Journal de la Société Française de Statistique
Vol. 160 No. 2 (2019)

An extension of MLDA to Three-way Contingency Tables

Titre: Une extension de l'analyse discriminante multiblocs aux tableaux de contingence ternaires

Philippe Casin[1]

Abstract: The aim of this paper is to propose an extension of Multiblock Linear Discriminant Analysis (MLDA) for analyzing a set of contingency tables which have been observed in different occasions and have the same number of rows and the same number of columns. This extension, Multiblock Linear Discriminant Analysis of Three-way Contingency Tables (MLDA-TCT, is midway between correspondence analysis and linear discriminant analysis; MLDA-TCT computes one or several variables for each data table, such that these variables take into account relationships between rows and columns of the contingency tables in one hand, and in the other hand, take into account relationships between contingency tables.

Résumé : L'objet de cet article est de proposer une extension de l'analyse factorielle discriminante de tableaux multiples à la description d'un ensemble de tableaux de contingence qui ont été observés à différentes occasions et qui ont le même nombre de lignes et le même nombre de colonnes. Cette méthode, MLDA-TCT, est un compromis entre l'analyse factorielle des correspondances et l'analyse discriminante linéaire. MLDA-TCT détermine une ou plusieurs variables auxiliaires pour chaque tableau de données, de telle manière que ces variables prennent en compte à la fois les relations entre les lignes et les colonnes des tableaux de contingence et les relations entre les tableaux de contingence.

Keywords: linear discriminant analysis, canonical analysis, correspondence analysis, bi-partitioned data table, Three-way contingency table
Mots-clés : analyse discriminante, analyse canonique, analyse des correspondances, tableau bi-partitionné, tableau de contingence ternaire
AMS 2000 subject classifications: 62H30, 62h25, 62-07

1. Introduction

Canonical Correlation Analysis (CCA; Hotelling (1936)) describes the relationship between two sets of variables observed on the same individuals. When each of these two sets of variables is constituted by the indicators of a categorical variable, CCA is the Correspondence Analysis (CA,Benzécri (1980); Black et al. (1998)) of the contingency table which displays the frequency distribution of the two categorical variables.

These two categorical variables may be observed at different times, for instance imports and exports tables calculated for different countries or different years, or at different occasions, for instance statistics by age, by level of education and by gender; a set of these contingency tables which have all the same number of columns and the same number of rows is named three-way contingency table.

[1] Université de Lorraine.
E-mail: philippe.casin@univ-lorraine.fr

 ISSN: 2102-6238

Methods for analyzing a set of contingency data tables have first been developed by French statisticians (Leclerc (1975); Saporta (1976); Cazes (1981)) and consist of correspondence analysis of the juxtaposition or of the sum of contingency data tables. Alternative factorial methods are MFACT method (Escofier and Pagès (1994); Bécue-Bertaut and Pages (2004); Kostov et al. (2013)), STATIS method (Lavit (1988); Vallejo-Arnadela et al. (2007)) or Simultaneous Analysis (Zarraga and Goitisolo (2002, 2003, 2009)).

Three-way contingency table is a particular case of multiple contingency table; specific methods have been developed to analyze three-way contingency tables (Kroonenberg and Lombardo (1999); Lombardo (2011); Kateri and Petros Dellaportas (2012); Beh and Lombardo (2014); Kang et al. (2015); Aktas (2016); Beh et al. (2018); Taneichi and Toyama (2019)).

Recently, methods for analyzing data which are structured both in blocks of variables and in groups of individuals (for an overview of these methods, see Tenenhaus and Tenenhaus (2014)) have been applied to sets of contingency tables ; on one hand, generalizations of STATIS (Vallejo-Arnadela et al. (2007); Sabatier et al. (2013)), have been developped in order to take into account of both structure of each data table and evolutions in the different time points or occasions. In the other hand, among the other methods of multivariate analysis of multiblock and multigroup data, see (Eslami et al. (2014); Bougeard et al. (2017); Kang et al. (2015); Bougeard et al. (2011, 2018))

Multiblock Linear Discriminant Analysis (MLDA, Casin (2015, 2017)), a method for analyzing bi-partitioned data tables, computes one or several new variables for each data table, such that these new variables take into account both relationships between sets of variables and canonical correlation and relationships between each block of variables and the partition of individuals in groups.

The aim of this paper is to extend principles of MLDA to three-way contingency table ; this new method, named MLDA-TCT, points out the differences between several contingency tables.

The organization of the paper is as follows: Section 2 introduces the problem and Section 3 defines the notation. In section 4, MLDA-TCT is introduced; MLDA-TCT is compared with others methods in Section 5. Section 6 is concened by plots and interpretation of results. A concrete application of CMLDA-TCT is given in Section 7. Section 8 concludes.

2. The problem

Let us consider a set of K contingency tables C_k, for $k = 1, \ldots, K$. These tables have the rows and columns in common, but their row margins and column margins are different.

The problem is to highlight the main differences between K contingency tables C_k, $k = 1, \ldots, K$, where C_k is a $r \times m$ matrix and classifies n_k individuals with respect to two categorical variables X_k and Y_k.

Let G be the categorical variable (with n rows and K columns) which describes the partition of the $n = \sum_{k=1}^{K} n_k$ individuals into K situations.

Generally (Cazes (1981); Bécue-Bertaut and Pages (2004); Vallejo-Arnadela et al. (2007); Zarraga and Goitisolo (2003)) categories of the variable G are not explicitly considered as active variables and G is only used to reinforce the interpretation of the results of methods of analyzing

 ISSN: 2102-6238

sets of contingency tables: categories of G are represented on the axes, but are excluded from the construction of these axes.

On the opposite, here, in order to point out differences between contingency data tables, G is an active variable: G is the logical response variable to the other variables which are the rows and columns of contingency tables C_k. Consequently, the discriminant power of G is the first criterion of determination of axes. The second criterion of determination of axes is the quality of the description of contingency tables: axes must provide the best possible description of the relationship between rows and columns of the average contingency table $\frac{1}{K}\sum_{k=1}^{K} C_k$.

To carry on this work, it is first necessarily to introduce the following proper representation of the data.

3. The data and their representation

3.1. Notation

As mentionned $C_k = X_k^{'} Y_k$ is a $r \times m$ matrix and classifies n_k individuals, with respect to two categorical variables X_k and Y_k. X_k (resp. Y_k) is a $n_k \times r$ (resp. $n_k \times m$) matrix and describes the partition of n_k individuals into r (resp. m) groups ; its r (resp. m) columns (called indicators) are dummy variables: a value of 1 indicates that the individual belongs to the group, a value of 0 that it does not. n denotes the total sum of individuals: $n = \sum_{k=1}^{K} n_k$.

Let us consider the following data tables X and Y with n rows and respectively r and m columns. The columns of these two matrices are dummy variables. X and Y are defined as follows:

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_k \\ \vdots \\ X_K \end{bmatrix} \text{ and } Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_k \\ \vdots \\ Y_K \end{bmatrix}$$

Z is the matrix $Z = \Big[X, Y\Big]$.

G_k is the matrix with n rows and K columns whose all values equal 0 except those of the $k-th$ column which are all equal to 1. Then, the categorical variable G (with n rows and K columns) describes the partition of the n individuals into K situations:

$$G = \begin{bmatrix} G_1 \\ \vdots \\ G_k \\ \vdots \\ G_K \end{bmatrix}$$

 ISSN: 2102-6238

3.2. *Representation of the data*

Data may be represented as shown below:

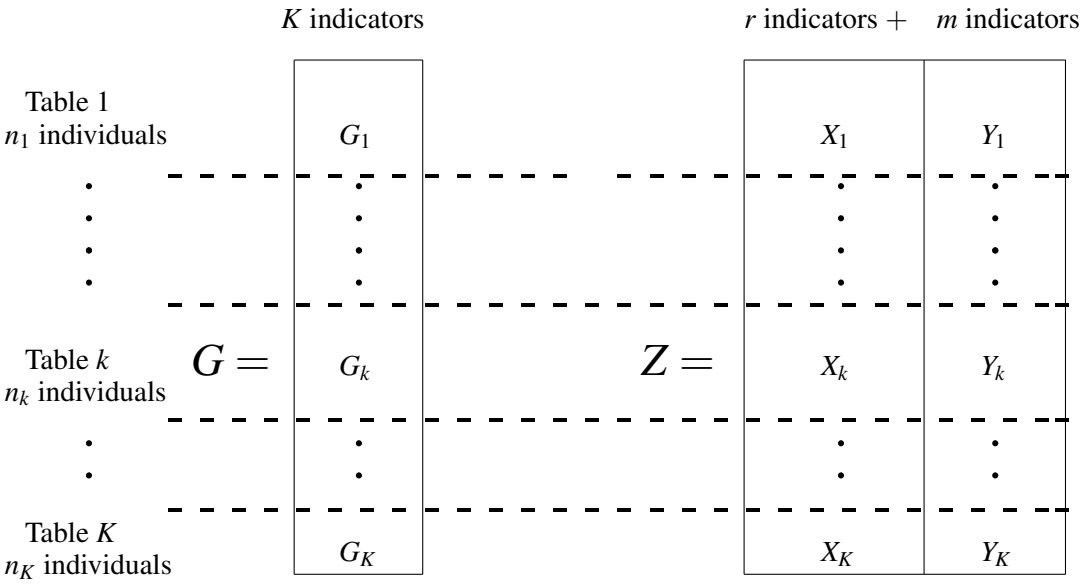

FIGURE 1. *Representation of the data*

Then $V_{GX} = G'X$ (resp. $V_{GY} = G'Y$) is the $K \times m$ (resp. $K \times r$) contingency table which classifies the n individuals with respect to the K situations and the r (resp. m) columns of the indicators of the variable X (resp. Y).

V_{GX} (resp. V_{GY}) is the $K \times r$ data table of the row (resp. column) margins of contingency tables C_k. $V_{XY} = X'Y$ is the sum of the matrices C_k: $V_{XY} = \sum_{k=1}^{K} C_k$. $V_{GG} = G'G$ (resp. $V_{XX} = X'X$, $V_{YY} = Y'Y$) is the $K \times K$ (resp.$r \times r$, $m \times m$) diagonal matrix of marginal sums of the K data tables G_k(resp X, Y).

V_{XG} (resp. V_{YG}) is the $r \times K$ (resp.$m \times K$) matrix whose k-th column is the absolute frequency of categories of X (resp. Y).

Let W_X (resp. W_Y, W_{X_j}, W_{Y_j}) be the space spanned by columns of X (resp. Y, G, X_j, Y_j) and P_X (resp. P_Y, P_G, P_{X_j}, P_{Y_j}) be the orthogonal projector onto W_X(resp. W_Y, W_G, W_{X_j}, W_{Y_j}).

Matrices are represented in upper-case letters and vectors in lower-case letters.

 ISSN: 2102-6238

4. Multiblock Linear Discriminant Analysis of Three-way Contingency Tables

4.1. Criteria used by MLDA-TCT

The principle of MLDA (Casin (2015, 2017)) is describing both the relationship between the blocks of variables (here, X and Y), using Carroll's Canonical Generalized Analysis criterion (Carroll (1968)), and the relationship between these blocks and the categorical variable (here, G) which describes the partition in groups of individuals, using Fisher's discriminant analysis criterion(Fisher (1936)).

MLDA is not directly applicable to a set of contingency tables. That is the reason why a proper representation of the data has been introduced in the previous section. In this way, MLDA can be extented and used to simultaneously analyze contingency tables C_k, $k = 1,\ldots,K$. This new method is named Multiblock Linear Discriminant Analysis of Three-way Contingency Tables (MLDA-TCT).

4.1.1. The first step

At the first stage, MLDA-TCT computes an auxiliary variable z^1, $z^1 \in R^n$, such that:

- $R^2(z^1, z_X^1)$ and $R^2(z^1, z_Y^1)$ which are correlation coefficients between z^1 and its projections onto the spaces W_X and W_Y are as high as possible (it is Carroll's criterion)

- μ_X^1 and μ_Y^1, which are correlation ratios between these variables z_X^1 and z_Y^1 and the categorical variable G, are as high as possible (it is Fisher's criterion).

More specifically, MLDA optimizes a compromise between these two criteria and computes z^1 such that $R^2(z^1, z_X^1)\mu_X^1 + R^2(z^1, z_Y^1)\mu_Y^1$ has a maximum value, subject to $\|z^j\|^2 = 1$.

4.1.2. Constraints of orthogonalization

Here, the aim is to describe X and Y and consequently the contingency table $X'Y = V_{XY} = \sum_{k=1}^{K} C_k$, or, in other words, the aim is to describe the average contingency data table $\frac{1}{K}\sum_{k=1}^{K} C_k$. Then, constraints are defined in order to compute an orthogonal basis of spaces W_X and W_Y and consequently, at the $j-th$ step, the auxiliary variables z_X^j (resp. z_Y^j) must be orthogonal to the previous auxiliary variables z_X^s (resp. z_Y^s) for $s = 1,\ldots,j-1$.

Consequently X^j (resp. Y^j) is the data table corresponding to the subspace of X (resp. Y) orthogonal to z_X^s (resp. z_Y^s) for $s = 1,\ldots,j-1$.

Z^j is the super matrix $Z^j = \left[X^j, Y^j\right]$, and $z^j = Z^j a^j$, where

$$a^j = \begin{bmatrix} a_X^j \\ a_Y^j \end{bmatrix}$$

Then, X^j (resp. Y^j) is the matrix of residuals of regression of X^{j-1} (resp. Y^{j-1}) on $z_X^j = X^{j-1}a_X^j$ (resp. $z_Y^j = Y^{j-1}a_Y^j$).

 ISSN: 2102-6238

4.1.3. The other steps

The search continues beyond the first step. At the $j-th$ step, z_X^j (resp.z_Y^j) is the projection of z^j onto W_X (resp. W_Y); z^j is the solution of:

$$\begin{cases} \text{Max } R^2(z^j, z_X^j)\mu_X^j + R^2(z^j, z_Y^j)\mu_Y^j \\ \text{subject to } \|z^j\|^2 = 1 \\ \text{subject to for } s = 1, \ldots, j-1 \; R(z_X^j, z_X^s) = 0 \\ \text{subject to for } s = 1, \ldots, j-1 \; R(z_Y^j, z_Y^s) = 0 \end{cases}$$

4.2. The solution

4.2.1. The first step

At the first step (Casin (2015, 2017)), z^1 is the first unit variance eigenvector of $P_X P_G P_X + P_Y P_G P_Y$. Let us denote $M_X = (X'X)^{-1}X'G(G'G)^{-1}G'X(X'X)^{-1} = V_{XX}^{-1}V_{XG}V_{GG}^{-1}V_{GX}V_{XX}^{-1}$, $M_Y = V_{YY}^{-1}V_{YG}V_{GG}^{-1}V_{GY}V_{YY}^{-1}$ and let M be the following matrix

$$M = \begin{bmatrix} M_X & 0 \\ 0 & M_Y \end{bmatrix}$$

then z^1 is the first unit variance eigenvector of ZMZ'. In other words, z^1 is the first principal component issued of a PCA with respect to the metric defined by the matrix M.

In practice, n, the total number of individuals of the K contingency tables, often takes large values and diagonalization of a large dimension square matrix is not easy. Let us consider A^1,the following $(m+r) \times 2K$ matrix:

$$A^1 = \begin{bmatrix} A_X^1 & 0 \\ 0 & A_Y^1 \end{bmatrix}$$

where $A_X^1 = V_{XX}^{-1}V_{XG}V_{GG}^{-0.5}$ and $A_Y^1 = V_{YY}^{-1}V_{YG}V_{GG}^{-0.5}$; $V_{GG}^{-0.5}$ is the diagonal matrix of squared root of marginal frequencies of G. Consequently, $A_X^1 A_X^{1'} = M_X$, $A_Y^1 A_Y^{1'} = M_Y$ and $A^1 A^{1'}$=M.

Let v^1 be the first eigenvector of $A^{1'}Z'ZA^1$ then $A^{1'}Z'ZA^1 v^1 = \lambda^1 v^1$, $ZA^1A^{1'}Z'ZA^1v^1 = \lambda^1 ZA^1 v^1$ and finally $ZMZ'z^1 = \lambda^1 z^1$.

Consequenty v^1 is the first eigenvector of a square matrix of order $2K$, which is much smaller than n, and $z^1 = ZA^1v^1$.

Let $a^1 = A^1 v^1$ be the column vector with $m+r$ rows and a_X^1 (resp. a_Y^1) the element of this vector, corresponding to the data table X (resp. Y):

$$a^1 = \begin{bmatrix} a_X^1 \\ a_Y^1 \end{bmatrix}$$

z_X^1 is the orthogonal projection of z^1 and equals:

$z_X^1 = P_X z^1 = X(X'X)^{-1}X'Za^1 = X(X'X)^{-1}X'(Xa_X^1 + Ya_Y^1)$ and consequently $z_X^1 = X(a_X^1 + V_{XX}^{-1}V_{XY}a_Y^1) = Xb_X^1$. For the same reasons, $z_Y^1 = Y(a_Y^1 + V_{YY}^{-1}V_{YX}a_X^1) = Yb_Y^1$.

The first eigenvalue equals 2 corresponding to the trivial solution.

 ISSN: 2102-6238

4.3. The other steps

The problem is the same at the $j-th$ stage as at the first stage, except that the auxiliary variables z_X^j (resp. z_Y^j) must be orthogonal to the previous auxiliary variables z_X^s (resp. z_Y^s) for $s = 1, \ldots, j-1$.

Consequently z^j is the first unit variance eigenvector of $P_{X^j} P_G P_{X^j} + P_{Y^j} P_G P_{Y^j}$.

Let A^j be the following matrix

$$A^j = \begin{bmatrix} A_X^j & 0 \\ 0 & A_Y^j \end{bmatrix}$$

where $A_X^j = V_{X^j X^j}^{-1} V_{X^j G} V_{GG}^{-0.5}$ and $A_Y^j = V_{Y^j Y^j}^{-1} V_{Y^j G} V_{GG}^{-0.5}$ and let $Z^j = \left[X^j, Y^j\right]$.

Let v^j be the $j-th$ eigenvector of $A^{j'} Z^{j'} Z^j A^j$, and let $a^j = A^j v_j^1$ be the column vector with $m+r$ rows and a_X^j (resp. a_Y^j) the element of this vector, corresponding to the data table X^j (resp. Y^j):

$$a^j = \begin{bmatrix} a_X^j \\ a_Y^j \end{bmatrix}$$

then $z_X^j = X^j b_X^j$ where $b_X^j = a_X^j + V_{X_j X_j}^{-1} V_{X_j Y_j} a_Y^j$

and $z_Y^j = Y^j (a_Y^j + V_{Y_j Y_j}^{-1} V_{Y_j X_j} a_X^j) = Y^j j b_Y^j$ where $b_Y^j = a_Y^j + V_{Y_j Y_j}^{-1} V_{Y_j X_j} a_X^j$

4.4. The trivial solution

Let us consider the vector u whose all values equal 1. Because $P_G u = u$, $P_X u = u$ and $P_Y u = u$: $\|w_X^1\|^2 + \|w_Y^1\|^2 = 2u$, that means that the first eigenvalue equals 2, and consequently, z^j, z_X^j and z_Y^j, which are orthogonal to u are centered.

4.5. The maximum possible number of steps

The problem is to compute eigenvectors of $A'Z'ZA$. Since the rank of X (resp.Y) equals m (resp. r) and since there is a trivial solution, the vector u, whose all values equal 1, then the maximum possible number of steps of LDA equals $(\inf(m,r) - 1)$.

4.6. Weighting of data tables

An alternative approach consists of ponderating the data tables. Weights can be different from one table to an other. In particular, in order to accord equal importance to data tables, data weights can be equal to $1/n_k$, the inverse of the number of individuals. $1/n_k C_k$ is then a table of probability (the sum of all its elements equals 1) ; V_{GG} is the identity matrix, and the element at the intersection of the j-th row and the k column of V_{XG} is the marginal frequency of the j-category for the k-th data table.

 ISSN: 2102-6238

4.7. The algorithm

Computation of results is based on matrices of small dimensions, B, $X'G$, $Y'G$ and $G'G$ and does not need computation of projectors onto the spaces W_X and W_Y.

step 1: compute V_{XX}, V_{YY}, V_{XG}, V_{XY}, V_{YG}, V_{GG} and

$$A^1 = \begin{bmatrix} A_X^1 & 0 \\ 0 & A_Y^1 \end{bmatrix}$$

step 2: compute v^1, the first eigenvector of $A^{1'}V_{ZZ}A^1$ where ;

$$V_{ZZ} = Z'Z = \begin{bmatrix} V_{XX} & V_{XY} \\ V_{YX} & V_{YY} \end{bmatrix}$$

step 3: compute $b_X^1 = a_X^1 + V_{XX}^{-1}V_{XY}a_Y^1$ and $b_Y^1 = a_Y^1 + V_{YY}^{-1}V_{YX}a_X^1$ where

$$a^1 = A^1v^1 = \begin{bmatrix} a_X^1 \\ a_Y^1 \end{bmatrix}$$

step 4: compute $C_X^1 = Id - b_X^1(b_X^{1'}V_{XX}b_X^1)^{-1}b_X^{1'}V_{XX}$ and $C_{Y_k}^1 = Id - b_Y^1(b_Y^{1'}V_{YY}b_Y^1)^{-1}b_Y^{1'}V_{YY}$ and consider D_X^1 (resp. D_Y^1) the matrix whose columns are all columns of C_X^1 (resp. C_Y^1), except the last.

step 5: replace X in calculations by $X^1 = X_kD_X^1$ and Y by $Y^1 = Y_kD_Y^1$ and consequently replace V_{XX}, V_{XY} and V_{YY} respectively by $V_{X^1X^1} = D_X^{1'}V_{XX}D_X^1$, $V_{X^1Y^1} = D_X^{1'}V_{XY}D_Y^1$ and $V_{Y^1Y^1} = Y_kV_{YY}D_Y^1$, etc. Go to step 1. Etc...

step 6: computations stop after $(\inf(m,r)-1)$ steps.

5. Plots and interpretation of results

5.1. Orthogonality between the components

Let us consider two components z^r and z^s, and suppose, for convenience, that $r > s$.

$$z^s = \begin{bmatrix} z_X^s \\ z_Y^s \end{bmatrix}$$

Because $r > s$, z_X^s is a linear combination of columns of X^r orthogonal to $X^rb_X^r$, and then $z_X^s = X^rd_X^s$ where d_X^s is a column vector with m rows. Consequently $d_X^{s'}X^{r'}X^rb_Y^r = d_X^{s'}X^{r'}X^r(a_X^r + (X^{r'}X^r)^{-1}X^{r'}Y^ra_Y^r) = 0$, and then $d_X^{s'}X^{r'}X^ra_X^r + d_X^{s'}X^{r'}Y^ra_Y^r = 0$.

For the same reasons, $z_Y^s = Y^rd_Y^s$ and $d_Y^{s'}Y^{r'}Y^ra_Y^r + d_Y^{s'}Y^{r'}X^ra_X^r = 0$ where d_Y^s is a colmumn vector with r rows.

It follows that $d_X^{s'}X^{r'}X^ra_X^r + d_X^{s'}X^{r'}Y^ra_Y^r + d_Y^{s'}Y^{r'}Y^ra_Y^r + d_Y^{s'}Y^{r'}X^ra_X^r = 0$ and finally $(X^rd_X^s + Y^rd_Y^s)'(X^ra_X^r + Y^ra_X^r) = 0$. Consequently z^r and z^s are orthogonal.

5.2. Plots based on auxiliary variables

MLDA-JCT provides two different types of graphical representations:

 ISSN: 2102-6238

5.2.1. *Plots based on the discriminant analysis*

Like DA does, plots are based on variables z^j and both represent categories of variables X and variables Y. Moreover, these plots superimpose a representation of the K data tables whose coordinates are the mean of categories for the corresponding data table.

5.2.2. *Plots of the rows and columns*

MLDA-TCT provides an orthogonal basis of spaces spanned by data table X (resp. Y_k) and consequently a graphical representation of categories of rows (resp. columns) based on variables z_X^j (resp. z_Y^Y).

$z_X^j = Xd_X^j$ (resp. $z_Y^j = Y_k d_Y^j$), where d_X^j (resp. d_Y^j) is the $j-th$ vector of scores for the rows (resp. columns) of this average contingency table. The coordinate of a data table on an axis is the mean of coordinates of individuals of these data table: let $f_{h,s}^X$ (resp. $f_{h,t}^Y$) be the relative frequency of the $s-th$ (resp. $t-th$) category of the rows (resp. columns) for the $k-th$ data table X_k (resp. Y_k) equals $\sum_{s=1}^{m} f_{h,s}^X d_{X,s}^j$ (resp. $\sum_{t=1}^{r} f_{h,t}^Y d_{Y,t}^j$ where $d_{X,s}^j$ (resp. $d_{Y,t}^j$) is the score of the $s-th$ (resp. $t-th$) category of X (resp. Y) at the $j-th$ step.

5.3. Computing values of the criteria

Formulas for computing these values are as follows:

1. Discrimination using Fisher's criterion

   $\mu_X^1 = \frac{Var(w_X^1)}{Var(z_1^j)} = \frac{z_X^{1'} G V_{GG}^{-1} G' z_X^1}{z_X^{1'} z_X^1} = \frac{b_X^{1'} V_{XG} V_{GG}^{-1} V_{GX} b_X^1}{b_X^{1'} V_{XX} b_X^1}$

   Similarly,

   $\mu_Y^1 = \frac{b_Y^{1'} V_{YG} V_{GG}^{-1} V_{GY}^j b_Y^j}{b_Y^{j'} V_{YY} b_Y^j}$

   These two indicators are synthesized by the discrimination power of the variable z^j which equals:

   $\mu_z^1 = \frac{a^{j'} V_{ZG} V_{GG}^{-1} V_{GZ} a^j}{a^{j'} V_{ZZ} a^j}$

2. Correspondence Analysis using Carroll's criterion

   $R^2(z^1, z_X^1) = \frac{Var(z_X^1)}{Var(z^1)} = \frac{Var(X^1 b_X^1)}{Var(z^1)} = \frac{b_X^{1'} V_{XX} b_X^1}{a^{1'} V_{ZZ} a^1}$.

   Similar formulas are obtained for the data table Y:

   $R^2(z^1, z_Y^1) = \frac{b_Y^{1'} V_{YY} b_Y^1}{a^{1'} V_{ZZ} a^1}$.

   and these two indicators can be synthesized by the following indicator, which is the criterion to be maximized by CA:

   $R^2(z_X^1, z_Y^1) = \frac{(b_X^{1'} V_{XY} b_Y^1)^2}{(b_X^{1'} V_{XX} b_X^1)(b_Y^{1'} V_{YY} b_Y^1)}$

 ISSN: 2102-6238

5.4. Alternative constraints of orthogonalization

The auxiliary variables z_X^j (resp. z_Y^j) are orthogonal to each other and consequently MLDA-TCT provides an orthogonal basis of the space W_X (resp. W_Y); the following alternative constraints of orthogonalization can be considered: for $j \neq j'$, z^j is orthogonal to $(z^j)'$. Then z^j is the $j-th$ eigenvector of ZMZ' and is easily computed. But, this approach does not provide an orthogonal basis of spaces W_X and W_Y.

6. Comparison with other methods

6.1. Comparison with correspondence analysis and related methods

Both MLDA-TCT and CA provide an orthogonal basis of W_X and an orthogonal basis of W_Y and then describe relation between X and Y ; moreover, both MLDA-TCT and CA are Principal Component Analysis of the data table Z, but they are not associated with the same metric (Jollife (2002)). The metric associated with MLDA-TCT is, at the first step:

$$A^1 = \begin{bmatrix} A_X^1 & 0 \\ 0 & A_Y^1 \end{bmatrix}$$

where $A_X^1 = V_{XX}^{-1} V_{XG} V_{GG}^{-0.5}$ and $A_Y^1 = V_{YY}^{-1} V_{YG} V_{GG}^{-0.5}$; $V_{GG}^{-0.5}$ is the diagonal matrix of squared root of marginal frequencies of G.

Since the matrix associated with CA is:

$$B = \begin{bmatrix} B_X & 0 \\ 0 & B_Y \end{bmatrix}$$

where $B_X = V_{XX}^{-1}$ and $B_Y = V_{YY}^{-1}$.

Consequently, CA does not take into account the relationship between X and G on one hand, and the relationship between Y and G on the other hand.

For same reasons, methods based on a CA of the sum of the contingency tables or on a CA of the data table which juxtaposes contingency tables (Cazes (1981)) do not take into account relationship between G and the independent variables X and Y.

6.2. Discriminant Analysis

In contrast to LDA, MLDA-TCT determines an independent variable structured in two blocks and takes into account the correlation between z_X^j and z_Y^j which characterised these two blocks ; MlDA-TCT also provides an orthogonal basis of spaces described by columns of X and Y. The maximal number of steps equals $g-1$ for DA and $(\inf(m,r)-1)$ for MLDA-TCT.

6.3. Discriminant Correspondence Analysis

Discriminant Correspondence Analysis (DCA)(Leclerc (1975)) is the correspondence analysis of the data table $\left[G'X,\ldots,G'Y\right]$. This data table juxtaposes the contingency tables of margins of data tables X and Y for the K occasions and does not take into account interactions among the explanatory variables X and Y.

 ISSN: 2102-6238

6.4. Simultaneous Analysis, Multiple Factor analysis and STATIS

Simultaneous Analysis (Zarraga and Goitisolo (2002, 2003, 2009)) Multiple Factor analysis (Escofier and Pagès (1994); Bécue-Bertaut and Pages (2004)) and STATIS (Lavit (1988); Vallejo-Arnadela et al. (2007)) consist on a weighted principal component analysis of the data tables, the weights being first eigenvalues of CA of ad hoc contingency tables, and computed in a previous step. The common feature of all these methods is that they search for structures common to all contingency tables on study, whereas the main objective of MLDA-TCT is to determine main differences between these contingency tables.

6.5. Methods of analysis of multiblock and multigroup data

Tenenhaus and Tenenhaus (Tenenhaus and Tenenhaus (2014)) present an overview of techniques for analyzing data tables structured in blocks of variables or in groups of individuals ; these techniques are based on generalizations of canonical analysis or on principal components analysis ; Eslami and al. (Eslami et al. (2014)) analyze data which are structured both in groups and in blocks, but with multigroup Principal Components Analysis whereas Sabatier and al. (Sabatier et al. (2013)) uses LDA criterion but associated with STATIS approach.

Bougeard and al.(Bougeard et al. (2017)) consider a set of predictor variables organized into blocks, and a set of dependant variables using multiblock PLS and multiblock redundancy analysis (Eslami et al. (2014); Beh and Lombardo (2014)). Kang and all. (Kang et al. (2015)) use the Discriminative Least Squares Regression (DLSR, a method proposed by Xiang et al (Xiang et al. (2012)) and solutions are given by an iterative algorithm.

The main difference between all these methods and MLDA-TCT is that MLDA-TCT applies classical criterion of correspondance analysis of individuals for blocks and Fischer's classical criterion of discrimination for groups to the proper representation of data exposed in Section 3.

7. An application

7.1. The data

The population under study consists of 20 819 men and 12 282 women. For each gender, a data table relative to shoplifting among 350 Dutch stores and big textile shops (Israels (1987); Zarraga and Goitisolo (2002)) describes relationships between 9 classe groups ($0-11$, $12-14$, $15-17$, $18-20$, $21-29$, $30-39$, $40-49$, $50-64$, and $65+$) and 13 kinds of stolen objects (CLOThes, CLothing ACcessories, TOBAcco, WRITing accesories, BOOKs, RECOrds, HOUSehold accessories, SWEEts, TOYS, JEWEllery, PERFums, HOBBies and OTHErs). The items described by the columns of X are class groups since the stolen objects are described by columns of Y.

7.2. MLDA-TCT results

Because of the important difference in size between the two genders, each of the two data tables is weighted by the inverse of the total number of its individuals.

 ISSN: 2102-6238

MLDA-TCT computes inf(13,9) = 9 eigenvalues: the first eigenvalue equals 2.000 and corresponds to the trivial eigenvalue, the second eigenvalue equals 0.183, the third one and the fourth one respectively equal 0.012, 5.0010^{-04}, and the following eigenvalues are smaller than 10^{-13}.

j	λ^j	$R^2(z_X^j, z_Y^j)$	μ_Z^j	$R^2(z^j, z_X^j)$	μ_X^j	$R^2(z^j, z_Y^j)$	μ_Y^j
$j=1$	0.183	0.017	0.190	0.987	0.176	0.017	0.051
$j=2$	0.012	0.003	0.012	0.034	0.001	0.003	0.012

TABLE 1. *Values of criteria*

The discriminant power of higher-order components is close to 0. Comparatively, first eigenvalues, which can be compared to $R^2(z_X^j, z_Y^j)$, equal 0.321 , 0.046 , 0.004 , 0.002.

Discriminant analyzis results are represented on the first figure, where the first axis is the discrimant variable z^1 and the second axis the discriminant variable z^2:

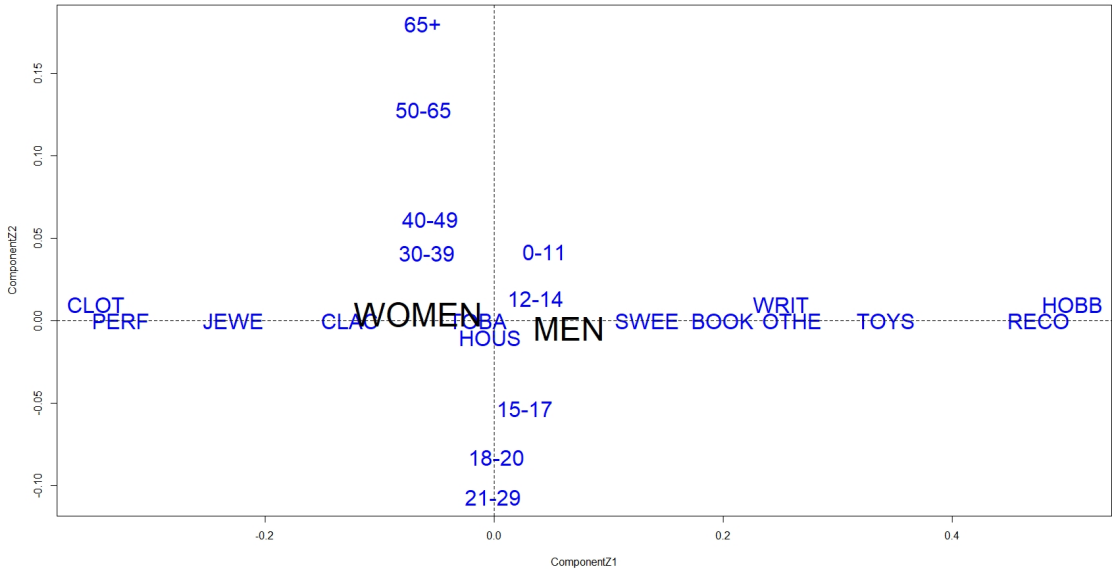

FIGURE 2. *Plot of components z^1 and z^2*

The second graphic provides a representation of variables z_X^1 and z_X^2:

and the representation of variables z_Y^1 and z_Y^2 is given by the following figure:

The discriminant power of the first axis equals 0.128 and the correlation between the two variables z_X^1 and z_Y^1 equals 0.187. CLOC, CLAC, 30-39, 40-49, 50-59, 65+ are on the left-hand side of the first axis, and correspond to an over-representation of this kind of stolen objects for women, and especially for oldest women ; on the opposite, RECO, WRIT, SWEE, BOOK, OTHE and HOBB, 0-11, 12-14, 15-17 are on the right-hand side and correspond to an over-representation of this kind of stolen objects for Men and especially youngest men.

The discriminant power of the second axis is much less important than those of the first axis: this second axis points out higher proportions of old shoplifters for Women, independently of kind object stolen, and higher proportions of young shoplifters for Men.

 ISSN: 2102-6238

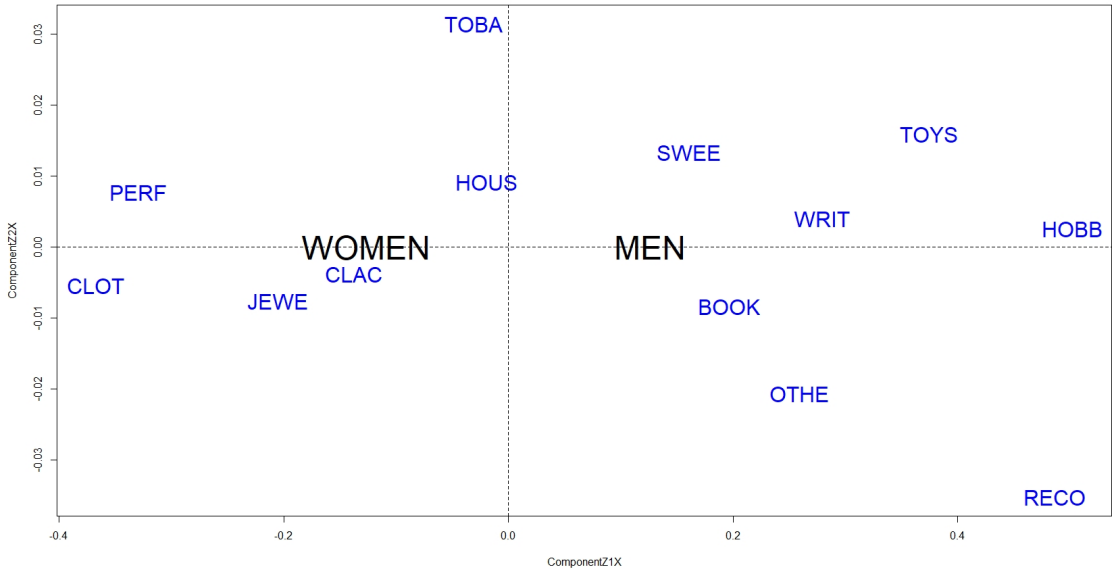

FIGURE 3. *Plot of components* z_X^1 *and* z_X^2

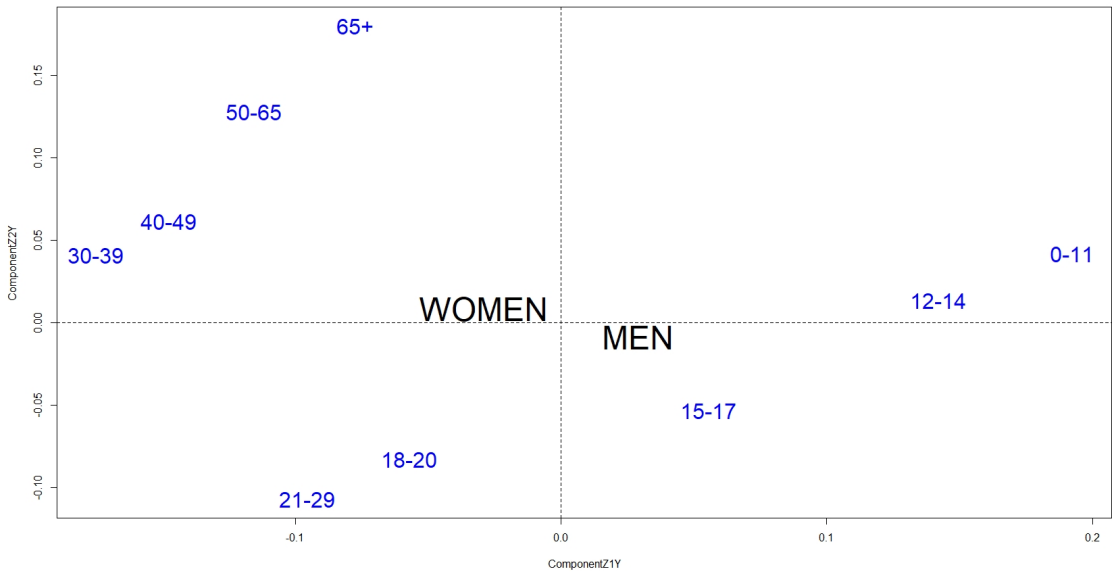

FIGURE 4. *Plot of components* z_Y^1 *and* z_Y^2

 ISSN: 2102-6238

7.3. Comparison with results from other methods

Results of separate CA of the two data tables show that first axis of CA of male data table and first axis of CA of female data table are close to each other: they both oppose 0-12 and 12-14, TOYS, SWEET, WRITE, in one side to middle age people and CLOTH in the other side. For men, the second axis highlights a low attraction for JEWE and RECO, for 0-12 and 65+, and a strong attraction for 15-17 whereas women of 15-17 have a strong attraction for CLOT and JEWE and women of 65+ a strong attraction for TOBA.

As mentioned by (Zarraga and Goitisolo (2002); Israels (1987)) considers CA of the data table which juxtaposes the two contingency tables: the first axis opposes 0-14 to middle age classes, and the second Women-JEWE and Women-CLOT to Men-HOBB. And the two first axes of CA of the sum of the two contingency tables provide results close to CA of male data table (there are twice as many men as women.)

Bécue-Bertaut and Pages (2004) introduce MFACT, an extension of CA which takes into account the particularities of each contingency table, ie differences between their margin and consequently computes weights of columns in such a way that the influence of each of these columns is comparable in a global analysis. Simultaneous analysis ((Zarraga and Goitisolo (2002, 2003, 2009)) is close to MFACT, but the allocation of weights attributed to each table is different. Zarraga and Goitisolo (2002) compare results of these two methods for the joint study of contingency tables of shoplifters. For the first axis, global representations are close to superposition of plots given by separate CA of the two data tables; the second axis of MFACT highlights the attraction for women from 21 to 39 to CLOT, and the second axis of simultaneous analysis the attraction for women to JEWE and RECO and the attraction for women over to 65 to TOBA.

First axes of all these methods point common elements between contingency tables whereas MLDA-TCT (which is partially based on a criterion of discrimination) highlights differences between these contingency tables (see previous section). Moreover, MLDA-TCT provides two supplementary graphics, and then a much accurate description of highlighted relationships between items of each data sets.

8. Concluding remarks

MLDA-TCT takes into account both relationships between the independent variables (as CA does) and specificities of each data table (as DA does). MLDA-TCT's rules for results interpretation are close to those of DA and to those of CA, and MLDA-TCT quantifies the importance of each independent variable in relation with the classification of individuals and the strength of relationships between these variables. It is worth noting that this new technique can easily be generalized to multi-way contingency data table, by adding other independent variables.

References

Aktas, S. (2016). Subsymmetry and asymmetry models for multiway square contingency tables with ordered categories. *Open mathematics research article*, 14:195–204.

Bécue-Bertaut, M. and Pages, J. (2004). A principal axes method for comparing multiple contingency tables: Mfact. *Computational Statistics and Data Analysis*, 45:481–503.

 ISSN: 2102-6238

Beh, J. and Lombardo, R. (2014). Symetrical and non-symetrical three-way correspondance analysis. *Correspondance Analysis: Theory, Pratice and New Strategies*, pages 481–503.

Beh, J., Lombardo, R., and Gianmarco, A. (2018). Correspondence analysis and the freeman-tukey statistic: A study of archaeological data. *Computational Statistics & Data Analysis*, 128:73 – 86.

Benzécri, J.-P. (1980). *L'analyse des données*. Dunod.

Black, W. C., Babin, B., Anderson, R. E., and Tatham, R. (1998). Multivariate data analysis. 5:207–219.

Bougeard, S., Abdi, H., Saporta, G., and Niang, N. (2017). Clusterwise analysis for multiblock component methods. *Advances in Data Analysis and Classificationses Classif*, pages 1–29.

Bougeard, S., Niang, S., Verron, T., and Bry, X. (2018). Current multiblock methods: Competition or complementarity? a comparative study in a unified framework. *Chemometrics and Intelligent Laboratory Systems*, 182:131 – 148.

Bougeard, S., Qannari, E., and Rose, N. (2011). Multiblock redudancy analysis: interpretation tools and application in epidemiology. *Journal of chemiometrics*, 25:467–475.

Carroll, J.-D. (1968). Generalization of canonical correlation analysis to three or more sets of variables. *Proceedings of the 76th annual convention of the Americain Psychological Association*, 3:227–228.

Casin, P. (2015). L'analyse factorielle discriminante de tableaux multiples. *Journal de la Société Française de Statistique*, 156:1–20.

Casin, P. (2017). Categorical multiblock linear discriminant analysis. *Journal of Applied Statistics*, pages 1–14.

Cazes, P. (1981). L'analyse de certains tableaux rectangulaires décomposés en blocs : généralisation des propriétés rencontrées dans l'analyse des correspondances multiples.iv.cas modèles. *Les cahiers de l'analyse des données*, VI(2):135–143.

Escofier, B. and Pagès, J. (1994). Multiple factor analysis (afmult package). *Computational Statistics and Data Analysis*, 18.1994:121–140.

Eslami, A., Qannari, E., Kohler, A., and Bougeard, S. (2014). Multivariate analysis of multiblock and multigroup data. *Chemometrics and Intelligent Laboratory Systems*, 133:63–69.

Fisher, R.-A. (1936). The use of multiple measurements in taxonomics problems. *Annals of eugenics*, 7-(2):179–188.

Hotelling, H. (1936). Relations between two sets of variants. *Biometrika*, 28:321–337.

Israels, A. (1987). *Eigenvalues techniques for qualitative data*. DSWO Press.

Jollife, I.-T. (2002). *Principal Component Analysis*. Springer.

Kang, M., Kim, D.-C., Liu, C., and Gao, J. (2015). Multiblock discriminant analysis for integrative genomic study. *Biomed Research International*, 2015:1–10.

Kateri, M. and Petros Dellaportas, P. (2012). Conditional symmetry models for three-way contingency tables. *Journal of Statistical Planning and Inference*, 142(8):2430 – 2439.

Kostov, B., Bécue-Bertaut, M., and Husson, F. (2013). Multiple factor analysis for contingency tables in the factominer package. *The R Journal*, 5:29–38.

Kroonenberg, P. and Lombardo, R. (1999). Nonsymmetric correspondence analysis: A tool for analysing contingency tables with a dependence structure. *Multivariate Behavioral Research*, 34(3):367–396.

Lavit, C. (1988). *Analyse conjointe de tableaux quantitatifs*. Masson, Paris.

Leclerc, A. (1975). L'analyse factorielle des correspondances sur juxtaposition de tableaux de contingence. *Revue de statistique appliquée*, 23-3:5–16.

Lombardo, R. (2011). Three-way association measure decompositions: The delta index. *Journal of Statistical Planning and Inference*, 141(5):1789 – 1799.

Sabatier, R., Vivien, M., and Reynès, C. (2013). Une nouvelle proposition, l'analyse discriminante multitableaux : Statis-lda. *Journal de la Société Française de Statistique*, 154:31–43.

Saporta, G. (1976). Liaison entre plusieurs ensembles de variables et codage de variables qualitatives. *Thèse, Université de Paris VI.*

Taneichi, N. Sekiya, Y. and Toyama, J. (2019). Transformed statistics for tests of conditional independence in j x k x l contingency tables. *Journal of Multivariate Analysis*, 171:193 – 208.

Tenenhaus, A. and Tenenhaus, M. (2014). Regularized generalizd canonical correlation analysis for multiblock and multigroup data analysis. *Journal of operational research*, 238:391–403.

Vallejo-Arnadela, A., Vincente-Villardon, J., and Gamindo-Villardon, M. (2007). Canonical-statis : Biplot analysis of multi-group structured data based on statis-act methodology. *Computational Statistics and Data Analysis*, 46:4193–4205.

Xiang, F., Nie, G., Meng, G., Pan, C., and Zhang, C. (2012). Discriminant least squares regression for multiclass

 ISSN: 2102-6238

classification and feature selection. *IEE transactions on neural networks and learning systems*, 23:1738–1754.

Zarraga, A. and Goitisolo, B. (2002). Méthode factorielle pour l'analyse simultanée de tableaux de contingence. *Revue de statistique appliquée*, 50(2):47–70.

Zarraga, A. and Goitisolo, B. (2003). Etude de la structure inter-tableaux a travers l'analyse simultanée. *Revue de statistique appliquée*, 51 (3):117–142.

Zarraga, A. and Goitisolo, B. (2009). Simultaneous analysis and multiple factor analysis for contingency tables: Two methods for the joint study of contingency tables. *Computational Statistics & Data Analysis*, 53:3171 – 3182.

 ISSN: 2102-6238