EXACT CONDITIONAL TESTING OF CERTAIN FORMS OF POSITIVE ASSOCIATION FOR BIVARIATE ORDINAL DATA

Francesco Bartolucci[1] and Luisa Scaccia[2]

[1] Dipartimento di Economia, Finanza e Statistica
Università di Perugia
Via A. Pascoli, 20 - 06123 Perugia
(e-mail: bart@stat.unipg.it)

[2] Dipartimento di Istituzioni Economiche e Finanziarie
Università di Macerata
Via Crescimbeni, 20 - 62100 Macerata
(e-mail: scaccia@unimc.it)

ABSTRACT. We describe an exact conditional approach to test for certain forms of positive association between two ordinal variables. The approach is based on maximizing a conditional version of the multinomial likelihood for the observed table given the row and column margins. This allows us to remove the uncertainty that typically arises in testing hypotheses on the association between two categorical variables due to the presence of nuisance parameters corresponding to the marginal distributions of the two variables. Conditional maximum likelihood estimates of the parameters are obtained through Markov chain Monte Carlo methods. The Pearson's chi-squared is used as test statistic. A p-value for this statistic is computed by simulation, when data are sparse, or by exploiting the asymptotic theory.

1 INTRODUCTION

In a recent paper, Bartolucci *et al.* (2001) proposed a general framework for fitting models for bivariate ordinal data incorporating certain forms of positive association, such as *Positive quadrant dependence* (PQD) and *Total positivity of order 2* (TP_2). They also showed that the deviance of any of these models with respect to the saturated model has null asymptotic distribution which belongs to the *chi-bar squared* family (Shapiro, 1988). This distribution may be used to perform a statistical test for the form of positive association of interest. A difficulty in performing this kind of test is that such a distribution depends on the marginal distributions of the two variables and this gives uncertainty to the result of the testing procedure. The problem may be overcome by conditioning the inference on the observed margins which represent sufficient statistics for the nuisance parameters describing the two marginal distributions. This approach is well-known in categorical data analysis and goes back to Fisher (1934, 1935).

A first attempt to implement a conditional approach to test for a certain form of positive association was already made by Bartolucci *et al.* (2001). However, they noticed that the distribution resulting from conditioning on the observed margins, known as *multivariate generalized hypergeometric*, is intractable whenever the sample size or the dimension of the table are moderately large. This is because computing the probability of a certain table requires enumerating all the possible tables with the same margins. Thus, they suggested an approximation based on maximizing the product multinomial likelihood, that derives from conditioning only on the row margin, under the constraint that the marginal distribution of the column variable is equal to the observed one. Another solution, which may be used regardless of the sample size and the dimension of the table, was proposed by Bartolucci and Scaccia (2004). In this paper we summarize this approach giving result to its main aspects.

In the approach of Bartolucci and Scaccia (2004), the parameters of the model incorporating a certain form of positive association are estimated by means of a *Monte Carlo Maximum Likelihood* algorithm (Geyer, 1991). In summary, this algorithm consists of maximizing a likelihood ratio, with respect to a fixed point of the parameter space, through the joint use of the importance sampling (Hammersley and Handscomb, 1964) and the Metropolis-Hastings algorithm (Metropolis *et al.*, 1953; Hastings, 1970). The maximization is performed by a constrained Fisher-scoring algorithm similar to that described in Bartolucci *et al.* (2001). The Pearson's chi-squared is used as discrepancy measure between nested models since, in this context, has proved to perform better than the likelihood ratio test. A p-value for this statistic is computed through Monte Carlo simulations, when the observed table is sparse, or by exploiting the asymptotic theory.

The paper is organized as follows. In Section 2 we describe a class of models for testing positive association between two ordinal variables. Conditional inference for these models is described in Sections 3 and 4. Finally, an application is illustrated in Section 5.

2 MODELS OF POSITIVE ASSOCIATION

Let A and B be two categorical variables with I and J categories, respectively; let p_{ij} be the joint probability of the ith category of A and the jth category of B and $\mathbf{p}$ be the IJ-dimensional column vector with elements p_{ij} arranged in appropriate order. Douglas *et al.* (1991) showed that 16 different types of log-odds ratio may be used to describe the association between A and B. These correspond to the logits used for the marginal distribution of A and those used for the marginal distribution of B which may be of type: *local* (l), *continuation* (c), *reverse continuation* (r) or *global* (g); see Bartolucci *et al.* (2001) for further details. For example, when local logits are used for A and global logits are used for B, *local-global* log-odds ratios result; these are defined as

$$\eta_{ij} = \log \frac{(\sum_{k\leq j} p_{ik})(\sum_{k>j} p_{i+1,k})}{(\sum_{k>j} p_{ik})(\sum_{k\leq j} p_{i+1,k})}, \qquad i = 1,\ldots,I-1, \quad j = 1,\ldots,J-1.$$

Different types of log-odds ratio determine different forms of positive association expressed through the constraint $\eta_{ij} \geq 0\ \forall i, j$. For instance, TP_2 may be expressed by using local log-odds ratios, while global log-odds ratios determine PQD, which is less stringent. Regardless of the type of log-odds ratios, independence corresponds to the constraint $\eta_{ij} = 0, \forall i, j$.

An interesting feature is that the vector $\boldsymbol{\eta}$ with elements η_{ij} arranged in appropriate order may be simply expressed as

$$\boldsymbol{\eta} = \mathbf{K} \log(\mathbf{M}\mathbf{p}), \tag{1}$$

where $\mathbf{K}$ is a matrix of contrasts and $\mathbf{M}$ is a marginalization matrix; for a detailed description see Bartolucci *et al.* (2001). Then, positive association may be expressed as $\boldsymbol{\eta} \geq \mathbf{0}$. However, further linear equality and/or inequality constraints on the log-odds ratios may be of interest. Therefore, we deal with a general class of models expressed as $\mathbf{C}\boldsymbol{\eta} = \mathbf{0}$, $\mathbf{D}\boldsymbol{\eta} \geq \mathbf{0}$.

3 CONDITIONAL MAXIMUM LIKELIHOOD ESTIMATION

For an observed $I \times J$ contingency table, let x_{ij} be the frequency in the (ij)th cell, $\mathbf{x}$ be the IJ-dimensional column vector with elements x_{ij} arranged as in $\mathbf{p}$ and $n = \sum_i \sum_j x_{ij}$ be the sample size. If we assume that the (unconditional) distribution of $\mathbf{x}$ is multinomial, the conditional distribution of $\mathbf{x}$, given the row and column margins, is *multivariate generalized hypergeometric*. Under this distribution, the probability of a table with frequencies $\mathbf{x}$ is

$$\pi(\mathbf{y};\boldsymbol{\theta}) = m(\mathbf{y}) \exp(\mathbf{y}'\boldsymbol{\theta})/c(\boldsymbol{\theta}),$$

where the vector $\mathbf{y}$ is obtained by removing from $\mathbf{x}$ the elements corresponding to the last row and the last column of the contingency table, $\boldsymbol{\theta}$ is the vector of canonical parameters $\theta_{ij} = \log\left[p_{ij}p_{IJ}/(p_{iJ}p_{Ij})\right]$, $i = 1,\ldots,I-1$, $j = 1,\ldots,J-1$, and $m(\mathbf{y})$is the multinomial factor. Finally, $c(\boldsymbol{\theta})$ is the normalizing constant given by the sum of $m(\mathbf{y})\exp(\mathbf{y}'\boldsymbol{\theta})$ over the set $\mathcal{Y}$ of all the possible frequency vectors $\mathbf{y}$ corresponding to tables with row and column margins equal to those of the observed one.

In most cases, we cannot compute exactly $c(\boldsymbol{\theta})$ and, therefore, the maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$ under the constraint $\boldsymbol{\eta} \in \mathcal{H}$, with $\mathcal{H} = \{\boldsymbol{\eta} : \mathbf{C}\boldsymbol{\eta} = \mathbf{0}, \mathbf{D}\boldsymbol{\eta} \geq \mathbf{0}\}$, where $\boldsymbol{\eta}$ is a vector of generalized log-odds ratios defined as in (1). However, we can compute this MLE by maximizing a suitable estimate of the likelihood ratio $\lambda(\boldsymbol{\theta};\mathbf{y}) = \log\left[\pi(\mathbf{y};\boldsymbol{\theta})/\pi(\mathbf{y};\tilde{\boldsymbol{\theta}})\right]$, were $\tilde{\boldsymbol{\theta}}$ is a fixed parameter vector appropriately chosen (Gelman and Meng, 1998). This estimate is obtained by a Monte Carlo technique known as *importance sampling* (Hammersley and Handscomb, 1964). In fact, $\lambda(\boldsymbol{\theta};\mathbf{y}) = \mathbf{y}'(\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}}) - \log[c(\boldsymbol{\theta})/c(\tilde{\boldsymbol{\theta}})]$, where $c(\boldsymbol{\theta})/c(\tilde{\boldsymbol{\theta}})$ is equal to the expected value, with respect to $\pi(\mathbf{y};\tilde{\boldsymbol{\theta}})$, of $\exp[\mathbf{y}'(\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}})]$. Then, this ratio may be estimated as $\frac{1}{T}\sum_{t=1}^{T} u_t$, where $u_t = \exp[\mathbf{y}_t'(\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}})]$ and $\mathbf{y}_1,\ldots,\mathbf{y}_T$ are T frequency vectors drawn from $\pi(\mathbf{y};\tilde{\boldsymbol{\theta}})$ by using the Markov chain Monte Carlo (MCMC) algorithm of Diaconis and Sturmfels (1998). Then, $\lambda(\boldsymbol{\theta};\mathbf{y})$ may be estimated as

$$\tilde{\lambda}(\boldsymbol{\theta};\mathbf{y}) = \mathbf{y}'(\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}}) - \log\frac{1}{T}\sum_{t=1}^{T} u_t,$$

and maximized, under the constraint $\boldsymbol{\eta} \in \mathcal{H}$, by means of a constrained version of the Fisher-scoring algorithm. At the $(s+1)$th step, this algorithm solves the problem

$$\min_{\boldsymbol{\eta}\in\mathcal{H}} (\tilde{\mathbf{v}}^s - \boldsymbol{\eta})'\tilde{\mathbf{F}}(\boldsymbol{\eta}^s)(\tilde{\mathbf{v}}^s - \boldsymbol{\eta}),$$

where $\tilde{\mathbf{v}}^s = \boldsymbol{\eta}^s + \tilde{\mathbf{F}}(\boldsymbol{\eta}^s)^{-1}\tilde{\mathbf{s}}(\boldsymbol{\eta}^s)$, with $\boldsymbol{\eta}^s$ denoting the estimate obtained at the end of the sth step, $\tilde{\mathbf{s}}(\boldsymbol{\eta})$ denoting the estimated score vector and $\tilde{\mathbf{F}}(\boldsymbol{\eta})$ denoting the estimated information matrix with respect to $\boldsymbol{\eta}$. These are given by $\tilde{\mathbf{s}}(\boldsymbol{\eta}) = \mathbf{J}(\boldsymbol{\eta})'[\mathbf{y} - \tilde{\boldsymbol{\mu}}(\boldsymbol{\theta})]$ and $\tilde{\mathbf{F}}(\boldsymbol{\eta}) = \mathbf{J}(\boldsymbol{\eta})'\tilde{\mathbf{V}}(\boldsymbol{\theta})\mathbf{J}(\boldsymbol{\eta})$, with $\mathbf{J}(\boldsymbol{\eta})$ being the Jacobian of the transformation from $\boldsymbol{\eta}$ to $\boldsymbol{\theta}$ and

$$\tilde{\boldsymbol{\mu}}(\boldsymbol{\theta}) = \sum_{t=1}^{T} w_t \mathbf{y}_t \quad \text{and} \quad \tilde{\mathbf{V}}(\boldsymbol{\theta}) = \sum_{t=1}^{T} w_t \mathbf{y}_t \mathbf{y}_t' - \tilde{\boldsymbol{\mu}}(\boldsymbol{\theta})\tilde{\boldsymbol{\mu}}(\boldsymbol{\theta})',$$

where $w_t = u_t/\sum_t u_t$, being the estimates of, respectively, the conditional expected value and the variance of $\mathbf{y}$ based on the same MCMC algorithm mentioned above. Note that under independence, i.e. when $\boldsymbol{\theta} = \mathbf{0}$, we may exactly compute the probability of a certain table and the moments of the hypergeometric distribution.

The above algorithm requires a preliminary choice of $\tilde{\boldsymbol{\theta}}$. A sensible strategy is to set $\tilde{\boldsymbol{\theta}}$ equal to the canonical parameter vector corresponding to $\tilde{\boldsymbol{\eta}}$, the unconditional estimate of $\boldsymbol{\eta}$ under the constraint $\boldsymbol{\eta} \in \mathcal{H}$ and the further constraint that the marginal probabilities are equal to the observed relative frequencies; $\tilde{\boldsymbol{\eta}}$ may again be computed by using a constrained Fisher-scoring algorithm. We also use $\tilde{\boldsymbol{\eta}}$ as starting value for $\boldsymbol{\eta}$ in the maximization algorithm.

Finally, note that the algorithm may become unstable when the observed table contains null frequencies. We suggest replacing these frequencies with a negligible value (e.g. 10^{-6}), but using a larger value in computing $\tilde{\boldsymbol{\eta}}$. This allows us to reduce the risk that the corresponding cells of the tables sampled from $\pi(\mathbf{y}; \tilde{\boldsymbol{\theta}})$ are always null.

4 HYPOTHESIS TESTING

Let S denote the saturated model, H the model formulated by $\boldsymbol{\eta} \in \mathcal{H}$ and H_0 that formulated by $\boldsymbol{\eta} \in \mathcal{H}_0 = \{\boldsymbol{\eta} : \mathbf{C}\boldsymbol{\eta} = \mathbf{0}, \mathbf{D}\boldsymbol{\eta} = \mathbf{0}\}$, i.e. by turning all the inequality into equality constraints. Obviously, H_0 is nested in H, which, in turn, is nested in S. Let $\dot{\boldsymbol{\eta}}$, $\hat{\boldsymbol{\eta}}$ and $\bar{\boldsymbol{\eta}}$ be the MLE computed by using the algorithm in Section 3 under the models S, H and H_0, respectively, and $\dot{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\theta}}$ and $\bar{\boldsymbol{\theta}}$ be the corresponding values of the canonical parameter vector. For measuring the discrepancy between nested models, we make use of the Pearson's chi-squared statistic, X^2, which, in the present context, is more reliable than the likelihood ratio statistic; see the discussion in Bartolucci and Scaccia (2004). When referred to models S and H_0, this statistic may be expressed as

$$X^2 = (\mathbf{x} - \bar{\boldsymbol{\nu}})' \text{diag}(\bar{\boldsymbol{\nu}})^{-1} (\mathbf{x} - \bar{\boldsymbol{\nu}}),$$

where $\bar{\boldsymbol{\nu}}$ is the conditional expected value of $\mathbf{x}$ corresponding to $\bar{\boldsymbol{\eta}}$. Note that X^2 may be expressed as the sum of two components. The first one, $X_1^2 = (\mathbf{x} - \hat{\boldsymbol{\nu}})' \text{diag}(\hat{\boldsymbol{\nu}})^{-1} (\mathbf{x} - \hat{\boldsymbol{\nu}})$, where $\hat{\boldsymbol{\nu}}$ denotes the expected value of $\mathbf{x}$ corresponding to $\hat{\boldsymbol{\eta}}$, measures the discrepancy between S and H. The second one, $X_2^2 = X^2 - X_1^2$, measures that between H and H_0. Although the difference between X^2 and X_1^2 might be negative, the chance of this is very small, especially if an adequate number of MCMC samples is used for the parameter estimation.

We can associate p-values to X_1^2 and X_2^2 to decide weather to reject H in favor of S, implying that the positive association of interest does not hold, or H_0 in favor of H, having in this way a directed test for independence which has more power against a narrower set of alternatives. As usual, these p-values are computed under H_0 that, in our context, normally corresponds to the independence model; see Dardanoni and Forcina (1998) and Bartolucci *et al.* (2001). Under independence, we have that $\boldsymbol{\eta} = \mathbf{0}$ and, given the row and column margins, this uniquely determines the distributions of X_1^2 and X_2^2. This is the major advantage of the conditional approach with respect to the unconditional one where such distributions depend on the nuisance parameters.

In case of small samples, we compute p-values by a standard Monte Carlo simulation: we draw a certain number of tables under H_0 and then compute the p-values for X_1^2 and X_2^2 as the proportions of tables with larger values of the two statistics than the observed table. When H_0 corresponds to the independence model, we draw the tables using an exact Monte Carlo algorithm described in Diaconis and Sturmfels (1998). In the presence of large samples, instead, we can rely on the asymptotic theory. Under H_0, the asymptotic distribution of X_1^2 is $\bar{\chi}^2(\mathbf{F}_0^{-1}, \mathcal{H}^o)$ and that of X_2^2 is $\bar{\chi}^2(\mathbf{F}_0^{-1}, \mathcal{H})$, where $\bar{\chi}^2(\mathbf{\Sigma}, \mathcal{C})$ denotes the *chi-bar squared* distribution with parameters $\mathbf{\Sigma}$ (a variance-covariance matrix) and $\mathcal{C}$ (a convex cone); see Shapiro (1988). Moreover, $\boldsymbol{F}_0$ is the information matrix under H_0 and $\mathcal{C}^o$ denotes the dual of the cone $\mathcal{C}$. The survival function of $\bar{\chi}^2(\mathbf{\Sigma}, \mathcal{C})$ is given by $Pr(\bar{\chi}^2(\mathbf{\Sigma}, \mathcal{C}) \geq z) = \sum_{i=0}^{m} w_i(\mathbf{\Sigma}, \mathcal{C}) Pr(\chi_i^2 \geq z)$, where χ_i^2 is a chi-squared random variable with i degrees of freedom, $w_i(\mathbf{\Sigma}, \mathcal{C})$, $i = 0, 1, \ldots, m$, are weights depending on $\mathbf{\Sigma}$ and $\mathcal{C}$, and m is the size of the squared matrix $\mathbf{\Sigma}$. Computation of the weights $w_i(\mathbf{\Sigma}, \mathcal{C})$ is a difficult numerical problem, unless m is less than 4; however, accurate estimates can be easily obtained by simulation (Dardanoni and Forcina, 1998).

As shown by Bartolucci and Scaccia (2004), the framework described above may also be used for the analysis of bivariate tables stratified according to one or more discrete explanatory variables, such as gender or educational level.

5 An application

The data in Table 1 are taken from a survey conducted by the Department of Energy; for a detailed description see Simonoff (1987). These data concern a sample of 147 female employees having the Bachelor (but not higher) degree who are cross-classified by monthly salary and years since degree. The table is quite sparse, with an average cell frequency equal to 2.70 and 28% of the cells having null frequencies.

	Years Since Degree								
Salary	0–2	3–5	6–8	9–11	12–14	15–17	18–23	24–29	30+
950–1350	7	1	1	0	0	0	0	2	0
1351–1750	10	6	5	3	0	1	1	1	0
1751–2150	12	14	7	1	4	2	2	1	2
2151–2550	0	1	8	3	3	3	5	0	4
2551–2950	0	0	3	2	0	6	5	2	7
2951–3750	1	0	1	0	1	1	6	0	2

Table 1. Salary and years since degree for a sample of 147 female employees.

We fitted two models of positive association, based on log-odds ratios of type l and c, respectively, obtaining the results reported in Table 2. We can conclude that the data conform to the form of positive association expressed through log-odds ratios of type cc (corresponding to logits of type continuation for both variables): the Pearson's chi-squared statistic between the saturated and the constrained model is $X_1^2 = 17.415$ with a simulated p-value of 0.6240. This implies that also PQD holds for these data. On the other hand, TP_2 has to be rejected since $X_1^2 = 68.542$ with a simulated p-value equal to 0.0020. Also, the independence model has to be rejected with a p-value for X_2^2 smaller than 10^{-4}. In conclusion, a certain degree of positive association exists between the variables, i.e. females with more years since degree have a better chance of getting higher salary. However, having rejected TP_2, the association is not so strong as expected. This is a more precise conclusion than that of Simonoff (1987), who simply recognized the existence of positive association, without specifying its strength.

Type of log-odds ratios	X_1^2	Monte Carlo p-value	Asym. p-value	X_2^2	Monte Carlo p-value	Asym. p-value
cc	17.415	0.6240	0.6000	94.920	0.0000	0.0000
ll	68.542	0.0020	0.0007	43.792	0.0000	0.0000

Table 2. Pearson's chi-squared statistic for positive association (X_1^2) and independence model (X_2^2).

Finally, the simulated p-values are always very close to the asymptotic ones and then the asymptotic theory seems to provide a reasonable approximation even for sparse tables.

REFERENCES

BARTOLUCCI, F., FORCINA, A., DARDANONI, V. (2001): Positive quadrant dependence and marginal modelling in two-way tables with ordered margins. *Journal of the American Statistical Association*, *96*, 1497–1505.

BARTOLUCCI, F., SCACCIA, L. (2004): Testing for positive association in contingency tables with fixed margins. *Computational Statistics and Data Analysis*, *47*, 195–210.

DARDANONI, V., FORCINA, A. (1998): A unified approach to likelihood inference on stochastic orderings in a nonparametric context. *Journal of the American Statistical Association*, *93*, 1112–1123.

DIACONIS, P., STURMFELS, B. (1998): Algebraic algorithms for sampling from conditional distributions. *Annals of Statistics*, *98*, 363–397.

DOUGLAS, R., FIENBERG, S.E., LEE, M.T., SAMPSON, A.R., WHITAKER, L.R. (1991): Positive Dependence Concepts for Ordinal Contingency Tables. In: H.W. Block, A.R. Sampson, and T.H. Savits: (Eds.) *Topics in Statistical Dependence*, Hayward, CA, 189–202.

FISHER, R.A. (1934): *Statistical Methods for Research Workers*, (originally published 1925, 14th edn. 1970), Edinburgh: Oliver and Boyd.

FISHER, R.A. (1935): The logic of inductive inference (with discussion). *Journal of the Royal Statistical Society*, *98*, 39–82.

GELMAN, A., MENG, X.L. (1998): Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statistical Science*, *13*, 163–185.

GEYER, C.J. (1991): Markov Chain Monte Carlo Maximum Likelihood. In: E.M. Keramidas: (Eds.) *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, Interface Foundation of North America, Fairfax Station, VA, 156–163.

HAMMERSLEY, J., HANDSCOMB, D. (1964): *Monte Carlo Methods*, Methuen, London.

HASTINGS, W.K. (1970): Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*, 97–109.

METROPOLIS, N., ROSENBLUTH, A.W., ROSENBLUTH, M.N., TELLER, A.H., TELLER, E. (1953): Equations of state calculations by fast computing machines. *Journal of Chemistry Physics*, *21*, 1087–1091.

SHAPIRO, A. (1988): Towards a unified theory of inequality constrained testing in multivariate analysis. *International Statistical Review*, *56*, 49–62.

SIMONOFF, J.S. (1987): Probability estimation via smoothing in sparse contingency tables with ordered categories. *Statistics and Probability Letters*, *5*, 55–63.