Proceedings of the Challenges for Analysis of the Economy, the Businesses, and Social Progress

International Scientific Conference

Szeged, November 19-21, 2009

ISBN 978-963-06-9558-9

Edited by
Péter Kovács
Katalin Szép
Tamás Katona

Published by
Unidocument Kft.
www.e-document.hu

Szeged, 2010

UNIVERSITY OF SZEGED
FACULTY OF ECONOMICS AND
BUSINESS ADMINISTRATION

HUNGARIAN CENTRAL
STATISTICAL OFFICE

Decomposition of the Gray-Williams "tau" in main and interaction effects by ANOVA in three-way contingency table

Luigi D'Ambra[1] - Antonello D'Ambra[2] - Pasquale Sarnacchiaro[3]

The identification of meaningful relationships between two or more categorical variables is an important, and ongoing, element to the analysis of contingency tables. It involves detecting categories that are similar and/or different to other categories. Correspondence analysis can be used to detect such relationships by providing a graphical interpretation of the association between the variables, and it is especially useful when it is known that this association is of a symmetric nature. (Greenacre 1984), (Lebart et al. 1984).

In this paper, we will explore the Gray-Williams index when used as the measure of association in non-symmetrical correspondence analysis (NSCA). It will be shown that, by concatenating a predictor variable of a three-way contingency table, the two measures are equivalent. The paper will analyse the sum of squares for nominal data partitioning the Sum of squares for main effects and the interaction in the sense of analysis of variance giving an orthogonal decomposition of Gray Williams index .

Keywords: Three-way contingency table, The Gray-Williams measure of association, Catanova, Main effects Interaction

1. Introduction

The identification of meaningful relationships between two or more categorical variables is an important, and ongoing, element to the analysis of contingency tables. It involves detecting categories that are similar and/or different to other categories. Correspondence analysis can be used to detect such relationships by providing a graphical interpretation of the association between the variables, and it is especially useful when it is known that this association is of a symmetric nature. (Greenacre 1984), (Lebart et al. 1984).

There are many real-life applications where it is not appropriate to perform classical correspondence analysis because of the obvious asymmetry of the association between the variables. In these cases non-symmetrical correspondence

[1] Luigi D'Ambra , Dept. Biological Sciences, Università di Napoli Federico II (Italy)
[2] Antonello D'Ambra, Department of Management and Accounting (Italy)
[3] Pasquale Sarnacchiaro, University of Naples Federico II (Italy)

analysis can be considered. (D'Ambra-Lauro 1989, 1992), (Gimaret et al. 1998) and (Kroonenberg–Lombardo 1999).

The key difference between the symmetrical and non-symmetrical versions of correspondence analysis rests in the measure of association used to quantify the relationship between the variables. For a two-way, or multi-way, contingency table, the Pearson chi-squared statistic is commonly used when it can be assumed that the categorical variables are symmetrically related. However, for a two-way table, it may be that one variable can be treated as a predictor variable and the second variable can be considered a response variable. For such a variable structure, the Pearson chi-squared statistic is not an appropriate measure of association. Instead one may consider the Goodman-Kruskal tau index. Where there are more than two cross-classified variables, multivariate versions of the Goodman-Kruskal tau index can be considered. These include Marcotorchino's index (Marcotorchino 1985) and Gray-Williams' indices (Gray-Williams 1975), (Anderson-Landis 1980)

In this paper, we will explore the Gray-Williams index when used as the measure of association in non-symmetrical correspondence analysis (NSCA). It will be shown that, by concatenating a predictor variable of a three-way contingency table, the two measures are equivalent. The paper will analyse the sum of squares for nominal data partitioning the Sum of squares for main effects and the interaction in the sense of analysis of variance giving an orthogonal decomposition of Gray Williams index .

This paper is divided into six further sections. In Section 2 we consider the measure of association for two asymmetric cross-classified categorical variables. In Section 3 we provide a description of NSCA where the Goodman-Kruskal tau index is used as a measure of asymmetric association. This section also offers two tools that can be used to delve deeper into the source of association using this index. One is the C-statistic based on the work of Light-Margolin (1971), and the other is confidence circles. This latter tool was discussed in some detail for symmetrical, or classical, correspondence analysis of nominal variables by Lebart et al. (1984).
The Gray-Williams measure of complete association and its link to the Goodman Kruskal tau index when concatenating a predictor variable is discussed in section 4. In section 5 we analyse the interaction between the predictor variables and we present an orthogonal decomposition of Gray-Williams "Multiple" τ in which we have the part of main effects and the part of interaction.
A case study we present in the section 6, some final consideration ended the paper.

2. Measuring Non-SymmetricAssociation

Suppose we consider the cross classification of n individuals/units according to two categorical variables, X_1 and Y, that form a two-way contingency table, N. Let X_1 be the column variable that consists of c categories, and Y be the row variable consisting of r categories. Denote the *(i, j)*th cell entry by $\mathrm{n_{ij}}$ for i = 1, 2, . . . , r and j = 1, 2, . . . , c, and the *(i, j)*th joint proportion by $p_{ij} = n_{ij} / n$ so that $\sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij} = 1$.

Define the *i*th row marginal proportion by $p_{i\bullet} = \sum_{j=1}^{c} p_{ij}$ and define the *j*th column marginal proportion by $p_{\bullet j} = \sum_{i=1}^{r} p_{ij}$.

The chi-squared statistics commonly used as a means of formally measuring the departure from independence between $\mathrm{X_1}$ and Y. By considering this statistic, it is assumed that there is a symmetric relationship between the two variables. However, there are many situations where the association between two categorical variables is not symmetric.

Suppose there exist an asymmetric association between two categorical variables such that $\mathrm{X_1}$ is treated as a predictor variable and Y is the response variable. Therefore, a more appropriate measure of their association is to adjust the chi-squared statistic and consider instead

$$n\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(p_{ij} - p_{i\bullet}p_{\bullet j}\right)^2}{p_{\bullet j}} \tag{1}$$

This measure was proposed by Goodman-Kruskal (1954) as a means of measuring the proportional reduction in error (PRE) in the prediction of the response variable given a predictor variable. Mirkin (2001,) also discussed these measures for nominal variables, as did Light-Margolin (1971) in the context of ANOVA for contingency tables. Therefore, suppose we let $\pi_{ij} = p_{ij} / p_{\bullet j} - p_{i\bullet}$ be the difference between the unconditional prediction of the *i*th row category, $\mathrm{p_{i\bullet}}$, and the conditional prediction of that, given the *j*th column category, $p_{ij} / p_{\bullet j}$. Given the *j*th column category, if it does not contribute to the predictability of the *i*th row category, then $\pi_{ij} = 0$. Formal procedures can be adopted to measure the predictability of the row response categories given the information in the column categories by considering the Goodman - Kruskal (1954) tau index

$$\tau = \sum_{i=1}^{r}\sum_{j=1}^{c} p_{\bullet j}\left(\frac{p_{ij}}{p_{\bullet j}} - p_{i\bullet}\right)^2 \Bigg/ \left(1 - \sum_{i=1}^{r} p_{i\bullet}^2\right) = \tau_{num} \Bigg/ \left(1 - \sum_{i=1}^{r} p_{i\bullet}^2\right) \tag{2}$$

Light-Margolin (1971) also considered such a measure and described it as "the proportion of total variation in the response variable which is accounted for by knowledge of the grouping [predictor] variable".

The numerator of (2), τ_{num}, can be alternatively expressed as (1) divided by the sample size n, and is bounded by the interval [0, 1]. When the distribution of each of the response (row) categories across each of the columns is identical to the overall marginal proportion, such that $p_{ij} / p_{\bullet j} = p_{i\bullet}$, there is no relative increase in predicability of the row variable and thus τ is zero. Note that zero predictability also implies no association (ie independence) between the two categorical variables. When $\tau = 1$, there is perfect predictability of the response categories (rows) given the predictor categories (columns).

3. Non-Symmetrical Correspondence Analysis: testing and confidence circles

The measure of the departure from independence of the *(i, j)*th cell of the two-way contingency table, N, when there is an asymmetric association between two categorical variables, can be quantified by the π_{ij} that is defined in Section 2. To obtain characteristics and low-dimensional summaries of the structure of this association, NSCA involves applying a singular value decomposition (SVD) to π_{ij} so that

$$\pi_{ij} = \frac{p_{ij}}{p_{\bullet j}} - p_{i\bullet} = \sum_{m=1}^{M} a_{im}\lambda_m b_{jm} \tag{3}$$

where $M = \min(r, c) - 1$ and λ_m is the *m*th singular value of π_{ij} for $m = 1, \ldots, M$. The quantities a_{im} and b_{jm} are, respectively, the elements of the singular vectors $\mathbf{a}_{\text{m}}$ and $\mathbf{b}_{\text{m}}$ associated with the *i*th row and *j*th column categories and have the property

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

By considering the decomposition (3), the numerator of the Goodman-Kruskal tau index can be decomposed so that

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

When performing NSCA (Beh-D'Ambra 2010), we can graphically depict the association between the row and column categories by plotting along the *m*th dimension of the non-symmetrical correspondence plot the row and column profile coordinates

$$f_{im} = a_{im}\lambda_m \qquad \text{and} \qquad g_{jm} = b_{jm}\lambda_m .$$

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

$$\tau_{num} = \sum_{i=1}^{r}\sum_{m=1}^{M} f_{im}^2 = \sum_{j=1}^{c}\sum_{m=1}^{M} p_{\bullet j}\, g_{jm}^2 .$$

Therefore, points lying at a distance from the origin of the plot indicate that these categories contribute more to τ_{num} than those points that lie near the origin. Also, if column points lie close to the origin, these categories do not contribute to the predictability of the response variable. If predictor (row) points lie close to the origin, these categories are not affected by any variation in the predictor variable.

The Goodman-Kruskal tau index is a good measure for determining the predictability of the rows given the columns. However, as Agresti (1990) indicated, a low value of τ does not mean that there is a "low" association between the two variables. While τ is an appropriate measure of the predictability, the statistic cannot, in its current form, be used to formally test for association. Instead such tests are carried out using the C-statistic of Light-Margolin (1971)

$$C = (n-1)(r-1)\tau = (n-1)(r-1)\sum_{m=1}^{M} \lambda_m^2 \Bigg/ \left(1 - \sum_{i=1}^{r} p_{i\bullet}^2\right) \tag{4}$$

Under the hypothesis of zero predictability ($H_0 : \Pi_{ij} = 0$), Light - Margolin (1971) showed that the C-statistic is asymptotically chi-squared distributed with $(r-1)(c-1)$ degrees of freedom. These authors introduced this statistic when deriving an analysis of variance procedure for contingency tables, commonly referred to as CATANOVA (Categorical Analysis of Variance).

When the variables of a two-way contingency table are considered to be symmetrically related, as is in the case for classical correspondence analysis, Lebart et al. (1984) presented the idea of confidence circles to identify those categories that contribute to the hypothesis of independence and those that do not. These circles are similar to the regions that Mardia et al. (1982, p. 346) derived for canonical analysis. Ringrose (1992, 1996) also explored the use of these types of circles for correspondence analysis, although a bootstrap procedure was employed for their construction. When categorical variables are ordinal in nature showed that the radii of these circles are identical to those of Lebart et al. (1984). However the confidence circles derived for use in symmetrical correspondence analysis are not applicable for NSCA. Here we present the radii lengths of confidence circles for NSCA.

Suppose that a two-way contingency table consists of row and column variables asymmetrically structured in the manner described in Section 2. The C–statistic of (4) can be expressed in terms of the predictor (row) coordinates such that

$$C = (n-1)(r-1)\sum_{j=1}^{c}\sum_{m=1}^{M} p_{\bullet j} g_{jm}^2 \bigg/ \left(1 - \sum_{i=1}^{r} p_{i\bullet}^2\right) \sim \chi^2 .$$

For the *j*th column (predictor) coordinate,

$$(n-1)(r-1)\sum_{m=1}^{M} p_{\bullet j} g_{jm}^2 \bigg/ \left(1 - \sum_{i=1}^{r} p_{i\bullet}^2\right) \sim \chi_{(r-1)}^2 .$$

Therefore

$$\sum_{m=1}^{M} p_{\bullet j} g_{jm}^2 \sim \chi_{(r-1)}^2 \left(1 - \sum_{i=1}^{r} p_{i\bullet}^2\right) \bigg/ ((n-1)(r-1)) .$$

Since, for higher dimensions, the coordinates will be close to zero (as the singular values associated with these dimensions are generally relatively close to zero), the relationship between the *j*th column coordinates for the first two dimensions of a two-dimensional non-symmetrical correspondence plot is

$$p_{\bullet j} g_{j1}^2 + p_{\bullet j} g_{j2}^2 = \chi^2_{(2)} \left(1 - \sum_{i=1}^{r} p_{i\bullet}^2\right) \Big/ (n-1)(r-1).$$

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

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

Therefore, the 95% confidence circle for the *j*th column coordinate in the two-dimensional non-symmetrical correspondence plot has a radius of length

$$o_j = \sqrt{5.99\left(1 - \sum_{i=1}^{r} p_{i\bullet}^2\right) \Big/ \left(p_{\bullet j}(n-1)(r-1)\right)} . \tag{5}$$

Note that (5) depends on the *j*th marginal proportion classified into that category. Thus, if there is a very small number of classifications made in the *j*th predictor category, its radius length associated with this category will be relatively large. Similarly, for a relatively large classification, the radius length will be relatively small. Since we are interested in the predictability of the row categories given the column categories, confidence circles will only be constructed for the predictor variable.

Careful attention must be given to the interpretation of these regions. They do not suggest that a point has any significant link with an axis, since the axes have no direct interpretation (other than to graphically depict the proportion of the association between the variables it reflects). Overlapping regions may provide some indication as to the level of association between intra-variable categories but they do not provide formal evidence that such an association exists, although employing the *uncertainty circles* of Gabriel (1995) can provide such insight. The real strength of the confidence circles described here lies in their ability to reflect the significance of a particular predictor category in accounting for the level of predictability on a response variable. If the origin is enclosed within the confidence circle of a predictor category, then that category does not contribute to the predictability of the response variable. Similarly, if the origin falls outside of a confidence circle, then that particular predictor category does contribute to the predictability of the response variable. Such conclusions can be made keeping in mind the level of significance used to construct these circular regions.

4. Multiple NSCA – The Gray-Williams index

Suppose we consider the cross classification of n individuals/units according to three categorical variables X_1, X_2 and Y, that form a three-way contingency table, N. Let X_1 be the second (column) variable that consists of c categories, X_2 be the third (tube) variable that consists of t categories, and Y be the first (row) variable consisting of r categories. The terminology "tube" is used to be consistent with much of the discussion that has been made on multiple categorical data analysis; for example, Kroonenberg (1989) uses the expression. One may also consider X_2 to be a stratifying variable. The resulting contingency table is therefore of size $r \times c \times t$. Here we consider the relationship between the three variables to be asymmetric, in that Y is the response variable and depends on the two predictor variables X_1 and X_2.

To measure the asymmetric association of the three variables, one may consider multivariate extensions of the Goodman-Kruskal tau index.(Anderson-Landis 1980) Two examples include Marcotorchino's index and Lombardo's index. Another measure, with which we will concern ourselves here, is the Gray-Williams index (Gray-Williams 1975).

Let $\pi_{ijk} = p_{ijk} / p_{\bullet jk} - p_{i\bullet\bullet}$, for i = 1, 2, . . . r, j = 1, 2, . . . c and k = 1, 2, . . . , t, be the difference between the unconditional marginal proportion of the *i*th response category, $p_{i\bullet\bullet}$, and the (conditional) prediction of the *i*th response given the joint proportion of the two predictor variables, $p_{ijk} / p_{\bullet jk}$. Gray-Williams (1975) proposed an extension of the Goodman-Kruskal tau index for three categorical variables where the proportional reduction in error for the prediction of the response (row) variable can be measured by considering

$$\tau_{GW} = \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{t} p_{\bullet jk}\left(\frac{p_{ijk}}{p_{\bullet jk}} - p_{i\bullet\bullet}\right)^2 \Bigg/ \left(1 - \sum_{i=1}^{r} p_{i\bullet\bullet}^2\right) = \tau_{GWnum} \Bigg/ \left(1 - \sum_{i=1}^{r} p_{i\bullet\bullet}^2\right) \quad (6)$$

Just as was done for the Goodman-Kruskal tau index of (2), the numerator of the Gray-Williams index, τ_{GWnum}, will be the focus of our discussion here, since the denominator is independent of the any of the joint cell proportions of the table N.

To determine the structure of the dependence between three categorical variables (one criterion variable and two predictor variables), one may consider a three-way extension of the SVD of (3):

$$\pi_{ijk} = \sum_{m=1}^{M} a_{im} \lambda_m b_{jm} c_{km}$$

where

$$\sum_{i=1}^{r} a_{im} a_{im'} = \begin{cases} 1, & m = m' \\ 0, & m \neq m' \end{cases}, \qquad \sum_{j=1}^{c} p_{\bullet j \bullet} b_{jm} b_{jm'} = \begin{cases} 1, & m = m' \\ 0, & m \neq m' \end{cases}$$

and

$$\sum_{k=1}^{t} p_{\bullet\bullet k} c_{km} c_{km'} = \begin{cases} 1, & m = m' \\ 0, & m \neq m' \end{cases}.$$

This approach is analogous to the PARAFAC/CANDECOMP models independently considered by Harshman (1970) and Carroll-Chang (1970) (Faber et al 2003). An alternative approach was considered by Lombardo-Carlier-D'Ambra (1996). For their approach, π_{ijk} (constructed to reflect the variation in predictability as measured by the Marcotorchino index) is decomposed using the Tucker3 decomposition (Tucker 1966).

Another method of decomposition that can be considered is

$$\pi_{ijk} = \frac{p_{ijk}}{p_{\bullet jk}} - p_{i\bullet\bullet} = \sum_{m=1}^{M} a_{im} \lambda_m b_{jkm} \tag{7}$$

where $M = \min(r, c + t) - 1$ and a_{im} and b_{jkm} are subject to the constraints

$$\sum_{i=1}^{r} a_{im} a_{im'} = \begin{cases} 1, & m = m' \\ 0, & m \neq m' \end{cases} \qquad \text{and} \qquad \sum_{j=1}^{c} \sum_{k=1}^{t} p_{\bullet jk} b_{jkm} b_{jkm'} = \begin{cases} 1, & m = m' \\ 0, & m \neq m' \end{cases}$$

respectively. This approach is called Multiple Non Symmetrical Correspondence Analysis (MNSCA).

The generalised singular values, λ_m, are again arranged in descending order such that $1 > \lambda_1 > \lambda_2 > \cdots > \lambda_M > 0$. The value a_{im} is an element of the singular vector $\mathbf{a}_m$ and is associated with the *i*th row response category. Similarly the value b_{jkm} is an element of the joint singular vector $\mathbf{b}_m$ of length *ct* and is associated with the joint association between the two predictor variables. The calculation of these quantities can be easily performed, not through any modification of the SVD procedure of (3), but by simply concatenating one of the predictor variables to form a two-way table.

To demonstrate this point, suppose we transform the $r \times c \times t$ contingency table N in such a way that the tube predictor variable is concatenated so that N is of size $\mathrm{r} \times \mathrm{ct}$.

For the *k*th *(k = 1, 2, . . . , t)* $r \times c$ submatrix, the Goodman-Kruskal tau numerator, $\tau_{num|k}$, is

$$\tau_{num|k} = \sum_{i=1}^{r}\sum_{j=1}^{c} p_{\bullet jk}\left(\frac{p_{ijk}}{p_{\bullet jk}} - p_{i\bullet\bullet}\right)^2 .$$

Aggregating each of these t measures of asymmetry yields

$$\begin{aligned}\sum_{k=1}^{K}\tau_{num|k} &= \sum_{i=1}^{r}\sum_{j=1}^{c} p_{\bullet j1}\left(\frac{p_{ij1}}{p_{\bullet j1}} - p_{i\bullet\bullet}\right)^2 + \cdots + \sum_{i=1}^{r}\sum_{j=1}^{c} p_{\bullet jt}\left(\frac{p_{ijt}}{p_{\bullet jt}} - p_{i\bullet\bullet}\right)^2 \\ &= \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{t} p_{\bullet jk}\left(\frac{p_{ijk}}{p_{\bullet jk}} - p_{i\bullet\bullet}\right)^2\end{aligned} \tag{8}$$

Which is the numerator of the Gray-Williams index, τ_{GWnum}, defined by (6).

For a concatenated three-way contingency table with response (row) marginal proportions $\{\mathrm{p}_{1\bullet}, \ldots, \mathrm{p}_{\mathrm{r}\bullet}\}$ and predictor (column) marginal proportions $\{\mathrm{p}_{\bullet 11}, \mathrm{p}_{\bullet 21}, \ldots, \mathrm{p}_{\bullet \mathrm{ct}}\}$, equation (8) is equivalent to the Goodman-Kruskal tau index. This is apparent since

$$\begin{aligned}\tau_{num} &= \sum_{i=1}^{r}\left[p_{\bullet 11}\left(\frac{p_{i11}}{p_{\bullet 11}} - p_{i\bullet\bullet}\right)^2 + p_{\bullet 21}\left(\frac{p_{i21}}{p_{\bullet 21}} - p_{i\bullet\bullet}\right)^2 + \cdots + p_{\bullet ct}\left(\frac{p_{ict}}{p_{\bullet ct}} - p_{i\bullet\bullet}\right)^2\right] \\ &= \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{t} p_{\bullet jk}\left(\frac{p_{ijk}}{p_{\bullet jk}} - p_{i\bullet\bullet}\right)^2 \\ &= \tau_{GWnum} .\end{aligned}$$

Therefore, the Gray-Williams index for the $\mathrm{r} \times \mathrm{c} \times \mathrm{t}$ contingency table, N, is equivalent to the Goodman-Kruskal tau index when concatenating a predictor variable. When performing a NSCA for a three-way contingency table, the influence of the predictor variables X_1 and X_2 on the response variable Y may therefore be made by considering the SVD of the concatenated data. By considering the concatenated contingency table, NSCA can be applied to obtain profile coordinates

$$f_{im}^{*} = a_{im}^{*}\lambda_{m}^{*} \qquad \text{and} \qquad g_{(jk)m}^{*} = b_{(jk)m}^{*}\lambda_{m}^{*} \tag{9}$$

Here, a_{im}^{*} is the *i*th element of the *m*th singular vector associated with the rows of the concatenated table. Similarly $b_{(jk)m}^{*}$ is the *(j, k)*th element of the *m*th singular vector associated with the columns of the concatenated table, and λ_{m}^{*} is the *m*th singular value.
By considering (9) for the concatenated NSCA, the numerator of the Gray-Williams index may be expressed as the weighted sum of squares of these coordinates so that

$$\tau_{GWnum} = \sum_{i=1}^{r}\sum_{m=1}^{M}\left(f_{im}^{*}\right)^{2} = \sum_{j=1}^{c}\sum_{k=1}^{t}\sum_{m=1}^{M} p_{\bullet jk}\left(g_{(jk)m}^{*}\right)^{2} = \sum_{m=1}^{M}\left(\lambda_{m}^{*}\right)^{2}.$$

For the application of confidence circles we consider Gray-Williams τ and the C-statistic of Anderson –Landis

$$C = (n-1)(r-1)\tau = (n-1)(r-1)\sum_{m=1}^{M}\lambda_{m}^{2} \Big/ \left(1 - \sum_{i=1}^{r} p_{i\bullet}^{2}\right).$$

Under the hypothesis of zero predictability ($H_0 : \Pi_{ijk} = 0$), Anderson Landis (1980) showed that the C-statistic is asymptotically chi-squared distributed with $(r-1)(ct-1)$ degrees of freedom. Therefore the radius can be computed as shown in formula (5).

5. Analysis of interaction term

Interaction effects represent the combined effects of predictor variables on the response variable. When interaction effects are present, the impact of one predictor variable depends on the level of the other predictor, in other words it means that interpretation of the main effects is incomplete or misleading.

In case of no interaction effect, a difference in level between the two lines would indicate a main effect of predictor variable.

Many texts stipulate that you should interpret the interaction first. If the interaction is not significant, you can then examine the main effects without needing to qualify the main effects because of the interaction. If the interaction is significant, you cannot examine the main effects because the main effects do not tell the complete story. It seems that it makes more sense to tell the simple story first and then the more complex story. In the two-way case, we prefer to examine each of the main effects first and then the interaction.

Regarding MNSCA, in order to consider the different effects of the predictor variables (main and interaction effects) on response variable, our approach starts from the exact reconstruction formula of the contingency table using eigen values and coordinates, particularly

$$p_{ijk} = p_{.jk}\left[p_{i.} + \sum_{m=1}^{M}\left(\frac{1}{\sqrt{\lambda_m}}\right)a_{im}g_{jkm}\right] \tag{10}$$

The coordinates g_{jkm} (for $j = 1, 2, \ldots c$ and $k = 1, 2, \ldots, t$) include the main effects (j,k) and the interaction (jXk), the a_{im} (for $i = 1, 2, \ldots r$) are the row coordinates (response variable) and λ_m is the eigenvalue with $m = \min[(r-1);(ct-1)]$ equal to the rank of matrix.

We replace in formula (10) the coordinates g_{jkm} ($m=1....M$) with the functions h_{jm} and w_{km} obtained by two way analysis of variance without interaction.

$$\hat{p}_{ijk} = p_{.jk}\left[p_{i.} + \sum_{m=1}^{M}\left(\frac{1}{\sqrt{\lambda_m}}\right)f_{im}(h_{jm} + w_{km})\right] \tag{11}$$

This new matrix $\hat{P}$ represents the dependence between categories of rows and columns after the elimination of the interaction effect. Performing a MNSCA on $\hat{P}$ we improve the interpretation of the main effect, to be more precise we represent only the effect of the prediction variables on the response variable purified to the interaction between predictors.

The choice of this functions yields the following orthogonal decomposition:

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{t}\left[\left(\frac{P_{ijk}}{P_{\bullet jk}}-P_{i\bullet\bullet}\right)^{2}p_{.jk}\right]=\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{t}\left[\left(\frac{\hat{p}_{ij,k}}{P_{\bullet jk}}-P_{i\bullet\bullet}\right)^{2}p_{.jk}\right]+$$

$$+\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{t}\left[\left(\frac{p_{ijk}}{p_{\bullet jk}}-\frac{\hat{p}_{ij,k}}{p_{\bullet jk}}\right)^{2}p_{\bullet jk}\right]$$

The ***Gray Williams "Multiple" numerator*** = ***Main effects*** + ***Interaction***

It easy to verify that the matrix $\hat{P}$, with general terms $\hat{p}_{ij,k}$, and P have the same column and row marginals. Moreover considering the matrix $\overline{P}=P-\hat{P}$ and performing a MNSCA, we compute the analysis of the interaction between predictors without the main effects.

It is possible to show that if in our approach we use one-way analysis of variance instead of two way analysis of variance without interaction, we get the solution proposed by Takane-Jung (2009) based on linear constraints on the predictor categories (Takane-Shibayama 1991). In this last case the Gray-Williams multiple τ is decomposed in two components: the formes gives the Goodman-Kruskal numerator and the other gives partial τ_{num} Gray-Williams.

6. Case study

In this section, we present a detailed application of the proposed method. The case study pertains to the analysis of a 5 x 6 x 4 contingency table obtained cross-classifying subjects by mathematical score at University, teaching method used and final grade at school (independence variables). The data collected are placed in a bivariate table (table 1).

Table 1. Cross classification of students in term of mathematical score at University (criterion variable), teaching method used and final grade at school (predictor variables)

	Method	A	A	A	A	A	A
	School score	School very low score (---)	School low score (--)	School middle low score (-)	School middle high score (+)	School high score (++)	School very high score (+++)
Mathematic score very low	(--)	12	10	9	8	6	5
Mathematic score low	(-)	7	18	8	3	6	4
Mathematic score middle	(±)	1	8	10	5	1	1
Mathematic score high	(+)	2	8	6	12	16	5
Mathematic score very high	(++)	5	2	9	6	12	25
	Method	B	B	B	B	B	B
	School score	School very low score (---)	School low score (--)	School middle low score (-)	School middle high score (+)	School high score (++)	School very high score (+++)
Mathematic score very low	(--)	12	10	5	5	4	4
Mathematic score low	(-)	5	7	7	4	2	2
Mathematic score middle	(±)	1	6	6	4	3	2
Mathematic score high	(+)	5	4	3	17	11	3
Mathematic score very high	(++)	3	3	2	5	9	28

	Method	C	C	C	C	C	C
	School score	School very low score (---)	School low score (--)	School middle low score (-)	School middle high score (+)	School high score (++)	School very high score (+++)
Mathematic score very low	(--)	2	3	4	1	3	2
Mathematic score low	(-)	4	5	8	2	4	5
Mathematic score middle	(±)	5	6	9	6	7	8
Mathematic score high	(+)	5	8	13	13	22	23
Mathematic score very high	(++)	20	22	30	35	40	45
	Method	D	D	D	D	D	D
	School score	School very low score (---)	School low score (--)	School middle low score (-)	School middle high score (+)	School high score (++)	School very high score (+++)
Mathematic score very low	(--)	1	1	3	2	1	1
Mathematic score low	(-)	2	2	5	4	2	2
Mathematic score middle	(±)	2	2	7	7	8	9
Mathematic score high	(+)	5	6	16	12	18	21
Mathematic score very high	(++)	12	20	24	28	38	44

Source: own creation

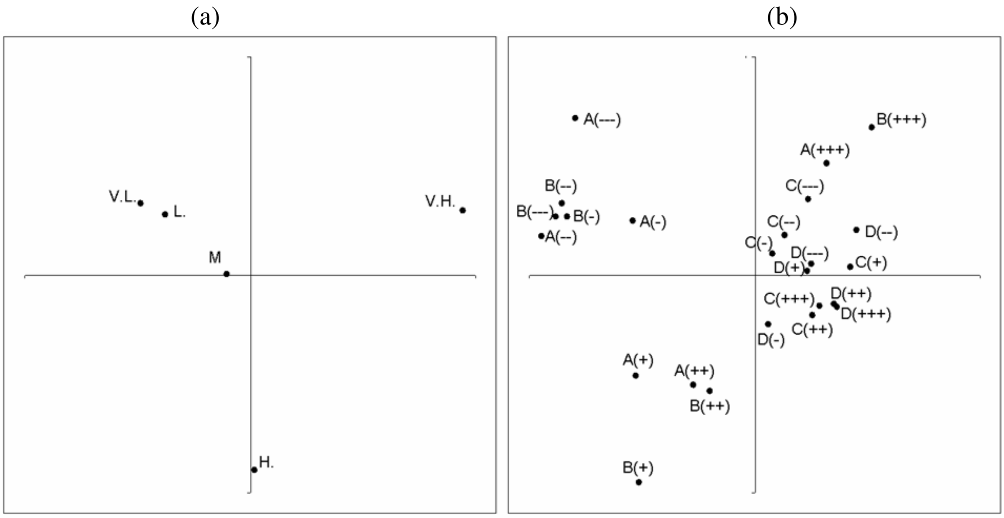
The first variable is a response or criterion variable and it has five classes of ordered categories: 18-20 (VL), 21-23 (L), 24-26 (M), 27-29 (H), 30 (VH). The second and third variables are predictor variables. The variable Final grade at Italian school are grouped in six categories: 60-64 (---), 65-69 (--), 70-79 (-), 80-89 (+), 90-94 (++), 95-100 (+++). Teaching methods has four categories: Technological tools - Projector/video/slide (A), Problem solving Brainstorming (B), Direct Teaching (C) and Lecture (D). The symbol in parentheses are the label in graphic representations.

In order to analyze the statistical dependence of mathematical score at University from teaching method used and final grade at school we perform a

MNSCA. We represent the two dimensional configuration as the best solution in Figure1, particularly in the left side we plot the criterion categories, in the right side we project the modalities of the predictor variables.

In order to selection the dimension we use permutation test. It works in the following way: first, on compute the singular value (SV) from the original data set. Then, the columns of predictor variables are randomly permuted, and SV's are computed from the permuted data set. The largest SV from the permuted data set is compared with that from the original data set. To test the statistical significance of the SV from the original data set, we repeat the same procedure K time (with K very great) and count how many times the former is larger than the latter. If this count is smaller than Kα (where α is the prescribed significance level), the largest SV being tested is significantly different from 0. Each subsequent SV can be tested in the same way after eliminating the effect of the preceding SV's. In our case we found two axis significative.

Figure 1. Classical MNSCA (a) row coordinates (b) column coordinates (total inertia explained 91,51%)



Source: own creation

The predictive power of a particular predictor category on a particular criterion category can be evaluated by the magnitude of the inner product between the two vectors representing the two categories. For example overlapping the two plots we can remark that A (+++) and B(+++) are closest to VH (the highest mathematical score). This means that the students having the highest final grade at school and used as teaching method A and B have achieved highest mathematical score.

In table 2, for each category has been computed the radius of confidence circle and the distance from the origin of the axes. The decision rule is: if the radius is greater than the distance then the category is significant

Table 2. Radius of the Confidence Circle and Distance from the origin of the axes (* Category statistical significant at the 5% level).

	Distance from origin	**Radius**
A(---)	0,453	0,200*
A(--)	0,476	0,153*
A(-)	0,279	0,160*
A(+)	0,299	0,178*
A(++)	0,204	0,162*
A(+++)	0,223	0,164*
B(---)	0,447	0,204*
B(--)	0,438	0,190*
B(-)	0,424	0,217*
B(+)	0,384	0,176*
B(++)	0,189	0,193
B(+++)	0,330	0,166*
C(---)	0,158	0,173
C(--)	0,087	0,157
C(-)	0,049	0,130
C(+)	0,214	0,138*
C(++)	0,140	0,119*
C(+++)	0,151	0,114*
D(---)	0,128	0,221
D(--)	0,235	0,187*
D(-)	0,074	0,140
D(+)	0,118	0,143
D(++)	0,182	0,127*
D(+++)	0,188	0,118*

Source: own creation

The classical MNSCA is based on the decomposition of Gray Williams "Multiple" τ including together main effects and interaction term. In order to know the statistical significance of the single main effect and of the interaction, we can use the factorial representation analysis of variance of nominal data (Onukogu 1984). The results are summarized in table 2.

Table 3. CATANOVA table

Source	SS	C-statistic	dof	p-value
Final grade at school (1)	0,021	22,75	5	0,0002
Teaching methods (2)	0,029	31,26	3	0,0000
Interaction (3) Final grade at school *Teaching methods	0,018	19,98	15	0,0447
Between (1) + (2) + (3)	0,068	73,99	23	0,0000
Within	0,652		1053	
Total	0,720		1076	

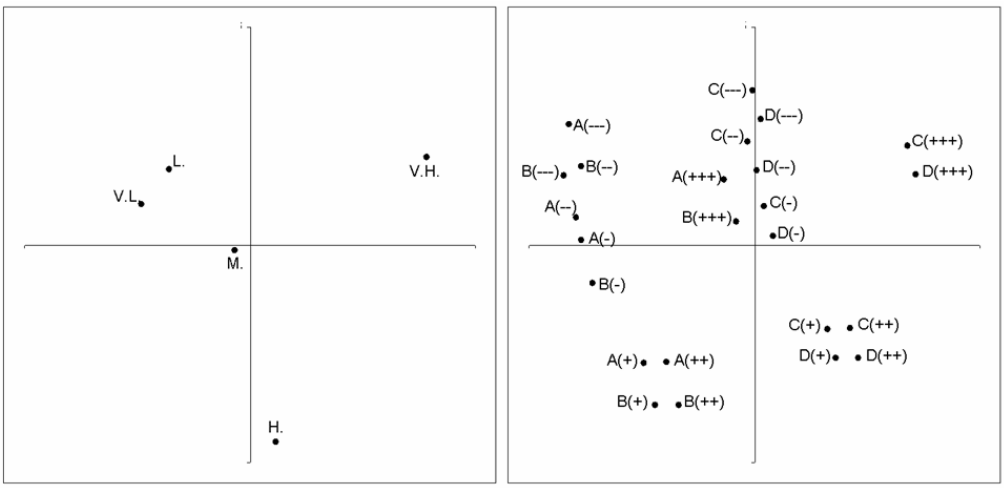
Source: own creation

All the sources of variation are statistically significant, therefore the levels of Matematical Score depends on the final grade at school, on teaching methods and on their interaction. In classical MNSCA the effect of interaction could make unclear the interpretation of the axis.

Following the approach proposed in section 5, we can separate the effect of the main sources of variation from the interaction. Particularly in figure 2 and 3 we represent the main effects and the interaction term respectively. Following the procedure presented previously, in both cases the dimension selected is composed by two axis significative.

Figure 2. MNSCA only main effects (a) row coordinates (b) column coordinates (total inertia explained 94,83%)

(a) (b)



Source: own creation

In figure 2, we note that overlapping the two plots the predictor categories closest to VH (the highest mathematical score) are C (+++) and D (+++), moreover it seems that the teaching methods more effective are C and D because all categories of Final grade at school are closest to the criterion categories VH and H.

Considering only the main effects, in table 4 for each category has been computed the radius of confidence circle and the distance from the origin of the axes. The decision rule is: if the radius is greater than the distance then the category is significant

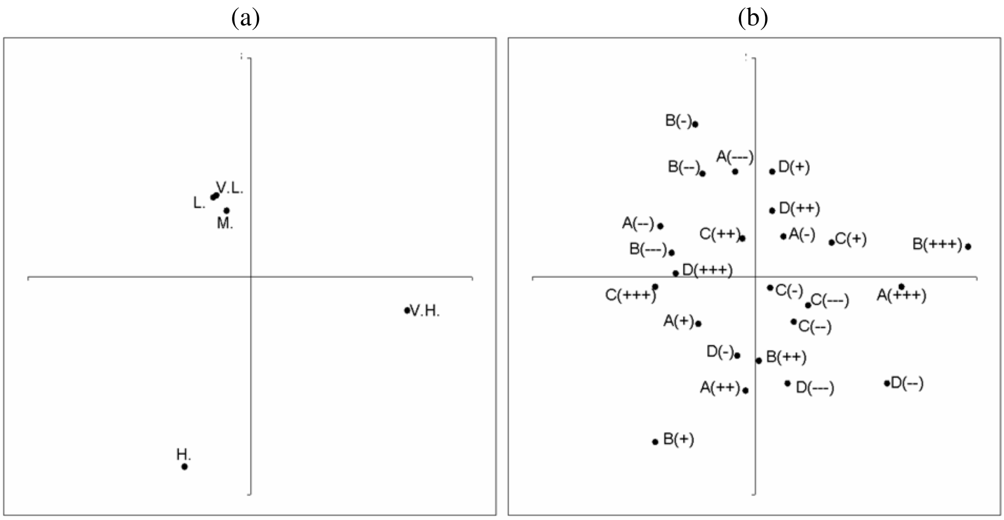
Table 4. Radius of the Confidence Circle and Distance from the origin of the axes (* Category statistical significant at the 5% level) for main effects

	Distance from origin	**Radius**
A(---)	0,339	0,200*
A(--)	0,341	0,153*
A(-)	0,308	0,160*
A(+)	0,213	0,178*
A(++)	0,176	0,162*
A(+++)	0,070	0,164
B(---)	0,312	0,204*
B(--)	0,317	0,190*
B(-)	0,288	0,217*
B(+)	0,207	0,176*
B(++)	0,174	0,193
B(+++)	0,036	0,166
C(---)	0,106	0,173
C(--)	0,072	0,157
C(-)	0,032	0,130
C(+)	0,142	0,138*
C(++)	0,180	0,119*
C(+++)	0,281	0,114*
D(---)	0,087	0,221
D(--)	0,051	0,187
D(-)	0,034	0,140
D(+)	0,164	0,143*
D(++)	0,201	0,127*
D(+++)	0,292	0,118*

Source: own creation

These results are supported by the plot of interaction in which we can observe how the students with the highest score in mathematic are those that had the best Final grade at school and used A and B teaching methods.

Figure 3. MNSCA only interaction effect (a) row coordinates (b) column coordinates (total inertia explained 86,34%)



Source: own creation

In this paper we have presented a method of analysing two complementary parts of the predictive relationships between the columns and rows of a three way contingency table, one part can be explained by main effect of predictive categories on criterion variable and the other represents the effect of interaction between the two predictive variables on the criterion variable. The usefulness of the method is shown by a study regarding the statistical dependence of mathematical score at University from teaching method used and final grade at school.

References

Agresti, A. 1990: *Analysis of Categorical Data*. Wiley, New York.

Anderson, R. J. - Landis, J. R. 1980: CATANOVA for multidimensional contingency tables: Nominal-scale response. *Comm. Statist. Theory Methods,* 9, 1191-1206. p.

Beh, E. J. - D'Ambra, L. 2009: Non symmetrical correspondence analysis with concatenation and linear constraints. *Australian and New Zealand Journal of statistics.* (to appear 2010)

Caroll, J. D. - Chang, J.-J. 1970: Analysis of individual differences in multi-dimensional scaling via an N-way generalisation of "Eckart-Young" decomposition. *Psychometrika* 35, 283-319. p.

D'Ambra, L. – Lauro, N. 1989: Non symmetrical analysis of three way contingency tables. In Coppi, R. - Bolasco, S. (eds): *Multiway Data Analysis*. North Holland, 301–315. p.

D'Ambra, L. - Lauro, N. 1992: Exploratory non symmetric data analysis. *Italian J. Appl. Stats.* 4, 511-529. p.

Faber, N. M. - Bro, R. - Hopke, P. K. 2003: Recent developments in CANDECOMP/PARAFAC algorithms: a critical review. *Chemo. Intell. Lab. Syst.,* 65, 119-137. p.

Gabriel, K. R. 1995: MANOVA biplots for two-way contingency tables. In Krzanowski, W. J. (eds): *Recent Advances in Descriptive Multivariate Analysis*. Oxford University Press, 227 – 268. p.

Gimaret-Carpentier, C. - Chessel, D. - Pascal, J.-P. (1998): Non-symmetric correspondence analysis: an alternative for species occurrences data. *Plant Ecology* 138, 97-112. p.

Goodman, L. A. - Kruskal, W. H. 1954: Measures of association for cross-classifications. *J. Amer. Statist. Assoc.,* 49, 732-764. p.

Gray, L. N. - Williams, J. S. 1975: Goodman and Kruskals tau b: Multiple and partial analogs. *American Statistical Association (Proceedings of the Social Statistics Section)*, 444-448. p.

Greenacre, M. J. 1984: *Theory and Applications of Correspondence Analysis*. Academic Press, London.

Harshman, R. A. 1970: Foundations of the PARAFAC procedure: models and conditions for an "explanatory" multi-modal factor analysis. *UCLA Working Paper in Phonetics,* 16, 1-84. p.

Kroonenberg, P. 1989: Singular value decompositions of interactions in three-way contingency tables. In Coppi, R. - Bolasco, S. (eds): *Multiway Data Analysis*. Amsterdam, North-Holland, 169 – 184. p.

Kroonenberg, P. - Lombardo, R. 1999: Nonsymmetric correspondence analysis: A tool for analysing contingency table with a dependence structure. *Multiv. Beh. Res.,* 34, 367–397. p.

Lebart, L. - Morineau, A. - Warwick, K. M. 1984: *Multivariate Descriptive Statistical Analysis*. Wiley, New York .

Light, R. J. - Margolin, B. H. 1971: An analysis of variance for categorical data. *J. Amer. Statist. Assoc.,* 66, 534-544. p.

Lombardo, R. - Carlier, A. - D'Ambra, L. 1996 : Nonsymmetric correspondence analysis for three-way contingency tables. *Methodologica,* 4, 59-80. p.

Marcotorchino, F. 1985: *Utilisation des Comparaisons par Paires en Statistique de Contingences: Parte III*. Report # F 071, Etude du Centre Scientifique, IBM, France.

Mardia, K. V. - Kent, J. T. - Bibby, J. M. 1982: *Multivariate Analysis*. Academic Press, London.