood ratio test still looks very conservative introducing some spurious interactions amongst the variables again even for the case of a small correlation coefficient.

TABLE 1

Simulation results. Sample size: 13000

generating model	ρ	Selected model	2-step procedure		Chi square procedure		mean cell size	average number of categories		
			Num. Iter.	%	Num. Iter.	%		1st dim.	2nd dim.	3rd dim.
$X_1X_2{:}X_2X_3$	0.2	$X_1X_2{:}X_1X_3{:}X_2X_3$			126	25.2	3.3	15.9	15.8	15.8
		$X_1X_2{:}X_2X_3$	500	100	374	74.8				
	0.5	$X_1X_2{:}X_1X_3{:}X_2X_3$			500	100	3.3	15.9	15.9	15.8
		$X_1X_2{:}X_2X_3$	500	100						
	0.8	$X_1X_2{:}X_1X_3{:}X_2X_3$			499	99.8	2.9	15.9	18.4	15.9
		$X_1X_2{:}X_2X_3$	500	100	1	0.2				
$X_1X_2{:}X_3$	0.2	$X_1X_2{:}X_1X_3$			23	4.6	3.3	15.8	15.8	15.8
		$X_1X_2{:}X_1X_3{:}X_2X_3$			4	0.8				
		$X_1X_2{:}X_2X_3$			23	4.6				
		$X_1X_2{:}X_3$	500	100	450	90				
	0.5	$X_1X_2{:}X_1X_3$			22	4.4	3.3	15.9	15.8	15.8
		$X_1X_2{:}X_1X_3{:}X_2X_3$			3	0.6				
		$X_1X_2{:}X_2X_3$			18	3.6				
		$X_1X_2{:}X_3$	500	100	457	91.4				
	0.8	$X_1X_2{:}X_1X_3$			18	3.6	3.3	15.9	15.9	15.8
		$X_1X_2{:}X_1X_3{:}X_2X_3$			2	0.4				
		$X_1X_2{:}X_2X_3$			19	3.8				
		$X_1X_2{:}X_3$	500	100	461	92.2				

A last slot of simulations not reported here has been done using the model $X_1X_2{:}X_3$, for generating samples of size 17000. The two-step procedure acts correctly selecting the generating model. As one might have expected given the smaller dimensionality of the sufficient statistics of the model which generates the tables, the chi square procedure works better than in the case of data generated through the model $X_1X_2{:}X_2X_3$.. A slightly better behaviour than in the simulations reported in table 1 was also found for all of the considered values of the correlation coefficient, but again this is not an unexpected result given the higher sample size.

TABLE 2

Simulation results. Sample size: 17000

Generating model	ρ	Selected Model	2-step procedure		Chi square procedure		Mean cell size	average number of categories		
			Num. Iter.	%	Num. Iter.	%		1st dim.	2nd dim.	3rd dim.
$X_1X_2{:}X_2X_3$	0.2	$X_1X_2{:}X_1X_3{:}X_2X_3$			166	33.2	4.2	16.1	16.0	16.0
		$X_1X_2{:}X_2X_3$	500	100	334	66.8				
	0.5	$X_1X_2{:}X_1X_3{:}X_2X_3$			500	100	4.1	16.1	16.1	16.0
		$X_1X_2{:}X_2X_3$	500	100						
	0.8	$X_1X_2{:}X_1X_3{:}X_2X_3$			500	100	3.6	16.1	18.6	16.1
		$X_1X_2{:}X_2X_3$	500	100						

5. CONCLUSIONS

In this paper the capability of a two-step smoothing technique in addressing the multivariate structure of ordered categorical data is investigated. The proposed procedure seems to work rather well compared to a standard asymptotic technique in the context of sparse data. In particular it seems to work better the more extreme the sparseness is and the more complex the interaction structure behind the data.

Some issues however remain to be investigated.

It is well known that kernel estimator suffers the so called "course of dimensionality" i.e. the need of progressively larger sample size in higher dimensions to achieve comparable accuracy. A consequence is that in "very high dimensions local neighbourhoods are empty and neighbourhoods that are not empty are almost sure not local" (Simonoff, 1996). Therefore this technique could not be suitable for extremely high multidimensional situations. Other smoothing estimators that suffer less this problem may be evaluated in this paradigm.

As underlined by Cressie and Read (1988), the distribution of the test statistics under asymptotic sparse conditions is unknown and hard to be worked out. In this paper a fixed threshold is used. Another possibility to get around the problem could be a computational approach. The bootstrap (Davison and Hinkley, 2000) could be usefully applied to this context. In particular the resampling procedure could take advantage by the smoothing itself and the samples may be drawn from the smoothed table. Sampling from the smoothed distribution instead of from the empirical one is known as "smoothed bootstrap" (Hall *et al.*, 1989, Simonoff, 1996). The smoothed bootstrap in the context of sparse categorical data is used for instance in Borgoni and Provasi (2001). The computational requirement however may become cumbersome in a very high dimensional spaces.

Dipartimento di Statistica RICCARDO BORGONI
Università degli Studi di Milano-Bicocca

REFERENCES

M. AERTS, I. AUGUSTYNS, P. JANSSEN, (1997), *Smoothing sparse multinomial data using local polynomial fitting*. "Journal of Nonparametric Statistics", **8**, pp. 127-147.

A. AGRESTI (1984), *Analysis of ordinal categorical data*, Wiley, New York.

A. AGRESTI (1992), *A survey of exact inference for contingency tables* (with discussion) "Statistical Science", 7, pp. 131-177.

A. AGRESTI (2002), *Categorical data analysis* (2nd edition), Wiley, New York.

A. AGRESTI, C.R. MEHTA, N.R. PATIL (1992), *Exact inference for contingency tables with ordered categories*. "Journal of the American Statistical Association" 85, 410, pp. 453-458.

A. AGRESTI, M.C. YANG (1987), *An empirical investigation of some effects of sparseness in contingency tables*, "Computational Stat. & Data Analysis", 5, pp. 9-21.

Y.M.M. BISHOP, S.E. FIEMBERG, P.W. HOLLAND (1975), *Discrete multivariate analysis*. MIT Press, Cambridge.

A.W. BOWMAN, P. HALL, D.M. TITTERINGTON (1984), *Cross validation in nonparametric estimation of probabilities and probabilities density*, "Biometrika" 71, pp. 341-51.

R. BORGONI (1998), *Modelli strutturali basati sull'entropia per l'analisi di tabelle di contingenza sparse*, unpublished doctoral thesis, University of Padua.

R. BORGONI, P. PALMITESTA, C. PROVASI (1998), *Multivariate nonparametric methods based on information theory for the analysis of longitudinal categoric data*, "Metron", LVI. (1-2), pp. 189-203.

R. BORGONI, C. PROVASI (2001), *Nonparametric estimation methods for sparse contingency tables*, S. BORRA, R. ROCCI, M. VICHI, M. SCHADER (eds), *Advances in classification and data analysis*, Springer, pp. 277-282.

D. CONTINI, G. LOVISON (1993), *The effect of marginal disuniformity on the χ^2 approximation to the distribution of Pearson's X^2 in sparse contingency tables.* "Computational Statistics and Data Analysis", 16, 2, pp. 185-199.

N. CRESSIE, T.C. READ (1988), *Goodness-of-fit statistics for discrete multivariate data*, Springer, Berlin.

I. CSISZÀR, (1975), *I-divergence geometry of probability distributions and minimization problems*, "Annals of Probability", 3, pp. 146-158.

A.C. DAVISON, D.V. HINKLEY (1999), *Bootstrap methods and their application*, Cambridge University Press.

J.N. DARROCH, D. RATCLIFF (1972), *Generalised iterative scaling for loglinear models*, "The Annals of Mathematical Statistics", 43, pp. 1470-1480.

J. DONG, J.S. SIMONOFF (1994), *The construction and properties of boundary kernels for smoothing sparse multinomials,* "Journal of Computational and Graphical Statistics", 3, pp. 57-66.

J. DONG, J.S. SIMONOFF (1995), *A geometric combination estimator for d-dimensional ordinal sparse contingency tables*, "The Annals of Statistics"*,* 23, pp. 1143-1159.

P. GIUDICI (1998), *Smoothing sparse contingency tables: a graphical Bayesian approach*, "Metron", LVI, 1-2.

D.V. GOKHALE, S. KULLBACK (1978), *The information in contingency table*, Marcel Dekker, New York.

J.J. FORSTER, J. W. MCDONALD, P.W.F. SMITH (1996), *Monte Carlo exact conditional tests for log-linear and logistic models.* "Journal of the Royal Statistical Society", Series B, 58, pp. 445-453.

J.J. FORSTER, J. W. MCDONALD, P.W.F. SMITH (2002), *Markov chain Monte Carlo exact inference for binary and multinomial logistic regression models.* "Statistics and Computing", to appear.

S.J. HABERMAN (1977), *Log-linear models and frequency tables with small expected cell counts*, "Annals of Statistics", 5, pp. 1148-1169.

P. HALL, T.J. DI CICCIO, J.P. ROMANO (1989), *On smoothing and the bootstrap*, "The Annals of Statistics", 17, 2, pp. 692-704.

P. HALL, D.M. TITTERINGTON (1987), *On smoothing sparse multinomial data*, "Australian Journal of Statistics", 39, pp. 19-37.

M.E. JOHNSON (1987), *Multivariate statistical simulation.* Wiley, New York.

M.C. JONES (1993), *Simple boundary correction for kernel density estimation,* "Statistics and Computing", 3, pp. 135-146.

S. KULLBACK, (1968), *Information theory and statistics,* 2nd edition, New York: Dover Books.

K. KRIPPENDORFF, (1986), *Information theory: structural models for qualitative data.* Sage Pub, Beverly Hills.

S.P. LIN (1982), *Automatic model selection in contingency tables*, "Applied Statistics", 31, pp. 317-326.

T. PAPAIOANNOU (1985), *Measures of information*, "Encyclopaedia of Statistical Sciences", 5, pp. 391-397.

C.H. SIM (1993), *Generation of poisson and gamma random variate vectors with given marginals and covariance matrix*, "Journal of Statistical Computation and Simulation", 47, pp. 1-10.

J.S. SIMONOFF (1983), *A penalty function approach to smoothing large sparse contingency tables*, "Annals of Statistics", 11, pp. 208-218.

J.S. SIMONOFF (1986), *Jackknifing and boostrapping goodness of fit statistics in parse multinomials*, "Journal of the American Statistical Association", 81, pp. 1005-11.

J.S. SIMONOFF (1995), *Smoothing categorical data*, "Journal of Statistical Planning and Inference", 47, pp. 41-69.

J.S. SIMONOFF (1996), *Smoothing methods in statistics*, Springer Berlin.

E.S. SOOFI (1994), *Capturing the intangible concept of information*, "Journal of American Statistical Association", 89, pp. 1243-1254.

F.W. STEUTEL, K. VAN HARN (1979), *Discrete analogous of self-decomposability and stability*, "Annals of Probability", 7, pp. 893-99.

C. TARANTOLA, P. DALLAPORTAS (2001), *Statistical methods for the analysis of contingency tables by using conditional independence relations*, C. PROVASI, (ed) *Modelli complessi e metodi computazionali intensive per la stima e la previsione*, pp. 169-174.

M. VON DAVIER (1997), *Bootstrapping goodness-of-fit statistics for sparse categorical data. Results of a Monte Carlo study*, "Methods of Psychological Research Online", http:/www.mpr-online.de/, 2, pp. 29-48.

RIASSUNTO

Una procedura di lisciamento a due passi per l'analisi di tabelle di contingenza sparse con marginali ordinati

Nell'analisi statistica di fenomeni economici, demografici o sociali l'interesse è spesso rivolto all'individuazione della struttura multivariata che soggiace al fenomeno osservato. In contesti multidimensionali spesso può accadere che il numero delle celle presenti nella tabulazione congiunta di più variabili categoriali sia molto elevato rispetto alla numerosità campionaria producendo una frequenza media di cella bassa o perfino nulla. Diversi autori hanno proposto metodi basati su tecniche di lisciamento per analizzare dati categoriali in condizioni di sparsità delle osservazioni, ma poco è stato fatto per valutare se tali tecniche possono essere d'aiuto nell'individuare la struttura multivariate dei dati. Il presente lavoro mostra come metodi di lisciamento, combinati con opportune misure sviluppate nell'ambito della teoria dell'informazione, possono fornire vantaggi nell'analisi di dati categoriali caratterizzati da sparsità.

SUMMARY

A two-step smoothing procedure for the analysis of sparse contingency tables with ordered categories

Assessing the multivariate structure of data is often the aim of the statistical analysis of economical, demographic and social phenomena. In many situations in the analysis of categorical data it may happen that the number of cells can be close to, or even greater than, the number of observations at hand resulting in very small or even zero cell counts. In this case a contingency table is usually referred to as a sparse table. In this sort of situation the optimal properties of the usual statistical procedures may break down. Several authors investigated the use of smoothing methods for sparse count data but a little was done to evaluate if these methods can be helpful in discovering the multivariate structure of the data. This paper shows as the joint use of smoothing techniques and information measures may improve the analysis in a multivariate sparse context.