Journal of Applied Statistics

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/cjas20

Studying the dependence between ordinal-nominal categorical variables via orthogonal polynomials

Rosaria Lombardo\textsuperscript{a}, Eric J. Beh\textsuperscript{b} & Antonello D'Ambra\textsuperscript{a}

\textsuperscript{a} Economics Faculty, Second University of Naples, Corso Gran Priorato di Malta, Capua, CE, 81043, Italy
\textsuperscript{b} School of Mathematical and Physical Sciences, University of Newcastle, Callaghan, NSW, 2308, Australia

Available online: 17 Jan 2011

To cite this article: Rosaria Lombardo, Eric J. Beh & Antonello D'Ambra (2011): Studying the dependence between ordinal-nominal categorical variables via orthogonal polynomials, Journal of Applied Statistics, 38:10, 2119-2132

To link to this article: http://dx.doi.org/10.1080/02664763.2010.545118

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: http://www.tandfonline.com/page/terms-and-conditions

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan, sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
Studying the dependence between ordinal-nominal categorical variables via orthogonal polynomials

Rosaria Lombardo\textsuperscript{a*}, Eric J. Beh\textsuperscript{b} and Antonello D’Ambra\textsuperscript{a}

\textsuperscript{a}Economics Faculty, Second University of Naples, Corso Gran Priorato di Malta, - Capua, CE 81043, Italy; \textsuperscript{b}School of Mathematical and Physical Sciences, University of Newcastle, Callaghan, NSW, 2308 Australia

(Received 8 April 2010; final version received 9 November 2010)

In situations where the structure of one of the variables of a contingency table is ordered recent theory involving the augmentation of singular vectors and orthogonal polynomials has shown to be applicable for performing symmetric and non-symmetric correspondence analysis. Such an approach has the advantage of allowing the user to identify the source of variation between the categories in terms of components that reflect linear, quadratic and higher-order trends. The purpose of this paper is to focus on the study of two asymmetrically related variables cross-classified to form a two-way contingency table where only one of the variables has an ordinal structure.

Keywords: ordered categorical variables; non-symmetric correspondence analysis; bivariate moment decomposition; singular value decomposition; orthogonal polynomials

1. Introduction

For a two-way contingency table, identifying a meaningful category dependence structure between variables can be made by considering a non-symmetric correspondence analysis (NSCA) of the data (see, for example [12,18]). NSCA is a useful tool for graphically depicting the relationship between variables when one is assumed to be logically dependent from the other one. This technique involves decomposing the non-symmetric measure of association given by the tau index [15,16,20], using singular value decomposition (SVD). In the case where both variables are ordered a special partition of the tau index using what is referred to as bivariate moment decomposition (BMD [8,13]) can be considered (doubly ordered NSCA, [22]). From this analysis, we obtain a confirmatory approach to NSCA and a visual summary of the dependence structure between the categories of two ordered variables using orthogonal polynomials [14]. This partition,
and the graphical summary that can be derived from it, has the advantage of allowing the user to decompose the total inertia into meaningful components. The methodology presented in this paper, referred to as singly ordinal non-symmetric correspondence analysis (SONSCA) permits one to combine the summaries obtained from SVD and BMD of the data. It allows the user to visualize and identify the primary causes (if they exist) of the dependent relationship between the categories of two variables where one of them is assumed to be an ordinal categorical variable. Both the total inertia (sum of the squared singular values) and the partial inertia (each squared singular value) can therefore be expressed by components that reflect within- and between-variable variation in terms of location, dispersion and higher-order moments. There are many other measures of association that may be considered for a singly ordered two-way contingency table - see, for example, [7,28]. However, these measures do not incorporate a graphical analysis to describe the association between categorical variables.

This paper is divided into six further sections. A brief description of classical NSCA will be made in Section 2, while Section 3 will discuss the details of its singly ordered variant. Within this section, the recurrence relation used to generate the orthogonal polynomials will be given as well as the plotting system for visualising the association in the contingency table. Section 4 will provide some insight into the interpretation of the correspondence plot, including intra-variable distances, while Section 5 will present ways to formally test for association in asymmetrically related categorical variables. Section 6 will illustrate, by way of an example, the application of the technique, and some final remarks will be made in Section 7.

2. Classical non-symmetrical correspondence analysis

When a one-way dependence, or asymmetric, structure exists between the categories of a two-way contingency table, NSCA is an appropriate method to use for identifying the nature of this relationship. NSCA can be viewed as an extension of symmetric, or classical, correspondence analysis (CA) to a more general linear context by adopting a particular metric.

Suppose we have an \( I \times J \) contingency table \( N \) that cross-classifies \( n \) individuals/units. In this paper, \( I \) is the number of row categories that form a response variable, and \( J \) is the number of column categories that form a predictor variable. Denote the number of entries in the \((i, j)\)th cell by \( n_{ij} \), for \( i = 1, \ldots, I \) and \( j = 1, \ldots, J \), and the proportion allocated in this cell as \( p_{ij} = n_{ij}/n \).

Let \( p_i = \sum_{i=1}^I p_{ij} \) be the proportion of individuals/units classified into the \( i \)th row and let \( p_j = \sum_{j=1}^J p_{ij} \) be the column marginal proportion for the \( j \)th category. Let \( l \) and \( p_j \) be the general elements of the metrics in the space \( \mathbb{R}^I \) and \( \mathbb{R}^J \), respectively.

When identifying the asymmetric association between ordinal/nominal-scale variables of a two-way contingency table using NSCA one may visualize this association in a low-dimensional space. NSCA is based on the optimization of the numerator of the measure of predictability that belongs to the class of measures based on the proportional reduction of error in prediction (2). Suppose we define \( \pi_{ij} = p_{ij}/p_j - p_i \) to be the difference between the unconditional row marginal prediction \( p_i \) and the conditional prediction \( p_{ij}/p_j \) for the \((i, j)\)th cell of the table. In order to analyse the increase in predictability of the response variable given the presence of the predictor variable, the Goodman–Kruskal tau index

\[
\tau = \frac{\sum_{i=1}^I \sum_{j=1}^J p_{ij}(p_{ij}/p_j - p_i)^2}{1 - \sum_{i=1}^I p_i^2} = \frac{N\tau}{1 - \sum_{i=1}^I p_i^2}.
\]

can be used [16]. Since the denominator of \( \tau \) measures the overall error in prediction and does not depend on the predictor categories, we focus our attention on the
numerator

\[ N_\tau = \sum_{i=1}^{I} \sum_{j=1}^{J} p_{i*} \left( \frac{p_{ij}}{p_{i*}} - p_{i*} \right)^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} p_{ij} \pi_{ij}^2 \]  

(1)

which is the absolute measure of overall predictability of the rows given the columns.

In classical NSCA [12,18,24] the full rank graphical representation of row and column categories can be found by first considering the generalized SVD of \( \pi_{ij} \) such that

\[ \pi_{ij} = \sum_{s=1}^{S} \lambda_s a_{is} b_{js} \]

for \( S = \min(I, J) - 1 \) where \( \lambda_s \) is the \( s \)th singular value of \( \pi_{ij} \) and can be expressed by

\[ \lambda_s = \sum_{i=1}^{I} \sum_{j=1}^{J} p_{i*} \pi_{ij} a_{is} b_{js}. \]

Here \( a_s = \{ a_{is} : i = 1, \ldots, I \} \) and \( b_s = \{ b_{js} : j = 1, \ldots, J \} \) are the orthonormal singular vectors in a unit metric and a weighted metric, respectively, so that

\[ \sum_{i=1}^{I} a_{is} a_{is'} = \begin{cases} 1 & s = s', \\ 0 & s \neq s', \end{cases} \quad \text{and} \quad \sum_{j=1}^{J} b_{js} b_{js'} = \begin{cases} 1 & s = s', \\ 0 & s \neq s'. \end{cases} \]

Furthermore, it follows that \( N_\tau \) can be expressed in terms of sum of squares of the singular values such that

\[ N_\tau = \sum_{s=1}^{S} \lambda_s^2. \]  

(2)

For the graphical presentation of the association between the predictor and response categories plotting systems can be derived. Under NSCA, we can define the coordinates for the \( i \)th response (row) and \( j \)th predictor category (column) for the \( s \)th dimension of a correspondence plot by

\[ f_{is} = a_{is} \lambda_s, \quad g_{js} = b_{js} \lambda_s, \]

respectively. The measure of variation, \( N_\tau \), can be expressed as the sum of square of these coordinates by

\[ N_\tau = \sum_{i=1}^{I} \sum_{s=1}^{S} f_{is}^2 = \sum_{j=1}^{J} \sum_{s=1}^{S} p_{i*} g_{js}^2. \]

This infers that those categories positioned far from the origin contribute more to \( N_\tau \) while those that lie close to the origin indicate their lack of contribution to the hypothesis of predictability of the rows given the columns.

3. Singly ordinal non-symmetrical correspondence analysis

The classical approach of NSCA, as described above, is applicable for nominal-scale categorical variables. However, when one of the categorical variables exhibits an ordinal structure we refer to the methodology discussed in the subsequent sections as singly ordered NSCA (SONSCA).
In the case where the predictor variable is ordered we refer to the analysis as SONSCA1, while the analysis for an ordinal-scale response variable is referred to as SONSCA2. The SONSCA approach relies on the partition of $N_\tau$ using a mixed decomposition, which is the amalgamation of elements obtained from SVD and BMD. Such a decomposition for mixed contingency tables requires the generation of orthogonal polynomials for the ordered categorical variable, and so the following section will define a recurrence relation to calculate them.

### 3.1 Orthogonal polynomials

The set of polynomials for each variable of a contingency table, based on the simple recurrence formula of Emerson [14], are orthogonal over a set of points with arbitrary spacing and weighting. The arbitrary spacing permits the use of polynomials in experiments where the categories are unequally spaced on the independent variable, or where there exists an unequal number of observations at the different treatment levels. The decomposition (BMD) of the transformed contingency table using Emerson’s polynomials can be considered to be equivalent to the Gram–Schmidt method of orthogonality by a set of linearly independent vectors. Recent applications of these polynomials have been carried out in ordered symmetric and non-symmetric correspondence analysis [1–3,8,9,21,22,25].

Suppose that our contingency table, $N$, consists of ordered column categories and this ordinal structure is reflected by the set of column scores $\{s(j) : j = 1, 2, \ldots, J\}$. Setting $b_{j(1)}^* = 0$ and $b_{j(0)}^* = 1$, and using the predetermined set of scores, the column orthogonal polynomials of generic degree $v$ are calculated using the following general recurrence formula:

$$b_{j(v)}^* = S_v[(s(j) - T_v)b_{j(v-1)}^* - V_v b_{j(v-2)}^*],$$

where

$$T_v = \sum_{j=1}^{J} p_{\bullet \cdot} s(j) b_{j(v-1)}^* b_{j(v-1)}^2,$$

$$V_v = \sum_{j=1}^{J} p_{\bullet \cdot} s(j) b_{j(v-1)}^* b_{j(v-2)}^* b_{j(v-2)}^2$$

and

$$S_v = \left\{ \sum_{j=1}^{J} p_{\bullet \cdot} s(j)^2 b_{j(v-1)}^* b_{j(v-1)}^2 - T_v^2 - V_v^2 \right\}^{-1/2},$$

for $v = 0, 1, \ldots, J - 1$. These polynomials have the property

$$\sum_{j=1}^{J} p_{\bullet \cdot} b_{j(v)}^* b_{j(v')}^* = \begin{cases} 1 & v = v', \\ 0 & v \neq v' \end{cases}$$

and are orthogonal with respect to the column metric $p_{\bullet \cdot}$ in $\mathbb{R}^J$.

Consider the set of column polynomials $\{b_{j(1)}^* : j = 1, \ldots, J\}$. When ordinal scores are used in their calculation, these polynomials have a linear structure. Therefore, any quantity that involves this set of polynomials describes the linear behaviour of the column variable. Similarly, $\{b_{j(2)}^* : j = 1, \ldots, J\}$ has a quadratic structure and shows this behaviour of the column variable. These polynomials therefore reflect sources of variation between the row categories in
terms of the location and dispersion moments. Moments greater than the location or dispersion can be considered by taking into account values of \( u \) and \( v \) greater than 2.

When the row categories have an ordered structure, a similar set of polynomials, \( \{a^*_{i(u)} : u = 1, \ldots, I - 1, i = 1, \ldots, I\} \), can be derived. Alternative scoring schemes, such as midrank scores and Nishisato scores [2,27,31], can be used to reflect the ordinal structure of a variable. However, for simplicity, this paper will consider the use of natural scores. For the column variable consisting of \( J \) ordered categories, the natural scores, \( \{s(j) = j : j = 1, \ldots, J\} \), are used.

### 3.2 SONSCA for an ordered predictor variable – SONSCA1

When a two-way contingency table with only one ordered variable is considered, a decomposition using a combination of singular vectors and orthogonal polynomials can be applied. For SONSCA1, \( N_\tau \) can be partitioned using orthogonal polynomials for the ordered column categories and singular vectors for the unordered row categories. This is akin to the decomposition of the Pearson chi-squared statistic for singly ordered contingency tables [3]. As a consequence of this partition of \( N_\tau \), each eigenvalue (i.e. the inertia associated to each axis) can be partitioned into components that reflect linear, quadratic and higher-order sources of variation. Thus, it is possible to partition the measure of the absolute increase in predictability for a row category given a column, and visualize linear, quadratic or higher-order trends. By considering SONSCA using singular vectors for the nominal-scale row categories and orthogonal polynomials for the ordinal-scale column categories, \( N_\tau \) can be decomposed so that

\[
N_\tau = \sum_{i=1}^{I} \sum_{j=1}^{J} p_{ij} \pi_{ij}^2 = \sum_{s=1}^{S} \sum_{v=1}^{I-1} z_{s(v)}^2
\]  

\[ (3) \]

where

\[
z_{s(v)} = \sum_{i=1}^{I} \sum_{j=1}^{J} p_{ij} \pi_{ij} a_{is} b^*_v.
\]  

\[ (4) \]

By considering Equation (4), the left singular vectors \( a_i = \{a_{is} : i = 1, \ldots, I\} \) are associated with the row categories, while the set of orthogonal polynomials \( b^*_v = \{b^*_j : j = 1, \ldots, J\} \) are associated with the ordered columns. For example, when \( v = 1 \), \( z_{s(1)} \) describes the linear, or location, column component on the inertia of the \( s \)th principal axis. Similarly, when \( v = 2 \), \( z_{s(2)} \) describes the quadratic, or dispersion, column component on this inertia.

For SONSCA1, the graphical presentation of the association between, and within, the predictor and response variables can be made by considering the row and column principal coordinates

\[
\tilde{f}_{iv} = \sum_{s=1}^{S} a_{is} z_{s(v)}, \quad \tilde{g}_{js} = \sum_{v=1}^{J-1} b^*_j z_{s(v)}
\]

so that

\[
N_\tau = \sum_{i=1}^{I} \sum_{v=1}^{I-1} \tilde{f}_{iv}^2 = \sum_{j=1}^{J} \sum_{s=1}^{S} p_{ij} \tilde{g}_{js}^2.
\]

The above equations show that, just as it was with the classical approach to NSCA, those categories positioned close to the origin indicate their contribution to the null hypothesis of no predictability. Similarly those coordinates situated at a far distance from the origin highlight the importance of those categories to the dependence relationship which may exist between the two variables.
Suppose we consider $N_\tau$, which can be calculated by considering Equations (2) and (3). By evaluating these expressions so that

$$N_\tau = \sum_{s=1}^{S} \sum_{v=1}^{J-1} z_{s(v)}^2 = \sum_{s=1}^{S} \lambda_s^2$$

the square of the $s$th largest singular value for classical NSCA can be decomposed into linear, dispersion and higher-order column components such that

$$\lambda_s^2 = \sum_{v=1}^{J-1} z_{s(v)}^2$$

For example, the square of the first singular value (when $s = 1$) can be partitioned into linear, quadratic and higher-order moments such that

$$\lambda_1^2 = z_{1(1)}^2 + z_{1(2)}^2 + \cdots + z_{1(J-1)}^2$$

By aggregating across the row indices of $z_{s(v)}$, components are obtained that reflect various sources of variation. It is evident that the sum of the column component values is equal to the inertia of the contingency table. In general the $s$th column component for the association is $\sum_{v=1}^{J-1} z_{s(v)}^2$. These components allow the researcher to examine the linear, and non-linear sources of variation for the ordered categorical variable. Generally the $s$th order column component is $\sum_{s=1}^{S} z_{s(v)}^2$ and its statistical significance with each axis can be determined using the C-statistic (CATANOVA test, [11,23,29]).

### 3.3 SONSCA for an ordered response variable – SONSCA2

When the response variable consists of ordinal-scale categories, but the column responses are considered to be nominal, the numerator of $\tau$ can be partitioned using the row orthogonal polynomials $(a_{i(u)}^n)$ and the column singular vectors $(b_{js})$ such that

$$N_\tau = \sum_{u=1}^{I-1} \sum_{s=1}^{S} \tilde{z}_{(u)s}^2,$$

where

$$\tilde{z}_{(u)s} = \sum_{i=1}^{I} \sum_{j=1}^{J} p_{•j} a_{i(u)}^n \tau_{ij} b_{js}.$$

The row and column profile coordinates may therefore be defined as

$$\tilde{f}_{is} = \sum_{u=1}^{I-1} a_{i(u)}^n \tilde{z}_{(u)s}, \quad \tilde{g}_{ju} = \sum_{s=1}^{S} \tilde{z}_{(u)s} b_{js}.$$  

With this system of coordinates, it can be shown that

$$N_\tau = \sum_{i=1}^{I} \sum_{s=1}^{S} \tilde{f}_{is}^2 = \sum_{j=1}^{J-1} \sum_{u=1}^{I-1} p_{•j} \tilde{g}_{ju}^2.$$  

In SONSCA2 the inertia associated with the $s$th axis is given by

$$\lambda_s^2 = \sum_{u=1}^{I-1} \tilde{z}_{(u)s}^2.$$
so that the singular values can be partitioned into terms that summarize the variation in the row categories using location, dispersion and higher-order components.

4. Graphical representation in SONSCA

4.1 Centering of coordinates

A feature of using the SONSCA row and column principal coordinates is that they are centred about the centroid of the correspondence plot. That is, for SONSCA1, the mean distance of the $I$ row coordinates from the origin, with respect to the unit metric, is

\[
\sum_{i=1}^{I} \tilde{f}_{iv} = \sum_{i=1}^{I} \sum_{s=1}^{S} a_{is} \tilde{z}_{s(v)} = 0.
\]

Similarly, the column principal coordinates are centred around the origin, with respect to its metric $p_{•j}$. That is

\[
\sum_{j=1}^{J} p_{•j} \tilde{g}_{js} = \sum_{j=1}^{J} \sum_{v=1}^{J-1} b_{j(v)}^{*} \tilde{z}_{s(v)} = 0.
\]

4.2 Distance of two predictor categories

One of the primary reasons for considering the NSCA of a contingency table is that a graphical summary of the asymmetric variables can be made. Such a summary allows the researcher to identify those column categories which help to predict the outcome of the row (response) categories. For the predictor (column) variable, by observing the squared distances between the profiles of the $j$th and $j’$th column categories, we can identify those predictor categories that are similar or different. For SONSCA1, the Euclidean squared distance between the $j$th and $j’$th column category is

\[
d_{2}^{2}(j, j’) = \sum_{i=1}^{I} \left[ \left( \frac{p_{ij}}{p_{•j}} - \frac{p_{ij’}}{p_{•j’}} \right) \right]^{2} = \sum_{s=1}^{S} \left( \tilde{g}_{js} - \tilde{g}_{j’s} \right)^{2}.
\]

Therefore, two predictor categories that have a similar level of predictability on the response categories will be situated closely to one another in the correspondence plot. Similarly, column coordinates that are positioned at a distance from one another indicate that they have a different impact on the response categories. Therefore, the plotting system used by considering the coordinates above ensures that the property of distributional equivalence, as discussed in literature [17,19], holds.

Similarly, the squared Euclidean distance between the $i$th and $i’$th row categories is

\[
d_{2}^{2}(i, i’) = \sum_{v=1}^{I-1} (\tilde{f}_{iv} - \tilde{f}_{i’v})^{2}.
\]

The distance along each dimension of the correspondence plot can also be measured. For the $v$th dimension, the squared distance of the $i$th and $i’$th row profile coordinates is

\[
d_{2}^{2}(i, i'|v) = (\tilde{f}_{iv} - \tilde{f}_{i’v})^{2}.
\]

Therefore, two row categories situated closely to one another along a particular axis will contribute equally to the row component associated with that axis. For example, consider the difference in
the location of the predictor variables, i.e. when $v = 1$. When $d^2_i(i, i' | v = 1) \approx 0$ then the $i$th and $i'$th response categories have a similar location. Similarly, $d^2_i(i, i' | v = 2) \approx 0$ indicates that these categories have a similar spread.

### 4.3 Distance of predictor categories from the origin

The distance of the $j$th predictor category from the origin is an important part in identifying where the strength, or lack, of association exists between the two variables. Recall, that from classical NSCA, those column points situated close to the origin indicate a lack of predictability for the rows. Similarly, those columns that greatly influence the prediction of the row categories will be situated at some distance from the origin. For SONSICA1, consider the squared Euclidean distance of the $j$th predictor category from the origin

$$d^2_J(j, 0) = \sum_{i=1}^{I} \left( \frac{p_{ij}}{p_{\cdot j}} - \frac{p_{i \cdot}}{p_{\cdot \cdot}} \right)^2 = \sum_{s=1}^{S} g^2_{js}.$$

This distance reflects the influence of the $j$ column category to $N_\tau$ since

$$N_\tau = \sum_{j=1}^{J} p_{\cdot j}d^2_J(j, 0).$$

### 4.4 Interpretation of axes

In order to display the dependence structure between the row and column categories a correspondence plot can be constructed using the profile coordinates defined above. Generally the first (location) and second (dispersion) components will be the most statistically significant and so the correspondence plot will only need to consist of the first and second axes to reflect the variation in the profiles. However, sometimes higher-order components can be statistically significant. The interpretation of the axes generated from using orthogonal polynomials provides for a meaningful description of how categorical responses vary from one another within a variable. In particular, the construction of the correspondence plot based on the results from using the polynomials ensures that the product moment correlation, or its non-linear equivalents (see [13]) reflects the nature of association in the plot.

For ordered NSCA, since the $s$th axis reflects any difference in the distribution of profiles in terms of the $v$th (or $u$th) moment, the axis can graphically summarize those categories that are different in terms of their location, dispersion or higher order moment. It can also identify those with a similar $v$th (or $u$th) order moment. For points that are spread along a particular axis, this indicates the large magnitude of the component associated with the axis. Similarly if all of the row coordinates are clustered near the origin, then the distribution of the predictor does not exhibit a significant deviation from the overall distribution.

### 5. Confirmatory NSCA: tests of association

In order to determine the significance of the tau index we compute the C-statistic (CATANOVA test statistic), which allows the user to identify the strength of association between the two asymmetric variables. The C-statistic, originally considered by Light and Margolin [20,26], is defined as

$$C = (n - 1)(I - 1) \frac{N_\tau}{1 - \sum_{i=1}^{I} p_{i \cdot}^2}.$$
and is a random variable from the chi-squared distribution with \((I - 1)(J - 1)\) degrees of freedom at the \(\alpha\) level of significance, \(\chi^2_{\alpha(I-1)(J-1)}\). Note that the link between the Goodman–Kruskal tau index and the CATANOV A test statistic was successively pointed out by Sarndal [30] and Margolin and Light [26].

When either the predictor or response variable is ordered, \(N_t\) can be partitioned by considering either SONSCA1 or SONSCA2. For SONSCA1, the C-statistic can be expressed in terms of \(z_{s(v)}\) so that

\[
C = \frac{(n - 1)(I - 1)}{1 - \sum_{i=1}^{I} p^2_i} \sum_{s=1}^{S} \sum_{v=1}^{J-1} z_{s(v)}^2 \sim \chi^2_{\alpha(I-1)(J-1)}.
\]

By letting

\[
\hat{z}_{s(v)} = \frac{z_{s(v)}}{\sqrt{\frac{(n - 1)(I - 1)}{1 - \sum_{i=1}^{I} p^2_i}}},
\]

then

\[
C = \sum_{s=1}^{S} \sum_{v=1}^{J-1} \hat{z}_{s(v)}^2.
\]

It can be shown the \(\hat{z}_{s(v)}\) are asymptotically standard normally distributed random variables [11].

The C-statistic can also be expressed as a weighted sum of squares of the profile coordinates. This allows one to detect those predictor categories that do not contribute to the predictability of the row categories. Suppose we consider the approach to NSCA outlined by SONSCA1. The C-statistic can be written in terms of the column profile coordinates so that

\[
C = \frac{(n - 1)(I - 1)}{1 - \sum_{i=1}^{I} p^2_i} \sum_{j=1}^{J} \sum_{s=1}^{S} p_{\bullet j} g_{js}^2 \sim \chi^2_{\alpha(I-1)(J-1)}.
\]

Therefore, by rearranging this expression for the \(j\)th column coordinate

\[
\sum_{s=1}^{S} p_{\bullet j} g_{js}^2 \sim \chi^2_{\alpha(I-1)} \frac{1 - \sum_{i=1}^{I} p^2_i}{(n - 1)(I - 1)}.
\]

Since for the higher dimensions, the coordinates will generally be close to zero (since the singular values associated with these dimensions are generally close to zero) the squared Euclidean distance of the \(j\)th predictor category from the origin may be expressed as

\[
g_{j1}^2 + g_{j1}^2 = \chi^2_{\alpha(2)} \frac{1 - \sum_{i=1}^{I} p^2_i}{p_{\bullet j}(n - 1)(I - 1)}
\]

for some \(\alpha\) level of significance. Therefore, if we consider \(\alpha = 0.05\) then the impact of the \(j\)th predictor category in predicting the row (response) categories can be made by considering whether the origin falls within the circular region with given radius

\[
r_j^2 = \sqrt{\frac{5.99(1 - \sum_{i=1}^{I} p^2_i)}{p_{\bullet j}(n - 1)(I - 1)}}.
\]

These regions are referred to as confidence circles [6]. Lebart et al. [19] gave similar circular regions for the classic, or symmetric, approach to correspondence analysis.
Table 1. Cross-classification of analgesic drugs and effectiveness levels.

<table>
<thead>
<tr>
<th>Effectiveness</th>
<th>Analgesic drug</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
</tr>
<tr>
<td>Poor</td>
<td>5</td>
</tr>
<tr>
<td>Fair</td>
<td>1</td>
</tr>
<tr>
<td>Good</td>
<td>10</td>
</tr>
<tr>
<td>VeryG</td>
<td>8</td>
</tr>
<tr>
<td>Excellent</td>
<td>6</td>
</tr>
</tbody>
</table>

6. Example: drug data

In this section, we show the benefits of using orthogonal polynomials to graphically depict and summarize the association between ordinal/nominal variables.

The data in Table 1 concern the effectiveness of four analgesic drugs (the labels of which are randomly assigned, A, B, C and D, respectively) on 121 hospital patients. This contingency table has been analysed in the past; see, for example, [4,10]. However these analyses consider the relationship between the variables to be symmetric, when in fact it seems more appropriate to consider the association asymmetric so that the effectiveness of a drug is logically dependent on what drug is being tested. To measure the drug effectiveness, a five-point scale is used that consists of the categories poor, fair, good, very good, and excellent. These categories are labelled as Poor, Fair, Good, VeryG and Excellent, respectively, in Table 1. We can therefore see that Table 1 consists of ordered row (effectiveness) categories and nominal column (drug) categories. Thus, the aim of our study here is to determine, given the drugs studied, how effective they are in reducing pain. We shall be taking into account the ordinal structure of the Effectiveness variable.

The Goodman–Kruskal tau of Table data 2 is 0.089 and is statistically significant since its associated C-statistic of 42.78 has a $p$-value < 0.0001. Therefore, the type of drug being analysed does impact upon its effectiveness. A graphical representation of this asymmetric association is given by Figure 1.

Figure 1 shows that given that a patient is administered with either drug A or drug B that drug will be very effective. This can be seen from the close proximity of these predictor categories (drugs) to the response categories Excellent and VeryG. Similarly, given that a patient is administered with drugs C and D, their effectiveness is rated as Good to Poor.

It can be seen by viewing the axes of Figure 1 that the first two dimensions of Figure 1 account for more than 97.8% of the asymmetric association between the two variables, as measured by $\tau$. The first axis accounts for the majority (75.4%) of this association, while the second axis explains 22.5% of the association. It may also be noted that the third axis contributes very little to $\tau$ (2.2%) and the fourth axis has virtually a zero contribution (Table 2). Superimposed on the configuration of Figure 1 are the 95% confidence circles for each predictor categories. It shows that all of the drugs make a significant impact upon the effectiveness, as their confidence circles do not include the origin.

Since the effectiveness of the drugs is measured on an ordinal scale, we can perform SONSAC2.

To determine how the predictor categories vary in terms of their location, spread, and higher-order moments, across each of the axis, we can partition the C-statistic into the $z_{(u)}s$ and $z_{(u)}^2$ values, given in Tables 2 and 3, respectively. By considering the $z_{(u)}s$ values in Table 2, the dominant source of variation for the first squared singular value is due to the linear component since

$$\lambda_1 = (0.189)^2 + (0.112)^2 + (-0.066)^2 + (-0.027)^2 = 0.053.$$
Figure 1. Classical NSCA plot of Table 1 with 95% confidence circles for each predictor category.

Table 2. $z_{(u)s}$-values obtained from SONSAC2 of Table 1.

<table>
<thead>
<tr>
<th>$z_{(u)s}$</th>
<th>$s = 1$</th>
<th>$s = 2$</th>
<th>$s = 3$</th>
<th>$s = 4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u = 1$</td>
<td>0.189</td>
<td>-40.005</td>
<td>0.021</td>
<td>0.000</td>
</tr>
<tr>
<td>$u = 2$</td>
<td>0.112</td>
<td>0.043</td>
<td>-0.031</td>
<td>0.000</td>
</tr>
<tr>
<td>$u = 3$</td>
<td>-0.066</td>
<td>0.092</td>
<td>0.010</td>
<td>0.000</td>
</tr>
<tr>
<td>$u = 4$</td>
<td>-0.027</td>
<td>-0.074</td>
<td>-0.007</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Table 3. C-statistic = $z^2_{(u)s}$ illustration.

<table>
<thead>
<tr>
<th>C-statistics</th>
<th>$s = 1$</th>
<th>$s = 2$</th>
<th>$s = 3$</th>
<th>$s = 4$</th>
<th>Total</th>
<th>$p$-value</th>
<th>% inertia of component</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location ($u = 1$)</td>
<td>21.59</td>
<td>0.01</td>
<td>0.26</td>
<td>0.00</td>
<td>21.86</td>
<td>0.00</td>
<td>51.1%</td>
</tr>
<tr>
<td>Dispersion ($u = 2$)</td>
<td>7.59</td>
<td>1.12</td>
<td>0.58</td>
<td>0.00</td>
<td>9.29</td>
<td>0.03</td>
<td>21.7%</td>
</tr>
<tr>
<td>Skewness ($u = 3$)</td>
<td>2.63</td>
<td>5.14</td>
<td>0.06</td>
<td>0.00</td>
<td>7.83</td>
<td>0.05</td>
<td>18.3%</td>
</tr>
<tr>
<td>Kurtosis ($u = 4$)</td>
<td>0.43</td>
<td>3.35</td>
<td>0.03</td>
<td>0.00</td>
<td>3.81</td>
<td>0.28</td>
<td>8.9%</td>
</tr>
<tr>
<td>Total</td>
<td>32.24</td>
<td>9.61</td>
<td>0.93</td>
<td>0.00</td>
<td>42.78</td>
<td>0.00</td>
<td>100%</td>
</tr>
<tr>
<td>% inertia of axes</td>
<td>75.4</td>
<td>22.5</td>
<td>2.2</td>
<td>0.00</td>
<td>100.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Similarly, the dominant source of the principal inertia associated with the second axis is due to the cubic (skewness) component since

$$\lambda_2 = (-0.005)^2 + (0.043)^2 + (0.092)^2 + (-0.074)^2 = 0.016.$$  

To check the statistical significance of these components consider Table 3. By observing the sixth column of Table 3 we can see that the linear, quadratic and cubic components are significant at the 5% level of significance (each at three degrees of freedom). However, the situation is different when we look at the partial inertia explained by each axis. In particular for the first axis, only the linear (location) and quadratic (dispersion) components are significant. The location component
is the main source of variation between the drug categories. The first axis (when $s = 1$) is largely dominated by the location component of the variable and, to a lesser extent, by the dispersion component. For the second axis, only the third (skewness) and fourth (kurtosis) components are significant; for the third and fourth axes there are no statistically significant components.

Figure 2 shows the singly ordered correspondence plot obtained by performing SONSCA2 on Table 1. An important feature of Figure 2 is that it depicts the predictability of the levels of effectiveness given the analgesic drugs considered. Since not all the components play a role in determining the structure of the association, the positions of the predictor (drug) categories of Figures 1 and 2 are not identical. A distinct difference between the two plots is the varying position of the predictor categories mainly along the second axis, although there are small differences along the first axis as well. The different coordinates of the drug categories along the second axis is reflected by the significant dispersion component in Table 3. By taking into account the ordered structure of the effectiveness (row variable) in Figure 2, we can see that the higher levels of effectiveness, VeryG and Excellent, are associated with drug A and drug B, respectively, as in Figure 1. However, the association between the lower level of effectiveness appears different, since drug C and drug D behave in a similar manner in terms of location and spread, they result very close to each other (shifted to the bottom left quadrant of the plot) and their effectiveness is almost Good rather than Fair or Poor. By observing the distance of each category from the origin in Figure 2, drug B is the furthest away from the origin and so is less likely than the other drugs to contribute to the independence, on the other side, the response category Poor is the nearest to the origin and so the worst explained category.

7. Conclusion

The classical approach to NSCA can be generalized to include any type of decomposition of the tau numerator. Recently, new developments in CA and NSCA have allowed the researcher to incorporate the structure of two or more ordered categorical variables [21,22,25] by considering a BMD of $\pi_{ij}$. The focus of this paper has been to study NSCA for cases where only one variable is
ordered, i.e. SONSCA. This involves the mixed decomposition of $\pi_{ij}$ [3,5]. While the mathematical structure of the classical and ordered techniques remain very similar, they yield very different, but still informative, interpretations of the association between predictor and response variables. For example, the benefit of considering a mixed decomposition allows the user to determine the predictability of a row variable given a column variable (SONSCA1), or vice versa (SONSCA2). Identifying different, or similar, categories of a response variable can be made in terms of components that reflect their difference in location, spread, etc. This provides the user with additional information concerning the behaviour and structure of the association in the data that the classical analysis does not.

References


