UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE MEDELLÍN
FACULTAD DE MINAS

DYNA
http://dyna.medellin.unal.edu.co/

Missing data in multiple correspondence analysis under the available data principle of the NIPALS algorithm

Andrés Felipe Ochoa-Muñoz [a], Víctor Manuel González-Rojas [a] & Campo Elías Pardo [b]

[a] *Escuela de Estadística, Facultad de Ingenierías, Universidad del Valle, Cali, Colombia. andres.ochoa@correounivalle.edu.co, victor.m.gonzalez@correounivalle.edu.co*
[b] *Departamento de Estadística, Facultad de Ciencias, Universidad Nacional, Bogotá, Colombia. cepardot@unal.edu.co*

Received: June 27th, 2019. Received in revised form: October 24th, 2019. Accepted: November 13th, 2019

Abstract
Multiple correspondence analysis (MCA) in the presence of missing data is usually performed by removing the records that have missing or not available (NA) data; sometimes, an entire row or column of a data matrix is removed, which is not ideal because relevant information on an individual or variable of the study is lost. In some cases, it is assumed that the missing data are a category of the qualitative variable, resulting in a greater variance dispersion in the new axes. Possible solutions to this problem can be the imputation of the missing data or using an algorithm suited to the presence of this type of data. This work is focused on performing the MCA method in the presence of missing data, without using imputation techniques, by using the available data principle of the nonlinear estimation by iterative partial least squares (NIPALS) algorithm [25].

Keywords: multiple correspondence analysis; missing data; NIPALS; available data principle.

Datos faltantes en análisis de correspondencias múltiples bajo el principio de datos disponibles del algoritmo NIPALS

Resumen
El Análisis de Correspondencias Múltiples (ACM) en presencia de datos faltantes usualmente se trabaja eliminando los registros en donde exista el dato faltante o no disponible (NA), algunas veces se elimina toda la fila o toda la columna de la matriz de datos, lo cual no es adecuado ya que al realizarlo se pierde información relevante sobre algún individuo o variable del estudio. En algunos otros casos, se asume que el dato faltante es una categoría de la variable cualitativa, trayendo como consecuencia mayor dispersión de varianza en los nuevos ejes. Una solución para esta situación puede ser la imputación del dato faltante o utilizar un algoritmo que permita trabajar con la presencia de éste tipo de datos. Este trabajo se centra en realizar el método ACM en presencia de datos faltantes sin acudir a técnicas de imputación, para esto se utiliza el principio de datos disponibles del algoritmo NIPALS [25].

Palabras clave: análisis de correspondencias múltiples; datos faltantes; NIPALS; principio de datos disponibles.

1. Introduction

Currently, when a phenomenon is studied, measurements of different variables are taken over many observation units, generating large volumes of data. Multivariate statistical methods are appropriate in these situations because they consider the existing relationships between variables [2]. In some circumstances, these variables are qualitative, and a method that is frequently used for extracting information from these types of variables is the multiple correspondence analysis (MCA) technique. However, this method only works with complete information, that is, it does not allow the presence of missing data.

MCA is widely used in the analysis of surveys with

How to cite: Ochoa Muñoz, A.F, González Rojas, V.M, and Pardo, C.E, Missing Data in Multiple Correspondence Analysis under the Available Data Principle of the NIPALS Algorithm. DYNA, 86(211), pp. 249-257, October - December, 2019.

© The author; licensee Universidad Nacional de Colombia.

questions that must be answered with only one of several options [13]. The answers to these types of questions generate qualitative variables (nominal or ordinal), each of which are associated with splitting the individuals (disjoint groups of individuals). When a question is not answered, nonresponse (NR) or missing (NA) data are generated.

A table, which is the object of MCA analysis, has the statistical units in rows and the qualitative variables in columns. Each statistical unit, called an "individual", assumes only one category of each variable. A table that is analyzed with MCA has as many columns as variables, which indicate the categories assumed by the individuals. Because this table does not have a numerical meaning, it is transformed into a table of individuals by categories, where each qualitative variable generates as many columns as it has categories. This table is called a complete disjunctive table (CDT) because for each row within the columns of each variable there is only a single value of one, which indicates the category assumed, and the remaining columns are zero (see Table 1). The theoretical approach of the MCA starts from the CDT.

The MCA is the correspondence analysis (CA) of the CDT, which has very unique properties that are lost in the presence of missing data. Van der Heijden and Escofier [23] compare several methods including missing passive, missing passive modified margin, missing single, and missing multiple.

The missing passive method is equivalent to performing a correspondence analysis of the incomplete disjunctive table (IDT), which is called that because in the case of a nonresponse for a variable, the row has zeros in all the columns of the categories of that variable [7,14]. The missing passive modified margin method [7] is proposed to recover most of the MCA properties.

The missing single method, one of the most used, consists of creating a category for each variable with missing data. This option is usually managed by the analyst, who recodes the data prior to introducing them into an MCA program [23].

Currently, there are other authors working with missing data using the nonlinear estimation by iterative partial least squares (NIPALS) algorithm in multivariate analysis [1,18,19,22], and others working on the data imputation approach with the expectation maximation (EM) algorithm [3,11,12]. It is not exactly known which approach generates better results; however, works have been found that compare them to principal component analysis (PCA) [24].

In the case of MCA with missing data, authors Josse et al. [12] have worked with the EM algorithm approach, experiencing difficulties in the imputation process of the complete disjunctive table; it assigns one to the higher-frequency categories and presents some convergence problems. However, there are no known works or ideas that attempt to work with MCA under NIPALS. For this reason, this research proposal will generate more knowledge on how to process missing data with MCA.

In this work, it is proposed to use the NIPALS algorithm by Wold et al. [25] to perform MCA with the available data, that is, without the imputation of the missing data. The proposed method, called multiple correspondence analysis under the available data principle (MCAadp), evaluates the influence of this type of data on the factorial axes, the descriptive power (percentage of applied variance), and the inertia generated in each component, among others. This procedure can be seen in more detail in [15].

The MCAadp method is illustrated with the *DogBreeds* database of the *FactoClass* library of the R software [4,17]. It begins with the complete database and missing data are randomly generated in different percentages, i.e., 5%, 10% up to 50%.

The following contains a summary of the MCA, NIPALS algorithm, iterative MCA (iMCA), and proposed MCAadp methods.

2. Methodologies

In this section, the MCA method and the NIPALS algorithm are theoretically presented. The NIPALS algorithm is used to work in the presence of missing data, using the available data principle. In addition, each method, their optimization processes, the matrix to diagonalize, the concept of inertia, the eigenvalues, eigenvectors, and additional concepts that are relevant to the multivariate analysis are explained. This section also refers to the existing relationships between methods, especially that the MCA is a PCA of a matrix transformed into weighed profiles [13,21]. In the last subsection of this section, the method for the imputation of the data based on the EM algorithm for the MCA is explained. The MCA is presented below.

2.1. Multiple Correspondence Analysis (MCA)

The principles of this method can be credited to Guttman [8], Burt [5], and Hayashi [9]. MCA is used in the analysis of tables of individuals described by qualitative variables and to study the associations between different categories of variables being studied [13,16]. MCA is a generalization of CA, defined as an CA of the complete disjunctive table Z, where the number one is assigned to the category assumed by the individual, and zero to the category that was not selected, as observed in Table 1.

F is obtained from matrix Z, with the following general term:

$$f_{ij} = z_{ij}/ns\ ,\ f_{i.} = \frac{1}{n}, and\ f_{.j} = z_{.j}/ns \tag{1}$$

Table 1:
Complete Disjunctive Table Z_{ij}

Individual	Z_a		Z_b			Z_c			Z_i
1	1	0	1	0	0	1	0	0	s
2	1	0	1	0	0	1	0	0	s
.	.	.	.	.	.	.	.	.	s
.	.	.	.	.	.	.	.	.	s
n	1	0	0	1	0	0	1	0	s
$Z_{.j}$	$Z_{.1}$	$Z_{.2}$	$Z_{.3}$	$Z_{.4}$	$Z_{.5}$	.	.	$Z_{.p}$	ns

Source: The Authors

where s is the number of qualitative variables and n is the number of individuals.

2.1.1. Maximization and matrix to diagonalize

The MCA geometric goal is to find a new system of orthogonal axes u_α, where the inertia I of the cloud of individuals is projected such that the first axes concentrate most of the inertia in decrescent order. In this manner, the factorial coordinates ψ_α are obtained; they are the projection of the individuals over the space generated by u_α, where $\alpha = 1,2,\dots,p-s$. It is important to mention that the diagonal matrices $M_n = [\ddots\ 1/f_{i.}\ \ddots]$ and $M_p = [\ddots\ 1/f_{.j}\ \ddots]$ correspond to the metrics associated with the individuals and the categories. The inertia associated with the space of individuals is $I = \psi' {M_n}^{-1}\psi$, with $\psi = M_n F M_P u$, that is, $I = u' M_p F' M_n F M_p u$, which is the amount to maximize under the constraint $u' M_p u = 1$.

The Lagrangian solution leads to the system of eigenvalues and eigenvectors $Su = \lambda u$, with $S = F' M_n F M_p$ and $u' M_p S u = I = \lambda$, which correspond to the highest eigenvalue.

Matrix S is not necessarily symmetric; therefore, it does not guarantee that the eigenvectors are orthonormal. Observe from the previous system that ${M_p}^{1/2} S u = \lambda {M_p}^{1/2} u$, as follows:

$$ {M_p}^{1/2} F' M_n F {M_p}^{1/2} {M_p}^{1/2} u = \lambda {M_p}^{1/2} u \tag{2} $$

Instead of diagonalizing S, matrix S^* is diagonalized, as follows:

$$ S^* = M_p^{\frac{1}{2}} F' M_n F M_p^{\frac{1}{2}} \ under\ w'w = 1 \tag{3} $$

The eigenvectors $w = {M_p}^{1/2} u$ are orthogonal and are associated with the λ-eigenvalues of S^*. Note the following:

$$ S^* = S_o' S_o \ con\ S_o = {M_n}^{1/2} F {M_p}^{1/2} \Rightarrow \tag{4} $$

$$ S^* w = \lambda w $$

In this way, the relationship MCA has with PCA is observed, where MCA is a PCA of the symmetric matrix S^* [13,21]. Similarly, matrix $T^* = S_o S_o'$ is diagonalized in space R^n.

This scheme is very important because it is also followed for the solution with missing data, using the available data principle.

2.1.2. Total inertia of the cloud of categories

The inertia of the cloud of categories p is as follows:

$$ I = \sum_{j=1}^{p} I_j = \sum_{q=1}^{s} I_q = \frac{p}{s} - 1 \tag{5} $$

where I_j is the inertia contribution of a category and I_q is the inertia associated with a variable, i.e., the inertia of the subcloud of its categories.

2.1.2.1. Inertia contribution of a category

To calculate the inertia contribution associated with a category j, its weight is considered, i.e., the marginal column $f_j = z_j/ns$ and its distance to the center of gravity. In this way, the inertia by modality I_j is as follows:

$$ I_j = f_{.j} d^2(j,G) = \frac{z_{.j}}{ns}\left(\frac{n}{Z_{.j}} - 1\right) = \frac{1}{s}\left(1 - \frac{z_{.j}}{n}\right) \tag{6} $$

The inertia contribution of a category is higher if there is low frequency in the data set.

2.1.2.2. Inertia by variable

The inertia due to a variable (subtable) q is an increasing function of its number of categories p_q. The inertia by variable I_q is calculated as follows:

$$ I_q = \sum_{j}^{p_q} I_j = \frac{1}{s}(p_q - 1) \tag{7} $$

In the presentation of the MCAadp method, how to calculate the inertia expressions in the presence of missing data will be emphasized.

2.2. Nonlinear Estimation by Iterative Partial Least Square (NIPALS)

The NIPALS was proposed by Wold and is the basis of the partial least squares (PLS) regression [20]. It essentially performs a decomposition of the data matrix into singular values by iterative sequences of orthogonal projections (geometric concept of regression) obtained as point products. When the database is complete, there is an equivalence with the PCA results, and it can also work with missing data and obtain estimations from the reconstituted data matrix.

For the data matrix $Z_{n,p}$ of range a, whose columns $Z_1,\dots,Z_p$ are assumed to be centered or standardized, the decomposition derived from the PCA allows the reconstitution by $Z = \sum_\alpha^p \psi_\alpha u_\alpha'$, where ψ_α is the α-th principal component and u_α is the eigenvector associated with axis α [1]. Then, it is possible to make the reconstitution by individuals or variables, where $Z_j = \sum_\alpha^p \psi_\alpha u_{\alpha j}$ for $j = 1,\dots,p$ and $Z_i = \sum_\alpha^p \psi_\alpha u_{\alpha i}$ for $i = 1,\dots,n$.

The algorithm begins by taking the first column of Z_0 as the first principal component ψ_1. Then, a series of deflated tables will be constructed, called $Z_\alpha = Z_0 - \psi_\alpha u_\alpha$, which allow the cycle to restart and the remaining components

(orthogonal) ψ_2, .., ψ_p, and their respective eigenvectors $u_1, \ldots, u_p$ to be obtained.

Shown in the following subsections is the pseudocode of the algorithm when the data matrix is complete. As observed in stage 2.2.1, $u_{\alpha j}$ represents, prior to the normalization, the coefficient (slope) of the regression of $Z_{\alpha-1,j}$ over component ψ_α.

2.2.1. NIPALS algorithm pseudocode

Stage 1: $Z_o = Z_h$
Stage 2: $\alpha = 1,2, \ldots, p$
Stage 2.1: ψ_α = 1st first column of $Z_{\alpha-1}$
Stage 2.2: Repeat until convergence of u_α
Stage 2.2.1: $u_\alpha = \frac{Z'_{\alpha-1}\psi_\alpha}{\psi'_\alpha\psi_\alpha}$
Stage 2.2.2: Normalize U_α to 1
Stage 2.2.3: $\psi = \frac{Z_{\alpha-1}u_\alpha}{u'_\alpha u_\alpha}$
Stage 2.3: $Z_\alpha = Z_{\alpha-1} - \psi_\alpha\, u_\alpha{}'$ (ensures orthogonality)
Next, α

2.2.2. Available data principle

This principle refers to some operations between vectors, omitting the missing data and working with the available matched points; that is, if there are two vectors with NA, $\langle x, y\rangle$ can be found using the available data principle [15].

$$X = \begin{pmatrix} x_1 \\ NA \\ x_3 \\ \vdots \\ x_n \end{pmatrix} \quad Y = \begin{pmatrix} NA \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} \tag{8}$$

Then:
$\langle x, y\rangle = \sum_{xy\ existent} x_i\, y_i = x_3y_3 + x_4y_4 + \cdots + x_ny_n$
Note that the same result is obtained if the NA are replaced with zeros.

2.2.3. NIPALS missing data algorithm pseudocode

Stage 1: $Z_0 = Z_h$
Stage 2: $\alpha = 1,2, \ldots, \alpha$
Stage 2.1: $\psi_\alpha = 1^{st}$ first column of $Z_{\alpha\text{-}1}$
Stage 2.2: Repeat until convergence of u_α
Stage 2.2.1: For j=1,2,...,p

$$u_{\alpha j} = \frac{\sum_{i:Z_{ji}et\psi_{\alpha i}existent} Z_{\alpha-1,ji}\ \psi_{\alpha i}}{\sum_{i:Z_{ji}et\psi_{\alpha i}existent} \psi^2_{\alpha i}} \tag{9}$$

Stage 2.2.2: Normalize u_α to 1
Stage 2.2.3: For i = 1,2,...,n

$$\psi_{\alpha i} = \frac{\sum_{i:Z_{ji}existent} Z_{\alpha-1,ji}\, U_{\alpha j}}{\sum_{i:Z_{ji}existent} u^2_{\alpha j}} \tag{10}$$

Stage 2.3: $Z_\alpha = Z_{\alpha-1} - \psi_\alpha u'_\alpha$

The main characteristic of NIPALS is that it works with a series of point products as a sum of products of the matched elements. This allows it to work with missing data by adding the available data in each operation. Geometrically, the procedure considers the omitted elements falling over the regression straight line; they are not leverage points [20].

The pseudocode of the NIPALS algorithm with missing data contains stages 2.2.1 and 2.2.3, where the slopes of the lines of the least squares from the origin of the point cloud over the available data are calculated. $u_{\alpha j}$ and $\psi_{\alpha i}$ must capture, in their positions j and i, the missing data characteristic given by Z_{ij} [1].

2.3. Iterative MCA for missing data (MCA-EM)

The iterative MCA via EM (iMCA or EM-MCA) was proposed by Josse [11]. This method is based on the EM-PCA, where the missing data are estimated by average values, and then the distances between the original data S_o and the estimated data ψu are minimized, such that the iMCA uses the following loss function:

$$\ell = \| w(S_0 - \psi u') \|^2 = \sum_{i=1}^{n}\sum_{j=1}^{p} w_{ij}(S_{0_{ij}} - \psi_{i\alpha}u'_{\alpha j})^2 \tag{11}$$

where $S_0 = M_n^{1/2} F M_p^{1/2}$, $Z_{n.p}$ is the complete disjunctive table, $\psi_{n.q}$ is the factorial coordinates, $u_{\ p.q}$ is the eigenvector in $R^p, \alpha = 1, 2, \ldots, q\ (q < p - s)$, and w is an indicator variable (0 = NA; 1 = observed value). As in EM-PCA, this method minimizes the loss function associated with the complete data [12]. Presented in the following subsection is the pseudocode associated with the iMCA method.

2.3.1. iMCA algorithm pseudocode

1. Initiation L = 0: Z^0

The missing data are replaced by the proportion of ones in the complete disjunctive table Z_{ij}. The replacement of the missing data must add one per variable, which makes the marginal per row equal to s, as in the complete data.

Example: A = 0.4, *B* = 0.3, *C* = 0.3

2. Step L

2.1 Perform a singular decomposition of matrix $S_0 = Mn^{1/2}FMp^{1/2}$ (ψ and u are obtained here).

2.2 Perform the reconstitution of matrix $\hat{S}_o = \widehat{\psi u'}$, using the q dimensions ($q < p - s$) found by generalized cross-validation.

2.3 Perform the reconstitution of the complete disjunctive table $\hat{Z} = M_n^{-1/2}(\widehat{S_o} * ns)M_p^{-1/2}$. Here, the missing values are imputed with the reconstitution and the observed values of Z are the same.

Steps 2.1, 2.2, and 2.3 are repeated until convergence, where one is assigned to the highest-frequency category in $\hat{Z}$

Table 2:
Example of a complete disjunctive table $Z^*_{n.p}$ with NA

Individual	Gender		Religion			Race		
1	1	0	1	0	0	0	0	1
2	1	0	NA	NA	NA	1	0	0
3	NA	NA	0	1	0	NA	NA	NA
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	1	0	0	1	0	0	1	0

Source: The Authors

and zero is assigned to the remaining categories.

3. MCAadp: Multiple Correspondence Analysis under the available data principle

The MCAadp method in the presence of missing data is presented in this section. Methods to obtain the eigenvalues and eigenvectors in spaces R^n and R^p are shown. From these results, the transition relations are considered for finding the components in each space. In addition, the expressions of Total Inertia, Inertia by question, and Inertia by category are presented. With this method, the proposal by Wold [25] and the multiple correspondence analysis are adapted to work with missing data [15]. The MCAadp method is presented below.

3.1. Presentation of the MCAadp Method

To perform MCAadp, first, a disjunctive table with missing data $Z^*_{n.p}$ is constructed, as presented in Table 2.

$$\text{Zij} = \begin{cases} 1 \text{ if row 1 assumes category j} \\ 0 \text{ otherwise} \\ \text{NA if the cell has missing data} \end{cases}$$

Second, the relative frequency matrix $F^*_{n.p} = Z^*_{n.p}/k^*$ is calculated. In this process, it is important to consider that when building F^*, ${Z^*}/{k^*}$ is performed, where $k^* = ns^* = \sum Z^*_i = \sum Z^*_{.j}$ ($s^* = \sum Z^*_i/n$); it is obtained by adding by rows or columns to obtain the corresponding marginals, this time with the available data.

Note that if the NA are replaced with zeros, the same result is obtained in the marginals and in the sum of the table.

3.2. MCAadp in space R^p

Based on the relationship between MCA and PCA, the matrix to diagonalize is as follows:

$$S^* = {M_p^*}^{1/2} F^{*\prime} M_n^* F^* {M_p^*}^{1/2} \text{ with the constraint}$$
$$u' {M_p'^*}^{1/2} {M_p^*}^{1/2} u = 1 \quad (12)$$
$$w = {M_p^*}^{1/2} u;\ w'w = 1$$

In the diagonalization process of matrix S^* is the following system of eigenvalues and eigenvectors:

$$S^* w = \lambda w \quad (13)$$

Then, S^* contains the submatrices $S^*_o = M_n^{*1/2} F^* M_p^{*1/2}$, such that $S^* = S_o'^* S_o^*$. It is important to mention that M_n^* and M_p^* are obtained with the available data and correspond to the metric matrices for the row and columns, respectively. These are diagonal matrices that contain said weights in their diagonal. In detail, the matrices have the following structure:

$$M_p^* = [\ddots\ 1/f_{.j}^*\ \ddots]; M_n^* = [\ddots\ 1/f_{i.}^*\ \ddots] \quad (14)$$

Thus, matrix $S^*_o = M_n^{*1/2} F^* M_p^{*1/2}$ is obtained, which does not contain missing data, since the available data principle is considered when performing the point products.

Finally, for matrix S^*_o, a decomposition into singular values is executed iteratively, as performed by the NIPALS algorithm.

3.3. MCAadp pseudocode

1. The disjunctive table is constructed with NA ($Z_{ij}{}^*$)
2. $F^* = \frac{z^*}{k^*}$ is constructed ($k^* = ns^*$; available. $s^* = \frac{\sum z_i^*}{n}$)
3. Matrix $S_0{}^*$ is constructed using the available data principle as follows:

$$S_0^* = {M_n^*}^{1/2} F^* {M_p^*}^{1/2} \quad (15)$$

where $M_p^* = [\ddots\ 1/f_{i.}^*\ \ddots]$ *and* $M_n^* = [\ddots\ 1/f_{i.}^*\ \ddots]$.

4. An NIPALS (nonstandardized) is applied to matrix $S_0{}^*$.

3.4. MCAadp in space R^n

In section 3.2, the method was presented in the space associated with the variable R^p; this same scheme can be identified in the cloud of individuals where matrix $T^*_{n.n}$ is constructed, which is the matrix to diagonalize in this space. The construction of matrix T^* is performed considering the available data principle, given that $F^*_{n.p}$ contains NA records.

If T is diogonalized, then we have the following system of eigenvalues λ and eigenvectors v:

$$Tv = \lambda v \quad (16)$$

Because T is not necessarily symmetric, it does not have orthonormal eigenvectors. The following transformation is performed:

$$r = {M_n^*}^{1/2} v, \text{ such that } r'r = 1. \quad (17)$$

Then:

$$M_n^{*1/2} T^* v = \lambda M_n^{*1/2} v \quad (18)$$

$$M_n^{*1/2} F^* M_p^* F^{*\prime} M_n^{*1/2} M_n^{*1/2} v = \lambda M_n^{*1/2} v \quad (19)$$

$$T^* r = \lambda r;\ r'r = 1 \quad (20)$$

It is important to mention that matrix $T^*_{n.n}$ is constructed considering the available data principle and that from this matrix, the eigenvalues λ and eigenvectors r are found.

A more important situation in this procedure is that the eigenvalues λ in spaces R^p are R^n are equivalent, which makes the transition relations valid and provides coordinates ψ and φ.

3.5. Transition relations

As mentioned above, the MCAadp method guarantees that the eigenvalues in spaces R^n and R^p are equivalent; in this way, we can relate the coordinates of one space with the coordinates of another, considering the following expressions:

$$\varphi_\alpha = \frac{1}{\sqrt{\lambda}} M_P^* S_0^{*\prime} \psi_\alpha;\ \psi_\alpha = \frac{1}{S^* \sqrt{\lambda_\alpha}} S_0^* \varphi_\alpha \tag{21}$$

3.6. Components in R^n and R^p

To perform the calculations of components $\psi_{n.p}$ in R^p, the point product of matrix S_o^* with the eigenvector associated with the space of variables $w = M_p^{*1/2} u$ is performed. To calculate components $\varphi_{p.p}$ in R^n, the point product of matrix T_o^* with the eigenvector associated with the space of individuals $r = M_n^{*1/2} v$ is performed. Based on these calculations, the following expressions are obtained:

$$\psi = M_n^{*1/2} S_0^* w = M_n^* F^* M_p^* u \tag{22}$$

$$\varphi = M_p^{*1/2} T_0^* r = M_p^* F^{*\prime} M_n^* v;\ \text{where } T_0^* = S_0^{*\prime} \tag{23}$$

3.7. Inertia expressions for available data

Presented in this section are the expressions of Total Inertia, Inertia by category, and Inertia by question, such that these expressions consider that the available data principle s^* is used; it is the marginal estimated by row and is replaced in each Inertia expression. The new expressions are presented as follows:

Total Inertia: This inertia depends on the existing number of data with NA. If there are more, then s^* is smaller and thus the Total Inertia increases.

$$I^* = \frac{p}{s^*} - 1;\ \text{where } s^* = \frac{\sum z_i^*}{n} \tag{24}$$

Inertia by category: (25)

$$I_j^* = \frac{1}{s^*}\left(1 - \frac{Z_{.j}^*}{n}\right)$$

Inertia by variable: (26)

$$I_q^* = \frac{1}{s^*}\left(p_q - \left(\frac{Z_{.j}^*}{n}\right)\right)$$

where p_q is the number of categories per question q.

4. Equivalence between the results of the MCAadp and the MCA of incomplete tables

In the MCAadp proposal, the sums of the rows of the disjunctive table with missing data are replaced by the constant $s^* = \frac{\sum z_i^*}{n}$ to calculate the inertias. In addition, the sums and point products of the disjunctive table with missing data are equivalent to the same operations with the incomplete disjunctive table (i.e., with 0 instead of NA). Thus, the MCAadp coincides with the MCA method for an incomplete disjunctive table [7].

In these methods, the MCA properties are retained. The advantage of the MCAadp is derived from the NIPALS algorithm, which sequentially obtains the factors, leading to shorter calculation times if only the axes that will be analyzed are obtained.

5. Application

The DogBreeds database contains 27 breeds and six qualitative variables: Size (SIZ), Weight (WEI), Speed (SPE), Intelligence (INT), Affection (AFF), and Aggressiveness (AGR). Each variable has two or three categories, as illustrated in the FactoClass package [17]**.** SIZ has three categories: big, medium (med), and small. WEI has three categories: heavy, medium (med), and light. VEL has three categories: fast, medium (med), and slow. Variable INT has three categories: high, medium (med), and low. AFF and AGR each have two categories: high and low. It is important to mention that these qualitative variables are used as indicator variables. In this manner, the study is performed in a data matrix $\boldsymbol{Z_{ij}}$, which contains one, zero, or NA, depending on if the category is present or absent or if there are missing data. It is important to mention that matrix $\boldsymbol{Z_{ij}}$ is of dimension $\boldsymbol{n * p}$, where n represents individuals and p represents categories.

5.1. Missing data simulation in the study case

The work is performed with the first six variables and missing data will be randomly assigned to the matrix. In the following structure of the R software, m corresponds to the position where the missing data is located and a corresponds to the percentg of missing data values of the total

```
fmd <- function(Xo,a)
{
    X. <- as.matrix(Xo)
    n <- nrow(X.); p <- ncol(X.); N <- n*p
    m <- sample(N, round(a*N,0)) ; d <- length(m)
for(j in 1:d){
    X.[m[j]] <- NA
}
return(X.)
}
```

5.2. Simulation study

Table 3 shows the proposed simulation scenarios, which have a missing completely at random (MCAR) mechanism. At the same time, the scenarios when the entire data matrix has 1, 2, or 3 NA per row are considered. In the first three scenarios, the marginal per row $\boldsymbol{Z_i}$ is constant for every $\boldsymbol{i}$. However, when there are from 0 to 1, 0 to 2, and 0 to 3 NA per row, the marginal per Z_i is not constant for every $\boldsymbol{i}$. It is important to mention that the marginal per column is not constant for the $\boldsymbol{j}$ categories.

It is also important to mention that the NA percentage is calculated based on the total number of records ($27 * 6$). In this article, the work is performed with a maximum of 50% of missing data; the NA percentage in the last three scenarios is randomly generated. To study the presence of NA in a more general and random manner, different simulations were performed, as shown in Table 4, where the records per individual will have a maximum of three NA per row (not exceeding 50%). One thousand matrices are generated for each NA percentage: 5%, 10%... 50%, i.e., there will be 7000 simulated matrices and they are compared with the complete data case.

5.3. Statistical analysis of the simulation scenarios

First, the matrix with complete data must be analyzed to see how each of the following indicators behave:

- Eigenvalues λ and eigenvectors u
- Components ψ and φ in R^n and R^p
- Total Inertia, Inertia by category, and Inertia by question
- Descriptive power $(\lambda_1 + \lambda_2)/\Sigma\lambda$
- Factorial planes
- Orthogonality in the components and orthonormality in the eigenvectors

With this starting point, the same indicators are analyzed for each of the scenarios proposed in Table 3. In each of these analyses, it is identified if the inertia expressions agree with the theory with complete data. Then, in section 6, a comparison is performed between the MCAadp and the imputation method, where the scheme of Table 4 is used and where m is equal to the number of matrices to simulate the structure.

Table 3:
Simulation scenarios.

NA structure	Number of NA	% NA
MCAR	*1 NA per row*	*16.70%*
MCAR	*2 NA per row*	*33.30%*
MCAR	*3 NA per row*	*50.00%*
MCAR	*0:1 NA per row*	*9.26%*
MCAR	*0:2 NA per row*	*13.58%*
MCAR	*0:3 NA per row*	*27.16%*

Source: The Authors

Table 4:
Simulation scenarios for the descriptive power analysis

NA structure	Methods	% NA	m
0:3 NA per row	MCAadp iMCA	5%	1000
		10%	1000
		15%	1000
		20%	1000
		25%	1000
		30%	1000
		50%	1000

Source: The Authors

Table 5:
Results of inertia and descriptive power

	Number of NA per row	$\sum\lambda_\alpha$	Total Inertia	Descriptive power
Method	Complete data	1.667	1.667	0.519
MCAadp	1 NA	2.200	2.200	0.409
	2 NA	3.000	3.000	0.347
	3 NA	4.333	4.333	0.286
	0:1 NA	1.936	1.938	0.476
	0:2 NA	2.065	2.086	0.437
	0:3 NA	2.670	2.661	0.375
iMCA	1 NA	1.667	1.667	0.538
	2 NA	1.667	1.667	0.607
	3 NA	1.667	1.667	0.668
	0:1 NA	1.667	1.667	0.541
	0:2 NA	1.667	1.667	0.501
	0:3 NA	1.667	1.667	0.539

Source: The Authors

The code developed in the R software using the MCAadp method can be viewed at the following website:

https://github.com/AndresOchoaRSA/MCAapd

To perform the factorial planes, the *s.label()* function of the *ade4* library was used [6]. For the analysis with the imputation method, the *missMDA* library was used [10]. The R software version used is 3.6.1.

6. Results

Presented in Table 5 is a comparison of the previous scenarios with the imputation method proposed by Josse, J. et al. [12]. It is observed that the Total Inertia with the MCAadp is higher compared to the complete data. With the imputation method (iMCA), the Total Inertia is the same as in the complete data. This occurs because the imputation delivers an imputed complete matrix, where the inertia expressions are the same. However, it is observed that the descriptive power in the MCAadp method decreases, whereas with the imputation method it increases as a function of the number of NA. This may be considered a drawback of the imputation method, since an increase in descriptive power implies that having more NA records would be a favorable situation; however, it is expected that by having more NA in a data matrix, the representation performed will lose its descriptive power.

Figure 1 illustrates four first factorial planes of these analyses: MCA with complete data and MCAadp with 10%, 20% and 30% of NA. It is observed how as the missing data increase, some typologies and features that were found in the complete data are lost. However, Figure 2 shows the comparison between the complete data and matrices with missing data where the iMCA was used. It is observed that some typologies are also lost when the number of missing

data points increases. This type of analysis with the factorial planes becomes more difficult because it is a visual analysis and there should be an indicator describing the characteristics of the variables and individuals in these two axes. The proposed indicator is the descriptive power $(\lambda_1 + \lambda_2)/\sum\lambda$; with such an indicator, the percentage of variance explained in those two axes is found.

For this simulation case, Figure 3 shows that as the amount of missing data increases, the descriptive power decreases with the MCAadp method. Conversely, with the iterative MCA, as the missing data percentage increases, the descriptive power decreases. The previous situation is considered to be inconsistent because when there is a larger amount of missing data, the inertia relationships present in the data set should be more difficult to explain. It is important to mention that the line of reference corresponds to the descriptive power with complete data, which is 0.5198.

Note that for the case of MCA with complete data, the marginal per row is equal to s and in the case of MCAadp it is s^*, where $s > s^* \rightarrow I < I^*$, i.e., the inertia I^* in the MCAadp is higher than in MCA; however, as observed in Figure 3, the inertia in the first factorial plane decreases as a function of the NA percentage.

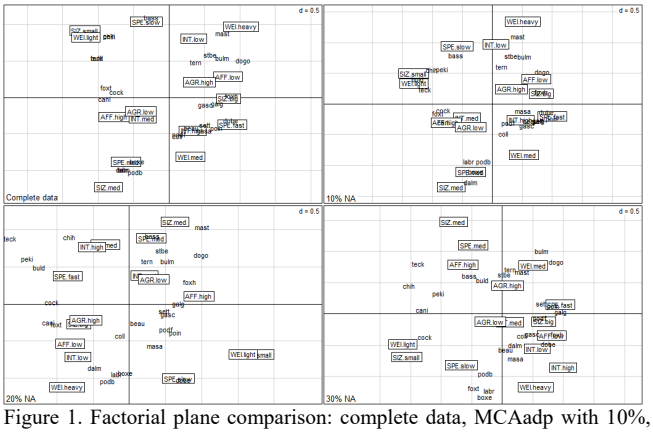

Figure 1. Factorial plane comparison: complete data, MCAadp with 10%, 20%, and 30% of NA.
Source: The Authors

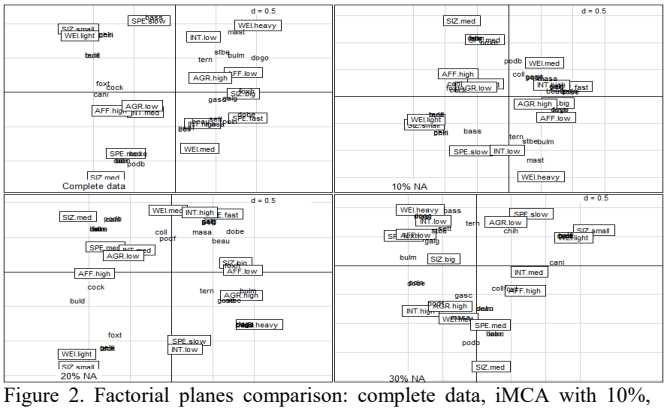

Figure 2. Factorial planes comparison: complete data, iMCA with 10%, 20%, and 30% of NA.
Source: The Authors

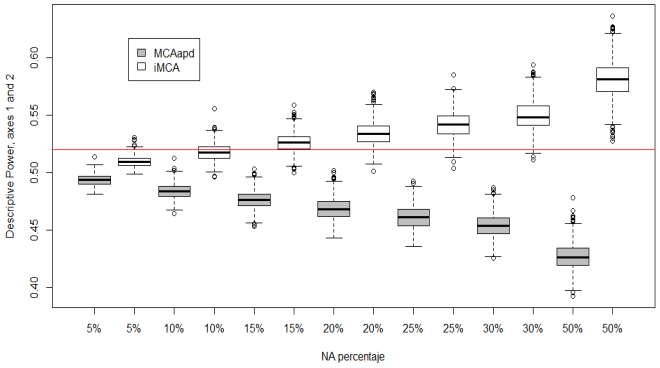

Figure 3. Behavior of the descriptive power as a function of the NA percentage with MCAadp and iterative MCA (iMCA)
Source: The Authors

7. Conclusions

Based on the results, the MCAadp presents a practical and efficient solution because its programming is simple and it has the interesting properties of orthogonality in the components, orthonormality in the eigenvectors, and equivalence in the eigenvalues in R^p and R^n, among others [15].

The MCAadp is an alternative solution to the missing data problem. It resorts to imputation techniques, but the user can use the method to perform imputation via the reconstitution of the matrix. In comparison with the iterative MCA method, the MCAadp presents higher consistency in terms of descriptive power, because at a higher number of NA, the descriptive power is expected to decrease. However, it is important to consider comparisons with other methods such as the regularized iterative MCA method [11] or the MCA with multiple imputations [3].

In a Masters thesis [15], work was also done with a higher-dimension data set (tea consumption data), and the same results were found regarding the simulation process, i.e., the descriptive power decreases as the missing data percentage increases when using MCAadp. For future works, it would be interesting to perform a cluster analysis and a multiple factorial analysis for qualitative variables, both with missing data via PLS, and adapt them to current libraries of the R software.

References.

[1] Aluja, T. and González, V.M., Gnm-nipals: general nonmetric-nonlinear estimation by iterative partial least squares, Revista de Matemática Teoría y Aplicaciones, 21(1), pp. 85-106, 2014.

[2] Aluja, T. and Morineau, A., Aprender de los datos: el análisis de los componentes principales: una aproximación desde el data mining, EUB, Barcelona, España, 1999. ISBN: 84-8312-022-4

[3] Audigier, V., Husson, F. and Josse, J., Mimca: multiple imputation for categorical variables with multiple correspondence analysis, pp. 1-30, 2015. arXiv preprint, arXiv:1505.08116v

[4] Brefort, A., Letude des races canines a partir de leurs caracteristiques qualitatives, Groupe HEC - Jouy en Josas, 1982.

[5] Burt, C., The factorial analysis of qualitative data, ritish, Journal of Statistical Psychology, 3(3), pp. 166-185, 1950.

[6] Dray, S., Dufour, A.B. et al., The ade4 package: implementing the duality diagram for ecologists, Journal of Statistical Software, 22(4), pp. 1-20, 2007. DOI: 10.18637/jss.v022.i04
[7] Escofier, B., Traitement es questionnaires avec non réponse, analyse des correspondances avec marge modifiée et analyse multicanonique avec contrainte, PhD Thesis, INRIA, 1981.
[8] Guttman, L., The quantification of a class of attributes: a theory and method of scale construction, The prediction of personal adjustment, 1941, pp. 319-348.
[9] Hayashi, C., Theory and examples of quantification.(II), in: Proc. of the Institute of Statist. Math, Vol. 4, pp. 19-30, 1956.
[10] Josse, J. and Husson, F., missMDA: a package for handling missing values in multivariate data analysis. Journal of Statistical Software, 70(1), pp. 1-31, 2016. DOI: 10.18637/jss.v070.i01
[11] Josse, J., Chavent, M., Liquet, B. and Husson, F., Handling missing values with regularized iterative multiple correspondence analysis, Journal of Classification, 29(1), pp. 91-116, 2012. DOI: 10.1007/s00357-012-9097-0
[12] Josse, J. and Husson, F., Handling missing values in exploratory multivariate data analysis methods, Journal de la Société Française de Statistique, [online]. 153(2), pp. 79-99, 2012. Available at: https://hal.archives-ouvertes.fr/hal-00811888
[13] Lebart, L., Morineau, A. and Piron, M., Statistique exploratoire multidimensionnelle, Dunod, Paris, 1997.
[14] Meulman, J., Homogeneity analysis of incomplete data, Vol. 1, DSWO Press, 1982.
[15] Ochoa-Muñoz, A.F. y González-Rojas, V.M., Análisis de correspondencias múltiples en presencia de datos faltantes: el principio de datos disponibles del algoritmo NIPALS (ACMpdd). MSc. Tesis, Universidad del Valle, Cali, Colombia, 2018.
[16] Pardo, C.E. and Cabardas, G., Métodos estadísticos multivariados en investigación social, en: cursillo del Simposio de Estadística, Santa Martha Colombia, 2001.
[17] Pardo, C.E., and Del Campo, P.C., Combinación de métodos factoriales y de análisis de conglomerados en R: el paquete FactoClass. Revista Colombiana de Estadística, 30(2), pp. 231-245, 2007. DOI: 10.15446/rce
[18] Russolillo, G., Partial least squares methods for non-metric data, PhD. Thesis, Università degli Studi di Napoli Federico II, Napoli, Italy, 2009.
[19] Sanchez, G., PLS path modeling with R. Trowchez Editions. Berkeley, USA, [online]. 2013.Available at: http://www.gastonsanchez.com/PLS _Path_Modeling_with_R.pdf
[20] Tenenhaus, M., La régression PLS, théorie et pratique, Editions Technip, 1998.
[21] Trejos, J., Castillo, W. and González, J., Análisis Multivariado de Datos. Métodos y Aplicaciones. Editorial UCR, 2014.
[22] Trinchera, L., Squillacciotti, S. and Esposito Vinzi, V., Pls typological path modeling: a model-based approach to classification, Proceedings of KNEMO, 2006, 87 P.
[23] Van der Heijden, P.G.M. and Escofier, B., Multiple correspondence analysis with missing data. In: Recherches sur lAnalyse des Correspondances, 2003, pp. 152-170.
[24] Vitelleschi, M. y Quaglino, B., Modelos PCA a partir de conjuntos de datos con información faltante, MSc. Tesis, Facultad de Ciencias Exactas y Naturales, Universidad Nacional de Catamarca, Argentina, 2009.
[25] Wold, H. et al., Estimation of principal components and related models by iterative least squares, Multivariate Analysis, 1, pp. 391-420, 1966.

A.F. Ochoa-Muñoz, is BSc. in Statistician in 2014, and a MSc. in Statistics in 2018, all of them from the University of Valle, Colombia. He works as a statistics professor, and his topics of interest are statistical modeling and multivariate analysis.
ORCID: 0000-0002-0003-1347

V.M. Gonzalez-Rojas, is BSc. in Statistician in 1990, MSc. in 2003, and Dr. in Statistics. He obtained his Dr. degree in 2014 from the Polytechnic University of Catalonia, Spain and has been working as an associate professor of the School of Statistics at the University of Valle, Colombia for nearly 18 years. His main topics of interest are Multivariate Data Analysis and Time Series.
ORCID: 0000-0002-6526-7879

C.E.Pardo-Turriago, is BSc. in Chemical Engineer in 1981, MSc. in Science – Statistics in 1992, and Dr. in Science – Statistics in 2012, all of them from the Universidad Nacional de Colombia. He has worked as a statistics professor at the Universidad Nacional de Colombia, Bogotá, in the field of multivariate analysis since 1994.
ORCID: 0000-0001-6464-1905

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE MEDELLÍN
FACULTAD DE MINAS

Área Curricular de Ingeniería Administrativa e Ingeniería Industrial

Oferta de Posgrados

Doctorado en Ingeniería - Industria y Organizaciones
Maestría en Ingeniería Administrativa
Maestría en Ingeniería Industrial
Especialización en Gestión Empresarial
Especialización en Ingeniería Financiera

Mayor información:

E-mail: acia_med@unal.edu.co
Teléfono: (57-4) 425 52 02