Statistica & Applicazioni
Vol. III, n. speciale 1, 2005
pp. 21-28

Three-Way Ordinal Non Symmetrical Correspondence Analysis for the Evaluation of the Patient Satisfaction

Eric J. Beh[*]

Biagio Simonetti[†]

Luigi D'Ambra[†]

Summary: *In some recent articles, emphasis has been given to the partition of the Goodman-Kruskal's tau index using orthogonal polynomials for the study of the non symmetrical relations in three-way contingency tables. New graphical techniques that consider such a partition and allow for the analysis of asymmetric relationships have been proposed, including three-way ordinal non symmetrical correspondence analysis (Simonetti, 2003). Such a procedure takes into account the presence of an ordinal predictor and response variables. In this paper we demonstrate the applicability of such a technique for the patient satisfaction evaluation.*

Keywords: *Correspondence Analysis, Orthogonal Polynomials, Patient Satisfaction.*

1. Introduction

The evaluation of the performance of various aspects of the health care industry is an important aspect when monitoring the requirements of the societies health needs. From a statistical point of view, such monitoring often requires carrying out surveys and questionnaires. Such sources of important information usually consist of multiple questions with a variety of possible responses. One popular method of determining important relationships among variables of a categorical nature is to consider Non Symmetrical Correspondence Analysis (NSCA, D'Ambra, Lauro, 1989). Often, for many studies, the structure of categorical variables is of an ordinal nature, and the classical approach to NSCA does not guarantee that the structure of these variables is maintained. This paper looks at the application of the method of three-way ordinal NSCA (Beh, Simonetti, D'Ambra, 2005)

[*] University of Western Sydney, Australia (email: e.beh@uws.edu.au).
[†] University of Naples Federico II (email: simonett@unina.it; dambra@unina.it).

to the evaluation of patient satisfaction on data collected in hospitals in Naples.

2. The Measure in Dependence in Multi-way Tables

For the study of the relationship between variables in two-way contingency tables, the most common measure of asymmetry is the Goodman-Kruskal tau index (Goodman & Kruskal, 1954). For multi-way tables one may consider the index of Marcotorchino (1984). Here we will focus on this index and consider its partition into location, dispersion and higher order components. For more information on this partition, refer to Beh *et al*. (2005). Consider an $I \times J \times K$ three-way contingency table, N, where the (i, j, k)-th cell entry is given by n_{ijk}, for i=1,…,I, j=1,…,J, k=1,…,K. Let the grand total of N be n and the relative frequency of n_{ijk} be $p_{ijk} = n_{ijk} / n$. Define the i-th row marginal proportion by $p_{i\bullet\bullet}$ and define the j-th column marginal proportion as $p_{\bullet j\bullet}$. Similarly, let the k-th tube marginal proportion be denoted as $p_{\bullet\bullet k}$. The conditional probability that an individual/unit classified into row i, given that it belongs to column j and tube k, is $p_{ijk} / (p_{\bullet j\bullet} p_{\bullet\bullet k})$. Also, let $\pi_{ijk} = p_{ijk} / (p_{\bullet j\bullet} p_{\bullet\bullet k}) - p_{i\bullet\bullet}$ be the difference between the conditional prediction $p_{ijk} / (p_{\bullet j\bullet} p_{\bullet\bullet k})$ and the unconditional marginal prediction of $p_{i\bullet\bullet}$. The proportion of reduction of error in the prediction of the response variable can be calculated using Marcotorchino index

$$\tau_M = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} p_{\cdot j\cdot} p_{\cdot\cdot k}\left(\frac{p_{ijk}}{p_{\cdot j\cdot} p_{\cdot\cdot k}} - p_{i\cdot\cdot}\right)^2}{1 - \sum_{i=1}^{I} p_{i\cdot\cdot}^2} = \frac{\tilde{\tau}_M}{1 - \sum_{i=1}^{I} p_{i\cdot\cdot}^2} \qquad (1)$$

where $\tilde{\tau}_M$ can be expressed as the weighted sum of squares of π_{ijk} so that

$$\tilde{\tau}_M = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} p_{\cdot j\cdot} p_{\cdot\cdot k} \pi_{ijk}^2 \qquad (2)$$

3. Marcotorchino Index Decomposition

For a two-way asymmetric contingency table, D'Ambra, Beh and Amenta (2005) partitioned the measure of predicability using orthogonal polynomials. In a similar manner π_{ijk} can be decomposed so that

$$\pi_{ijk} = \sum_{u=1}^{I-1}\sum_{v=1}^{J-1}\sum_{w=1}^{K-1} a_u(i)\, b_v(j)\, c_w(k)\, Z_{uvw} \,. \qquad (3)$$

Here $a_u(i)$, $b_v(j)$ and $c_w(k)$ are polynomials with properties

$$\sum_{i=1}^{I} a_u(i)\, a_{u'}(i) = \begin{cases} 1, & u = u' \\ 0, & u \neq u' \end{cases} \quad \sum_{j=1}^{J} p_{\bullet j \bullet}\, b_v(j)\, b_{v'}(j) = \begin{cases} 1, & v = v' \\ 0, & v \neq v' \end{cases}$$

$$\sum_{k=1}^{K} p_{\bullet\bullet k}\, c_w(k)\, c_{w'}(k) = \begin{cases} 1, & w = w' \\ 0, & w \neq w' \end{cases}$$

where $$Z_{uvw} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} p_{ijk}\, a_u(i)\, b_v(j)\, c_w(k)$$

are akin to the generalised correlations of Davy, Rayner and Beh (2003).

In fact Beh, Simonetti and D'Ambra (2005) show that the numerator of the Marcotorchino index can partitioned into these terms such that

$$\tilde{\tau}_M = \sum_{u=1}^{I-1}\sum_{v=1}^{J-1} Z_{uv0}^2 + \sum_{u=1}^{I-1}\sum_{w=1}^{K-1} Z_{u0w}^2 + \sum_{v=1}^{J-1}\sum_{w=1}^{K-1} Z_{0vw}^2 + \sum_{u=1}^{I-1}\sum_{v=1}^{J-1}\sum_{w=1}^{K-1} Z_{uvw}^2 \,. \qquad (4)$$

The term Z_{uvw} is a measure of the deviation from the (u, v, w)-th trivariate moment of the three variables from what would be expected under the hypothesis of complete predictability of the row variable given the column and tube variables. For simplicity equation (4) can be alternatively expressed by:

$$\tilde{\tau}_M = \tilde{\tau}_{IJ} + \tilde{\tau}_{IK} + \tilde{\tau}_{JK} + \tilde{\tau}_{IJK} \,.$$

It can be shown that the first term, $\tilde{\tau}_{IJ}$, is the numerator of the Goodman-Kruskal index of the first (row response) and second (column predictor) variables formed by aggregating over the tube categories. Similarly, the second term, $\tilde{\tau}_{IK}$, is the numerator of the Goodman-Kruskal index of the first (row response) and third (tube predictor) variables formed by aggregating over the column categories. The third term, $\tilde{\tau}_{JK}$, involves the

partition of the chi-square index of the two predictors. The last term $\tilde{\tau}_{IJK}$ is related to the Pearson chi-squared of the column (predictor) and tube (predictor) variables by first aggregating across the row categories.

4. Test of Significance

The problem with considering the tau statistic is that it is not suitable for formally testing whether there exists (or not) an association between two or more variables. To do so, D'Ambra, Beh and Amenta (2005) multiplied each term in (4) by (I - 1)(n - 1) to obtain a C-statistic:

$$C = (I-1)(n-1)\tilde{\tau}_{IJ} + (I-1)(n-1)\tilde{\tau}_{IK} + (I-1)(n-1)\tilde{\tau}_{JK} + \\ + (I-1)(n-1)\tilde{\tau}_{IJK} = C_{IJ} + C_{IK} + C_{JK} + C_{IJK}$$

The first term, C_{IJ}, is equivalent to the C-statistic of Light and Margolin (1971) for the row (response) and column (predictor) variables. This measure can be compared with the statistic obtained from the chi-squared distribution with (I - 1)(J - 1) *d.f.* Therefore, after aggregating the tube categories, C_{IJ} can be used to determine if there is a significant asymmetric association between the row and column categories. Similarly, when compared with the chi-squared statistic obtained from the distribution with (I - 1)(K - 1) *d.f.*, C_{IK} can be used to formally test for association between the row and tube categories. The trivariate term can be treated in the same manner. The numerator of Marcotorchino index, $\tilde{\tau}_M$, can be used as a global measure association between the three variables by comparing C-statistic against a chi-squared statistic with *(I - 1)(J - 1) + (I - 1)(K - 1) + (J - 1)(K - 1) + (I - 1)(J - 1)(K- 1) d.f.* Even when one of the terms of $\tilde{\tau}_M$ is shown to provide no evidence of association between the response variable and at least one of the predictor variables, there may still exist significant sources of association between the variables by looking more closely at the four terms in (4). To determine whether any of these components are significant, the partition of the C-statistic can be made by considering the sum of squares of:

$$\tilde{\tilde{Z}}_{uvw} = Z_{uvw}\sqrt{\frac{(I-1)(n-1)}{1-\sum_{i=1}^{I} p_{ijk}^2}}$$

which are asymptotically standard normally distributed.

5. Graphical Representation

To graphically describe the relationship between the row (response) variable and the column (explanatory) variable, one may consider the following coordinates for the *i*-th row, *j*-th column and *k*-th tube category

$$f_{im} = \sum_{u=1}^{I-1} a_u(i) Z_{umm} \; ; g_{jm} = \sum_{v=1}^{J-1} b_v(j) Z_{mvm} \; ; h_{km} = \sum_{w=1}^{K-1} c_w(k) Z_{mmw} \qquad (5)$$

respectively. These will allow for a joint representation of all three variables on one low-dimensional space. In fact, if we consider an adjustment to f_{im} and g_{jm} above, namely $\tilde{f}_{im} = \sum_{u=1}^{I-1} a_u(i) Z_{um0}$ and $\tilde{g}_{jm} = \sum_{v=1}^{J-1} b_v(j) Z_{mv0}$ these may be used to graphically represent the association between the row (response) and column (predictor) categories. The problem with the coordinates of (5) is that their weighted sum of squares is not $\tilde{\tau}_{IJK}$, but only an unsaturated version of it that includes only the linear-by-linear, quadratic-by-quadratic etc. measures of association. To overcome this problem one may consider instead the coordinates

$$f_{ivw} = \sum_{u=1}^{I-1} a_u(i) Z_{uvw} \; ; g_{juw} = \sum_{v=1}^{J-1} b_v(j) Z_{uvw} \; ; h_{kuv} = \sum_{w=1}^{K-1} c_w(k) Z_{uvw} \qquad (6)$$

6. Analysis of Patient Satisfaction

The analysed data were collected from a study undertaken in a hospital in Naples. The aim of the study was to determine the factors that influence the level of satisfaction of the patients recovered in different wards. The questionnaire used was based on the Servqual model. Previous analysis on this data (D'Ambra L. *et al*., 2004) shows that the most important factors explaining the *Overall Satisfaction* (S1, S2, S3, S4) are the *Cleanliness* (C1, C2, C3, C4) of the hospital and the *Quality of Management* (Q1, Q2, Q3, Q4). The evaluation, or level of satisfaction, of the patients were recorded on a 4-point scale: poor [1], fair [2], good [3] and excellent [4]. An initial analysis of the data shows that the Marcotorchino numerator is $\tilde{\tau}_M = 0.4374$. With a corresponding *C* statistic of 2228.48 there is ample evidence to indicate that the patients satisfaction is influenced by the two predictor variables.

Table 1. *Partition of the Marcotorchino Numerator Index and contribution (as a percentage) of single components*

Component	$\tilde{\tau}_{IJ}$	$\tilde{\tau}_{IK}$	$\tilde{\tau}_{JK}$	$\tilde{\tau}_{IJK}$	$\tilde{\tau}_M$

Value	0.0795	0.1301	0.1647	0.0631	0.4374
% Contribution	18.17	29.75	37.65	14.43	100

Table 1 shows that of the two predictor variables, the tube variable (*Quality of Management)* is a more dominant factor for determining a patient's satisfaction than the cleanliness of the hospital.

This is evident since $\tilde{\tau}_{IK}$ contributes more than $\tilde{\tau}_{IJ}$ to the Marcotorchino numerator. However, both of the predictor variables significantly influence the response variable. This is because the *p*-value of the *C*-statistic for $\tilde{\tau}_{IJ}$ and $\tilde{\tau}_{IK}$ are both zero. In fact, the largest association between the three variables exists between the two predictor variables – *Quality of Management* and *Cleanliness* - since $\tilde{\tau}_{JK}$ contributes to more than a third of the association explained by $\tilde{\tau}_{M}$. Therefore, it seems appropriate to consider a doubly ordered non-symmetrical correspondence analysis of these variables. Figure 1(left) gives the correspondence plot from such an analysis. It shows that those patients who responded positively to the cleanliness of the hospital also responded favourably to the quality of management. For a simultaneous analysis of all three variables, Figure 1(right) shows a joint low-dimensional plot of the three variables in terms of the location and dispersion components. The plotting system used was that of (5) for $m = 1$ and 2. The general configuration of *Quality* and *Cleanliness* categories in the two plots of Figure 1 are fairly similar. However it is clear that the relative positions of Q1 and Q2 in the two plots are very different. This is because the first axis in Figure 1 only considers the trivariate location association between the three variables. Similarly the second axis includes only the trivariate dispersion association. These axes therefore ignore all the contributions made by the linear-by-quadratic and other "non-diagonal" component values. For this reason we adopt instead the coordinate system of (6) yielding a series of correspondence plots. By observing the significant Z_{uvw} values, the linear source of variation is the most important. So we will obtain a set of two linear plots – one representing the association between the response variable and column predictor variables taking into account the location differences in the tubes. The total inertia is quantified by $\tilde{\tau}_{M|w=1} = 0.0272$ and represents 43.11% of $\tilde{\tau}_{IJK}$ (when focusing only on the row and column association). The correspondence plot of this association is given by Figure 2 (left). The association between the response variable and the tube predictor variable can be measured by taking into account the location difference of the columns. In this case the total inertia is quantified by $\tilde{\tau}_{M|v=1} = \sum_{i=1}^{I}\sum_{w=1}^{K-1} f_{i1w}^{2} = \sum_{k=1}^{K}\sum_{u=1}^{I-1} p_{\bullet\bullet k} h_{ku1}^{2} = 0.0354$ and accounts for 56.04%

of the association between these variables for $\tilde{\tau}_{IJK}$. The plot of this association is given by Figure 2(right). One may also obtain a similar plot for the two predictor variables given the location differences in the response.

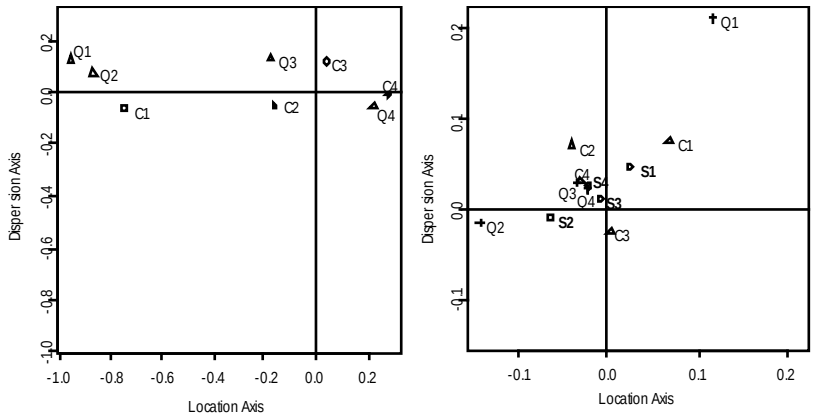

Figure 1. *(left) Doubly ordered symmetric correspondence plot of the two predictor variables; (right) Joint representation of the three variables*

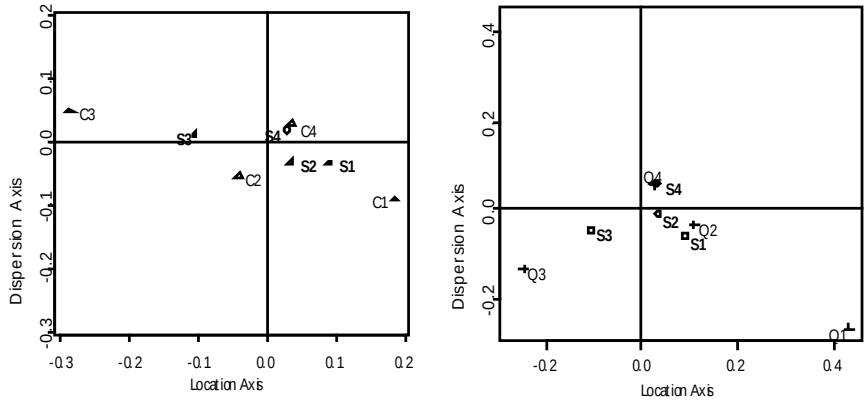

Figure 2. *Ordinal Correspondence Plot of Overall Satisfaction - (left) and Cleanliness given location difference in the Quality of Management; (right) and Quality of Management given location difference in the Cleanliness*

Figure 2 shows that patients who were satisfied with their stay in hospital were also satisfied with the cleanliness of the hospital and the quality of their management. Figure 2(left) shows that there is a difference in the location of the first two variables, but there is very little difference in the spread of the categories, when taking into account the third variable. However Figure 2(right) shows that there appears to be a significant difference in the way the quality of management is spread from one another.

7. Conclusions

The method described above is useful when the three variables consist of ordinal sets of categories. There are cases when only one, or two, variables consist of ordinal categories and so the partition described above can be generalised in several different ways. In terms of health care, any of these options are available to the researcher, but describe in different ways the relationship between the categories. The analysis described here graphically shows how the two predictor variables impact upon the satisfaction of the patients.

References

Beh E. J., Simonetti B., D'Ambra L. (2005). Partitioning Marcotorchino's index for ordinal three-way contingency tables, (in review).

D'Ambra L., Beh E. J., Amenta P. (2005). CATANOVA for two-way contingency tables with ordinal variables using orthogonal polynomials, *Communications in Statistics (Theory and Methods)*, **34**, 1755-1769.

D'Ambra L., Lauro N. C. (1989). Non symmetrical correspondence analysis for three-way contingency table, In: R. Coppi and S. Bolasco (Eds.), *Multiway Data Analysis*, pp. 301-315. Elsevier, Amsterdam.

D'Ambra L. *et al*. (2004). Analisi statistica multivariata per la valutazione della patient satisfaction, In A. Pagano, G. Vittadini, *Qualità e Valutazione delle strutture sanitarie*, Etas, Milano.

Davy P. D., Rayner J. C. W., Beh E. J. (2003). Generalised correlations and Simpson's Paradox, In *Current Research on Modelling, Data Mining and Quantitative Techniques*, (Eds. P. Pemajayantha, R. Mellor, S. Peiris, and J. R. Rajasekara), pp. 63-73.

Emerson P. L. (1968). Numerical construction of orthogonal polynomials from a general recurrence formula, *Biometrics*, **24**, 696-701.

Goodman. L. A., Kruskal W. H. (1954). Measures of association for cross-classifications, *Journal of the American Statistical Association*, **49**, 732-764.

Light R. J., Margolin B. H. (1971). An analysis of variance for categorical data, *Journal of the American Statistical Association*, **66**, 534-544.

Marcotorchino F. (1984). Utilisations des comparisons par paires en statistique des contingences, *CS IBM, Paris*, Part I and II no.F *071*, no. F *069*.

Simonetti B. (2003). *Ordinal and Non Ordinal Non Symmetric Correspondence Analysis for Three-way Tables in Sensorial Analysis* (in Italian) (PhD Thesis), University of Naples (Federico II), Italy.