NOVA

University of Newcastle Research Online

nova.newcastle.edu.au

Beh, Eric J.; D'Ambra, Luigi "Some interpretative tools for non-symmetrical correspondence analysis" Journal of Classification Vol. 26, Issue 1, p. 55-76 (2009)

Available from: <http://dx.doi.org/10.1007/s00357-009-9025-0>

The final publication is available at www.springerlink.com

Accessed from: <http://hdl.handle.net/1959.13/1054338>

Some Interpretative Tools for Non-Symmetrical Correspondence Analysis

ERIC J. BEH

School of Computing and Mathematics, University of Western Sydney, Australia;

LUIGI D'AMBRA

Department of Mathematics and Statistics, University of Naples, Italy

Abstract

Non-symmetrical correspondence analysis (NSCA) is a very practical statistical technique for the identification of the structure of association between asymmetrically related categorical variables forming a contingency table. This paper considers some tools that can be used to numerically and graphically explore in detail the association between these variables and include the use of confidence regions, the establishment of the link between NSCA and the analysis of variance of categorical variables, and effect of imposing linear constraints on a variable.

Keywords:

CATANOVA ; Confidence circles; Goodman-Kruskal tau index; Linear constraints; Non-symmetrical correspondence analysis.

1. INTRODUCTION

Identifying meaningful associations that may exist between two or more categorical variables is an important, and ongoing, element to the analysis of contingency tables. It involves detecting categories of one variable that are similar and/or different to other categories of the same, or of a different, variable. Generally, this can be done by first isolating those categories that do, and do not, exhibit characteristics consistent with the departure from independence.

For a graphical analysis of these departures, correspondence analysis is a very popular tool that provides a visual perspective to identifying the strength of dependence between the rows and columns of a two-way contingency table. When two categorical variables are assumed to have a symmetric two-way relationship, simple correspondence analysis can be performed. For example, refer to Greenacre (1984), Lebart, Morineau and Warwick (1984), and Beh (2004) for details of this classical approach. When it is evident that there is a one-way relationship between the variables, then non-symmetrical correspondence analysis is more appropriate. For details refer to D'Ambra and Lauro (1989, 1992) and Kroonenberg and Lombardo (1998). In this case, the direction of the association between the variables is an important issue and implies that the researcher can not use Pearson chi-squared statistic. Instead, the analysis of asymmetrically associated categorical variables can be made using the Goodman-Kruskal or, equivalently, the Light-Margolin, index. For such a case, Singh (1996) showed that the CATANOVA method (using the C-statistic of Light and Margolin, 1971) should be given preferential use in practical

applications since, in small samples (generally where $n < 5$), the C-statistic provides a better approximation than the chi-squared statistic under the hypothesis of zero association.

This paper explores the use of various tools that can be used in conjunction with NSCA and does so in four further sections. Section 2 provides a brief description of classical non-symmetrical correspondence analysis and compares it with the more popular symmetric approach. Section 3 gives eight tools that help to better identify how the one-way dependence between two categorical variables is structured. Section 4 provides a simple demonstration of some of these tools by considering their application to two real-life contingency tables. Some concluding remarks are made in Section 5.

2 NON-SYMMETRICAL CORRESPONDENCE ANALYSIS

2.1 The Goodman-Kruskal tau Index

Consider an $I \times J$ two-way contingency table, N , where the (i, j) th cell entry is denoted by n_{ij} for $i = 1, 2, \dots, I$ and $j = 1, 2, \dots, J$. Let the grand total of N be n and

the (i, j) th joint proportion by $p_{ij} = n_{ij} / n$ so that $\sum_{i=1}^I \sum_{j=1}^J p_{ij} = 1$. Define the i th row

marginal proportion by $p_{i\bullet} = \sum_{j=1}^J p_{ij}$ and define the j th column marginal proportion as

$p_{\bullet j} = \sum_{i=1}^I p_{ij}$. Furthermore, let $\pi_{ij} = \frac{p_{ij}}{p_{\bullet j}} - p_{i\bullet}$ be the differences between the

unconditional marginal prediction $p_{i\bullet}$ and the conditional prediction of the i th response (given column responses), $p_{ij} / p_{\bullet j}$. Therefore, we treat the row variable as a response variable given the presence of the column, or predictor, variable. Goodman and Kruskal (1954) proposed that the proportional reduction in error for the prediction of the response (row) variable can be measured by considering

$$\tau = \frac{\sum_{i=1}^I \sum_{j=1}^J p_{\bullet j} \left(\frac{p_{ij}}{p_{\bullet j}} - p_{i\bullet} \right)^2}{1 - \sum_{i=1}^I p_{i\bullet}^2} = \frac{\tau_{num}}{1 - \sum_{i=1}^I p_{i\bullet}^2}.$$

When, for each row, the distribution of the response categories across each of the columns is identical to the overall marginal proportion there is no relative increase in predictability and thus τ is zero. Similarly, $\tau = 1$ only when there is perfect predictability of the response categories (rows) given the predictor categories (columns). D'Ambra and Lauro (1989) showed that the diagonal elements of $\Pi D_j \Pi^T$ can be interpreted as measures of covariance between the two variables.

Here Π is the matrix whose (i, j) th element is $\pi_{ij} = \frac{p_{ij}}{p_{\bullet j}} - p_{i\bullet}$ and D_j is a diagonal matrix where the (j, j) th element is $p_{\bullet j}$. Similarly, if we consider the metric

$$D_i = \frac{1}{\sum_{j=1}^J p_{\bullet j} \left(\frac{p_{ij}}{p_{\bullet j}} - p_{i\bullet} \right)^2},$$

the trace of $D_i^{1/2} \Pi D_j \Pi^T D_i^{1/2}$ is the index proposed by Tallur (1986) in the context of cluster analysis.

2.2 Classical NSCA

Non-symmetrical correspondence analysis involves applying the singular value decomposition to π_{ij} so that

$$\pi_{ij} = \frac{p_{ij}}{p_{\bullet j}} - p_{i\bullet} = \sum_{m=1}^M a_{im} \lambda_m b_{jm}, \quad (1)$$

where $M = \min(I, J) - 1$ and λ_m is the m th singular value of π_{ij} for $m = 1, \dots, M$.

The quantities a_{im} and b_{jm} are singular vectors associated with the i th row and j th column categories, respectively, and have the property

$$\sum_{i=1}^I a_{im} a_{im'} = \begin{cases} 1, & m = m' \\ 0, & m \neq m' \end{cases} \quad \sum_{j=1}^J p_{\bullet j} b_{jm} b_{jm'} = \begin{cases} 1, & m = m' \\ 0, & m \neq m' \end{cases}.$$

Therefore, the numerator of the Goodman and Kruskal tau index can be decomposed so that

$$\tau_{num} = \sum_{m=1}^M \lambda_m^2.$$

When performing NSCA we can graphically depict the association between the response (row) categories and predictor (column) categories by plotting, along the m th dimension of the non-symmetrical correspondence plot, the row and column profile coordinates

$$f_{im} = a_{im} \lambda_m \quad (2)$$

$$g_{jm} = b_{jm} \lambda_m. \quad (3)$$

If one considers using these coordinates, then it must be kept in mind that they are not guaranteed to be centered about the origin of the correspondence plot (a useful property underlying the coordinates from classical, symmetrical, correspondence analysis). However, with respect to the unit and $p_{\bullet j}$ metrics, the row and column coordinates, respectively, are closely related to the numerator of the tau index through

$$\tau_{num} = \sum_{i=1}^I \sum_{m=1}^M f_{im}^2 = \sum_{j=1}^J \sum_{m=1}^M p_{\bullet j} g_{jm}^2 . \quad (4)$$

Therefore, points lying at a distance from the origin of the plot make more of a contribution to τ_{num} than those points that lie near the origin.

2.3 Distances in NSCA

One of the main reasons for considering NSCA is that a graphical summary of the data can be made. A plot obtained from performing NSCA allows the user to identify response and/or predictor categories that are relatively similar or different based on their proximity from one another. Such proximities can be measured by observing the squared distance between π_{ij} for each i (or j). Suppose, for example, we are interested in identifying those predictor categories that are similar, or different. It can be easily shown that the squared Euclidean distance between the j th and j' th predictor categories can be measured by

$$d_j^2(j, j') = \sum_{i=1}^I \left(\frac{p_{ij}}{p_{\bullet j}} - \frac{p_{ij'}}{p_{\bullet j'}} \right)^2 = \sum_{i=1}^I (\pi_{ij} - \pi_{ij'})^2 .$$

Similarly, the distance between two predictor categories can be measured by

$$d_j^2(i, i') = \sum_{j=1}^J \frac{1}{p_{\bullet j}} (p_{ij} - p_{i'j})^2 - (p_{i\bullet} - p_{i'\bullet})^2 .$$

These distance measures show that the proximity of two row (or two column) categories indicates the extent to which those categories are similar, or different. Predictor categories that are similarly distributed across the response variable will be positioned close to one another in a correspondence plot. Conversely, column coordinates situated at a distance from one another are different. This characteristic ensures that the property of distributional equivalence holds – see, for example, Lebart, Morineau and Warwick (1984, pg 34).

The proximity of a particular response category from a predictor category indicates the importance of the relationship between them – this is generally dictated by the large number of classifications made into the corresponding cell of N . In contrast to this, a response category that is at a distance from a particular predictor category indicates that there are almost no observations in that cell. However, the user is warned that care must be taken when making inferences based on inter-variable distances.

If it is of interest to measure the distance of a predictor category from the origin (since a point close to the origin indicates its lack of contribution to the association of the categories), then the squared distance of column j from the origin is

$$d_j^2(j, 0) = \sum_{i=1}^I \pi_{ij}^2 = \sum_{u=1}^{I-1} g_{ju}^2 .$$

Therefore, this distance reflects the influence of the j th column category to the N_τ statistic. This property is evident since equation (4) implies that

$$\tau_{num} = \sum_{j=1}^J p_{\bullet j} d_j^2(j, 0).$$

2.4 Comparison of Classical Correspondence Analysis and Non Symmetrical Correspondence Analysis

NSCA can be viewed as an adaptation of centered principal components analysis on the j th column profile $p_{ij} / p_{\bullet j}$ with a weight of $p_{\bullet j}$ and a metric of one. A major difference between non-symmetrical correspondence analysis and the symmetric version is that classical correspondence analysis can be viewed as a principal component analysis of the data with same weight but with a metric of $1/p_{i\bullet}$, the reciprocal of the marginal proportion of the i th response variable. If one views the difference between the two approaches from this perspective, it is apparent that the choice of metric in classical correspondence analysis is not suitable in some practical situations. The reason is that if the metric $1/p_{i\bullet}$ is small, the i th response category becomes a dominant source of information when determining the strength of association between the two categorical variables. Comparing the total inertia of classical correspondence analysis

$$\Phi^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(p_{ij} - p_{i\bullet} p_{\bullet j})^2}{p_{i\bullet} p_{\bullet j}} = \sum_{i=1}^I \frac{1}{p_{i\bullet}} \sum_{j=1}^J \left(\frac{p_{ij}}{\sqrt{p_{\bullet j}}} - \frac{p_{i\bullet}}{\sqrt{p_{\bullet j}}} \right)^2$$

with the numerator of the Goodman-Kruskal tau index, τ_{num} , which can be rewritten as

$$\sum_{i=1}^I \sum_{j=1}^J \left(\frac{p_{ij}}{\sqrt{p_{\bullet j}}} - \frac{p_{i\bullet}}{\sqrt{p_{\bullet j}}} \right)^2 \quad ,$$

we see that the main difference is in the weighting by $1/p_{i\bullet}$. Therefore, Φ^2 is highly sensitive to small marginal proportions in the response variable. Since the numerator of the Goodman-Kruskal tau index is not sensitive in these situations, it can be used as a valid alternative to Φ^2 in these cases.

3. SOME HELPFUL TOOLS

The classical approach to performing NSCA is a very useful procedure for understanding the asymmetric association of a categorical predictor variable and a categorical response variable. In this section we comment on some common, and not so common, tools that can be used to better understand the nature of the association of the variables.

3.1 Formally Testing Predictability

The Goodman-Kruskal index is a measure of the predictability of the row (response) categories given the column (predictor) categories. The statistic is zero when the row categories cannot be predicted based on the distribution of the column categories, and is one when there is perfect predictability of the response variable. However, Agresti (1996) indicates that low values of τ do not mean that there is a “low” association between the two variables. While τ is an appropriate measure of the predictability of the rows given the columns, the statistic cannot, without modification, be used to formally test for association. Instead such tests are performed using the C-statistic of Light and Margolin (1971), defined as

$$C = (n-1)(I-1)\tau = \frac{(n-1)(I-1)}{1 - \sum_{i=1}^I p_{i\bullet}^2} \sum_{m=1}^M \lambda_m^2, \quad (5)$$

which is, under the null hypothesis, $H_0 : P_{ij} = P_{i\bullet}$ for all $j = 1, 2, \dots, J$, asymptotically chi-squared distributed with $(I-1)(J-1)$ degrees of freedom. Light and Margolin (1971) introduced this statistic when deriving an analysis of variance procedure for contingency tables, commonly referred to as CATANOVA. For the analysis of ordinal response and predictor variables, D’Ambra, Beh and Amenta (2005) proposed a method of partitioning the Goodman-Kruskal tau index that takes into consideration this structure. From a NSCA point of view, Lombardo, Beh and D’Ambra (2007) proposed a graphical approach to visualize the association between asymmetrically associated categorical variables. However, we will not delve any further into methods that consider the structure of ordinal categorical variables.

3.2 Alternative Graphical Displays

The most utilized feature of any symmetric, or asymmetric, correspondence analysis is the graphical summary, or correspondence plot, that can be obtained. For non-symmetrical correspondence analysis the coordinates for the response and predictor categories defined by (2) and (3), respectively, is enough to visually describe their relationship. However, many other coordinate systems can also be considered. For example, the row and column coordinates of Goodman (1986) for the classical approach can be adopted for NSCA by defining the row and column coordinates

$$\tilde{f}_{im} = a_{im} \lambda_m^\gamma \quad \tilde{g}_{jm} = b_{jm} \lambda_m^{1-\gamma}$$

for some $0 \leq \gamma \leq 1$. Scaling λ_m in this manner means that the correspondence plot behaves in much the same way as a biplot for asymmetrically related categorical data. Consider the classical correspondence analysis of a two-way contingency table. For the analysis of quantitative variables Gabriel (1971) describes three sets of coordinates that coincide with $\gamma = 1/2$, $\gamma = 1$ and $\gamma = 0$. For the analysis of categorical data, the assignment of these values has also been described as “symmetric”, “row isometric” and “column isometric” factorizations, respectively (Lombardo, Carlier and D’Ambra, 1996). When there exists an asymmetric association between the two variables, these coordinates can be considered and the

factorizations apply. In fact, there seems to be some advantage in considering these coordinates instead of (2) and (3), since the measure of predictability of the (i, j) th cell can be expressed in terms of these coordinates such that

$$\pi_{ij} = \sum_{m=1}^M \tilde{f}_{im} \tilde{g}_{jm} .$$

This bilinear form suggests that one may obtain a simple geometric interpretation of the points associated with the i th row and j th column in a correspondence plot.

However, if one wishes to compare only the response categories, given the predictor categories, a simple graphical summary (Figure 1) can be obtained by plotting π_{ij} vs j ; which can be termed the P-plot. It allows for a quick graphical comparison of the response categories across the different predictor categories. Those π_{ij} that are consistently close to zero indicate that the predictor categories are not good at predicting the response categories. The P-plot, just like the correspondence plot, has the advantage that it allows for a direct comparison of the response categories by considering their proximity from one another. As a result, if $\pi_{ij} \geq 0$, for $j = 1, \dots, J$, then the conditional distribution of the i th response category given the j th predictor category is greater than what is expected under H_0 (i.e. the marginal distribution of the j th predictor category). Similarly, if $\pi_{ij} \leq 0$ then the i th response category given the j th predictor variable is less than what is expected under H_0 . Distance measures between π_{ij} and $\pi_{ij'}$ can also be obtained. For example,

$$d_i(j, j') = \pi_{ij} - \pi_{ij'} = \frac{p_{ij}}{p_{\bullet j}} - \frac{p_{ij'}}{p_{\bullet j'}}$$

is the distance between the j th and j' th column profile at the i th response category so that $d_i^2(j, j') = \sum_{i=1}^I d_i^2(j, j')$. Note that Choulakian and Allard (1998) proposed a

similar type of plot, but looked at the cumulative sum of $\pi_{ij} - \sum_{m=1}^i \pi_{mj}$, and plotted this against j , obtaining a Z-plot. Such a plot was proposed to visually describe the variation between categories of one variable while taking into consideration the ordinal structure of a second variable.

For a more detailed, or simply a different, graphical summary the researcher may decide that other graphical summaries are more appropriate. There are a multitude of other techniques that can be used to visualize the asymmetric association between the rows and columns. Mosaic displays (Friendly, 1994, 1999) can be considered, as can the Chernoff face (Chernoff, 1973) and Andrews' curve (Andrews, 1972). In fact, the Andrews' curve has been used in conjunction with correspondence analysis before – see for example Rovin (1994), and Khattree and Naik (2002a, 2002b).

3.3 Dimensionality of the Solution

When one obtains a correspondence plot, the next step is to determine the appropriate number of dimensions to use.

Recall that τ_{num} may be expressed as the sum of the squared singular values. Therefore the ratio λ_m^2 / τ_{num} represents the proportion of the total inertia explained by the m th axis. Alternatively, the squared singular values can be divided by $1 - \sum_i p_{i\bullet}^2$ to determine those axes that have a large contribution to the Goodman-Kruskal tau index. In some cases a large number of dimensions may be needed to obtain an adequate graphical perspective of the nature of the association between a predictor and a response variable. However, in general, higher dimensions are unlikely to contribute significantly to the predictability of the rows given the columns. Generally, two or three dimensions are adequate. Where correspondence plots require more than three dimensions, alternative multi-dimensional graphical tools should be used – for example, one may consider Andrews' curves.

An alternative approach to deciding the adequate number of dimensions to use is to consider the following. When a contingency table consists of J column categories, and if the data were purely random with no significant dependent structure, then each axis should account for, on average, $100/(J-1) = b\%$ of τ_{num} . Likewise, if there are I row categories, the average axis should account for $100/(I-1) = a\%$ of τ_{num} . The choice of an influential dimension can be made by choosing those axes whose singular value exceeds $\psi = \frac{1}{\min(I, J) - 1} \tau_{num}$.

3.4 Some Classical Descriptive Measures

The relative importance of a predictor or response category to the measure of predictability can be easily calculated by considering the following ratios. The contribution of the i th response category to τ_{num} can be measured by considering the relative response contribution

$$RRC_i = \frac{\sum_{j=1}^J p_{\bullet j} \pi_{ij}^2}{\tau_{num}} = \frac{\sum_{m=1}^M f_{im}^2}{\tau_{num}},$$

while the j th relative predictor contribution can be measured by

$$RPC_j = \frac{p_{\bullet j} \sum_{i=1}^I \pi_{ij}^2}{\tau_{num}} = \frac{p_{\bullet j} \sum_{m=1}^M g_{jm}^2}{\tau_{num}}.$$

Similarly the relative contribution of the m th axis can be expressed by

$$RA_m = \frac{\lambda_m^2}{\tau_{num}} .$$

Absolute contributions can be calculated to indicate the proportion of inertia explained by each category in relation to each principal axis. The contribution of the i th category to the m th axis can be obtained by means of the following ratios

$$CTA(i, m) = f_{im}^2 / \lambda_m , \quad CTA(j, m) = p_{\bullet j} \frac{g_{jm}^2}{\lambda_m} .$$

The *absolute* contribution allows one to define *explicative points* whose absolute contribution $CTA(i, m)$ (for $i = 1, \dots, I$) or $CTA(j, m)$ (for $j = 1, \dots, J$) are distinctly higher than the contributions of other points. For example, the I response categories whose contributions are higher than the average of the whole contribution are considered as *explicative*.

3.5 The Quality of Representation

Determining the *quality* of the graphical representation of a particular row or column category provides additional richness to the interpretation of the relationships that may exist in the contingency table. Suppose that a low-dimensional correspondence plot consists of M^* dimensions, where $M^* < M$. The *quality* of a point is defined as the ratio of the squared distance of the point from the origin in the M^* -dimensional subspace to the squared distance from the origin in the space defined by the maximum number of dimensions, M . A low quality means that the number of dimensions chosen for the subspace (M^*) does not adequately represent that column or row category. The squared correlation for any column measures the degree of association between that column and a particular axis. This quantity may be further interpreted by taking into account the sign of the co-ordinates on each axis.

3.6 Confidence Circles

When the variables of a two-way contingency table are considered to be symmetrically related, as is in the case of classical correspondence analysis, Lebart, Morineau and Warwick (1984) presented the simple idea of confidence circles to identify those categories that are consistent with what is expected under the hypothesis of independence. These circles are similar to those of Mardia, Kent and Bibby (1982) derived for canonical analysis. Others to have explored the use of these circles include Ringrose (1992, 1996) who focused on a bootstrap procedure for their construction. When categorical variables are ordinal in nature, Beh (2001) showed that the radii of these circles are identical to those of Lebart et al (1984). However, for NSCA the construction of these circles has not been considered in the literature. By preserving the spirit and intent of the circles obtained by Lebart, Morineau and Warwick (1984), in this section we look at the derivation of confidence circles for NSCA. Due to the asymmetric structure of the association between the variables, only the confidence circles for the predictor variable are obtained and will provide an indication of those predictor categories that significantly impact upon the response categories.

Suppose that a two-way contingency table consists of row (response) and column (predictor) categories that are asymmetrically related. The C – statistic of (5) above can be expressed in terms of the predictor profile coordinates such that

$$C = \frac{(n-1)(I-1)}{\left(1 - \sum_{i=1}^I p_{i\bullet}^2\right)} \sum_{j=1}^J \sum_{m=1}^M p_{\bullet j} g_{jm}^2 \sim X_{(I-1)(J-1)}^2 .$$

For the j th column coordinate

$$\frac{(n-1)(I-1)}{\left(1 - \sum_{i=1}^I p_{i\bullet}^2\right)} \sum_{m=1}^M p_{\bullet j} g_{jm}^2 \sim X_{(J-1)}^2 .$$

Therefore

$$\sum_{m=1}^M p_{\bullet j} g_{jm}^2 \sim X_{(J-1)}^2 \frac{\left(1 - \sum_{i=1}^I p_{i\bullet}^2\right)}{(n-1)(I-1)} . \quad (6)$$

For higher dimensions, the coordinates will generally close to zero (since the singular values associated with these dimensions are generally relatively close to zero). Therefore the relationship between the j th predictor coordinate for the first two dimensions of a two-dimensional non-symmetrical correspondence plot is

$$p_{\bullet j} g_{j1}^2 + p_{\bullet j} g_{j2}^2 = \frac{X_{(2)}^2 \left(1 - \sum_{i=1}^I p_{i\bullet}^2\right)}{(n-1)(I-1)} .$$

At the 5% level of significance, this can be expressed as

$$g_{j1}^2 + g_{j2}^2 = \left[\sqrt{\frac{5.99 \left(1 - \sum_{i=1}^I p_{i\bullet}^2\right)}{p_{\bullet j} (n-1)(I-1)}} \right]^2 .$$

where 5.99 is the value of the cumulative chi-squared statistic with 2 degree of freedom at the 0.05 level of significance. Therefore, the 95% confidence circle for the j th predictor coordinate in the two-dimensional non-symmetrical correspondence plot has a radius

$$r_j^J = \sqrt{\frac{5.99 \left(1 - \sum_{i=1}^I p_{i\bullet}^2\right)}{p_{\bullet j} (n-1)(I-1)}} . \quad (7)$$

Note that (7) depends on the j th marginal proportion classified into that category. Thus, for a very small classification in the j th predictor category, the radius will be relatively large. Similarly, for a relatively large classification, the radius will be relatively small. By considering these confidence circles it can be identified which (if any) predictor categories significantly contribute to the dependence structure between the row and column variables. If the origin lies within a confidence circle, then the associated predictor category does not significantly influence the outcome of the response categories. On the other hand, if the origin does not include the origin, then those predictor categories do significantly influence the outcome of the response categories.

3.7 Outliers in NSCA

A contingency table may contain one or more “outlying” row and/or column categories. The impact of these outliers when drawing conclusions about the dependence between two variables is that they tend to dominate the interpretation of one or more of the axes. In particular, their dominance may be substantial in higher dimensions. Therefore, the remaining row and/or column points tend to be tightly clustered in the plot and thus can make interpretations difficult. There have been several contributions on the identification of outliers (also referred to as “influence”) in correspondence analysis. For example, refer to Kim (1992, 1994), Benasseni (1993) and Bar-Hen and Mortier (2004).

Potential outliers may be detected by identifying rows or columns that have a high absolute co-ordinate. They may be treated as supplementary points with NSCA rerun without these points being allowed to determine the nature or orientation of the axes.

Based on the results of (7), if

$$\xi_j = \max \left\{ \left| g_{jm} \sqrt{\frac{p_{\bullet j}(n-1)(I-1)}{1 - \sum_{i=1}^I p_{i\bullet}^2}} \right|, m = 1, \dots, M \right\} > 1.96$$

the j th predictor category influences the structure of the response variable. That is, by considering this measure, those predictor categories that provide a significant prediction of the response can be identified. This result can be obtained by considering the Euclidean distance of the j th predictor category from the origin when the profile coordinate g_{jm} is considered.

In terms of outlier detection, one may consider the j th predictor category as an outlier if ξ_j exceeds 2.58. This value corresponds to the extreme (upper) 1% of categories that are not consistent with what is expected under the hypothesis of independence.

3.8 NSCA with Linear Constraints

As we have demonstrated above, NSCA is a useful tool for obtaining a graphical representation of the dependence structure between a response variable and a

predictor variable. This visual summary may be improved upon if additional information concerning the structure of the variables is available prior to performing the analysis. By incorporating this information through linear constraints on the row and/or column singular vectors, a representation of the data may be obtained.

Recall that classical NSCA involves decomposing π_{ij} according to (1) above. In matrix notation this may be expressed as $Z = \Pi = AD_\lambda B^T$ where $A^T A = I$, $B^T D_J B = I$, $D_J = \text{diag}(p_{\cdot j})$ and D_λ is a diagonal matrix with singular values arranged in descending order. We can incorporate the additional information regarding the structure of the predictor and response categories by defining the linear constraints as

$$G^T A = 0 \quad \text{and} \quad H^T B = 0$$

Here, G is an $I \times K$ matrix of rank K and H is a $J \times L$ matrix of rank L that defines the linear constraints on the response and predictor variable, respectively. These constraints apply to the first K dimensions for the row categories and the first L dimensions of the column categories. For example, if linear constraints were imposed only on the singular values for the first two dimensions of a correspondence plot then $K = L = 2$. Böckenholt and Böckenholt (1990) defined a non-iterative procedure for performing correspondence analysis with linear constraints. These constraints ensure that the average of the row and weighted average column scores equal zero. Following their work, NSCA can be considered by applying singular value decomposition to the matrix

$$\left[I - G(G^T G)^{-1} G^T \right] Z \left[I - D_J^{1/2} H(H^T D_J H)^{-1} H^T D_J^{1/2} \right] = M D_\omega S^T,$$

where $M^T M = S^T D_J S = I$ and D_ω is a diagonal matrix of the singular values ω_m arranged in descending order. For the m th singular value, other linear constraints may be imposed by computing the rank-one reduced matrix

$$Z_m = (I - \mathbf{a}_m \mathbf{a}_m^T) Z (I - \mathbf{b}_m \mathbf{b}_m^T),$$

where $\mathbf{a}_m = \{a_{im} : m = 1, \dots, I\}$ and $\mathbf{b}_m = \{b_{jm} : m = 1, \dots, J\}$ are singular vectors corresponding to first m singular values, and substituting Z_m for Z above. More generalized work on implementing linear constraints can be seen in Takane and Shibayama (1991) and Takane, Yanai and Mayekawa (1991).

4. EXAMPLES

4.1 Infant Response to Attachment of Mother

Consider the two-way contingency table of Table 1 that cross-classifies a mothers attachment to her child, and the child's response to their mother's level of attachment. The data is based on an extensive study of mother-child attachment conducted by van IJzendoorn (1995), while Kroonenberg and Lombardo (1999) considered the data in

their exposition of NSCA. The four column categories are a result of the Adult Attachment Interview (George, Kaplan and Main, 1985) while the four response categories are observed from the Ainsworth Strange Situation (Ainsworth, Blehar, Waters and Wall, 1978).

Table 1: Cross-classification of the attachment classification of a mother and her infant

<i>Infant Response</i>	<i>Mother's Attachment Classification</i>			
	<i>Dismissing</i>	<i>Autonomous</i>	<i>Preoccupied</i>	<i>Unresolved</i>
<i>Avoidant</i>	62	29	14	11
<i>Secure</i>	24	210	14	39
<i>Resistant</i>	3	9	10	6
<i>Disorganised</i>	19	26	10	62

Since the mother's level of attachment to their child will influence how the child responds, the column variable (*Mother's Attachment Classification*) will be treated as a predictor variable, while the row variable (*Infant Response*) will be treated as a response variable. Therefore one may perform NSCA on Table 1, as Kroonenberg and Lombardo (1999) did. We shall begin our analysis of the data by applying SVD to each of the π_{ij} values derived from the table. Doing so, the Goodman-Kruskal tau index is $\tau = 0.19897$ where its numerator is $\tau_{num} = 0.12589$. Therefore the C-statistic is 326.51 and is highly statistically significant (p-value = 0.001) indicating that there is a strong asymmetric relationship between the two variables. However, it is unclear how the predictor variable influences the outcome of the response variable.

Table 2: Absolute and relative contribution of each row category to τ_{num}

	<i>Avoidant</i>	<i>Secure</i>	<i>Resistant</i>	<i>Disorganised</i>	Total
Absolute	0.03507	0.06012	0.00244	0.02826	0.12589
Relative (%)	27.85	47.76	1.94	22.45	100

Table 3: Absolute and relative contribution of each column category to τ_{num}

	<i>Dismissing</i>	<i>Autonomous</i>	<i>Preoccupied</i>	<i>Unresolved</i>	Total
Absolute	0.04418	0.04225	0.00745	0.03201	0.12589
Relative (%)	35.10	33.56	5.91	25.43	100

A preliminary analysis of the major contributors to τ_{num} reveals that *Dismissing*, *Autonomous* and *Secure* each contribute to over a third (relative to the other categories within their variable) of the statistic. Table 2 and Table 3 show the absolute and relative contribution of each category to τ_{num} .

A preliminary graphical summary of how a child responds to the level of attachment a mother has towards him/her can be seen from the P-plot of Figure 1. This figure shows that all four child classification levels appear to be influenced by different

mother classifications. Although a *Resistant* child appears to be positively associated with mothers whose attachment levels are classified as *Preoccupied* while a *Secure* child is positively associated by mothers with an *Autonomous* attachment classification. These mothers do not tend to lead to *Resistant*, *Avoidant* or *Disorganised* children. This finding is influenced by the relatively large cell occupied by (*Secure*, *Autonomous*)

Figure 1: P-plot of Table 1

Of course, Tables 2 and 3, or Figure 1, give no simultaneous indication of the relationship between the two variables and their impact upon τ_{num} . To gain a more informative summary a non-symmetrical correspondence plot may be constructed. This plot, given by Figure 2, shows that, given the marginal classification of the column categories, *Unresolved* and *Secure* also contribute to the significance of τ_{num} .

Figure 2 also shows that *Resistant* does not contribute to the association between the attachment between the mother and her infant child. This effect is not surprising since it contributes only 1.94% of τ_{num} (see Table 2). Figure 2 also shows that a mother's attachment that is classified as *Autonomous* is associated with a child's *Secure* attachment. Such a conclusion is apparent by considering the P-plot of Figure 1. However, while not reflected in the P-plot, or Tables 2 and 3, the non-symmetrical correspondence plot of Figure 2 graphically accounts for 97.63% of the asymmetrical association that exists between the response and predictor variables.

Figure 2: Non-symmetrical Correspondence Plot of Table 1

To determine which predictor categories are consistent with what is expected under the hypothesis of zero predictability, we can superimpose onto Figure 2 the predictor confidence circles. By considering (4), the radii for each of the column radii lengths are

$$r^J = (0.10825, 0.06796, 0.16238, 0.10356)^T$$

Figure 3 gives the non-symmetrical correspondence plot with these 95% confidence circles included.

The confidence circles in Figure 3 show those predictor categories (*Mother Attachment Classification*) that are significant in influencing the attachment classification of the children – all of the predictor (column) categories influence in some way the outcome of the response (row) categories. In particular, an *Autonomous* mother will certainly lead to a *Secure* child and it is the only predictor for that child. It also appears that a *Preoccupied* mother is likely to lead to an *Avoidant* child.

Figure 3: Non-symmetrical Correspondence Plot of Table 1, with the 95% confidence circles for both the row and column categories.

Note that for Table 1 $\psi = 0.3333 \times 0.12589 = 0.04196$ while each of the squared singular values are $\lambda_1^2 = 0.08739$, $\lambda_2^2 = 0.03552$ and $\lambda_3^2 = 0.00298$. Therefore it appears that only one dimension (accounting for 69.42%) is enough to represent the relationship. However since the inclusion of a second dimension is very straight forward, and improves the summary by accounting a further 28.21% of τ_{num} , it is included in the plot. When there are a large number of categories in each variable ψ becomes a very important numerical summary for the correspondence plot.

In terms of “outliers”, $\xi_1 = 9.142899$ (*Dismissing*), $\xi_2 = 10.36975$ (*Autonomous*), $\xi_3 = 3.532511$ (*Preoccupied*) and $\xi_4 = 7.726272$ (*Unresolved*). Therefore, all the predictor categories are significant in their influence on the response variable – a conclusion verified by considering the confidence circles in Figure 3. However *Preoccupied* is the least dominant - this is evident since the radius of its confidence circle is the largest of the four categories. Using the “outlier” measure of $\xi > 2.58$, all the categories may be considered “outliers”. However, it is more reasonable to say that all of the mother attachment (column) categories greatly influence (with 99% confidence) the outcome of their child’s classification (row categories).

4.2 Application of NSCA with linear constraints

Consider now Table 4 which was considered in the analysis of Böckenholt and Böckenholt (1990). It cross-classifies 3184 individuals according to their religion, years of formal education and general attitude towards abortion where each of these variables consists of three categories. The data was also considered by Haberman (1979, pg 399) who treated it as a 3×9 contingency table by concatenating the *Religion* and *Education* variables, and we shall perform our analysis of Table 4 in the same manner.

Table 4: Cross-classification of 3184 individuals according to religion, years of education and attitude towards abortion

Attitude Towards Abortion	<i>Religion</i>								
	Northern Protestant			Southern Protestant			Catholic		
	≤ 8	9 - 12	≥ 13	≤ 8	9 - 12	≥ 13	≤ 8	9 - 12	≥ 13
Positive	49	293	244	27	134	138	25	172	93
Neutral	46	140	66	34	98	38	40	103	57
Negative	115	277	100	117	167	73	88	312	135
Total	210	713	410	178	399	249	153	587	285

For Table 4, the numerator of the Goodman-Kruskal index is 0.03106. The C-statistic of Table 4 is therefore 310.5165, and when compared with the theoretical chi-squared statistic with 16 degrees of freedom indicates that there is ample evidence to conclude that a person's religion and number of years of education can influence their attitude towards abortion (p-value < 0.001).

Suppose we define the matrix H as

$$H = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ -1 & 0 & 1 & 1 \\ 1 & 0 & -1 & 1 \\ 0 & 0 & -1 & 1 \\ -1 & 0 & -1 & 1 \\ 0 & 1 & 0 & -2 \\ 0 & 0 & 0 & -2 \\ 0 & -1 & 0 & -2 \end{pmatrix}$$

The first column of the matrix H indicates that a comparison of the age groups " ≤ 8 " (indicated by the 1) and " ≥ 13 " (indicated by the -1) is being made for Northern and Southern protestants. Since we are not interested in the classification of the Catholics

the last three elements in the first column are zero. Similarly, the second column looks at the age groups “ ≤ 8 ” and “ ≥ 13 ” for only the Catholics while the third column allows for a comparison of the two protestant groups. The last column compares those that are protestant with the Catholics. Implementing the procedure described in Section 3.8 yields the correspondence plot of Figure 4.

Figure 4: Non-symmetrical Correspondence Plot of Table 4 with Linear Constraints

Figure 4 shows that those Catholics who participated in the study are more homogeneous than the Protestants towards their attitude to abortion, when taking into account the three levels of educational background. Generally the Catholics in the study tend to have a more negative opinion of abortion than the Protestant groups, although those with at least thirteen years of education tended to remain fairly neutral on the subject. Figure 4 also shows that Protestants, both Southern and Northern, who are educated with at least 13 years of formal study tend to have a positive attitude towards abortion. However the two Protestant groups are quite different in their attitude for those with between nine and twelve years of formal education. Southern Protestants who fall into this educational category tend to have a fairly neutral attitude towards abortion while their Northern Protestants counterparts have a more positive attitude.

If one compares Figure 4 with the non-symmetrical correspondence plot without imposing linear constraints (Figure 5), it can be seen that, just as was the case with linear constraints considered, the attitude towards abortion by the Catholics in the study is very different to the two Protestant groups for all educational backgrounds. However the years of education amongst the Catholics in the study suggests that their correspondence is quite different when compared to the configuration of points in

Figure 4. While Figure 4 shows that Catholics with ≤ 8 years of education and between nine and twelve years of education have a similar attitude towards abortion, without imposing the constraints suggests that their attitude is very different. Also, Figure 5 suggests that Catholics with ≤ 8 years of formal education are associated more with a neutral attitude than what Figure 4 shows.

Figure 5: Non-symmetrical Correspondence Plot of Table 4 without Linear Constraints

5 Discussion

In this paper we have discussed the development and application of various statistical tools that can be used for performing classical NSCA. These tools can be used to further explore the nature of the association between two asymmetric categorical variables. However further work on these technique should be explored. While this paper discusses the analysis of two-way contingency tables, future research should focus on the development of these techniques for the analysis of multivariate categorical variables. In fact, we may consider the application of these procedures to the partition of multivariate versions of the Goodman-Kruskal tau index, including the Marcotorchino index (Marcotorchino, 1985) or the Gray-Williams statistics (Gray and Williams, 1975). The adaptation of these tools for the analysis of asymmetric association between the variables of an ordinal two-way and multi-way contingency table can also be considered.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous referees for their comments and suggestions during the preparation of this paper.

References

- AINSWORTH, M. D., BLEHAR, M. C., WATERS, E. and WALL, S. (1978), *Patterns of Attachment: A Psychological Study of the Strange Situation*, Hillsdale, NJ: Erlbaum.
- AGRESTI, A. (1996), *An Introduction to Categorical Data Analysis*, New York: Wiley.
- ANDREWS, D. F. (1972), "Plots of high-dimensional data", *Biometrics*, 28, 125-136.
- BAR-HEN, A. and MORTIER, F. (2004), "Influence and sensitivity measures in correspondence analysis", *Statistics*, 38, 207-215.
- BEH, E. J. (2001), "Confidence circles for correspondence analysis using orthogonal polynomials", *Journal of Applied Mathematics and Decision Sciences*, 5, 35-45.
- BEH, E. J. (2004), "Simple correspondence analysis: A bibliographic review", *International Statistical Review*, 72, 257-284.
- BENASSEN, J. (1993), "Perturbational aspects in correspondence analysis", *Computational Statistics and Data Analysis*, 15, 393-410.
- BOCKENHOLT, U. and BOCKENHOLT, I. (1990) "Canonical analysis of contingency tables with linear constraints", *Psychometrika*, 55, 633-639.
- CHERNOFF, H. (1973), "The use of faces to represent points in K-dimensional space graphically", *Journal of the American Statistical Association*, 68, 361-368.
- CHOULAKIAN, V. and ALLARD, J. (1998), "The Z-plot: A graphical procedure for contingency tables with an ordered response variable", In *Visualization of Categorical Data*, (Blasius, J. and Greenacre, M. eds), Academic Press: London, 99-105.
- D'AMBRA L. and LAURO, N. (1989), "Non symmetrical analysis of three way contingency tables". In *Multiway Data Analysis*, Eds. R. Coppi and S. Bolasco, North Holland, 301 –315.
- D'AMBRA, L. and LAURO, N. (1992), "Exploratory non symmetric data analysis", *Italian Journal of Applied Statistics*, 4, 511-529.
- D'AMBRA, L. BEH, E. J. and AMENTA, P. (2005), "CATANOVA for two-way contingency tables with ordinal variables using orthogonal polynomials", *Communications in statistics (Theory and Methods)*, 34, 1755-1769.
- FRIENDLY, M. (1994), "Mosaic displays for multi-way contingency tables", *Journal of the American Statistical Association*, 89, 190-200.
- FRIENDLY, M. (1999), "Extending mosaic displays: Marginal, conditional, and partial views of categorical data", *Journal of Computational and Graphical Statistics*, 8, 373-395.
- GABRIEL, K. R. (1971), "The biplot graphic display of matrices with application to principal component analysis", *Biometrika*, 58, 453-467.
- GEORGE, C., KAPLAN, N. and MAIN, M. (1985), *Adult Attachment Interview*, Unpublished manuscript, University of California, Berkeley.
- GOODMAN, L. A. (1986), "Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables", *International Statistical Review*, 54, 243-309.
- GOODMAN, L. A. and KRUSKAL, W. H. (1954), "Measures of association for cross-classifications", *Journal of the American Statistical Association*, 49, 732-764.

- GRAY, L. N. and WILLIAMS, J. S. (1975), "Goodman and Kruskals tau b: Multiple and partial analogs", *American Statistical Association (Proceedings of the Social Statistics Section)*, 444 – 448.
- GREENACRE, M. J. (1984), *Theory and Application of Correspondence Analysis*, London: Academic Press.
- HABERMAN, S. J. (1979) *Analysis of Quantitative Data (Volume 2)*, New York: Academic Press..
- KHATTREE, R., and NAIK, D. N. (2002a), "Andrews plots for multivariate data: Some new suggestions and applications", *Journal of Statistical Planning and Inference*, 100, 411-425.
- KHATTREE, R., and NAIK, D. N. (2002b), "Association in contingency tables, correspondence analysis, and (modified) Andrews plots", In *Goodness-of-Fit Tests and Validity of Models* (Balakrishnan, N., Huber, C., Nikulin, M. and Mesbah, M. eds), Boston: Birkhauser, p. Chapter 23.
- KIM, H. (1992), "Measures of influence in correspondence analysis", *Journal of Statistical Planning and Inference*, 40, 201-217.
- KIM, H. (1994), "Influence functions in multiple correspondence analysis", *Korean Journal of Applied Statistics*, 7, 69-74.
- KROONENBERG, P. and LOMBARDO, R. (1999), „Nonsymmetric correspondence analysis: A tool for analysing contingency table with a dependence structure“, *Multivariate Behavioral Research*, 34, 367 – 397.
- LEBART, L., MORINEAU, A. and WARWICK, K. M. (1984) *Multivariate Descriptive Statistical Analysis*, New York: John Wiley & Sons.
- LIGHT, R. J. and MARGOLIN, B. H. (1971), "An analysis of variance for categorical data", *Journal of the American Statistical Association*, 66, 534-544.
- LOMBARDO, R., BEH, E. J. and D'AMBRA, L. (2007), "Non-symmetric correspondence analysis with ordinal variables", *Computational Statistics and Data Analysis*, 52, 566-577.
- LOMBARDO, R., CARLIER, A. and D'AMBRA, L. (1996), "Nonsymmetric correspondence analysis for three-way contingency tables", *Methodologica*, 4, 59-80.
- MARCOTORCHINO, F. (1985), *Utilisation des Comparaisons par Paires en Statistique de Contingencies: Parte III*, Report # F 071, Etude du Centre Scientifique, IBM, France.
- MARDIA, K. V., KENT, J. T., and BIBBY, J. M. (1982), *Multivariate Analysis*, London: Academic Press.
- RINGROSE, T. J. (1992), "Bootstrapping and correspondence analysis in archaeology", *Journal of Archaeological Science*, 19, 615-629.
- RINGROSE, T. J. (1996), "Alternative confidence regions for canonical variate analysis", *Biometrika*, 83, 575-587.
- ROVAN, J. (1994), "Visualizing solutions in more than two dimensions", In *Correspondence Analysis in the Social Sciences*, (Greenacre, M. and Blasius, J. eds), London: Academic Press, 210-229.
- SINGH, B. (1996), "On CATANOVA method for analysis of two-way classified nominal data", *Sankhya B*, 58, 379--388
- TAKANE, Y. and SHIBAYAMA, T. (1991), "Principal component analysis with external information on both subjects and variables", *Psychometrika*, 56, 97-120.

- TAKANE, Y., YANAI, H. and MAYEKAWA, S. (1991), Relationships among several methods of linearly constrained correspondence analysis, *Psychometrika*, 56, 667-684.
- TALLUR, B. (1986), "Un nouvel indece d'agregation en classification ascendant hierarchique". *Revue de Statistique Appliquée* . 34 (2), 25-35.
- VAN IJZENDOORN, M. H. (1995), "Adult attachment representations, parental responsiveness, and infant attachment. A meta-analysis on the predictive validity of the Adult Attachment Interview", *Psychological Bulletin*, 117, 387-403.